2016 ICSA APPLIED STATISTICS SYMPOSIUM

Atlanta, GA
June 12-15, 2016

Published by: International Chinese Statistical Association
Photographer for the front cover: Diana Li

International Chinese Statistical Association

Applied Statistics Symposium

2016

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 12 - 15, 2016

Hyatt Regency Atlanta

Atlanta, Georgia, USA

Organized by

International Chinese Statistical Association

Contents

Welcome

2016 ICSA Applied Statistics Symposium
June 12-15, Hyatt Regency Atlanta, Atlanta, Georgia, USA

Welcome to the 2016 International Chinese Statistical Association (ICSA) Applied Statistical Symposium!

This year we gather here for the 25th ICSA annual symposium. The theme of this symposium is the **Challenge of Big Data and Applications of Statistics**, which is in recognition of the advent of big data era. The executive and organizing committees have been working diligently to put together a strong and comprehensive program including nine short courses, three keynote lectures, 134 scientific sessions, one student paper session and exciting social events. Our scientific program includes keynote lectures from renowned statisticians Dr. Bin Yu (University of California, Berkeley), Dr. David Madigan (Columbia University) and Dr. Paul Albert (National Institute of Health), and invited, topic contributed, contributed talks reflecting recent advances and challenges in statistics, business statistics and biostatistics, which are closely related to Big Data analysis. As suggested by the 2015 ICSA President Dr. Wei Shen, we are delighted to inform that during the conference we will also offer a free career service for graduate students.

With your full support, this symposium attracts more than 650 statisticians working in academia, government, and industry from all over the world. We hope that the symposium offers great opportunities for learning, networking and recruiting, and that you will receive inspirations from old research ideas and develop new ones. Social events in this 2016 ICSA Symposium include mixer (Sunday, June 12 evening), banquet (Tuesday, June 14 – banquet speaker will be Dr. Michael Eriksen) and a trip to the Stone Mountain (Wednesday, June 15th, 1-6pm). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

The city of Atlanta enjoys a mild climate throughout the year, and is accessible from most cities across the North America. Downtown Atlanta provides numerous opportunities for dining, shopping and lodging, etc. In addition, Atlanta offers world-class attractions, including the World of Coca Cola, the CNN Center, and the Centennial Olympic Park that are within walking distance from our event hotel Hyatt Regency Atlanta. It is our sincere hope you have great opportunities to experience the wonderful activities during your stay in Atlanta.

Thanks for coming to the 2016 ICSA Applied Statistics Symposium in Atlanta!

Yichuan Zhao, on behalf of
2016 Applied Statistics Symposium Executive and Organizing committees



ICSA OFFICERS and COMMITTEES (January-December 2016)

EXECUTIVES

President: Mei-Ling Ting Lee
Past President: Wei Shen
President-elect: T. Tony Cai
Executive Director (2014-2016): Zhezhen Jin
Treasurer (2016-2018): Hongliang Shi

BOARD of DIRECTORS

Cong Chen (2014-2016)
Zhen Chen (2014-2016)
Chuhsing Kate Hsiao (2014-2016)
Bingyi Jing (2014-2016)
Mengling Liu (2014-2016)
Ying Zhang (2014-2016)
Shuangge Ma (2015- 2017)
Dejun Tang (2015-2017)
Lilly Yue (2015-2017)
Ying Yuan (2015-2017)
Tian Zheng (2015-2017)
Alan Chiang (2016-2018)
Yanyuan Ma (2016-2018)
Lu Tian (2016-2018)
Bo Yang (2016-2018)
Hao Helen Zhang (2016-2018)
Sheng Luo (2016, Biometrics Section Representative)

STANDING COMMITTEES

Program Committee: Gang Li (Chair, 2016), Naitee Ting (2014-2016), Dejun Tang (2014-2016), Faming Liang (2014-2016), Yichuan Zhao (2015-2017), Ying Zhang (2015-2017), Ming-Hui Chen (2015-2017), Yi Li (2016-2018), Lanju Zhang (2016-2018), Hongmei Jiang (2016- 2018).

Finance Committee: Hongliang Shi (Chair, 2016-2018), Linda Yau (2016-2018), Rochelle Fu (2015-2017) , Zhezhen Jin (Ex-Officio 2014-2016).

Nomination and Election Committee: Mei-Cheng Wang (Chair, 2016), Joanna Shih (2016-2018), Tsai-Hung Fan (2016-2018), Bing-Shun Wang (2014-2016), Cong Chen (2015-2016), Ying Wei (2015-2016), Haoda Fu (2015-2017), Yuan Ji (2015-2017), Li-Shan Huang (2015-2017), Xuewen Lu (2015-2017)

Publication Committee: Yi-Ching Yao (Chair, 2016), Chunming Zhang (2015-2017), Kelly Zou (2015-2017), Xin He (Editor of Bulletin), Mei-Cheng Wang (Co-Editor of SIB), Hongyu Zhao (Co-Editor of SIB), Hsin-Cheng Huang (Co-Editor of S. Sinica), Ruey Tsay (Co-Editor of S. Sinica), Zhiliang Ying (Co-Editor of S. Sinica), Zhezhen Jin (Ex-Officio), Jiahua Chen (Co-Editor of ICSA book series), Din Chen (Co-Editor of ICSA book series).

Membership Committee: Tony Jianguo Sun (Chair, 2016), Jiajuan Liang (2016-2018), Samuel Wu (2016-2018), Bo Fu (2016-2018), Caixia Li (2014-2016), Jianxin Pan (2014-2016), Zhigen Zhao (2015-2017), Lanju Zhang (2015-2017).

Awards Committee: Heping Zhang (Chair, 2014-2016), Mei-Chiung Shih (2014-2016), Shu-Yen Ho (2015-2017), Aiyi Liu (2015- 2017), Ming Yen Cheng (2016-2018), Tsung-Chi Cheng (2016-2018).

IT Committee: Jun Yan (Chair, 2016), Chengsheng Jiang, Lixin (Simon) Gao, Don Sun, Ruth Whitworth, Hongtu Zhu, Zhezhen Jin (Ex-Officio).

Archive Committee: Lili Yu (Chair, 2016), Smiley Cheng, Shein-Chung Chow, Nancy Lo

Lingzi Lu Award Committee (ASA/ICSA): Ivan Chan (Chair, 2014-2016, ICSA), Gang Li (2014-2016, ICSA), Haonan Wang (2016-2017), Eric Kolaczyk (2015, ASA), Victoria Romberg (2014-2016, ASA).

ICSA Representative to JSM Program Committee: Yi Li (2017), Ying Zhang (2016)

AD HOC COMMITTEES

2016 Applied Statistics Symposium Committee Yichuan Zhao (Chair)

2016 JSM Local Committee Min Yang (Chair)

Investment Ad Hoc Committee Linda Yau (Chair)

BIOMETRICS SECTION: Sheng Luo (Chair, 2016)

CHAPTERS

ICSA-Canada Chapter: Changbao Wu (Chair), Wendy Lou (Treasurer/Secretary)

ICSA - New England Chapter: Huyuan Yang (Chair)

ICSA – Midwest Chapter: Lanju Zhang (Chair)

ICSA Shanghai Committee (Term 10/2013 to 10/2016): Dejun Tang (Chair, Novartis, China)

Office of ICSA

Lili Yu, Congjian Liu, Ruth Whitworth, Karl Peace
Jiann-Ping Hsu College of Public Health, Georgia Southern University

Conference Committees

Executive Committee

Yichuan Zhao	Chair, Georgia State University
Zhezhen Jin	Columbia University
Jian Chen	SunTrust, Atlanta
Nelson Chen	Treasurer, Emory University
Yi Li	University of Michigan
Peter Song	University of Michigan
Gang Li	Johnson & Johnson

Local Committee

Jian Chen	Co-Chair, SunTrust, Atlanta
Yichuan Zhao	Co-Chair, Georgia State University
Nelson Chen,	Treasurer, Emory University
Jie Chen	Augusta University
Yijuan Hu	Emory University
Ping Ma	University of Georgia
Yajun Mei	Georgia Institute of Technology
Sherry Ni	Kennesaw State University
Jing Zhang	Georgia State University

Program Committee

Yichuan Zhao	Chair, Georgia State University
Li-Shan Huang	National Tsinghua University, Taiwan
Xiaoming Huo	Georgia Institute of Technology
Qi Jiang	Amgen
Gang Li	University of California, Los Angeles
Mei-Ling Ting Lee	University of Maryland, College Park
Aiyi Liu	NIH
Shiyao Liu	Genentech
Qi Long	Emory University
Wenbin Lu	North Carolina State University
Xiaolong Luo	Celgene
James Pan	Janssen R&D
Liang Peng	Georgia State University
Yongming Qu	Eli Lilly and Company
Tony Sun	University of Missouri
Fei Tan	Indiana University-Purdue University Indianapolis
Samuel Wu	University of Florida
Haonan Wang	Colorado State University
Yujun Wu	Sanofi
Li-an Xu	Bristol-Myers-Squibb
Yunling Xu	FDA
Bo Yang	Vertex

Lijian Yang	Suzhou University
Lilly Yue	FDA
Jin-Ting Zhang	National University of Singapore
Wei Zhang	Boehringer Ingelheim Pharmaceuticals, Inc.
Jun Zhao	AbbVie
Yunpeng Zhao	George Mason University
Hongjian Zhu	University of Texas School of Public Health

Program Book Committee

Zhezhen Jin	Chair, Columbia University
Xiaoyi Min	Georgia State University
Yunxiao Chen (student)	Columbia University

Student Paper Award Committee

Yi Li	Chair, University of Michigan
Guang Cheng	Purdue University
Yixin Fang	New York University
Haoda Fu	Eli Lilly and Company
Steve Qin	Emory University
Hongkun Wang	Georgetown University

Short Course Committee

Peter Song	Chair, University of Michigan
Yi Pan	CDC
Limin Peng	Emory University
Lihong Qi	University of California, Davis
Song Yang	NIH
Zhigang Zhang	Memorial Sloan Kettering Cancer Center

Fund Raising Committee

Gang Li	Chair, Johnson & Johnson
Jason Liao	Novartis
Zhaoling Meng	Sanofi
Hongliang Shi	Blueprint Medicines
Li Zhu	Amgen

Webmaster

Haitao Huang	Georgia State University
Tu Tran	Georgia State University
Chengsheng Jiang	University of Maryland, College Park

Symposium Sponsors

The 2016 ICSA Applied Statistics Symposium is supported by a financial contribution of the following sponsors:

Georgia State University

EMORY UNIVERSITY

THE UNIVERSITY OF GEORGIA.

DEPARTMENT OF STATISTICS

FRANKLIN COLLEGE OF ARTS AND SCIENCES

AUGUSTA UNIVERSITY

MEDICAL COLLEGE OF GEORGIA

DEPARTMENT OF BIOSTATISTICS AND EPIDEMIOLOGY

ims

Georgia State University | COLLEGE OF ARTS & SCIENCES

KENNESAW STATE UNIVERSITY

College of Science and Mathematics

Department of Statistics and Analytical Sciences

Center for BEHAVIORAL NEUROSCIENCE

ASA

AMERICAN STATISTICAL ASSOCIATION

Promoting the Practice and Profession of Statistics

MERCK

Brightech

THE LOTUS GROUP

Institute of Biomedical Sciences, Georgia State University

Department of Mathematics and Statistics, Georgia State University

The organizing committees greatly appreciate the support of the above sponsors.

The 2016 ICSA Applied Statistics Symposium Exhibitor

Wiley

The Lotus Group

Eli Lilly and Company

Hospira, a Pfizer company

Center for Statistical Science at Georgia Tech

Statistics = data science?

-C.F. Jeff Wu

The Stewart School of Industrial & Systems Engineering and the School of Mathematics have joined forces to form the Center for Statistical Science at Georgia Tech. The focus of this interdisciplinary center is to leverage knowledge and resources to extend Georgia Tech's reach in research and teaching. The Center was established to be the home of statistical work at Georgia Tech.

#1 ranked Graduate Program for the 26th consecutive year

according to U.S. News & World Report

Statistical science consists of a body of tools, concepts, and algorithms for collecting, analyzing, and interpreting data. It is an integral part of ISyE, where engineers are trained in understanding the system by collecting and analyzing data. Learning statistics enables one to efficiently and cost-effectively gather data and build empirical models to describe and understand the performance of the system that can be used for prediction, control, optimization, and decision-making.

Research Areas:

- Bayesian Statistics
- Data Mining
- Design of Experiments
- Nonparametric Methods
- Regression Analysis
- Statistical Computing
- Statistical Inference
- Time Series Analysis

Application Areas

- Quality and Reliability Engineering
- Health Informatics
- Logistics
- Advanced Manufacturing

For more information or to apply, visit
statistics.gatech.edu

Georgia Tech Center for Statistical Science
Interdisciplinary Program

Transportation to the Hotel

Shuttle Service

Visit the SuperShuttle customer service booth at Hartsfield-Jackson International Airport and ask for the shuttle to the Hyatt Regency Atlanta Downtown. SuperShuttle Atlanta is the official 24 hour share-ride shuttle service provider. You can book your reservation by calling 1-800-Blue-Van (258-3826) or visit www.supershuttle.com. The rate to and from the airport one way is $16.50 and $33.00 round trip per person. A SuperShuttle van will pick you up at the designated time, bring you to the airport and drop you off right outside your airline terminal!

Limousine Service

The Concierge can arrange a limousine service to pick up guests from Atlanta Airport at a specified gate for $70.00 for a town car, $95.00 for a SUV or $155.00 for a stretch limousine.

Taxi Service

To/From airport fee: $30.00 plus $2.00 per each additional person. Plus, $1.50 Flag Drop Fee. Fares originating from a business and concluding at a business within the zone of downtown are at a rate of $10 ($2 for each additional person).

MARTA

MARTA (Metro Atlanta Rapid Transit Authority): $2.50 per ride + $1.00 for a Breeze Card to gain train access. Rail system runs approx. every 10 minutes. Hyatt Regency Atlanta is connected to Peachtree Center Train Station via Peachtree Center Mall. 20 minutes from Airport. To get to Hyatt, take MARTA to the Peachtree Center Station and exit Northeast towards Peachtree Center Mall.
For details and other means of transportation, visit http://atlantaregency.hyatt.com/en/hotel/our-hotel/transportation.html.

Travel to/in Atlanta

By plane

Atlanta's principal airport is Hartsfield-Jackson Atlanta International Airport (ATL), located about 8 miles south of downtown. Hartsfield-Jackson was for many years the world's busiest airport, and is a major hub for Delta Air Lines and Southwest Airlines.

By Car

Atlanta is at the intersection of several major highways: I-75, I-85, and I-20.

By Train

Atlanta is served by Amtrak. Amtrak's Crescent train runs daily and serves New York, Philadelphia, Baltimore, Washington, Charlotte, Gainesville, Birmingham and New Orleans. More Amtrak routes can be found here.

The Atlanta Amtrak station is located at 1688 Peachtree St. N.W., which is a few miles north of downtown.

Local Transportation by MARTA

Atlanta is served by MARTA (Metropolitan Atlanta Rapid Transit Authority), +1 404-848-4711, which operates both rapid rail and bus networks in the city of Atlanta.

Hyatt Regency Atlanta

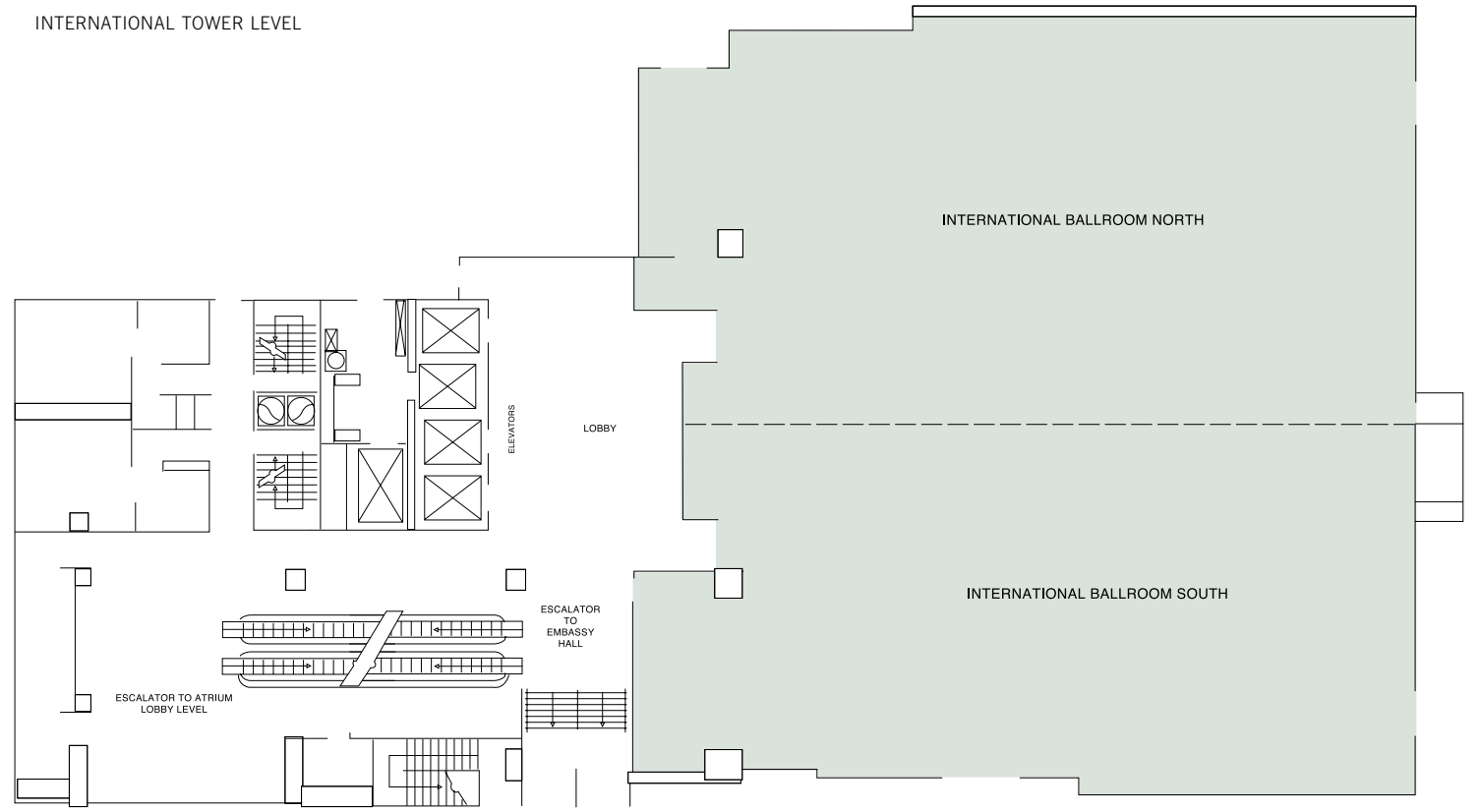
DIRECTIONS

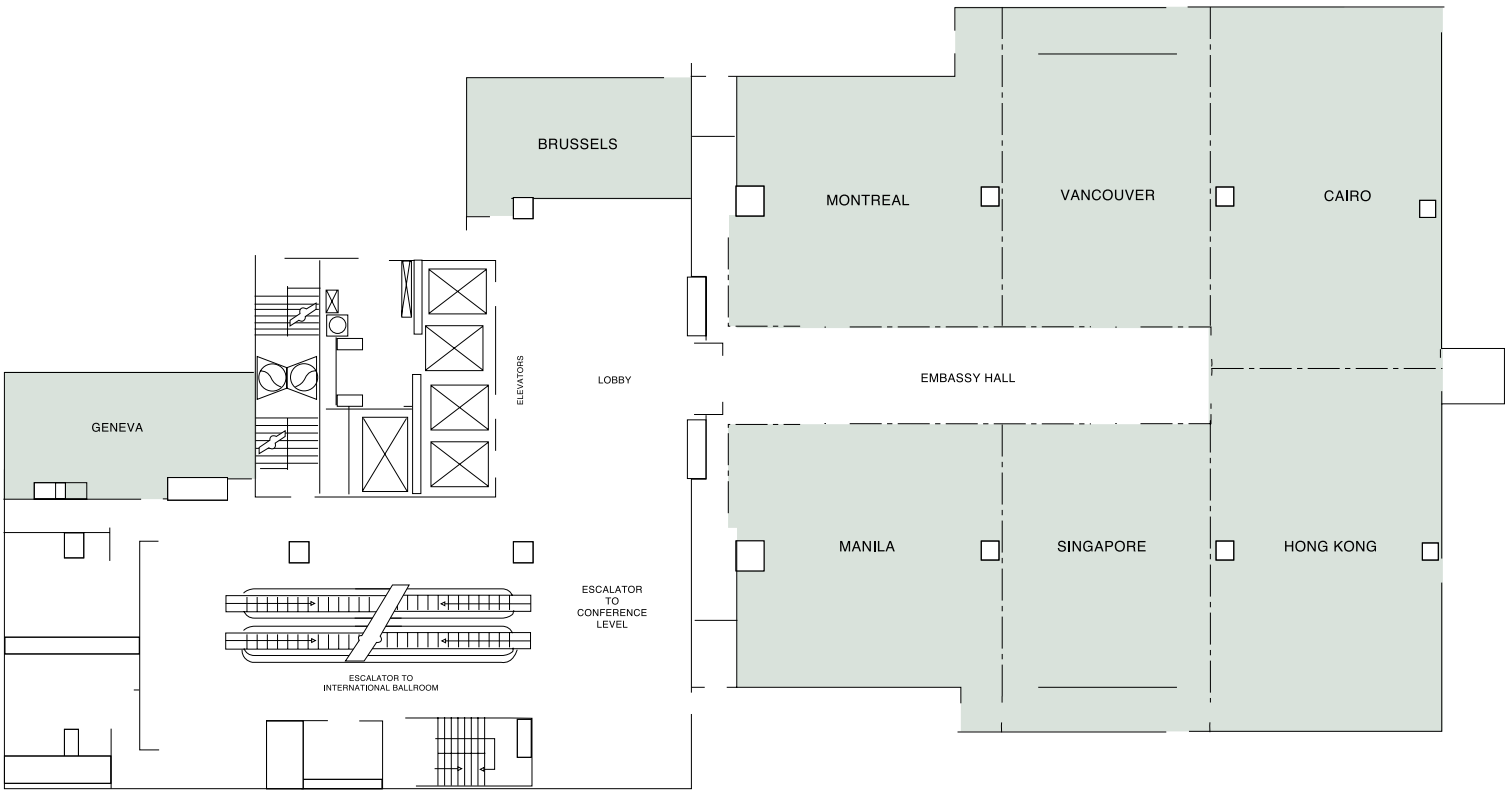
From Hartsfield-Jackson Int'l Airport (13 miles): Take 75 / 85 North to exit 248C (on right) and International Blvd. Turn left. Proceed to the third traffic light. Turn right on Peachtree Center Ave. Motor Lobby entrance is one block on left.

EXHIBIT LEVEL

Hyatt Regency Atlanta

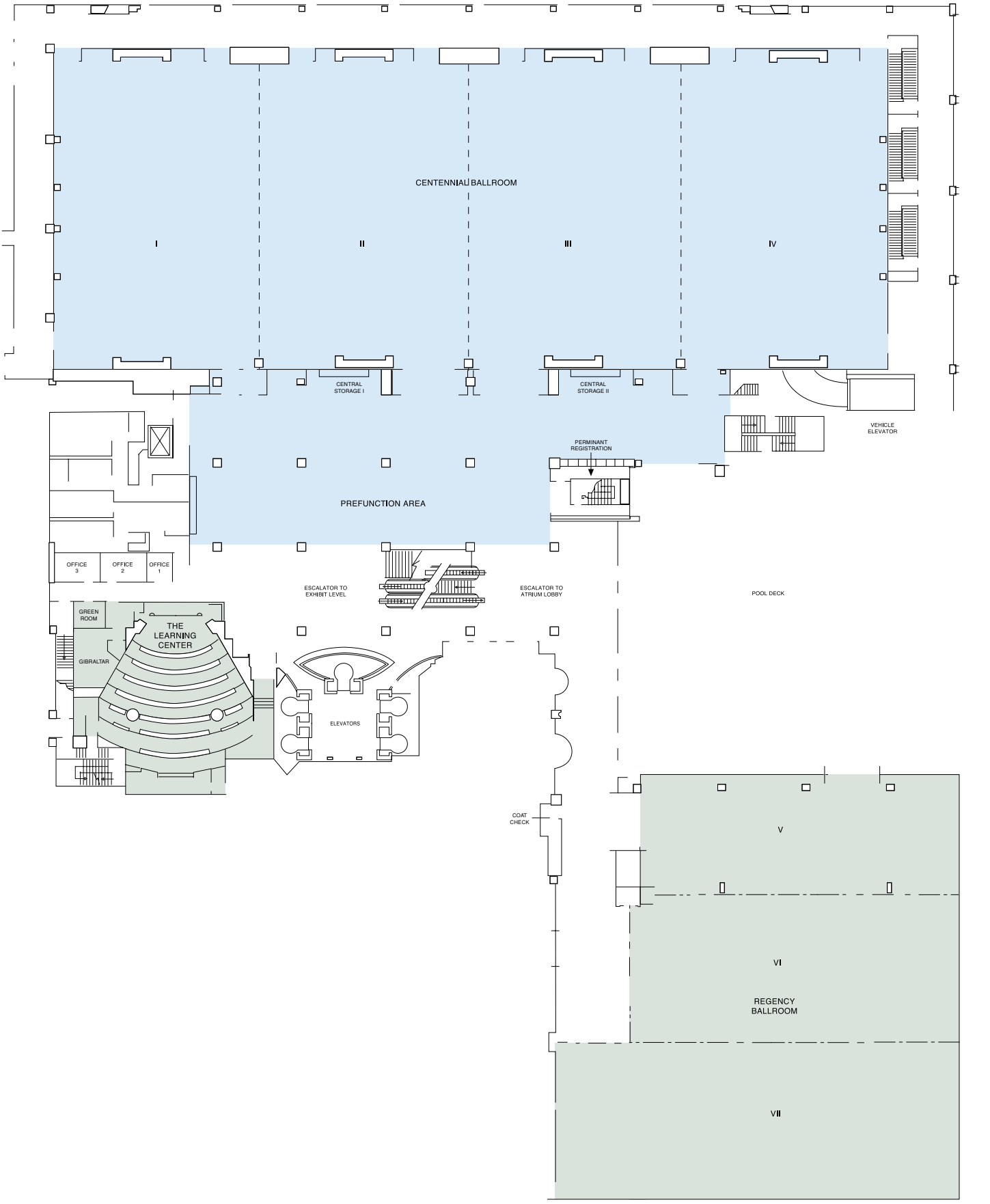
INTERNATIONAL TOWER LEVEL





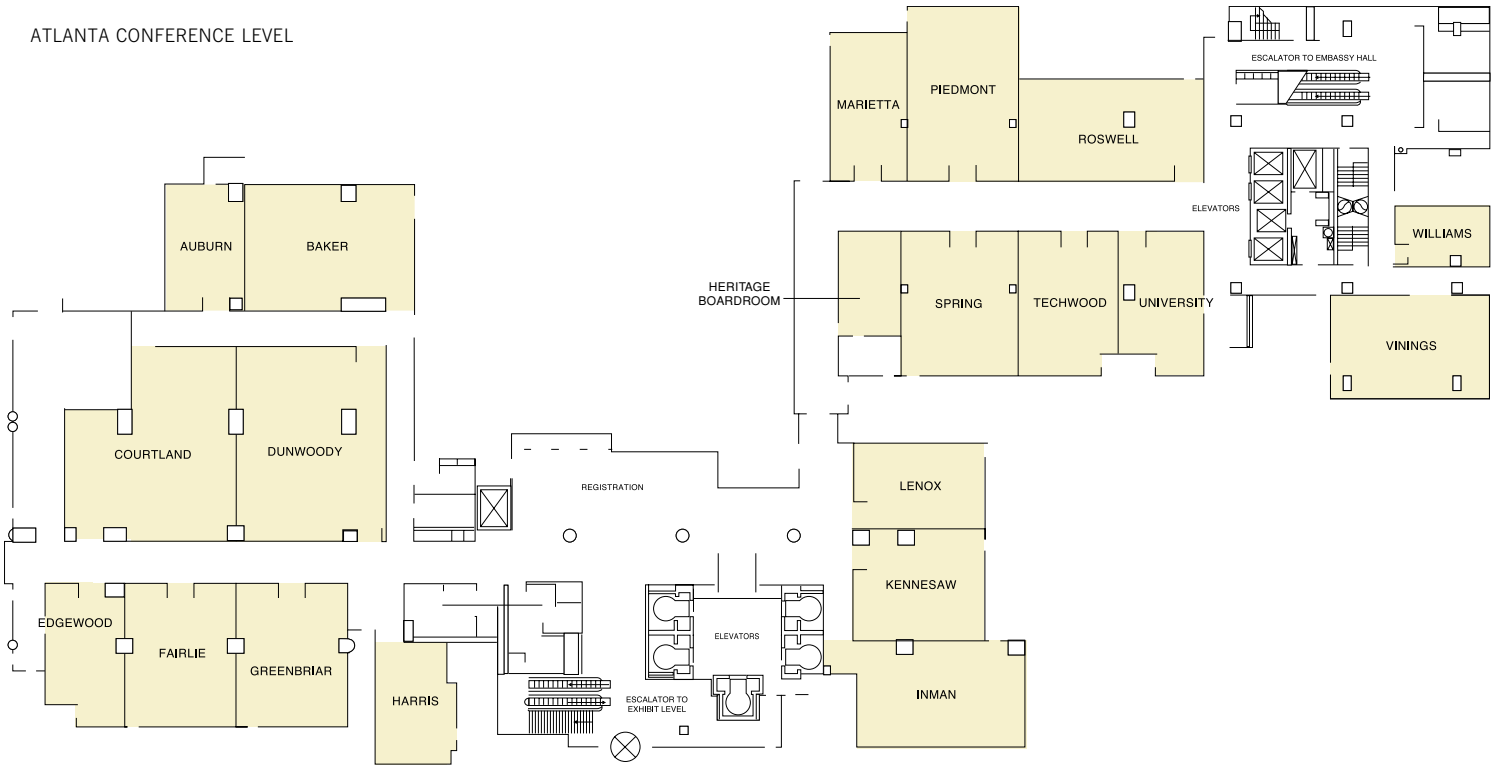
Hyatt Regency Atlanta

BALLROOM LEVEL

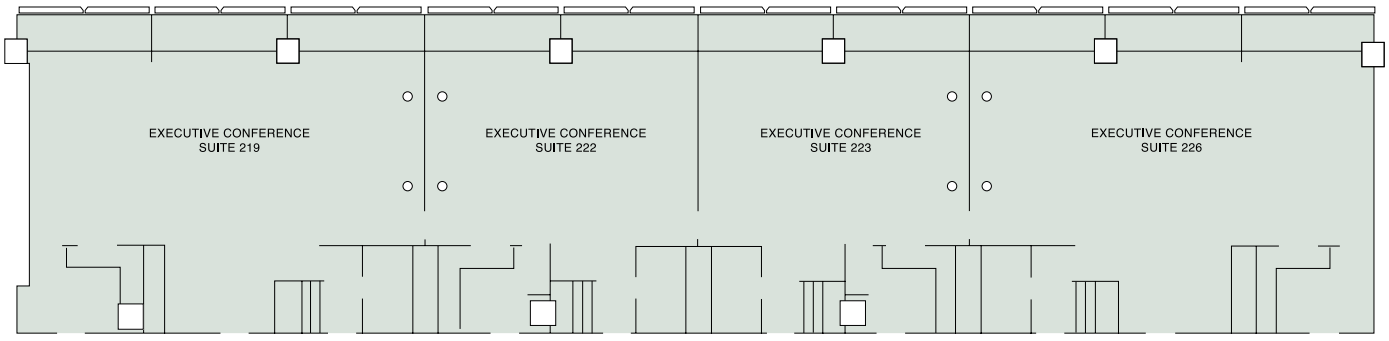


Hyatt Regency Atlanta

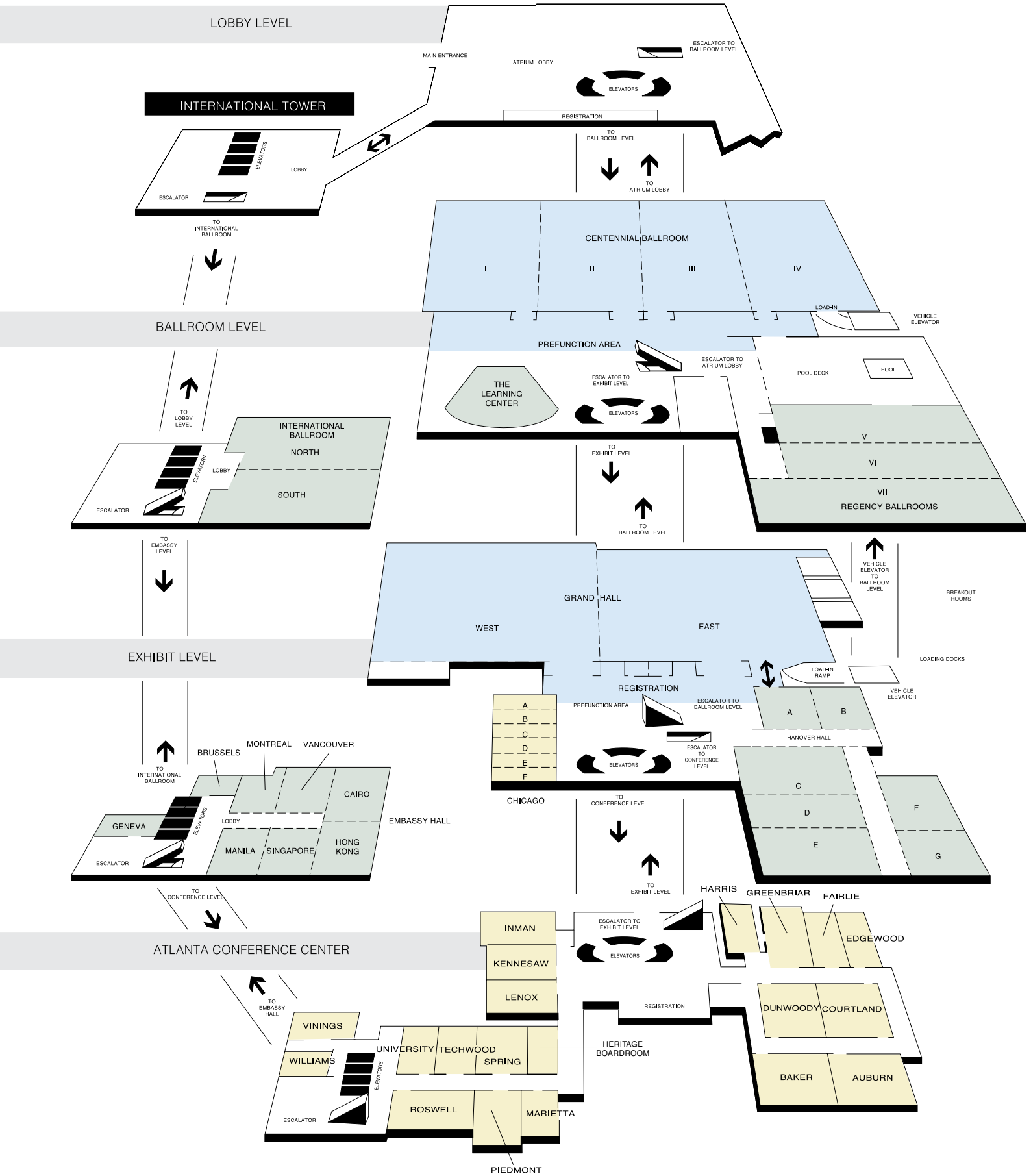
ATLANTA CONFERENCE LEVEL



EXECUTIVE SUITES



Hyatt Regency Atlanta



07.09

Sunday June 12, 2016

Time	Room	Session
8:00 AM - 6:00 PM	International Foyer	Registration
7:00 AM - 8:45AM		Breakfast
9:45 AM – 10:15 AM		Break
8:00 AM - 5:00 PM	Vinings	***Short Course:*** *Quantile Regression Methods for Applied Statistics*
8:00 AM - 5:00 PM	Roswell	***Short Course:*** *Advancing Drug Development through Precision Medicine and Innovative Adaptive Designs: Concepts, Rationale, and Case Studies*
8:00 AM - 5:00 PM	University	***Short Course:*** *Applied Nonlinear Statistical Methods*
8:00 AM - 12:00 PM	Piedmont	***Short Course:*** *Propensity Scores and Causal Inference*
8:00 AM - 12:00 PM	Spring	***Short Course:*** *Noninferiority Testing in Clinical Trials: Issues and Challenges*
8:00 AM - 12:00 PM	Techwood	***Short Course:*** *Statistical Methods and Software for Multivariate Meta-Analysis*
12:00 PM - 1:00 PM		***Lunch for Registered Full-Day Short Course Attendees***
1:00 PM - 5:00 PM	Piedmont	***Short Course:*** *Analysis of Complex Omics Data Using System-Based Approaches*
1:00 PM - 5:00 PM	Spring	***Short Course:*** *Analysis of Large and Complex Networks*
1:00 PM - 5:00 PM	Techwood	***Short Course:*** *Applied Meta-Analysis Using R*
2:45 PM - 3:15 PM		Break
6:00 PM - 8:30 PM	International North	***ICSA Board Meeting (Invited Only)***
7:00 PM - 9:00 PM	International South	***Opening Mixer***

Monday June 13, 2016

7:30 AM - 6:00 PM	International Foyer	Registration
7:00 AM – 8:45AM		Breakfast
8:00 AM - 8:30 AM	International Ballroom	***Opening Ceremony***: Yichuan Zhao, Mark Becker, Mei-Ling Ting Lee, David Morganstein
8:30 AM - 9:30 AM	International Ballroom	***Keynote I:*** Bin Yu, UC Berkeley
9:30 AM - 10:00 AM		Break
10:00 AM -11:40 PM	*See program*	***Parallel Sessions***
11:40 PM - 1:30 PM		Lunch on own
1:30 PM - 3:10 PM	*See program*	***Parallel Sessions***
3:10 PM - 3:30 PM		Break
3:30 PM - 5:10 PM	*See program*	***Parallel Sessions***

Tuesday June 14, 2016

7:30 AM - 5:30 PM	International Foyer	Registration
7:00 AM - 8:45 AM		Breakfast
8:00 AM - 9:00 AM	International Ballroom	***Keynote II:*** David Madigan, Columbia University
9:00 AM - 9:10 AM		Break
9:10 AM - 10:10 AM	International Ballroom	***Keynote III:*** Paul Albert, National Institute of Health
10:10 AM - 10:30 AM		Break
10:30 AM - 12:10 PM	*See program*	***Parallel Sessions***
12:10 PM - 1:30 PM		Lunch on own
1:30 PM - 3:10 PM	*See program*	***Parallel Sessions***
3:10 PM - 3:30 PM		Break
3:30 PM - 5:10 PM	*See program*	***Parallel Sessions***
6:30 PM - 10:00 PM	Off site	***Banquet (Banquet Speaker:*** Michael Eriksen, Georgia State University***)***

Wednesday June 15, 2016

8:30 AM - 1:00 PM	International Foyer	Registration
7:30 AM – 9:00 AM		Breakfast
8:30 AM - 10:10 AM	*See program*	***Parallel Sessions***
10:10 AM - 10:30 AM		Break
10:30 AM - 12:10 PM	*See program*	***Parallel Sessions***
1:00 PM – 6:00 PM		Excursion(fee event)

Bin Yu

Chancellor's Professor at the University of California, Berkeley

Bin Yu is Chancellor's Professor in the Departments of Statistics and of Electrical Engineering & Computer Science at the University of California at Berkeley. Her current research interests focus on statistics and machine learning theory, methodologies, and algorithms for solving high-dimensional data problems. Her group is engaged in interdisciplinary research with scientists from genomics, neuroscience, and remote sensing. She is Member of the National Academy of Sciences and Fellow of the American Academy of Arts and Sciences. She was a Guggenheim Fellow in 2006, and President of IMS (Institute of Mathematical Statistics) in 2013-2014. She is a Fellow of IMS, ASA, IEEE and AAAS. She has served or is serving on numerous journal editorial boards including those of JMLR, Annals of Statist, and JASA. She is on SAB of IPAM and BOT of ICERM.

Title: Unveiling the mysteries in spatial gene expression

Location and Time: International Ball Room, Monday June 13, 8:30-9:30 AM

Abstract: Genome-wide data reveal an intricate landscape where gene activities are highly differentiated across diverse spatial areas. These gene actions and interactions play a critical role in the development and function of both normal and abnormal tissues. As a result, understanding spatial heterogeneity of gene networks is key to developing treatments for human diseases. Despite the abundance of recent spatial gene expression data, extracting meaningful information remains a challenge for local gene interaction discoveries. In response, we have developed staNMF, a method that combines a powerful unsupervised learning algorithm, nonnegative matrix factorization (NMF), with a new stability criterion that selects the size of the dictionary. Using staNMF, we generate biologically meaningful Principle Patterns (PP), which provide a novel and concise representation of Drosophila embryonic spatial expression patterns that correspond to pre-organ areas of the developing embryo. Furthermore, we show how this new representation can be used to automatically predict manual annotations, categorize gene expression patterns, and reconstruct the local gap gene network with high accuracy. Finally, we discuss on-going crispr/cas9 knock-out experiments on Drosophila to verify predicted local gene-gene interactions involving gap-genes. An open-source software is also being built based on SPARK and Fiji.

This talk is based on collaborative work of a multi-disciplinary team (co-lead Erwin Frise) from the Yu group (statistics) at UC Berkeley, the Celniker group (biology) at the Lawrence Berkeley National Lab (LBNL), and the Xu group (computer science) at Tsinghua Univ.

David Madigan

Professor, Department of Statistics, Executive Vice President and Dean of the Faculty of Arts and Sciences, Columbia University

David Madigan is a Professor of Statistics at Columbia University in New York City where he also serves as Executive Vice President and Dean of the Faculty of Arts & Sciences. He received a bachelor's degree in Mathematical Sciences and a Ph.D. in Statistics, both from Trinity College Dublin. He has previously worked for AT&T Inc., Soliloquy Inc., the University of Washington, Rutgers University, and SkillSoft, Inc. He has over 150 publications in such areas as Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance and probabilistic graphical models. He is an elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. He recently completed terms as Editor-in-Chief of Statistical Science and as Editor-in-Chief of Statistical Analysis and Data Mining - the ASA Data Science Journal.

Title: Observational Studies: Promise and Peril

Location and Time: International Ball Room, Tuesday June 14, 8:00-9:00 AM

Abstract: Threats to the validity of observational studies on the effects of interventions raise questions about the appropriate role of such studies in decision making. Nonetheless, scholarly journals in fields such as medicine, education, and the social sciences feature many such studies, often with limited exploration of these threats, and the lay press is rife with news stories based on these studies. Consumers of these studies rely on the expertise of the study authors to conduct appropriate analyses, and on the thoroughness of the scientific peer-review process to check the validity, but the introspective and ad hoc nature of the design of these analyses renders appropriate interpretation of such studies challenging at best. In this talk I will provide an overview of the current state of the art in observational studies with a particular focus on healthcare. I will describe some current promising research directions.

Paul Albert

Senior investigator and Branch chief of DESPR's Biostatistics and Bioinformatics Branch, National Institute of Health

Dr. Albert is Senior Investigator and Chief of the Biostatistics & Bioinformatics Branch in the Division of Intramural Population Health Research at the Eunice Kennedy Shriver National Institute of Child Health and Human Development. During Dr. Albert's 27 year career at the NIH his primary research has been in the areas of longitudinal data analysis, diagnostic testing, and the analysis of complex biomarker studies. His collaborative and methodological research has spanned many application areas including neurology, psychiatry, cardiac disease, cancer, and most recently fetal medicine. Dr. Albert has published over 290 peer reviewed papers in leading statistical and medical journals including papers in the Journal of the American Statistical Association, the Journal of the Royal Statistical Society Series A and C, Biometrika, Biometrics, and Biostatistics. He is a fellow of the American Statistical Association and serves as an associate editor of Biometrics, Statistics in Medicine, Statistics in the Biosciences, and Fertility and Sterility

Title: Statistical Challenges in Obstetrics: Predicting poor pregnancy outcomes from multivariate Ultrasound Fetal Growth data

Location and Time: International Ball Room, Tuesday June 14, 9:10-10:10 AM

Abstract: There are many analytical issues in the area of obstetrics that present challenges for statisticians working in this exciting area. We will review the general area and then focus on the problem of predicting poor pregnancy outcomes such as small-for-gestational age or preterm birth from longitudinal biomarkers collected longitudinally during pregnancy. We will begin by presenting simple two-stage estimation procedures that approximates a full maximum-likelihood approach for predicting a binary event from multivariate longitudinal growth data (Albert, Statistics in Medicine, 2012). Subsequently, we will present a class of joint models for multivariate growth curve data and a binary event that accommodates a flexible skewed error distribution for the ultrasound measurements and an asymmetric link function relating the longitudinal to the binary process (Kim and Albert, In Press at Biometrics). Finally, we will present a tree-based approach for identifying subgroups of women who have an enhanced predictive accuracy for predicting a binary event from fetal growth data (Foster, et al., JRSS-A, 2016).

Michael Eriksen

Regents' Professor and Founding Dean of School of Public Health, Georgia State University

Michael Eriksen is Regents' Professor and founding Dean of the School of Public Health at Georgia State University. He is also director of Georgia State University's Tobacco Center of Regulatory Science (TCORS) and the Center of Excellence in Health Disparities Research (CoEx). Prior to his current positions, Eriksen served as a senior advisor to the World Health Organization in Geneva and was the longest-serving director of the Centers for Disease Control and Prevention's Office on Smoking and Health (1992-2000). Previously, Eriksen was director of behavioral research at the M.D Anderson Cancer Center. He has recently served as an advisor to the Bill & Melinda Gates Foundation, the Robert Wood Johnson Foundation, the American Legacy Foundation, and the CDC Foundation.

Eriksen has published extensively on tobacco prevention and control and is the lead author of The Tobacco Atlas. He has served as an expert witness on behalf of the US Department of Justice and the Federal Trade Commission in litigation against the tobacco industry. He is editor-in-chief of Health Education Research and has been designated as a Distinguished Cancer Scholar by the Georgia Cancer Coalition. He is a recipient of the WHO Commemorative Medal on Tobacco or Health and a Presidential Citation for Meritorious Service, awarded by President Bill Clinton. Eriksen is a past president and Distinguished Fellow of the Society for Public Health Education, and has been a member of the American Public Health Association for over 40 years.

Title: Tobacco Use in China: Great Challenges and Greater Opportunities

Time and Location: Golden House Restaurant, 1600 Pleasant Hill Rd, Duluth, 30096, Tuesday June 14 6:30 PM

Abstract: China is the epi-center of global tobacco issues, and data are essential in fully understanding the public health challenges and opportunities. One-third of the world's smokers are Chinese, and there are more smokers in China, almost exclusively men, than there are people in the United States. China is also the world's largest grower of tobacco leaves, the biggest manufacturer of cigarettes, and is controlled by a monopoly – the China National Tobacco Corporation.

While there are great challenges, there are even greater opportunities. In the United States, there are more ex-smokers than there are current smokers. In China, only about 10% of "ever smokers" have quit, so there is the potential for hundreds of millions of Chinese men to quit smoking. The public health question is really not "if," but "when." In the US, it has taken over 50 years to reduce smoking rates in half. Given the dynamism of China, hopefully similar results will not take half a century. The importance of data in reducing smoking in China will be presented with special reference to The Tobacco Atlas http://www.tobaccoatlas.org and our work in 22 Chinese cities funded by the Gates Foundation and Pfizer, Inc. http://ctp.publichealth.gsu.edu/.

ASA Bio-pharmaceutical Awards

Lijia Wang, Emory University
- Title: A Latent Class Modeling Approach for Predicting Kidney Obstruction in the Absence of a Gold Standard
- Time: Tuesday, June 14th. 1:30 PM - 3:10 PM
- Session 85: Student Award Session (GREENBRIAR, ATLANTA CONFERENCE CENTER -LL3)

Xiaolu Zhu, University of Illinois at Urbana-Champagne
- Title: Individualizing Drug Dosage with Longitudinal Data
- Time: Tuesday, June 14th. 1:30 PM - 3:10 PM
- Session 85: Student Award Session (GREENBRIAR, ATLANTA CONFERENCE CENTER -LL3)

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

Xinrui Zhang, University of Florida
- Title: Internal pilot design for repeated measures
- Time: Monday, June 13th, 3:30 PM - 5:10 PM
- Session 44: J.P Hsu Invited Session on Biostatistical and Regulatory Sciences (INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)

ICSA Student Paper Awards

Fei Gao, University of North Carolina – Chapel Hill
- Title: Semiparametric Estimation of the Accelerated Failure Time Model with Partly Interval-censored Data
- Time: Tuesday, June 14th. 1:30 PM - 3:10 PM
- Session 85: Student Award Session (GREENBRIAR, ATLANTA CONFERENCE CENTER -LL3

Yi Lu, Ohio State University
- Title: Bayesian Registration of Functions with a Gaussian Process Prior
- Time: Tuesday, June 14th. 1:30 PM - 3:10 PM
- Session 75: New development in function data analysis (TECHWOOD, ATLANTA CONFERENCE CENTER LL3)

Ziyi Li, Emory University
- Title: Incorporating Biological Information in Sparse PCA with Application to Genomic Data
- Time: Tuesday, June 14th. 1:30 PM - 3:10 PM
- Session 85: Student Award Session (GREENBRIAR, ATLANTA CONFERENCE CENTER -LL3)

So Young Park, North Carolina State University
- Title: Optimal Design for Sparse Functional Data
- Time: Tuesday, June 14th. 3:30 PM – 5:30 PM
- Session 96: Methods of integrating novel functional data with multimodal measurements (TECHWOOD, ATLANTA CONFERENCE CENTER - LL3

Kevin Lee, The Pennsylvania State University
- Title: Nonparametric mixture of Gaussian graphical models, with applications to ADHD imaging data
- Time: Tuesday, June 14th. 3:30 PM - 5:10 PM
- Session 93: Recent developments of graphical model for big data (EDGEWOOD, ATLANTA CONFERENCE CENTER LL3

1. Quantile Regression Methods for Applied Statistics

Presenter: Marco Geraci, University of South Carolina

Course length: One day

Outline/Description
Quantile regression (QR) is a statistical technique for the analysis of conditional quantile functions -- models in which quantiles of the conditional distribution of the response variable (e.g., the median or the 5th centile) are modeled as functions of a set of predictors. QR is capable of handling complex effects on the location, scale and shape of a distribution. Unlike mean regression, no parametric assumption about the shape of the error distribution is needed.

This course will provide an introduction to principles of and methods for QR analysis. It will cover the basics as well as more advanced methods (models for count and binary data, clustered data, transformation-based models for nonlinear and bounded responses). Problems will be motivated by applications in different fields, including epidemiology and clinical research (e.g., growth charts modeling, assessment of quantile treatment effects). Examples with data analyses will be illustrated using R software.

References
R. Koenker (2005). Quantile Regression, New York: Cambridge University Press.

L. Hao and D. Q. Naiman (2007). Quantile Regression, London: SAGE.

About the presenter
Dr. Geraci is Associate Professor of Biostatistics at the University of South Carolina (USC). His research interests are in quantile inference, random-effects models, missing data, and statistical computing. He taught courses on quantile regression at University College London, the Royal Statistical Society (RSS) and USC, and gave seminars on related topics in numerous universities. Dr. Geraci has published peer-reviewed articles in statistics, cancer epidemiology, physical activity, gastroenterology, nuclear medicine and higher education. He is the developer of Linear Quantile Mixed Models (LQMM), that is, QR models for clustered data. He also authored three statistical software packages, including `lqmm`, an R package for LQMM methods, and `Qtools`, a collection of utilities for quantile inference. In 2010, Dr. Geraci was awarded Chartered Statistician by the RSS. He is Statistical Editor for the *Journal of Child Health Care* (SAGE Publications) and member of the Editorial Board for the magazine *Significance* (Wiley on behalf of the RSS and the American Statistical Association).

2. Advancing Drug Development through Precision Medicine and Innovative Adaptive Designs: Concepts, Rationale, and Case Studies

Presenters: Sandeep Menon, and Weidong Zhang, Pfizer, Inc, and Jing Wang, Gilead Siences, Inc.

Course Length: One day

Outline/Description:
Precision medicine has paved the way for a new era of delivering tailored treatment options to patients according to their biological profiles. In combination with innovative adaptive design, this has presented drug developers unprecedented opportunities to engage novel thinking to accelerate drug discovery. In the first part of this course, step-by-step introductions to basic biology and genetics will be presented, and is followed by overviews of cutting edge technologies such as microarray and next generation sequencing technologies that have been widely used to generate omics data. Built on the basic knowledge of biology and omics data, key concepts of precision medicine studies and strategies of how in practice this novel approach can be applied to drug discovery will be discussed. In addition, statistical considerations and challenges posed in omics data such as data normalization, statistical modeling and interpretation will also be discussed. Examples are case studies from the instructors' work and from medical literature. The second part of this course will introduce a few concepts on Adaptive and Bayesian Designs focusing on biomarker-adaptive approach that utilize biomarker information that may be identified from precision medicine studies. Bayesian methods for design and monitoring for randomized clinical trials will be presented.

References
Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods.

About the presenters
Dr. Weidong Zhang is a trained statistician and a statistical geneticist. At Pfizer, he leads the efforts in precision medicine and clinical biomarker study focusing on omics data. In addition, he is a lead statistician in multiple Phase 1 and Phase 2 clinical programs. Weidong is passionate about drug discovery and precision medicine. 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 obtained his PhD degree in Statistical Genetics and MS degree in Statistics both from the University of Wisconsin-Madison.

Dr. Sandeep Menon is currently the Vice President and Head of the Statistical Research and Consulting Center at Pfizer Inc. and also holds adjunct faculty positions at Boston University and Tufts University School of Medicine. His group, located at different Pfizer sites globally, provides scientific and statistical leadership and consultation to the Global Head of Statistics and senior Pfizer management in discovery, clinical development, legal, commercial, and marketing. His responsibilities also include providing a strong presence for Pfizer in regulatory and

professional circles to influence content of regulatory guidelines and their interpretation in practice. Previously he held positions of increased responsibility and leadership where he was in charge of all the biostatistics activities for the entire portfolio in his unit, spanning from discovery (target) through proof-of-concept studies for supporting immunology and autoimmune disease, inflammation and remodeling, rare diseases, cardiovascular and metabolism, and center of therapeutic innovation. He was responsible for overseeing biostatistical aspects of more than 40 clinical trials, more than 25 compounds, and 20 indications. He is a core member of the Pfizer Global Clinical Triad (Biostatistics, Clinical, and Clinical Pharmacology) Leadership team. He has been in the industry for over a decade, and prior to joining Pfizer he worked at Biogen Idec and Aptiv Solutions. His research interests are in adaptive designs and personalized medicine. He has several publications in top-tier journals and recently coauthored and coedited Clinical and Statistical Considerations in Personalized Medicine and Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods. He is an active member of the Biopharmaceutical Section of the American Statistical Association (ASA), serving as associate editor of ASA journal Statistics in Biopharmaceutical Research (SBR) and as a core member of the ASA Samuel S. Wilks Memorial Medal Committee. He is the co-chair of the sub-team under the cross industry DIA-sponsored Adaptive Design Scientific Working Group (ADSWG) on the Role of Adaptive Designs in Personalized Medicine, member of the biomarker identification sub-team formed under the currently existing multiplicity working group sponsored by the Society for Clinical Trials, and an invited program committee member at the Biopharmaceutical Applied Statistics Symposium (BASS). He is on the editorial board for the Journal of Medical Statistics and Informatics and on the advisory board for the MS in biostatistics program at Boston University

Dr. Jing Wang received her PhD degree in biostatistics at Boston University in 2013. She is currently a senior biostatistician at Gilead Sciences, Inc. At Gilead, she has participated in the NDA submission for Fixed-Dose Combination of Sofosbuvir/Velpatasvir for Treatment of Hepatitis C as a lead biostatistician, and she is currently working on several global Phase 3 studies for HCV treatment. Her research interests are in adaptive designs, personalized medicine and multi-regional clinical trial designs. She has published her work in multiple statistical journals and books.

3. Applied Nonlinear Statistical Methods

Presenter: Timothy E. O'Brien, Loyola University Chicago

Course length: One Day

Outline/Description
Researchers often recognize that nonlinear regression models are more applicable for modelling their physical and medical processes than are linear ones for several important reasons. Nonlinear models usually fit their data well and often in a more parsimonious manner (typically with far fewer model parameters). Also, nonlinear models and the corresponding model parameters are usually more scientifically meaningful. But selecting an efficient experimental design; choosing, fitting and interpreting an appropriate nonlinear model; and deriving and interpreting confidence intervals for key model parameters present practitioners with fundamental and important challenges.

This course covers linear regression and generalized linear models (e.g. logistic regression), Gaussian nonlinear models, and generalized nonlinear models, focusing on applications. Illustrations are given from bioassay, relative potency and drug or similar compound synergy useful in biomedical (including pharmacokinetics) and environmental sciences (including toxicology). Caveats are discussed regarding convergence, diagnostics, and the inadequacy of standard (Wald) confidence intervals provided by most software packages. Extensions to bivariate situations (e.g., for both efficacy and safety of drugs) and censored (survival) analysis are also provided, as are implications for experimental design. Implementation using the SAS and R statistical software packages is discussed.

References
O'Brien, T.E., Intermediate Methods in Applied Statistics and Biostatistics, forthcoming Springer text.

About the presenter
Tim O'Brien is a tenured professor and graduate director at Loyola University Chicago. He is a four-time Fulbright awardee, Loyola Master Teacher awardee, Best Paper awardee, and 2011-12 Jefferson Science Fellow (US State Dept., USAID), and a member of ASA, ENAR, RSS (elected) and ISI (elected). His PhD dissertation (NCSU, 1993) and research focuses on robust optimal design, generalized and normal nonlinear modelling especially in assessing potency and synergies, and predictive analytics. Dr. O'Brien has published over 50 research articles and book chapters, and engages in extensive statistical consulting in industry (Amgen, BMS, BASF, Chiron, Glaxo, J&J, Janssen, Novartis), research institutes (IDI, INRA, INSERM, NIH, USAID), national and international universities, and various I/NGO's (Battelle, PHPT). Chosen as an American Statistical Association Travelling Short Course, this course has been given 21 times in 12 countries to over 650 participants.

4. Statistical Methods and Software for Multivariate Meta-Analysis

Presenters: Haitao Chu, University of Minnesota, and Yong Chen, University of Pennsylvania

Course length: Half Day

Outline/Description
Comparative effectiveness research is aiming at informing health care decisions concerning the benefits and risks of different

diagnosis and treatment options. The growing number of assessment instruments and treatment options for a given condition, as well as the rapid escalation in their costs, has generated the increasing need for scientifically rigorous comparisons of diagnostic tests and multiple treatments in clinical practice via multivariate meta-analysis. The overall goal of this short course is to give an overview of the cutting-edge and robust multivariate meta-analysis methods to enhance the consistency, applicability, and generalizability for meta-analysis of diagnostic tests and multiple treatment comparisons. A number of case studies with detailed annotated SAS and R codes will be presented.

The outline is as follows: (1) Methods and Software for Meta-analysis of Diagnostic Tests (2 hour): (a) When the reference test can be considered as a gold standard (1 hour): Bivariate general linear mixed and generalized linear mixed models; Trivariate model that accounts for disease prevalence; Alternative parameterizations. (b) When the reference test cannot be considered as a gold standard (1 hour): Random effects models; Bayesian HROC Models; Unification of the previous two models. (2) Methods and Software for Meta-analysis of Multiple Treatments Comparisons (2 hour): (a) Contrast-based network meta-analysis: (1 hour); (b) Arm-based network meta-analysis: (1 hour).

Textbook/References
It will be based on co-instructors' recent multiple publications.

About the presenters
Dr. Chu's methodology research focuses on comparative effectiveness research, meta-analysis and diagnostic test accuracy studies, and he has been working on meta-analysis for more than ten years. He has published over 125 peer-reviewed manuscripts, including over 25 on systematic reviews and meta-analysis in top ranked statistical and medical journals such as *JASA, Biometrics, SIM, SMMR, Clinical Trials, JNCI, AIDS* and *AJE*, and coauthored three R packages pcnetmeta, mmeta and xmeta. Dr. Chu's research on meta-analysis has been supported by FDA, AHRQ, NIAID, NIDCR and NLM.

Dr. Yong Chen has been working on the field of meta-analysis since 2008. He has published over 50 peer-reviewed manuscripts, including 15 on statistical methods for meta-analysis in top statistical journals such as *Biometrics, JRSS-C, SIM* and *SMMR*. He is one of the main contributors R mmeta and xmeta. His research has been supported by both AHRQ and NIH.

5. Analysis of Complex Omics Data Using System-Based Approaches

Presenter: Shuangge Ma, Yale University

Course Length: Half day

Outline/Description
Omics data on complex diseases are now commonly encountered by researchers in academia, biopharmaceutical companies, and government agencies. They are featured with extremely high dimensionality and complex interconnections among measurements, and are a unique but popular type of big data. For the analysis of complex omics data, classic statistical approaches, which focus on individual and groups of variables, are insufficient. Instead, it is necessary to take a system perspective. In the recent literature, system-based methods have been developed, however, have not been systematically introduced. This short course targets to fill this gap. In this course, I will survey the newly developed system-based methods for the analysis of complex omics data, with an emphasis on methodological development and applications. Topics to be covered will include: (1) Background of omics data, with brief discussions on experiments (that generate such data), data collection, processing, and management. (2) A brief survey of the existing methods; Discussions on their insufficiency. (3) System-based analysis approaches: (3.a) Principle of system-based analysis; (3.b) Multiple system-based methods that address different aspects of omics data analysis. There will be an emphasis on network-based analysis. Both single-layer and multiplex network analysis methods will be discussed; (3.c) Brief discussions on implementations, software packages, and examples; (3.d) Pros and cons of system-based analysis approaches. (4) Discussions on unsolved problems and future directions.

References
Slides and reading materials will also be made available.

About the presenter
Shuangge Ma is associate professor of biostatistics. He is a Fellow of ASA and Elected Member of ISI. His research interests include genetic epidemiology, statistical genetics, semiparametric analysis, and survival analysis. He has published over 100 journal articles, 14 abstracts, and 2 books. He has offered over 100 presentations at major conferences and university seminars. He has been conducting research on the analysis of omics data for over ten years. On the system-based analysis, he has published in *Bioinformatics, Genetic Epidemiology, Annals of Statistics, Briefings in Bioinformatics, BMC Medical Genomics*, and other journals.

6. Applied Meta-Analysis Using R

Presenters: Yan Ma, The George Washington University, and Ding-Geng (Din) Chen, University of North Carolina at Chapel Hill.

Course length: Half day

Outline/Description
Meta-analysis is the art and science of synthesizing information from diverse sources to draw a more effective inference. With rising cost of medical and health research, many clinical studies are carried out with small sample sizes resulting in low power for detecting clinically useful effect size. This phenomenon has increased the chance of producing conflicting results from different studies. By pooling the estimated effect sizes from each of the component studies through metaanalytic methods, information from larger number of patients and increased power is expected. Because of its important impact, choice of methods for performing meta-analysis has come under scrutiny and development of novel methods and software has become a rapidly growing field. This half-day short course is then designed to present an overview of meta-analysis from the theoretical meta-analysis models to their implementations in the popular public available free software R with a variety of examples. These examples are compiled from real applications in medical and public health literature and the analyses are illustrated by a step-by-step fashion using appropriate R packages and functions. Topics to be discussed will include: (1) An overview of systematic review and meta-analysis with examples of published meta-analyses in medicine. (2) An introduction of R packages for meta-analysis. (3) Fixed-effects and random-effects meta-analysis and statistical methods for identifying and quantifying between-study heterogeneity. (4) Meta-analysis for continuous data and binary data (odds ratio, risk ratio and risk difference). (5) Meta-analysis for rare events: continuity correction and methods without using continuity correction. (6) Reporting guidelines. This course is designed to familiarize attendees with statistical techniques and R packages specific to meta-analysis and enable them to follow the logic and R implementation to analyze their own research data. Prior programming knowledge of R is a plus but not required for this course. All course contents are covered in the two books written by the instructors: "Applied Meta-Analysis Using R" and "Innovative Statistical Methods for Public Health Research".

References
Chen D, Peace KE. Applied Meta-Analysis with R. Chapman & Hall/CRC Biostatistics Series, 2013.

Ma Y, Zhang W, Chen D. Meta-analytic Methods for Public Health Research. Innovative Statistical Methods for Public Health Research. Springer ICSA Book Series in Statistics (Chen D and Wilson J Eds, 2015)

About the presenters
Dr. Yan Ma is an Associate Professor of Biostatistics at the George Washington University and was an Assistant Professor at Weill Medical College of Cornell University. He has published over 80 peer-reviewed papers and recently received an R01 award from AHRQ for development of statistical methods and health disparities research. In 2010, he was a recipient of Statistics in Epidemiology Young Investigator Award by the American Statistical Association. In 2012, Dr. Ma and his collaborators won the esteemed Team Science Award from the Association for Clinical Research Training, American Federation for Medical Research, Association for Patient Oriented Research, and Society for Clinical and Translational Science. Dr. Ma has been invited to give research talks for his work in meta-analysis at several national conferences and academic institutions. He developed and taught a new course "Applied meta-analysis with R" at the George Washington University and recently published with Dr. Chen together a book chapter on meta-analysis in Springer/ICSA Book Series in Statistics.

Dr. Din Chen is Wallace H. Kuralt Distinguished Professor at the University of North Carolina at Chapel Hill and was the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. He has more than 100 referred professional publications and co-authored six books on clinical trial methodology and public health applications. Professor Chen was honored with the "Award of Recognition" in 2014 by the Deming Conference Committee for highly successful advanced biostatistics workshop tutorials at 4 successive Deming conferences on 4 different books that he has written or co-edited. In 2013, he was invited to give a short course at the twentieth Annual Biopharmaceutical Applied Statistics Symposium (BASS XX, 2013) for his contribution in meta-analysis and received a "Plaque of Honor" for his short course.

7. Noninferiority Testing in Clinical Trials: Issues and Challenges

Presenter: Tie-Hua Ng, Food and Drug Administration

Course length: Half day

Outline/Description:
The objective of a noninferiority (NI) trial is to show that the test treatment or the experimental treatment is not inferior to the standard therapy or the active control by a small margin known as the NI margin. This short course elaborates the rationale of choosing the NI margin as a small fraction of the therapeutic effect of the active control as compared to placebo in testing of the NI hypothesis of the mean difference with a continuous outcome. This NI margin is closely related to M1 and M2, the NI margins discussed in the FDA draft guidance on NI clinical trials issued in March of 2010.

This short course also covers fundamental concepts related to NI trials, such as assay sensitivity, constancy assumption, discounting and preservation. As time permits, this short course (i) explains the differences between fixed-margin and synthesis methods, (ii) addresses the issues of switching between superiority and noninferiority, (iii) discusses gold-standard design and the equivalency of three or more treatment groups, (iv) investigates the roles of intention-to-treat and per-protocol analyses, and (v) presents an extended example of thrombolytic therapies.

References
Ng, T-H. (2014). Noninferiority Testing in Clinical Trials: Issues and Challenges. Boca Raton, FL: Chapman & Hall/CRC. (http://www.crcpress.com/product/isb n/9781466561496)

About the presenter:
Tie-Hua Ng, Ph.D. (Statistics from the University of Iowa in 1980) held several positions before joining the Food and Drug Administration (FDA) in 1987. He left the FDA in 1990 to work for the Henry M. Jackson Foundation. In 1995, he returned to the FDA, Center for Biologics Evaluation and Research (CBER). He is currently a team leader supporting the Office of Blood Research and Review within CBER. Over the past 23 years, he had made numerous presentations at professional meetings and published extensively in the area of active controlled/noninferiority studies. He offered four half-day short courses from 2009 through 2012. As a result of these efforts, he is the sole author of a book entitled "Noninferiority Testing in Clinical Trials: Issues and Challenges" published in December 2014.

8. Propensity Scores and Causal Inference

Presenters: Wei Pan, Duke University, and Haiyan Bai, University of Central Florida

Course length: Half day

Outline/Description:
Based on the causal inference framework, through lectures and hands-on activities, this shot course will introduce basic concepts of propensity score methods, including matching, stratification, and weighting; and the use of R packages such as MatchIt. Packages of propensity score methods in SAS, Stata, and SPSS will be also briefly introduced. This course is appropriate for applied statisticians in industries, government, and academia. Participants will learn why and when we need propensity score methods for making causal inference from observational studies on big data and how to perform propensity score methods using available software packages on samples from large-scale national databases.

References
Pan, W., & Bai, H. (Eds.). (2015). Propensity score analysis: Fundamentals and developments. New York, NY: The Guilford Press

About the presenters
Dr. Wei Pan is an Associate Professor and Director of the Research Design and Statistics Core in the School of Nursing at Duke University. His research work focuses on causal inference (propensity score analysis, 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 and edited a book, entitled "Propensity Score Analysis: Fundamentals and Developments." Dr. Pan has also organized and taught many professional development courses on propensity score methods at national conventions of major professional associations.

Dr. Haiyan Bai is an Associate Professor of Quantitative Research Methodology at the University of Central Florida. Her interests include propensity score analysis, resampling methods, research design, measurement and evaluation, and the applications of statistical methods in the educational and behavioral sciences. She has published a book on resampling methods as well as numerous articles in refereed journals, and has served on the editorial boards of several journals. Dr. Bai has also given over 10 professional development courses and organized many workshops on propensity score methods at national and international conferences.

9. Analysis of Large and Complex Networks

Presenter: Zheng Tracy Ke, University of Chicago

Course length: Half day

Outline/Description:
The emergence of large online social networks (e.g., Facebook, LinkedIn) has motivated a great deal of research interest in network data analysis. This short course gives an overview of frontier problems and methods for large networks, especially large social networks.

1. Overview of the area of social networks. Identify interesting scientific problems. Discuss potential scientific impacts.
2. Models/methods/theory. Review static and dynamic network models (with a focus on the stochastic block model and its variants). Overview recent methods for problems such as community detection and link prediction (e.g., the method SCORE for community detection). Discuss theory and mathematical tools.
3. Real data analysis. Showcase and discuss a data set of the coauthorship and citation networks of statisticians. It is based on about 60,000 published papers in 20 top statistics journals in 1980-2015.

References
E. Kolaczyk and C. Gabor (2014). Statistical analysis of network data with R. *Springer*

A. Goldenberg, et al. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*

P. Ji and J. Jin (2016). Coauthorship and Citation Networks for Statisticians. *Annals of Applied Statistics*, to appear
M. Newman (2010). Networks: an introduction. *Oxford University Press*.

About the presenter

Dr. Ke is currently Assistant Professor of University of Chicago, Department of Statistics. In collaboration with Pengsheng Ji (University of Georgia) and Jiashun Jin (Carnegie Mellon University), she is collecting and cleaning a data set of coauthor/citation relationship of statisticians. The project has two phases. In Phase I, Ji and Jin have collected a smaller data set (4 journals in a 10-year period), and the results will appear in Annals of Applied Statistics as a discussion paper. In Phase II, the team will collect a data set 15-20 times larger than that of Phase I.

2016 ICSA Applied Statistics Symposium Career Service

Dear All:

Thank you for registering for the 2016 ICSA applied statistics symposium in Atlanta.

The 2016 symposium will offer short courses, invited sessions, topic contributed, contributed sessions, as well as opportunities for networking and recruiting.

Several firms and recruiting agencies will be on-site during the ICSA symposium. We are pleased to offer free career service during the 2016 ICSA Symposium. You are encouraged to utilize our career service.

If you have an interest in a career in statistics, we encourage you to utilize the ICSA Career Placement Service. This is offered to you free of charge. To get started, please submit your resumes to ICSA Symposium Career Service Coordinator - Haoda Fu, Eli Lilly, fu_haoda@lilly.com by June 1. You may also contact employers of interest who will be onsite during the symposium.

Opening Mixer:

Sunday, June 12, 7-9 PM, Hyatt Regency Atlanta, International Ballroom South

Banquet:

Tuesday June 14, 6:30-10:00 PM, Golden House Restaurant

Address: 金源大酒楼, 1600 Pleasant Hill Rd, Duluth, GA 30096, Phone: 770-921-2228

Buses leave at **5:15PM**, **5:40PM**.

Banquet speaker: Prof. Michael Eriksen (Georgia State University).

Karaoke and dance performance by Atlanta Professional Dance of Academy (APDA), Atlanta.

Cost: $50 per guest, $25 student/junior guest.

Excursion:

A trip to Stone Mountain Park, Wednesday, June 15, 1-6 PM

Departure location and time: Hyatt Regency Atlanta, Wednesday, June 15, 1:00 PM

The bus will leave at 1pm at the entrance of Hyatt Regency Atlanta and the estimated return time from Stone Mountain is 6pm. Tour participation is limited. The trip will include admittance into Stone Mountain. It offers free access to Stone Mountain Park public areas including picnic areas, hiking and nature trails, children's playground, walk-up trail, Quarry Exhibit, the Confederate Hall Historical & Environmental Education Center and the Laser show Spectacular. Other activities such as sunset adventure, Atlanta braves & theme park combo deals, Georgia Aquarium & Stone Mountain Park: Seas to Mountain Combo are at your own expenses.

Price: $30 per guest and $20 per junior guest.

The 10th ICSA International Conference
Shanghai, China, December 19-22, 2016

The 10th ICSA International Conference will be held at Xuhui campus of Shanghai Jiao Tong University (SJTU), Shanghai, China, during December 19-22, 2016. The theme of this conference is to promote global growth of modern statistics in the 21st century. The purpose of this conference is to bring statisticians from all over the world to Shanghai, China, which is the financial, trade, information and shipping center of China, to share cutting-edge research, discuss emerging issues in the field of modern probability and statistics with novel applications, and network with colleagues from all parts of the world.

James O. Berger of Duke University, **Tony Cai** of University of Pennsylvania, **Kai-Tai Fang** of Beijing Normal University – Hong Kong Baptist University United International College (UIC), **Zhi-Ming Ma** of the Academy of Math and Systems Science, CAS, **Marc A. Suchard** of the UCLA Fielding School of Public Health and David Geffen School of Medicine at UCLA, **Lee-Jen Wei** of Harvard University, and **C. F. Jeff Wu** of Georgia Institute of Technology will deliver keynote presentations. There will be a special session in honor of the receipt(s) of the second Pao-Lu Hsu award. In addition, there will be over 200 invited sessions. All participants including invited speakers are responsible for paying registration fees and booking hotel rooms directly from the hotels listed on the conference website.

The scientific program committee of the 2016 ICSA International Conference, is co-chaired by Ming-Hui Chen of University of Connecticut, Zhi Geng of Peking University, and Gang Li of University of California at Los Angeles. For conference logistics, please directly contact Dong Han and Weidong Liu, the co-chairs of the local organizing committee. All inquiries should be sent to Ms. Limin Qin at qinlimin@sjtu.edu.cn.

More detailed information is available at http://www.math.sjtu.edu.cn/conference/2016icsa/.

All of you are welcome to participant in this important ICSA conference and to visit Shanghai, one of beautiful and historical cities in the world during December 19-22, 2016.

Announcement: 2017 ICSA Applied Statistics Symposium

You are cordially invited to the 2017 ICSA Applied Statistics Symposium in Chicago!

The ICSA Applied Statistics Symposium will be held from Sunday, June 25 to Wednesday, June 28, 2017 at the Hilton Chicago Downtown in Chicago, Illinois. Please send any inquiry to Dr. Lanju Zhang (Lanju.zhang@abbvie.com).

The city of Chicago enjoys a beautiful Spring, Summer, and Fall throughout the year, and is easily accessible from most cities across North America. Downtown Chicago provides a great variety of dining, accommodation and shopping, etc. In addition, it offers world-class attractions, including the Art Institute of Chicago, Willis Tower - Skydeck Chicago, the Museum Campus featuring Shedd Aquarium, Field Museum and Adler Planetarium; and the Millennium Park, all within walking distance from the Hilton Chicago. 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 15, 2016.

Call for Student Paper Award Applications

Up to eight student award winners (five Student Travel Awards, one Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper, and possible two ASA Biopharmaceutical Awards) will be selected. Each winner will receive a plaque or certificate, an award for travel and registration reimbursement up to $1,000 or a cash award of $550, whichever is bigger, as well as free registration for a short course. The deadline for applications is March 15, 2017.

Call for Short Course Proposals

Anyone who is interested in giving a one-day or half-day short course at 2017 ICSA Applied Statistics Symposium is welcome to submit a short-course proposal to Dr. Jun Zhao (zhao.jun@abbvie.com). The submission deadline is December 31, 2016.

Executive Committee

- Lanju Zhang, Chair of Organizing Committee, AbbVie Inc
- Xuming He, Program Committee Chair, University of Michigan
- Yuan Ji, Program Book Chair, University of Chicago
- Hongmei Jiang, Local Committee Chair, Northwestern University
- Xuan Liu, Treasurer, AbbVie Inc
- Lingsong Zhang, Student Paper Award Chair, Purdue University
- Jun Zhao, Short Course Chair, AbbVie Inc
- Bo Yang, Fund Raising Chair, Vertex Pharmaceuticals
- Wei Shen, Strategic Advisor

ICSA DINNER at 2016 JSM in Chicago, IL

The ICSA will hold the annual members meeting and award ceremony on August 3 (Wednesday) at 6pm in Waldorf room at Hilton Chicago Hotel during JSM in Chicago, IL. The annual ICSA banquet will follow the members meeting at MingHin Cuisine, 2168 S Archer Ave., Chicago, IL 60616, (312)808-1999, http://www.minghincuisine.com/. MingHin is a Cantonese fusion restaurant located in Chinatown. It is about 1 mile away from McCormick Place. It can be reached via CTA BUS #21 (get on McCormick Place Convention Center and get off at Cermak-Chinatown Red Line station). This restaurant features a cozy setting, superior cuisine and elegant decor. The banquet menu will include MingHin BBQ Combination Platter/ Special Soup of the Day/ Mixed Seafood in Nest/ Salt and Pepper Shrimp/ Beef with Chinese Broccoli/ Pan Fried Chilean Seabass/ Lobster with Ginger and Green Onion/ Japanese Tofu Topped with Mushroom/ Stir-Fried Bok Choy with Garlic/ Beijing Duck (2 course).

Scientific Program (June 13th - June 15th)

Monday, June 13. 8:00 AM - 9:30 AM

Opening Ceremony and Keynote session I (*Keynote*)
Room: International Ballroom
Organizers: ICSA 2016 organizing committee.
Chair: Yichuan Zhao, Conference Chair.

8:00 AM Welcome
Mark Becker, President of the Georgia State University
Mei-Ling Ting Lee, 2016 ICSA President
David Morganstein, 2015 ASA President

8:30 AM Keynote lecture I
Bin Yu. University of California, Berkeley

9:30 AM Floor Discussion.

Monday, June 13. 10:00 AM-11:40 AM

Session 1: Biostatistics in Medical Applications (*Invited*)
Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3
Organizer: Mei-Ling Ting Lee, University of Maryland.
Chair: Benjamin Haaland, Georgia Institute of Technology.

10:00 AM Patient-Centered Pragmatic Clinical Trials: What, Why, When, How?
Sally Morton. University of Pittsburgh

10:25 AM Robust Methods for Treatment Effect Calibration, with Application to Non-Inferiority Trials
♦*Zhiwei Zhang*[1], *Lei Nie*[1], *Guoxing Soon*[1] *and Zonghui Hu*[2]. [1]FDA [2]NIH

10:50 AM A Comparison Study of Fixed and Mixed Models for Gene Level Association Studies of Complex Traits
♦*Ruzong Fan*[1], *Jeesun Jung* [2], *Chi-yang Chiu*[1], *Daniel E. Weeks*[3], *Alexander F. Wilson*[4], *Joan E. Bailey-Wilson*[4] *and Christopher I. Amos*[5]. [1]NICHD, NIH [2]National Institute on Alcohol Abuse and Alcoholism [3]University of Pittsburgh [4]NHGRI, NIH [5] Geisel School of Medicine at Dartmouth

11:15 AM Methods for biomarker combination in presence of missing gold standard
♦*Danping Liu*[1], *Ashok Chaurasia*[2] *and Zheyu Wang*[3]. [1]National Institutes of Health [2]NIH and University of Waterloo [3]Johns Hopkins University

11:40 AM Floor Discussion.

Session 2: Geometric Approaches in Functional Data Analysis (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zhengwu Zhang, Statistical and Applied Mathematical Sciences Institute.
Chair: Sebastian Kurtek, The Ohio State University.

10:00 AM Bayesian uncertainty quantification in partial differential equation models
Prithwish Bhaumik. The University of Texas at Austin

10:25 AM Statistical Analysis of Trajectories on Riemannian Manifolds
♦*Jingyong Su*[1] *and Anuj Srivastava*[2]. [1]Texas Tech University [2]Florida State University

10:50 AM Fast Functional Genome Wide Association Analysis of Surface-based Imaging Genetic Data
♦*Chao Huang and Hongtu Zhu*. Dept. Biostatistics/ UNC at Chapel Hill

11:15 AM Shape Analysis of Trees Using Elastic Curve Geometry and Side Branch Permutation
♦*Adam Duncan*[1], *Eric Klassen*[2] *and Anuj Srivastava*[1]. [1]Florida State University, Dept. of Statistics [2]Florida State University, Dept. of Mathematics

11:40 AM Floor Discussion.

Session 3: Advance in Statistical Genomics and Computational Biology (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zuoheng Wang, Yale University.
Chair: Zuoheng Wang, Yale University.

10:00 AM A Novel and Efficient Study Design to Test Parent-of-Origin Effects in Family Trios
♦*Rui Feng and Xiaobo Yu*. University of Pennsylvania

10:25 AM A Mathematical Population Genetics Model for Cancer Gene Detection
Amei Amei. University of Nevada, Las Vegas

10:50 AM Machine learning application in gene regulation and its high performance computing on GPU platform
♦*Zhong Wang*[1], *Lauren Choate*[2], *Tinyi Chu*[3] *and Charles Danko*[1]. [1]College of Veterinary Medicine, Cornell University [2]Molecular Biology and Genetics, Cornell University [3]Computational Biology, Cornell University

11:15 AM Impact of genotyping errors on statistical power of association tests in genomic analyses
Lin Hou. Tsinghua University

11:40 AM Floor Discussion.

Session 4: New Developments in Biomedical Research and Statistical Genetics (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizers: Qingxia Chen, Vanderbilt University; Lihong Qi, University of California, Davis.
Chair: Lihong Qi, University of California, Davis.

10:00 AM Treatment Effect Estimate and Model Diagnostics with Two-way Time-Varying Treatment Switching
♦*Qingxia Chen*[1], *Fan Zhang*[2], *Ming-Hui Chen*[2] *and Xiuyu Cong*[3]. [1]Vanderbilt University [2]University of Connecticut [3]Boehringer Ingelheim Pharmaceuticals

10:25 AM Detecting regulatory relationships between DNA alterations and gene/protein expressions
♦*Jie Peng*[1]*, Chris Conley* [1] *and Pei Wang*[2]. [1]UC Davis [2]Icahn School of Medicine at Mount Sinai

10:50 AM Pre-conditioning method for risk prediction with application to ED acute heart failure patients
♦*Dandan Liu, Cathy Jenkins, Sean Collins, Alan Storrow and Frank Harrell*. Vanderbilt University Medical Center

11:15 AM On Nonsmooth Estimating Functions via Jackknife Empirical Likelihood
Jinfeng Xu. University of Hong Kong

11:40 AM Floor Discussion.

Session 5: Change-Point Problems and their Applications (III) (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizers: Jie Chen, Augusta University; Yajun Mei, Georgia Institute of Technology.
Chair: Yajun Mei, Georgia Institute of Technology.

10:00 AM Sequential surveillance of structural breaks in firms' credit rating migrations
♦*Haipeng Xing and Ke Wang*. Stony Brook University

10:25 AM Detection of Changes monitored at Random Time Points
Marlo Brown. Niagara University

10:50 AM 1 D-ary Sequential Tests of Circular Error Probability
♦*Yan Li*[1] *and Yajun Mei*[2]. [1]East China Normal University [2]Georgia Institute of Technology

11:15 AM Sequential event detection in networked Hawkes process
Shuang Li, ♦*Yao Xie, Mehrdad Farajtabar and Le Song*. Georgia Institute of Technology

11:40 AM Floor Discussion.

Session 6: Statistical and Computational Analysis of High-Throughput RNA Sequencing Data (*Invited*)
Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Liang Chen, University of Southern California.
Chair: Hongkai Ji, John Hopkins University.

10:00 AM On the correlation analysis of RNA-seq data
Yinglei Lai. The George Washington University

10:25 AM MSIQ: JOINT MODELING OF MULTIPLE RNA-SEQ SAMPLES FOR ACCURATE ISOFORM QUANTIFICATION
Wei Vivian Li[1]*, Anqi Zhao*[2]*, Shihua Zhang*[3] *and* ♦*Jingyi Jessica Li*[1]. [1]Department of Statistics, UCLA [2]Department of Statistics, Harvard University [3]Chinese Academy of Sciences

10:50 AM Statistical Models for Single Cell RNA Sequencing
Cheng Jia, Yuchao Jiang, Mingyao Li and ♦*Nancy Zhang*. University of Pennsylvania

11:15 AM Robust statistical analysis of RNA-seq data
Maoqi Xu and ♦*Liang Chen*. University of Southern California

11:40 AM Floor Discussion.

Session 7: Emerging Statistical Methods for Longitudinal Data (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Lan Xue, Oregon State University.
Chair: Lily Wang, Iowa State University.

10:00 AM Functional Multiple Indicators, Multiple Causes Measurement Error Models
♦*Carmen Tekwe, Roger Zoh, Fuller Bazer and Raymond Carroll*. Texas A&M University

10:25 AM Efficient quantile marginal regression for longitudinal data with dropouts
♦*Hyunkeun Cho*[1]*, Hyokyoung Hong*[2] *and Mi-Ok Kim*[3]. [1]Western Michigan University [2]Michigan State University [3]Cincinnati Children's Hospital Medical Center

10:50 AM Partially Linear Additive Quantile Regression in Ultra-high Dimension
Ben Sherwood[1] *and* ♦*Lan Wang*[2]. [1]Johns Hopkins University [2]University of Minnesota

11:15 AM Structural Nonparametric Methods for Estimation, Prediction and Tracking with Longitudinal Data
♦*Colin Wu and Xin Tian*. National Heart, Lung and Blood Institute, NIH

11:40 AM Floor Discussion.

Session 8: Empirical Bayes, Methods and Applications (*Invited*)
Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)
Organizer: Linda Zhao, University of Pennsylvania.
Chair: Wenjiang Fu, University of Houston.

10:00 AM Empirical Bayes methods and nonparametric mixture models via nonparametric maximum likelihood
♦*Lee Dicker*[1]*, Sihai Zhao*[2] *and Long Feng*[1]. [1]Rutgers [2]UIUC

10:25 AM Block-Linear Empirical Bayes Estimation of a Heteroscedastic Normal Mean
♦*Asaf Weinstein*[1]*, Zhuang Ma*[2]*, Lawrence Brown*[2] *and Cunhui Zhang*[3]. [1]Stanford University [2]University of Pennsylvania [3]Rutgers University

10:50 AM Nonparametric empirical Bayes approach to integrative high-dimensional classification
Sihai Zhao. University of Illinois at Urbana-Champaign

11:15 AM Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective
♦*Roger Koenker*[1] *and Jiaying Gu*[2]. [1]U. of Illinois [2]U. of Toronto

11:40 AM Floor Discussion.

Session 9: Bayesian Methods for Complex Data. (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Debdeep Pat, Florida State University.
Chair: Jing Zhang, Georgia State University.

10:00 AM Bayesian Neural Networks for Personalized Medicine
Faming Liang. University of Florida

10:25 AM Bayesian Regression Trees for High Dimensional Prediction and Variable Selection
Antonio Linero. Florida State University

10:50 AM Novel Statistical Frameworks for Analysis of Structured Sequential Data
♦*Abhra Sarkar and David Dunson*. Duke University

11:15 AM Bayesian Multiple Classification with Frequent Pattern Mining
♦*Wensong Wu and Tan Li*. Florida International University

11:40 AM Floor Discussion.

Session 10: Statistics and its Applications (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Marianthi Markatou, University at Buffalo.
Chair: Jason Liao, Merck.

10:00 AM Cardiovascular clinical trials in the 21st century: Pros and Cons of an inexpensive paradigm
Nancy Geller. National Heart, Lung, and Blood Institute, NIH

10:25 AM Optimality of Training/Test Size and Resampling Effectiveness in Cross-Validation
♦*Georgios Afendras and Marianthi Markatou*. SUNY at Buffalo

10:50 AM A Zero-inflated Poisson Model for Species Quantification Based on Shotgun Metagenomic Data
Tony Cai, Hongzhe Li and ♦*Jing Ma*. University of Pennsylvania

11:15 AM Floor Discussion.

Session 11: Recent Developments of High-Dimensional Hypothesis Testing. (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Jun Li, Kent State University.
Chair: Yuan Jiang, Oregon State University.

10:00 AM CONDITIONAL MEAN AND QUANTILE DEPENDENCE TESTING IN HIGH DIMENSION
♦*Xianyang Zhang*[1], *Shun Yao*[2] *and Xiaofeng Shao*[2]. [1]Texas A&M University [2]University of Illinois at Urbana-Champaign

10:25 AM Principal Component based Adaptive-weight Burden Test for Quantitative Trait Associations
♦*Xiaowei Wu*[1] *and Dipankar Bandyopadhyay*[2]. [1]Virginia Tech [2]Virginia Commonwealth University

10:50 AM Homogeneity Test of Covariance Matrices and Change-Points Identification with High-Dimensional Longitudinal Data
♦*Pingshou Zhong*[1] *and Runze Li* [2]. [1]Michigan State University [2]Penn State University

11:15 AM A neighborhood-assisted test for high-dimensional mean vector
♦*Jun Li*[1], *Yumou Qiu*[2] *and Song Xi Chen*[3]. [1]Kent State University [2]University of Nebraska-Lincoln [3]Iowa State University

11:40 AM Floor Discussion.

Session 12: Advanced Methodologies in Analyzing Censoring Data (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Jiajia Zhang, University of South Carolina.
Chair: Jiajia Zhang, University of South Carolina.

10:00 AM Jackknife Empirical Likelihood for Linear Transformation Models with Censored Data
Hanfang Yang [1], *Shen Liu*[1] *and* ♦*Yichuan Zhao*[2]. [1]Renmin University of China [2]Georgia State University

10:25 AM Quantile Residual Life Regression with Longitudinal Biomarker Measurements for Dynamic Prediction
♦*Ruosha Li*[1], *Xuelin Huang*[2] *and Jorge Cortes*[2]. [1]Univ of Texas Health Science Center at Houston [2] The University of Texas MD Anderson Cancer Center

10:50 AM Promotion time cure model with nonparametric form of covariate effects
Pang Du[1] *and* ♦*Tianlei Chen*[2]. [1]Virginia Tech [2]Celgene

11:15 AM Bayesian semiparametric regression models for interval-censored data
♦*Xiaoyan Lin, Lianming Wang and Bo Cai*. University of South Carolina

11:40 AM Floor Discussion.

Session 13: Recent Advancement in Adaptive Design of Early Phase Clinical Trials by Accounting for Schedule Effects or Using Other Approaches (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yisheng Li, The University of Texas MD Anderson Cancer Center.
Chair: Tu Xu, AbbVie.

10:00 AM A Subgroup Cluster Based Bayesian Adaptive Design for Precision Medicine
Wentian Guo[1], ♦*Yuan Ji*[2] *and Daniel Catenacci*[3]. [1]Fudan University, Shanghai, China [2]Northshore University/University of Chicago [3]University of Chicago Medical Center

10:25 AM Phase I design for locating schedule-specific maximum tolerated doses
Nolan Wages. University Of Virginia

10:50 AM TITE-CRM method incorporating cyclical safety data with application to oncology phase I trials
Bo Huang. Pfizer

11:15 AM A dose-schedule-finding design for phase I/II clinical trials
Beibei Guo[1], ♦*Yisheng Li*[2] *and Ying Yuan*[2]. [1]Louisiana State University [2]University of Texas MD Anderson Cancer Center

11:40 AM Floor Discussion.

Session 14: Contemporary Statistical Methods for Complex Data (*Invited*)
Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3
Organizer: Li-Shan Huang, National Tsing Hua University.
Chair: Chunming Zhang, University of Wisconsin.

10:00 AM Bridging density functional theory and big data analytics with applications
Chien-Chang Chen[1]*, Hung-Hui Juan*[2]*, Meng-Yuan Tsai*[2] *and* ♦*Henry Horng-Shing Lu*[2]. [1]National Central University, Taiwan [2]National Chiao Tung University, Taiwan

10:25 AM Nonparametric divergence-based flow cytometric classification
♦*Ollivier Hyrien and Andrea Baran*. University of Rochester

10:50 AM Untangle the Structural and Random Zeros in Statistical Modelling
♦*Hua He*[1]*, Wan Tang*[2]*, Wenjuan Wang*[3]*, Naiji Lu*[4] *and Ding-Geng Chen*[5]. [1]Tulane University [2]Tulane University [3]Brightech International, LLC [4]Huazhong University of Science and Technology [5]University of North Carolina, Chapel Hill

11:15 AM Covariance Structures and Estimation for Axially Symmetric Spatial Processes on the Sphere
♦*Haimeng Zhang*[1]*, Chunfeng Huang*[2] *and Scott Robeson*[2]. [1]University of North Carolina - Greensboro [2]Indiana University - Bloomington

11:40 AM Floor Discussion.

Session 15: Recent Development in Time-to-Event Data Analysis (*Invited*)
Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3
Organizers: Qing Yang, Duke University; Gang Li, University of California, Los Angeles.
Chair: Ying Ding, University of Pittsburgh.

10:00 AM A NPMLE Approach for Extended Cure Rate Model with Left Truncation and Right-Censoring
♦*Jue Hou and Ronghui Xu*. University of California, San Diego

10:25 AM A Semiparametric Joint Model for Longitudinal and Survival Data in End-of-Life Studies
♦*Zhigang Li*[1]*, HR Frost*[1]*, Tor Tosteson*[1]*, Lihui Zhao*[2]*, Lei Liu*[2]*, Huaihou Chen*[3] *and Marie Bakitas*[4]. [1]Dartmouth College [2]Northwestern University [3]University of Florida [4]The University of Alabama at Birmingham

10:50 AM Group variable selection in survival and competing risks model
♦*Kwang Woo Ahn, Natasha Sahr, Anjishnu Banerjee and Soyoung Kim*. Medical College of Wisconsin

11:15 AM Sample Size for Joint Testing Cause-Specific Hazard and Overall Hazard in the Presence of Competing
♦*Qing Yang*[1]*, Wing Kam Fung*[2] *and Gang Li*[3]. [1]Duke University [2]University of Hong Kong [3]University of California, Los Angeles

11:40 AM Floor Discussion.

Session 16: Statistics and Big Data (*Invited*)
Room: BAKER, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yichuan Zhao, Georgia State University.
Chair: Xiaoyi Min, Georgia State University.

10:00 AM Jackknife empirical likelihood inference for AFT models
♦*Xue Yu and Yichuan Zhao*. Georgis State University

10:25 AM Jackknife Empirical Likelihood for the Concordance Correlation Coefficient
♦*Anna Moss and Yichuan Zhao*. Georgia State University

10:50 AM Jackknife Empirical Likelihood for the Mean Difference of Two Zero Inflated Skewed Populations
♦*Faysal Satter and Yichuan Zhao*. Georgia State University

11:15 AM Rank-based estimating equation with non-ignorable missing responses under empirical likelihood
♦*Huybrechts F Bindele*[1] *and Yichuan Zhao*[2]. [1]University of South Alabama [2]Georgia State University

11:40 AM Floor Discussion.

Session 17: Recent Advances in Design Problems (*Contributed*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Chair: Eugene (Yijian) Huang, Emory University.

10:00 AM Considerations for Pediatric Trial Designs and Analyses
Meehyung Cho, ♦*Zhiying Qiu, Jenny Ye, Hui Quan and Peng-Liang Zhao*. Sanofi US

10:15 AM Dose-finding Designs Incorporating Toxicity and Efficacy
♦*Jun Yin*[1]*, Monia Ezzalfani*[2]*, Dan Sargent*[1] *and Sumithra Mandrekar*[1]. [1]Mayo Clinic [2]Institut Curie

10:30 AM Statistical Considerations in the Design and Analysis of Reference Database for OCT Device
Haiwen Shi. FDA/CDRH

10:45 AM Design Considerations for Non-randomized Medical Device Clinical Studies
♦*Heng Li, Vandana Mukhi and Yun-Ling Xu*. FDA/CDRH

11:00 AM Bayesian Analysis of Disease Modification using Doubly-Randomized Delay-Start Matched Control Design
♦*Ibrahim Turkoz*[1] *and Marc Sobel*[2]. [1]Janssen Research&Development, LLC [2]Temple University

11:15 AM A prospective-retrospective study design to access the clinical performance of an in-vitro diagnosti
Shiling Ruan. Allergan Plc

11:30 AM Floor Discussion.

Monday, June 13. 1:30 PM - 3:10 PM

Session 18: New Statistical Computing using R (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Mei-Ling Ting Lee, University of Maryland.
Chair: Ruzong Fan, NIH/NICHD.

1:30 PM Rank-Based Tests for Clustered Data with R Package clusrank
♦*Yujing Jiang*[1]*, Jun Yan*[1] *and Mei-Ling Ting Lee*[2]. [1]University of Connecticut [2]University of Maryland

1:55 PM The R Package "threg" to Implement Threshold Regression: A model for time-to-event survival data
Mei-Ling Ting Lee. University of Maryland

2:20 PM Threshold Regression for Interval-Censored Data for Cure-Rate or without Cure-Rate Model
♦*Man-Hua Chen*[1] *and Mei-Ling Ting Lee*[2]. [1]Tamkang University [2]University of Maryland

2:45 PM Discussant: Jing Zhang, University of Maryland

3:10 PM Floor Discussion.

Session 19: Statistical Modeling and Inference on Complex Biomedical Data (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Lynn Lin, Pennsylvania State University.
Chair: Lynn Lin, Pennsylvania State University.

1:30 PM A Generalized Estimating Equations Framework for the Analysis of Intracellaur Cytokine Staining Data
♦*Amit Meir*[1], *Raphael Gottardo*[2] *and Greg Finak*[2]. [1]University of Washington [2]Hutch Cancer Research Center

1:55 PM Haplotyping and quantifying allele-specific expression at gene and gene isoform level by Hybrid-Seq
Benjamin Deonovic[1], *Yunhao Wang*[2], *Jason Weirather*[3] *and* ♦*Kin Fai Au*[3]. [1]Department of Biostatistics, University of Iowa [2]University of Chinese Academy of Sciences [3]Department of Internal Medicine, Univ. of Iowa

2:20 PM Intrinsic Noise in Nonlinear Gene Regulation Inference
♦*Chao Du*[1] *and Wing.H. Wong*[2]. [1]University of Virginia [2]Stanford University

2:45 PM A Kernel-Based Approach to Covariate Adjustment for Causal Inference
♦*Yeying Zhu*[1], *Jennifer Savage*[2] *and Debashis Ghosh*[3]. [1]University of Waterloo [2]Pennsylvania State University [3]Colorado School of Public Health

3:10 PM Floor Discussion.

Session 20: Statistical Advances in Omics Data Integration (*Invited*)
Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Cen Wu, Kansas State University.
Chair: Feifei Xiao, University of South Carolina.

1:30 PM Genomic Determination Index
♦*Cheng Cheng, Wenjian Yang, Robert Autry, Steven Paugh and William Evans*. St. Jude Children's Research Hospital

1:55 PM graph-GPA: A graphical model to prioritizing GWAS results by integrating pleiotropy
♦*Dongjun Chung*[1], *Hang Kim*[2] *and Hongyu Zhao*[3]. [1]Medical University of South Carolina [2]University of Cincinnati [3]Yale University

2:20 PM Pathway based integrative study of omics data for predicting cancer prognosis in TCGA melanoma
♦*Yu Jiang*[1], *Xingjie Shi*[2], *Qing Zhao*[3], *Cen Wu*[4] *and Shuangge Ma*[5]. [1]University of Memphis [2]Nanjing University of Finance and Economics [3]Merck Research Laboratories [4]Kansas State University [5]Yale University

2:45 PM Joint Precision Matrix Estimation with Sign Consistency
♦*Yuan Huang, Qingzhao Zhang and Shuangge Ma*. Yale University

3:10 PM Floor Discussion.

Session 21: Statistical Preprocessing of Deep Sequencing Data (*Invited*)
Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)
Organizers: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center; Xiangqin Cui, University of Alabama at Birmingham.
Chair: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.

1:30 PM FACETS: Cellular Fraction and Copy Number Estimates from Tumor Sequencing
♦*Venkatraman Seshan and Ronglai Shen*. MSKCC

1:55 PM Normalization issues in single cell RNA sequencing
♦*Zhijin Wu*[1] *and Hao Wu*[2]. [1]Brown University [2]Emory University

2:20 PM Preprocessing issues with epigenetic assays based on sequencing
Kasper Hansen. Johns Hopkins University

2:45 PM Exploring Immune Repertoire Sequencing Data
♦*Xiangqin Cui, Amanda Hall and Roslyn Mannon*. University of Alabama at Birmingham

3:10 PM Floor Discussion.

Session 22: Change-Point Problems and their Applications (I) (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizers: Jie Chen, Augusta University; Yajun Mei, Georgia Institute of Technology.
Chair: Jie Chen, Augusta University.

1:30 PM Jump Information Criterion for Estimating Jump Regression Curves
Peihua Qiu. University of Florida

1:55 PM Monitoring Sparse Contingency Table in Multivariate Categorical Process
Dongdong Xiang. East China Normal University

2:20 PM Modeling the Next Generation Sequencing Read Count Data for DNA Copy Number Variants Study
♦*Tieming Ji*[1] *and Jie Chen*[2]. [1]University of Missouri at Columbia [2]Augusta University

2:45 PM Changepoint Detection in Categorical Time Series with Application to Hourly Sky-cloudiness Condition
♦*QiQi Lu*[1] *and Xiaolan Wang*[2]. [1]Virginia Commonwealth University [2]Environment and Climate Change Canada

3:10 PM Floor Discussion.

Session 23: Order-restricted Statistical Inference and Applications (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Heng Wang, Michigan State University.
Chair: Ping-Shou Zhong, Michigan State University.

1:30 PM Variable and Shape Selection using the Cone Information Criterion
♦*Mary Meyer and Xiyue Liao*. Colorado State University

1:55 PM Constrained Statistical Inference in Linear Mixed Models with Applications
♦*Casey Jelsema*[1] *and Shyamal Peddada*[2]. [1]West Virginia University [2]National Institute of Environmental Health Science

2:20 PM Testing for uniform stochastic orderings via empirical likelihood under right censoring
Hammou ELBARMI. Baruch College, The City University of New York

2:45 PM A test of order-restricted means for data with imputation
♦*Heng Wang and Ping-Shou Zhong*. Michigan State University

3:10 PM Floor Discussion.

Session 24: Recent Advances of the Statistical Research on Complicated Structure Data (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Tao Yu, National University of Singapore.
Chair: Yaji Xu, Food and Drug Administration.

1:30 PM Integrative analysis of datasets with different resolutions reveals consistent genetic effects
Yuan Jiang. Oregon State University

1:55 PM False discovery rate estimation with covariates
Kun Liang. University of Waterloo

2:20 PM Using a monotonic density ratio model to find the asymptotically optimal combination of multiple dia
♦*Baojiang Chen*[1]*, Pengfei Li*[2]*, Jing Qin*[3] *and Tao Yu*[4]. [1]University of Nebraska Medical Center [2]University of Waterloo [3]NIH [4]National University of Singapore

2:45 PM Variable selection in the presence of nonignorable missing data
Jiwei Zhao. State University of New York at Buffalo

3:10 PM Floor Discussion.

Session 25: Innovative Methods for Modeling and Inference with Survival Data (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ruosha Li, The University of Texas School of Public Health.
Chair: Lan Wang, University of Minnesota.

1:30 PM Statistical Inference on length-biased data with semiparametric accelerated failure time models
♦*Jing Ning*[1]*, Jing Qin*[2] *and Yu Shen*[1]. [1]The University of Texas MD Anderson Cancer [2]National Institutes of Health

1:55 PM Association analysis of gap times with multiple causes
Xiaotian Chen, ♦*Yu Cheng, Ellen Frank and David Kupfer*. University of Pittsburgh

2:20 PM Restoration of Monotonicity Respecting in Dynamic Regression
Yijian Huang. Emory University

2:45 PM A General Semiparametric Accelerated Failure Time Model Imputation Approach for Censored Covariate
♦*Ying Ding*[1]*, Shengchun Kong*[2] *and Shan Kang*[3]. [1]University of Pittsburgh [2]Purdue University [3]Robert Bosch LLC, Research and Technology Center

3:10 PM Floor Discussion.

Session 26: Nonparametric Methods for Neural Spike Train Data (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Wei Wu, Florida State University.
Chair: Wei Wu, Florida State University.

1:30 PM Calibrating nonparametric inference of monosynaptic connections from spike train recordings
♦*Asohan Amarasingham and Jonathan Platkiewicz*. City University of New York

1:55 PM Nonparametric Methods for Decoding Rat Hippocampal Neuronal Ensemble Spikes
Zhe Chen. New York University School of Medicine

2:20 PM Receptive field models of multiunit activity and the decoding of hippocampal replay events
Uri Eden. Boston University

2:45 PM Nonparametric discriminative filtering for neural decoding
♦*Michael C. Burkhart, David M. Brandman, Carlos Vargas-Irwin and Matthew T. Harrison*. Brown University

3:10 PM Floor Discussion.

Session 27: Subgroup Identification/Analysis in Clinical Trials (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Jianxiong(George) Chu, Food and Drug Administration(FDA).
Chair: Rajesh Nair, FDA.

1:30 PM Assessing consistency of treatment effect across subgroups in a large randomized study on dual anti-platelet therapy
Joseph Massaro. Boston University/Harvard Clinical Research Inst.

1:55 PM Issues in Subgroup Analyses
♦*Weishi Yuan, Kun He and Rajeshwari Sridhara*. FDA

2:20 PM Subgroup Analysis: Issues and Possible Improvement
♦*Lu Cui and Tu Xu*. AbbVie

2:45 PM Discussion: Led by Chenguang Wang, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

3:10 PM Floor Discussion.

Session 28: Model Selection/Averaging and Objective Assessment. (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yuhong Yang, University of Minnesota.
Chair: Shuhui Yu , National University of Kaohsiung .

1:30 PM On model selection from a finite family of possibly misspecified models
Ching-Kang Ing. Institute of Statistical Science, Academia Sinica

1:55 PM Model selection confidence sets by likelihood ratio testing
Davide Ferrari. University of Melbourne

2:20 PM Kernel Estimation and Model Combination in a Bandit Problem with Covariates
♦*Wei Qian*[1] *and Yuhong Yang*[2]. [1]Rochester Institute of Technology [2]University of Minnesota

2:45 PM Cross Assessment Towards a More Reproducible Model Selection
Yuhong Yang. University of Minnesota

3:10 PM Floor Discussion.

Session 29: Advances and Applications in Methods for Comparative Effectiveness Research (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Li Li, Eli Lilly and Company.
Chair: Yu Kong, Eli Lilly and Company.

1:30 PM An overview of methods for controlling unmeasured confounders in comparative observational research
♦*Xiang Zhang and Douglas Faries*. Eli Lilly and Company

1:55 PM Generalized propensity score matching with multiple treatments: An application
♦*Zhanglin Cui, Lisa Hess, Robert Goodloe, Gebra Carter and Douglas Faries*. Eli Lilly and Company

2:20 PM Considerations of the Appropriate Number of Strata in Stratified Randomized Clinical Trials
♦*Bo Fu, Su Chen, Yao Li, Ziqian Geng, Yijie Zhou, Lu Cui and Lois Larsen*. AbbVie

2:45 PM Discussant: Carol Yen-Chin Lin, CDC

3:10 PM Floor Discussion.

Session 30: Statistical Genetics (*Invited*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Organizer: Xiaowei Wu, Virginia Tech.
Chair: Xiaowei Wu, Virginia Tech.

1:30 PM Integrating competing but complementary genetic association tests derived from the same data
Lei Sun. University of Toronto

1:55 PM Detecting Schizophrenia Genes via a Two-Sample Test for High-Dimensional Covariance Matrices
♦*Lingxue Zhu*[1]*, Jing Lei*[1]*, Bernie Devlin*[2] *and Kathryn Roeder*[1]. [1]Carnegie Mellon University [2]University of Pittsburgh

2:20 PM Kernel machine association testing for longitudinally-measured quantitative phenotypes
♦*Zuoheng Wang*[1]*, Zhong Wang*[2]*, Xinyu Zhang*[1] *and Ke Xu*[1]. [1]Yale University [2]Cornell University

2:45 PM Adaptive genetic association analysis of multiple traits adjusting for population structure
Duo Jiang. Oregon State University

3:10 PM Floor Discussion.

Session 31: Recent Advances in Statistical Methods for Challenging Problems in Neuroimaging Applications (*Invited*)
Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ying Guo, Emory University.
Chair: Yikai Wang, Emory University.

1:30 PM Statistical method for neuron network recovery
Chunming Zhang. University of Wisconsin-Madison

1:55 PM Community Detection and Clustering via G-models with an Application to fMRI
Florentina Bunea[1]*, Christophe Giraud*[2] *and* ♦*Xi Luo*[3]. [1]Cornell University [2]Universite Paris Sud [3]Brown University

2:20 PM Parsimonious tensor response regression with applications to neuroimaging analysis
Xin Zhang. Florida State University

2:45 PM A Distributional Independent Component Analysis approach for fMRI data
♦*Subhadip Pal*[1]*, Ying Guo*[1] *and Jian Kang*[2]. [1]Emory University [2]University of Michigan

3:10 PM Floor Discussion.

Session 32: Advances in Pharmacogenomics and Biomarker Development (*Invited*)
Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3
Organizer: Hui-Rong Qian, Eli Lilly and Company.
Chair: Alan Chiang, Eli Lilly and Company.

1:30 PM A Few statistical issues in determining sample's positivity using Flow Cytometry and ELISpot assay
Shuguang Huang. Stat4ward LLC

1:55 PM Quantitative Reproducibility Analysis for Identifying Reproducible Targets from High-Throughput Expe
♦*Wenfei Zhang*[1]*, Ying Liu*[2] *and Yuefeng Lu*[1]. [1]Sanofi [2] Sanofi

2:20 PM Region based approach with guided weights for Illumina 450K BeadChip methylation analysis
♦*Yushi Liu, Chunlao Tang and James Scherschel*. Eli Lilly and Company

2:45 PM Advances and issues in RNAseq expression data analysis
♦*Hui-Rong Qian, Phil Ebert and John Calley*. Eli Lilly and Company

3:10 PM Floor Discussion.

Session 33: Model Assessment for Complex Dependence Structure (*Invited*)
Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zhengyuan Zhu, Iowa State University.
Chair: Lily Wang, Iowa State University.

1:30 PM Empirical likelihood tests for alternative spatial dependence structures in Markov random fields
♦*Mark Kaiser*[1]*, Daniel Nordman*[2] *and Yeon-jung Seo*[1]. [1]Iowa State University [2]Iowa State Univesity

1:55 PM Global patterns of lightning properties using spatial point process models on a sphere
Mikyoung Jun. Texas A&M University

2:20 PM A sparse areal mixed model for multivariate outcomes
John Hughes. University of Minnesota

2:45 PM Discussant: Zhengyuan Zhu, Iowa State University

3:10 PM Floor Discussion.

Session 34: Semiparametric Methods in Biostatistics (*Invited*)
Room: BAKER, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yanyuan Ma, University of South Carolina.
Chair: Wenhui Sheng, University of West Georgia.

1:30 PM Generalized Accelerated Failure Time Spatial Frailty Model for Arbitrarily Censored Data
Haiming Zhou[1], *Timothy Hanson*[2] *and* ♦*Jiajia Zhang*[2]. [1]Northern Illinois University [2]University of South Carolina

1:55 PM Graphical Modeling of Biological Pathways in Genome-wide Association Studies
♦*Yujing Cao and Min Chen*. University of Texas at Dallas

2:20 PM Semiparametric regression analysis for multiple-disease group testing data
♦*Dewei Wang, Peijie Hou and Joshua Tebbs*. University of South Carolina

2:45 PM Penalized spline mixed effects model with random time shift; an application to labor curves
Caroline Mulatya and ♦*Alexander McLain*. University of South Carolina

3:10 PM Floor Discussion.

Monday June 13. 3:30 PM - 5:10 PM

Session 35: Recent Research of Omics Data by Young Investigators (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ruzong Fan, NIH/NICHD.
Chair: Wei Chen, Children's Hospital of Pittsburgh at The University of Pittsburgh.

3:30 PM GMMAT: logistic mixed models to control for population stratification and relatedness in genetic ass
♦*Han Chen*[1], *Chaolong Wang*[2], *Matthew Conomos*[3], *Adrienne Stilp*[3], *Cathy Laurie*[3], *Ken Rice*[3] *and Xihong Lin*[1]. [1]Harvard T.H. Chan School of Public Health [2]Genome Institute of Singapore [3]University of Washington

3:55 PM Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples
Qi Yan. University of Pittsburgh

4:20 PM Meta-analysis of Quantitative Pleiotropic Traits at Gene Level with Multivariate Functional Linear M
♦*Chi-Yang Chiu and Ruzong Fan*. NIH

4:45 PM Floor Discussion.

Session 36: New Methods with Large and Complex Data (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Cheng Yong Tang, Temple University.
Chair: Yisheng Li, The University of Texas MD Anderson Cancer Center.

3:30 PM Blessing of Massive Scale: A Total Cardinality Constrained Approach for Spatial Graphical Model
♦*Ethan Fang, Han Liu and Mengdi Wang*. Princeton University

3:55 PM Trace pursuit for model-free variable selection with matrix-valued predictors
Yuexiao Dong. Temple University

4:20 PM Testing independence with high-dimensional correlated samples
♦*Xi Chen* [1] *and Weidong Liu* [2]. [1]New York University [2]Shanghai Jiaotong University

4:45 PM Precision Matrix Estimation by Inverse Principal Orthogonal Decomposition
♦*Cheng Yong Tang*[1] *and Yingying Fan*[2]. [1]Temple University [2]University of Southern California

5:10 PM Floor Discussion.

Session 37: Can Linearly Dependent Confounders Be Estimated? ¨C The Case of Age-Period-Cohort and Beyond (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Wenjiang Fu, University of Houston.
Chair: Shuangge Ma, Yale University.

3:30 PM The Great Society, Reagan's Revolution, and Generations of Presidential Voting
♦*Yair Ghitza*[1] *and Andrew Gelman*[2]. [1]Catalist [2]Columbia University

3:55 PM Confusions about the APC confounding. What have we missed? How can we do better?
Wenjiang Fu. Department of Mathematics, University of Houston

4:20 PM Bias Correction in Modeling Complex Rate Data – How and Why?
Martina Fu. Stanford University

4:45 PM Alternative Approach to the Identifiability Problem–Finding the Truth through Smoothing
♦*Shujiao Huang and Wenjiang Fu*. University of Houston

5:10 PM Floor Discussion.

Session 38: Challenges and Methods in Biomarker Research and Medical Diagnosis (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Danping Liu, NIH/NICHD.
Chair: Sandra Safo, Emory University.

3:30 PM Regulatory Perspective and Case Studies on Biomarker Validation of Companion Diagnostics
♦*Jingjing Ye and Gene Pennello*. FDA

3:55 PM Better Use of Family History Data to Predict Breast Cancer Risk
♦*Shanshan Zhao*[1], *Yue Jiang*[2] *and Clarice Weinberg*[1]. [1]National Institute of Environmental Health Science [2]University of North Carolina at Chapel Hill

4:20 PM Prediction of longitudinal biomarkers on recurrence events in the presence of a terminal event

♦*Ming Wang, Cong Xu and Vern M. Chinchilli*. Penn State Hershey Medical Center

4:45 PM Estimation of Diagnostic Accuracy of a Biomarker When the Gold Standard Is Measured with Error

Mixia Wu[1]*, Dianchen Zhang*[1] *and* ♦*Aiyi Liu*[2]. [1]Beijing University of Technolog [2]NICHD/NIH

5:10 PM Floor Discussion.

Session 39: Change-Point Problems and their Applications (II) (*Invited*)

Room: VININGS, ATLANTA CONFERENCE CENTER - LL3

Organizers: Jie Chen, Augusta University; Yajun Mei, Georgia Institute of Technology.

Chair: Yajun Mei, Georgia Institute of Technology.

3:30 PM Majority versus Consensus Decision Making in Decentralized Sequential Change Detection.

♦*Georgios Fellouris and Sourabh Banerjee*. University of Illinois, Urbana-Champaign

3:55 PM On robustness of N-CUSUM stopping rule in a Wiener disorder problem

♦*Hongzhong Zhang*[1]*, Neofytos Rodosthenou*[2] *and Olympia Hadjiliadis*[3]. [1]Columbia University [2]Queen Mary University of London [3]City University of New York

4:20 PM On the Optimality of Bayesian Change-Point Dtection

♦*Dong Han*[1]*, Fugee Tsung*[2] *and Jinguo Xian*[1]. [1]Dept. of Statistics, Shanghai Jiao Tong Univ. [2]Dept. of IELM, Hong Kong Univ. of Sc. & Techn.

4:45 PM Estimating the Number of States in Hidden Markov Models via Marginal Likelihood

Yang Chen[1]*, Cheng-Der Fuh*[2]*,* ♦*Chu-Lan Kao*[2] *and Samuel Kou*[1]. [1]Harvard University [2]National Central University

5:10 PM Floor Discussion.

Session 40: Statistical Issues in Analysis and Interpretation of Human Drug Abuse Study Data (*Topic Contributed*)

Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3

Organizer: Wei Liu, US Food and Drug Administration.

Chair: Qianyu Dang, US Food and Drug Administration.

3:30 PM Some Review Issues in Design and Statistical Analysis of Human Drug Abuse Potential Studies

Wei Liu. FDA CDER

3:50 PM Statistical Approaches and Issues in HAP Studies

Kelsey Brown[1]*,* ♦*Reilly Reis*[1] *and Michael Smith*. [1]PRA Health Sciences

4:10 PM Common statistical issues in drug development for major psychiatric disorders

Thomas Birkner. Food and Drug Administration

4:30 PM Floor Discussion.

Session 41: Analysis of Multi-Type Data (*Invited*)

Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3

Organizer: Bertrand Clarke, University of Nebraska-Lincoln.

Chair: Yao Xie, Georgia Institute of Technology.

3:30 PM Bayesian Models and Analysis of High-dimensional Multiplatform Genomics Data

Sounak Chakraborty. University of Missouri-Columbia

3:55 PM Kernel machines for -omics data integration

♦*Dominik Reinhold, Junxiao Hu, Katerina Kechris and Debashis Ghosh*. University of Colorado Denver

4:20 PM Big Data Regression for Predicting Genome-wide Functional Genomic Signals

Weiqiang Zhou, Ben Sherwood, Zhicheng Ji, Fang Du, Jaiwei Bai and ♦*Hongkai Ji*. Johns Hopkins Bloomberg School of Public Health

4:45 PM Prioritizing causal SNPs through integrating phenotype, genotype, omics and functional annotations

♦*Qi Zhang*[1]*, Constanza Rojo*[2] *and Sunduz Keles*[2]. [1]University of Nebraska Lincoln [2]University of Wisconsin Madison

5:10 PM Floor Discussion.

Session 42: New Advances in Quantile Regression (*Invited*)

Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3

Organizer: Lan Wang, University of Minnesota.

Chair: Ruosha Li, University of Texas School of Public Health.

3:30 PM Estimation and Inference of Quantile Regression Under Biased Sampling

♦*Gongjun Xu*[1]*, Tony Sit*[2]*, Lan Wang*[1] *and Chiung-Yu Huang*[3]. [1]University of Minnesota [2]The Chinese University of Hong Kong [3]Johns Hopkins University

3:55 PM A Quantile Approach for Fractional Data

♦*Hyokyoung (Grace) Hong*[1] *and Huixia Wang*[2]. [1]Michigan State University [2]George Washington University

4:20 PM High dimensional censored quantile regression

♦*Qi Zheng*[1]*, Limin Peng*[2] *and Xuming He*[3]. [1]University of Louisville [2]Emory University [3]University of Michigan

4:45 PM An alternative formulation of functional partial quantile linear regression and its properties

♦*Dengdeng Yu, Linglong Kong and Ivan Mizera*. University of Alberta

5:10 PM Floor Discussion.

Session 43: Advances and Challenges in Time-to-Event Data Analysis (*Invited*)

Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3

Organizer: Xu Zhang, University of Mississippi Medical Center.

Chair: Ruiyan Luo, Georgia State University.

3:30 PM Analysis of dependently truncated data in Cox framework

♦*Xu Zhang*[1]*, Yang Liu*[2] *and Ji Li*[3]. [1]University of Mississippi Medical Center [2]Centers for Disease Control and Prevention [3]University of Oklahoma Health Sciences Center

3:55 PM To MICE or not to MICE? A study of multiple imputation strategies for Accelerated Failure Time Model
♦*Lihong Qi*[1]*, Ying-Fang Wang*[2]*, Rongqi Chen*[1] *and Yulei He*[3]. [1]University of California Davis [2]The California State University [3]CDC

4:20 PM Surviving joint models: Improving the survival subcomponent of joint longitudinal-survival models
Michael Griswold. Center of Biostatistics, Univ MS Medical Center

4:45 PM The proportional odds cumulative incidence model for competing risks
Frank Eriksson[1]*, Jianing Li*[2]*, Thomas Scheike*[1] *and* ♦*Mei-Jie Zhang*[3]. [1]University of Copenhage [2]Merck [3]Medical College of Wisconsin

5:10 PM Floor Discussion.

Session 44: Jiann-Ping Hsu invited Session on Biostatistical and Regulatory Sciences (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Lili Yu, Georgia Southern University.
Chair: Timothy O'Brien, Loyola University Chicago.

3:30 PM A homoscedasticity test for the Accelerated Failure Time model
♦*Lili Yu*[1]*, Liang Liu*[2] *and Din Chen*[3]. [1]Georgia Southern University [2]University of Georgia [3]University of North Carolina at Chapel Hill

3:55 PM Prospective Validation of the National Field Triage Guidelines: Challenges of a Probability Stratifi
♦*Rongwei (Rochelle) Fu and Craig Newgard*. OHSU

4:20 PM From Statistical Power to Statistical Assurance: Time for the Paradigm Change in Clinical Trial Desi
♦*Ding-Geng Chen*[1] *and Shuyen Ho*[2]. [1]University of North Carolina at Chapel Hill [2]PAREXEL, Durham, NC 27709, USA

4:45 PM Internal pilot design for repeated measures
♦*Xinrui Zhang and Yueh-Yun Chi*. University of Florida

5:10 PM Floor Discussion.

Session 45: Complex Data Analysis: Theory and Methods (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Feng Yang, Columbia University.
Chair: Lucy Xia, Stanford University.

3:30 PM Linear hypothesis testing in high-dimensional one-way MANOVA
♦*Jin-Ting Zhang, Jia Guo and Bu Zhou*. National University of Singapore

3:55 PM Estimation of sparse directed acyclic graphs through a lasso framework and its applications
Sung Won Han and ♦*Judy Zhong*. New York University

4:20 PM Graph-Guided Banding for Covariance Estimation
Jacob Bien. Cornell University

4:45 PM Flexible Spectral Methods for Community Detection
♦*Pengsheng Ji*[1]*, Jiashun Jin*[2] *and Tracy Ke*[3]. [1]University of Georgia [2]Carnegie Mellon University [3]University of Chicago

5:10 PM Floor Discussion.

Session 46: Fundamentals and Challenges in Subgroup Identification and Analysis (*Invited*)
Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizers: Qi Jiang, Amgen; Rui (Sammi) Tang, Vertex.
Chair: Qi Jiang, Amgen.

3:30 PM A Visualization Method Measuring the Performance of Biomarkers for Guiding Treatment Decisions and Subgroup Identification
♦*rui tang*[1]*, hui yang*[2] *and jing huang*[3]. [1]vertex [2]amgen [3]veracyte

3:55 PM What can we learn from subgroup analysis in randomized controled trials?
Xin Zhao. Janssen Pharmaceutical

4:20 PM Subgroup Analysis to Assess Benefit:Risk
♦*Steven Snapinn and Qi Jiang*. Amgen

4:45 PM Discussant: Bo Huang, Pfizer

5:10 PM Floor Discussion.

Session 47: Joint Model and Applications (*Invited*)
Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3
Organizer: Hongbin Zhang, City University of New York.
Chair: Qing Yang, Duke University.

3:30 PM Joint inference of GLMM and NLME with Informative Censoring with Application in HIV/AIDS
♦*Hongbin Zhang*[1] *and Lang Wu*[2]. [1]City University of New York [2]University of British Columbia

3:55 PM Joint-modelling of discrete, continuous and semi-continuous data with applications
Renjun Ma. University of New Brunswick, Canada

4:20 PM Two-Step and Likelihood Methods for HIV Viral Dynamic Models with Covariate Measurement Errors
♦*WEI LIU*[1] *and LANG WU*[2]. [1]YOKR UNIVERSITY [2]UNIVERSITY OF BRITISH COLUMNIA

4:45 PM Dynamic modeling and inference for event detection
Hongyu Miao. School of Public Health, UTHealth

5:10 PM Floor Discussion.

Session 48: Recent Development in Functional Data Analysis and Applications (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Wenbin Lu, North Carolina State University.
Chair: Jiajia Zhang, University of South Carolina.

3:30 PM Dynamic Functional Mixed Models for Child Growth
Andrew Leroux[1]*,* ♦*Luo Xiao*[2]*, Will Checkley*[1] *and Ciprian Crainiceanu*[1]. [1]Johns Hopkins University [2]North Carolina State University

3:55 PM Nested Hierarchical Functional Data Modeling for Root Gravitropism Data

Yehua Li. Iowa State University

4:20 PM Functional and imaging data in precision medicine

♦*Todd Ogden*[1]*, Adam Ciarleglio*[2]*, Thaddeus Tarpey*[3] *and Eva Petkova*[2]. [1]Columbia University [2]New York University [3]Wright State University

4:45 PM A New Method on Flexible Combination of Multiple Diagnostic Biomarkers

♦*Tu Xu*[1]*, Junhui Wang*[2]*, Yixin Fang*[3] *and Alan Rong*[4]. [1]Abbvie Inc. [2]City University of Hong Kong [3]New York University [4]Astellas Pharma

5:10 PM Floor Discussion.

Session 49: Recent Development of Bayesian High Dimensional Modeling, Inference and Computation (*Invited*)

Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)

Organizer: Qifan Song, Purdue University.

Chair: Arman Sabbaghi, Purdue University.

3:30 PM A new double empirical Bayes approach for high-dimensional problems

Ryan Martin. University of Illinois at Chicago

3:55 PM NEARLY OPTIMAL BAYESIAN SHRINKAGE FOR HIGH DIMENSIONAL REGRESSION

♦*Qifan Song*[1] *and Faming Liang*[2]. [1]Purdue Univ. [2]Univ. of Flordia

4:20 PM Scalable Bayesian Variable Selection for Structured High-dimensional Data

Changgee Chang, Suprateek Kundu and ♦*Qi Long*. Emory University

4:45 PM Prediction risk for global-local shrinkage regression

♦*Anindya Bhadra*[1]*, Jyotishka Datta*[2]*, Yunfan Li*[1]*, Nicholas Polson*[3] *and Brandon Willard*[3]. [1]Purdue University [2]Duke University [3]University of Chicago

5:10 PM Floor Discussion.

Session 50: On Clinical Trials with a High Placebo Response (*Invited*)

Room: BAKER, ATLANTA CONFERENCE CENTER - LL3

Organizer: Pilar Lim, Johnson & Johnson.

Chair: Pilar Lim, Johnson & Johnson.

3:30 PM Placebo Effects and Sequential Parallel Comparison (SPC) Design

Eiji Ishida. FDA/CDER/OTS/OB/DBI

3:55 PM An Unbiased Estimator of the Two-Period Treatment Effect in Doubly Randomized Delayed-Start Designs

♦*Yihan Li*[1]*, Yanning Liu*[2]*, Qing Liu*[2] *and Pilar Lim*[2]. [1]AbbVie [2]Janssen Research and Development

4:20 PM Accounting for High Placebo Response Rates in Clinical Trials

Pilar Lim, ♦*Akiko Okamoto and George Chi*. Janssen Research & Development

4:45 PM Discussant: Gheorghe Doros, Boston University

5:10 PM Floor Discussion.

Session 51: Statistical Learning Methods (*Contributed*)

Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3

Chair: Jun Li, Kent State University.

3:30 PM High-Dimensional Hypothesis Testing With the Lasso

♦*Sen Zhao, Ali Shojaie and Daniela Witten*. University of Washington

3:45 PM A simultaneous variable selection and clustering method for high-dimensional multinomial regression

♦*Sheng Ren*[1]*, Emily Kang*[1] *and Jason Lu*[2]. [1]University of Cincinnati [2]Cincinnati Children's Hospital Research Foundation

4:00 PM Lasso-type Network Community Detection within Latent Space

♦*Shiwen Shen and Edsel Pena*. University of South Carolina

4:15 PM Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data

♦*Chiyu Gu*[1]*, Subharup Guha*[1] *and Veera Baladandayuthapani*[2]. [1]University of Missouri [2]MD Anderson Cancer Center

4:30 PM Sparsity and Error Analysis of Empirical Feature-Based Regularization Schemes

♦*Xin Guo*[1]*, Jun Fan*[2] *and Ding-Xuan Zhou*[3]. [1]The Hong Kong Polytechnic University [2]University of Wisconsin-Madison [3]City University of Hong Kong

4:45 PM Compressed Covariance Matrix Estimation With Automated Dimension Learning

♦*Gautam Sabnis*[1]*, Debdeep Pati*[2] *and Anirban Bhattacharya*[3]. [1]Florida State University [2]Florida State University [3]Texas A&M University

5:00 PM Floor Discussion.

Tuesday, June 14. 8:00 AM - 10:10 AM

Keynote session II (*Keynote*)

Room: International Ballroom

Organizers: ICSA 2016 organizing committee.

Chair: Zhezhen Jin, Columbia University and Mei-Ling Ting Lee, University of Maryland.

8:00 AM Keynote lecture II

David Madigan. Columbia University

9:10 AM Keynote lecture III

Paul Albert. National Institute of Health

10:10 AM Floor Discussion.

Tuesday, June 14. 10:30 AM - 12:10 PM

Session 52: Novel Design and/or Analysis for Phase 2 Dose Ranging Studies (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Grace Li, Eli Lilly and Company.
Chair: Grace Li, Eli Lilly and Company.

10:30 AM Role of Biomarkers and Quantitative Models in Dose Ranging Trials
Yaning Wang. FDA

10:55 AM Achieving multiple objectives in a Phase 2 dose-ranging study
Ming-Dauh Wang. Eli Lilly and Company

11:20 AM Optimizing Oncology Combination Therapy by Integrating Multiple Information into Dose Escalation
Xiaowei Guan. Bristol Myers Squibb

11:45 AM Discussant: Alan Chiang, Eli Lilli and Company

12:10 PM Floor Discussion.

Session 53: Lifetime Data Analysis (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Mei-Ling Ting Lee, University of Maryland.
Chair: Mei-Ling Ting Lee, University of Maryland.

10:30 AM The effect of the timing of treatment confounded by indication and screening on cancer mortality
Alex Tsodikov. University of Michigan

10:55 AM Dynamic risk prediction models for data with competing risks
♦*Chung-Chou Chang*[1] *and Qing Liu*[2]. [1]University of Pittsburgh [2]Novartis Pharmaceutical Corporation

11:20 AM Modeling Gap Times in Panel Count Data with Informmative Observation Times: Assessing Spontaneous Labor in Women
♦*Rajeshwari Sundaram and Ling Ma*. Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH

11:45 AM Integrated analysis of multidimensional omics data for prognosis
Shuangge Ma. Yale University

12:10 PM Floor Discussion.

Session 54: Innovative Methods for Clinical Trial Designs (*Invited*)
Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Hongjian Zhu, The University of Texas School of Public Health.
Chair: Ruosha Li, The University of Texas School of Public Health.

10:30 AM Generalized Efron's Biased Coin Design and its Theoretical Properties
YANQING HU. West Virginia University and Incyte

10:55 AM Efficient Algorithms for Extended Two-stage Adaptive Designs for Phase II Clinical Trials
Seongho Kim[1] *and* ♦*Weng Kee Wong*[2]. [1]Karmanos Cancer Institute, School of Medicine, Wayne State University [2]UCLA

11:20 AM Optimal Sequential Enrichment Designs for Phase II Clinical Trials
Yong Zang[1] *and* ♦*Ying Yuan*[2]. [1]Florida Atlantic University [2]MD Anderson Cancer Center

11:45 AM New covariate-adjusted response-adaptive designs for precision medicine
♦*Feifang Hu, Fan Wang and Wanying Zhao*. George Washington University

12:10 PM Floor Discussion.

Session 55: Recent Advances in Statistical Methods for Alzheimer's Disease Studies (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Jing Qian, University of Massachusetts, Amherst.
Chair: Xiaoyi Min, Georgia State University.

10:30 AM Alzheimer's Disease Neuroimaging Initiative: statistical challenges and solutions
♦*Sharon Xie*[1]*, Matthew White*[2] *and Jarcy Zee*[3]. [1]University of Pennsylvania [2]Boston Children's Hospital [3]Arbor Research Collaborative for Health

10:55 AM Adjusting for dependent truncation with inverse probability weighting
♦*Jing Qian*[1] *and Rebecca Betensky*[2]. [1]University of Massachusetts Amherst [2]Harvard University

11:20 AM Capturing Change in Cognition Using Data from Two Cohort Studies of Aging and Dementia
Lei Yu. Rush Alzheimer's Disease Center

11:45 AM A BAYESIAN FUNCTIONAL LINEAR COX REGRESSION MODEL IN ALZHEIMER'S Disease study
♦*Eunjee Lee*[1]*, Hongtu Zhu*[1]*, Dehan Kong*[1]*, Yalin Wang*[2]*, Kelly Sullivan Giovanello*[1] *and Joseph G Ibrahim*[1]. [1]University of North Carolina at Chapel Hill [2]ARIZONA STATE UNIVERSITY

12:10 PM Floor Discussion.

Session 56: High Dimensional Model and Prediction (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Rui Feng, University of Pennsylvania.
Chair: Rui Feng, University of Pennsylvania.

10:30 AM Comparison of common variable selection approaches for accuracy and predictive values involving corr
♦*Wei-Ting Hwang, Rengyi (Emily) Xu, Clementina Mesaros and Ian Blair*. University of Pennsylvania

10:55 AM Integration of high-dimensional genomic and imaging features for risk prediction of lung cancers
Fenghai Duan. Brown University

11:20 AM A statistical framework for eQTL mapping with imprinting effect detection using RNA-seq data
♦*Feifei Xiao*[1]*, Guoshuai Cai*[2]*, Jianzhong Ma*[3] *and Chirstopher Amos*[2]. [1]University of South Carolina [2]Dartmouth College [3]University of Texas Health Science Center

11:45 AM Variable screening via quantile partial correlation
Shujie Ma. University of California, Riverside

12:10 PM Floor Discussion.

Session 57: Statistical Methods for Medical Research using Real-World Data (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizers: Haoda Fu, Eli Lilly and Company; Wei Shen, Eli Lilly and Company.
Chair: Haoda Fu , Eli Lilly and Company.

10:30 AM Gauging drug-outcome associations by leveraging EMR and existent knowledge
Jia Zhan, ♦*Xiaochun Li and Changyu Shen*. Indiana University School of Medicine

10:55 AM Statistical methods for drug safety using Electronic Health Records
♦*Ying Wei*[1]*, Daniel Backenroth*[1]*, Ying Li*[2]*, Alex Belloni*[3] *and Carol Friedman*[1]. [1]Columbia University [2]IBM [3]Duke University

11:20 AM Generate Individualized Treatment Decision Tree Algorithm with Application to EMR
Kevin Doubleday[1]*, Haoda Fu*[2] *and* ♦*Jin Zhou*[1]. [1]University of Arizona [2]Lilly Corporate Center

11:45 AM Addressing Unmeasured Confounding in Comparative Observational Research
♦*Wei Shen, Douglas Daries and Xiang Zhang*. Eli Lilly and Company

12:10 PM Floor Discussion.

Session 58: New Developments on High Dimensional Learning (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yichao Wu, North Carolina State University.
Chair: Wei Liu, York University.

10:30 AM Robust High-dimensional Data Analysis Using a Weight Shrinkage Rule
♦*Xiaoli Gao*[1]*, Bin Luo*[1] *and Yixin Fang*[2]. [1]University of North Carolina at Greensboro [2]New York University

10:55 AM Inverse Methods for Sufficient Forecasting Using Factor Models
♦*Lingzhou Xue*[1]*, Jianqing Fan*[2]*, Wei Luo*[3] *and Jiawei Yao*[4]. [1]Penn State University [2]Princeton University [3]CUNY - Baruch College [4]Citadel LLC

11:20 AM Estimation and Variable Selection for Single-index Cox Proportional Hazard Regression
Peng Zeng. Auburn University

11:45 AM An augmented ADMM algorithm with application to the generalized lasso problem
Yunzhang Zhu. The Ohio State University

12:10 PM Floor Discussion.

Session 59: Emerging Statistical Theory in Analyzing Complex Data (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ping Ma, University of Georgia.
Chair: Ping Ma, University of Georgia.

10:30 AM Using an Event-History with Risk-Free Model to Study Genetics of Alcohol Dependence
Hsin-Chou Yang[1]*, I-Chen Chen*[2]*, Yuh-Chyuan Tsay*[1]*, Zheng-Rong Li*[1]*, Chun-houh Chen*[1]*, Hai-Guo Hwu*[3] *and* ♦*Chen-Hsin Chen*[1]. [1]Institute of Statistical Science, Academia Sinica [2]Dept of Biostatistics, University of Kentucky [3]Dept of Psychiatry, National Taiwan University

10:55 AM A general approach to categorizing a continuous scale according to an ordinal outcome
♦*Limin Peng*[1]*, Amita Manatunga*[1]*, Ming Wang*[2]*, Ying Guo*[1] *and AKM Rahman*[1]. [1]Emory University [2]Penn State University

11:20 AM Multivariate Fay-Herriot Hierarchical Bayesian Estimation of Small Area Means with Measurement Error
Serena Arima[1]*, William Bell*[2]*,* ♦*Gauri Datta*[3]*, Carolina Franco*[2] *and Brunero Liseo*[1]. [1]University of Rome, La Sapienza [2]U.S. Census Bureau [3]U.S. Census Bureau, University of Georgia

11:45 AM Optimal Estimation for Quantile Regression with Functional Response
♦*Xiao Wang*[1]*, Zhengwu Zhang*[2]*, Linglong Kong*[3] *and Hongtu Zhu*[4]. [1]Purdue University [2]SAMSI [3]University of Alberta [4]University of North Carolina

12:10 PM Floor Discussion.

Session 60: Survival and Cure Rate Modeling (*Invited*)
Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3
Organizer: Li-Shan Huang, National Tsing Hua University.
Chair: Xu Zhang, University of Mississippi Medical Center.

10:30 AM Evaluating Utility Measurement from Recurrent Marker Processes in the Presence of Competing Terminal
Mei-Cheng Wang. Johns Hopkins University

10:55 AM Tests for stochastic ordering under biased sampling
♦*Hsin-wen Chang*[1]*, Hammou El Barmi*[2] *and Ian McKeague*[3]. [1]Academia Sinica [2]The City University of New York [3]Columbia University

11:20 AM On relative importance in the effect of two exposures
Xinhua Liu and ♦*Zhezhen Jin*. Department of Biostatistics, Columbia University

11:45 AM Semiparametric Accelerated Failure Time Models with Missing Covariates
♦*Shuai Chen and Menggang Yu*. University of Wisconsin - Madison

12:10 PM Floor Discussion.

Session 61: Semiparametric Statistical Methods for Complex Data (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Heng Lian, University of New South Wales.
Chair: Nelson Chen, Emory University.

10:30 AM ESTIMATION OF NON-CROSSING QUANTILE SURFACES
Chen Dong[1], *Shujie Ma*[2], *Liping Zhu*[3] *and* ♦*Xingdong Feng*[1]. [1]Shanghai University of Finance and Economics [2]University of California-Riverside [3]Renmin University of China

10:55 AM Composite Estimation for Single-Index Model with Responses Subject to Detection Limits
Yanlin Tang[1], ♦*Huixia Wang*[2] *and Hua Liang*[2]. [1]Tongji University [2]The George Washington University

11:20 AM Efficient Estimation of Partially Linear Models for Spatial Data over Complex Domains
♦*Lily Wang*[1], *Guannan Wang*[2], *Ming-Jun Lai*[3] *and Lei Gao*[1]. [1]Iowa State University [2]College of William and Mary [3]The University of Georgia

11:45 AM A Regression Model for the General Trend Analysis of Bivariate Panel Data in Continuous Scale
♦*Yi Ran Lin*[1] *and Wei Hsiung Chao*[2]. [1]Institute of Statistical Science, Academia Sinica [2]Dept. of appl. math., National Dong Hwa University

12:10 PM Floor Discussion.

Session 62: Recent Advances on Multiple Fronts of Statistical Analysis for Genomics Data (*Invited*)
Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3
Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.
Chair: Lihong Qi, University of California, Davis.

10:30 AM A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer
Binghui Liu[1], *Xiaotong Shen*[2] *and* ♦*Wei Pan*[2]. [1]Northeast Normal University [2]University of Minnesota

10:55 AM A novel tail dependence measure to quantify the reproducibility and quality of sequencing experiments
Tao Yang and ♦*Qunhua Li*. Penn State University

11:20 AM Large scale multiple testing for clustered signals
♦*Hongyuan Cao*[1] *and Wei Biao Wu* [2]. [1]University of Missouri-Columbia [2]University of Chicago

11:45 AM A Cautionary Note on using Cross-validation for Molecular Classification
♦*Li-Xuan Qin and Huei-Chung Huang*. Memorial Sloan Kettering Cancer Center

12:10 PM Floor Discussion.

Session 63: Multi-regional Clinical Trials (MRCT): Statistical Challenges, Trial Design Approaches, and Other Aspects (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ran Liu, AbbVie.
Chair: Ran Liu, AbbVie.

10:30 AM A Few Considerations on Regional Differences in MRCT
♦*Bo Yang*[1] *and Yijie Zhou*[2]. [1]Vertex [2]AbbVie

10:55 AM Inconsistency and drop-minimum data analysis
Fei Chen[1], ♦*Gang Li*[1] *and Gordon Lan*. [1]J&J

11:20 AM Regional Efficacy Assessment in Multi-Regional Clinical Development
Yijie Zhou. AbbVie

11:45 AM New Method of Borrowing Information from Outside Regional Data for Analyzing the Local Regional Data
♦*Takahiro Hasegawa*[1], *Lu Tian*[2], *Brian Claggett*[3] *and Lee-Jen Wei*[4]. [1]Shionogi & Co., Ltd. [2]Stanford University School of Medicine [3]Brigham and Women's Hospital [4]Harvard University

12:10 AM Floor Discussion.

Session 64: High Dimensional Inference: Methods and Applications (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Haipeng Shen, University of Hong Kong.
Chair: Xin Qi, Georgia State University.

10:30 AM Localized-Variate PCA for Multivariate Functional Data
Robert Krafty. University of Pittsburgh

10:50 AM Nonparametric Screening under Conditional Strictly Convex Loss
Xu Han. Temple University

11:10 AM Dimension Reduction for Big Data Analysis
Dan Shen. University of South Florida

11:30 AM Distance-Based Methods for Analyzing Data from 16S rRNA Microbiome Studies
Glen Satten. Centers for Disease Control and Prevention

11:50 AM Bootstrap-Based Measures of Uncertainty for EEG Artifact Detection using ICA
♦*Rachel Nethery and Young Truong*. University of North Carolina at Chapel Hill

12:10 PM Floor Discussion.

Session 65: Modern Advancements in High-Dimensional Functional Data (*Invited*)
Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)
Organizer: Guannan Wang, College of William & Mary.
Chair: Guannan Wang, College of William & Mary.

10:30 AM Single-index Models for Function-on-Function Regression
♦*Guanqun Cao*[1] *and Lily Wang*[2]. [1]Auburn University [2]Iowa State University

10:55 AM Longitudinal Regression for Time-varying Functional Covariate
♦*Md Nazmul Islam and Ana-Maria Staicu*. North Carolina State University

11:20 AM A rotate-and-solve procedure for high dimensional classification
Ning Hao. The University of Arizona

11:45 AM Multivariate Spatio-Temporal Models for High-Dimensional Areal Data
♦*Scott Holan, Jonathan Bradley and Christopher Wikle*. University of Missouri

12:10 PM Floor Discussion.

Session 66: Bayesian Approaches for Medical Product Evaluation (*Invited*)
Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yunling Xu, US Food and Drug Administration.
Chair: Yihan Li, AbbVie.

10:30 AM BEANZ: A Web-Based Software for Bayesian Analysis of Heterogeneous Treatment Effect
Chenguang Wang. Johns Hopkins University

10:55 AM Bayesian Models to Leverage Data from Early Visits in an Adaptive Design
Anna McGlothlin. Berry Consultants

11:20 AM Incorporation of stochastic engineering models as a prior in Bayesian medical device trials
Tarek Haddad[1], ♦*Adam Himes*[1], *Laura Thompson*[2], *Telba Irony*[2] *and Rajesh Nair*[2]. [1]Medtronic [2]FDA

11:45 AM Evaluation of treatment efficacy using a Bayesian mixture piecewise linear model
♦*Lili Zhao*[1], *Dai Feng, Brian Neelon and Marc Buyse*. [1]Biostatistics, University of Michigan

12:10 PM Floor Discussion.

Session 67: New Methods for Complex Data (*Invited*)
Room: BAKER, ATLANTA CONFERENCE CENTER - LL3
Organizer: Rui Song, North Carolina State University.
Chair: Marianthi Markatou, University at Buffalo.

10:30 AM Jackknife Empirical Likelihood for the Gini Correlation
♦*YONGLI SANG*[1], *XIN DANG*[2] *and YICHUAN ZHAO*[3]. [1]UNIVERSITY OF MISSISSIPPI [2]UNIVERSITY [3]GEORGIA STATE UNIVERSITY

10:55 AM Complex-valued wavelet lifting and applications
♦*Marina Knight*[1], *Jean Sanderson*[2], *Matt Nunes*[3] *and Piotr Fryzlewicz*[4]. [1]University of York, UK [2]University of Sheffield, UK [3]University of Lancaster, UK [4]London School of Economics, UK

11:20 AM Neyman-Pearson (NP) Classification and NP-ROC
♦*Xin Tong* [1], *Yang Feng*[2] *and Jingyi Li*[3]. [1]University of Southern California [2]Columbia University [3]University of California, Los Angeles

11:45 AM Reduced-Rank Linear Discriminant Analysis
♦*Yue Niu*[1], *Ning Hao*[1] *and Bin Dong*[2]. [1]University of Arizona [2]Peking University

12:10 PM Floor Discussion.

Session 68: Recent Advances in Regression Analysis (*Contributed*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Chair: Wenjiang Fu, University of Houston.

10:30 AM Classification of Paper Citation Trajectories Through Functional Poisson Regression Model
♦*Ruizhi Zhang*[1], *Jian Wang*[2] *and Yajun Mei*[1]. [1]Georgia Institute of Technology [2]KU Leuven

10:45 AM Kernel Ridge Regression under Random Projection: Computational-and-Statistical Trade-off
♦*Meimei Liu*[1], *Zuofeng Shang*[2] *and Guang Cheng*[1]. [1]Department of Statistics, Purdue University [2]Department of Math Science, Binghamton University

11:00 AM Solving the Identifiability Problem with the Lasso Regularization in Age-period-cohort Analysis
♦*Beverly Fu*[1] *and Wenjiang Fu*[2]. [1]Okemos High School [2]University of Houston

11:15 AM A general framework for the regression analysis of pooled biomarker assessments
♦*Yan Liu, Christopher McMahan and Colin Gallagher*. Clemson University

11:30 AM Permutation inference distribution for linear regression and related models
♦*Qiang Wu and Paul Vos*. East Carolina University

11:45 AM Comparison of Classical and Quantile Regression Methods for Modeling Childhood Obesity
♦*Gilson Honvoh*[1], *Roger Zoh*[2], *Hongwei Zhao*[2], *Mark Benden*[2], *Guoyao Wu*[3] *and Carmen Tekwe*[2]. [1]Texas A&M School of Public Health [2]Texas A&M School of Public Health [3]Texas A&M University

12:00 PM Floor Discussion.

Tuesday, June 14. 1:30 PM - 3:10 PM

Session 69: Recent Development in Dose Finding Studies (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Dong Xi, Novartis.
Chair: Dong Xi, Novartis.

1:30 PM A Simple and Efficient Statistical Approach for Designing an Early Phase II Clinical Trial
Yaohua Zhang[1], *Qiqi Deng*[2], *Susan Wang*[2] *and* ♦*Naitee Ting*[2]. [1]University of Connecticut [2]Boehringer-Ingelheim Pharmaceuticals, Inc.

1:55 PM Sample Size Consideration based on Quantitative Decision Framework in Dose Finding Studies
♦*Huilin Hu and Dong Xi*. Novartis Pharmaceuticals Corporation

2:20 PM Improving dose finding studies with MCP-Mod design consideration
♦*Kuo-mei Chen and Jose Pinheiro*. Jassen Research & Development, LLC

2:45 PM Discussant: H.M. James Hung, FDA

3:10 PM Floor Discussion.

Session 70: New Advances in High Dimensional and Complex Data Analysis (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Pengsheng Ji, University of Georgia.
Chair: Pengsheng Ji, University of Georgia.

1:30 PM Probing the Pareto Frontier of Computational-Statistical Tradeoffs
Han Liu. Princeton University

1:55 PM Factor Adjusted Graphlet Screening
♦*Tracy Ke and Fan Yang*. University of Chicago

2:20 PM Truth, Knowledge, P-Values, Bayes, & Inductive Inference
Edel Pena. University of South Carolina

2:45 PM Regression in heterogeneous problems
Hanwen Huang. UGA

3:10 PM Floor Discussion.

Session 71: Design of Experiments I (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Abhyuday Mandal, University of Georgia.
Chair: John Stufken, Arizona State University.

1:30 PM Minimax designs using clustering
♦*Simon Mak and V. Roshan Vengazhiyil*. Georgia Institute of Technology

1:55 PM Optimal Experimental Designs for Nonlinear Conjoint Analysis
Mercedes Esteban-Bravo[1], ♦*Agata Leszkiewicz*[2] *and Jose M. Vidal-Sanz*[1]. [1]Universidad Carlos III de Madrid [2]Georgia State University

2:20 PM Obtaining locally D-optimal designs for binary response experiments via Particle Swarm Optimization
♦*Joshua Lukemire*[1], *Abhyuday Mandal*[2] *and Weng Kee Wong*[3]. [1]Emory University [2]University of Georgia [3]University of California at Los Angeles

2:45 PM Floor Discussion.

Session 72: Recent Advancement about Adaptive Design in all Phases of Clinical Trial (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizers: Qi Jiang, Amgen ; Rui (Sammi) Tang, Vertex.
Chair: Rui (Sammi) Tang, Vertex.

1:30 PM Key Statistical Issues in Adaptive Design in Oncology Trials Application
♦*Qi Jiang and Chunlei Ke*. Amgen

1:55 PM Early phase trial designs
Sumithra Mandrekar. Mayo Clinic

2:20 PM Panel:
Ying Yuan, The University of Texas MD Anderson Cancer Center
Ouhong Wang, Amgen Asia R&D Center
Qi Jiang, Amgen
Sumithra J. Mandrekar, Mayo Clinic

2:45 PM Floor Discussion.

Session 73: Recent Advances on Statistical Analysis of Safety and/or Efficacy Endpoints in Clinical Trials (*Invited*)
Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)
Organizer: Zhigang Zhang, Memorial Sloan-Kettering Cancer Center.
Chair: Zhigang Zhang, Memorial Sloan-Kettering Cancer Center.

1:30 PM On the Restricted Mean Survival Time Curve in Survival Analysis
♦*Lihui Zhao*[1], *Brian Claggett*[2], *Lu Tian*[3], *Hajime Uno*[4], *Lorenzo Trippa*[5] *and Lee-Jen Wei*[5]. [1]Northwestern University [2]Harvard Medical School [3]Stanford University [4]Dana-Farber Cancer Institute [5]Harvard University

1:55 PM Bayesian methods for meta-analysis combining randomized-controlled and single-arm studies
♦*Jing Zhang*[1], *Chia-Wen Ko*[2], *Lei Nie*[2], *Yong Chen*[3] *and Ram Tiwari*[2]. [1]University of Maryland [2]U.S. Food and Drug Administration [3]University of Pennsylvania

2:20 PM Quantifying treatment benefit in molecular subgroups to assess a predictive biomarker.
♦*Alexia Iasonos and Jaya Satagopan*. Memorial Sloan Kettering Cancer Center

2:45 PM Bayesian proportional hazards model for interval censored data in cancer clinical trials
♦*Ai Ni, Zhigang Zhang and Mithat Gonen*. Memorial Sloan Kettering Cancer Center

3:10 PM Floor Discussion.

Session 74: New Statistical Methods for Analysis of Large-Scale Genomic Data (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Shujie Ma, University of California, Riverside.
Chair: Shujie Ma, University of California, Riverside.

1:30 PM Statistical Inference for Time Course RNA-Seq Data
Xiaoxiao Sun[1], *Dalpiaz David*[2], *Di Wu*[3], *Jun Liu*[3] *and* ♦*Ping Ma*[1]. [1]University of Georgia [2]University of Illinois at Urbana-Champaign [3]Harvard University

1:55 PM Nonparametric regularized regression for network construction and taxa selection
♦*Wenchuan Guo*[1], *Zhenqiu Liu*[2] *and Shujie Ma*[1]. [1]University of California, Riverside [2]Cedars-Sinai Medical Center

2:20 PM Statistical inference of allele-specific contacts from high-throughput chromosome conformation data
♦*Wenxiu Ma*[1], *Xinxian Deng*[2], *Vijay Ramani*[2], *Zhijun Duan*[2], *Jay Shendure*[2], *William Noble*[2] *and Christine Disteche*[2]. [1]University of California Riverside [2]University of Washington

2:45 PM Metagenomics Binning via sequential Monte Carlo (SMC) method
Xinping Cui, Chen Gao and ♦*Wei Cui*. University of California, Riverside

3:10 PM Floor Discussion.

Session 75: New Development in Function Data Analysis (*Invited*)

Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ruiyan Luo, Georgia State University.
Chair: Matt Hayat, Georgia State University.

1:30 PM Nonlinear function on function regression model
♦*Xin Qi and Ruiyan Luo*. Georgia State University

1:55 PM Bayesian Registration of Functions with a Gaussian Process Prior
♦*YI LU, Sebastian Kurtek and Radu Herbei*. The Ohio State University

2:20 PM Fast Bayesian inference for complex, ultra-high dimensional functional data
♦*Hongxiao Zhu*[1], *Fengrong Wei*[2] *and Xiaowei Wu*[1]. [1]Virginia Tech [2]University of West Georgia

2:45 PM Function on function regression with thousands of predictive curves
♦*Ruiyan Luo and Xin Qi*. Georgia State University

3:10 PM Floor Discussion.

Session 76: Some Recent Developments in Robust High-dimensional Data Analysis (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Xiaoli Gao, University of North Carolina at Greensboro.
Chair: Xiaoli Gao, University of North Carolina at Greensboro.

1:30 PM Fast Community Detection in Complex Networks with a K-Depths Classifier
Yahui Tian and ♦*Yulia Gel*. University of Texas at Dallas, USA

1:55 PM ROCKET: Robust Confidence Intervals via Kendall's Tau for Transelliptical Graphical Models
♦*Mladen Kolar*[1] *and Rina Foygel Barber*[2]. [1]The University of Chicago Booth School Of Business [2]University of Chicago

2:20 PM Model diagnostics and robust estimation in low-rank models
Kun Chen. University of Connecticut

2:45 PM Variable selection for partially linear models via learning gradients
Lei Yang[1], ♦*Yixin Fang*[1], *Junhui Wang* [2] *and Yongzhao Shao*[1]. [1]New York University School of Medicine [2]City University of Hong Kong

3:10 PM Floor Discussion.

Session 77: Recent Advances in Statistical Methods for Handling Missing Data (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Qi Long, Emory University.
Chair: Domonique W. Hodge, Emory University.

1:30 PM Combining IRT with Multiple Imputation to Crosswalk between Health Assessment Questionnaires
Chenyang Gu and ♦*Roee Gutman*. Brown University

1:55 PM A new framework for addressing selection bias due to missing data in EHR-based research
♦*Sebastien Haneuse*[1] *and Michael Daniels*[2]. [1]Harvard T.H. Chan School of Public Health [2]University of Texas - Austin

2:20 PM Imputing cost data that are missing not at random in SEER-Medicare linked data
Rebecca Andridge. The Ohio State University College of Public Health

2:45 PM Nonparametric Imputation for non-ignorable missing data
Domonique Hodge[1], *Chiu-Hsieh Hsu*[2] *and* ♦*Qi Long*[1]. [1]Emory University [2]The University of Arizona

3:10 PM Floor Discussion.

Session 78: Statistical Research in Clinical Trials (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Xiaohong Huang, Vertex Pharmaceutical.
Chair: Lei Hua, Vertex Pharmaceutical.

1:30 PM A Statistical Approach for Clinical Operations in a Long-term Schizophrenia Study
Jun Zhao. AbbVie

1:55 PM Sensitivity Analyses for the Primary Efficacy Endpoint for Kalydeco R117H sNDA Submission
♦*Lan Lan*[1] *and Mei-Hsiu Ling*[2]. [1]Vertex [2]Sr. Director

2:20 PM Estimating Optimal Treatment Regimes via Subgroup Identification in Randomized Control Trials and Observational Studies
Haoda Fu. Eli Lilly and Company

2:45 PM Discussant: Zhaoling Meng, Sanofi

3:10 PM Floor Discussion.

Session 79: Modeling and Analyzing Medical Device Data (*Invited*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Organizer: Mei-Cheng Wang, Johns Hopkins University.
Chair: Mei-Cheng Wang, Johns Hopkins University.

1:30 PM A two-stage model for wearable device data
♦*Jiawei Bai, Yifei Sun, Ciprian Crainiceanu and Mei-Cheng Wang*. Johns Hopkins University

1:55 PM A functional data analysis framework for accelerometry data
♦*Chongzhi Di*[1], *David Buchner*[2], *Andrea LaCroix*[3] *and Ross Prentice*[1]. [1]Fred Hutchinson Cancer Research Center [2]University of Illinois Urbana-Champaign [3]Unviersity of California at San Diego

2:20 PM Variable selection in the concurrent functional linear model
♦*Jeff Goldsmith*[1] *and Joseph Schwartz*[2]. [1]Columbia University, Department of Biostatistics [2]Columbia University Medical Center

2:45 PM Accelerometers, physical activity, and conditional random fields
♦ *John Staudenmayer and Evan Ray*. UMass-Amherst

3:10 PM Floor Discussion.

Session 80: Data Ming and Big Data Analysis (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yichuan Zhao, Georgia State University.
Chair: Hubert Chen, University of Georgia.

1:30 PM Wisdom of Crowds: Meta-analysis of Gene Set Enrichment Studies Utilizing Isoform Expression
♦*Xinlei Wang, Lie Li and Guanghua Xiao*. Southern Methodist University

1:55 PM Statistical Learning and Likelihood Methods as Educational Tools
Timothy OBrien. Loyola University Chicago

2:20 PM A multi-resolution functional ANOVA model for emulation of large-scale, high-dimensional simulations
Chih-Li Sung, Wenjia Wang and ♦*Benjamin Haaland*. Georgia Tech, ISyE

2:45 PM Big Data: How Do We Stay Relevant?
David Morganstein. Westat, Inc.

3:10 PM Floor Discussion.

Session 81: Missing Data and Multiple Imputation (*Invited*)
Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yang Liu, Centers for Disease Control and Prevention.
Chair: Yang Liu, Centers for Disease Control and Prevention.

1:30 PM Quantitative assessment of exposure to fecal contamination for young children in Accra, Ghana
♦*Yuke Wang*[1]*, Peter Teunis*[2]*, Christine Moe*[1]*, Clair Null*[3]*, Suraja Raj*[1]*, Kelly Baker*[4] *and Habib Yakubu*[1]. [1]Emory University [2]RIVM, Netherlands [3]Mathematica Policy Research [4]The University of Iowa

1:55 PM Comparison of two approaches for imputing a composite categorical variable
♦*Yi Pan, Ruiguang Song, Yulei He, Qian An and Guoshen Wang*. Centers for Disease Control and Prevention

2:20 PM Multiple Imputation of Missing Linkage to Care Data for CDC-funded HIV Testing Program Data in 2015
♦*Guoshen Wang, Yi Pan, Puja Seth, Ruiguang Song and Lisa Belcher*. Centers for Disease Control and Prevention

2:45 PM Evaluation of imputation methods for Tdap vaccination among pregnant women, Internet panel survey
♦*Helen Ding*[1]*, Carla Black*[2]*, Srivastav Anup*[3] *and Greby Stacie*[2]. [1]CFD Research Corporation [2]CDC [3]LEIDO

3:10 PM Floor Discussion.

Session 82: Statistical Innovations in the Analysis of Metagenomic Data (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yuan Jiang, Oregon State University.
Chair: Yuan Jiang, Oregon State University.

1:30 PM New development in alignment-free genome and metagenome comparison
Fengzhu Sun. University of Southern California

1:55 PM Informative Approach in Differential Analysis on Time Course Microbial Studies
♦*Lingling An and Dan Luo*. University of Arizona

2:20 PM Regression modeling of microbiome data integrating the phylogenetic tree
♦*Jun Chen*[1] *and Jian Xiao*[2]. [1]Mayo Clinic [2]Mayo linic

2:45 PM Regression analysis with compositional covariates
Hongmei Jiang. Northwestern University

3:10 PM Floor Discussion.

Session 83: Statistical Methods for Network Analysis (*Invited*)
Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yunpeng Zhao, George Mason University.
Chair: Yunpeng Zhao, George Mason University.

1:30 PM Co-clustering of nonsmooth graphons
David Choi. Carnegie Mellon University

1:55 PM Network Reconstruction From High Dimensional Ordinary Differential Equations
♦*Shizhe Chen, Ali Shojaie and Daniela Witten*. University of Washington

2:20 PM Measuring Influence in Twitter Ecosystems Using a Counting Process Modeling Framework
Shawn Mankad. Cornell University

2:45 PM Network Inference from Grouped Observations Using Star Models
♦*Charles Weko*[1] *and Yunpeng Zhao*[2]. [1]Department of Defense [2]George Mason University

3:10 PM Floor Discussion.

Session 84: Robust EM Methods and Algorithms (*Invited*)
Room: BAKER, ATLANTA CONFERENCE CENTER - LL3
Organizer: Xin Dang, University of Mississippi.
Chair: Xinwei Deng, Virginia Tech.

1:30 PM Robust rank-based EM algorithm and trimmed BIC criterion
Xin Dang. University of Mississippi

1:55 PM Robust Expectation Maximization Algorithm for Mixture Models
♦*Yichen Qin*[1] *and Carey Priebe*[2]. [1]University of Cincinnati [2]Johns Hopkins University

2:20 PM Robust estimation of clusters & mixture models based on trimming and constraints
Luis Angel García-Escudero[1]*, Alfonso Gordaliza*[1]*, Francesca Greselin*[2] *and* ♦*Agustin Mayo-Iscar*[1]. [1]Universidad de Valladolid [2]University of Milano Bicocca

2:45 PM Robust mixture regression by EM algorithm
Chun Yu[1]*,* ♦*Weixin Yao*[2] *and Kun Chen*[3]. [1]Jiangxi University of Finance and Economics [2]University of California, Riverside [3]University of Connecticut

3:10 PM Floor Discussion.

Session 85: Student Award Session (*Invited*)
Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3
Organizer: ICSA 2016 Student Paper Award Committee.
Chair: Haitao Chu, University of Minnesota.

1:30 PM Incorporating Biological Information in Sparse PCA with Application to Genomic Data
♦*Ziyi Li, Sandra Safo and Qi Long*. Emory University

1:55 PM Semiparametric Estimation of the Accelerated Failure Time Model with Partly Interval-censored Data
♦*Fei Gao, Donglin Zeng and Danyu Lin*. University of North Carolina at Chapel Hill

2:20 PM A Latent Class Modeling Approach for Predicting Kidney Obstruction in the Absence of a Gold Standard
♦*Lijia Wang*[1]*, Qi Long*[1]*, Andrew Taylor*[2] *and Amita Manatunga*[1]. [1]Biostatistics and Bioinformatics, Emory University [2]Radiology and Imaging Sciences, Emory University

2:45 PM Individualizing Drug Dosage with Longitudinal Data
♦*Xiaolu Zhu and Annie Qu*. University of Illinois, Urbana-Champaign

3:10 PM Floor Discussion.

Tuesday, June 14. 3:30 PM - 5:10 PM

Session 86: Bayesian Approaches in Drug Development: How Many Things Can We Accomplish? (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Cristiana Mayer, Janssen R&D, Johnson & Johnson.
Chair: Cristiana Mayer, Janssen R&D, Johnson & Johnson.

3:30 PM Bayesian approach in Proof-of-Concept in drug development: a case study
Michael Lee. Janssen

3:55 PM A Bayesian approach to evaluate program success for programs with multiple trials
♦*Meihua Wang and Guanghan(Frank) Liu*. Merck & Co.

4:20 PM Bayesian Benefit-Risk Assessments for Medical Products
Telba Irony. Center for Biologics Evaluation and Research - FDA

4:45 PM Discussant: Rong (Rachel) Chu, Agensys, Inc.

5:10 PM Floor Discussion.

Session 87: The Keys to Career Success as a Biostatistician (*Invited*)
Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)
Organizers: Jason Liao, Merck; Helena Fan, The Lotus Group.
Chair: Lei Wang, The Lotus Group.

3:30 PM The Keys to Career Success as a Biostatistician
Lisa Lupinacci. Merck & Co., Inc.

3:50 PM Making Sense of Data: Challenges of Being a Non-Clinical Biostatistician
♦*Binbing Yu and Harry Yang*. MedImmune

4:10 PM You Are Statistically Significant–A career path for biostatisticians to make a difference
Helena Fan. The Lotus Group LLS

4:30 PM Discussant: Bo Yang, Vertex

4:50 PM Discussant: Wei Shen, Eli Lilli and Company

5:10 PM Floor Discussion.

Session 88: Design of Experiments II (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Abhyuday Mandal, University of Georgia.
Chair: Weng Kee Wong, University of California, Los Angeles.

3:30 PM Partial Aliasing Relations in Mixed Two- and Three-Level Designs
Arman Sabbaghi. Purdue University Department of Statistics

3:55 PM Designing covering arrays for testing statistical software
♦*Ryan Lekivetz and Joseph Morgan*. JMP Division of SAS

4:20 PM On Orthogonality through the block factor
Sunanda Bagchi. Indian Statistical Institute, Bangalore

4:45 PM Using fractional factorials to obtain efficient designs for certain generalized linear models
John Stufken. Arizona State University

5:10 PM Floor Discussion.

Session 89: Statistical Phylogenetics (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Liang Liu, University of Georgia.
Chair: Liang Liu, University of Georgia.

3:30 PM Bayesian Analysis of Evolutionary Divergence with Genomic Data Under Diverse Demographic Models
♦*Yujin Chung and Jody Hey*. Temple University

3:55 PM Displayed trees do not determine distinguishability under the network multispecies coalescent
♦*James Degnan*[1] *and Sha Zhu*[2]. [1]University of Canterbury [2]Oxford University

4:20 PM Likelihood Estimation of Large Species Trees Using the Coalescent Process
Arindam RoyChoudhury. Columbia University

4:45 PM Mechanistic Models for the Retention of Duplicate Genes in Phylogenetic Birth-Death Processes
David Liberles. Temple University

5:10 PM Floor Discussion.

Session 90: Adaptive Methods and Regulation for Clinical Trials (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zhengjia (Nelson) Chen, Emory University.
Chair: Zhengjia (Nelson) Chen, Emory University.

3:30 PM Group sequential trial design under parametric survival model
♦*Jianrong Wu* [1] *and Xiaoping Xiong*[2]. [1]St. Jude Children's Research Hospital [2]St. Jude Children's Research Hospital

3:55 PM A Robust Bayesian Dose-Finding Design for Phase I/II Clinical Trials
♦*Suyu Liu*[1] *and Valen Johnson*[2]. [1]The UT MD Anderson Cancer Center [2]Texas A&M University

4:20 PM Continual reassessment method with multiple toxicity constraints
♦*Bin Cheng and Shing Lee*. Columbia University

4:45 PM The challenges and opportunities of adaptive designs in medical device studies
♦*Jie (Jack) Zhou, Hong (Laura) Lu and Xiting (Cindy) Yang*. FDA Center for Devices and Radiological Health

5:10 PM Floor Discussion.

Session 91: Recent Developments of Nonparametric Methods for High-Dimensional Data (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Guanqun Cao, Auburn University.
Chair: Guanqun Cao, Auburn University.

3:30 PM Variable Selection for Semiparametric Geospatial Models
♦*Guannan Wang*[1] *and Lily Wang*[2]. [1]College of William & Mary [2]Iowa State University

3:55 PM Optimal Prediction for Functional Linear Regression with A Functional Response
Xiaoxiao Sun[1], *Xiao Wang*[2], *Ping Ma*[1] *and* ♦*Pang Du*[3]. [1]University of Georgia [2]Purdue University [3]Virginia Tech

4:20 PM Ultra-high Dimensional Additive Partial Linear Models
♦*Xinyi Li, Li Wang and Dan Nettleton*. Iowa State University

4:45 PM Weighing Schemes for Functional Data
♦*Xiaoke Zhang*[1] *and Jane-Ling Wang*[2]. [1]University of Delaware [2]University of California, Davis

5:10 PM Floor Discussion.

Session 92: The Analysis of Complex Time-to-Event Data (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ronghui Xu, University of California, San Diego.
Chair: Pengsheng Ji, University of Georgia.

3:30 PM The Analysis of Spontaneous Abortion with Left Truncation, Partly Interval Censoring and Cure Rate
♦*Yuan Wu*[1] *and Ronghui Xu*[2]. [1]Duke University [2]UCSD

3:55 PM Residual-Based Model Diagnosis Methods for Mixture Cure Models
♦*Yingwei Peng*[1] *and Jeremy Taylor*[2]. [1]Queen's University [2]University of Michigan

4:20 PM A New Coefficient of Determination for Regression Models
Chun Li. Case Western Reserve University

4:45 PM The Analysis and Desgin of Cancer Clinical Trials Based on Progression Free Survival
♦*Leilei Zeng*[1], *Yidan Shi*[1], *Lan Wen*[2] *and Richard Cook* [1]. [1]University of Waterloo [2]MRC Biostatistics Units

5:10 PM Floor Discussion.

Session 93: Recent Developments of Graphical Model for Big Data (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ningtao Wang, University of Texas School of Public Health.
Chair: Lingzhou Xue, Pennsylvania State University.

3:30 PM Estimation and Inference for Dynamic Network Models for High-Dimensional Time Course Data
Hulin Wu. University of Texas Health Science Center at Houston

3:55 PM Nonparametric mixture of Gaussian graphical models, with applications to ADHD imaging data
♦*Kevin Lee and Lingzhou Xue*. Pennsylvania State University

4:20 PM Consistent Estimation of Curved Exponential-Family Random Graph Models with Local Dependence and Growing Neighborhoods
Michael Schweinberger. Rice University

4:45 PM Composite Likelihood Inference on Stochastic Block Model for Big Networks
Ningtao Wang. University of Texas School of Public Health

5:10 PM Floor Discussion.

Session 94: Recent Developments in Statistics (*Invited*)
Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zhezhen Jin, Columbia University.
Chair: Zhezhen Jin, Columbia University.

3:30 PM Simulating Longer Vectors of Correlated Binary Random Variables via Multinomial Sampling
Justine Shults. University of Pennsylvania

3:55 PM Stochastic Analysis of the Cost-Effectiveness Frontier
♦*Daniel Heitjan*[1] *and Yu Lan*[2]. [1]SMU/UTSW [2]SMU

4:20 PM On the uncertainty of data extrapolation in pediatric drug development
Alan Chiang. Eli Lilly and Company

4:45 PM The Role of Kernels in Data Analysis: A Statistical Perspective
Marianthi Markatou. University at Buffalo

5:10 PM Floor Discussion.

Session 95: Recent Advances in Biomarker Evaluation and Risk Prediction (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Dandan Liu, Vanderbilt University.
Chair: Dandan Liu, Vanderbilt University.

3:30 PM Efficient Epidemiological Study Designs for Quantitative Longitudinal Data
Jonathan Schildcrout. Vanderbilt University

3:55 PM Least Squares Regression Methods for Clustered ROC Data with Discrete Covariates
♦*Liansheng Tang*[1], *Wei Zhang*[2], *Qizhai Li*[2], *Xuan Ye*[3] *and Leighton Chan*[4]. [1]NIH and George Mason University [2]Chinese Academy of Sciences [3]George Mason University and FDA [4]National Institute of Health

4:20 PM Genetic-based Prediction of Vaccine-derived Poliovirus Circulation
♦*Kun Zhao, Jaume Jorba, Jane Iber, Qi Chen, Kelley Bullard, Olen Kew and Cara Burns*. Centers for Disease Control and Prevention

4:45 PM Novel diagnostic accuracy analysis for competing risks outcomes with ROC surface

♦*Song Zhang and Yu Cheng*. University of Pittsburgh

5:10 PM Floor Discussion.

Session 96: Methods of Integrating Novel Functional Data with Multimodal Measurements (*Invited*)

Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3

Organizer: Haochang Shou, University of Pennsylvania.

Chair: Gen Li, Columbia University.

3:30 PM Three-Part Joint Modeling Methods for Complex Functional Data

♦*Haocheng Li*[1]*, John Staudenmayer*[2]*, Tianying Wang*[3] *and Carroll Raymond*[3]. [1]University of Calgary [2]University of Massachusetts [3]Texas A&M University

3:55 PM Understanding the time-varying associations between two functional measurements

♦*Haochang Shou*[1]*, Simon Simon Vandekar*[1]*, Lihong Cui*[2]*, Vadim Zipunnikov*[3] *and Kathleen Merikangas*[2]. [1]University of Pennsylvania [2]National Institutes of Mental Health [3]Johns Hopkins University

4:20 PM Prediction Models of Dimentia Transition using longitudinal structural brain images

♦*Seonjoo Lee*[1]*, Liwen Wu*[2] *and Yaakov Stern*[2]. [1]NYSPI and Columbia University [2]Columbia University

4:45 PM Optimal Design for Sparse Functional Data

♦*So Young Park*[1]*, Luo Xiao*[1]*, Jayson Wilbur*[2] *and Ana-Maria Staicu*[1]. [1]North Carolina State University [2]METRUM Research Group

5:10 PM Floor Discussion.

Session 97: Recent Statistical Methodology Developments for Medical Diagnostics (*Invited*)

Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3

Organizers: Dongliang Wang, SUNY Upstate Medical University; Jingjing Yin, Georgia Southern University.

Chair: Robert L. Vogel, Georgia Southern University.

3:30 PM Combining large number of weak biomarkers based on AUC

♦*Li Yan*[1]*, Lili Tian*[2] *and Song Liu*[1]. [1]RoswellPark Cancer Institute [2]SUNY University at Buffalo

3:55 PM Comparing two correlated diagnostic tests based on joint testing of the AUC and the Youden index

♦*Jingjing Yin*[1]*, Lili Tian*[2] *and Hani Samawi*[1]. [1]Biostatistics, Georgia Southern University [2]Biostatistics, University at Buffalo

4:20 PM Empirical Likelihood Confidence Regions in the Evaluation of Medical Tests with Verification Bias

♦*Gengsheng Qin*[1] *and Binhuan Wang*[2]. [1]Georgia State University [2]New York University

4:45 PM MLEs for diagnostic measures of accuracy by log-linear models for correction of workup bias.

♦*Haresh Rochani, Hani Samawi, Robert Vogel and Jingjing Yin*. Jian Ping Hsu College of Public Health

5:10 PM Floor Discussion.

Session 98: Combining Information in Biomedical Studies (*Invited*)

Room: FAIRLIE, ATLANTA CONFERENCE CENTER - LL3

Organizers: Xiaoyi Min, Georgia State University; Wei Yang, University of Pennsylvania.

Chair: Rui Feng, University of Pennsylvania.

3:30 PM Inference Using Combined Longitudinal and Survival Data in CKD

Wei Yang. University of Pennsylvania

3:55 PM An adaptive Fisher's method for combining information across samples

♦*Xiaoyi Min*[1]*, Chi Song*[2] *and Heping Zhang*[3]. [1]Georgia State University [2]Ohio State University [3]Yale University

4:20 PM Bayesian Latent Hierarchical Model for Transcriptomic Meta-Analysis

Zhiguang Huo[1], ♦*Chi Song*[2] *and George Tseng*[1]. [1]University of Pittsburgh [2]The Ohio State University

4:45 PM Genetic Effect and Association Test for Covariance Heterogeneity in Multiple Trait Comorbidity

Yuan Jiang[1], ♦*Yaji Xu*[2] *and Heping Zhang*[3]. [1]Oregon State University [2]Food and Drug Administration [3]Yale University

5:10 PM Floor Discussion.

Session 99: New Frontiers in Genomics and Precision Medicine (*Invited*)

Room: COURTLAND, ATLANTA CONFERENCE CENTER - LL3

Organizer: Chad He, Fred Hutchinson Cancer Research Center.

Chair: Dengdeng?Yu, University of Alberta.

3:30 PM Estimating interactions between a treatment and a large number of genomic features

James Dai. Fred Hutchinson Cancer Research Center, Seattle

3:55 PM New approaches for genetic association mapping with large-scale genetic data in diverse populations

♦*Timothy Thornton and Caitlin McHugh*. University of Washington

4:20 PM A Statistical Framework for Pathway and Gene Identification from Integrative Analysis

♦*Quefeng Li*[1]*, Menggang Yu*[2] *and Sijian Wang*[2]. [1]University of North Carolina, Chapel Hill [2]University of Wisconsin, Madison

4:45 PM Integrated analysis of DNA methylation and gene expression data in human aging

♦*Karen Conneely*[1]*, Elizabeth Kennedy*[1]*, Alicia Smith*[2]*, Elisabeth Binder*[3] *and Kerry Ressler*[4]. [1]Department of Human Genetics, Emory University [2]Department of Psychiatry, Emory University [3]Department of Psychiatry, Max Planck Institute [4]Department of Psychiatry, Harvard University

5:10 PM Floor Discussion.

Session 100: New Developments in BFF Inferences in the Era of Data Science (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Min-ge Xie, Rutgers University.
Chair: Jeff Simonoff, New York University.

3:30 PM Fiducial Inference: Fisher's Big Blunder or Big Bang?
♦*Keli Liu*[1] *and Xiao-Li Meng*[2]. [1]Stanford University [2]Harvard University

3:55 PM Applications of the Poisson Dempster-Shafer Model (DSM) and the General Univariate DSM for Inference of Infection Time from Acute HIV-1 Genomes
Paul Edlefsen. Fred Hutchinson Cancer Research Center

4:20 PM Fusion Learning: combining inferences using data depth and confidence distribution
♦*Dungang Liu*[1]*, Regina Liu*[2] *and Min-ge Xie*[2]. [1]University of Cincinnati [2]Rutgers University

4:45 PM Discussant: Ryan Martin, University of Illinois at Chicago

5:10 PM Floor Discussion.

Session 101: Topics in Statistics (*Contributed*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Chair: Xiaoli Gao, The University of North Carolina at Greensboro.

3:30 PM Modelling cumulative effects of air pollution on respiratory illnesses
Xingfa Zhang. Guangzhou University China

3:45 PM Statistical Analysis of cochlear shape
♦*jean michel loubes, jose braga and laurent risser*. université de toulouse

4:00 PM Improving Online Education by Understanding Demographic Effects on Student Success and Retention
James Monroe. Kennesaw State University

4:15 PM A Mixed Variance Component Model for Quantifying the Elasticity Modulus of Nanomaterials
♦*Angang Zhang*[1] *and Xinwei Deng*[2]. [1]Merck [2]Virginia Tech

4:30 PM The power of Rmarkdown and RStudio IDE: Lessons Learned Teaching R in Public Health and Nursing
Melinda Higgins. Emory University

4:45 PM Different estimations of time series models and application for foreign exchange in emerging markets
Jingjing Wang. Student

5:00 PM Floor Discussion.

Session 102: Topics in Biostatistics (*Contributed*)
Room: BAKER, ATLANTA CONFERENCE CENTER - LL3
Chair: Yong Chen, University of Pennsylvania.

3:30 PM A Maximum Likelihood Approach for Non-invasive Cancer Diagnosis Using Methylation Profiling of Blood
♦*Carol Sun*[1] *and Wenyuan Li*[2]. [1]Oak Park High School [2]University of Southern California

3:45 PM Parametric Bootstrap in Meta-analyses to Construct Cls for Event Rates and Differences in Event Rate
♦*Gaohong Dong*[1]*, Jennifer Ng*[2]*, Steffen Ballerstedt*[3] *and Marc Vandemeulebroecke*[3]. [1]iStats Inc. and Infotree Service Inc. [2]Novartis Pharmaceuticals Corporation [3]Novartis Pharma AG

4:00 PM Estimation of Energy Expenditure
Shan Yang . Merck & Co., Inc.

4:15 PM Meta-Analysis of Rare Binary Events in Treatment Groups with Unequal Variability
♦*Lie Li*[1]*, Ou Bai and Xinlei Wang*. [1]Southern Methodist University

4:30 PM Combining Evidence of Regional Treatment Effects under Discrete Random-Effects Model (DREM) in MRCT
♦*Hsiao-Hui Tsou*[1]*, K. K. Gordon Lan*[2]*, Chi-Tian Chen*[1]*, H.M. James Hung*[3]*, Chin-Fu Hsiao*[1] *and Jung-Tzu Liu*[1]. [1]National Health Research Institutes [2]Janssen Pharmaceutical Companies of Johnson & John [3]US Food and Drug Administration

4:45 PM Meta-analysis with incomplete multinomial data: An application to tumor response in cancer patients
♦*Charity J. Morgan, Pooja Ghatalia and Guru Sonpavde*. University of Alabama at Birmingham

5:00 PM Floor Discussion.

Wednesday, June 15. 8:30 AM - 10:10 AM

Session 103: New Development on Missing Data Problems (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Sixia Chen, University of Oklahoma.
Chair: Jing Qian, University of Massachusetts.

8:30 AM Multiply robust imputation procedures for the treatment of item nonresponse in surveys
♦*Sixia Chen*[1] *and David Haziza*[2]. [1]University of Oklahoma [2]University of Montreal

8:55 AM Empirical Likelihood Methods for Complex Surveys with Data Missing-by-Design
♦*Changbao Wu, Min Chen and Mary Thompson*. University of Waterloo

9:20 AM Pseudo-population bootstrap methods for imputed survey data
♦*David Haziza*[1]*, Zeinab Mashreghi*[2] *and Christian Léger*[1]. [1]Université de Montréal [2]University of Winninpeg

9:45 AM Lack-of-fit testing of a regression model with response missing at random
Xiaoyu Li. Auburn University

10:10 AM Floor Discussion.

Session 104: Advances in Ultra High Dimensional Data Analysis (*Invited*)
Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)
Organizer: Guoqing Diao, George Mason University.
Chair: Yuexiao Dong, Temple University.

8:30 AM Divergence based Inference for High-dimensional Regression Problems: Uncertainty Assessment, Robustn

Anand Vidyashankar. George Mason University

8:55 AM Tests for Nonparametric Interactions Using Random Forests

♦*Giles Hooker*[1] *and Lucas Mentch*[2]. [1]Cornell University [2]SAMSI

9:20 AM High Dimensional Multivariate Testing with Applications in Neuroimaging

♦*Bret Hanlon*[1] *and Nagesh Adluru*[2]. [1]University of Wisconsin Statistics [2]Waisman Laboratory for Brain Imaging and Behavior

9:45 AM Conditional Variable Screening in High-Dimensional Binary Classification

♦*Guoqing Diao*[1] *and Jing Qin*[2]. [1]Department of Statistics, George Mason University [2]National Institutes of Health, NIAID

10:10 AM Floor Discussion.

Session 105: Mixture regression: new methods and applications (*Topic Contributed*)

Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)

Organizer: Tapio Nummi, University of Tampere, Finland.

Chair: Chi Song, Ohio State University.

8:30 AM Order Dependence of Hypersphere Decomposition for Covariance Matrix

♦*Qingze Li and Jianxin Pan*. University of Manchester, UK

8:50 AM Testing of multivariate spline growth model

♦*Jyrki Mottonen*[1] *and Tapio Nummi*[2]. [1]University of Helsinki [2]University of Tampere

9:10 AM Labor market attachment in early adulthood: A trajectory analysis approach

♦*Janne Salonen*[1]*, Tapio Nummi*[2]*, Antti Saloniemi*[2] *and Pekka Virtanen*[2]. [1]Finnish Centre for Pensions [2]University of Tampere

9:30 AM A semiparametric mixture regression model for longitudinal data

♦*Tapio Nummi*[1]*, Janne Salonen*[2]*, Lasse Koskinen*[1] *and Jianxin Pan*[3]. [1]University of Tampere, Finland [2]The Finnish Centre of Pensions, Finland [3]University of Manchester, UK

9:50 AM Discussant: Tim O'Brien, Loyola University Chicago

10:10 AM Floor Discussion.

Session 106: Spatial and Spatio-temporal Statistical Modeling and their Applications (*Invited*)

Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3

Organizer: Emily L. Kang, University of Cincinnati.

Chair: Ganggang Xu, Binghamton University, The State University of New York.

8:30 AM Disease Risk Estimation by Combining Case Control Data with Aggregated Information on Population at Risk

♦*Xiaohui Chang*[1]*, Rasmus Waagepetersen*[2]*, Herbert Yu*[3]*, Xiaomei Ma*[4]*, Theodore Holford*[4]*, Rong Wang*[4] *and Yongtao Guan*[5]. [1]College of Business, Oregon State University [2]Dept of Mathematical Sciences, Aalborg University [3]Epidemiology Program, University of Hawaii Cancer [4]Yale School of Public Health [5]Dept of Management Science, University of Miami

8:55 AM Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process

♦*Jonathan Bradley*[1]*, Christopher Wikle*[2] *and Scott Holan*[2]. [1]Florida State University [2]University of Missouri

9:20 AM Changes in Spatio-temporal Precipitation Patterns in Changing Climate Conditions

♦*Won Chang*[1]*, Michael Stein*[1]*, Jiali Wang*[2]*, Rao Kotamarthi*[2] *and Elisabeth Moyer*[1]. [1]University of Chicago [2]Argonne National Laboratory

9:45 AM Estimating the Health Effects of Ambient Air Pollution Accounting for Spatial Exposure Uncertainty

♦*Howard Chang, Yang Liu and Stefanie Sarnat*. Emory University

10:10 AM Floor Discussion.

Session 107: Recent Development in Sufficient Dimension Reduction and Variable Selection (*Invited*)

Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3

Organizer: Qin Wang, Virginia Commonwealth University.

Chair: Qin Wang, Virginia Commonwealth University.

8:30 AM Variable selection via additive conditional independence

♦*Kuang-Yao Lee*[1]*, Bing Li*[2] *and Hongyu Zhao*[1]. [1]Yale University [2]Pennsylvania State University

8:55 AM A BAYESIAN APPROACH FOR ENVELOPE MODELS

Kshitij Khare[1]*, Subhadip Pal*[2] *and* ♦*Zhihua Su*[1]. [1]University of Florida [2]Emery University

9:20 AM Pseudo Estimation for High Dimensional Data

♦*Wenbo Wu*[1] *and Xiangrong Yin*[2]. [1]University of Oregon [2]University of Kentucky

9:45 AM On the second-order inverse regression methods for a general type of elliptical predictors

Wei Luo. Baruch College

10:10 AM Floor Discussion.

Session 108: New Approaches in Dimension Reduction for Modern Data Applications (*Invited*)

Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3

Organizer: Xin Zhang, Florida State University.

Chair: Xin Zhang, Florida State University.

8:30 AM Generalized Mahalanobis Depth in Point Process Data

Shuyi Liu and ♦*Wei Wu*. Florida State University

8:55 AM On the Estimation of Ultra-High Dimensional Semiparametric Gaussian Copula Models

Qing Mai. Florida State University

9:20 AM Supervised Integrative Principal Component Analysis
♦*Gen Li*[1] *and Sungkyu Jung*[2]. [1]Columbia University, Department of Biostatistics [2]Pittsburgh University, Department of Statistics

9:45 AM A modern optimization perspective on iterative proportional scaling for large tables and count data
♦*Yiyuan She and Shao Tang*. Florida State University

10:10 AM Floor Discussion.

Session 109: Deep Dive on Multiplicity Issues in Clinical Trials. (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizers: Weihua Tang, Biogen; Li-An Xu, BMS.
Chair: Weihua Tang, Biogen.

8:30 AM Use of intermediate endpoint in Phase II/III adaptive designs
♦*Xiaoyun (Nicole) Li and Cong Chen*. Merck& Co.

8:55 AM Controlling Overall Type I Error with Hierarchical Testing Procedure: something not obvious
Li-an Xu. Bristol-Myers Squibb

9:20 AM Power and sample size calculation in graphical approaches
♦*Dong Xi, Willi Maurer, Ekkehard Glimm and Frank Bretz*. Novartis

9:45 AM Discussant: Steve Bai, FDA

10:10 AM Floor Discussion.

Session 110: Adaptive Designs in Clinical Trials (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Samuel S. Wu, University of Florida.
Chair: Samuel S. Wu, University of Florida.

8:30 AM Identifying Main Effects in Multi Factor Clinical Trials
♦*Abhishek Bhattacharjee*[1] *and Samuel Wu*[2]. [1]Department of Statistics, University of Florida [2]Department of Biostatistics, University of Florida

8:55 AM Interval Estimation in Multi-stage Drop-the-losers Designs
Xiaomin Lu[1], ♦*Ying He*[2] *and Samuel Wu*[1]. [1]University of Florida [2]Clarkson University

9:20 AM Graphical Approach to Multiple Test Procedures in 2×2 Factorial Designs
♦*Xiaomin Lu*[1], *John Kairalla*[1], *Hui Zeng*[2] *and Samuel Wu*[1]. [1]University of Florida [2]Pharmaceutical Product Development (PPD)

9:45 AM Classification of Subtypes of Cancers Using Neural Networks and Gene Expression Data
Lan Gao. The University of Tennessee at Chattanooga

10:10 AM Floor Discussion.

Session 111: Shape Analysis in Applications (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizers: Charles Hagwood, National Institute of Standards and Technology; Anuj Srivastava, Florida State University.
Chair: Jin-Ting Zhang, National University of Singapore.

8:30 AM Non-Euclidean Classification of Medically Imaged Objects via SRNF
Xiaoming Dong. Florida State University

8:55 AM A Novel Geometric Approach For Semiparametric Density Estimation
♦*Sutanoy Dasgupta*[1], *Debdeep Pati and Anuj Srivastava*. [1]Florida State University

9:20 AM An Analytical Approach for Computing Shape Geodesics
♦*Charles Hagwood*[1], *Javier Bernal*[1] *and Gunay Dogan*[2]. [1]NIST [2]Theiss Research and NIST

9:45 AM Shape metrology and its applications in Medicine and Forensics
Z.Q. John Lu. National Institute of Standards and Technology

10:10 AM Floor Discussion.

Session 112: Bias Reduction and Subgroup Identification in Observational Studies (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Stan Young, CGStat LLC.
Chair: Yichuan Zhao, Georgia State University.

8:30 AM Local Control Strategy: Can Current Indoor Radon Levels Cause Lung Cancer Mortality?
♦*Robert L. Obenchain*[1] *and S. Stanley Young*[2]. [1]Risk Benefit Statistics [2]Omicsoft

8:55 AM Time Series Smoother
♦*Cheng You*[1], *Dennis Lin*[1] *and S. Stanley Young*[2]. [1]The Pennsylvania State University [2]CGStat

9:20 AM Reliability of a Meta-analysis of Observational Studies
Kumer Das[1], *Adam Fadhli-Theis*[2], ♦*Allen Heller*[3] *and S. Stanley Young*[4]. [1]Lamar University [2] Lamar University [3]Pharma Study Design LLC [4] CGStat LLC

9:45 AM Discussant: Stan Young, CGStat LLC

10:10 AM Floor Discussion.

Session 113: Advance in Statistical Method on Complex Data and Applications in Statistical Genomics (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizers: Shiyao Liu, Genentech; Wen Zhou, Colorado State University.
Chair: Shiyao Liu, Genentech.

8:30 AM Joint Estimation of Multiple Dependent Gaussian Graphical Models with Applications to Mouse Genomics
♦*Yuying Xie*[1], *Yufeng Liu*[2] *and William Valdar*[2]. [1]Michgan State University [2]University of North Carolina at Chapel Hill

8:55 AM Identification of Pairwise Informative Features for Clustering Data with Growing Dimension
♦*Lulu Wang, Wen Zhou and Jennifer Hoeting*. Colorado State University

9:20 AM Detect chromatin interaction from multiple Hi-C datasets by hierarchical hidden Markov random model
♦*Zheng Xu*[1], *Ming Hu*[2] *and Yun Li*[1]. [1]University of North Carolina at Chapel Hill [2]New York University School of Medicine

9:45 AM Floor Discussion.

Session 114: Statistical Issues in EHR Data (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yong Chen, University of Pennsylvania.
Chair: Rui Duan, University of Pennsylvania.

8:30 AM Comparative effectiveness of dynamic treatment strategies using electronic health records and the pa
Miguel Hernán. Harvard TH Chan School of Public Health

8:55 AM Accounting for Error and Misclassification in Time to Event Analyses Using EHR-derived Endpoints
♦*Rebecca Hubbard*[1]*, Weiwei Zhu*[2]*, Le Wang*[1] *and Jessica Chubak*[2]. [1]University of Pennsylvania [2]Group Health Research Institute

9:20 AM Phenotype validation in Electronic Health Record based genetic association studies
Lu Wang, Jason Moore, Scott Damrauer and ♦*Jinbo Chen*. University of Pennsylvania

9:45 AM Robust methods for association studies in EHR data
Jing Huang, Rui Duan and ♦*Yong Chen*. University of Pennsylvania

10:10 AM Floor Discussion.

Session 115: Recent Advances in Integrative Analysis of Omics Data (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Sandra Safo, Emory University.
Chair: Eun Jeong Min, Emory University.

8:30 AM Exploratory Factorization of Multi-Source Data
♦*Eric Lock and Michael OConnell*. University of Minnesota, Division of Biostatistics

8:55 AM A Bayesian approach for the integrative analysis of omics data: A kidney cancer case study
♦*Thierry Chekouo*[1]*, Francesco Stingo*[1]*, James Doecke*[2] *and Kim-Anh Do*[1]. [1]U.T MD Anderson Cancer Center [2]CSIRO Australia

9:20 AM Integrative analysis of multiple omics data with biological information.
♦*Sandra Safo, Shuzhao Li and Qi Long*. Emory University

9:45 AM A Full Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-type Omics data
Qianxing Mo. Baylor College of Medicine

10:10 AM Floor Discussion.

Session 116: Survival Analysis (*Contributed*)
Room: HANOVER C, EXHIBIT LEVEL - LL2
Chair: Chen-Hsin Chen, Academia Sinica.

8:30 AM A frailty model for recurrent events of the same type during alternating restraint and non-restraint
♦*Xiaoqi Li*[1]*, Yong Chen*[2] *and Ruosha Li*[3]. [1]Baylor College of Medicine [2]University of Pennsylvania [3]University of Texas School of Public Health

8:45 AM Variable Selection for Interval-Censored Survival Data Under the Proportional Hazards Model
♦*Tyler Cook*[1] *and Jianguo Sun*[2]. [1]University of Central Oklahoma [2]University of Missouri

9:00 AM The Spike-and-Slab lasso Cox Model for Survival Prediction and Associated Genes Detection
♦*Zaixiang Tang*[1] *and Nengjun Yi*[2]. [1]Department of biostatistics, Soochow University [2]Department of Biostatistics,UAB

9:15 AM Variable Selection for Mixture and Promotion Time Cure Rate Models
♦*Abdullah Masud and Zhangsheng Yu*. Indiana University

9:30 AM Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach
♦*Xinyan Zhang*[1]*, Yan Li*[1]*, Tomi Akinyemiju*[1]*, Akinyemi I. Ojesina*[1]*, Phillip Buckhaults*[2]*, Bo Xu*[3] *and Nengjun Yi*[1]. [1]University of Alabama at Birmingham [2]The University of South Carolina [3]Southern Research Institute

9:45 AM Application of Heavy-Tailed Probability Distributions on Financial Markets
Xing Yang. Jackson State University

10:00 AM Floor Discussion.

Session 117: Advances in Hypothesis Testing (*Contributed*)
Room: HANOVER D, EXHIBIT LEVEL - LL2
Chair: Jing Zhang, Georgia State University.

8:30 AM Methods for f lack of fit test without replicates
Tyler George. Central Michigan University

8:45 AM Generalized Inference for the Reliability of of Stress-Strength Model for the Inverse Exponentials
Sumith Gunasekera. The University of Tennessee at Chattanooga

9:00 AM A New Approach to Multiple Testing of Grouped Hypotheses
Yanping Liu, Sanat Sarkar and ♦*Zhigen Zhao*. Temple University

9:15 AM Non-zero tests on heterogeneity variance in the quantitative synthesis of cancer detection biomarker
Hanfei Xu and ♦*Kepher Makambi*. Georgetown University

9:30 AM A Group of New F-Tests for Multiple Mean Comparisons and Their Applications in Medical Research
Jiajuan Liang. University of New Haven, U.S.A.

9:45 AM Testing the Presence of Significant Covariates Through Conditional Marginal Regression
♦*Yanlin Tang, Huixia Wang and Emre Barut*. The George Washington University

10:00 AM Floor Discussion.

Session 118: Statistical Analysis of Complex Data I (*Contributed*)
Room: HANOVER E, EXHIBIT LEVEL - LL2
Chair: Ye Liang, Oklahoma State University.

8:30 AM Spatial Data Fusion for large non-Gaussian Remote Sensing Datasets
♦*Hongxiang Shi and Emily Kang*. University of Cincinnati

8:45 AM INCORPORATING GEOSTROPHIC WIND INFORMATION FOR IMPROVED SPACE-TIME SHORT-TERM WIND SPEED FORECASTING

♦*Xinxin Zhu*[1]*, Kenneth Bowman and Marc Genton*. [1]Merial

9:00 AM LESA: Longitudinal Elastic Shape Analysis of Brain Subcortical Structures

♦*Zhengwu Zhang*[1]*, Anuj Srivastava*[2]*, Jingwen Zhang*[3]*, Martin Styner*[3]*, Weili Lin*[3] *and Zhu Hongtu*[3]. [1]SAMSI [2]Florida State University [3]The University of North Carolina at Chapel Hill

9:15 AM Model based heritability scores for high-throughput sequencing data

Pratyaydipta Rudra[1]*,* ♦*W. Jenny Shi*[2]*, Brian Vestal*[1]*, Pamela Russel*[1]*, Laura Saba*[3] *and Katerina Kechris*[1]. [1]University of Colorado School of Public Health [2]University of Colorado School of Medicine [3]University of Colorado Skaggs School

9:30 AM Robust Functional Linear Models

♦*Melody Denhere*[1]*, Nedret Billor*[2] *and Huybrechts Bindele*[3]. [1]University of Mary Washington [2]Auburn University [3]University of South Alabama

9:45 AM Floor Discussion.

Session 119: Nonparametric and Semiparametric Methods (*Contributed*)

Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3

Chair: Yi-Ran Lin, Academia Sinica.

8:30 AM Covariate Adjusted Cross Ratio Estimation

♦*Ran Liao*[1]*, Tianle Hu*[2] *and Sujuan Gao*[1]. [1]Indiana University [2]Eli Lilly

8:45 AM Predicting the Timing of the Final Event in a Clinical Trial using the Bayesian Bootstrap and beyond

♦*Marc Sobel*[1] *and Ibrahim Turkoz*[2]. [1]Temple University [2]Janssen Research and Development

9:00 AM Modeling Hourly Electricity Demand Using Spline-Based Nonparametric Transfer Function Models

Jun Liu. Georgia Southern University

9:15 AM A Differences in Differences Approach for Bias Reduction in Semiparametric Models

♦*Chan Shen*[1] *and Roger Klein*[2]. [1]UT MD Anderson Cancer Center [2]Rutgers University

9:30 AM Exploiting Variance Reduction Potential in Local Gaussian Process Search

♦*Chih-Li Sung*[1]*, Robert Gramacy*[2] *and Benjamin Haaland*[1]. [1]Georgia Institute of Technology [2]The University of Chicago

9:45 AM Semi-Parametric Estimation for Multivariate Skew-Elliptical Distributions

Jing Huang. European School of Management and Technology

10:00 AM Floor Discussion.

Wednesday, June 15. 10:30 AM-12:10 PM

Session 120: Recent Advances in Network Data Inference (*Invited*)

Room: INTERNATIONAL SOUTH, INTERNATIONAL TOWER (LL1)

Organizer: Emma Jingfei Zhang, University of Miami.

Chair: Haoda Fu, Eli Lilly and Company.

10:30 AM Network Inference From Time Varying Grouped Observations

Yunpeng Zhao. George Mason University

10:55 AM Studying the Communication Network on Facebook of French Election with Spectral Contextualization

♦*Yilin Zhang, Karl Rohe and Chris Wells*. University of Wisconsin-Madison

11:20 AM Graph-limit Enabled Fast Computation for Fitting Exponential Random Graph Models to Large Networks

♦*Ran He*[1] *and Tian Zheng*[2]. [1]Columbia University, Bell Labs [2]Columbia University

11:45 AM A Hypothesis Testing Framework for Modularity Based Network Community Detection

♦*Jingfei Zhang*[1] *and Yuguo Chen*[2]. [1]University of Miami [2]University of Illinois at Urbana-Champaign

12:10 PM Floor Discussion.

Session 121: Recent Developments in Design of Experiments (*Invited*)

Room: INTERNATIONAL NORTH, INTERNATIONAL TOWER (LL1)

Organizer: Wei Zheng, Indiana University-Purdue University Indianapolis.

Chair: John Stufken, Arizona State University.

10:30 AM Optimal designs for the two-dimensional interference model

Wei Zheng and ♦*Heng Xu*. Indiana University Purdue University Indianapolis

10:55 AM Maximum Empirical Likelihood Estimation in U-statistics-based General Estimating Equations

♦*Lingnan Li and Hanxiang Peng*. Indiana University-Purdue University, Indianapolis

11:20 AM Bayesian D-Optimal Design of Experiments with Quantitative and Qualitative Responses

Lulu Kang[1]*,* ♦*Xinwei Deng*[2] *and Ran Jin*[2]. [1]Illinois Institute of Technology [2]Virginia Tech

11:45 AM BAYESIAN SEQUENTIAL DATA COLLECTION FOR CALIBRATING QUALITATIVE VARIABLES

♦*Qiong Zhang and yongjia song*. Virginia Commonwealth University

12:10 PM Floor Discussion.

Session 122: Design and Analysis of Traditional Therapy Studies (*Invited*)

Room: GREENBRIAR, ATLANTA CONFERENCE CENTER - LL3

Organizer: Chenguang Wang, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University.
Chair: Chenguang Wang, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University.

10:30 AM Assess the Accuracy of Diagnosis and Syndrome Differentiation Results Made by Traditional Medicine P
♦*Zheyu Wang*[1] *and Andrew Zhou*[2]. [1]Johns Hopkins University [2]University of Washington

10:55 AM Evaluating Traditional Chinese Medicine using Modern Clinical Design and Statistical Methodology
Lixing Lao[1], ♦*Yi Huang*[2], *Chiguang Feng*[3], *Brian Berman*[3] *and Ming Tan*[4]. [1]The Univ. of Hong Kong, School of Chinese Medicine [2]University of Maryland, Baltimore County [3]University of Maryland, School of Medicine [4]Georgetown University, Medical Center

11:20 AM Acupuncture and Prevention of Chronic Postsurgical Pain: From Experimental Study to Clinical Trial
Jiangang Song. Shanghai University of Traditional Chinese Medicine

11:45 AM Discussant: Sejong Bae, University of Alabama at Birmingham

12:10 PM Floor Discussion.

Session 123: Statistical Analysis of Structural Morphology and Functional Measures in Medical Studies (*Invited*)
Room: PIEDMONT, ATLANTA CONFERENCE CENTER - LL3
Organizer: Chafik Samir, University of Clermont Ferrand, France.
Chair: Jean-Michel Loubes, University of Toulouse.

10:30 AM Statistical Shape Analysis of Anatomical Structures Using Square-Root Normal Fields
Sebastian Kurtek. The Ohio State University

10:55 AM Statistical Shape Analysis of Biological Morphology
Shantanu Joshi. University of California Los Angeles

11:20 AM Statistical Analysis and Simulations of Soft Tissue Shapes in Medical Studies
♦*Chafik Samir and Thomas Deregnaucourt*. University of Clermont Auvergne

11:45 AM FUNCTIONAL CLUSTERING BASED ON PROJECTION ESTIMATION WITH APPLICATION TO BLOOD-SAMPLE SPECTRUM
♦*Anne-Françoise Yao*[1] *and Gerald GREGORI* [2]. [1]Lab. Mathematics Clermont-Ferrand Univ(France) [2]MIO, Aix-Marseille University(France)

12:10 PM Floor Discussion.

Session 124: Adaptive Randomization: Recent Advances in Theory and Practice (*Invited*)
Room: TECHWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yanqing Hu, West Virginia University.
Chair: Yanqing Hu, West Virginia University.

10:30 AM Utility of Outcome Adaptive Randomization for Multisite Comparative Trials
♦*Mi-Ok Kim, Chunyan Liu and Nusrat Harun*. Cincinnati Children's Hospital Medical Center

10:55 AM OPTIMAL FLEXIBLE SAMPLE SIZE DESIGN WITH ROBUST POWER
♦*Lanju Zhang*[1], *Lu Cui*[1] *and Bo Yang*[2]. [1]AbbVie Inc [2]Vertex Pharmaceuticals

11:20 AM Comparing efficiency, randomness of adaptive treatment allocation procedures in clinical trials
Li-Xin Zhang . Zhejiang University

11:45 AM Covariate-adjusted response-adaptive deigns and their statistical inference
Wanying Zhao. George Washington University

12:10 PM Floor Discussion.

Session 125: ROC Analysis and Estimation of Large Covariance Matrices (*Invited*)
Room: INMAN, ATLANTA CONFERENCE CENTER - LL3
Organizer: Gengsheng (Jeff) Qin, Georgia State University.
Chair: Gengsheng (Jeff) Qin, Georgia State University.

10:30 AM Statistical Inferences for the Partial Youden Index
♦*Chenxue Li*[1], *Jinyuan Chen*[2] *and Gengsheng Qin*[3]. [1]Georgia State Univerisity [2]Lanzhou Univeristy [3]Georgia State University

10:55 AM Jackknife empirical likelihood confidence regions for ROC curve with non-ignorable verification bias
♦*Haiqi Wang*[1], *Gengsheng Qin*[1] *and Jinyuan Chen*[2]. [1]Georgia State University [2]Lanzhou University

11:20 AM Jackknife empirical likelihood inference for the pairs of sensitivity,specificity and cut-point
♦*Jinyuan Chen*[1], *Haiqi Wang*[2] *and Gengsheng Qin*[2]. [1]Lanzhou University [2]Georgia State University

11:45 AM Tuning parameter selection in regularized estimations of large covariance matrices
Yixin Fang[1], ♦*Binhuan Wang*[1] *and Yang Feng*[2]. [1]NYU School of Medicine [2]Columbia University

12:10 PM Floor Discussion.

Session 126: Integrating Modeling and Simulation (M&S) in the Drug Development (*Invited*)
Room: KENNESAW, ATLANTA CONFERENCE CENTER - LL3
Organizer: Zhaoling Meng, Sanofi.
Chair: Yuke Wang, Emory University.

10:30 AM Integrating pharmacometrics and statistics to accelerate early clinical development: a case study
♦*Bret Musser*[1], *James Bolognese*[2], *Ghassan Fayad*[1], *Yue Shentu*[1], *Lori Mixson*[1], *Nitin Patel*[2] *and Jaydeep Bhattacharyya*[2]. [1]Merck & Co., Inc [2]Cytel

10:55 AM Applications of Bayesian Modeling and Simulation Methodology in the Pediatric Drug Development
♦*Chyi-Hung Hsu and Steven Xu*. Janssen Research & Development

11:20 AM Application of population-based modeling and simulation for dose justification in clinical developme
♦*Emmanuel Chigutsa and Johan Wallin*. Global PKPD, Eli Lilly and Company

11:45 AM PK/PD modeling of recurrent events and CTS in optimizing Phase 3 dose selection
♦*Zhaoling Meng, Tao Sheng, Lei Ma, Qiang Lu, Dimple Patel and Hui Quan*. sanofi

12:10 PM Floor Discussion.

Session 127: Recent Developments in Statistical Learning of Complex Data (*Invited*)
Room: VININGS, ATLANTA CONFERENCE CENTER - LL3
Organizer: Ganggang Xu, Binghamton University, State University of New York.
Chair: Qifan Song, Purdue University.

10:30 AM Noncrossing Ordinal Classification of Complex Data
Xingye Qiao. Binghamton University

10:55 AM A Conditional Dependence Measure with Applications in Undirected Graphical Models
Jianqing Fan[1], *Yang Feng*[2] *and* ♦*Lucy Xia*[3]. [1]Princeton University [2]Columbia University [3]Stanford University

11:20 AM Provable Sparse Tensor Decomposition and Its Application to Personalized Recommendation
♦*Wei Sun*[1], *Junwei Lu*[2], *Han Liu*[2] *and Guang Cheng*[3]. [1]Yahoo [2]Princeton [3]Purdue

11:45 AM Joint Estimation of Multiple Undirected Graphs with Covariates
♦*Peng Wang*[1] *and Xiaotong Shen*[2]. [1]University of Cincinnati [2]University of Minnesota

12:10 PM Floor Discussion.

Session 128: Nonclinical Statistical Applications in Pharmaceutical Industry (*Invited*)
Room: LENOX, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yanbing Zheng, AbbVie.
Chair: Yuanyuan Duan, AbbVie.

10:30 AM Statistical Experiences on Qualification of a Screening Assay
Jorge Quiroz. Merck Research Laboratories

10:55 AM Frequentist and Bayesian Simulations Using Random Coefficients Model to Set Shelf-Life Specification
♦*Richard Montes*[1] *and David LeBlond*[2]. [1]Hospira, a Pfizer company [2]CMC Statistics

11:20 AM A Bayesian Approach to Analytical Biosimilarity Assessment
♦*Yanbing Zheng and Lanju Zhang*. AbbVie Inc.

11:45 AM Bioequivalence evaluation of sparse sampling data using bootstrap resampling method
Meiyu Shen. Food and Drug Administration

12:10 PM Floor Discussion.

Session 129: Recent Advances in Analysis of Interval-Censored Failure Time data and Longitudinal Data (*Invited*)
Room: SPRING, ATLANTA CONFERENCE CENTER - LL3
Organizer: Yang Li, University of North Carolina-Charlotte.
Chair: Xiaoyi Min, Georgia State University.

10:30 AM Maximum Likelihood Estimation for the Proportional Odds Model with Partly Interval-Censored Data
♦*Liang Zhu*[1], *Dingjiao Cai*[2], *Yimei Li*[3], *Xingwei Tong*[2], *Jianguo Sun*[4], *Deo Kumar Srivastava*[3] *and Melissa M. Hudson*[3]. [1]University of Texas Health Science Center [2]Beijing Normal University [3]St. Jude Children's Research Hospital [4]University of Missouri-Columbia

10:55 AM Case-cohort studies with interval-censored failure time data
♦*Qingning Zhou, Haibo Zhou and Jianwen Cai*. University of North Carolina at Chapel Hill

11:20 AM Bayesian Nonparametric Inference for Panel Count Data with Dependent Observation Times
♦*Ye Liang*[1] *and Yang Li*[2]. [1]Oklahoma State University [2]University of North Carolina - Charlotte

11:45 AM Semiparametric varying-coefficient regression analysis of recurrent events
♦*Yang Li*[1], *Yanqing Sun*[1] *and Li Qi*. [1]University of North Carolina at Charlotte

12:10 PM Floor Discussion.

Session 130: Survival Analysis and its Applications (*Invited*)
Room: EDGEWOOD, ATLANTA CONFERENCE CENTER - LL3
Organizer: Haihong Li, Biogen.
Chair: Haihong Li, Biogen.

10:30 AM Estimating the optimal treatment regime based on restricted mean lifetime
♦*Min Zhang*[1] *and Baqun Zhang* [2]. [1]University of Michigan [2]Renmin University

10:55 AM Parameter estimation in survival models with missing failure indicators in EMR by incorporating Expe
♦*Li Li and Tianle Hu*. Eli Lilly and Company

11:20 AM Survival trees for left-truncated / right-censored data, with application to time-varying covariates
Wei Fu and ♦*Jeffrey Simonoff*. New York University

11:45 AM Computerized Multistage testing
Duanli Yan. Educational Testing Service

12:10 PM Floor Discussion.

Session 131: Recent Developments in Nonparametric Statistics and their Applications (*Invited*)
Room: UNIVERSITY, ATLANTA CONFERENCE CENTER - LL3
Organizer: Xiangrong Yin, University of Kentucky.
Chair: Pengsheng Ji, University of Georgia.

10:30 AM Tensor sufficient dimension reduction
Shanshan Ding. University of Delaware

10:55 AM Sufficient Dimension Reduction via Distance Covariance
Wenhui Sheng. University of West Georgia

11:20 AM A subsampled double bootstrap for massive data
Srijan Sengupta[1], *Stanislav Volgushev*[2] *and* ♦*Xiaofeng Shao*[1]. [1]University of Illinois at Urbana-Champaign [2]Cornell University

11:45 AM A new class of measures for independence test with its application in big data
♦*Xiangrong Yin and Qingcong Yuan*. University of Kentucky

12:10 PM Floor Discussion.

Session 132: Statistical Methods for Medical Research (*Invited*)
Room: ROSWELL, ATLANTA CONFERENCE CENTER - LL3
Organizer: Chang?Yu, Vanderbilt University.
Chair: Mei-Jie Zhang, Medical College of Wisconsin.

10:30 AM Variable screening for classification with errors in the class labels
Guanhua Chen. Vanderbilt University

10:55 AM Robust Tests for Genetic Association Analysis Incorporating Genotyping Uncertainty
♦*Juan Ding, Wenjun Xiong, Junjian Zhang and You Su*. Guangxi Normal University

11:20 AM Performance Evaluation of Propensity Score Methods for Multi-level Treatments
♦*Hui Nian*[1]*, Juan Ding*[2]*, Chang Yu*[1]*, William Wu*[1]*, Richard Shelton*[2]*, William DUpont*[1] *and Pingsheng Wu*[2]. [1]Department of Biostatistics, Vanderbilt University [2]Department of Medicine, Vanderbilt University

11:45 AM Batch Effects on Design and Analysis of Equivalence and Non-inferiority Studies
♦*Jason Liao*[1] *and Ziji Yu*[2]. [1]Merck [2]University of Rochester

12:10 PM Floor Discussion.

Session 133: Statistical Analysis of Complex Data II (*Contributed*)
Room: MARIETTA, ATLANTA CONFERENCE CENTER - LL3
Chair: Hongbin Zhang, City University of New York.

10:30 AM A Mixed Effects Model for analyzing Complex AIDS Clinical Data
Tao Wang. School of Mathematics Yunnan Normal University CN

10:45 AM Simulated Data for SNP Set Association Tests in Family Samples
♦*Hung-Chih Ku*[1] *and Chao Xing*[2]. [1]DePaul University [2]University of Texas Southwestern Medical Center

11:00 AM A Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations
♦*Chen Chen, Min Zhang and Dabao Zhang*. Purdue University

11:15 AM Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams
♦*Kun Liu, Ruizhi Zhang and Yajun Mei*. Georgia Institute of Technology

11:30 AM Phase-amplitude functional framework for analyzing RNA sequencing data with point process filtering
♦*Sergiusz Wesolowski*[1]*, Daniel Vera*[2] *and Wei Wu*[3]. [1]Prefer not to mention [2]Center of Genomics, Florida State Univ [3]Department of Statistics, Florida State Univ

11:45 AM Assessment of drug combination effects using mixed-effects model
Shouhao Zhou. UT - MD Anderson Cancer Center

12:00 PM Floor Discussion.

Session 134: Statistical Genetics (*Contributed*)
Room: HANOVER C, EXHIBIT LEVEL - LL2
Chair: Jie Chen, Augusta University.

10:30 AM Kernel-based Nonparametric Testing in High-dimensional Data with Applications to Gene Set Analysis
♦*Tao He*[1]*, Ping-Shou Zhong*[2]*, Yuehua Cui*[2] *and Vidyadhar Mandrekar*[2]. [1]San Francisco State University [2]Michigan State University

10:45 AM Optimal Filtering to Increase Detections of Differentially Expressed Genes in Microarray Data
♦*Zixin Nie and Kun Liang*. University of Waterloo

11:00 AM A unified X-chromosome genetic association test accounting for different XCI processes
♦*Jian Wang, Robert Yu and Sanjay Shete*. UT MD Anderson Cancer Center

11:15 AM Structured Sparse Co-Inertia Analysis with Applications to Genomics and Metabolomics Data
♦*Eun Jeong Min, Sandra Safo and Qi Long*. Emory University

11:30 AM Testing for gene-gene interaction in case-control GWAS
Zhongxue Chen. Indiana University Bloomington

11:45 AM A Set-Valued System Model for Secondary Trait Genetic Association Analysis in Case-Control Studies
Wenjian Bi. St. Jude Children's Research Hospital

12:00 PM Floor Discussion.

Session 135: Topics in Statistics II (*Contributed*)
Room: HANOVER D, EXHIBIT LEVEL - LL2
Chair: Guoqing Diao, George Mason University.

10:30 AM Semiparametric Inference via Sparsity-Induced Kriging for Massive Spatial Datasets
♦*Pulong Ma and Emily Kang*. University of Cincinnati

10:45 PM Analyzing of mutation of TNBC cell cytoplasmic level using Bayesian partition methods
♦*Guanhao Wei*[1]*, Jing Zhang*[1]*, Remus Osan*[1] *and Ritu Aneja*[2]. [1]Dept. of Math and Stat,Georgia State University [2]Dept. of Biology,Georgia State University

11:00 AM BAYESIAN METHOD FOR CAUSAL INFERENCE IN MULTIVARIATE TIME SERIES WITH APPLICATION ON SALES DATA
BO NING. North Carolina State University

11:15 AM PCAN: Probabilistic Correlation Analysis of Two Non-normal Data Sets
Roger Zoh. Texas A&M University School of Public Health

11:30 AM Intensity Estimation in Poisson Process with Compositional Noise
♦*Glenna Gordon, Wei Wu and Anuj Srivastava*. Florida State University, Department of Statistics

11:45 AM The application of association rules mining in vehicle crash study for Mississippi coastal areas
♦*Zhao Ma, Feng Wang and Ningning Wang*. Jackson State University

12:00 PM Floor Discussion.

Abstracts

Session 1: Biostatistics in Medical Applications

Patient-Centered Pragmatic Clinical Trials: What, Why, When, How?

Sally Morton
University of Pittsburgh
scmorton@pitt.edu

What are patient-centered pragmatic clinical trials, and why is this study design becoming increasingly commonplace? Primarily, healthcare reform has focused attention on patient-centered comparative effectiveness research that seeks to answer "What healthcare treatment works best, for whom, and under what circumstances?" from the unique perspectives of patients, families, caregivers, clinicians, and policy-makers. Consequently, the gap between available and necessary evidence for clinical decisions, including the cost of producing such evidence, demands new analytic approaches. This talk will address pragmatic clinical trial design, data, and analysis opportunities for statisticians. Topics will include cluster randomization and data networks, using Patient-Centered Outcomes Research Institute (PCORI) studies as examples. Challenges for the statistics discipline beyond these methodological issues, for example implications for collaboration and training, will also be raised.

Robust Methods for Treatment Effect Calibration, with Application to Non-Inferiority Trials

♦*Zhiwei Zhang*[1]*, Lei Nie*[1]*, Guoxing Soon*[1] *and Zonghui Hu*[2]
[1]FDA
[2]NIH
zhiwei.zhang@fda.hhs.gov

In comparative effectiveness research, it is often of interest to calibrate treatment effect estimates from a clinical trial to a target population that differs from the study population. One important application is an indirect comparison of a new treatment with placebo on the basis of two separate clinical trials: a non-inferiority trial comparing the new treatment with an active control and a historical trial comparing the active control with placebo. The available methods for treatment effect calibration include an outcome regression (OR) method based on a regression model for the outcome and a weighting method based on a propensity score (PS) model. This article proposes new methods for treatment effect calibration: one based on a conditional effect (CE) model and two doubly robust (DR) methods. The first DR method involves a PS model and an OR model, is asymptotically valid if either model is correct, and attains the semiparametric information bound if both models are correct. The second DR method involves a PS model, a CE model and possibly an OR model, is asymptotically valid under the union of the PS and CE models, and attains the semiparametric information bound if all three models are correct. The various methods are compared in a simulation study and applied to recent clinical trials for treating human immunodeficiency virus infection.

A Comparison Study of Fixed and Mixed Models for Gene Level Association Studies of Complex Traits

♦*Ruzong Fan*[1]*, Jeesun Jung* [2]*, Chi-yang Chiu*[1]*, Daniel E. Weeks*[3]*, Alexander F. Wilson*[4]*, Joan E. Bailey-Wilson*[4] *and Christopher I. Amos*[5]
[1]NICHD, NIH
[2]National Institute on Alcohol Abuse and Alcoholism
[3]University of Pittsburgh
[4]NHGRI, NIH
[5] Geisel School of Medicine at Dartmouth
fanr@mail.nih.gov

In association studies of complex traits, fixed effect regression models are usually used to test for association between traits and major gene loci. In recent years, variance-component tests based on mixed models were developed for region-based genetic variant association tests. In the mixed models, the association is tested by a null hypothesis of zero variance via a sequence kernel association test (SKAT), its optimal unified test (SKAT-O), and a combined sum test of rare and common variant effect (SKAT-C). Although there are some comparison studies to evaluate the performance of mixed and fixed models, there is no systematic analysis to determine when the mixed models perform better and when the fixed models perform better. Here we evaluated, based on extensive simulations, the performance of the fixed and mixed model statistics, using genetic variants located in 3, 6, 9, 12, and 15 kb simulated regions. We compared the performance of three models: (i) mixed models which lead to SKAT, SKAT-O, and SKAT-C, (ii) traditional fixed effect additive models, and (iii) fixed effect functional regression models. To evaluate the type I error rates of the tests of fixed models, we generated genotype data by two methods: (i) using all variants; (ii) using only rare variants. We found that the fixed effect tests accurately control or have low false rates. We performed simulation analyses to compare power for two scenarios: (i) all causal variants are rare; (ii) some causal variants are rare and some are common. Either one or both of the fixed effect models performed better than or similar to the mixed models except when (1) the region sizes are 12 and 15 kb and (2) effect sizes are small. Therefore, the assumption of mixed models could be satisfied and SKAT/SKAT-O/SKAT-C could perform well if the number of causal variants is large and each causal variant contributes a small amount to the traits (i.e., polygenes). In major gene association studies, we argue that the fixed effect models perform better or similarly to mixed models in most cases since some variants should affect the traits relatively large. In practice, it makes sense to perform analysis by both the fixed and mixed effect models and to make a comparison, and this can be readily done using our R codes and the SKAT packages.

Methods for biomarker combination in presence of missing gold standard

♦*Danping Liu*[1]*, Ashok Chaurasia*[2] *and Zheyu Wang*[3]
[1]National Institutes of Health
[2]NIH and University of Waterloo
[3]Johns Hopkins University
danping.liu@nih.gov

In biomarker evaluation, the true disease condition may be missing because it is expensive or harmful to ascertain. It is well-known that the complete-case analysis often leads to biased estimation of the diagnostic accuracy, known as "verification bias". When multiple biomarkers exist for diagnosing or predicting the disease condition, an important research question is how to combine these markers to improve the diagnostic accuracy. Existing methods, such as logistic regression or maximization of the area under ROC curv

Session 2: Geometric Approaches in Functional Data Analysis

Bayesian uncertainty quantification in partial differential equation models

Prithwish Bhaumik
The University of Texas at Austin
prithwish.bhaumik@utexas.edu

Often the probability density of a generalized linear model contains a function indexed by the parameter vector θ. The function is unknown. But it is known to satisfy a system of partial differential equations (PDE's) which typically do not have an analytical solution. Our objective is to infer on θ. Bhaumik and Ghosal (2015) and Bhaumik and Ghosal (2016) considered this problem in the context of ordinary differential equations and additive model. In this paper we first approximate the function using a random series based on B-spline basis functions. A prior is put on the coefficients of the basis functions. A posterior on θ is induced from those of the coefficients using two approaches namely two-step approach and numerical solution based approach. In two-step approach the posterior on the parameter is induced by minimizing an integrated weighted squared distance between the derivative of the spline representation and the derivative suggested by the PDEs. Although this approach is computationally fast, the Bayes estimator is not asymptotically efficient. This drawback is remedied in the second approach where we maximize the expected log-likelihood constructed using the numerical solution of the PDE. Under both approaches we establish a Bernstein-von Mises theorem which assures that Bayesian uncertainty quantification matches with the frequentist one.

Statistical Analysis of Trajectories on Riemannian Manifolds

♦*Jingyong Su*[1] *and Anuj Srivastava*[2]
[1]Texas Tech University
[2]Florida State University
jingyong.su@ttu.edu

In this research we propose to develop a comprehensive framework for registration and analysis of manifold-valued processes. Functional data analysis in Euclidean spaces has been explored extensively in literature. But we study a different problem in the sense that functions to be studied take values on nonlinear manifolds, rather than in vector spaces. Manifold-valued data appear frequently in shape and image analysis, computer vision, biomechanics and many others. If the data were contained in Euclidean space, one would use standard Euclidean techniques and there has been a vast literature on these topics. However, the non-linearity of the manifolds requires development of new methodologies suitable for analysis of manifold-valued data. We propose a comprehensive framework for joint registration and analysis of multiple manifold-valued processes. The goals are to take temporal variability into account, derive a rate-invariant metric and generate statistical summaries (sample mean, covariance etc.), which can be further used for registering and modeling multiple trajectories.

Fast Functional Genome Wide Association Analysis of Surface-based Imaging Genetic Data

♦*Chao Huang and Hongtu Zhu*
Dept. Biostatistics/ UNC at Chapel Hill
chaohuang.stat@gmail.com

More and more large-scale imaging genetic studies are being widely conducted to collect a rich set of imaging, genetic, and clinical data to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. Several major big-data challenges arise from testing millions of genome-wide associations with functional signals sampled at millions of locations in the brain from thousands of subjects. In this talk, we are presenting a Fast Functional Genome Wide Association Study (FFGWAS) framework to carry out whole-genome analyses of multimodal imaging data. FFGWAS consists of three components including (1) a multivariate varying coefficient model for modeling the relation between multiple functional imaging responses and a set of covariates (both genetic and non-genetic predictors), (2) a global sure independence screening (GSIS) procedure for reducing the dimension from a very large scale to a moderate scale, and (3) a detection procedure for detect significant cluster-locus pairs. We also successfully applied FFGWAS to a large-scale imaging genetic data analysis of ADNI data with 708 subjects, 30,000 vertices on hippocampal surface, and 501,584 SNPs.

Shape Analysis of Trees Using Elastic Curve Geometry and Side Branch Permutation

♦*Adam Duncan*[1]*, Eric Klassen*[2] *and Anuj Srivastava*[1]
[1]Florida State University, Dept. of Statistics
[2]Florida State University, Dept. of Mathematics
a.duncan@stat.fsu.edu

Many objects in nature consist of branching structures such as blood vessel networks, lung airway paths, and neuron structures. These tree-shaped objects vary in both topology and geometry which makes them difficult to analyze with existing statistical techniques. A recent kind of method incorporates both geometry and topology by defining equivalence relations under certain permutations of branches. In the quotient space of such relations, geometric attributes of branches can be directly compared, giving a metric on the shape space of trees. For general tree topologies, this approach runs into prohibitive computational cost due to the combinatorial explosion of possible permutations. We propose a method which avoids this problem by restricting the class of allowed topologies in the following way. We consider trees which consist of: (1) a main branch viewed as a parameterized curve in $\mathbb{R}^3$, and (2) some number of secondary branches– also parameterized curves in $\mathbb{R}^3$ – which emanate from the main branch at arbitrary points. The combinatorial problem of matching side branches is reduced to a linear assignment problem which is efficiently solved by well-known algorithms. We then define a metric between trees with an elastic Riemannian metric between individual pairs of matched branches. The method is illustrated on a set of axonal tree structures taken from confocal microscope images of Drosophila Melanogaster neurons. We show examples of geodesics paths, Karcher means, and modes of variability, as well as cross-validated classification between mutated and wild type groups.

Session 3: Advance in Statistical Genomics and Computational Biology

A Novel and Efficient Study Design to Test Parent-of-Origin Effects in Family Trios

♦*Rui Feng and Xiaobo Yu*
University of Pennsylvania
ruifeng@upenn.edu

Increasing evidence has shown that both genes and their parent-of-origins can play a major role in the risk of various prenatal and neonatal diseases. Statistical models that accommodate parent-of-origin effects can not only improve the power to detect gene-disease association, but also help recover heritability undetected in traditional genome-wise association studies. In many studies, children's

DNA samples were collected for initial testing and their parents were willing to contribute their own DNAs for further investigations. Nowadays, next-generation sequencing (NGS) is often done on children's DNA samples to search for common and novel genetic variants. However, the cost of NGS of the whole family is generally not affordable to all, making the assessment of parental-of-origin effects impossible. Motivated by the reality, we propose a new study design collecting data from children's NGS in combination with the parental genotypes obtained from the traditional genotyping array, at a fractional cost of sequencing all family members. We develop a powerful and efficient likelihood-based method, which allows accurate phasing of children's alleles and thus enables estimating parent-of-origin effect of children's alleles on the disease phenotype. Our method also incorporates missing genotypes or missing parent that often happen in genetic studies. We use simulations to evaluate the performance of our method and compare it with existing genotype-based methods to detect parent-of-origin effects. With even a regular read length of 100 base pairs, our method shows advantage over that using SNP data and achieves power closer to that of the optimal test.

A Mathematical Population Genetics Model for Cancer Gene Detection

Amei Amei
University of Nevada, Las Vegas
amei.amei@unlv.edu

Mutation frequencies can be modeled as a Poisson random field (PRF) to estimate speciation times and the degree of selection on newly arisen mutations. This approach provides a quantitative theory for comparing intraspecific polymorphism with interspecific divergence in the presence of selection and can be used to estimate population genetic parameters. Due to a built-in structure of the species divergence time, a recently developed time-dependent PRF model is especially suitable for estimating selective effects of more recent mutations such as the mutations that have occurred in the human genome. By analyzing the estimated distribution of the selection coefficients at each individual gene, we are able to detect a gene or a group of genes that might have related to the diagnosed cancer. Moreover, the estimate of the species divergence time will provide useful information regarding the occurrence time of the cancer. The disease analyzed in this study is a type of cytogenetically normal myelodysplatic syndrome (CN-MDS) which affects the bone marrow and blood. The DNA was extracted from bone marrow from two patients using a kit purchased from Qiagen (DNeasy Blood and Tissue kit). The time-dependent PRF model was first applied to 621 genes in chromosome one and estimated the mean selection coefficient of a newly arisen non-synonymous mutation that is observed as polymorphism and the two types of mutation rates for each gene. Based on our results, there are three groups with 33 genes whose mean selection coefficients are large negative values and hence can be treated as a pool of cancer related genes from Chromosome one. To get a complete set of candidate genes, we will apply the model to aligned DNA sequences from the rest of the 22 chromosomes including chromosome X.

Machine learning application in gene regulation and its high performance computing on GPU platform

♦*Zhong Wang*[1]*, Lauren Choate*[2]*, Tinyi Chu*[3] *and Charles Danko*[1]
[1]College of Veterinary Medicine, Cornell University
[2]Molecular Biology and Genetics, Cornell University
[3]Computational Biology, Cornell University
wzhy2000@hotmail.com

Machine learning methods have been applied to a broad range of areas within genetics and genomics, especially for the analysis of genome sequencing data sets. We present how to combine the statistical model and machine learning methods to identify the transcriptional regulatory elements from the large scale and complex sequence data set. Also R packages developed by our lab are introduced to assist in the practice of data analysis of chip-seq, RNA-seq or GRO-seq. Besides the packages of computational biology, a general R package for SVM on GPU platform is implemented and released for the big data.

Impact of genotyping errors on statistical power of association tests in genomic analyses

Lin Hou
Tsinghua University
houl@tsinghua.edu.cn

A key step in genomic studies is to assess high throughput measurements across millions of markers for each participant's DNA, either using microarrays or sequencing techniques. Accurate genotype calling is essential for downstream statistical analysis of genotype-phenotype associations, and next generation sequencing (NGS) has recently become a more common approach in genomic studies. How the accuracy of variant calling in NGS-based studies affects downstream association analysis has not, however, been studied using empirical data in which both microarrays and NGS were available. In this article, we investigate the impact of variant calling errors on the statistical power to identify associations between single nucleotides and disease, and on associations between multiple rare variants and disease.

Session 4: New Developments in Biomedical Research and Statistical Genetics

Treatment Effect Estimate and Model Diagnostics with Two-way Time-Varying Treatment Switching

♦*Qingxia Chen*[1]*, Fan Zhang*[2]*, Ming-Hui Chen*[2] *and Xiuyu Cong*[3]
[1]Vanderbilt University
[2]University of Connecticut
[3]Boehringer Ingelheim Pharmaceuticals
cindy.chen@vanderbilt.edu

Treatment switching frequently occurs in clinical trials due to ethical reasons. Intent-to-treat analysis without adjusting for switching yields biased and inefficient estimates of the treatment effects. In this paper, we propose a class of semiparametric semi-competing risks transition survival models to accommodate two-way time-varying switching to post-study therapy. Theoretical properties of the proposed model are examined. An efficient expectation-maximization algorithm is derived and implemented in existing software to obtain the maximum likelihood estimates along with model diagnostic tools. Simulation studies are conducted to demonstrate the validity of the model. The proposed method is further applied to data from a clinical trial with patients having recurrent or metastatic squamous-cell carcinoma of the head and neck.

Detecting regulatory relationships between DNA alterations and gene/protein expressions

♦*Jie Peng*[1]*, Chris Conley* [1] *and Pei Wang*[2]
[1]UC Davis
[2]Icahn School of Medicine at Mount Sinai
jiepeng108@gmail.com

Motivated by network construction in integrative genomics, we developed a sparse conditional graphical model called Spacemap.

Spacemap integrates two high dimensional data types through a penalized multivariate regression framework in the high-dimension-low-sample-size regime. We apply spacemap to model regulatory relationships between DNA alterations and gene/protein expressions using a breast cancer data set. The performance of spacemap is evaluated by simulation studies. Strategies for model fitting and model selection will also be discussed.

Pre-conditioning method for risk prediction with application to ED acute heart failure patients

♦*Dandan Liu, Cathy Jenkins, Sean Collins, Alan Storrow and Frank Harrell*

Vanderbilt University Medical Center
dandan.liu@vanderbilt.edu

Accurate risk prediction is very important in acute care settings as efficient resource allocation becomes critical. Conventional methods for developing prediction models include pre-specification, principal component analysis, and step-wise regression. When the scope of predictors greatly exceeds the sample size these methods do not always perform well. The 'pre-conditioning' method offers a potential alternative to classic methods for high dimensional settings. It separates risk prediction and model selection, utilizing the correlation structure of the predictors. Leveraging on a multi-site study on acute heart failure patients admitted to the emergency medicine, the pre-conditioning method was compared to other conventional methods in predicting 30-day adverse event. Assessment on model performance was based on Somer's Dxy, area under the curve (AUC) of Receiver Operating Characteristic (ROC) curves, and calibration curves.

On Nonsmooth Estimating Functions via Jackknife Empirical Likelihood

Jinfeng Xu

University of Hong Kong
xujf@hku.hk

n many applications, the parameters of interest are estimated by solving non-smooth estimating functions with U-statistic structure. Because the asymptotic covariances matrix of the estimator generally involves the underlying density function, resampling methods are often used to bypass the difficulty of non-parametric density estimation. Despite its simplicity, the resultant-covariance matrix estimator depends on the nature of resampling, and the method can be time-consuming when the number of replications is large. Furthermore, the inferences are based on the normal approximation that may not be accurate for practical sample sizes. In this paper, we propose a jackknife empirical likelihood-based inferential procedure for non-smooth estimating functions. Standard chi-square distributions are used to calculate the p-value and to construct confidence intervals. Extensive simulation studies and two real examples are provided to illustrate its practical utilities.

Session 5: Change-Point Problems and their Applications (III)

Sequential surveillance of structural breaks in firms' credit rating migrations

♦*Haipeng Xing and Ke Wang*

Stony Brook University
xing@ams.sunysb.edu

Recent studies have shown that firms' credit rating migration process is not stationary and may have structural breaks. Assuming the generator of probability transition matrices of firms' credit rating to be piecewise constant and the jump time of generator corresponds to the structural break time in the pattern of firms' rating migrations, we study several types of sequential surveillance rules for early detection. The surveillance rules we investigated includes the Shewhart control chart, an generalized likelihood ratio (GLR) detection rule for a single change-point with unknown pre- and post-change transition matrices, a detection rule based on an extension of Shiryaev's Bayes single change-point model, and a detection rule for multiple unknown structural breaks. We provide theoretical discussion and extensive simulations to compare the performance of these rules. We further use these rules to online detect structural breaks in firms' credit rating migrations based on U.S. firms' rating record from 1986 to 2012.

Detection of Changes monitored at Random Time Points

Marlo Brown

Niagara University
mbrown@niagara.edu

We look at a Poisson process where the arrival rate changes at some unknown time point. We monitor this process only at certain time points. At each time point, we count the number of arrivals that happened in that time interval. In previous work, it was assumed that the time intervals were fixed in advanced. We relax this assumption to assume that the time intervals that the process is monitored is also random. For a loss function consisting of the cost of late detection and a penalty for early stopping, we develop, using dynamic programming, the one and two steps look ahead Bayesian stopping rules. We then compare various observation schemes to determine the best model. We provide some numerical results to illustrate the effectiveness of the detection procedures.

1 D-ary Sequential Tests of Circular Error Probability

♦*Yan Li*[1] *and Yajun Mei*[2]

[1]East China Normal University

[2]Georgia Institute of Technology
yli@stat.ecnu.edu.cn

In the context of evaluating a system's precision quality, one wants to utilize observed two-dimensional data to quickly test the hypotheses on the circular error probability (CEP) or the probability of non-conforming, which refer to the chance of the system hitting or missing a pre-specified disk target, respectively. We propose to develop efficient sequential tests based on the D-ary quantization of the observed two-dimensional data. We not only analyze the asymptotic performance limits of non-stationary D-ary quantization in the CEP context, but also investigate the optimal stationary D-ary quantizers from both theoretical and practical viewpoints. Our results suggest that the sequential probability ratio test (SPRT) based on the stationary triplet quantization is often efficient and makes operational sense in practice.

Sequential event detection in networked Hawkes process

Shuang Li, ♦*Yao Xie, Mehrdad Farajtabar and Le Song*

Georgia Institute of Technology
su06ee@gmail.com

We present a sequential change-point detection procedure for a networked Hawkes process, which is commonly used to model social network activities with self- and mutual excitations. The Hawkes process consists of a sequence of discrete events data that occur at non-uniform times on nodes and edges of a network. Change-point in the influence matrix of the networked Hawkes process usually represents an excitation caused by an event. Our method is based on the generalized likelihood ratio statistic for point processes, and

it achieves weak signal detection by aggregating local statistics over time and networks. The computation of the detection statistic only requires aggregation over local neighborhood in the network, and hence it is scalable to the size of the network. Moreover, the detection statistic can be updated using an EM-like algorithm to include new samples and hence it can efficiently handle streaming data. We derive highly accurate theoretical characterization of the false-alarm-rate, using the change-of-measure technique and the mean field characterization for the multi-dimensional Hawkes process. We demonstrate the good performance of our algorithm via numerical examples and real-world twitter and memetracker datasets.

Session 6: Statistical and Computational Analysis of High-Throughput RNA Sequencing Data

On the correlation analysis of RNA-seq data

Yinglei Lai
The George Washington University
ylai@gwu.edu

Genome-wide RNA sequencing (RNA-seq) data have been increasingly collected during the recent years. A correlation analysis of RNA-seq data can be important for us to understand how genes interact with each other. RNA-seq data are generally count-type observations. Furthermore, many genes have multiple isoforms. Therefore, it can be challenging to conduct a correlation analysis of RNA-seq data. We propose a multivariate approach for the correlation analysis of RNA-seq data. Our simulation study demonstrates the advantage of our method. We use the RNA-seq data collected by The Cancer Genome Atlas (TCGA) project to illustrate our method.

MSIQ: JOINT MODELING OF MULTIPLE RNA-SEQ SAMPLES FOR ACCURATE ISOFORM QUANTIFICATION

Wei Vivian Li[1]*, Anqi Zhao*[2]*, Shihua Zhang*[3] *and* ♦*Jingyi Jessica Li*[1]
[1]Department of Statistics, UCLA
[2]Department of Statistics, Harvard University
[3]Chinese Academy of Sciences
jli@stat.ucla.edu

Next-generation RNA sequencing (RNA-seq) technology has been widely used to assess full-length RNA isoform abundance in a high-throughput manner. RNA-seq data offer insight into gene expression levels and transcriptome structure, enabling us to better understand the regulation of gene expression and fundamental biological processes. Accurate quantification of RNA isoforms from RNA-seq data is a challenging computational task due to the information loss in sequencing experiments. Recent accumulation of multiple RNA-seq data sets from the same biological condition provides new opportunities to improve the isoform quantification accuracy. However, existing statistical or computational methods for multiple RNA-seq samples either pool the samples into one sample or assign equal weights to the samples in estimating isoform abundance. These methods ignore the possible heterogeneity in the quality and noise levels of different RNA-seq samples, and could have biased and unrobust estimates. In this article, we develop a method named "joint modeling of multiple RNA-seq samples for accurate isoform quantification"(MSIQ) for more accurate and robust isoform quantification, by integrating multiple RNA-seq samples under a Bayesian framework. Our method aims to (1) identify the informative group of samples with homogeneous quality and (2) improve isoform quantification accuracy by jointly modeling multiple RNA-seq samples with more weights on the informative group. We show that MSIQ provides a consistent estimator of isoform abundance, and demonstrate the accuracy and effectiveness of MSIQ compared to alternative methods through simulation studies on D. melanogaster genes. We justify MSIQ's advantages over existing approaches via application studies on real RNA-seq data of human embryonic stem cells and brain tissues. We also perform a comprehensive analysis on how the isoform quantification accuracy would be affected by RNA-seq sample heterogeneity and different experimental protocols.

Statistical Models for Single Cell RNA Sequencing

Cheng Jia, Yuchao Jiang, Mingyao Li and ♦*Nancy Zhang*
University of Pennsylvania
nzh@wharton.upenn.edu

Traditional gene expression measurements were made on bulk tissue samples containing populations of cells. Recent laboratory advances have made possible the measurement of RNA levels in single cells. This new frontier offers exciting challenges and opportunities. I will describe some of these challenges and propose a new error model for single cell RNAseq that explicitly addresses the technical issues of dropout, amplification artifact, and cell size confounding. I will show that this model can be used in differential expression analysis. I will also show how addressing these technical biases allows us to better characterize the stochasticity of gene transcription. This is joint work with Cheng Jia, Yuchao Jiang, and Mingyao Li.

Robust statistical analysis of RNA-seq data

Maoqi Xu and ♦*Liang Chen*
University of Southern California
liang.chen@usc.edu

The individual sample heterogeneity is one of the biggest obstacles in biomarker identification for complex disease such as cancers. Current statistical models to identify differentially expressed genes between disease and control groups often overlook the substantial human sample heterogeneity. Meanwhile traditional nonparametric tests lose detailed data information and sacrifice the analysis power, although they are distribution free and robust to heterogeneity. Here, we propose a weighted empirical likelihood-ratio test (WELT) with a mean-variance relationship constraint for the differential expression analysis of RNA-seq. As a distribution-free nonparametric model, WELT handles individual heterogeneity by estimating an empirical probability for each observation without making any assumption about read-count distribution. It also incorporates a constraint for the read-count overdispersion which is widely observed in RNA-seq data. WELT demonstrates a significant improvement over existing methods such as edgeR, DESeq, T-tests, Wilcoxon tests, and the classic empirical likelihood-ratio test (ELT) when handling heterogeneous groups. It will significantly advance the transcriptomics studies of cancers and other complex disease.

Session 7: Emerging Statistical Methods for Longitudinal Data

Functional Multiple Indicators, Multiple Causes Measurement Error Models

♦*Carmen Tekwe, Roger Zoh, Fuller Bazer and Raymond Carroll*
Texas A&M University
cdtekwe@sph.tamhsc.edu

Energy expenditure is often used in studying obesity to approximate the amount of energy expended by the body to perform routine bodily functions. It is not directly observable, however, it is a latent construct with multiple indicators such as respiratory quotient, volumetric oxygen consumption and volumetric carbon diox-

ide production. Metabolic rate is used to assess the body's ability to perform metabolic processes. Metabolic rate is often approximated by heat production plus some error. Obesity development involves an imbalance between dietary energy intake and whole body energy expenditure. In studying this pathway, a variation of the multiple indicators, multiple cause measurement error (MIMIC ME) models can be applied. In this paper, we define the functional MIMIC ME model by extending the linear MIMIC ME model to allow longitudinal responses that appear as curves over a given period of time. The mean curves are modeled using basis splines and functional principal components. We also propose a novel approach to identifying the classical measurement error associated with approximating true metabolic rate by heat production based on functional principal components. In addition to defining the FMIMIC ME model, we estimate the model parameters using the EM algorithm. We also provide a discussion of the model's identifiability. The model is applied to study the relationship between metabolic rate and the multiple indicators of energy expenditure among Zucker diabetic fatty rats. Results from a brief simulation study are also provided.

Efficient quantile marginal regression for longitudinal data with dropouts

♦*Hyunkeun Cho*[1]*, Hyokyoung Hong*[2] *and Mi-Ok Kim*[3]
[1]Western Michigan University
[2]Michigan State University
[3]Cincinnati Children's Hospital Medical Center
hyunkeun.cho@wmich.edu

In many biomedical studies independent variables may affect the conditional distribution of the response differently in the middle as opposed to the upper or lower tail. Quantile regression evaluates diverse covariate effects on the conditional distribution of the response with quantile specific regression coefficients. In this talk, we develop an empirical likelihood inference procedure for longitudinal data that accommodates both the within-subject correlations and informative dropouts under missing at random mechanisms. We borrow matrix expansion idea of quadratic inference function and incorporate the within-subject correlations under an informative working correlation structure. The proposed procedure does not assume the exact knowledge of the true correlation structure nor does it estimate the parameters of the correlation structure. Theoretical results show that the resulting estimator is asymptotically normal and more efficient than one attained under a working independence correlation structure. We expand the proposed approach to account for informative dropouts under missing at random mechanisms. The methodology is illustrated by empirical studies and a real life example of HIV data analysis.

Partially Linear Additive Quantile Regression in Ultra-high Dimension

Ben Sherwood[1] *and* ♦*Lan Wang*[2]
[1]Johns Hopkins University
[2]University of Minnesota
wangx346@umn.edu

We consider a flexible semiparametric quantile regression model for analyzing high dimensional heterogeneous data. This model has several appealing features: (1) By considering different conditional quantiles, we may obtain a more complete picture of the conditional distribution of a response variable given high dimensional covariates. (2) The sparsity level is allowed to be different at different quantile levels. (3) The partially linear additive structure accommodates nonlinearity and circumvents the curse of dimensionality. (4) It is naturally robust to heavy-tailed distributions. We approximate the nonlinear components using B-spline basis functions. We first study estimation under this model when the nonzero components are known in advance and the number of covariates in the linear part diverges. We then investigate a nonconvex penalized estimator for simultaneous variable selection and estimation. We derive its oracle property for a general class of nonconvex penalty functions in the presence of ultra-high dimensional covariates under relaxed conditions. To tackle the challenges of nonsmooth loss function, nonconvex penalty function and the presence of nonlinear components, we combine a recently developed convex-differencing method with modern empirical process techniques. Monte Carlo simulations and an application to a microarray study demonstrate the effectiveness of the proposed method. We also discuss how the method for a single quantile of interest can be extended to simultaneous variable selection and estimation at multiple quantiles.

Structural Nonparametric Methods for Estimation, Prediction and Tracking with Longitudinal Data

♦*Colin Wu and Xin Tian*
National Heart, Lung and Blood Institute, NIH
wuc@nhlbi.nih.gov

Longitudinal analysis has three important objectives in biomedical studies: (a) estimating the time-varying population-average and subject-specific covariate effects on the outcome process of interest; (b) predicting the future subject-specific outcome trajectories; (c) evaluating the tracking abilities of important risk factors and health outcomes. Because longitudinal data (which is often referred as functional data) consist repeatedly measured outcome and covariate processes over time, they can be used to accomplish the above three objectives simultaneously. Popular parametric methods for longitudinal analysis, such as the generalized mixed-effects models, are often too restrictive and unrealistic for real applications because of their modeling assumptions. On the other hand, nonparametric models without any structural assumptions could be computationally infeasible and difficult to interpret. We present in this talk some structural nonparametric methods to accomplish the above three objectives, namely estimation, prediction and tracking, based on a class of nonparametric mixed-effects models. Our methods, which use either local kernel-type smoothing or global smoothing via B-splines, have the appropriate model flexibility and computational feasibility, and are useful to answer many scientific questions which could not be properly addressed by parametric or unstructured nonparametric regression models. We demonstrate the application of our methods through a long-term epidemiological study of pediatric cardiovascular risk factors and a series of simulation studies. Asymptotic developments of our methods suggest that the convergence rates of our smoothing estimators depend on the number of subjects as well as the numbers of repeated measurements.

Session 8: Empirical Bayes, Methods and Applications

Empirical Bayes methods and nonparametric mixture models via nonparametric maximum likelihood

♦*Lee Dicker*[1]*, Sihai Zhao*[2] *and Long Feng*[1]
[1]Rutgers
[2]UIUC
ldicker@gmail.com

Empirical Bayes methods have a long and rich history in statistics, and are particularly well-suited for many problems involving heterogeneous and high-dimensional data. Nonparametric maximum likelihood (NPML) is one elegant approach to empirical Bayes that

has been studied since the 1950s and is closely related to the analysis of nonparametric mixture models. However, implementation and analysis of NPML-based methods for empirical Bayes is notoriously difficult. Recent computational breakthroughs have greatly simplified the implementation of NPML-based methods. Leveraging these recent advances, we have developed a variety of promising and flexible new methods involving NPML for empirical Bayes. In this talk we will discuss these methods, along with a variety of applications, including (i) classification of cancer patients using microarray data, (ii) predicting batting averages for Major League Baseball players, and (iii) online estimation of blood glucose density for diabetes. This talk is based on joint work with Sihai Dave Zhao (UIUC) and Long Feng (Rutgers).

Block-Linear Empirical Bayes Estimation of a Heteroscedastic Normal Mean

♦*Asaf Weinstein*[1]*, Zhuang Ma*[2]*, Lawrence Brown*[2] *and Cun-hui Zhang*[3]
[1]Stanford University
[2]University of Pennsylvania
[3]Rutgers University
asafw@stanford.edu

We revisit a classic problem: X_i $N(\theta_i, V_i)$ indep, V_i known, $i = 1, ..., n$, and the goal is to estimate the (nonrandom) means θ_i under sum of squared errors. When the variances are all equal, linear empirical Bayes estimators which model the true means as i.i.d. random variables lead to (essentially) the James-Stein estimator, and have strong frequentist justifications. In the heteroscedastic case such empirical Bayes estimators are less adequate if the V_i and θ_i are empirically dependent. We suggest a new empirical Bayes procedure which groups together observations with similar variances and applies a spherically symmetric estimator to each group separately. Our estimator is exactly minimax and at the same time asymptotically achieves the risk of a stronger oracle than the usual one. The motivation for the new estimator comes from extending a compound decision theory argument from equal variances to unequal variances.

This is joint work with Larry Brown, Zhuang Ma and Cun-Hui Zhang.

Nonparametric empirical Bayes approach to integrative high-dimensional classification

Sihai Zhao
University of Illinois at Urbana-Champaign
sdzhao@illinois.edu

An emerging problem in integrative genomics is the development of high-dimensional classifiers trained on one dataset that can also incorporate information from existing results of previous studies. Such integrative methods should be able to outperform standard classifiers trained only using a single dataset. However, general principles for how separate datasets can be integrated for predictive modeling are lacking. This talk introduces an empirical Bayes framework for integrative classification. Under this framework the optimal integrative classifier is given by the Bayes classifier, which is implemented using nonparametric maximum likelihood methods to estimate unknown prior distributions.

Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective

♦*Roger Koenker*[1] *and Jiaying Gu*[2]
[1]U. of Illinois
[2]U. of Toronto
rkoenker@uiuc.edu

Empirical Bayes methods for Gaussian compound decision problems involving longitudinal data are considered. The new convex optimization formulation of the nonparametric (Kiefer-Wolfowitz) maximum likelihood estimator for mixture models is employed to construct nonparametric Bayes rules for compound decisions. The methods are first illustrated with some simulation examples and then with an application to models of income dynamics. Using PSID data we estimate a simple dynamic model of earnings that incorporates bivariate heterogeneity in intercept and variance of the innovation process. Profile likelihood is employed to estimate an AR(1) parameter controlling the persistence of the innovations. We find that persistence is relatively modest, when we permit heterogeneity in variances. Evidence of negative dependence between individual intercepts and variances is revealed by the nonparametric estimation of the mixing distribution, and has important consequences for forecasting future income trajectories.

Session 9: Bayesian Methods for Complex Data.

Bayesian Neural Networks for Personalized Medicine

Faming Liang
University of Florida
faliang@ufl.edu

Complex diseases such as cancer have often heterogeneous responses to treatment, and this has attracted much interest in developing individualized treatment rules to tailor therapies to an individual patient according to the patient-specific characteristics. In this talk, we discuss how to use Bayesian neural networks to achieve this goal, including how to select disease related features. The theoretical properties of Bayesian neural networks is studied under the small-n-large-P framework, and simulation is done using the parallel stochastic approximation Monte Carlo algorithm on a multicore computer. The performance of the proposed approach is illustrated via simulation studies and real data examples.

Bayesian Regression Trees for High Dimensional Prediction and Variable Selection

Antonio Linero
Florida State University
arlinero@stat.fsu.edu

A recent stream of research concerns the construction of decision tree ensembles which are motivated by a generative probabilistic model, the most influential method being the Bayesian additive regression trees model of Chipman et al., (2010). Generally speaking, it is well-known that unmodified techniques involving ensembles of regression trees, such as random forests, have performance which degrades sharply as the number of irrelevant predictors increases. We discuss various aspects of the use of Bayesian regression tree models, with emphasis on high dimensional prediction and variable selection under sparsity assumptions. We show that the Bayesian additive regression tree model is not robust to the presence of irrelevant predictors. This is demonstrated in simulations and suggested by a connection to kernel methods. We examine this connection to kernel methods to gain insight into the operating features of Bayesian additive regression trees and, motivated by this connection to kernel-based methods, we propose a novel model we refer to as Dirichlet additive regression trees which is robust to the presence of irrelevant predictors. Unlike Bayesian additive regression trees, our model provides useful posterior inclusion probabilities for each predictor, and performs variable selection without artificially restricting the number of trees used in the ensemble. The Dirichlet additive regression trees prior is shown to overcome these

deficiencies of Bayesian additive regression trees, and to outperform existing methods on both variable selection and prediction tasks.

Novel Statistical Frameworks for Analysis of Structured Sequential Data

♦*Abhra Sarkar and David Dunson*
Duke University
abhra.stat@gmail.com

We are developing a broad array of novel statistical frameworks for analyzing complex sequential data sets. Our research is primarily motivated by a collaboration with neuroscientists trying to understand the neurological, genetic and evolutionary basis of human communication using bird and mouse models. The data sets comprise structured sequences of syllables or 'songs' produced by animals from different genotypes under different experimental conditions. Simple first order Markov chains are insufficiently flexible to learn complex serial dependency structures and systematic patterns in the vocalizations, an important goal in these studies. To this end, we have developed a sophisticated nonparametric Bayesian approach to higher order Markov chains building on probabilistic tensor factorization techniques. Our proposed method is of very broad utility, with applications not limited to analysis of animal vocalizations, and provides new insights into the serial dependency structures of many previously analyzed sequential data sets arising from diverse application areas. Our method has appealing theoretical properties and practical advantages, and achieves substantial gains in performance compared to previously existing methods. Our research also paves the way to advanced automated methods for more sophisticated dynamical systems, including higher order hidden Markov models that can accommodate more general data types.

Bayesian Multiple Classification with Frequent Pattern Mining

♦*Wensong Wu and Tan Li*
Florida International University
wenswu@fiu.edu

In this presentation we consider a two-class classification problem, where the goal is to predict the class membership of M units based on the values of high-dimensional categorical predictor variables as well as both the values of predictor variables and the class membership of other N independent units. We focus on applying generalized linear regression models with Boolean expressions of categorical predictors. We consider a Bayesian and decision-theoretic framework, and develop a general form of Bayes multiple classification function (BMCF) with respect to a class of cost-weighted loss functions. In particular, the loss function pairs such as the proportions of false positives and false negatives, and (1-sensitivity) and (1-specificity), are considered. The best Boolean expressions are selected by a two-step data driven procedure, where the candidates are first selected by Apriori Algorithm, an efficient algorithm for detecting association rules and frequent patterns, and the final expressions are selected by Bayesian model selection or averaging. The results will be illustrated via simulations and on a Lupus diagnosis dataset.

Session 10: Statistics and its Applications

Cardiovascular clinical trials in the 21st century: Pros and Cons of an inexpensive paradigm

Nancy Geller
National Heart, Lung, and Blood Institute, NIH
gellern@nhlbi.nih.gov

The expense of large cardiovascular clinical trials with cardiovascular event endpoints has led to attempts to simplify trials to make them less complex and easier to implement. The new paradigm is not-too-large, simple, and inexpensive. Innovations include use of registries to find eligible participants and lower per subject costs, simplified data collection, and use of surrogate endpoints rather than cardiovascular events in order to decrease sample size and complete trials more quickly. Several examples will be given to illustrate the pros and cons of this new paradigm.

Optimality of Training/Test Size and Resampling Effectiveness in Cross-Validation

♦*Georgios Afendras and Marianthi Markatou*
SUNY at Buffalo
gafendra@buffalo.edu

Biomarker identification and validation for accurate characterization of cancer subtypes and other disease is a very important aspect of modern medicine. Cross Validation (CV) is used extensively in this context. An important question in CV is whether rules can be established to allow "optimal" sample size selection of the training/test set, for fixed values of the total sample size n. We study the cases of random CV and k-fold CV. We begin by defining the resampling effectiveness of random CV estimators of the generalization error and study its relation to "optimal" training sample size selection. We then define "optimality" via simple statistical rules that amount to selecting the optimal training sample size via minimization of the variance of the test set error. We show that in a broad class of loss functions the optimal training sample size equals half of the total sample size, independently of the data distribution and the data analytic task. We discuss optimal selection of the number of folds in k-fold CV and address the case of classification via logistic regression, substantiating our claims theoretically and empirically. We contrast our results with standard practice in the use of CV.

A Zero-inflated Poisson Model for Species Quantification Based on Shotgun Metagenomic Data

Tony Cai, Hongzhe Li and ♦*Jing Ma*
University of Pennsylvania
jinma@upenn.edu

The development of next-generation technologies has made possible quantification of the composition of the Human microbiome using direct DNA sequencing. Existing methods such as MetaPhlAn quantify the relative abundances of species based only on the number of mapped reads, marker lengths and the total number of mapped reads, but ignore the excess zero counts and the nonuniform effects of markers. In this work, we propose a Zero-inflated Poisson model to quantify microbial abundances based on species-specific markers. Our model takes into account high sparsity and overdispersion of the metagenomic data as well as the marker-specific effects when normalizing the sequencing counts to obtain more accurate quantification of abundances. We present examples to illustrate the effectiveness of our method.

Session 11: Recent Developments of High-Dimensional Hypothesis Testing.

CONDITIONAL MEAN AND QUANTILE DEPENDENCE TESTING IN HIGH DIMENSION

♦*Xianyang Zhang*[1]*, Shun Yao*[2] *and Xiaofeng Shao*[2]
[1]Texas A&M University
[2]University of Illinois at Urbana-Champaign

zhangxiany@stat.tamu.edu

Motivated by applications in biological science, we propose a novel test to assess the conditional mean dependence of a response variable on a large number of covariates. Our procedure is built on the martingale difference divergence recently proposed in Shao and Zhang (2014), and it is able to detect certain type of departure from the null hypothesis of conditional mean independence without making any specific model assumptions. Theoretically, we establish the asymptotic normality of the proposed test statistic under suitable assumption on the eigenvalues of a Hermitian operator, which is constructed based on the characteristic function of the covariates. To account for heterogeneity within the data, we further develop a testing procedure for conditional quantile independence at a given quantile level and provide an asymptotic justification. Empirically, our test of conditional mean independence delivers comparable results to the competitor, which was constructed under the linear model framework, when the underlying model is linear. It significantly outperforms the competitor when the conditional mean admits a nonlinear form. The proposed test is employed to analyze a microarray data set on Yorkshire Gilts to detect significant gene ontology terms which are significantly associated with the thyroid hormone.

Principal Component based Adaptive-weight Burden Test for Quantitative Trait Associations

♦*Xiaowei Wu*[1] *and Dipankar Bandyopadhyay*[2]
[1]Virginia Tech
[2]Virginia Commonwealth University
xwwu0808@gmail.com

High-throughput sequencing techniques, when applied to screen samples from pedigrees or with population structure, yield genotype data with complex correlations attributed to both familial relation and linkage disequilibrium. Accounting for such correlations in genetic association analysis can improve power. To better assess gene-based association, we developed PC-ABT, a novel principal component based adaptive-weight burden test. This method employs "data-driven" weights in a retrospective, mixed-model burden test framework, thus makes full use of genotypic correlations. By adjusting the number of principal components that make major contributions to genetic association, PC-ABT is able to achieve maximized test statistic while controlling degree of freedom of the null distribution, thus overcomes the deficiencies of existing adaptive burden or kernel tests. Simulation studies demonstrated that, while keeping type I error well controlled, PC-ABT is generally more powerful than fixed-weight burden test and family-based SKAT in various scenarios. We illustrated the application of PC-ABT by an analysis of fasting glucose associated genes with common and rare variants from the Framingham Heart study.

Homogeneity Test of Covariance Matrices and Change-Points Identification with High-Dimensional Longitudinal Data

♦*Pingshou Zhong*[1] *and Runze Li* [2]
[1]Michigan State University
[2]Penn State University
pszhong@stt.msu.edu

High-dimensional longitudinal data such as time-course microarray data are now widely available. One important feature of such data is that, for each individual, high- dimensional measurements are repeatedly collected over time. Moreover, these measure- ments are spatially and temporally dependent which, respectively, refers to dependence within each particular time point and among different time points. This paper focuses on testing the homogeneity of covariance matrices of high-dimensional measurements over time against the change-point type alternatives. We allow the dimension of measurements (p) to be much larger than the number of individuals (n). Specifically, a test statistic for the equivalence of covariance matrices is proposed and the asymptotic normality is established. In addition to testing, an estimator for the location of the change point is given whose rate of convergence is established and shown to depend on p, n and the signal-to- noise ratio. The proposed method is extended to locate multiple change points by applying a binary segmentation approach, which is shown to be consistent under some mild conditions. The proposed testing procedure and change-point identification methods are able to accommodate both spatial and temporal dependences. Simulation studies and an application to a time-course microarray data set are presented to demonstrate the performance of the proposed method.

A neighborhood-assisted test for high-dimensional mean vector

♦*Jun Li*[1]*, Yumou Qiu*[2] *and Song Xi Chen*[3]
[1]Kent State University
[2]University of Nebraska-Lincoln
[3]Iowa State University
jli49@kent.edu

Although many tests have been proposed to remedy the classical Hotelling's T2 test in high dimensional setting, the corresponding test statistics are constructed by excluding the sample covariance matrix, because of its noninvertibility. To exploit advantageous effect of data dependence, we propose a novel Neighborhood-Assisted (NA) test with test statistic obtained by replacing the inverse of sample covariance matrix in Hotelling's T2 with the regularized estimator through banding the Cholesky factor. Because of regression interpretation of the Cholesky factor, the proposed NA test explores neighborhood dependence by regressing each component of the random vector to its nearest predecessors, and thus can be more powerful than other tests without taking dependence into account. Most importantly, the NA test is robust to a wide range of dependence in the sense that its implementation does not rely on any structural assumption of the unknown covariance matrix. Simulation and case studies are given to demonstrate the performance of the proposed NA test.

Session 12: Advanced Methodologies in Analyzing Censoring Data

Jackknife Empirical Likelihood for Linear Transformation Models with Censored Data

Hanfang Yang [1]*, Shen Liu*[1] *and* ♦*Yichuan Zhao*[2]
[1]Renmin University of China
[2]Georgia State University
yichuan@gsu.edu

A class of linear transformation models with censored data was proposed by Cheng et al. (1995) as a generalization of Cox models in survival analysis. This paper develops inference procedure for regression parameters based on jackknife empirical likelihood approach. We can show that the limiting variance is not necessary to estimate and the Wilk's theorem can be obtained. Jackknife empirical likelihood benefits from the simpleness in optimization using jackknife pseudo-value. In our simulation studies, the proposed method is compared with the existing methods, such as the traditional empirical likelihood in terms of coverage probability.

Quantile Residual Life Regression with Longitudinal Biomarker Measurements for Dynamic Prediction

♦*Ruosha Li*[1]*, Xuelin Huang*[2] *and Jorge Cortes*[2]

[1]Univ of Texas Health Science Center at Houston
[2] The University of Texas MD Anderson Cancer Center
ruosha.li@uth.tmc.edu

Residual life is of great interest to patients with life-threatening disease. It is also important for clinicians who estimate prognosis and make treatment decisions. Quantile residual life has emerged as a useful summary measure of the residual life. It has many desirable features, such as robustness and easy interpretation. In many situations, the longitudinally collected biomarkers during patients' follow-up visits carry important prognostic value. In this work, we study quantile regression methods that allow for dynamic predictions of the quantile residual life, by flexibly accommodating the post-baseline biomarker measurements in addition to the baseline covariates. We propose unbiased estimating equations that can be solved via existing L-1 minimization algorithms. The resulting estimators have desirable asymptotic properties and satisfactory finite-sample performance. We apply our method to a study of chronic myeloid leukemia to demonstrate its usefulness as a dynamic prediction tool.

Promotion time cure model with nonparametric form of covariate effects

Pang Du[1] *and* ♦*Tianlei Chen*[2]
[1]Virginia Tech
[2]Celgene
tlchenvt@gmail.com

Survival data with a cured portion are commonly seen in clinical trials. Motivated from a biological interpretation of cancer metastasis, promotion time cure model is a popular alternative tool to the mixture cure rate model for analyzing such data. The existing promotion cure models all assume a restrictive parametric form of covariate effects, which can be incorrectly specified especially at the exploratory stage. In this talk, we present a nonparametric approach to modeling the covariate effects under the framework of promotion time cure model. The covariate effect function is estimated by smoothing splines via the optimization of a penalized profile likelihood. Point-wise interval estimates are also derived from the Bayesian interpretation of the penalized profile likelihood. Simulations show excellent performance of the proposed nonparametric method which is then applied to a melanoma study.

Bayesian semiparametric regression models for interval-censored data

♦*Xiaoyan Lin, Lianming Wang and Bo Cai*
University of South Carolina
lin9@mailbox.sc.edu

Interval-censored time to event data commonly occur in many fields such as demographical, epidemiological, and medical studies. In these studies, participants usually undergo periodical observations or examinations, and the time of event is not observed exactly but is known to fall within some interval. Analyzing such interval-censored data is challenging due to the complicated data structure and censoring scheme. In this poster, we present three recently proposed Bayesian methods (Lin and Wang, 2010; Wang and Lin, 2011; Lin et al., 2014) for analyzing interval-censored data under the semiparametric Probit, proportional odds, and proportional hazards models. The approaches share two common strategies. First, they all adopt monotone splines (Ramsay 1988) to model certain unknown baseline nondecreasing functions, and therefore to produce smooth estimates of baseline functions. Second, they all adopt certain data augmentation to facilitate Bayesian computation. Unlike many existing Bayesian methods in the literature, our developed Gibbs samplers are efficient and easy to execute because they do not require imputing any unobserved failure times or contain any complicated Metropolis-Hastings steps.

Session 13: Recent Advancement in Adaptive Design of Early Phase Clinical Trials by Accounting for Schedule Effects or Using Other Approaches

A Subgroup Cluster Based Bayesian Adaptive Design for Precision Medicine

Wentian Guo[1], ♦*Yuan Ji*[2] *and Daniel Catenacci*[3]
[1]Fudan University, Shanghai, China
[2]Northshore University/University of Chicago
[3]University of Chicago Medical Center
koaeraser@gmail.com

In precision medicine, a patient is treated with targeted therapies that are predicted to be effective based on the patient's baseline characteristics such as biomarker profiles. Oftentimes, patient subgroups are unknown and must be learned through inference using observed data. We present SCUBA, a Subgroup ClUster based Bayesian Adaptive design aiming to fulfill two simultaneous goals in a clinical trial, 1) to report multiple subgroup-treatment pairs (STPs) and 2) to precisely allocate patients to their desirable treatments. Using random partitions and semi-parametric Bayesian models, SCUBA provides coherent and probabilistic assessment of potential patient subgroups and their associated targeted therapies. Each STP can then be used for future confirmatory studies as precision treatment. Through extensive simulation studies, we present an application of SCUBA to an innovative clinical trial in gastroesphogeal cancer.

Phase I design for locating schedule-specific maximum tolerated doses

Nolan Wages
University Of Virginia
nwages@virginia.edu

The majority of methods for the design of Phase I trials in oncology are based upon a single course of therapy, yet in actual practice it may be the case that there is more than one treatment schedule for any given dose. Therefore, the probability of observing a dose-limiting toxicity (DLT) may depend upon both the total amount of the dose given, as well as the frequency with which it is administered. The objective of the study then becomes to find an acceptable combination of both dose and schedule. It may be of interest to find multiple MTD's, one for each schedule, for further testing for efficacy in a Phase II setting. In this talk, we present a two-dimensional dose-finding method that extends the continual reassessment method to account for the location of an MTD within each schedule being investigated. Operating characteristics are demonstrated through simulation studies, and some brief discussion of implementation and available software is also provided.

TITE-CRM method incorporating cyclical safety data with application to oncology phase I trials

Bo Huang
Pfizer
bo.huang@pfizer.com

Delayed dose limiting toxicities (i.e. beyond first cycle of treatment) is a challenge for phase I trials. The time-to-event continual reassessment method (TITE-CRM) is a Bayesian dose-finding design to address the issue of long observation time and early patient drop-out. It uses a weighted binomial likelihood with weights assigned to observations by the unknown time-to-toxicity distribution,

and is open to accrual continually. To avoid dosing at overly toxic levels while retaining accuracy and efficiency for DLT evaluation that involves multiple cycles, we propose an adaptive weight function by incorporating cyclical data of the experimental treatment with parameters updated continually (Huang and Kuan, 2014). This provides a reasonable estimate for the time-to-toxicity distribution by accounting for inter-cycle variability and maintains the statistical properties of consistency and coherence. Case studies are presented using the proposed design. Design calibrations for the clinical and statistical parameters are conducted to ensure good operating characteristics. Simulation results show that the proposed TITE-CRM design with adaptive weight function yields significantly shorter trial duration, does not expose patients to additional risk, is competitive against the existing weighting methods, and possesses some desirable properties. We also share our experience in the trial conduct by effectively using a Dose Escalation Steering Committee (DESC).

A dose-schedule-finding design for phase I/II clinical trials

Beibei Guo[1], ♦*Yisheng Li*[2] *and Ying Yuan*[2]
[1]Louisiana State University
[2]University of Texas MD Anderson Cancer Center
ysli@mdanderson.org

Dose finding methods aiming at identifying an optimal dose of a treatment with a given schedule may be at a risk of misidentifying the best treatment for patients. We propose a phase I/II clinical trial design to find the optimal dose-schedule combination. We define schedule as the method and timing of administration of a given total dose in a treatment cycle. We propose a Bayesian dynamic model for the joint effects of dose and schedule. The model proposed allows us to borrow strength across dose-schedule combinations without making overly restrictive assumptions on the ordering pattern of the schedule effects. We develop a dose-schedule finding algorithm to allocate patients sequentially to a desirable dose-schedule combination, and to select an optimal combination at the end of the trial. We apply the proposed design to a phase I/II clinical trial of a γ-secretase inhibitor in patients with refractory metastatic or locally advanced solid tumors, and we examine the operating characteristics of the design through simulations.

Session 14: Contemporary Statistical Methods for Complex Data

Bridging density functional theory and big data analytics with applications

Chien-Chang Chen[1], *Hung-Hui Juan*[2], *Meng-Yuan Tsai*[2] *and* ♦*Henry Horng-Shing Lu*[2]
[1]National Central University, Taiwan
[2]National Chiao Tung University, Taiwan
hslu@stat.nctu.edu.tw

The framework of the density functional theory (DFT) reveals both strong suitability and compatibility for investigating large-scale systems in the Big Data regime. By technically mapping the data space into physically meaningful bases, the article provides a simple procedure to formulate global Lagrangian and Hamiltonian density functionals to circumvent the emerging challenges on large-scale data analyses. Then, the informative features of mixed datasets and the corresponding clustering morphologies can be visually elucidated by means of the evaluations of global density functionals. Simulation results of data clustering illustrated that the proposed methodology provides an alternative route for analyzing the data characteristics with abundant physical insights. For a comprehensive demonstration in a high dimensional problem without prior ground truth, the developed density functionals were also applied on the post-process of magnetic resonance imaging (MRI) and better tumor recognitions can be achieved on the T1 post-contrast and T2 modes. It is appealing that the post-processing MRI using the proposed DFT-based algorithm would benefit the scientists in the judgment of clinical pathology and the applications of high dimensional biomedical image processing. Eventually, successful high dimensional data analyses reveal that the proposed DFT-based algorithm has the potential to be used as a framework for investigations of large-scale complex systems.

Nonparametric divergence-based flow cytometric classification

♦*Ollivier Hyrien and Andrea Baran*
University of Rochester
ollivier_hyrien@urmc.rochester.edu

Flow cytometry is a single-cell assay routinely used in clinical settings for disease diagnosis. The construction of supervised classifiers using flow cytometry data often requires two steps. In the first step, candidate features, for example clusters, are extracted from the data. In the second step, a decision rule is built using supervised learning techniques. When the number of candidate features is large, regularization and variable selection procedures are invoked to eliminate unnecessary features. In this talk, we present a different approach to flow cytometric classification in which class assignment is performed by summarizing the evidence provided by the data that a given phenotype belongs to each of the available classes by means of an f-divergence. The proposed approach eliminates the construction and selection of candidate features. It makes no distributional assumptions about phenotypes, and automatically integrates predictive patterns in the classifier. We present properties of classifiers that rely on pairwise dissimilarities and offer a comparison with centroid-based classifiers. We find that the curse of dimensionality is efficiently addressed via dimensionality reduction.

Untangle the Structural and Random Zeros in Statistical Modelling

♦*Hua He*[1], *Wan Tang*[2], *Wenjuan Wang*[3], *Naiji Lu*[4] *and Ding-Geng Chen*[5]
[1]Tulane University
[2]Tulane University
[3]Brightech International, LLC
[4]Huazhong University of Science and Technology
[5]University of North Carolina, Chapel Hill
hhe2@tulane.edu

We will propose a new approach to address the problems resulting from ignoring the difference between structural and random zeros in zero-inflated count explanatory variables while performing regression analysis. Without paying close attention to these structural zeros in explanatory variables, the estimates for parameters in the model can be biased and results in misleading conclusions. Authors have shown such issue through intensive simulation studies in this manuscript. As a solution, including an indicator of the structural zero in the regression model as a predictor could dramatically solve the bias issue. However, such indicator is often unobserved in practice. We will propose a new approach that is simply based on the structure of hierarchical model to address the bias issue when the indicator of structural zero is unobserved. The estimation is conducted using maximum likelihood estimation. The real data from NHANES are used to illustrate the utilization of the proposed approach.

Covariance Structures and Estimation for Axially Symmetric

Spatial Processes on the Sphere

♦*Haimeng Zhang*[1]*, Chunfeng Huang*[2] *and Scott Robeson*[2]
[1]University of North Carolina - Greensboro
[2]Indiana University - Bloomington
h_zhang5@uncg.edu

In this presentation, I will discuss a class of axially symmetric spatial processes in the analysis of global scale data, whereby their covariance function depends on differences in longitude alone. In addition, a simplified representation of a valid axially symmetric process and its covariance function will be presented. Further construction of parametric covariance models for axially symmetric processes will be explored, and some preliminary results on the estimation will be discussed.

Session 15: Recent Development in Time-to-Event Data Analysis

A NPMLE Approach for Extended Cure Rate Model with Left Truncation and Right-Censoring

♦*Jue Hou and Ronghui Xu*
University of California, San Diego
j7hou@ucsd.edu

The analysis of spontaneous abortion (SAB) data collected via observational studies in pregnancy demands modification to the traditional cure-rate setting (Farewell, 1982). Such data has a observable 'cured' portion as the survivors at the well-defined finite upper bound of failure time. The data is also subject to left truncate in addition to right-censoring because women may enter or withdraw from a study any time during their pregnancy. Left truncation in particular causes unique bias in the presence of a cured portion. In our paper, we extend the classical cure rate model to accommodate such data by proposing a conditional nonparametric maximum likelihood approach (NPMLE). To tackle the computational challenge brought by left truncation, we develop a rapid algorithm for NPMLE inspired by the "ghost copy" EM from Qin et al (2011), using existing glm and coxph solvers. A closed form variance estimator for EM is derived following Louis (1982). Under weaker assumptions, we prove the consistency of the resulting estimator involving an unbounded baseline hazard. We then show the asymptotic normality with stronger assumptions. Simulation results are presented to illustrate finite sample performance. We present the analysis of the motivating SAB study to illustrate the power of addressing both occurrence and timing of failure times in practice.

A Semiparametric Joint Model for Longitudinal and Survival Data in End-of-Life Studies

♦*Zhigang Li*[1]*, HR Frost*[1]*, Tor Tosteson*[1]*, Lihui Zhao*[2]*, Lei Liu*[2]*, Huaihou Chen*[3] *and Marie Bakitas*[4]
[1]Dartmouth College
[2]Northwestern University
[3]University of Florida
[4]The University of Alabama at Birmingham
Zhigang.Li@dartmouth.edu

A unique feature of end-of-life (EOL) research studies, such as hospice/palliative care studies, is the short life expectancy of participants which makes it important to study the terminal trajectory of longitudinal data in these studies. Survival data are also commonly seen in such studies and strongly correlated with longitudinal data. Without appropriate handling of the correlation, estimate for longitudinal trajectory will be inefficient with distorted clinical interpretation. Unfortunately, censoring of the survival times occurs frequently and makes it challenging to account for the correlation. To address these issues, we propose a novel semiparametric statistical approach for jointly modeling longitudinal and survival data in EOL studies. There are two sub-models in our joint modeling approach: a semiparametric mixed effects model for the longitudinal data and a Cox model for the survival data. Regression splines method with natural cubic B-splines are used to estimate the nonparametric curves and AIC is employed to select knots. Inference for quality-adjusted life years is provided as well using this approach. Performance of the model is assessed through simulation and we also apply the model to a recently completed randomized clinical trial in palliative care research to establish a standard modeling approach in the field of EOL research.

Group variable selection in survival and competing risks model

♦*Kwang Woo Ahn, Natasha Sahr, Anjishnu Banerjee and Soyoung Kim*
Medical College of Wisconsin
kwooahn@mcw.edu

We propose an adaptive group bridge method, enabling simultaneous variable selection both within and between groups, for survival and competing risks data. The adaptive group bridge is applicable to independent and clustered data. It also allows the number of variables to diverge as the sample size increases. We show that our new method possesses the oracle property, including variable selection consistency at group and within-group levels. We also show superior performance in simulation study over several competing approaches. A real bone marrow transplant data is also illustrated.

Sample Size for Joint Testing Cause-Specific Hazard and Overall Hazard in the Presence of Competing

♦*Qing Yang*[1]*, Wing Kam Fung*[2] *and Gang Li*[3]
[1]Duke University
[2]University of Hong Kong
[3]University of California, Los Angeles
qing.yang@duke.edu

It has been well recognized that in the presence of competing risks, the effects of a variable on the time to the occurrence of a particular type of failure, say type 1 failure, is not completely characterized by the type 1 cause-specific hazard (CSH1) alone. Additional quantities such as the cumulative incidence function for type 1 failure, the overall hazard (OH) due to any cause, and the cause specific hazard due other causes other than 1, need to be considered jointly. In this talk, we consider sample size determination for jointly testing CSH1 and OH based on the joint tests recently developed by Li and Yang [1]. These pair of quantities correspond to important study endpoints such as the disease specific survival and overall survival, which are frequently used as co-primary endpoints in clinical trials. Simulations are used to illustrate potential savings on sample size using the joint tests as compared to the Bonferroni adjustment method. An R package has been developed to implement our methods. We illustrate our methods and the potential sample size saving of the joint tests over the Bonferroni method through simulations and the 4-D (Die Deutsche Diabetes Dialyse Studie) clinical trial.

Session 16: Statistics and Big Data

Jackknife empirical likelihood inference for AFT models

♦*Xue Yu and Yichuan Zhao*
Georgis State University
xyu6@gsu.edu

Accelerated failure time (AFT) model is a parametric and useful model that provides an alternative to the commonly used proportional hazards models on the survival function, and more importantly, the responses are subject to censoring. However, it is difficult to compute the estimators of regression parameters because the most widely used rank-based estimating equations are not smoothing. Brown and Wang (2007) use an induced smoothing approach that smooths the estimating functions in order to obtain point and variance estimators. In this paper, we use jackknife empirical likelihood (JEL) method to make statistical inference for AFT models without computing the limiting variance. We also compare the JEL method and the existing normal approximation (NA) methods, and the extensive simulation results suggest that the JEL method provides valid inferences for AFT models. We also apply proposed method to two real data sets.

Jackknife Empirical Likelihood for the Concordance Correlation Coefficient

♦*Anna Moss and Yichuan Zhao*
Georgia State University
smoss8@student.gsu.edu

The concordance correlation coefficient (CCC) is a common measure of reproducibility or agreement between data values in paired samples. Confidence intervals and hypothesis tests of the CCC using normal approximations (NA) have been shown to have poor coverage for highly skewed distributions. This study applies the jackknife empirical likelihood (JEL) to confidence intervals for the CCC and compares coverage probability and interval length for JEL and NA methods. Data are simulated for values of the CCC between 0.25 - 0.95 from normal and non-normal distributions of varying skewness. Simulation results showed that JEL methods perform better than the NA methods particularly with data from skewed distributions.The JEL methods have the widest confidence intervals in most cases. Application of JEL methods are illustrated by evaluating concordance between self-reported and clinically measured body weight and height from the National Health and Nutrition Examination Survey (NHANES).

Jackknife Empirical Likelihood for the Mean Difference of Two Zero Inflated Skewed Populations

♦*Faysal Satter and Yichuan Zhao*
Georgia State University
faysal1.618@live.com

Jackknife empirical likelihood (JEL) method was proposed to construct non-parametric confidence interval for the mean difference of two independent zero inflated skewed populations. These confidence intervals were compared with confidence interval assuming normality. Simulation studies are carried out. Finally, the method was implemented on a real life data.

Rank-based estimating equation with non-ignorable missing responses under empirical likelihood

♦*Huybrechts F Bindele*[1] *and Yichuan Zhao*[2]
[1]University of South Alabama
[2]Georgia State University
hbindele@southalabama.edu

In this talk, we will consider a general regression model with responses missing not at random. We will consider a rank-based estimating equation of the regression parameter from which a rank-based estimator will be derived. Based on its asymptotic normality property, a consistent sandwich estimator of the corresponding asymptotic covariance matrix will discussed. In order to overcome the under coverage issue of the normal approximation procedure, the empirical likelihood based on the rank-based gradient function will be introduced, and its asymptotic distribution established. Extensive simulation experiments under different settings of error distributions with different missing mechanisms will be considered, and the simulation results will show that the proposed empirical likelihood approach has better performance in terms of coverage probability and average length of confidence intervals for the regression parameters compared with the normal approximation approach and its least-squares counterpart. A real data example will be provided to illustrate our methods.

Session 17: Recent Advances in Design Problems

Considerations for Pediatric Trial Designs and Analyses

Meehyung Cho, ♦*Zhiying Qiu, Jenny Ye, Hui Quan and Peng-Liang Zhao*
Sanofi US
zhiying.qiu@sanofi.com

Pediatric trials are often conducted to obtain extended marketing exclusivity or to satisfy regulatory requirements. There are many challenges in designing and analyzing pediatric trials arising from special ethical issues and the relatively small accessible patient population. The application of conventional phase 3 trial designs to pediatrics is generally not realistic in some therapeutic areas. In this presentation we review regulatory guidance and existing research in pediatrics. We then examine different approaches for designing a pediatric trial and analyzing outcomes. We consider weighted combination methods utilizing available adult data such as James-Stein shrinkage estimates, empirical shrinkage estimates and Bayesian methods. We also consider the idea of concept of consistency used in multi-regional trials and apply to design and analysis of a pediatric trial. The performance of these methods is assessed through simulation.

Dose-finding Designs Incorporating Toxicity and Efficacy

♦*Jun Yin*[1]*, Monia Ezzalfani*[2]*, Dan Sargent*[1] *and Sumithra Mandrekar*[1]
[1]Mayo Clinic
[2]Institut Curie
Yin.Jun@mayo.edu

Recent development of cancer immunotherapies and molecularly targeted agents (MTAs) require innovative approaches beyond the traditional phase I designs commonly used for cytotoxic agents. Specifically, immunotherapies and MTAs often demonstrate reduced rates of severe toxic reactions, however their adverse effects may persist or even accumulate over multiple ongoing treatment cycles, as opposed to the acute toxic reactions from cytotoxic agents that often occur in the first treatment cycles. In addition, since maximizing the dose of immunotherapies and MTAs may present no further clinical benefit compared to an intermediate dose, it is important to also consider efficacy endpoints in dose assignment, so that an intermediate dose with sufficient efficacy can be identified as the recommended phase 2 dose. Under these circumstances, the focus of phase I trials has thus shifted from finding the maximum tolerated dose (MTD) to identifying the maximum effective dose (MED). The MED is quantified by a measure of efficacy (in addition to safety) that can be measured reliably within a short period of time, for example, tumor shrinkage, change in expression levels of a marker, and etc. We review two directions in novel model-based designs utilizing both toxicity and efficacy in phase I: (1) Bivariate analysis of toxicity and efficacy in joint models; (2) Trivariate analysis on ordinal toxicity and efficacy outcome. In addition, since most

of these designs consider both toxicity and efficacy as binary, we investigate joint modeling of a continuous toxicity score and efficacy measure from multiple treatment cycles. Simulation studies have demonstrated favorable performance of these model-based designs; Practical considerations and challenges that limit the implementation of these designs will also be discussed.

Statistical Considerations in the Design and Analysis of Reference Database for OCT Device

Haiwen Shi
FDA/CDRH
Haiwen.Shi@fda.hhs.gov

Optical Coherence Tomography (OCT) is a medical imaging technology invented in 1991. Since its invention, the OCT has been applied for imaging of eyes and has had largest impact in ophthalmology. The output measures from OCT have been used for diagnosis and monitoring of retinal diseases such as glaucoma. To better understand and explain the measures, some OCT devices are associated with a reference database, which consists of apparently normal subjects. The OCT measures on these normal subjects comprise a baseline distribution. Some percentiles of the distribution, e.g. 1st and 5th percentiles can be used as reference limits, with which the measure from a new patient can be compared to find the possibility of exposing to certain eye disease. In this talk, I will discuss some statistical considerations in the design and analysis reference database for OCT device. It is not uncommon that the OCT measures are influenced by some covariates, e.g. age, disc size etc. Thus they should be adjusted by these covariates. I will talk about how to determine the relevant covariates and different methods of adjusting them. Specifically, I will go through three methods, stratified method, multiple-variable linear regression, and quantile regression. The pros and cons of the three methods will be compared and elaborated.

Design Considerations for Non-randomized Medical Device Clinical Studies

♦*Heng Li, Vandana Mukhi and Yun-Ling Xu*
FDA/CDRH
heng.li@fda.hhs.gov

A Non-randomized study design encompasses studies without a control group and studies that use patient-level control arm data from previous studies or registries. We will share some key elements of the use of objective performance criteria (OPC) and performance goal (PG) in studies without a control group. I will also share some key elements of the process of submitting an investigational device exemption (IDE) of a non-randomized clinical study that involves the use of propensity score (PS) methodology.

Bayesian Analysis of Disease Modification using Doubly-Randomized Delay-Start Matched Control Design

♦*Ibrahim Turkoz*[1] *and Marc Sobel*[2]
[1]Janssen Research&Development, LLC
[2]Temple University
iturkoz@its.jnj.com

Randomized Delayed-Start designs introduce major drawbacks for evaluating disease modifying agents. A high percentage of subjects in the control arm are expected to drop out before entering into the delayed-start period due to the long treatment period necessary to show the effects on disease progression. This results in major drawbacks for existing analytic approaches. The proposed study design resolves issues of randomized delayed-start trials and makes it feasible to establish a disease modification indication. The proposed innovation is a hybrid of randomized and epidemiologic design that introduces a run-in period and the second randomization for delayed-start. The run-in period is used for a matched control analysis. In the delayed-start period, the objective is to show that the late initiation of treatment does not permit the same level of recovery as experienced by those subjects who have been on the drug from the start. The "failure to catch up" concept is evaluated using hierarchical priors. Bayesian methodology is a natural fit for establishing disease modification compared to the traditional non-inferiority margins or parallel lines approach.

A number of models are developed to characterize disease modification. These models are roughly divided into two groups. The first group of models has a growth curve describing evolution with slope parameters which are the same for all subjects. This group of includes both simple linear and simple non-linear spline models. The second group has a growth curve describing subject specific evolution. This group of models includes linear random intercept and linear random intercept&random slope models, and their spline counterparts. We compare the models with regards to both how well they fit the current data and how well they can predict the future.

A prospective-retrospective study design to access the clinical performance of an in-vitro diagnosti

Shiling Ruan
Allergan Plc
ruan_shiling@allergan.com

In vitro diagnostic products (IVD's) are those reagents, instruments, and systems intended for use in diagnosis of disease or other conditions, including a determination of the state of health, in order to cure, mitigate, treat, or prevent disease or its sequelae. Clinical performance studies are designed to evaluate whether the IVD test is suitable for the study objective, the intended use, and its intended population when this cannot be addressed with the analytical performance data, literature and/or experience gained by routine diagnostic. The clinical performance studies can be either observational studies or interventional studies. In an interventional study, the test results obtained during the study may influence patient management decisions and may be used to guide treatments, whereas in an observational study, the test results obtained during the study are not used for patient management and do not impact treatment decisions. This talk presents a prospective-retrospective observational study design to evaluate the clinical performance of an IVD test intended to guide patient treatment. The considerations on choosing between the interventional and observational study design are discussed, together with the statistical considerations (hypothesis and analysis) of the proposed design.

Session 18: New Statistical Computing using R

Rank-Based Tests for Clustered Data with R Package clusrank

♦*Yujing Jiang*[1]*, Jun Yan*[1] *and Mei-Ling Ting Lee*[2]
[1]University of Connecticut
[2]University of Maryland
yujing.jiang@uconn.edu

Rank based tests are popular distribution-free alternatives to the popular one-sample and two-sample t-tests. For independent data, they are available in the base package of R. For clustered data, several rank-based tests that accounts for the within-cluster dependence are recently developed, but no existing package provides all of them at one place with ease of access. In package clusrank, the latest developments are implemented and wrapped under a unified high-level user-friendly function. Different methods are dispatched based on an input formula, offering great flexibility in handling various

situations such as unbalanced cluster sizes, presence of ties, stratified data, and heterogeneous group members from the same cluster. Methods based on both asymptotics for large samples and permutation for small samples are available. We present some details about the implemented methods and a comprehensive comparison of them in a simulation study.

The R Package "threg" to Implement Threshold Regression: A model for time-to-event survival data

Mei-Ling Ting Lee
University of Maryland
mltlee@umd.edu

I will introduce the R package "threg", which implements the estimation procedure of an intuitive model based on the first time that the sample path of a Wiener process hits a boundary threshold. The threshold regression methodology is well suited to applications involving survival and time-to-event data, it is as an important alternative to the Cox model without the need of proportional hazards assumption. This new package includes four functions: the "threg", the hazard ratio function "hr", and "predict" and "plot" for object

Threshold Regression for Interval-Censored Data for Cure-Rate or without Cure-Rate Model

♦*Man-Hua Chen*[1] *and Mei-Ling Ting Lee*[2]
[1]Tamkang University
[2]University of Maryland
mchen@mail.tku.edu.tw

The threshold regression (TR) technique bases on the inverse Gaussian distribution can deal with nonproportional hazards is useful alternative to the Cox proportional hazards model to analyze a first-hitting-time (FHT) survival data. The R package threg (Xiao 2013) has been designed for implements the estimation procedure of a threshold regression model (TR) with right censored data. Because of usefulness in real applications, subjects commonly examining or observing periodically in medical follow up studies, we consider the interval censored data for TR and a plug-in cure rate option. The R package thregI is completed with two data sets, one of a breast cancer study (bcos) and NASA's Hypobaric Decompression Sickness Data Bank (hdsd). Eight relevant functions have been programmed.

Session 19: Statistical Modeling and Inference on Complex Biomedical Data

A Generalized Estimating Equations Framework for the Analysis of Intracellaur Cytokine Staining Data

♦*Amit Meir*[1]*, Raphael Gottardo*[2] *and Greg Finak*[2]
[1]University of Washington
[2]Hutch Cancer Research Center
amitmeir@uw.edu

Intracellular Cytokine Staining (ICS) - a type of cytometry experiment used to measure cytokine production at the single cell level - is an important measure used in immune monitoring and vaccine development. A well-known challenge in analyzing flow cytometry data is that they are prone to batch and technical variation, but also produce many correlated features (cell subsets). These effects are often ignored; cell subsets are treated independently, counts are modeled as proportions, and batch effects are not systematically accounted for. We propose a generalized estimating equation modeling framework for analyzing cytometry count data, allowing for the screening of cell populations while accounting for both technical and biological nuisance factors. Immune system responses tend to vary significantly among individuals, such that some experience dramatic shift in cellular composition (ie cell subset counts) in response to stimuli while others experience little or no change. To account for the possible variable treatment effect, we assume a regression mixture model and estimate it using a marginal EM algorithm adapted for use with generalized estimating equations. We demonstrate our methodology by applying it to experimental assays measuring cytokine expression at the single-cell level.

Haplotyping and quantifying allele-specific expression at gene and gene isoform level by Hybrid-Seq

Benjamin Deonovic[1]*, Yunhao Wang*[2]*, Jason Weirather*[3] *and* ♦*Kin Fai Au*[3]
[1]Department of Biostatistics, University of Iowa
[2]University of Chinese Academy of Sciences
[3]Department of Internal Medicine, Univ. of Iowa
kinfai-au@uiowa.edu

The haplotype phase problem is to find the true combination of genetic variants on a single chromosome from individuals. Furthermore, haplotypes of a gene can be expressed non-equally, a phenomenon known as allele-specific expression (ASE). Haplotype phasing and quantification of ASE are essential for studying the association between genotype and disease. No existing method solves these two intrinsically linked problems together. Rather, most current strategies have great dependence on known haplotypes or family trio data. Herein, we present a novel method, IDP-ASE, which utilizes a Bernoulli mixture model for RNA-seq data and MCMC to derive the most likely set of haplotypes, phase each read to a haplotype, and estimate ASE. Our model leverages the strengths of both Second Generation Sequencing (SGS) and Third Generation Sequencing (TGS). The long read length of TGS data facilitates phasing, while the accuracy and depth of SGS data facilitates estimation of ASE. Moreover, IDP-ASE is capable of estimating ASE at both the gene and isoform level. We present the performance of IDP-ASE on simulation data and apply it to data from various real data sets which harbor extensive ASE events.

Intrinsic Noise in Nonlinear Gene Regulation Inference

♦*Chao Du*[1] *and Wing.H. Wong*[2]
[1]University of Virginia
[2]Stanford University
cd2wb@virginia.edu

Cellular intrinsic noise plays an essential role in the regulatory interactions between genes. Although a variety of quantitative methods are used to study gene regulation system, the role of intrinsic noises has largely been overlooked. Using the Kolmogorov backward equation (master equation), we formulate a causal and mechanistic Markov model. This framework recognizes the discrete, nonlinear and stochastic natures of gene regulation and presents a more realistic description of the physical systems than many existing methods. Within this framework, we develop an associated moment-based statistical method, aiming for inferring the unknown regulatory relations. By analyzing the observed distributions of gene expression measurements from both unperturbed and perturbed steady-states of gene regulation systems, this method is able to learn valuable information concerning regulatory mechanisms. This design allows us to estimate the model parameters with a simple convex optimization algorithm. We apply this approach to a synthetic system that resembles a genetic toggle switch and demonstrate that this algorithm can recover the regulatory parameters efficiently and accurately.

A Kernel-Based Approach to Covariate Adjustment for Causal

Inference

♦*Yeying Zhu*[1]*, Jennifer Savage*[2] *and Debashis Ghosh*[3]
[1]University of Waterloo
[2]Pennsylvania State University
[3]Colorado School of Public Health
yeyingzhu12@gmail.com

One of the commonly used approaches to the causal analysis of observational data is matching, which involves taking treated subjects and finding comparable control subjects who have either similar covariate values and/or propensity score values. An important goal is to achieve balance in the covariates among the treatment groups. In this talk, we introduce the concept of distributionally balance preserving which requires the distribution of the covariates to be the same in different treatment groups. We also propose a new balance measure called kernel distance, which is the empirical estimate of the probability metric defined in the reproducing kernel Hilbert spaces. Compared to the traditional balance metrics, the kernel distance measures the difference in the two multivariate distributions instead of the difference in the finite moments of the distributions. We then incorporate kernel distance into genetic matching, the state-of-the-art matching procedure. Simulation results show that the kernel distance is the best indicator of bias in the estimated casual effect compared to several commonly used balance measures and it can improve the performance of genetic matching. We apply the proposed approach to analyze the Early Dieting in Girls study. The study indicates that mothers' overall weight concern increases the likelihood of daughters' early dieting behavior.

Session 20: Statistical Advances in Omics Data Integration

Genomic Determination Index

♦*Cheng Cheng, Wenjian Yang, Robert Autry, Steven Paugh and William Evans*
St. Jude Children's Research Hospital
cheng.cheng@stjude.org

In this presentation we introduce and discuss a new concept called Genomic Determination Index (GeDI), to address the question of how much variability in a phenotype can be determined by large sets of diverse genomic factors totalling to a number of millions. In a way similar to heritability, GeDI is a measurement of the proportion of the phenotype variance attributable to the variations in a set of genomic factors under study. No existing large-scale sparse regression method can effectively address this problem. A method to estimate GeDI is presented. This method consists of two parts: a step-wise regression with forward variable selection, and a permutation analysis for bias correction. The entire development will be illustrated and evaluated by a diverse dataset from a study of ex vivo sensitivity of acute lymphoblastic leukaemia cells to glucocorticoid treatment.

graph-GPA: A graphical model to prioritizing GWAS results by integrating pleiotropy

♦*Dongjun Chung*[1]*, Hang Kim*[2] *and Hongyu Zhao*[3]
[1]Medical University of South Carolina
[2]University of Cincinnati
[3]Yale University
chungd@musc.edu

Results from Genome-Wide Association Studies (GWAS) have shown that complex diseases are often affected by many genetic variants with small or moderate effects. Identification of these risk variants remains a very challenging problem. Hence, there is a need to develop more powerful statistical methods to leverage available information to improve upon traditional approaches that focus on a single GWAS dataset. Our study was motivated by the accumulating evidence suggesting that different complex diseases share common risk bases, i.e., pleiotropy. In this presentation, I will discuss our novel statistical approach, graph-GPA, to increase statistical power to identify risk variants through joint analysis of multiple GWAS data sets using a graphical modeling approach. Moreover, graph-GPA provides a parsimonious representation of genetic relationship among phenotypes, which is especially powerful when an increasing number of phenotypes are jointly studied. I will discuss the power of graph-GPA with the simulation studies and its application to real GWAS datasets.

Pathway based integrative study of omics data for predicting cancer prognosis in TCGA melanoma

♦*Yu Jiang*[1]*, Xingjie Shi*[2]*, Qing Zhao*[3]*, Cen Wu*[4] *and Shuangge Ma*[5]
[1]University of Memphis
[2]Nanjing University of Finance and Economics
[3]Merck Research Laboratories
[4]Kansas State University
[5]Yale University
yjiang4@memphis.edu

The progression of cutaneous melanoma is a complex process involving genetic and epigenetic changes in multiple pathways. Approaches that focused on multiple types of (epi)genetic measurement and pathway analysis would help to comprehensively describe the biological processes and identify prognostic markers. The multiple (epi)genetic measurements from The Cancer Genome Atlas (TCGA) cutaneous melanoma makes it possible for us to model the prognosis comprehensively. In the current study, we conduct integrated pathway analysis using dimension reduction and group variable selection methods. It has been found that integration of clinical variables with genomic measurements has improved performance in prediction of cancer survival. Our study identifies pathways that may play a significant role in melanoma prognosis. The identified pathways are worth of further investigation as biomarkers for melanoma progression and potential therapeutic targets for treatments.

Joint Precision Matrix Estimation with Sign Consistency

♦*Yuan Huang, Qingzhao Zhang and Shuangge Ma*
Yale University
yuan.huang@yale.edu

Gaussian graphical model gains popularity for inferring the relationship between random variables as the precision matrix has a natural interpretation of conditional dependence. When dimension is high, analysis of a single dataset is often unsatisfactory and joint modeling provides an efficient tool to study the common sparsity structure. To this end, sparse group penalization method is widely adopted where the entries at the same place across the matrices are considered as a group and penalty applies on the magnitude. In practice, it may be difficult to interpret the result if a group has different signs across different studies. We propose a joint modeling approach that encourages sparse parameter estimates with sign similarity within a group. Simulation studies show that the proposed method outperforms existing methods in a variety of settings. The theoretical properties are also investigated.

Session 21: Statistical Preprocessing of Deep Sequencing Data

FACETS: Cellular Fraction and Copy Number Estimates from Tumor Sequencing

♦*Venkatraman Seshan and Ronglai Shen*
MSKCC
seshanv@mskcc.org

DNA sequencing of the whole exome or a panel of cancer genes are used to get information on somatic changes to the DNA. The sequencing reads can be used to obtain both the total and allele specific copy number information using both total and allele specific coverage depths. This in turn can be used to estimate the cellular fraction of the tumor as a whole as well as the cellular fractions of all the somatic changes. In this talk I will present the FACETS algorithm which adapts the Circular Binary Segmentation (CBS) algorithm of array based copy number data and allele specific copy number extension in PSCBS.

Normalization issues in single cell RNA sequencing

♦*Zhijin Wu*[1] *and Hao Wu*[2]
[1]Brown University
[2]Emory University
zwu@stat.brown.edu

Single cell RNA-seq (scRNA-seq) is a recently developed technology that enables the transcriptomic profiling at single cell level, providing information for inter-cellular transcriptomic heterogeneity, and adding another layer to the studies of gene expression. This technology has gained tremendous interests since its introduction, and has been applied to profile highly heterogeneous samples such as cancer, brain and immune cells. Similar to data from other high-throughput technologies, scRNA-seq data are affected by substantial technical and biological artifacts, maybe more so due to the low amount of starting materials and more sample preparation. Some unique features of scRNA-seq data include the use of "unique molecule identifier" (UMI) that makes quantification at transcripts without amplification effects possible, and the almost universal use of External RNA Controls Consortium (ERCC) controls. These provide internal controls and reveal issues in scRNA-seq data that otherwise could have been hidden. In addition, they can serve as basis for normalization. We present systematic biases and artifacts identified in early scRNA-seq data and strategies for normalization.

Preprocessing issues with epigenetic assays based on sequencing

Kasper Hansen
Johns Hopkins University
kasperdanielhansen@gmail.com

Various molecular assays can be combined with second generation sequencing to measure epigenetic marks genome wide. However, molecular manipulation prior to sequencing, or choices in the processing pipeline, can introduce artifacts in the generated data, and these artifacts needs to be corrected to achieve valid inference. We will illustrate this with the following examples (1) enrichment issues with single cell ATAC sequencing, (2) identifying and correcting for mapping bias in whole-genome bisulfite sequencing data, and (3) identification of biological signal in input controls in ChIP sequencing.

Exploring Immune Repertoire Sequencing Data

♦*Xiangqin Cui, Amanda Hall and Roslyn Mannon*
University of Alabama at Birmingham
xcui@uab.edu

The human immune system produces extraordinary diverse antigen receptors to recognize and combat antigen stimuli. A new application of next generation sequencing is to survey the immune system by sequencing the mRNA production of the complementarity determining region 3 (CDR3) of antigen receptors in immune cells, such as T cells and B cells. The goal is to obtain all the antigen receptors being expressed in the immune cells to form an immune Repertoire. For data analysis, the DNA or Amino Acids sequences are often aligned to the germline sequences of CDR3 to determine the source of each diversity-contributing element and then define the deletion/addition as well as the somatic mutations in comparison with the germline sequences. The antigen receptor sequences that share the same germline sources are called a clonetype. The diversity of the repertoire is often visualized and quantified. The unique properties of immune Repertoire sequence data are that the function of each clonetype is unknown and there is little sharing of clonetypes across individuals. The challenge is how to associated immune Repertoire data with clinical outcomes to identify biomarkers or functional clonetypes. We have been exploring an immune Repertoire dataset from kidney transplant patients. Our preliminary findings will be discussed.

Session 22: Change-Point Problems and their Applications (I)

Jump Information Criterion for Estimating Jump Regression Curves

Peihua Qiu
University of Florida
pqiu@ufl.edu

Jump information criterion for estimating jump regression curves
Peihua Qiu Department of Biostatistics University of Florida
Nonparametric regression analysis when the regression function is discontinuous has broad applications. Existing methods for estimating a discontinuous regression curve usually assume that the number of jumps in the regression curve is known beforehand, which is unrealistic in certain cases. Although there is some existing research on estimation of a discontinuous regression curve when the number of jumps is unknown, this problem is still mostly open because such a research often requires assumptions on other related quantities such as a known minimum jump size. In this research, we try to solve the problem by proposing a jump information criterion, which consists of a term measuring the fidelity of the estimated regression curve to the observed data and a penalty related to the number of jumps and jump sizes. Then, the number of jumps can be determined by minimizing our criterion. Theoretical and numerical work shows that this method works well in practice. This is a joint research with Dr. Zhiming Xia.

Monitoring Sparse Contingency Table in Multivariate Categorical Process

Dongdong Xiang
East China Normal University
ddxiang@sfs.ecnu.edu.cn

In modern statistic process control (SPC), more and more real applications involve multiple categorical quality characteristics, whose distribution can be displayed by contingency table. Traditional charting schemes are well developed for the tables with "small cell number and large sample size" by using the log-linear models to characterize the relationship each cell count and the levels of the categorical variables. When the number of categorical variables in-

creases, the number of cells in contingency table grows extremely fast so that most of the cell entries are very small or zeros counts. This is so-called sparse contingency table in the literature. In such situations, the traditional methodologies are inadequate to use. This paper is devoted to developing a monitoring method for sparse contingency table. In pahse I, a two-stage group lasso methods are developed to perform models selection and parameter estimation in high-dimension log-linear models. Secondly, the Chi-square testing statistics is adapted to the sparse contingency table. Our numerical results show that the novel charting schemes performs better than the existing control charts. Finally, a real data example is used to demonstrate the effectiveness of the proposed control chart.

Modeling the Next Generation Sequencing Read Count Data for DNA Copy Number Variants Study

♦*Tieming Ji*[1] *and Jie Chen*[2]

[1]University of Missouri at Columbia

[2]Augusta University
jit@missouri.edu

As one of the most recent advanced technologies developed for biomedical research, the next generation sequencing (NGS) technology has opened more opportunities for scientific discovery of genetic information. The NGS technology is particularly useful in elucidating a genome for the analysis of DNA copy number variants (CNVs). The study of CNVs is important as many genetic studies have led to the conclusion that cancer development, genetic disorders, and other diseases are usually relevant to CNVs on the genome. One way to analyze the NGS data for detecting boundaries of CNV regions on a chromosome or a genome is to phrase the problem as a statistical change point detection problem presented in the read count data. We therefore provide a statistical change point model to help detect CNVs using the NGS read count data. We use a Bayesian approach to incorporate possible parameter changes in the underlying distribution of the NGS read count data. Posterior probabilities for the change point inferences are derived. Extensive simulation studies have shown advantages of our proposed methods. The proposed methods are also applied to a publicly available lung cancer cell line NGS dataset, and CNV regions on this cell line are successfully identified.

Changepoint Detection in Categorical Time Series with Application to Hourly Sky-cloudiness Condition

♦*QiQi Lu*[1] *and Xiaolan Wang*[2]

[1]Virginia Commonwealth University

[2]Environment and Climate Change Canada
qlu2@vcu.edu

Changepoints are extremely important features to consider when homogenizing categorical time series and analyzing its trends and variations. The original sky-cloudiness conditions with 11 categories are reported hourly in Canada and hence exhibit seasonality and serial autocorrelation in nature. Annually aggregating a hourly series could help reduce the serial correlation and seasonality, but it will largely shorten the length of the time series. This talk introduces a changepoint detection method for periodic and serially correlated multinomial time series using a marginalized transition model. The serial dependence is described via a first-order Markov chain. The proposed stochastic model allows the likelihood-based inference. An application of this method is illustrated using a read sky-cloudiness data in Canada.

Session 23: Order-restricted Statistical Inference and Applications

Variable and Shape Selection using the Cone Information Criterion

♦*Mary Meyer and Xiyue Liao*
Colorado State University
meyer@stat.colostate.edu

The partial linear generalized additive model is considered, where the goal is to choose a subset of predictor variables and describe the component relationships with the response, in the case where there is very little a priori information. For each predictor, the user need only specify a set of possible shape or order restrictions. For example, the systematic component associated with a continuous predictor might be assumed to be increasing, decreasing, convex, or concave. The effect of a treatment variable might have a tree ordering or be unordered. A model selection method chooses the nature of the relationships as well as the variables. Given a set of predictors and shape or order restrictions, the maximum likelihood estimator for the constrained generalized additive model is found using iteratively re-weighted cone projections. The cone information criterion is used to select the best combination of variables and shapes.

Constrained Statistical Inference in Linear Mixed Models with Applications

♦*Casey Jelsema*[1] *and Shyamal Peddada*[2]

[1]West Virginia University

[2]National Institute of Environmental Health Science
casey.jelsema@mail.wvu.edu

In many applications researchers are interested in testing for inequality constraints in the context of linear fixed and/or mixed effects models. For example, a researcher may wish to test for an increasing response over increasing dose levels. Popular procedures such as ANOVA only test for differences and not trends or patterns in the means. Consequently, not only that they do not answer the underlying scientific question of interest, they could lose power to tests that are designed for testing inequality constraints.

While there exists a large body of literature for performing statistical inference under inequality constraints, user-friendly statistical software for implementing such methods is lacking. In this talk, we discuss constrained inference for linear fixed and/or mixed effects models using residual bootstrap, a general methodology that is reasonably robust to non-normality and can accommodate heteroscedasticity. This bootstrap based methodology, called CLME, is implemented in R using syntax similar to existing linear model packages. The CLME package also contains a graphical interface, enabling a researcher with minimal knowledge of R to easily access it.

We demonstrate the implementation of CLME package using a data obtained from the NIEHS Fibroid Growth Study.

Testing for uniform stochastic orderings via empirical likelihood under right censoring

Hammou ELBARMI
Baruch College, The City University of New York
hammou.elbarmi@baruch.cuny.edu

Empirical likelihood based tests for the presence of uniform stochastic ordering (or hazard rate ordering) among two univariate distributions functions are developed when the data are right censored in the one- and two-sample cases. The proposed test statistics are formed by taking the supremum of some functional of localized empirical likelihood test statistics. The null asymptotic distributions

of these test statistics are distribution-free and have simple representations in terms of a standard Brownian motion. Simulations show that the tests we propose outperform in terms of power the one sided log-rank test at practically all of the distributions that we consider. The stochastic ordering case will be shown to be a special case of our procedure. We illustrate our theoretical results using a real life example.

A test of order-restricted means for data with imputation

♦*Heng Wang and Ping-Shou Zhong*
Michigan State University
hengwang@msu.edu

Missing values appear very often in many applications, but the problem of missing values has not received much attention in testing order-restricted alternatives. Under the missing at random (MAR) assumption, we impute the missing values nonparametrically using kernel regression. For data with imputation, the classical likelihood ratio test designed for testing the order-restricted means is no longer applicable since the likelihood does not exist. This paper proposes a novel method for constructing statistics for testing means with an increasing order or a decreasing order based on jackknife empirical likelihood (JEL) ratio. It is shown that the JEL ratio statistic evaluated under the null hypothesis converges to a chi-bar-square distribution, which is the same as that of the classical likelihood ratio test statistic. Simulation study shows that our proposed test maintains the nominal level well under the null and has prominent power under the alternative. The test is robust for normally and non-normally distributed data. The proposed method is applied to an ADNI study for helping find out a biomarker for the diagnosis of the Alzheimer's disease.

Session 24: Recent Advances of the Statistical Research on Complicated Structure Data

Integrative analysis of datasets with different resolutions reveals consistent genetic effects

Yuan Jiang
Oregon State University
yuan.jiang@stat.oregonstate.edu

As advanced platforms produce data in a higher and higher resolution, a large collection of high-dimensional datasets are made available from different studies with possibly various resolutions. These multi-resolution datasets are incompatible in the sense that a predictor in a low-resolution dataset corresponds to a group of predictors in a high-resolution dataset. This incompatible but nested data structure poses new challenges to the existent statistical methods for data integration. This paper proposes a statistical regularization approach that can integratively analyze multiple high-dimensional datasets with different resolutions. This approach not only enables joint estimation of model parameters but also ensures consistent findings from multiple studies. Simulation studies illustrate the advantage of the proposed joint analysis in terms of its consistent findings and its enhanced statistical power compared to separate analyses. Meanwhile, an integrative analysis of multi-resolution genetic datasets shows the applicability of the proposed method to genetic association studies.

False discovery rate estimation with covariates

Kun Liang
University of Waterloo
liangkun1@gmail.com

Multiple testing becomes an increasingly important topic in high-dimensional statistical analysis. However, most commonly used false discovery rate estimation and control methods do not take covariates into consideration. To better estimate false discovery rate, we propose a novel nonparametric method which efficiently utilizes the covariate information. Our proposed method enjoys some desirable theoretical properties. In addition, we evaluate the performance of our proposed method over existing methods using simulation studies.

Using a monotonic density ratio model to find the asymptotically optimal combination of multiple dia

♦*Baojiang Chen*[1]*, Pengfei Li*[2]*, Jing Qin*[3] *and Tao Yu*[4]
[1]University of Nebraska Medical Center
[2]University of Waterloo
[3]NIH
[4]National University of Singapore
baojiang.chen@unmc.edu

With the advent of new technology, new biomarker studies have become essential in cancer research. To achieve optimal sensitivity and specificity, one needs to combine different diagnostic tests. The celebrated Neyman–Pearson lemma enables us to use the density ratio to optimally combine different diagnostic tests. In this paper, we propose a semiparametric model by directly modeling the density ratio between the diseased and nondiseased population as an unspecified monotonic nondecreasing function of a linear or nonlinear combination of multiple diagnostic tests. This method is appealing in that it is not necessary to assume separate models for the diseased and nondiseased populations. Further, the proposed method provides an asymptotically optimal way to combine multiple test results. We use a pool-adjacent-violation-algorithm to find the semiparametric maximum likelihood estimate of the receiver operating characteristic (ROC) curve. Using modern empirical process theory we show cubic root n consistency for the ROC curve and the underlying Euclidean parameter estimation. Extensive simulations show that the proposed method outperforms its competitors. We apply the method to two real-data applications.

Variable selection in the presence of nonignorable missing data

Jiwei Zhao
State University of New York at Buffalo
zhaoj@buffalo.edu

Variable selection methods are well developed for a completely observed data set in the past two decades. In the presence of missing values, those methods need to be tailored to different missing data mechanisms. In this paper, we focus on a flexible and generally applicable missing data mechanism, which contains both ignorable and nonignorable missing data mechanism assumptions. We show how the regularization approach for variable selection can be adapted to the situation under this missing data mechanism. The computational and theoretical properties for variable selection consistency are established. The proposed method is further illustrated by comprehensive simulation studies, for both low and high dimensional settings.

Session 25: Innovative Methods for Modeling and Inference with Survival Data

Statistical Inference on length-biased data with semiparametric accelerated failure time models

♦*Jing Ning*[1]*, Jing Qin*[2] *and Yu Shen*[1]
[1]The University of Texas MD Anderson Cancer

[2]National Institutes of Health
jning@mdanderson.org

Outcome dependent sampling bias arises when the observations are not randomly selected from the target population. Length-biased sampling is a special case of left-truncated data, and has been recognized in various applications. The analysis of length-biased data is complicated by informative right censoring due to the biased sampling mechanism, and consequently the techniques for conventional survival analysis are not applicable. In this talk, I will present our recent work to evaluate covariate effects on the failure times of the target population under the accelerated failure time model given the observed length-biased data.

Association analysis of gap times with multiple causes

Xiaotian Chen, ♦*Yu Cheng, Ellen Frank and David Kupfer*
University of Pittsburgh
yucheng@pitt.edu

We aim to close a methodological gap in analyzing durations of successive events that are subject to induced dependent censoring as well as competing-risk censoring. In the Bipolar Disorder Center for Pennsylvanians (BDCP) study, some patients who managed to recover from their symptomatic entry later developed a new depressive or manic episode. It is of great clinical interest to quantify the association between time to recovery and time to recurrence in patients with bipolar disorder. The estimation of the bivariate distribution of the gap times with independent censoring has been well studied. However, the existing methods cannot be applied to failure times that are censored by competing causes such as in the BDCP study. Bivariate cumulative incidence function (CIF) has been used to describe the joint distribution of parallel event times that involve multiple causes. To the best of our knowledge, however, there is no method available for successive events with competing-risk censoring. Therefore, we extend the bivariate CIF to successive events data, and propose nonparametric estimators of the bivariate CIF and the related conditional CIF. Moreover, an odds ratio measure is proposed to describe the cause-specific dependence, leading to the development of a formal test for independence of successive events. Simulation studies demonstrate that the estimators and tests perform well for realistic sample sizes, and our methods can be readily applied to the BDCP study.

Restoration of Monotonicity Respecting in Dynamic Regression

Yijian Huang
Emory University
yhuang5@emory.edu

Dynamic regression models, including the quantile regression model and Aalen's additive hazards model, are widely adopted to investigate evolving covariate effects. Yet lack of monotonicity respecting with standard estimation procedures remains an outstanding issue. Advances have recently been made, but none provides a complete resolution. In this talk, we propose a novel adaptive interpolation method to restore monotonicity respecting, by successively identifying and then interpolating nearest monotonicity-respecting points of an original estimator. Under mild regularity conditions, the resulting regression coefficient estimator is shown to be asymptotically equivalent to the original. Our numerical studies have demonstrated that the proposed estimator is much more smooth and may have better finite-sample efficiency than the original as well as, when available as only in special cases, other competing monotonicity-respecting estimators. Illustration with a clinical study is provided.

A General Semiparametric Accelerated Failure Time Model Imputation Approach for Censored Covariate

♦*Ying Ding*[1]*, Shengchun Kong*[2] *and Shan Kang*[3]
[1]University of Pittsburgh
[2]Purdue University
[3]Robert Bosch LLC, Research and Technology Center
yingding@pitt.edu

We consider a general framework to handle regression with censored covariate where the response can be any type including survival data. Multiple imputation (MI) is a popular technique to handle missing data which requires assumptions of compatibility between the imputation and substantive model to obtain valid estimates. With censored covariate, we propose a MI approach, namely, the semiparametric two-step importance sampling imputation (STISI) method, to tackle the problem where the censored covariate is imputed through a semiparametric accelerated failure time (AFT) model. This STISI method imputes the missing covariate from a semiparametric AFT model conditional on fully observed covariates with acceptance probability derived from the substantive model. This two-step procedure automatically ensures compatibility between the regression model and imputation model and takes the full advantage of semiparametric assumption in the imputation. Extensive simulations demonstrate that the STISI method yields valid estimates in all scenarios and outperforms some existing methods that are commonly used in practice. The method is illustrated by analyzing the urine arsenic data for patients from National Health and Nutrition Examination Survey (NHANES) (2003-2004).

Session 26: Nonparametric Methods for Neural Spike Train Data

Calibrating nonparametric inference of monosynaptic connections from spike train recordings

♦*Asohan Amarasingham and Jonathan Platkiewicz*
City University of New York
aamarasingham@ccny.cuny.edu

There has been recently a great deal of interest in "mapping the brain", namely in establishing the precise structural organization of neural microcircuits. It is thought that such a map will allow us to bridge the gap between single neuron physiology and animal cognitive behavior. High-density extracellular recordings observe simultaneously the activity of hundreds of neurons with millisecond precision in the behaving mammal. Neural connectivity is typically inferred from this recording type by extracting the spikes from the extracellular potentials and estimating pairwise correlations between spike trains. Ultimately, such methods still rely on the assumption that a correlation between two spike trains reflects a direct synaptic coupling. There is, however, no widely-accepted biophysical justification for this assumption, which has been criticized based on biophysical modeling work (Ostojic et al.) and is potentially critiqued by other lines of evidence as well. Lacking direct experimental access, numerical simulations of biophysical models of monosynaptic spike transfer, which are biologically faithful to some degree, are natural candidates to investigate this debate.

I will discuss our studies of the dynamical monosynapse model considered by Ostojic et al.. We show that a millisecond correlation can be exhibited with this model under realistic conditions, rendering the consideration of background network modulation irrelevant. Using this model, we validate the relevance of applying nonparametric statistical methods for detecting fine timescale interactions, based on conditional inference, for monosynapse detection from in

vivo spike data. We attempt to clarify the relationship between nonparametric statistical approaches to characterizing neural data and standard dynamical models of neurons, in this context.

Nonparametric Methods for Decoding Rat Hippocampal Neuronal Ensemble Spikes

Zhe Chen
New York University School of Medicine
zhe.chen3@nyumc.org

The parametric statistical methods are self-contained and elegant, and their practical implementation is simple and analytically tractable. However, their limitations lie in imposing prior assumptions regarding the underlying data. In contrast, nonparametric statistical methods are "distribution-free" and let data speak for themselves. Specifically, nonparametric methods, either Bayesian or non-Bayesian, avoid the subjectivity of choosing a specific parametric model or model parameters. In recent years, many nonparametric methods have been proposed or developed for neural spike train analysis. Examples of such include spike sorting, characterizing neuronal receptive fields ("neural encoding"), and reconstructing sensory or motor stimuli based on ensemble spike activity ("neural decoding").

In this talk, we will present some overview of our previous and recent work on applying nonparametric methods for decoding rat hippocampal neuronal ensemble spikes. In the first part, we present a statistical nonparametric framework for decoding unsorted hippocampal ensemble spikes. The new framework allows us to sidestep the time-consuming and error-prone spike sorting process and directly reconstruct animal's spatial position [1,2]. In the second part, we propose a Bayesian nonparametric method to uncover neural representations of sorted hippocampal ensemble spike activity during spatial navigation or during sleep [3,4]. Specifically, we propose a hierarchical Dirichlet process-hidden Markov model (HDP-HMM) and Markov chain Monte Carlo (MCMC) inference to tackle the estimation problem. It is worthy pointing out that these two methods enable us to decipher rat hippocampal population codes without explicit measure or estimate of neuronal receptive fields in the conventional sense. Together, these new nonparametric statistical methods provide powerful toolkits for neural spike train analysis.

References [1] Chen Z, Kloosterman F, Layton S, Wilson MA (2012). Transductive neural decoding for unsorted neuronal spikes of rat hippocampus. Proc. IEEE EMBC, pp. 1310-1313. [2] Kloosterman F, Layton S, Chen Z, Wilson MA (2014). Bayesian decoding using unsorted spikes in the rat hippocampus. J. Neurophysiol. 111(1):217-227. [3] Linderman S, Johnson MJ, Wilson MA, Chen Z (2016). A nonparametric Bayesian approach for uncovering rat hippocampal population codes during spatial navigation. J. Neurosci. Methods, 236:36-47. [4] Chen Z, Grosmark A, Penagos H, Wilson MA (2016). Uncovering representations of sleep-associated hippocampal ensemble spike activity. Scientific Reports, under revision.

Receptive field models of multiunit activity and the decoding of hippocampal replay events

Uri Eden
Boston University
tzvi@bu.edu

Traditionally, experiments designed to study the role of specific spike patterns in learning and memory tasks take one of two forms, 1) observational studies that characterize statistical properties of neural activity during such tasks or 2) interventional studies that broadly alter neural activities over an entire neural population or brain region. This work is part of a larger project to allow investigators to manipulate neural populations in a content-specific way, altering spiking activity related to certain learning and memory patterns while leaving activity related to other patterns intact.

One fundamental challenge of this work is to decode the information content of specific spike sequences in real-time. Previously, we have used point process theory to develop efficient decoding algorithms based on spike train observations. However these algorithms assume the spike trains have been accurately sorted ahead of time, which is not possible for real-time decoding. Here we present a new point process decoding algorithm that does not require multiunit signals to be sorted. We use the theory of marked point processes to characterize the relationship between the coding properties of multiunit activity and features of the spike waveforms . We develop a marked point process filter to compute the posterior distribution of a signal to decode given multiunit activity from a neural population. We first characterize the spiking activity of a neural population using the conditional intensity function for marked point processes. We then construct point process filters to iteratively calculate the full posterior density of a signal. We illustrate our approach with a simulation study as well as with experimental data recorded in the hippocampus of a rat performing a spatial memory task. We construct multiunit encoding models using both a kernel density estimator and a mixture of Gaussian model fit to the observed movement and waveform data. Our decoding framework is then used to reconstruct first the animal's position, and then a cognitive state related to memory replay, from unsorted multiunit spiking activity. We then compare the quality of fit of our decoding framework to that of a traditional spike-sorting and decoding framework. Our analyses show that the proposed decoding algorithm performs as well as or better than algorithms based on sorted single-unit activity. These results provide a mechanism for content-specific manipulations of neural population activity.

Nonparametric discriminative filtering for neural decoding

♦*Michael C. Burkhart, David M. Brandman, Carlos Vargas-Irwin and Matthew T. Harrison*
Brown University
michael_burkhart@brown.edu

Bayesian filtering algorithms, such as the Kalman filter, are routinely used to predict time-varying stimuli or behaviors from recorded brain signals. This work introduces a novel form of recursive Bayesian filtering that facilitates the use of probabilistic discriminative models in conjunction with filtering. It seems particularly useful in situations where the predicted variables are low dimensional and observable during training, but the observed brain states are high dimensional with possibly complex dynamics. The motivating application is neural prosthetics, where the hidden states are a small number of behavioral variables and the observables are a large number of brain signals.

Session 27: Subgroup Identification/Analysis in Clinical Trials

Assessing consistency of treatment effect across subgroups in a large randomized study on dual anti-platelet therapy

Joseph Massaro
Boston University/Harvard Clinical Research Inst.
jmm@bu.edu

A recent study assessed the effect of long-term dual-antiplatelet therapy (DAPT) on incident cardiovascular events in patients receiving drug-eluting stents. Over 9000 patients were given open-label DAPT for 12 months, after which they were randomized to one of two treatment groups: an additional 18 months of DAPT or 18 months of Placebo. Primary endpoints were time-to-stent thrombosis and time-to-a composite endpoint of major adverse cardiovascular and cerebrovascular events (MACCE; composite of death, stroke, and myocardial infarction). Subjects were enrolled into the trial by HCRI and from four post-marketing stent surveillance studies sponsored by stent manufacturers. Patients were enrolled from up to three regions (North America, Europe, and Australia/New Zealand). Thus, the assessments of consistency of randomized treatment difference across studies and across regions were of interest. We discuss our statistical approach to these assessments, and how/why we adjusted for differences in baseline characteristics across studies and regions through the use of propensity scores. We discuss interpretations and conclusions we made from our assessments of consistency.

Issues in Subgroup Analyses

♦*Weishi Yuan, Kun He and Rajeshwari Sridhara*
FDA
vivian.yuan@fda.hhs.gov

When analyzing clinical trial data, there are always various issues in subgroup analyses, including trial design, analysis methodology and multiple analyses. In many cases there is no consensus on how to evaluate and interpret these subgroups. In this talk we will present some examples to illustrate these issues.

Subgroup Analysis: Issues and Possible Improvement

♦*Lu Cui and Tu Xu*
AbbVie
lu.cui@abbvie.com

Confirmatory randomized clinical trial is designed to provide definitive information on the efficacy and safety of a new drug. While the outcome based on the overall population is the basis for the approval, sound subgroup outcome based on unbiased parameter estimate may provide additional evidence and insight of drug efficacy. This presentation is to revisit issues related to subgroup analysis. Use of extensive randomization stratification to minimize potential bias in subgroup analysis is investigated via statistical simulations, and the results are presented to illustrate the idea.

Session 28: Model Selection/Averaging and Objective Assessment.

On model selection from a finite family of possibly misspecified models

Ching-Kang Ing
Institute of Statistical Science, Academia Sinica
cking@stat.sinica.edu.tw

Model selection problems are usually classified into two categories according to whether the data generating process (DGP) is included among the family of candidate models. The first category assumes that the DGP belongs to the candidate family, and the objective of model selection is simply to choose this DGP with probability as high as possible. The second category assumes that the DGP is not one of the candidate models. In this case, one of the top concerns is to choose the model having the best prediction capability. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a key point of contention. In this article, we propose a misspecification-resistant information criterion (MRIC) to rectify this difficulty under the fixed-dimensional framework, which requires that the set of candidate models is fixed with the sample size. We prove the asymptotic efficiency of MRIC regardless of whether the true model belongs to the candidate family or not. We also illustrate MRIC's finite-sample performance using Monte Carlo simulation.

Model selection confidence sets by likelihood ratio testing

Davide Ferrari
University of Melbourne
dferrari@unimelb.edu.au

The traditional activity of model selection aims at discovering a single model superior to other candidate models. In the presence of pronounced noise, however, multiple models are often found to explain the same data equally well. To resolve this model selection ambiguity, we introduce model selection confidence sets (MSCSs) in the context of maximum likelihood estimation. A MSCS is defined by a list of models statistically equivalent to the true model at a user-specified level of confidence, thus extending the familiar notion of confidence intervals to the model-selection framework. We propose to construct MSCSs using the likelihood ratio test; our approach guarantees correct coverage probability of the true model when both sample size and model dimension increase. We derive conditions under which the MSCS contains all the relevant information about the true model structure. In addition, we propose natural statistics to measure importance of parameters in a principled way that accounts for the overall model uncertainty. When the overall space of feasible models is large, MSCSs is implemented by an adaptive stochastic search algorithm which samples MSCS models with high probability.

Kernel Estimation and Model Combination in a Bandit Problem with Covariates

♦*Wei Qian*[1] *and Yuhong Yang*[2]
[1]Rochester Institute of Technology
[2]University of Minnesota
wxqsma@rit.edu

Multi-armed bandit problem is an important optimization game that requires an exploration-exploitation tradeoff to achieve optimal total reward. Motivated from industrial applications such as online advertising and clinical research, we consider a setting where the rewards of bandit machines are associated with covariates, and the accurate estimation of the corresponding mean reward functions plays an important role in the performance of allocation rules. Under a flexible problem setup, we establish asymptotic strong consistency and perform a finite-time regret analysis for a sequential randomized allocation strategy based on kernel estimation. In addition, since many nonparametric and parametric methods in supervised learning may be applied to estimating the mean reward functions but guidance on how to choose among them is generally unavailable, we propose a model combining allocation strategy for adaptive performance. Simulations and a real data evaluation are conducted to illustrate the performance of the proposed allocation strategy.

Cross Assessment Towards a More Reproducible Model Selection

Yuhong Yang
University of Minnesota
yangx374@umn.edu

Reproducibility of research findings has become a central issue in

science and beyond. In data analysis, model selection uncertainty, especially in high-dimensional settings, is a major source that may contribute much to irreproducibility of statistical conclusions. In this talk, we advocate the approach of cross validation/assessment to come up with a proper weighting on the most plausible models. The weights can be used for rate-optimal model combining (model mis-specification permitted), for quantifying variable selection uncertainty, and for measuring variable importance. Theoretical results justify the proposed approach. Numerical examples provide strong support and also insight on the key matters.

Session 29: Advances and Applications in Methods for Comparative Effectiveness Research

An overview of methods for controlling unmeasured confounders in comparative observational research

♦*Xiang Zhang and Douglas Faries*
Eli Lilly and Company
zhang_xiang@lilly.com

Propensity score based methods (regression, weighting, matching, and stratification) provide useful statistical tools to assess the casual effect in comparative effectiveness research using real world /observational/ big data. However, casual inference based on these methods relies on the unprovable assumption of "no unmeasured confounding" To date, most research simply notes the violation of this assumption as a limitation of the research and no quantitative assessment of the potential impact of unmeasured confounding is performed. Over the past decades, many quantitative approaches to assessing the impact of unmeasured confounding have arisen, including pseudo randomization, negative controls, sensitivity analysis that quantifies the robustness of the analysis to varying levels of unmeasured confounding, and newer approaches that incorporate information external to the study and produce an adjusted estimate of the effect. To ensure appropriate use of information arising from such comparative observational research, analyses should include a thorough and quantitative assessment of the potential impact of unmeasured confounders. However, the many options provided by recently developed methods and the variety of research scenarios makes this a challenge to understand the optimal course of action.

The two main goals of this talk are to provide an overview of statistical methods for addressing unmeasured confounding and also introduce a best practice guidance. The best practice guidance will include a flowchart / decision tree approach to recommending analysis options given the study scenario and availability of information on the unmeasured confounders.

Generalized propensity score matching with multiple treatments: An application

♦*Zhanglin Cui, Lisa Hess, Robert Goodloe, Gebra Carter and Douglas Faries*
Eli Lilly and Company
cui_zhanglin@lilly.com

Objective: Conventional pairwise propensity score matching (PSM) may produce inconsistent results when applied to multiple treatments due to the lack of a robust region of common support from the samples of matched patients across comparisons. This research introduces generalized propensity score matching and contrasts it with conventional pairwise PSM in terms of the implementation steps, assessment of overlap of populations, and implications on inference and generalizability. This is demonstrated with a case study assessing the comparative effectiveness of common second-line treatments for lung cancer in the US.

Methods: Electronic medical record (EMR) data for lung cancer patients in the IMS Oncology database were used for this study. Patients diagnosed with lung cancer (ICD-9-CM 162.2-162.9) from 1/1/2007-6/30/2013 and who received at least two lines of treatment were included. Binomial and multinomial logistic regressions were used to estimate propensity scores in a conventional pairwise and a generalized fashion, respectively. For generalized PSM, a region of common support with sufficient overlap in the covariate distribution and minimum variance of the covariate space was identified. Generalized PSM with replacement was then conducted to construct counterfactual outcomes under each treatment for each patient. Absolute standardized differences (ASD) in covariates were used to assess balance among the treatments. A Cox proportional hazards model was used for survival analysis after conducting the generalized PSM, and the results were compared to conventional pairwise PSM. Bootstrapping was performed as a sensitivity analysis.

Results: The 5 most common lung cancer treatments were identified with a total sample size of 5,222 patients. Generalized PSM used 61.2% of the patient sample while the conventional pairwise PSM used 24.1-77.1% across the 10 comparisons. Perfect balance (ASD=0) in the generalized PSM and acceptable balance (ASD¡0.1) in the conventional pairwise PSM were achieved among the treatments on each covariate. While the range for median overall survival was similar for generalized PSM (5.6-8.9 months) and conventional pairwise PSM (5.6-9.5 months) among the top 5 treatments, the generalized PSM achieved statistical significance (p¡0.05) in 8 out of the 10 survival comparisons whereas conventional pairwise PSM achieved it in only 1. Similar results were obtained from bootstrapping. The noted differences arose from different matched patient samples and the size of the samples. Practical issues after assessing the overlap across multiple treatments will be discussed.

Conclusions: The generalized PSM allows for comparisons across multiple treatments using a region of common support while removing bias from observed covariates under the "no unmeasured confounding" assumption and may have potential applications in observational studies with multiple treatments.

Considerations of the Appropriate Number of Strata in Stratified Randomized Clinical Trials

♦*Bo Fu, Su Chen, Yao Li, Ziqian Geng, Yijie Zhou, Lu Cui and Lois Larsen*

AbbVie

bo.fu.stat@gmail.com

Stratified randomization is commonly used in clinical trials to balance treatment groups with respect to important factors that may impact trial outcome variables and consequently to reduce potential bias and confounding when assessing treatment differences. On the other hand, over-stratification is typically not recommended because too many strata can lead to empty cells, resulting in imbalanced size of different treatment groups and impacting statistical analysis. In this work we investigate what is the appropriate max number of strata for a given trial size when balancing the considerations of reduced bias and confounding versus the impact on other features of statistical inference such as variability, coverage rate, type I error and power. The investigation is done for both the overall trial analysis as well as subgroup analysis frameworks.

Session 30: Statistical Genetics

Integrating competing but complementary genetic association tests derived from the same data

Lei Sun
University of Toronto
sun@utstat.toronto.edu

In many scientific studies, different statistical tests are being proposed with competing claims about the performance in terms of power. The power of a given test depends on the nature of the alternatives. For example, in the phenotype-genotype association analyses of complex human traits, the class of location-tests is commonly used to detect phenotypic mean differences between genotype groups. However, complex genetic etiologies including GxG and GxE interactions can result in homoscedasticity, where the class of scale-tests would be more powerful. In another example where association between multiple rare genetic variants and a trait is of interest, two classes of tests have been proposed (Derkach et al. 2014, Statistical Science). The class of linear-tests (also known as burden-tests) is preferred if the majority of the variants under consideration are truly associated and have the same direction of effect, while the class of quadratic-tests (variance component-tests) are more powerful in other settings. To achieve robustness it is natural to combine the evidence for association from the two (or more) complementary tests, but how? The well-known Fisher's method, commonly used in meta-analyses, combines p-values of the same test applied to K independent samples. Here we propose to use it to combine p-values of different tests applied to the same sample. We first show that, in both settings, the two classes of tests are asymptotically independent of each other under the null hypothesis of no association, thus the resulting Fisher's test statistic has the asymptotic chi-squared distribution with 4 d.f. We then demonstrate via extensive simulations and applications (Type 1 Diabetes and Cystic Fibrosis complications) that the resulting class of joint test statistics is not only robust but could have better power than the individual tests (Derkach et al. 2013, Genetic Epidemiology; Soave et al. 2015, The American Journal of Human Genetics). We also compare Fisher's method to alternative ways of aggregating information such the minimal p-value approach, and discuss various caveats in practical implementation. This is joint work with graduate students Andriy Derkach and David Soave, and Professor Jerry Lawless, Dr. Andrew Paterson and Dr. Lisa Strug.

Detecting Schizophrenia Genes via a Two-Sample Test for High-Dimensional Covariance Matrices

♦*Lingxue Zhu*[1]*, Jing Lei*[1]*, Bernie Devlin*[2] *and Kathryn Roeder*[1]
[1]Carnegie Mellon University
[2]University of Pittsburgh
lzhu@cmu.edu

Scientists routinely compare gene expression levels in cases versus controls to determine genes associated with a disease. Similarly, detecting differences in co-expression among genes can be critical to understanding complex human disorders; however statistical methods have been severely limited due to the high dimensional nature of this problem. Here, we propose a sparse-Leading-Eigenvalue-Driven (sLED) test for high-dimensional two-sample covariance matrices. sLED provides a novel perspective that focuses on the spectrum of the differential matrix, and accommodates the sparse and weak signals in many gene expression data. We show that sLED achieves full power asymptotically under mild assumptions, and simulation studies verify that it outperforms other existing testing procedures in many biologically plausible scenarios. Applying sLED to the largest gene-expression dataset comparing Schizophrenia and control brains, we provide a novel list of potential risk genes and reveal intriguing patterns of gene interaction change in Schizophrenia subjects.

Kernel machine association testing for longitudinally-measured quantitative phenotypes

♦*Zuoheng Wang*[1]*, Zhong Wang*[2]*, Xinyu Zhang*[1] *and Ke Xu*[1]
[1]Yale University
[2]Cornell University
zuoheng.wang@yale.edu

Recent developments in high-throughput sequencing technologies have made it possible to search for both rare and common genetic variants associated with complex diseases. Many phenotypes in health studies are measured at multiple time points. The rich information on repeated measurements on each subject not only provides a more accurate assessment of disease condition, but also allows us to explore the genetic influence on disease onset and progression. However, most association tests mainly focus on a single time point. To address this limitation, we propose LSKAT (Longitudinal Sequence Kernel Association Test), a region-based variants association test for longitudinal data, which extends the SKAT method for a single measurement to repeated measurements. LSKAT uses several variance components to account for the within-subject correlation structure of the longitudinal data, and the contributions from all genetic variants (common and rare) in a region. Additionally, we propose another test LMSKAT (Longitudinal Multi-Kernel Association Test) which allows for the time-varing genetic effects by using multiple kernels to detect genes affecting the temporal trends of the trait. In simulation studies, we evaluate the performance of LSKAT and LMSKAT, and demonstrate that they have improved power, by making full use of multiple measurements, as comparing to previously proposed tests on a single measurement or average measurements for each subject. We apply LSKAT and LMSKAT to testing with body mass index in Framingham heart study.

Adaptive genetic association analysis of multiple traits adjusting for population structure

Duo Jiang

Oregon State University
jiangd@stat.oregonstate.edu

In genetic association mapping, the joint testing of multiple traits can boost power by aggregating the association signals across the traits and can reveal valuable insights into the pleiotropy of complex diseases. Recognizing the limitations of existing multiple-trait association tests, we propose a new method based on a mixed effects quasi-likelihood framework. It is applicable to an arbitrarily large number of phenotype variables, which are allowed to be either continuous, discrete or a combination of both. It effectively accounts for population structure and/or family relationships in the sample. We use a variance component test to achieve potential power gain over fixed effects multivariate tests, which tend to lose power as the number of tested traits grows. Moreover, our approach not only allows heterogeneity among the traits in terms of their association effects with the tested genetic variant, but is also able to adaptively adjust to the unknown configuration of these effects to obtain robust power. We evaluate our method and compare its performance with existing methods in simulation studies.

Session 31: Recent Advances in Statistical Methods for Challenging Problems in Neuroimaging Applications

Statistical method for neuron network recovery

Chunming Zhang
University of Wisconsin-Madison
cmzhang@stat.wisc.edu

In neurophysiological study, neural signals provide hidden information of interaction among neurons. The activity of a neuron's spike firing might effect the chance of another neuron to fire in a certain period of time. To study the functional connectivity, we apply some proper regularization method combined with some loss term which adapts to unspecified distributions. Simulation results show the effectiveness of the proposed method in detecting connectivities while cleaning out insignificant interactions at the same time. We apply this method to a real neurophysiological data set collected from part of the prefrontal cortex of rats during a designed experiment. The results provide some insight into the interaction network in that region.

Community Detection and Clustering via G-models with an Application to fMRI

Florentina Bunea[1]*, Christophe Giraud*[2] *and* ♦*Xi Luo*[3]
[1]Cornell University
[2]Universite Paris Sud
[3]Brown University
xi.rossi.luo@gmail.com

Functional MRI yields big and complex data. Clustering (or region of interest specification) is usually the first approach employed to reduce the dimensionality and to enhance interpretability. Classical clustering techniques, such as hierarchical clustering and k-means, are usually based on the closeness between two data points or between a data point to a cluster centroid. Such a definition of closeness may face challenges in clustering network or correlated data. In this talk, we examine clustering using a covariance matrix approach. We introduce a new class of models, the G-Models, for partitioning variables into communities with exchangeable behavior, defined on the whole covariance matrix. These models are motivated by three different but inter-related concepts: the exchangeability of variables, block covariance structures, and latent variable covariance matrices. These concepts will lead to the same clustering partitions under certain regularity conditions, and these parturitions are computed via our fast method and algorithm. Theoretical analysis shows that our method recovers the true partition exactly with high probability and is also minimax optimal. Numerical merits will be demonstrated using simulated data and a publicly available fMRI dataset.

Parsimonious tensor response regression with applications to neuroimaging analysis

Xin Zhang
Florida State University
henry@stat.fsu.edu

Aiming at abundant scientific and engineering data with not only high dimensionality but also complex structure, we study the regression problem with a multi-dimensional array (tensor) response and a vector predictor. Applications include, among others, comparing tensor images across groups after adjusting for additional covariates, which is of central interest in neuroimaging analysis. We propose parsimonious tensor response regression adopting a generalized sparsity principle. It models all voxels of the tensor response jointly, while accounting for the inherent structural information among the voxels. It effectively reduces the number of free parameters, leading to feasible computation and improved interpretation. We achieve model estimation through a nascent technique called the envelope method, which identifies the immaterial information and focuses the estimation based upon the material information in the tensor response. We demonstrate that the resulting estimator is asymptotically efficient, and it enjoys a competitive finite sample performance. We also illustrate the new method on real neuroimaging studies.
(Joint work with Dr. Lexin Li)

A Distributional Independent Component Analysis approach for fMRI data

♦*Subhadip Pal*[1]*, Ying Guo*[1] *and Jian Kang*[2]
[1]Emory University
[2]University of Michigan
subhadippal@gmail.com

Research advances in neuroimaging, genomics and pathophysiology hold great promises to revolutionize diagnosis and treatment for various disease including the mental health diseases. There is an emerging interest in conducting studies that focus on investigations of mechanisms of mental disorders. These investigations have led to an expanded depth of multimodal brain imaging data, genomic data and clinical assessments. This wealth of diverse datasets provides an unprecedented opportunity for crosscutting investigations that may offer new insights to understand mechanisms underlying diseases.
ICA is one of the most widely used source separation techniques for identifying hidden factors that underlie sets of random variables and measurements. This rapidly evolving technique has found successful applications in a wide range of scientific especially biomedical signaling (fMRI, optical imaging) and visual receptive fields. ICA has been shown to be a highly effective tool for dimension reduction as well as denoising and identification of latent source signals. In recent years, ICA has been widely applied in neuroimaging studies for investigating brain functional networks. Although widely applied in many scientific areas current ICA method have several major limitations. First, ICA was developed for continuous data and work on ICA for discrete data is restricted only for binary data. Also the standard ICA method lack the applicability for joint modelling of multimodal data. In this project we have developed a novel Distributional Independent Component Analysis (D-ICA) method that can overcome the aforementioned limitations of standard ICA by providing a unified framework for extracting source signals from diverse types of data. The new concept of D-ICA represent a paradigm shift to the traditional ICA method by performing ICA on the distributional level rather than on the observed data. We construct a Bayesian framework to analyze data and develop MCMC technique for the parameter estimation. We demonstrate the advantages of our methods over the existing method via simulation studies and we also successfully apply our method to fMRI real datasets.

Session 32: Advances in Pharmacogenomics and Biomarker Development

A Few statistical issues in determining sample's positivity using Flow Cytometry and ELISpot assay

Shuguang Huang
Stat4ward LLC
shu444@gmail.com

Flow cytometry and ELISpot assays serve as valuable tools for various aspects of the immunotherapy drug development process ranging from target discovery and characterization to evaluation of responses in a clinical setting. The assays allow for the identification of populations and subpopulations of cells by the expression of one or more antigens; results are then expressed as the percentage of the parent population expressing the antigen of interest. Most commonly, data are reported as percent positive cells which is established by setting a gate around a given cell population versus another.

Quite often a decision needs to be made on whether the observed difference between two samples (e.g. unstimulated vs stimulated) is "significant" which may be related to statistical significance and biological significance. For statistical significance, methods such as Fisher exact test, t-test, bootstrapping are often used; for biological significance, quite often some empirical rules are applied (e.g. 2-fold difference). Here in the talk I'd like to discussion another layer: the analytical significance, which is close-related to the assay's limit of detection and limit of quantification.

Quantitative Reproducibility Analysis for Identifying Reproducible Targets from High-Throughput Expe

♦*Wenfei Zhang*[1]*, Ying Liu*[2] *and Yuefeng Lu*[1]
[1]Sanofi
[2] Sanofi
wenfei.zhang@sanofi.com

High-throughput assays are widely used in biological research to select potential targets. One single high-throughput experiment can efficiently study a large number of candidates simultaneously, but is subject to substantial variability. Therefore it is scientifically important to performance quantitative reproducibility analysis to identify reproducible targets with consistent and significant signals across replicate experiments. A few methods exist, but all have limitations. We propose a new method for identifying reproducible targets. Considering a Bayesian hierarchical model, we show that the test statistics from replicate experiments follow a mixture of multivariate Gaussian distributions, with the one component with zero-mean representing the irreproducible targets. A target is thus classified as reproducible or irreproducible based on its posterior probability belonging to the reproducible components. We study the performance of our proposed method using simulations and a real data example. The proposed method is shown to have favorable performance in identifying reproducible targets compared to other methods.

Region based approach with guided weights for Illumina 450K BeadChip methylation analysis

♦*Yushi Liu, Chunlao Tang and James Scherschel*
Eli Lilly and Company
liu_yushi@lilly.com

DNA methylation is a well-known epigenetic phenomenon for understanding disease progression. The process could be viewed as the conversion from cytosine to 5-methylcytosine. It could either be repressive and permissive in terms of the expression regulation. Such process will change significantly during the disease progress. Currently, Illumina HumanMethylation450Chip methylation array provides a unique opportunity to detect this phenomenon genome widely. The methylation data analysis could provide insight into the disease mechanism and is important for future drug target identification. However, people are focusing on the probe level analysis traditionally. Such results may be complicated especially for the brain diseases. Here, we provide a way of meta-analysis to summarize the probe level results to gene level based on latent variable analysis. As an example, we used schizophrenia HumanMethylation450Chip data from one brain study. Based on the pathway and disease analysis, we have shown that the top ranked genes are enriched for schizophrenia/bipolar related genes.

Advances and issues in RNAseq expression data analysis

♦*Hui-Rong Qian, Phil Ebert and John Calley*
Eli Lilly and Company
qianhu@lilly.com

Measurement of whole-genome gene expression has been desired in biological research and was made practical decades ago by various microarray technology. In recent years next-generation sequencing (NGS) has gained more and more attention due to its flexibility of application and unbiased nature. However, large quantity of data and relatively less understanding of newer technology pose great challenge on experimental design, data processing, and data analysis. RNAseq uses NGS technology to quantify genome-wise gene expression. In this topic, we will focus on gene expression data from RNAseq and discuss several key issues we encounter in RNAseq experimental design and data analysis. We will also compare results from Affymetrix microarray and Illumina HiSeq RNAseq data. We hope that this discussion and our results promote good practice in design, data analysis, and interpretation of NGS studies.

Session 33: Model Assessment for Complex Dependence Structure

Empirical likelihood tests for alternative spatial dependence structures in Markov random fields

♦*Mark Kaiser*[1]*, Daniel Nordman*[2] *and Yeon-jung Seo*[1]
[1]Iowa State University
[2]Iowa State Univesity
mskaiser@iastate.edu

In spatial Markov random field models, specification of an appropriate dependence structure revolves around the selection of neighborhoods. A neighborhood structure implies conditional independencies that are difficult to diagnose and assess because, although they are influenced by distributional assignments, they are characteristics of a model that transcend the form of the full conditional distributions that constitute the model specification. For example, the use of Gaussian or log-Gaussian conditional distributions constitutes a different aspect of model behavior than do the conditional independencies implied by four-nearest versus eight-nearest neighborhood structures. We apply a spatial block empirical likelihood method to tackle this problem. The method has proven effective in distinguishing between different potential neighborhood structures and a test procedure based on it exhibits good power. The procedure is illustrated with two problems, choosing between four-nearest and eight-nearest neighborhoods for a Gaussian model, and choosing between two co-dependent Markov random fields and one bivariate Markov random field. This latter problem is motivated by a problem of modeling the spatial relation between wind speeds and power generation in a wind farm.

Global patterns of lightning properties using spatial point process models on a sphere

Mikyoung Jun
Texas A&M University
mjun@stat.tamu.edu

Methods for spatial and spatio-temporal point process models have been traditionally concerned with planar domain. In this talk, I will

present a method for spatial and spatio-temporal point patterns on a global scale. In particular, nonstationary structures of the point patterns, commonly found in environmental applications, will be modeled through flexible parametric models. Some of the proposed models will be applied to global lightning data and model assessment results will be presented. Computational techniques for dealing with large point pattern data will be discussed.

A sparse areal mixed model for multivariate outcomes

John Hughes
University of Minnesota
jphughesjr@gmail.com

Multivariate spatially aggregated data are common in many disciplines. When fitting spatial regressions for such data, one needs to account for dependence to ensure reliable inference for the regression coefficients. Traditional multivariate conditional autoregressive (MCAR) models offer a popular and flexible approach to modeling such data, but the MCAR models suffer from two major shortcomings: (1) bias and variance inflation due to spatial confounding, and (2) high-dimensional spatial random effects that make fully Bayesian inference for such models computationally challenging. We propose the multivariate sparse areal mixed model (MSAMM) as an alternative to the MCAR models. Since the MSAMM extends the univariate SAMM, the MSAMM alleviates spatial confounding and speeds computation by greatly reducing the dimension of the spatial random effects. We specialize the MSAMM to handle zero-inflated count data (which can be modeled as bivariate outcomes), and apply our zero-inflated model to a large Census dataset for the state of Iowa.

Session 34: Semiparametric Methods in Biostatistics

Generalized Accelerated Failure Time Spatial Frailty Model for Arbitrarily Censored Data

Haiming Zhou[1]*, Timothy Hanson*[2] *and* ♦*Jiajia Zhang*[2]
[1]Northern Illinois University
[2]University of South Carolina
jzhang@mailbox.sc.edu

Flexible incorporation of both geographical patterning and risk effects in cancer survival models is becoming increasingly important, due in part to the recent availability of large cancer registries. Most spatial survival models stochastically order survival curves from different subpopulations. However, it is common for survival curves from two subpopulations to cross in epidemiological cancer studies and thus interpretable standard survival models can not be used without some modification. Common fixes are the inclusion of time-varying regression effects in the proportional hazards model or fully nonparametric modeling, either of which destroys any easy interpretability from the fitted model. To address this issue, we develop a generalized accelerated failure time model which allows stratification on continuous or categorical covariates, as well as providing per-variable tests for whether stratification is necessary via novel approximate Bayes factors. The model is interpretable in terms of how median survival changes and able to capture crossing survival curves in the presence of spatial correlation. A detailed Markov chain Monte Carlo algorithm is presented for posterior inference and a freely available function frailtyGAFT is provided to fit the model in the R package spBayesSurv. We apply our approach to a subset of the prostate cancer data gathered for

Graphical Modeling of Biological Pathways in Genome-wide Association Studies

♦*Yujing Cao and Min Chen*
University of Texas at Dallas
mchen@utdallas.edu

Complex diseases are affected by a variety of genetic factors. Identifying disease-associated genes in GWAS can help us to better understand the genetic risk factors that may lead to the development of prevention and intervention solutions to fight these diseases. Traditional single marker based approaches in GWAS often lack adequate statistical power. To address the challenge, researchers are using existing knowledge of biological pathways to assist in the search of disease-associated genes, which motivated by the fact that a complex disease is jointly affected by multiple genes. Studies show that considering a single biological pathway within an association analysis has worked well to improve the power of detecting disease-associated genes. We propose to incorporate more than one biological pathway to further increased the power. Different biological pathways can have similar genes interacting in different ways. Considering multiple biological pathways can bring us more information about the functional relations between these genes. We propose a Bayesian framework to combine two or more biological pathways, and employ a graphical model based on Markov Random Field (MRF) to describe the topological structure of the combined biological pathways. Simulation studies show combining multiple pathways can improve the power of identifying the association status of genes.

Semiparametric regression analysis for multiple-disease group testing data

♦*Dewei Wang, Peijie Hou and Joshua Tebbs*
University of South Carolina
deweiwang@stat.sc.edu

Group testing, also known as pooled testing, has been widely implemented as an efficient means to lower cost for large-scaled infectious disease screening. With the use of assays, that detect multiple infections, screening practices now produce testing results for multiple diseases simultaneously. This new type of data is referred to as the multiple-disease group testing data. Recent advances in the group testing estimation literature that deal with this type of data have mainly focused on estimating disease prevalences; i.e., assuming the population is homogenous. The main goal of this work is to build a regression model that can produce interpretable statistical inference for each disease separately. The approach we took views the correlation among diseases as nuisance parameters that can be modeled by multiple single index models. We demonstrated the asymptotic properties of our estimator. Further, we investigate the finite sample performance of our proposed methodology through simulation and by applying it to a chlamydia and gonorrhea data obtained from the Infertility Prevention Project.

Penalized spline mixed effects model with random time shift; an application to labor curves

Caroline Mulatya and ♦*Alexander McLain*
University of South Carolina
mclaina@mailbox.sc.edu

Characterizing and predicting future labor curves is important to obstetrics since it can aid in labor management and decision making. However, modeling cervical dilation curves remain a challenge due to the fact that women's onset of labor is unknown. Often researchers take the time to full dilation (i.e.,10 cm) as the benchmark time and run time backwards. This approach has drawbacks as it does not include women whose cervical dilation does not get to 10 cm, and cannot be used for prospective prediction. In this

talk, we propose a penalized splines mixed effects model with random shift parameters for prospectively modeling labor curves. In the proposed model, we view each woman as having an unknown time-shift, which when adjusted for appropriately aligns her curve. Our proposed model uses flexible B-spline basis functions to capture the non-linear relationship between cervical dilation and time. Penalized smoothing splines are used to balance the complexity of the model with the goodness of fit. To incorporate the random time shift parameters, a Monte Carlo Expectation Maximization (MCEM) procedure is implemented. The random shift parameters were generated using a rejection sampling algorithm in E-step. In M-step, we propose an augmented data approach which allows convenient updating of the parameters via the nlme package in R. We demonstrate the proposed method through simulation studies and real data from Consortium of Safe Labor study. We also demonstrate the utility of the proposed approach in providing dynamic individualized predictions.

Session 35: Recent Research of Omics Data by Young Investigators

GMMAT: logistic mixed models to control for population stratification and relatedness in genetic ass

♦*Han Chen*[1]*, Chaolong Wang*[2]*, Matthew Conomos*[3]*, Adrienne Stilp*[3]*, Cathy Laurie*[3]*, Ken Rice*[3] *and Xihong Lin*[1]
[1]Harvard T.H. Chan School of Public Health
[2]Genome Institute of Singapore
[3]University of Washington
hanchen@hsph.harvard.edu

Linear mixed models have recently become popular in genetic association studies to account for population stratification and relatedness, and in practice they have been applied to both continuous and binary traits. We show that applying linear mixed models to binary traits may lead to incorrect type I error rates in the presence of population stratification. We develop GMMAT, a computationally efficient program for binary trait analysis using logistic mixed models, and show in both simulation studies and real data that it accounts for population stratification and relatedness effectively.

Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples

Qi Yan
University of Pittsburgh
qi.yan@chp.edu

The recent development of sequencing technology allows identification of association between the whole spectrum of genetic variants and complex diseases. Jointly testing for association between genetic variants and multiple correlated phenotypes may increase the power to detect causal genes in family-based studies, but familial correlation needs to be appropriately handled. Here we propose a novel approach, for multivariate family data using kernel machine regression (denoted as MF-KM), which is based on linear mixed model framework and can be applied to a large range of studies with different types of traits.

Meta-analysis of Quantitative Pleiotropic Traits at Gene Level with Multivariate Functional Linear M

♦*Chi-Yang Chiu and Ruzong Fan*
NIH
chiuchiyang@gmail.com

For a meta-analysis of multiple studies, multivariate functional linear models are developed to connect genetic variant data to multiple quantitative traits adjusting for covariates. The goal is to take the advantage of both meta-analysis and pleiotropy analysis in order to improve power and to carry out a unified association analysis of multiple studies and multiple traits of complex disorders. Three types of approximate F-distributions based on Pillai-Bartlett trace, Hotelling-Lawley trace, and Wilks's Lambda are introduced to test association between multiple quantitative traits and multiple genetic variants. Simulation analysis is performed to evaluate the false positive rates and power performance of the proposed models and tests. The proposed methods were applied to analyze lipid traits in eight European cohorts.

Session 36: New Methods with Large and Complex Data

Blessing of Massive Scale: A Total Cardinality Constrained Approach for Spatial Graphical Model

♦*Ethan Fang, Han Liu and Mengdi Wang*
Princeton University
ethanfangxy@gmail.com

We consider high-dimensional spatial graphical model estimation under a total cardinality constraint (i.e., the ℓ_0-constraint). Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this 'blessing of massive scale' phenomenon. Motivated by this result, we propose an efficient algorithm to solve the dual problem and prove that the solution achieves optimal statistical properties.

Trace pursuit for model-free variable selection with matrix-valued predictors

Yuexiao Dong
Temple University
ydong@temple.edu

As a novel model-free variable selection method, trace pursuit with vector-valued predictors is proposed in Yu, Dong and Zhu (2015). We extend trace pursuit to matrix-valued predictors in this paper. A unified sequential test approach is applied to three types of hypotheses, which include test for significant rows, test for significant columns, and test for significant sub-matrices. The effectiveness of the proposed methods are demonstrated through extensive numerical studies.

Testing independence with high-dimensional correlated samples

♦*Xi Chen* [1] *and Weidong Liu* [2]
[1]New York University
[2]Shanghai Jiaotong University
xichen@nyu.edu

Testing independence among a number of (ultra) high-dimensional random samples is an fundamental and challenging problem. By arranging n identically distributed p-dimensional random vectors into a p by n data matrix, we investigate the testing problem on independence among columns under the matrix-variate normal modeling of the data. We propose a computationally simple and tuning free test statistic, characterize its limiting null distribution, analyze the statistical power and prove its minimax optimality. As an important by-product of the test statistic, a ratio-consistent estimator for the quadratic functional of covariance matrix from correlated samples is developed.

Precision Matrix Estimation by Inverse Principal Orthogonal Decomposition

♦*Cheng Yong Tang*[1] *and Yingying Fan*[2]
[1]Temple University

[2]University of Southern California
yongtang@temple.edu

We consider a parsimonious approach for modeling a large precision matrix in a factor model setting. The approach is developed by inverting a principal orthogonal decomposition (IPOD) that disentangles the systematic component from the idiosyncratic component in the target dynamic system of interest. In the IPOD approach, the impact due to the systematic component is captured by a low-dimensional factor model. Motivated by practical considerations for parsimonious and interpretable methods, we propose to use a sparse precision matrix to capture the contribution from the idiosyncratic component to the variation in the target dynamic system. Conditioning on the factors, the IPOD approach has an appealing practical interpretation in the conventional graphical models for informatively investigating the associations between the idiosyncratic components. We discover that the large precision matrix depends on the idiosyncratic component only through its sparse precision matrix, and show that IPOD is convenient and feasible for estimating the large precision matrix in which only inverting a low-dimensional matrix is involved. We formally establish the estimation error bounds of the IPOD approach under various losses and find that the impact due to the common factors vanishes as the dimensionality of the precision matrix diverges. Extensive numerical examples including real data examples in practical problems demonstrate the merits of the IPOD approach in its performance and interpretability.

Session 37: Can Linearly Dependent Confounders Be Estimated? ¨C The Case of Age-Period-Cohort and Beyond

The Great Society, Reagan's Revolution, and Generations of Presidential Voting

♦*Yair Ghitza*[1] *and Andrew Gelman*[2]
[1]Catalist
[2]Columbia University
yghitza1@gmail.com

We build a generational model of presidential voting, in which long-term partisan presidential voting preferences are formed, in large part, through a weighted "running tally" of retrospective presidential evaluations, where weights are determined by the age in which the evaluation was made. Under the model, the Gallup Presidential Approval Rating time series is shown to be a good approximation to the political events that inform retrospective presidential evaluations. The political events of a voter's teenage and early adult years, centered around the age of 18, are enormously important in the formation of these long-term partisan preferences. The model is shown to be powerful, explaining a substantial amount of the macro-level voting trends of the last half century, especially for white voters and non-Southern whites in particular. We use a narrative of presidential political events from the 1940s to the present day to describe the model, illustrating the formation of five main generations of presidential voters.

Confusions about the APC confounding. What have we missed? How can we do better?

Wenjiang Fu
Department of Mathematics, University of Houston
wenjiangfu@hotmail.com

The linearly dependent age, period and birth cohort (APC) are well known confounders and their fixed effects have been believed deeply for decades non-estimable due to the parameter identification problem in the APC multiple classification model. Yet the importance of modeling simultaneous age, period and cohort effects has been emphasized repeatedly in the literature. Although this difficult parameter identification problem seems to be intractable, recent works provide solid evidence to address the problem, making it more promising than ever before to resolve this long term unsettled controversy. In this paper, I will first review the identification problem, then highlight some recent works, including the intrinsic estimator (Fu 2000, Fu 2015) and the smoothing cohort estimator (Fu 2008) and their properties, and point out why these novel approaches are promising using large sample theory and finite sample robustness. I will further explain what have been missed in previous work. Finally, I will present most recent work in hypothesis testing on the age, period and cohort effects between populations for statistical inference and illustrate with data in public health and economic studies.

Keywords: age-period-cohort, bias, confounding; consistent; intrinsic estimator; robust estimation.

Bias Correction in Modeling Complex Rate Data – How and Why?

Martina Fu
Stanford University
fumm95@stanford.edu

Modeling event rates can be complex in public health research or social studies to estimate temporal trends across a number of years. On one hand, the age-period-cohort (APC) models have a difficult identifiability problem, where biased estimation is often generated with popular approaches by specifying constraints based on investigator's belief. On the other hand, the estimation and comparison of summary rate through the direct age-standardization have been recently challenged on the selection of the standard population as the reference age structure. We study these two approaches and find that bias dominates in the estimation of parameters and temporal trend. We develop a bias correction method fo

Alternative Approach to the Identifiability Problem–Finding the Truth through Smoothing

♦*Shujiao Huang and Wenjiang Fu*
University of Houston
shujiao.h@gmail.com

We consider age-period-cohort (APC) model in social studies and chronic disease epidemiology on an a×p table with single observation in each cell, where rows, columns and diagonals represent age, period and birth cohort. The APC classification regression model suffers from an identifiability problem with multiple estimators having the same fitted values. Here we develop a two-stage smoothing-cohort model to address the identifiability problem. In stage 1, a smoothing cohort model yields a unique estimator with consistent estimation for age and period effects but not cohort effect. In stage 2, a non-contrast constraint is applied to age or period effect with estimates from stage 1. Three constraint selection methods are examined, including the largest ratio of estimates, the smallest variance of the estimate ratio and the smallest variance of linear combination of estimates. Our simulation results based on an a¿p data set show that the constraints on period effects outperform those on age effects. The constraint by the smallest variance of the period effect ratio yields the best estimation. We demonstrate our method with SEER cancer mortality data and sociology data.

Session 38: Challenges and Methods in Biomarker Research and Medical Diagnosis

Regulatory Perspective and Case Studies on Biomarker Validation of Companion Diagnostics

♦*Jingjing Ye and Gene Pennello*
FDA
jingjing.ye@fda.hhs.gov

A biomarker can be useful to provide information that is essential for the safe and effective use of a corresponding therapeutic product. For example, colorectal cancer patients who test negative for KRAS mutations are eligible for treatment with cetuximab. The assay or test to identify the biomarker should be analytically and clinically validated. In this talk, we will discuss the types and intended uses of the biomarkers, and the study designs and analyses to do the analytical and clinical validation on the biomarkers that meet the regulatory requirement. We will discuss case studies of several FDA approved the biomarkers.

Better Use of Family History Data to Predict Breast Cancer Risk

♦*Shanshan Zhao*[1]*, Yue Jiang*[2] *and Clarice Weinberg*[1]
[1]National Institute of Environmental Health Science
[2]University of North Carolina at Chapel Hill
shanshan.zhao@nih.gov

Family history is an important risk factor for many diseases, such as breast cancer. In statistical modeling, family history is usually used as a yes/no variable. However, if a female has one sister with breast cancer and she only has one sister, her risk of developing breast cancer is higher than those who has one diseased sister and several disease free sisters. Simply using family history as a yes/no variable will reduce the prediction power. We aim to develop a family history score that takes both number of diseased relatives and family size into consideration. This family history score is expected to improve prediction power. We apply this approach to the NIEHS Sister Study.

Prediction of longitudinal biomarkers on recurrence events in the presence of a terminal event

♦*Ming Wang, Cong Xu and Vern M. Chinchilli*
Penn State Hershey Medical Center
mwang@phs.psu.edu

Acute kidney injury (AKI) is an important clinical outcome to characterize the renal function, which could occur recurrently during the follow-up hospitalizations. There exist some observation evidences that the patient with more frequent episodes of AKI have higher risk of terminal event of death, indicating informative censoring mechanism. However, the urine/serum biomarkers for AKI recurrence are still in search or need validation. Our motivation study is from the Assessment, Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury (ASSESS-AKI) Consortium including AKI and non-AKI participants matched on major baseline confounders. Several potential biomarkers of interest are collected longitudinally in order to capture the renal disease progression. We propose a joint modeling of longitudinal biomarker, recurrent events and death to investigate their association based on the shared random-effect modeling. Maximum likelihood estimation and inference are obtained, and thereafter the predictive accuracy of the models with each individual biomarker are quantified using the expected prognostic observed cross-entropy (EPOCE) for comparison and identify the most promising candidate. Extensive simulation studies are provided to evaluate our proposal.

Estimation of Diagnostic Accuracy of a Biomarker When the Gold Standard Is Measured with Error

Mixia Wu[1]*, Dianchen Zhang*[1] *and* ♦*Aiyi Liu*[2]
[1]Beijing University of Technolog
[2]NICHD/NIH
liua@mail.nih.gov

New biomarkers continue to be developed for the purpose of diagnosis, and their diagnostic performances are typically compared with an existing gold standard used for the same purpose. Considerable amount of research has focused on receiver operating characteristic curves analysis when the gold standard is dichotomous. In the situation where the gold standard is measured on a continuous scale and dichotomization is not practically appealing, Obuchowski (2005) proposed an index to measure the accuracy of a continuous biomarker, which is essentially a linear function of the popular Kendall's tau. We consider the issue of estimating such an accuracy index when the continuous gold standard is measured with errors. We first investigate the impact of measurement errors on the accuracy index, and then propose methods to correct for the bias due to measurement errors. Simulation results show the effectiveness of the proposed estimator in reducing biases. The methods are exemplified with hemoglobin A1c measurements obtained from both the central lab and a local lab to evaluate the accuracy of the mean data obtained from the metered blood glucose monitoring against the centrally measured hemoglobin A1c from a behavioral intervention study for families of youth with type 1 diabetes.

Session 39: Change-Point Problems and their Applications (II)

Majority versus Consensus Decision Making in Decentralized Sequential Change Detection.

♦*Georgios Fellouris and Sourabh Banerjee*
University of Illinois, Urbana-Champaign
fellouri@illinois.edu

Suppose that multiple sensors monitor a system and at some unknown time there is anomaly in the environment that affects their observations. Assuming that each sensor is detecting the change locally, using the corresponding CUSUM statistic, we want to understand how to design the local decisions and also how to combine them in order to quickly detect the change, while controlling the global false alarm rate. While a first-order asymptotic analysis suggests that it is better to require consensus before raising an alarm, a second-order analysis reveals that a majority rule performs better in practice. This insight is verified by simulation experiments

On robustness of N-CUSUM stopping rule in a Wiener disorder problem

♦*Hongzhong Zhang*[1]*, Neofytos Rodosthenou*[2] *and Olympia Hadjiliadis*[3]
[1]Columbia University
[2]Queen Mary University of London
[3]City University of New York
hz2244@columbia.edu

We study a Wiener disorder problem of detecting the minimum of N change-points in N observation channels coupled by correlated noises. It is assumed that the observations in each dimension can have different strengths and that the change-points may differ from channel to channel. The objective is the quickest detection of the minimum of the N change-points. We adopt a min-max approach and consider an extended Lordens criterion, which is minimized

subject to a constraint on the mean time to the first false alarm. It is seen that, under partial information of the post-change drifts and a general nonsingular stochastic correlation structure in the noises, the minimum of N cumulative sums (CUSUM) stopping rules is asymptotically optimal as the mean time to the first false alarm increases without bound.

On the Optimality of Bayesian Change-Point Dtection

♦*Dong Han*[1]*, Fugee Tsung*[2] *and Jinguo Xian*[1]
[1]Dept. of Statistics, Shanghai Jiao Tong Univ.
[2]Dept. of IELM, Hong Kong Univ. of Sc. & Techn.
donghan@sjtu.edu.cn

By introducing suitable loss random variables of detection, we obtain optimal tests in terms of stopping time or alarm time for Bayesian change-point detection not only for a general prior distribution of change-points but also for observations being a Markov process. Moreover, the optimal (minimal) average detection delay is proved to be equal to 1 for any (possibly large) average run length to false alarm if there are at most finite possible change-points.

Estimating the Number of States in Hidden Markov Models via Marginal Likelihood

Yang Chen[1]*, Cheng-Der Fuh*[2]*,* ♦*Chu-Lan Kao*[2] *and Samuel Kou*[1]
[1]Harvard University
[2]National Central University
chulankao@gmail.com

We propose an estimator for the number of states in hidden Markov models based on marginal likelihood. We show that, by maximizing the marginal likelihood, one could have a consistent estimator for the number of states, and the convergence rate is further provided. For computational purpose, we also propose a method to efficiently estimate the marginal likelihood. Simulation studies are also provided.

Session 40: Statistical Issues in Analysis and Interpretation of Human Drug Abuse Study Data

Some Review Issues in Design and Statistical Analysis of Human Drug Abuse Potential Studies

Wei Liu
FDA CDER
Wei.Liu@fda.hhs.gov

In response to the growing drug abuse and misuse epidemic, FDA plays an important role to find solutions through regulation and public health strategies. In 2015, FDA issued a final guidance, "Guidance for Industry: Abuse-Deterrent Opioids - Evaluation and Labeling" to assist industry in developing opioid drug products with potentially abuse-deterrent properties. However, many abuse deterrent studies submitted to FDA were planned before the release of this guidance and FDA's recommendations were not implemented in the statistical analysis. In addition, there are merging statistical challenges in multi-faceted and often debated issue. Discussion of the statistical issues in analyzing these data among statisticians from both the FDA and the pharmaceutical industry is extremely important to the proper evaluation of abuse-deterrent effects. In this presentation, I will discuss some common statistical review issues in the design and analysis of abuse deterrent studies, particularly the primary and some secondary or supportive analyses as seen in new drug applications.

Statistical Approaches and Issues in HAP Studies

Kelsey Brown[1]*,* ♦*Reilly Reis*[1] *and Michael Smith*
[1]PRA Health Sciences
reisreilly@prahs.com

Human Abuse Potential (HAP) studies are designed to evaluate the abuse deterrent or potential characteristics of a drug. They provide an understanding of how well the drug is "liked"(or not "liked" compared to another drug in the same class or form of the drug to provide an estimate of the likelihood for a drug to be abused. The inferential analyses of the various abuse potential measures to evaluate these studies are continually evolving. Each approach comes with its own statistical benefits and issues. We will be presenting several methods utilized to determine abuse potential and will discuss various challenges that arise.

Common statistical issues in drug development for major psychiatric disorders

Thomas Birkner
Food and Drug Administration
Thomas.Birkner@fda.hhs.gov

The presentation will consist of three parts: 1) A brief description of the connections of a statistical reviewer of psychiatric drugs to human abuse potential (studies); 2) An overview of proposed methods to reduce the often observed large placebo response in depression and schizophrenia trials; and 3) The current practice and new developments in the treatment of missing data (i.e., estimands of interest) in statistical analysis.

Session 41: Analysis of Multi-Type Data

Bayesian Models and Analysis of High-dimensional Multiplatform Genomics Data

Sounak Chakraborty
University of Missouri-Columbia
chakrabortys@missouri.edu

In recent years the scale of omcis studies has expanded to measure and include multiple genomic features on a single patient, like gene expression, DNA methylation, gene mutation, copy number variation, promoter binding and protein expression. Combining and modeling multiple genome features coming from different data platforms is a big conceptual challenge and practical hurdle. However, it promises to improve the overall patient care and health. In this paper we propose to develop statistical nonlinear models to integrate genomic data from multiple platforms. Our models can incorporate the fundamental biological relationships that exist among the data sets obtained from different platforms and produce more accurate understanding of the functional responses. The proposed models are developed on the basis of Bayesian trees and Bayesian kernel machine models. Our methodologies are highly flexible in exploring, extracting, and analyzing complex biological systems and data sets from heterogeneous platforms. Combining all available genetic, pathological, and demographic information our models can dramatically improve the nature of clinical diagnosis and treatment of several human diseases.

Kernel machines for -omics data integration

♦*Dominik Reinhold, Junxiao Hu, Katerina Kechris and Debashis Ghosh*
University of Colorado Denver
dominik.reinhold@ucdenver.edu

In this talk, we will discuss a class of statistical methods for high-dimensional data that are termed kernel machines. They have been popularized in the machine learning context and have found tremendous utility in various genomics contexts recently. One key concept that arises is that of a metric, which is used to describe similarities

between pairs of observations. The similarities are represented by a kernel matrix. Given the availability of a metric for any particular data structure, a straightforward development of theory for testing of associations using kernel machines is available. The methodology is fairly generic and can be applied to a wide variety of fields. Popular kernels include the Gaussian and linear kernels. These kernels do not incorporate network structure that might be present in the data and that could give additional information on similarities between pairs. For the linear kernel, $k(x, y) = x'y$, a network linear kernel can be defined as $x'Ny$, where N represents the network structure. For the Gaussian kernel, $k(x, y) = exp(-a||x - y||^2)$, we present a generalization of the approach for the linear kernel. The idea is to consider a series expansion involving terms of the form $(x'y)^j$. These terms can then be replaced by $(x'Ny)^j$. Importantly, this strategy can be used for all kernels that have such a series expansion. We describe applications of kernel machines to Chronic Obstructive Pulmonary Disease (COPD) data, studying associations between phenotypes and genes (gene expression data). The network structure is obtained from Weighted Gene Co- expression Network Analysis (WGCNA) and from the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Big Data Regression for Predicting Genome-wide Functional Genomic Signals

Weiqiang Zhou, Ben Sherwood, Zhicheng Ji, Fang Du, Jaiwei Bai and ♦*Hongkai Ji*

Johns Hopkins Bloomberg School of Public Health
hji@jhu.edu

We develop BIRD, Big Data Regression, to handle the ultra-high-dimensional problem of predicting functional genomic signals on DNA using other data types. Applying BIRD to the Encyclopedia of DNA Element (ENCODE) data, we found that gene expression to a large extent predicts DNase I hypersensitivity (DH). We show that the predicted DH profile predicts transcription factor binding sites (TFBSs), prediction models trained using ENCODE data can be applied to Gene Expression Omnibus (GEO) samples to predict regulome, and one can use predictions as pseudo-replicates to improve the analysis of DNase-seq and ChIP-seq data. These analyses not only improve our understanding of the regulome-transcriptome relationship, but also illustrate that transcriptome-based prediction can provide a useful new approach for regulome mapping.

Prioritizing causal SNPs through integrating phenotype, genotype, omics and functional annotations

♦*Qi Zhang*[1]*, Constanza Rojo*[2] *and Sunduz Keles*[2]

[1]University of Nebraska Lincoln

[2]University of Wisconsin Madison
qi.zhang@unl.edu

In recent years, many large biomedical studies have been taken place, and the usually includes various types of data, such as clinical phenotype, genotype, gene expression and many epigenetic marks. The goals of such studies are usually identifying causal SNPs or markers of the phenotype of interests, and understand the molecular mechanism. The dominant work flow of quantitative analysis in such studies are usually performing *-QTL type of analysis on each of the data type, and study their overlaps with each other, and with functional annotation databases such as ENCODE tracks. However, such work flow ignores the natural association among different data types. In this project, we proposed a framework to prioritize SNPs/markers through joint analysis of these data. It integrates the multi-type data from the biomedical study under investigation and the functional annotation data in a principled way. We applied our method to human and mouse data, and the results lead to easier biological interpretation.

Session 42: New Advances in Quantile Regression

Estimation and Inference of Quantile Regression Under Biased Sampling

♦*Gongjun Xu*[1]*, Tony Sit*[2]*, Lan Wang*[1] *and Chiung-Yu Huang*[3]

[1]University of Minnesota

[2]The Chinese University of Hong Kong

[3]Johns Hopkins University
xuxxx360@umn.edu

Biased sampling occurs frequently in economics, epidemiology and medical studies either by design or due to data collecting mechanism. Failing to take into account the sampling bias usually leads to incorrect inference. We propose a unified estimation procedure and a computationally fast resampling method to make statistical inference for quantile regression with survival data under biased sampling schemes, including but not limited to the length-biased sampling, the case-cohort design and variants thereof. We establish the consistency and weak convergence of the proposed estimator as a process of the quantile level. We also investigate more efficient estimation using the generalized method of moments. The proposed method provides researchers and practitioners a convenient tool for analyzing data collected from various designs. Simulation studies and applications to real data sets are presented for illustration.

A Quantile Approach for Fractional Data

♦*Hyokyoung (Grace) Hong*[1] *and Huixia Wang*[2]

[1]Michigan State University

[2]George Washington University
hhong@stt.msu.edu

We focus on fractional data whose analysis poses challenges in two folds. First, the fractional data is bounded at zero and one. Therefore, conventional regression techniques assuming unrestricted range of the response distribution are not applicable. Second, mass points may occur at either or both of the extremes. We propose a new quantile regression method for analyzing fractional data. The proposed method avoids any restriction imposed on the interval [0, 1] and accounts for mass points by nonparametric transformation and censored quantile regression. The value of our proposed method is illustrated through simulation and the analysis of a data set of employee participation rates in 401(k) pension plans, where over 40% of the plans have a participation rate of exactly one.

High dimensional censored quantile regression

♦*Qi Zheng*[1]*, Limin Peng*[2] *and Xuming He*[3]

[1]University of Louisville

[2]Emory University

[3]University of Michigan
qi.zheng@louisville.edu

Quantile regression has emerged as a useful regression strategy to analyze heterogeneous covariate-response associations that are often encountered in practice. While most of current quantile regression for high dimensional covariates primary focuses on the complete data, the related development in dealing with censored survival (i.e. time-to-event) responses has been relatively sparse. We propose a new penalized censored quantile regression in the high dimensional survival data. Our two-step procedure adopts the perspective of globally concerned quantile regression (Zheng et al. [2015]), and sequentially investigates conditional quantiles over a continuum of quantile indices. In the first step, we incorporate

Lasso type L1 penalty functions into the stochastic integral based estimating equation for CQR to obtain a uniformly consistent estimator over the quantile region of interest. In the second step, we employ the Adaptive Lasso type penalties and uniform tuning parameter selectors to further reduce the bias induced by L1 penalties and the resulting estimator achieves improved estimation efficiency and model selection consistency. Our theoretical results also include the oracle rate of uniform convergence and weak convergence of the parameter estimators. Moreover, we use numerical studies to confirm our theoretical findings and illustrate the practical utility of our proposal.

An alternative formulation of functional partial quantile linear regression and its properties

♦*Dengdeng Yu, Linglong Kong and Ivan Mizera*
University of Alberta
dengdeng@ualberta.ca

The difficulty of deducing the asymptotic properties for functional partial quantile regression is mainly due to its iterative nature of bases formulation.

It used to have the same problem for functional partial least squares regression. However, Delaigle and Hall 2012 managed to bypass the discussion about such iterative formulation by using an alternative partial least squares formulation, based on the fact that there exists an equivalence between functional principal component space and functional partial least squares space, hence demonstrate consistency and establish convergence rates. Unfortunately, for functional PQR space we can not find such equivalence.

Hereby we propose an alternative partial quantile regression formulation for functional linear regression and establish the corresponding asymptotic properties.

Session 43: Advances and Challenges in Time-to-Event Data Analysis

Analysis of dependently truncated data in Cox framework

♦*Xu Zhang*[1]*, Yang Liu*[2] *and Ji Li*[3]
[1]University of Mississippi Medical Center
[2]Centers for Disease Control and Prevention
[3]University of Oklahoma Health Sciences Center
xzhang2@umc.edu

Truncation is a known feature of the bone marrow transplant (BMT) registry data, for which the survival time of a leukemia patient is left truncated by the transplant waiting time. Quasi-independence between the survival time variable and the transplant time variable is routinely assumed in analysis of the BMT registry data. It was recently noted that a longer waiting time was linked to poorer survival. A straightforward solution is the left-truncated version Cox model on the survival time with the transplant time as both the truncation variable and the covariate. We aimed at studying the probability of selection in this framework. We proposed the point estimator and derived its asymptotic distribution. Both truncated only data and censored and truncated data were generated in the simulation study. The proposed point and variance estimators showed good performance in various simulated settings. The bone marrow transplant registry data were analyzed as the illustrative example.

To MICE or not to MICE? A study of multiple imputation strategies for Accelerated Failure Time Model

♦*Lihong Qi*[1]*, Ying-Fang Wang*[2]*, Rongqi Chen*[1] *and Yulei He*[3]
[1]University of California Davis
[2]The California State University
[3]CDC
lhqi@ucdavis.edu

Complete Title: To MICE or not to MICE? A study of multiple imputation strategies for Accelerated Failure Time Models with Missing Covariates

Missing covariates often occur in biomedical studies with survival or time-to-event outcomes. Multiple imputation (MI) is an effective approach to this problem, especially with multiple incomplete covariates. Two main strategies for imputing multivariate incomplete data exist: joint modeling (JM) and fully conditional specification (FCS). JM is based on parametric statistical theory and is theoretically sound. However, the joint model may lack flexibility needed to represent typical data features, potentially leading to bias. In the past, the application of JM is also hindered by the lack of imputational machinery for the wide variety of multivariate models. FCS is a semi-parametric and flexible alternative that specifies the multivariate model by a series of conditional models, one for each incomplete variable. Compared with JM, FCS appears to have received more attentions recently due to its flexibility, seemingly simpler structure, and wide availability in software. Nevertheless, the theoretical properties of FCS are difficult to establish. In addition, practitioners tend to specify these conditional models in simple manners largely dictated by the software. Motivated by a study of hip fractures in Women's Health Initiative Cohort, we apply JM and FCS for multivariate incomplete covariates in accelerated failure time (AFT) models. In JM, we specify the joint model as the product of AFT model and general location model, implementing the imputation through the WinBUGS software. In FCS, we test a wide variety of specifications for the conditional models, implementing the imputation through the MICE software. Through a comprehensive simulation study, we investigate the performance of the two strategies when the corresponding imputation models are coherent or incoherent with the data-generating models. Simulation results show that in most of the cases, using the JM strategy yields satisfactory results. In addition, not all choices of the specifications for the conditional models in the FCS strategy yield good results: the ones based on simple specifications tend to produce suboptimal results in our settings. Therefore, despite the popularity of FCS among practitioners, we recommend reconsidering the JM strategy as an alternative, facilitated by the Bayesian computational software such as WinBUGS. We also warn against a mechanical use of FCS and suggest careful modeling of the conditional distributions among variables to ensure its good performance.

Surviving joint models: Improving the survival subcomponent of joint longitudinal-survival models

Michael Griswold
Center of Biostatistics, Univ MS Medical Center
mgriswold@umc.edu

Over the last two decades the development of joint models for simultaneous modelling of longitudinal and survival data has received substantial attention. These models provide an elegant paradigm for partnering time-to-event outcomes with longitudinal outcomes in order to account for common analytic issues such as embedding predictor trajectories within survival models and examining sensitivity to informative missingness. Generally, the survival model aspect of such joint models is treated fairly simply. We discuss extensions of the survival subcomponent of joint models that allow incorporation of time-varying predictors using estimation algorithms available in general-purpose software. We demonstrate these techniques using over 20 years of real-world data on cognitive decline and dementia from the Atherosclerosis Risk in Communities (ARIC) study.

The proportional odds cumulative incidence model for competing risks

Frank Eriksson[1]*, Jianing Li*[2]*, Thomas Scheike*[1] *and* ♦*Mei-Jie Zhang*[3]
[1]University of Copenhage
[2]Merck
[3]Medical College of Wisconsin
meijie@mcw.edu

We suggest an estimator for the proportional odds cumulative incidence model for competing risks data. The key advantage of this model is that the regression parameters have the simple and useful odds ratio interpretation. The model has been considered by many authors, but it is rarely used in practice due to the lack of reliable estimation procedures. We suggest new estimating procedures and show that their performance improve considerably on existing methods. We also suggest a goodness-of-fit test for the proportional odds assumption. We derive the large sample properties and provide estimators of the asymptotic variance. The method is illustrated by an application in a bone marrow transplant study and the finite-sample properties are assessed by simulations.

Session 44: Jiann-Ping Hsu invited Session on Biostatistical and Regulatory Sciences

A homoscedasticity test for the Accelerated Failure Time model

♦*Lili Yu*[1]*, Liang Liu*[2] *and Din Chen*[3]
[1]Georgia Southern University
[2]University of Georgia
[3]University of North Carolina at Chapel Hill
lyu@georgiasouthern.edu

The semiparametric accelerated failure time (AFT) model is the major linear model for survival data. 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 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 paper is then the first to propose a novel quasi-likelihood ratio test for the homoscedasticity assumption in the AFT model. The asymptotic property of this test is investigated theoretically along with simulation studies to show the satisfactory performance of this novel statistical test. A real dataset is used to demonstrate the application of this developed test.

Prospective Validation of the National Field Triage Guidelines: Challenges of a Probability Stratifi

♦*Rongwei (Rochelle) Fu and Craig Newgard*
OHSU
fur@ohsu.edu

Prospective Validation of the National Field Triage Guidelines: Challenges of a Probability Stratified Design and Verification Bias

Field triage plays an integral role in trauma systems by guiding Emergency Medical Services (EMS) personnel in identifying and transporting high-risk patients to major trauma centers. The national field trauma triage guidelines have been widely implemented in US trauma systems, but never prospectively validated. In a study to prospectively validate the guidelines as applied by out-of-hospital providers for identifying high-risk trauma patients, it is not feasible to abstract data from all samples. Therefore, we employed a probability stratified design to create a primary sample. In this talk, we would discuss how to address the challenges in the data and using the probability stratified design (missing data, varying weighting scheme, etc.) and in particular, how to address the issue of verification bias to properly estimate sensitivity and specificity. We found that the national field triage guidelines are relatively insensitive for identifying seriously injured patients and patients requiring early critical interventions, particularly among older adults.

From Statistical Power to Statistical Assurance: Time for the Paradigm Change in Clinical Trial Desi

♦*Ding-Geng Chen*[1] *and Shuyen Ho*[2]
[1]University of North Carolina at Chapel Hill
[2]PAREXEL, Durham, NC 27709, USA
dinchen@email.unc.edu

In biopharmaceuticals and biostatistics, a well designed clinical trial requires an appropriate sample size with adequate statistical power to address trial objectives. The statistical power is traditionally defined as the probability of rejecting the null hypothesis with a prespecified true clinical treatment effect. This power is a conditional probability conditioned on the true but actually unknown effect. In practice, this true effect is never fixed as a constant so a newly proposed alternative to this conventional statistical power is statistical assurance, by O'Hagan and Stevens (2001). The statistical assurance is a new paradigm in clinical trial design and is defined as the unconditional probability of rejecting the null hypothesis. It can then be obtained as an expected power where the expectation is based on the prior probability distribution of the unknown treatment effect, therefore it is a Bayesian concept. In this talk, we review the transition from conventional statistical power to assurance and discuss the computations of assurance using Monte-Carlo simulation-based approach.

Internal pilot design for repeated measures

♦*Xinrui Zhang and Yueh-Yun Chi*
University of Florida
xinrui@ufl.edu

Repeated measures of outcome are common in clinical trials and epidemiological studies. Designing studies with repeated measures requires accurate specifications of the variances and correlations in order to select an appropriate sample size. Underspecifying the variances leads to a sample size that is inadequate to detect a meaningful scientific difference, while overspecifying the variances results in an unnecessarily large sample size. Both lead to waste of resources and place study participants in unwarranted risk. We extend earlier work on internal pilot designs with the "Univariate Approach" to repeated measures. We provide approximate distributions of the final sample size and the test statistic for the final analysis. Extensive simulations examine the impact of misspecification of the covariance matrix and demonstrate the accuracy of the approximations in controlling the Type I error rate and achieving the target power. The proposed methods are illustrated by application to a longitudinal study assessing early antiretroviral therapy for youth living with HIV.

Session 45: Complex Data Analysis: Theory and Methods

Linear hypothesis testing in high-dimensional one-way MANOVA

♦*Jin-Ting Zhang, Jia Guo and Bu Zhou*
National University of Singapore
stazjt@nus.edu.sg

In recent years, with rapid development of data collecting technologies, high-dimensional data become increasingly prevalent. Much work has been done for hypotheses on mean vectors, especially for high-dimensional two-sample problems. Rather than considering a specific problem, we are interested in a general linear hypothesis testing (GLHT) problem on mean vectors of several populations, which include many existing hypotheses about mean vectors as special cases. A few existing methodologies on this important GLHT problem impose strong assumptions on the underlying covariance matrix so that the null distributions of the associated test statistics are asymptotically normal. In this paper, we propose a simple and adaptive L^2-norm based test for the GLHT problem. For normal data, we show that the null distribution of our test statistic is the same as that of a chi-square type mixture which is generally skewed. We give a sufficient and necessary condition such that the null distribution of our test statistic is asymptotically normal. However, this condition is not always satisfied in real data analysis. To overcome this difficulty, we propose to approximate the distribution of our test statistic using the well-known Welch-Satterthwaite chi-squared-approximation. The asymptotic and approximate power of our test is also investigated. The methodologies are then extended for non-normal data. Two simulation studies and a real data application are presented to demonstrate the good performance of our new test.

Estimation of sparse directed acyclic graphs through a lasso framework and its applications

Sung Won Han and ♦*Judy Zhong*
New York University
judy.h.zhong@gmail.com

Causal networks are conveniently presented by directed acyclic graphs (DAGs). To estimate DAGs from high dimensional data is challenging due to the large number of possible spaces of DAGs, the acyclicity constraint of the structures, the typically nonconvex objective functions, and the problem of equivalent classes from observational data. In this talk, we present an efficient two-stage algorithm to estimate sparse DAGs under adaptive L1-penalized likelihood objective function with the acyclicity constraint. Simulations are presented to demonstrate the efficiency and fexibility of the proposed method. Real data examples are discussed on gene regulatory networks.

Graph-Guided Banding for Covariance Estimation

Jacob Bien
Cornell University
jbien@cornell.edu

Reliable estimation of the covariance matrix is notoriously difficult in high dimensions. Numerous methods assume that the population covariance (or inverse covariance) matrix is sparse while making no particular structural assumptions on the desired sparsity pattern. A highly-related, yet complementary, literature studies the setting in which the measured variables have a known ordering, in which case a banded (or near-banded) population matrix is assumed. This work focuses on the broad middle ground that lies between the former approach of complete neutrality to the sparsity pattern and the latter highly restrictive assumption of having a known ordering. We develop a class of convex regularizers that is in the spirit of banding and yet attains sparsity structures that can be customized to a wide variety of applications.

Flexible Spectral Methods for Community Detection

♦*Pengsheng Ji*[1]*, Jiashun Jin*[2] *and Tracy Ke*[3]
[1]University of Georgia
[2]Carnegie Mellon University
[3]University of Chicago
psji@uga.edu

We propose a class of flexible spectral methods for community detection in directed and undirected networks. These methods extract the clustering information by taking the entry-wise ratios of the eigenvectors, and can be adapted for different purposes including exploring substructures, incorporating covariates. Some practical guidance about the choice of the number of communities will also be provided. Then we demonstrate using the statistician coauthorship and citation data collected by ourselves, and show a handful of meaningful communities, such as "high-dimensional data", "large-scale multiple testing", "Dimensional Reduction", "Objective Bayes" and "Theoretical Machine Learning", etc.

Session 46: Fundamentals and Challenges in Subgroup Identification and Analysis

A Visualization Method Measuring the Performance of Biomarkers for Guiding Treatment Decisions and Subgroup Identification

♦*rui tang*[1]*, hui yang*[2] *and jing huang*[3]
[1]vertex
[2]amgen
[3]veracyte
rui_tang@vrtx.com

We are in the era of personalized medicine, where the ultimate goal is to predict response to therapy based on patient characteristics. Numerous examples have demonstrated the importance of identifying the correct subgroup(s) of patients based on baseline characteristics for which a therapy is effective. Biomarkers that predict efficacy for a given drug therapy become increasingly important for treatment strategy and drug evaluation in personalized medicine. Methodology for appropriately identifying and validating such biomarkers is critically needed, though it is very challenging to develop, especially in trials of terminal diseases with survival endpoints. The marker-by-treatment predictiveness curve serves this need by visualizing the treatment effect on survival as a function of biomarker for each treatment. In this presentation, I will present the weighted predictiveness curve (WPC) method we proposed. Based on the nature of the data, it generates predictiveness curves by utilizing either parametric or nonparametric approaches. Especially for nonparametric predictiveness curves, by incorporating local assessment techniques it requires minimum model assumptions and provides great flexibility to visualize the marker-by-treatment relationship. WPC can be used to compare biomarkers and identify the one with the highest potential impact. Equally important, by simultaneously viewing several treatment-specific predictiveness curves across the biomarker range, WPC can also guide the biomarker based treatment regimens and help identify treatment effect heterogeneity and a subgroup of patients could benefit from the treatment during the discovery and exploratory phase of bringing a biomarker to clinical practice. The proposed method has recently been published in the Pharmaceutical Statistics journal.

What can we learn from subgroup analysis in randomized controled trials?

Xin Zhao
Janssen Pharmaceutical
xzhao121@gmail.com

Randomized trials provide the best evidence as to whether a treatment is, in general, beneficial. A prudent interpretation of trial re-

sults is to limit findings that will affect clinical decisions to overall treatment effects regarding primary endpoints that have been carefully planned, powered, and controlled for errors. Hence subgroup findings should generally be considered as just exploratory results. Subgroup analysis can be given some credence when it is limited to a small number of prespecified groups and when effects are large, consistent, duplicated in other studies, and clinically plausible. When subgroup effects are in the opposite direction of the overall results, the most prudent approach is to consider subgroup findings as hypotheses for another trial. In this talk, we will examine some case studies on subgroup analysis in randomized controled trials, share the utility and limitations, and provide some insight on how to conduct subgroup analysis in randomized trial setting.

Subgroup Analysis to Assess Benefit:Risk

♦*Steven Snapinn and Qi Jiang*

Amgen

ssnapinn@amgen.com

Evaluation of subgroups is a routine part of the analysis of nearly every large clinical trial. The purpose is to determine whether the effects of the treatment are consistent across the study population, or whether there are patient characteristics that can be used to predict which patients will experience a particularly large benefit or a particularly large harm (i.e., treatment-by-factor interactions). Unfortunately, the power to detect interactions is typically low, and there is a belief that subgroup analyses are far more likely to lead to false positive findings than to identify true interactions. However, it is important to note that the presence of an interaction depends on the scale on which the treatment effect is measured, and this has important implications for the assessment of benefit-risk in subgroups. Notably, when there is reason to believe that the relative effects of a treatment (e.g., the hazard ratio for a time-to-event endpoint) are consistent across subgroups, the absolute effects (e.g., the absolute risk reductions) are often highly variable. This is because there are often subgroups with greater and lesser disease severity, where event rates vary considerably; in this case, a constant hazard ratio across subgroups will lead to highly variable absolute risk reductions. In this chapter we argue that benefit-risk conclusions should be based on the absolute effects for benefits and harms, and, therefore, that subgroup analyses play a particularly important role.

Session 47: Joint Model and Applications

Joint inference of GLMM and NLME with Informative Censoring with Application in HIV/AIDS

♦*Hongbin Zhang*[1] *and Lang Wu*[2]

[1]City University of New York

[2]University of British Columbia

hongbin.zhang@sph.cuny.edu

In an HIV/AIDS study, we are often interested in the dynamics of viral load and CD4 counts over time during an antiretroviral treatment with the primary goal to understand their relationship and the interplay of a treatment option. Statistical analyses are challenging due to the fact that the viral load observations typically are left censored and measured with error. We propose a joint model for CD4 counts and viral load in which a non-linear mixed effects model (NLME) is used for the mis-measured viral load (as covariate process) and a generalized linear mixed effects model (GLMM) is used for the CD4 count (as response process) conditional on the true trajectory of the covariate process. Model parameters are jointly estimated by a Monte Carlo EM algorithm (MCEM) with Gibbs sampling. We compare the performance of our method to existing simpler methods via simulation and give an example based on a real data in which the methods do not yield the same inference on the direct (i.e., not mediated by viral load) effect of HIV antiretroviral treatment. Our simulation results suggest that our method leads to estimators with the least bias, more honest assessments of estimate uncertainty, and more accurate coverage rates.

Joint-modelling of discrete, continuous and semi-continuous data with applications

Renjun Ma

University of New Brunswick, Canada

renjun@unb.ca

Discrete, continuous and semi-continuous data are often clustered by subject and in medical studies. We propose to model these data of mixed types jointly using Tweedie models with distribution-free random effects. An orthodox best linear unbiased predictor approach has been developed in the estimation of our model. This approach is optimal for the regression parameters in the sense of Godambe. Our method is illustrated with application to medical data.

Two-Step and Likelihood Methods for HIV Viral Dynamic Models with Covariate Measurement Errors

♦*WEI LIU*[1] *and LANG WU*[2]

[1]YOKR UNIVERSITY

[2]UNIVERSITY OF BRITISH COLUMNIA

liuwei@mathstat.yorku.ca

HIV viral dynamic models have received much attention in the literature. Long-term viral dynamics may be modeled by semiparametric nonlinear mixed-effects (NLME) models, which incorporate large variation between subjects and auto-correlation within subjects and are flexible in modeling complex viral load trajectories. Time-dependent covariates may be introduced in the dynamic models to partially explain the between individual variations. In the presence of measurement errors and missing data in time-dependent covariates, we show that the commonly used two-step method may give approximately unbiased estimates but may under-estimate standard errors. We propose a two-stage bootstrap method to adjust the standard errors in the two-step method and a likelihood method.

Dynamic modeling and inference for event detection

Hongyu Miao

School of Public Health, UTHealth

Hongyu.Miao@uth.tmc.edu

Statistical inference of high-dimensional dynamic systems is a challenging task, mainly due to the curse of dimensionality. Here we consider gene regulatory networks as an example to detect significant regulatory interaction events, where the network size is on the order of 104 and the number of parameters is on the order of millions. While a few previous computational studies have claimed success in revealing genome-wide regulatory landscapes from temporal gene expression data, recent work suggests that these methods still suffer from the curse of dimensionality as network size increases to 100 or higher. Here we present a novel scalable algorithm for identifying genome-wide regulatory network structures. The highlight of our method is that its superior performance in fair comparison with other state-of-the-art approaches does not degenerate even for network size O(10^4), and is thus readily applicable to large-scale complex networks.

Session 48: Recent Development in Functional Data Analysis and Applications

Dynamic Functional Mixed Models for Child Growth

Andrew Leroux[1], ♦*Luo Xiao*[2], *Will Checkley*[1] *and Ciprian Crainiceanu*[1]
[1]Johns Hopkins University
[2]North Carolina State University
lxiao5@ncsu.edu

We propose a dynamic functional mixed effects model to predict the length of a child based on the past observations of length and weight and other baseline covariates. We model the effect of weight via a novel lagged time varying coefficient and develop reproduceable methods implemented in R. We conduct a systematic study comparing our approach with standard fixed effects models and nonlinear mixed effects models. We show that the dynamic functional model provides more accurate estimation and inference for the fixed effects as well as improved subject-level predictions.

Nested Hierarchical Functional Data Modeling for Root Gravitropism Data

Yehua Li
Iowa State University
yehuali@iastate.edu

In a Root Image Study in plant science, the rooting processes of seeds from various genotypes are recorded using digital cameras. The angular velocity of the root tip changes over time and hence modeled as functional data. Multiple seeds from the same genotypes are recorded using the same protocol under different camera setups. The data are collected from a large number of genotypes and have a natural genotype – camera file – seed, three-level nested hierarchical structure. The seeds are planted on different lunar days and an important scientific question is whether the moon phase has any effect on seed rooting. We allow the mean function of the angular velocity process to be dependent on lunar day and model the variation between genotypes, files and seeds respectively by hierarchical functional random effects with Karhunen-Lòve expansions. We estimate the covariance function of the functional random effects by a fast penalized tensor product spline approach, perform multi-level functional principal component analysis (FPCA) using the best linear unbiased predictor of the principal component scores, and improve the efficiency of mean estimation by iterative decorrelation. We choose the number of principal components using a conditional Akaike Information Criterion and test the lunar day effect using a generalized likelihood ratio (GLR) test statistic. We use a simple permutation procedure to evaluate the null distribution of the test statistic. Our simulation studies show that our estimation procedure and model selection criterion work well, the permutation based GLR tests based on FPCA enjoy the Wilks property and have better power than tests based on working independence test statistics. We also find significant moon phase effects in our motivating data.

Functional and imaging data in precision medicine

♦*Todd Ogden*[1], *Adam Ciarleglio*[2], *Thaddeus Tarpey*[3] *and Eva Petkova*[2]
[1]Columbia University
[2]New York University
[3]Wright State University
to166@columbia.edu

One goal in precision medicine is to make optimal patient-specific treatment decisions using available baseline data. One important application involves complex data in the treatment of neuropsychiatric disorders, including brain imaging or other functional data. Such data which can be incorporated into regression models of treatment outcome using functional data analytic techniques. Thus, obtaining useful and interpretable fits in such a situation generally involves some sort of dimension reduction or penalization, taking into account the particular (spatial) structure of the data. This talk will introduce the problems and describe some basic strategies that have proven effective in this context.

A New Method on Flexible Combination of Multiple Diagnostic Biomarkers

♦*Tu Xu*[1], *Junhui Wang*[2], *Yixin Fang*[3] *and Alan Rong*[4]
[1]Abbvie Inc.
[2]City University of Hong Kong
[3]New York University
[4]Astellas Pharma
tu.xu@abbvie.com

In medical research, it is common to collect information of multiple biomarkers to improve the accuracy of diagnostic tests. To evaluate the accuracy of a diagnostic test, the Youden index has been widely used in literature. Combining the measurements of these biomarkers into one single score is a popular practice to integrate the collected information, where the accuracy of the resultant diagnostic test is usually improved. Various parametric and nonparametric methods have been proposed to linearly combine biomarkers so that the corresponding Youden index can be optimized. Yet there seems to be little justification of enforcing such a linear combination. In this talk, we introduce a flexible approach that allows both linear and nonlinear combinations of biomarkers. The proposed approach formulates the problem in a large margin classification framework, where the combination function is embedded in a flexible reproducing kernel Hilbert space. Advantages of the proposed approach are demonstrated in a variety of simulated experiments as well as a real application to a liver disorder study.

Session 49: Recent Development of Bayesian High Dimensional Modeling, Inference and Computation

A new double empirical Bayes approach for high-dimensional problems

Ryan Martin
University of Illinois at Chicago
rgmartin@uic.edu

In high-dimensional problems, selecting a good prior—one that leads to a posterior with optimal concentration properties and efficient computation—can be a serious challenge. In this talk I will present a new kind of empirical Bayes approach that uses data in the prior in two ways: first, the prior is suitably centered on the data, and second, a regularization step is taken to prevent the greedy centering from driving the behavior of the posterior. In the context of a sparse high-dimensional linear model, a variety of posterior concentration results will be presented, along with simulation results that demonstrate the method's excellent performance. Extensions to other high-dimensional models, as well as nonparametric problems, will also be discussed.

NEARLY OPTIMAL BAYESIAN SHRINKAGE FOR HIGH DIMENSIONAL REGRESSION

♦*Qifan Song*[1] *and Faming Liang*[2]
[1]Purdue Univ.
[2]Univ. of Flordia
qfsong@purdue.edu

During the past decade, shrinkage priors have received much attention in Bayesian analysis of high-dimensional data. [2] established the posterior consistency for linear regression with shrinkage priors under the low-dimensional setting. In this paper, we study the problem under the high-dimensional setting. We show that if the shrinkage prior is heavy-tailed and allocates a sufficiently large probability mass in a very small neighborhood of zero, then the posterior consistency holds. While enjoying its advantages in the resulting posterior simulations, the shrinkage prior can lead to almost the same posterior contraction rate as the point-mass prior for recovering the true model. Our numerical results show that under posterior consistency, Bayesian methods can yield much better results in variable selection than the regularization methods, such as Lasso and SCAD. In addition, we study the asymptotic shape of the posterior distribution, and our result leads to a convenient way to quantify uncertainties of the regression coefficient estimates, which has been beyond the ability of regularization methods. As a by-product, we show that the Laplace prior can never lead to posterior consistency in the L1 -norm, which provides a theoretical explanation for the over-shrinkage phenomenon of Bayesian Lasso.

Scalable Bayesian Variable Selection for Structured High-dimensional Data

Changgee Chang, Suprateek Kundu and ♦*Qi Long*
Emory University
qlong@emory.edu

Variable selection for structured covariates lying on an underlying known graph is a problem motivated by practical applications, and has been a topic of increasing interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies. We propose an adaptive Bayesian shrinkage approach which incorporates prior network information by smoothing the shrinkage parameters for connected variables in the graph, so that the corresponding coefficients have a similar degree of shrinkage. We fit our model via a computationally efficient expectation maximization algorithm which scalable to high dimensional settings (p 100,000). Theoretical properties for fixed as well as increasing dimensions are established, even when the number of variables increases faster than the sample size. We demonstrate the advantages of our approach in terms of variable selection, prediction, and computational scalability via a simulation study, and apply the method to a cancer genomics study.

Prediction risk for global-local shrinkage regression

♦*Anindya Bhadra*[1]*, Jyotishka Datta*[2]*, Yunfan Li*[1]*, Nicholas Polson*[3] *and Brandon Willard*[3]
[1]Purdue University
[2]Duke University
[3]University of Chicago
bhadra@purdue.edu

Prediction performance in shrinkage regression suffers from two major difficulties: (i) the amount of relative shrinkage is monotone in the singular values of the design matrix and (ii) the amount of shrinkage does not depend on the response variables. Both of these factors can translate to a poor prediction performance, the risk of which can be explicitly quantified using Stein's unbiased risk estimate. We show that using a component-specific local shrinkage term that can be learned from the data under a suitable prior, in combination with a global shrinkage term, can alleviate both these difficulties and consequently, can result in an improved risk for prediction. Demonstration of improved prediction performance over competing approaches in a simulation study and in a real data set confirms the theoretical findings.

Session 50: On Clinical Trials with a High Placebo Response

Placebo Effects and Sequential Parallel Comparison (SPC) Design

Eiji Ishida
FDA/CDER/OTS/OB/DBI
eiji.ishida@fda.hhs.gov

A substantially high placebo response in a conventional parallel group placebo-controlled study may generate difficulty to yield efficacy evidence when the new treatment is effective. Perhaps because of this reason, there is growing interest in an SPC design that allows to incorporate placebo non-responders into an evaluation of efficacy of a new treatment. Despite the potential usefulness of this design, there seems to be continuing discussions on what statistical method is appropriate for analyzing trial data based on the SPC design. This presentation will review some backgrounds of the SPC design, briefly introduce some of the proposed statistical methodologies, and provide a generic way of evaluating trial efficacy data with the SPC design using simulated data.

An Unbiased Estimator of the Two-Period Treatment Effect in Doubly Randomized Delayed-Start Designs

♦*Yihan Li*[1]*, Yanning Liu*[2]*, Qing Liu*[2] *and Pilar Lim*[2]
[1]AbbVie
[2]Janssen Research and Development
yihan626@gmail.com

High placebo response rate is a common phenomenon in psychiatric clinical trials and is well known to be a major source of bias. Special trial designs, such as the sequential parallel design (SPD) by Fava et al (2003), the SPD with re-randomization (SPD-ReR) by Chen et al (2011), and the doubly randomized delayed-start design (DRDS) by Liu et al (2010, 2012), have been proposed to tackle this problem. These designs consist of two periods that aim at enriching the placebo non-responder population in the second period. The treatment effect over the duration of one period is then summarized by combining information from the first and second period. However, it is of interest to look at the treatment effect over the duration of two periods as well. Due to enrichment in the second period, a simple estimator of the treatment effect comparing subjects who stayed on active treatment for both periods vs. those on placebo for both periods would be biased. In this talk we will propose an unbiased estimator of the two-period treatment effect, based on the DRDS design.

Accounting for High Placebo Response Rates in Clinical Trials

Pilar Lim, ♦*Akiko Okamoto and George Chi*
Janssen Research & Development
aokamot@its.jnj.com

The basic reason for the failure of many standard randomized parallel placebo-controlled clinical trials with high placebo response rate is that the observed relative treatment difference only provides an estimate of the apparent treatment effect since the true treatment effect has been diminished by the presence of a substantial proportion of placebo responders in the population. Analogous to an active control trial, the true treatment effect cannot be measured by the relative treatment difference. An appropriate assessment of the true treatment effect is critical for making a risk/benefit analysis and

dosage recommendation. The primary purpose of this talk is to propose a method for adjusting the apparent treatment effect to account for the high placebo response rate within the framework of a doubly randomized delayed start design.

Session 51: Statistical Learning Methods

High-Dimensional Hypothesis Testing With the Lasso

♦*Sen Zhao, Ali Shojaie and Daniela Witten*
University of Washington
senz@uw.edu

We consider the problem of hypothesis testing in a high-dimensional linear model using the lasso. We show that, under some standard assumptions, the set of variables selected by the lasso is almost surely fixed, and contains all of the variables that have non-zero regression coefficients. These theoretical results are applied in order to justify restricting our attention to the set of features selected by the lasso. We then apply classical Wald and score tests on the reduced data set. Because the lasso-selected set is almost surely fixed, distribution truncation is not required in order to obtain asymptotically valid inference on the population regression coefficients; this is in contrast to the recently-proposed exact post-selection testing framework. We also establish connections between our proposals and the debiased lasso tests and investigate their differences. Finally, we perform extensive numerical studies in support of our methods.

A simultaneous variable selection and clustering method for high-dimensional multinomial regression

♦*Sheng Ren*[1]*, Emily Kang*[1] *and Jason Lu*[2]
[1]University of Cincinnati
[2]Cincinnati Children's Hospital Research Foundation
rensg@mail.uc.edu

We propose a new data-driven simultaneous variable selection and clustering method for high-dimensional multinomial regression. Unlike other grouping pursuit methods, for example regression with Graph Laplacian penalty, our method does not assume that moderate to highly correlated variables have similar regression coefficients or should belong to same clusters. Relaxing this assumption is practically meaningful when we have a multinomial response variable. For example, moderate to highly correlated expressed genes may associate with different subtypes of a disease. We propose a penalty function taking both regression coefficients and pairwise correlation into account for defining variables clusters. An algorithm with respect to this new penalty function is also developed, incorporating both convex optimization and clustering. We demonstrate the performance of our method via a simulation study and compare it with some other methods, showing that our method is able to yield correct variable clustering and to improve prediction performance. A real data example will also be presented.

Lasso-type Network Community Detection within Latent Space

♦*Shiwen Shen and Edsel Pena*
University of South Carolina
sshen@email.sc.edu

Community detection is one of the fundamental topics in the statistical network analysis. Many methods have been proposed in the literature to solve this problem, however, approaches are not feasible for large size network data. In this paper, we propose a method to shrink the distances among nodes inside a network using the lasso penalty, after projecting observed nodes into a prespecified latent space. We use the alternating direction method of multipliers (ADMM) algorithm in the estimation procedure to make the model be friendly to large size data. Simulation results and an example using open sourced data are provided in detail.

Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data

♦*Chiyu Gu*[1]*, Subharup Guha*[1] *and Veera Baladandayuthapani*[2]
[1]University of Missouri
[2]MD Anderson Cancer Center
cgz59@mail.missouri.edu

Omics datasets, which involve intrinsically different sizes and scales of high-throughput data, offer genome-wide, high-resolution information about the biology of lung cancer. One of the main goals of analyzing these data is to identify differential genomic signatures among samples under different treatments or biological conditions. e.g., treatment arms, tumor (sub)types, or cancer stages. We construct an encompassing class of nonparametric models called PDP-Seq that are applicable to mixed, heterogeneously scaled datasets. Each platform can choose from diverse parametric and nonparametric models including finite mixture models, finite and infinite hidden Markov models, Dirichlet processes, and zero and first order PDPs, which incorporate a wide range of data correlation structures. Simulation studies demonstrate that PDP-Seq outperforms many existing techniques in terms of accuracy of genomic signature identification. The pathway analysis identifies upstream regulators of many genes that are common genetic markers in multiple tumor cells.

Sparsity and Error Analysis of Empirical Feature-Based Regularization Schemes

♦*Xin Guo*[1]*, Jun Fan*[2] *and Ding-Xuan Zhou*[3]
[1]The Hong Kong Polytechnic University
[2]University of Wisconsin-Madison
[3]City University of Hong Kong
x.guo@polyu.edu.hk

We consider a learning algorithm generated by a regularization scheme with a concave regularizer for the purpose of achieving sparsity and good learning rates in a least squares regression setting. The regularization is induced for linear combinations of empirical features, constructed in the literatures of kernel principal component analysis and kernel projection machines, based on kernels and samples. In addition to the separability of the involved optimization problem caused by the empirical features, we carry out sparsity and error analysis, giving bounds in the norm of the reproducing kernel Hilbert space, based on a priori conditions which do not require assumptions on sparsity in terms of any basis or system. In particular, we show that as the concave exponent q of the concave regularizer increases to 1, the learning ability of the algorithm improves. Some numerical simulations for both artificial and real MHC-peptide binding data involving the ℓ^q regularizer and the SCAD penalty are presented to demonstrate the sparsity and error analysis.

Compressed Covariance Matrix Estimation With Automated Dimension Learning

♦*Gautam Sabnis*[1]*, Debdeep Pati*[2] *and Anirban Bhattacharya*[3]
[1]Florida State University
[2]Florida State University
[3]Texas A&M University
gss12b@my.fsu.edu

Constant technological advances have dramatically increased, in contemporary times, our ability to collect and store colossal amounts of data. From high-frequency finance and marketing datasets to genomic data arising from medical studies, we are faced

with problems of increasing dimension. In general, the focus of multivariate data analysis is to elucidate the underlying dependence among variables where the increasing dimension has deteriorating effect on the precision of inference of objects of interest. Assuming an inherent low dimensional structure, motivated by various scientific applications, is often adopted as the guiding light in the analysis. The existing low-rank estimation approaches via factor models and the probabilistic principal component analysis (PPCA) have proven to be successful in applications where the observations are concentrated around a low dimensional surface but a principled framework for choosing the structure dimension is lacking. We propose a new approach for high dimensional covariance matrix estimation based on a "compression-decompression" (C-D) mechanism. There are several lines of work that focus on performing high dimensional regression and covariance estimation using partial information embedded in the compressed measurements. Our work deviates from the majority of the work on compressive estimation in terms of: (i) a concrete principled way for choosing the random compressed dimension using Stein's Unbiased Risk Estimation (SURE) theory; (ii) computational efficiency and high scalability to massive covariance matrices.The (C-D) mechanism proceeds by projecting the high dimensional observations to a lower dimension to form compressed measurements and then decompressing them back to the original dimension. Our main idea is to use the sample covariance matrix of the compressed-decompressed data as an estimator of Σ instead of the sample covariance matrix $\hat{\Sigma}$. This is an entirely new way of regularization compared to l_q type penalties on various functionals of the covariance. The C-D estimator is integrated over all possible random projections to form the final estimator of the covariance matrix. Experimental simulation results demonstrate the efficacy of our method when the underlying observations are generated from an approximate low-dimensional structure.

Session 52: Novel Design and/or Analysis for Phase 2 Dose Ranging Studies

Role of Biomarkers and Quantitative Models in Dose Ranging Trials

Yaning Wang
FDA
yaning.wang@fda.hhs.gov

Dose selection is a challenging task in drug development. Since it is too costly to study multiple doses based on the long term efficacy endpoints that are used in phase 3 confirmative trials, it is quite common to rely on efficacy related biomarkers in dose ranging trial to select the optimal dose for further study in the confirmative trials. The traditional pair-wise comparison among doses requires a large sample size to derive the optimal dose. Instead, model-based analysis provides a more efficient method to analyze the data from the dose ranging trials. A case study will be provided to demonstrate how biomarkers and quantitative pharmacokinetic/pharmacodynamics models can be used to justify dose selection in drug development.

Achieving multiple objectives in a Phase 2 dose-ranging study

Ming-Dauh Wang
Eli Lilly and Company
md_wang@lilly.com

We present the design of a Phase 2 trial, where multiple objectives were critical for the end of Phase 2 decision. In addition to demonstration of the efficacy of the drug in the target disease population, a potential Phase 3 program would need to include trials for certain subpopulations both for a successful submission for approval and for the reason of commercial viability. Concerning selection of a dosing regimen in Phase 3, properties of the dose response over time, particularly the stability of the magnitude of response, were also of moment to be addressed in Phase 2. Besides, evidence of efficacy in various global regions would add to a positive end of Phase 2 decision. More advanced development programs of other drugs in this class all conducted multiple Phase 2 studies to address the above stated needs. Intended to narrow the development lag with the other programs, the Phase 2 study adopted the "one shot for all" approach, pursuing to illuminate all the above issues all at once. Statistical considerations for optimization of the design of such an "all-encompassing" Phase 2 trial, power calculation in particular, will be discussed. Final analysis of the trial to prepare the Phase 3 program for post-launch competition will be reviewed.

Optimizing Oncology Combination Therapy by Integrating Multiple Information into Dose Escalation

Xiaowei Guan
Bristol Myers Squibb
xiaowei.guan@bms.com

In oncology dose selection, the challenges are many: while keeping patients safe, the clinical trials should be small, adaptive and enable a quick assessment of which dose(s) are acceptably safe. Ideally, dose limiting events (DLTs) and responses (such as tumor reduction and pharmacodynamic (PD) endpoints) indicative of efficacy, as well as pharmacokinetic (PK) parameters should be considered in the dose escalation procedure, to balance clinical, statistical and operational aspects in a cost-effective and patient-sparing way. Since toxicity, PK and efficacy are available at different times, it is challenging to integrate different layers of clinical information quantitatively, especially in a dual-agent setting. Optimizing the totality of all available clinical information plays an important role in key clinical study design elements such as schedule, escalation strategy, targeted patient population, etc.

An outcome-adaptive Bayesian design which integrates multi-layer information is proposed to enable real-time adaptation in the combination setting. This approach allows for more effective optimization based on the totality of data to balance speed, reliability, efficacy and safety. Such efficient methods can accelerate the drug development processes and ultimately bring effective medicines to patients sooner.

Session 53: Lifetime Data Analysis

The effect of the timing of treatment confounded by indication and screening on cancer mortality

Alex Tsodikov
University of Michigan
tsodikov@umich.edu

In latent diseases such as cancer treatment occurs when the disease is diagnosed at a random stopping point in its stochastic process of progression. Screening procedures using biomarker-based diagnostic tests allow the disease to be detected and treated earlier with a different treatment that is appropriate for the earlier stage of the disease measured at diagnosis. To analyze whether screening is of benefit for cancer-specific mortality one has to take into account that treatment in such a scenario is dynamically confounded by the diagnostic process and the information about the disease that it reveals. Modeling approaches to the problem are discussed and a

model-based test for mortality benefit is proposed. The methods are illustrated by the analysis of screening trials.

Dynamic risk prediction models for data with competing risks

♦*Chung-Chou Chang*[1] *and Qing Liu*[2]
[1]University of Pittsburgh
[2]Novartis Pharmaceutical Corporation
changj@pitt.edu

Risk prediction via cause-specific cumulative incidence function (CIF) is of primary interest in medical decision making. To predict CIF more accurately during the disease course of a patient it will require researchers to dynamically incorporate historical and up-to-date time-dependent prognostic information into the estimation. For this purpose, we propose a landmark proportional subdistribution hazards (PSH) model and a more comprehensive landmark PSH supermodel as described in this study.

Our proposed models have several advantages over the currently available prediction models in addressing competing risks. First, our models are robust against violations of the PSH assumption and can directly predict the conditional CIFs bypassing the estimation of overall survival thus greatly simplify the prediction procedure. Second, our models can incorporate various types of time-dependent information including longitudinally repeatedly measured biomarkers, intermediate clinical events, and time-varying covariate effects. Third, our landmark PSH supermodel allows researchers to make predictions with a set of landmark points in only one step. Finally, our models are not computationally intensive and can be easily implemented using commercial statistical software packages.

The performance of our models was assessed via simulations and through analysis of data from a multicenter clinical trial for patients with estrogen-receptor-positive and node-negative breast cancer.

Modeling Gap Times in Panel Count Data with Informmative Observation Times: Assessing Spontaneous Labor in Women

♦*Rajeshwari Sundaram and Ling Ma*
Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH
sundaramr2@mail.nih.gov

Defining labor progression in women has been a long-standing challenge for obstetricians. Cervical dilation, as integer-valued measurement, is a key indicator of labor progression. Assessing the distribution of the time to per-unit increments of cervical dilation is of considerable interest in aiding obstetricians with better management of labor. Given that women are observed only intermittently for cervical dilation after they get admitted to hospital and that the observation frequency is very likely correlated to how fast/slow she dilates, one could view such data as panel count data with informative observation times and unknown time-zero. In this paper, we propose semiparametric proportional rate models for the cervical dilation process and observation process with a multiplicative subject-specific frailty variable capturing the correlation between the two processes. Inference procedures for the gap times between consecutive events are proposed for both the scenarios with known and unknown time-zero using maximum likelihood approach and estimating equations. The methodology is assessed through simulation study and the large sample property is established. A detailed analysis using the proposed methods is applied to the longitudinal cervical dilation data from the National Collaborative Perinatal Project as well as Consortium of Safe Labor.

Integrated analysis of multidimensional omics data for prognosis

Shuangge Ma
Yale University
shuangge.ma@yale.edu

For complex diseases, prognosis is of essential interest. Multiple types of omics measurements, including mRNA gene expression, methylation, copy number variation, SNP, and others, can have important implications for prognosis. The analysis of multidimensional omics data is challenging because of the high data dimensionality and, more importantly, because of the interconnections between different units of the same type of measurement and between different types of omics measurements. In our study, we have developed novel regularization-based methods, effectively integrated multidimensional data, and constructed prognosis models. Analyzing the TCGA data on several cancer types using the proposed method leads to biologically interpretable models with superior prognostic performance.

Session 54: Innovative Methods for Clinical Trial Designs

Generalized Efron's Biased Coin Design and its Theoretical Properties

YANQING HU
West Virginia University and Incyte
huyanqing135@gmail.com

In clinical trials with two treatment arms, Efron's biased coin design sequentially assigns a patient to the underrepresented arm with probability p¿0.5. Under this design the proportion of patients in any arm converges to 0.5, and the convergence rate is higher than under some other popular designs. The generalization of Efron's design to K¿2 arms and an unequal target allocation ratio (q1, ..., qK) can be found in some papers, most of which determine the allocation probabilities p's in a heuristic way. Nonetheless, it has been noted that by using inappropriate p's, the proportions of patients in the K arms can never converge to the target ratio. We develop general theory to answer the question what allocation probabilities ensure that the realized proportions under a generalized design converge to the target ratio (q1, ..., qK).

Efficient Algorithms for Extended Two-stage Adaptive Designs for Phase II Clinical Trials

Seongho Kim[1] *and* ♦*Weng Kee Wong*[2]
[1]Karmanos Cancer Institute, School of Medicine, Wayne State University
[2]UCLA
wkwong@ucla.edu

We develop a nature-inspired metaheuristic algorithm and call it discrete particle swarm optimization (DPSO) to find extended two-stage adaptive optimal designs for phase II trials with many parameters. These designs include Simon's design (1989) and those proposed by Lin and Shih (2004) as special cases. We show that DPSO not only frequently outperforms greedy algorithms, which are currently used to find such designs when there are only a few parameters; it is also capable of effectively solving adaptive design problems with many parameters that greedy algorithms cannot. In particular, we consider situations where a treatment seems promising in stage 1 but there is great uncertainty in its efficacy rate, and both drug development cost and ethics dictate that there be three pre-determined user-specified efficiency rates for possible testing at stage 2 given testing error rate constraints. Using a real adaptive trial for melanoma patients from the literature, we show one of the advantages of our extended 2-stage adaptive design strategy is that it can reduce the sample size by one-half compared with the one implemented.

Optimal Sequential Enrichment Designs for Phase II Clinical Trials

Yong Zang[1] *and* ♦*Ying Yuan*[2]
[1]Florida Atlantic University
[2]MD Anderson Cancer Center
yyuan@mdanderson.org

In the early phase development of molecularly targeted agents (MTAs), a commonly encountered situation is that the MTA is expected to be more effective for a certain biomarker subgroup, say marker-positive patients, but there is no adequate evidence to show that the MTA does not work for the other subgroup, i.e., marker-negative patients. After establishing that marker-positive patients benefit from the treatment, it is often of great clinical interest to determine whether the treatment benefit extends to marker-negative patients. The authors propose optimal sequential enrichment (OSE) designs to address this practical issue in the context of phase II clinical trials. The OSE designs evaluate the treatment effect first in marker-positive patients and then in marker-negative patients if needed. The designs are optimal in the sense that they minimize the expected sample size or the maximum sample size under the null hypothesis that the MTA is futile. An efficient, accurate optimization algorithm is proposed to find the optimal design parameters. One important advantage of the OSE design is that the go/no-go interim decision rules are specified prior to the trial conduct, which makes the design particularly easy to use in practice. A simulation study shows that the OSE designs perform well and are ethically more desirable than the commonly used marker-stratified design. The OSE design is applied to an endometrial carcinoma trial.

New covariate-adjusted response-adaptive designs for precision medicine

♦*Feifang Hu, Fan Wang and Wanying Zhao*
George Washington University
feifang@gwu.edu

Precision medicine is the systematic use of information pertaining to an individual patient to select or optimize that patient's preventative and therapeutic care. With today's modern technology and big data, it is much easier to identify important biomarkers that may associate with certain diseases and their treatments. To design a clinical trial for precision medicine, one should include these important biomarkers. In this talk, we propose a new family of covariate-adjusted response-adaptive designs, which incorporate these biomarkers as well as the responses. Some properties of the new designs are discussed.

Session 55: Recent Advances in Statistical Methods for Alzheimer's Disease Studies

Alzheimer's Disease Neuroimaging Initiative: statistical challenges and solutions

♦*Sharon Xie*[1]*, Matthew White*[2] *and Jarcy Zee*[3]
[1]University of Pennsylvania
[2]Boston Children's Hospital
[3]Arbor Research Collaborative for Health
sxie@mail.med.upenn.edu

Alzheimer's Disease (AD) Neuroimaging Initiative (ADNI) is a high profile multi-center study aiming to test whether serial magnetic resonance imaging, positron emission tomography, cerebral spinal fluid (CSF) biomarkers, other biomarkers, as well as clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early AD. In this talk, we highlight some statistical challenges in analyzing ADNI data. We mainly focus on the challenge when CSF sample is only available for a subset of the ADNI participants. The statistical power is limited when we examine time to significant disease progression featured by changes of CSF biomarkers if we can only use participants with the CSF samples. We present a new nonparametric approach to estimating survival distributions of time to abnormal CSF biomarkers by augmenting the missing CSF outcomes through the clinical diagnosis information. We demonstrate that the proposed estimator has little bias compared to the naive Kaplan-Meier survival function estimator, which uses only the clinical diagnosis, and more efficient with moderate missingness compared to the complete-case Kaplan-Meier survival function estimator, which uses only available CSF outcomes.

Adjusting for dependent truncation with inverse probability weighting

♦*Jing Qian*[1] *and Rebecca Betensky*[2]
[1]University of Massachusetts Amherst
[2]Harvard University
qian@schoolph.umass.edu

Many clinical trials and observational studies are conducted under complex sampling involving truncation. Ignoring the issue of truncation or incorrectly assuming quasi-independence can lead to bias and incorrect results. Currently available approaches for dependently truncated data are sparse. We present an inverse probability weighting method for estimating the survival function of a failure time subject to left truncation and right censoring. The proposed method allows adjusting for informative truncation due to variables affecting both event time and truncation time. Simulation studies show that the proposed method performs well in finite sample. We apply the proposed method to an Alzheimer's disease study.

Capturing Change in Cognition Using Data from Two Cohort Studies of Aging and Dementia

Lei Yu
Rush Alzheimer's Disease Center
lei_yu@rush.edu

Alzheimer's disease (AD) is the result of a sequence of pathophysiological events including Aβ deposition, synaptic dysfunction, tangle formation, and other biochemical and structural changes. Longitudinal trajectories of decline in cognition provide objective evidence about how the disease progresses over time and hold great promise to identify risk factors of AD that are not easily captured in conventional case-control studies. We leverage continuous longitudinal cognitive data from annual assessments up to 20 years from two cohort studies of aging and dementia, and characterize the profiles of late life cognitive decline. Different modeling techniques are pursued. First, we employ a random change point model to fit a piecewise linear trajectory with change point signaling the acceleration of cognitive decline; this approach permits examination of the differential effects of risk factors at different phases of the trajectory. Second, we apply a random effects mixture model to capture the heterogeneity of person-specific changes by identifying latent classes with distinct cognitive profiles; this allows for examination of the associations of risk factors with the classification. Third, we fit a continuous-time hidden Markov model to examine the impact of common brain pathologies on transitions from no cognitive impairment to mild impairment and finally dementia. These results suggest that (1) the trajectories of cognitive decline are most likely nonlinear; (2) there is considerable variability in cognitive trajectories that is not random, but rather is related to the burden of neu-

ropathology, as well as markers of cognitive and neural reserve, and (3) AD pathology alone nearly doubles the risk of developing cognitive impairment in late life, and comorbid pathologies further increase such risk.

A BAYESIAN FUNCTIONAL LINEAR COX REGRESSION MODEL IN ALZHEIMER'S Disease study

♦*Eunjee Lee*[1]*, Hongtu Zhu*[1]*, Dehan Kong*[1]*, Yalin Wang*[2]*, Kelly Sullivan Giovanello*[1] *and Joseph G Ibrahim*[1]
[1]University of North Carolina at Chapel Hill
[2]ARIZONA STATE UNIVERSITY
eunjee2@gmail.com

This talk discusses a Bayesian functional linear Cox regression model (BFLCRM) with both functional and scalar covariates. This new development is motivated by establishing the likelihood of conversion to Alzheimer's disease (AD) within mild cognitive impairment (MCI) patients. The aim of study is development of Bayesian survival models to quantify the effects of hippocampal morphology on the time to conversion from MCI to AD. Since hippocampal morphology takes the form of functional data, many classical survival models are not theoretically and computationally suitable. We propose the BFLCRM to incorporate functional covariates in a Bayesian survival model by applying functional principal component analysis (fPCA). Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. The BFLCRM was used to establish that functional covariates including hippocampus surface morphology and scalar covariates including brain MRI volumes, cognitive performance (ADAS-Cog), and APOE-ϵ4 status could accurately predict time to onset of AD. A simulation study was performed to evaluate the finite sample performance of the BFLCRM.

Session 56: High Dimensional Model and Prediction

Comparison of common variable selection approaches for accuracy and predictive values involving corr

♦*Wei-Ting Hwang, Rengyi (Emily) Xu, Clementina Mesaros and Ian Blair*
University of Pennsylvania
whwang@mail.med.upenn.edu

(Abstract Title) Comparison of common variable selection approaches for accuracy and predictive values involving correlated high-dimensional biomarker data using logistic regression models

Many variable selection methods can be applied to identify marker or marker combination in predicting outcomes such as disease status or treatment response. Evaluation of marker signature composition and its predictive values are essential. When the number of candidate biomarkers is large, regularization approaches such as LASSO, elastic-net or their extensions are increasing popular as a tool for variable selection. But it is unclear whether this type of method outperforms other simpler or traditional methods such as stepwise selection that based on AIC or uses principle components when the correlations between candidate markers are high to extremely high. In this project, we conducted a series of simulation study to understand the impact of the sample size, number of candidate biomarkers, presence of confounders, correlation among predictors and other, etc. on signature composition and model performance. The presented work will focus on binary disease status as response variable and continuous marker candidates as predictors through logistic regression. Illustration using a real world data for mesothelioma biomarker discovery will be presented.

Integration of high-dimensional genomic and imaging features for risk prediction of lung cancers

Fenghai Duan
Brown University
fduan@stat.brown.edu

Introduction: High-dimensional genomics, genetics and proteomics techniques have been widely used in cancer research for over two decades. Correspondingly, various genomic, genetic and proteomic signatures have been discovered in the cancer's diagnosis, prognosis and prediction. In recent years, large amount of quantitative imaging features extracted from non-invasive medical imaging (e.g., CT, PET, MRI) have been related to tumor phenotype. The aim of this talk is to explore various methods of integrating these two types of data in predicting lung cancer risk.

Methods: The National Lung Screening Trial (NLST) was a randomized screening trial that accrued over 53,000 older smokers to compare low-dose helical computed tomography (CT) relative to chest-x-ray screening in reducing lung cancer mortality. Half of accrued participants (about 26,000) underwent at least one CT screen. In addition, about 10,000 participants consented to have their specimens collected for the development of the NLST biorepository for lung cancer biomarker validation research. In this study, we used the imaging features extracted from the NLST CT-armed participants and the genetic signatures developed from the repository specimens to explore the optimum methods for integration of these biomarkers in the prediction of lung cancer risk.

Results and conclusions: Imaging and genetic signatures are developed to better understand tumor heterogeneity and predict risk of lung cancer diagnosis. The optimum integration of these events can further improve the screening performance and provide new insight into the potential of personalized targeted therapy.

A statistical framework for eQTL mapping with imprinting effect detection using RNA-seq data

♦*Feifei Xiao*[1]*, Guoshuai Cai*[2]*, Jianzhong Ma*[3] *and Chirstopher Amos*[2]
[1]University of South Carolina
[2]Dartmouth College
[3]University of Texas Health Science Center
xiaof@mailbox.sc.edu

Genomic imprinting is an important epigenetic phenomenon where the expression of certain genes depends on their parent-of-origin. Many imprinting genes are known to play important roles in human complex diseases such as diabetes, breast cancer and obesity. In recent years, array based eQTL studies have identified many regulatory variants that show associations with gene expression level. However, the necessity of family data to infer the phase information of non-informative markers has limited the identification of imprinting genes. The rapidly arising RNA-seq shed a light for this as its ability to provide information about the phase of a SNP. However, it has been repeated shown that RNA-seq data are overdispersed which brings challenge to the modeling of the gene expression profiling. Moreover, multicolinearlity occurs naturally when we are modeling multiple genetic components, such as additive, dominance and imprinting effects. To address these issues, we introduced a statistical framework to test the main allelic effects along with the parent-of-origin effect. We utilized an orthogonalization procedure which allowed for efficient imprinting effect detection whereas maintained the power to detect the main allelic effect from eQTLs. We conducted extensive simulations to demonstrate the statistical behavior of our proposed method. We also applied the models to

a large scale breast cancer study and revealed potential imprinted regulatory elements that control gene expression.

Variable screening via quantile partial correlation

Shujie Ma
University of California, Riverside
shujie.ma@ucr.edu

In quantile linear regression with ultra-high dimensional data, we propose an algorithm for screening all candidate variables and subsequently selecting relevant predictors. Specifically, we first employ quantile partial correlation for screening, and then we apply the extended Bayesian information criterion (EBIC) for best subset selection. Our proposed method can successfully select predictors when the variables are highly correlated, and it can also identify variables that make a contribution to the conditional quantiles but are marginally uncorrelated or weakly correlated with the response. Theoretical results show that the proposed algorithm can yield the sure screening set. By controlling the false selection rate, model selection consistency can be achieved theoretically. In practice, we proposed using EBIC for best subset selection so that the resulting model is screening consistent. Simulation studies demonstrate that the proposed algorithm performs well, and an empirical example is presented.

Session 57: Statistical Methods for Medical Research using Real-World Data

Gauging drug-outcome associations by leveraging EMR and existent knowledge

Jia Zhan, ♦*Xiaochun Li and Changyu Shen*
Indiana University School of Medicine
xiaochun@iu.edu

We address two issues in drug-outcome association studies. First, it has been recognized that electronic health records (EHR) databases may have systemic hidden biases, for example, failure or incomplete capture of exposure and covariates, such that confounding cannot be fully controlled. Consequently, risk estimates may be biased, resulting in the misguided assessment of the strength and direction of drug-outcome associations. Second, the distribution of the risk measures of a large number of drugs on market for a given outcome is unknown. Using acute myocardial infarction (AMI) as an example, we illustrate how the first issue can be addressed by calibrating the risk measures through drugs known to have no association with AMI in a population-level electronic medical records database (the Indiana Network for Patient Care). We then employ an empirical Bayes approach to address the second issue, which helps to improve the accuracy for the inference of the association of an individual drug with AMI. The study shows that without the hidden bias correction, 66.5%, 12.1% and 2.6% of the drugs included have a risk ratio for AMI greater than 1, 1.5 and 2, respectively, and that with the hidden bias correction, the proportions become 50.8%, 7.4% and 1.7%, respectively. For a new drug Aliskiren (RR=2.19), without the hidden bias correction, we gain 47% precision for the posterior estimate compared with the likelihood estimate; with the hidden bias correction, the precision gain is 48.8%. Our approach serves as a general strategy for pharmaco-epidemiology studies for either an individual drug-outcome pair or multiple drug-outcome pairs.

Statistical methods for drug safety using Electronic Health Records

♦*Ying Wei*[1]*, Daniel Backenroth*[1]*, Ying Li*[2]*, Alex Belloni*[3] *and Carol Friedman*[1]
[1]Columbia University
[2]IBM
[3]Duke University
yw2148@cumc.columbia.edu

Over the last decade, Electronic Health Records (EHRs) systems 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. Extensive effort has been dedicated to developing advanced clinical data processing and data management, to integrate patient data into a computable collection of rich longitudinal patient profiles. EHRs provide unprecedented opportunities for cohort-wide investigations and knowledge discovery.

One important use of EHR is post-market drug safely monitory. Adverse Drug Recaction(ADR) cause serious harm to patients, and result in huge financial burden in health care. A continuous post-marketing surveillance is hence crucial for patient safety. In recent years, EHRs are recognized as an important data sources for pharmacovigilance.One central challenge is to control confoundings. The traditional controlling methods including matching, stratification and regression adjustment fail to address the complexity of EHR data. Their recent adapted versions for big data have limited power and not robust enough for complex EHR data. We consider a new methods for estimating drug safety signals with a robust control over a large number of confounders in EHR. The results are illustrated using single hospital data at Columbia University focusing on four clinically important ADR: acute renal failure (ARF), acute liver failure (ALI), acute myocardial infarction (AMI), and upper gastrointestinal bleeding (GIB).

Generate Individualized Treatment Decision Tree Algorithm with Application to EMR

Kevin Doubleday[1]*, Haoda Fu*[2] *and* ♦*Jin Zhou*[1]
[1]University of Arizona
[2]Lilly Corporate Center
jzhou@email.arizona.edu

With new treatments and novel technology available, personalized medicine has become a key topic in the new era of healthcare. Traditional statistics methods for personalized medicine and subgroup identification primarily focus on single treatment or two arm randomized control trials (RCTs). With restricted inclusion and exclusion criteria, data from RCTs may not reflect real world treatment effectiveness. However, electronic medical records (EMR) offers an alternative venue. In this paper, we propose a general framework to identify individualized treatment rule (ITR), which connects the subgroup identification methods and ITR. It is applicable to both RCT and EMR data. Given the large scale of EMR datasets, we develop a recursive partitioning algorithm to solve the problem (ITR-Tree). A variable importance measure is also developed for personalized medicine using random forest. We demonstrate our method through simulations, and apply ITR-Tree to datasets from diabetes studies using both RCT and EMR data. Software package is available at https://github.com/jinjinzhou/ITR.Tree.

Addressing Unmeasured Confounding in Comparative Observational Research

♦*Wei Shen, Douglas Daries and Xiang Zhang*
Eli Lilly and Company
shen@lilly.com

While the use of real world / observational / big data for comparative effectiveness analyses has grown in recent years, causal infer-

ence from such data typically relies on the unprovable assumption of "no unmeasured confounders" To date, most research simply notes this assumption as a limitation of the research and no quantitative assessment of the potential impact of unmeasured confounding is performed. However, over the past decade many quantitative approaches to assessing the impact of unmeasured confounding have arisen. This includes approaches to quantifying the robustness of the analysis to varying levels of unmeasured confounding as well as newer approaches that incorporate information external to the study and produce an adjusted estimate of the effect. To ensure appropriate use of information arising from such comparative observational research, analyses should include a thorough and quantitative assessment of the potential impact of unmeasured confounders. However, the many options provided by recently developed methods and the variety of research scenarios makes this a challenge to understand the optimal course of action.

The two main goals of this talk are to 1) introduce a best practice guidance for addressing unmeasured confounding; 2) demonstrate how one can incorporate information obtained external from the research study to reduce bias caused by unmeasured confounding. The best practice guidance will include a flowchart / decision tree approach to recommending analysis options given the study scenario and availability of information. The second objective focuses on the common scenario where information on confounders exists in sources external to the particular study, such as the literature or other data bases. A Bayesian Twin Regression modeling approach will be presented that can incorporate information regarding confounders obtained from multiple external data sources and produce treatment effects adjusted for the additional information regarding key confounders.

Session 58: New Developments on High Dimensional Learning

Robust High-dimensional Data Analysis Using a Weight Shrinkage Rule

♦*Xiaoli Gao*[1]*, Bin Luo*[1] *and Yixin Fang*[2]
[1]University of North Carolina at Greensboro
[2]New York University
x_gao2@uncg.edu

In high-dimensional settings, a penalized least squares approach may lose its efficiency in both estimation and variable selection due to the existence of either outliers or heteroscedasticity. In this manuscript, we propose a novel approach to perform robust high-dimensional data analysis in a penalized weighted least square framework. Our main idea is to relate the irregularity of each observation to a weight vector and obtain the outlying status dataadaptively using a weight shrinkage rule. The proposed procedure result in an estimator with potential strong robustness and non-asymptotic consistency. We provide a unified link between the weight shrinkage rule and a robust M-estimation in general settings. We also establish the non-asymptotic oracle inequalities for the joint estimation of both the regression coefficients and weight vectors. These theoretical results allow the number of variables to far exceed the sample size. The performance of the proposed estimator is demonstrated in both simulation studies and real examples.

Inverse Methods for Sufficient Forecasting Using Factor Models

♦*Lingzhou Xue*[1]*, Jianqing Fan*[2]*, Wei Luo*[3] *and Jiawei Yao*[4]
[1]Penn State University
[2]Princeton University
[3]CUNY - Baruch College
[4]Citadel LLC
lxx6@psu.edu

We consider forecasting a single time series using a large number of predictors when a nonlinear forecasting function is present. The linear forecasting is very appealing due to its simplicity. However, it only reveals one dimension of the predictive power in the underlying factors. We develop the sufficient forecasting, which provides several sufficient predictive indices to deliver additional predictive power. The sufficient forecasting correctly estimates projections of the underlying factors even in the presence of an arbitrary and unknown forecasting function. Our work identifies the effective factors that have impacts on the forecast target when the target and the cross-sectional predictors are driven by different sets of common factors. We derive asymptotic properties for the estimate of the central subspace spanned by these projection directions as well as the estimates of the sufficient predictive indices. Our method and theory allow the number of predictors to be larger than the number of observations. We finally demonstrate that the sufficient forecasting improves upon the linear forecasting in both simulation studies and an empirical study of forecasting macroeconomic variables.

This talk is based on several joint works with Jianqing Fan, Wei Luo and Jiawei Yao.

Estimation and Variable Selection for Single-index Cox Proportional Hazard Regression

Peng Zeng
Auburn University
zengpen@auburn.edu

Cox proportional hazard regression has been widely used to model time-to-event data, where the hazard rate is modeled as the product of a nonparametric baseline function and a parametric form of covariates. Although convenient and easy to interpret, it may be subject to the risk of model misspecification. In this talk, we consider an extension of Cox regression, where the covariates have a multiplicative effect on the hazard rate via a nonparametric univariate function of a linear combination of covariates (single-index). This model is essentially an extension of the Cox regression to a single-index model. It achieves a good balance between interpretability and flexibility. We will discuss a novel estimation procedure for single-index Cox proportional hazard regression in high-dimensional data analysis. With a lasso-type penalty on the partial likelihood, the procedure can conduct variable selection simultaneously. The computing algorithm is efficient. Extensive simulation studies and real examples are used to demonstrate the good performance. The theoretical properties of the estimates are also explored.

An augmented ADMM algorithm with application to the generalized lasso problem

Yunzhang Zhu
The Ohio State University
zhu.219@osu.edu

In this talk, I will present a fast and stable algorithm for solving a class of linearly regularized statistical estimation problem. This type of problems arises in many statistical estimation procedures, such as high-dimensional linear regression with fused lasso regularization, convex clustering, and trend filtering, among others. We propose a so-called augmented alternating direction methods of multipliers (ADMM) algorithm to solve this class of problems. As compared to a standard ADMM algorithm, our proposal significantly reduces the amount of computation at each iteration while maintaining the same overall rate of convergence. The superior performance of the augmented ADMM algorithm will be demonstrated

on a generalized lasso problem. Finally, some possible extensions and interesting connections to two well-known algorithms in imaging literature will be discussed.

Session 59: Emerging Statistical Theory in Analyzing Complex Data

Using an Event-History with Risk-Free Model to Study Genetics of Alcohol Dependence

Hsin-Chou Yang[1], *I-Chen Chen*[2], *Yuh-Chyuan Tsay*[1], *Zheng-Rong Li*[1], *Chun-houh Chen*[1], *Hai-Guo Hwu*[3] *and* ♦*Chen-Hsin Chen*[1]

[1]Institute of Statistical Science, Academia Sinica

[2]Dept of Biostatistics, University of Kentucky

[3]Dept of Psychiatry, National Taiwan University

chchen@stat.sinica.edu.tw

Broadly-used case-control genetic association studies ignore possible later onsets for those currently unaffected subjects and assume all the affected subjects contribute equally regardless of their ages at onset. We thus utilize an event-history with risk-free model to simultaneously characterize alcoholism susceptibility and age at onset based on 65 independent non-Hispanic Caucasian males in the Collaborative Study on the Genetics of Alcoholism. After data quality control, 22 single nucleotide polymorphisms (SNPs) on 12 candidate genes are analyzed. The single-SNP analysis show that a SNP on DRD3 triggers alcoholism susceptibility; a SNP on GRIN2B and a SNP on NTRK2 delay, but a SNP on ALDH1A1 advances, ages at onset of alcoholism; and a SNP on DRD2 influences the range of onset ages. The multiple-SNP analysis reveals joint effects of 3 SNPs on DRD3, GRIN2B and DRD2 similar to the single SNP analyses with adjustment of habitual smoking status. Using the statistical visualization software Generalized Association Plots, the 5 genes revealed in this study form different gene clusters in the gene-pathway clustering analysis. This study gains a more comprehensive understanding of genetics of alcoholism than previous case-control studies.

A general approach to categorizing a continuous scale according to an ordinal outcome

♦*Limin Peng*[1], *Amita Manatunga*[1], *Ming Wang*[2], *Ying Guo*[1] *and AKM Rahman*[1]

[1]Emory University

[2]Penn State University

lpeng@emory.edu

In practice, disease outcomes are often measured in a continuous scale, and classification of subjects into meaningful disease categories is of substantive interest. To address this problem, we propose a general analytic framework for determining cut-points of the continuous scale. We develop a unified approach to assessing optimal cut-points based on various criteria, including common agreement and association measures. We study the nonparametric estimation of optimal cut-points. Our investigation reveals that the proposed estimator, though it has been ad-hocly used in practice, pertains to nonstandard asymptotic theory and warrants modifications to traditional inferential procedures. The techniques developed in this work are generally adaptable to study other estimators that are maximizers of nonsmooth objective functions while not belonging to the paradigm of M-estimation. We conduct extensive simulations to evaluate the proposed method and confirm the derived theoretical results. The new method is illustrated by an application to a mental health study.

Multivariate Fay-Herriot Hierarchical Bayesian Estimation of Small Area Means with Measurement Error

Serena Arima[1], *William Bell*[2], ♦*Gauri Datta*[3], *Carolina Franco*[2] *and Brunero Liseo*[1]

[1]University of Rome, La Sapienza

[2]U.S. Census Bureau

[3]U.S. Census Bureau, University of Georgia

gaurisdatta@gmail.com

Area-level models have been extensively used in small area estimation to produce model-based estimates of a population characteristic for small areas, when direct estimates obtained from surveys are modeled (see Fay and Herriot, 1979, JASA). Many surveys collect data on multiple variables which tend to be related. Joint modeling of multiple characteristics of correlated responses may lead to more precise small area estimates than separate univariate modeling of each characteristic would produce. Small area estimation models use auxiliary information to borrow strength from other areas and covariates related with a response variable or a response vector. Auxiliary variables are sometimes measured or obtained from surveys and are subject to measurement or sampling error. It was recognized by researchers that ignoring measurement error in the covariates and using standard solutions developed for covariates measured without error may not provide the correct solution to the problem. In fact it was demonstrated in univariate setup in small area estimation that this naive approach usually results in model-based small area estimators that would be more variable than the direct estimators when some of the covariate values in a small area are measured with substantial error (cf. Ybarra and Lohr, 2008, Biometrika; Arima, Datta and Liseo, 2015, Scand. J. Statist.). In this talk, we consider a multivariate Fay-Herriot model and develop Bayes small area estimates when one or more auxiliary variables are measured with error. We work out a hierarchical Bayesian analysis for the multivariate Fay-Herriot model with a functional measurement error treatment for the covariates measured with error. We apply the proposed methodology to two real examples.

Optimal Estimation for Quantile Regression with Functional Response

♦*Xiao Wang*[1], *Zhengwu Zhang*[2], *Linglong Kong*[3] *and Hongtu Zhu*[4]

[1]Purdue University

[2]SAMSI

[3]University of Alberta

[4]University of North Carolina

wangxiao@purdue.edu

Quantile regression with functional response and scalar covariates has become an important statistical tool for many neuroimaging studies. In this paper, we study optimal estimation of varying coefficient functions in the framework of reproducing kernel Hilbert space. Minimax rates of convergence under both fixed and random designs are established. We have developed easily implementable estimators which are shown to be rate-optimal. In the case of quantile regression with functional response, we prove that the divergence is exactly equal to the number of interpolated functional responses, which justifies the selection of the regularization parameter. We derive an efficient algorithm to compute the estimator based on the ADMM algorithm. Simulations and real data analysis are conducted to examine the finite-sample performance.

Session 60: Survival and Cure Rate Modeling

Evaluating Utility Measurement from Recurrent Marker Processes in the Presence of Competing Terminal

Mei-Cheng Wang
Johns Hopkins University
mcwang@jhu.edu

In follow-up or surveillance studies, marker data are frequently collected conditioning on the occurrence of recurrent events. In many situations, a marker measurement does not exist unless a recurrent event takes place. Examples include medical cost for inpatient or outpatient cares, length-of-stay for hospitalizations, and prognostic or quality-of-life measurement repeatedly measured at multiple infections related to a certain disease. A recurrent marker process, defined between a pre-specified time origin and a terminal event, is composed of recurrent events and repeatedly measured marker measurements. This talk considers nonparametric estimation of the mean recurrent marker process in the situation when the occurrence of terminal event is subject to competing risks. Statistical methods and inference are developed to address a variety of questions and applications, for the purposes of estimating and comparing the integrated risk in relation to recurrent events, marker measurements and time to the terminal event for different competing risk groups. A SEER-Medicare linked database is used to illustrate the proposed approaches. (*This is joint work with Yifei Sun)

Tests for stochastic ordering under biased sampling

♦*Hsin-wen Chang*[1]*, Hammou El Barmi*[2] *and Ian McKeague*[3]
[1]Academia Sinica
[2]The City University of New York
[3]Columbia University
hwchang@stat.sinica.edu.tw

In two-sample comparison problems it is often of interest to examine whether one distribution function majorizes the other, i.e. for the presence of stochastic ordering. This talk introduces a nonparametric test for stochastic ordering based on size-biased data, allowing the pattern of size bias to differ between the two samples. The test is formulated in terms of a maximally-selected local empirical likelihood statistic. A Gaussian multiplier bootstrap is devised to calibrate the test. A simulation study indicates that the proposed test outperforms an analogous Wald-type test, and that it provides substantially greater power than what is available when ignoring the sampling bias. The approach is illustrated using data on blood alcohol concentration and age of drivers involved in car accidents, in which size bias is present because the drunker drivers are more likely to be sampled. Further, younger drivers tend to be more affected by alcohol, so when comparing with older drivers, the analysis is adjusted for differences in the patterns of size bias.

On relative importance in the effect of two exposures

Xinhua Liu and ♦*Zhezhen Jin*
Department of Biostatistics, Columbia University
zj7@cumc.columbia.edu

To study relatively important effect of exposure variables on a health outcome with continuous measure, we propose to use a linear regression model with main predictor of weighted sum of standardized exposure variables that usually have non-negative correlation. The unknown weights typically range between zero and one that the exposure variable with a larger weight contributes more on the effect. We examined likelihood based tests, with and without ignoring constraint on model parameters in a two-stage analysis. At the first stage testing overall effect of the exposures, weights are treated as nuisance parameters which present only under the alternative hypothesis. Then Davies' method (1977, 1987) should be applied. As the likelihood ratio test (LRT) for re-parameterized linear model without constraint is easy to implement with existing statistical software and also unbiased, we compared the two approaches through a simulation study and found comparable empirical power with the two tests. At the second stage when there is evidence for the effect of exposures, one can use likelihood ratio test to detect unequal weights, where the maximum likelihood estimates of weights are subject to constraints. Our simulation study suggested that power of the test increase with size of the overall effect and sample size while decrease with increasing degree of correlation between the exposure variables. In application, we investigated relative importance of two neuro-toxicants, measured as blood Mn and As, in their effect on children's cognitive function adjusted for covariates.

Semiparametric Accelerated Failure Time Models with Missing Covariates

♦*Shuai Chen and Menggang Yu*
University of Wisconsin - Madison
schen264@wisc.edu

Semiparametric accelerated failure time model relates the logarithm of the failure time to covariates while leaving the error distribution unspecified. However, incomplete covariate data commonly occurs in survival analysis, which complicates the estimation. We propose a set of estimating functions for semiparametric accelerated failure time model with missing covariates, under censoring-ignorable missingness at random (CIMAR) mechanism (the missingness depends on survival time but not on censoring time). We first propose an estimating equation based on the inverse probability weighting technology using uncensored subjects. To improve efficiency, we further propose an estimating equation adopting information from censored subjects. With wisely chosen weights, the two estimating equations can be combined to yield a more efficient and consistent estimator. These equations can be solved easily with closed forms. Numerical studies demonstrate that our estimator works quite well in reasonably large samples.

Session 61: Semiparametric Statistical Methods for Complex Data

ESTIMATION OF NON-CROSSING QUANTILE SURFACES

Chen Dong[1]*, Shujie Ma*[2]*, Liping Zhu*[3] *and* ♦*Xingdong Feng*[1]
[1]Shanghai University of Finance and Economics
[2]University of California-Riverside
[3]Renmin University of China
feng.xingdong@mail.shufe.edu.cn

ESTIMATION OF NON-CROSSING QUANTILE SURFACES Though the theoretical properties of quantile regression have been extensively studied in the past three decades, in practice it is not unusual to obtain crossing quantile surfaces with regular approaches for estimating quantile functions at different quantile levels. The crossing quantile surfaces are intrinsically uninterpretable. To address this issue, we consider a semiparametric multi-index quantile regression subject to monotonicity restriction at different quantile levels. We first connect the semiparametric multi-index quantile regression model with a dimension-reducible model. Such a connection allows us to estimate the index coefficients consistently. The B-splines are then used to approximate the nonparametric function under the monotonicity restriction, which numerically corresponds to a constrained linear programming problem. To further improve the

computation efficiency, we estimate the B-spline coefficients based on a dual of linear programming. We assess the finite-sample performance of our proposed method through comprehensive simulations, and compare the prediction perfromance of different methods through an application to a real dataset.

Composite Estimation for Single-Index Model with Responses Subject to Detection Limits

Yanlin Tang[1], ♦*Huixia Wang*[2] *and Hua Liang*[2]
[1]Tongji University
[2]The George Washington University
judywang@gwu.edu

We propose a semiparametric estimator for single-index model with censored responses due to detection limits. In the presence of censoring, the mean function cannot be identified without any parametric distributional assumptions, but the quantile function is still identifiable at some quantile levels. To avoid parametric distributional assumption, we propose to fit censored quantile regression and combine information across quantile levels to estimate the unknown smooth link function and the index parameter in the single-index model. Under some regularity conditions, we show that the estimated link function achieves the nonparametric oracle convergence rate, and the estimated index parameter is asymptotically normal. The simulation study shows that the proposed estimator is competitive to the semiparametric least squares estimator based on the latent uncensored data when errors are from normal distribution, and much more efficient when errors are from heavy-tailed distributions. The practical value of the proposed method is demonstrated through the analysis of an human immunodeficiency virus antibody data set.

Efficient Estimation of Partially Linear Models for Spatial Data over Complex Domains

♦*Lily Wang*[1], *Guannan Wang*[2], *Ming-Jun Lai*[3] *and Lei Gao*[1]
[1]Iowa State University
[2]College of William and Mary
[3]The University of Georgia
lilywang@iastate.edu

We study the estimation of partially linear models for spatial data distributed over complex domains. Bivariate splines over triangulations are implemented to represent the nonparametric component on an irregular two-dimensional domain. This method does not require constructing finite elements or locally supported basis functions, allowing for an easier implementation of piecewise polynomial representations of various degrees and various smoothness over an arbitrary triangulation. A penalized least squares method is proposed to estimate the model via QR decomposition. The estimators of the parameters are proved to be asymptotically normal under some regularity conditions. The estimator of the bivariate function is consistent, and its rate of convergence is also established. The proposed method enables us to construct confidence intervals and permits inference for the parameters. The performance of the estimators is evaluated by two simulation examples and a real data analysis.

A Regression Model for the General Trend Analysis of Bivariate Panel Data in Continuous Scale

♦*Yi Ran Lin*[1] *and Wei Hsiung Chao*[2]
[1]Institute of Statistical Science, Academia Sinica
[2]Dept. of appl. math., National Dong Hwa University
yriln@stat.sinica.edu.tw

In many epidemiologic cohort studies, the underlying bivariate response process in continuous scale may experience natural variation over time, as is the case with systolic and diastolic blood pressures. The blood pressure fluctuates quickly in response to physical activity, diet and stress, and changes slowly in response to ageing and metabolic condition. Instead of the short-term variation of blood pressure, it is often more of interest to study the general trend and evolution of blood pressure in the long run and to identify factors that are associated with the movement toward higher values hence leading to hypertension. To provide such a general trend information about the evolution of the response process, we developed a regression model based on the bivariate Ornstein-Uhlenbeck process. The time-varying covariates are incorporated via assuming them to be piecewise constants between two successive observation time points. The resultant model can be seen as an analogue of the local equilibrium distribution model for ordinal response (Kosorok and Chao, 1996) in the setting of multiple continuous responses. In addition to the general trend information, the proposed model has the capacity of assessing the serial correlation for each response and the mutual dependence among the responses. A generalized estimating equations approach was developed for the parameters of interest. Asymptotic properties of the estimators were established and validated by simulation studies. The estimating method and relevant model diagnostic tools are illustrated by using the blood pressure data arising in the Cardiovascular Disease Risk Factor Twotownship Study (CVDFACTS) which is a community-based cohort study carried out in Taiwan. The proposed regression method can be also applied to study the evolution of multiple continuous biomarkers with possible correlation in research of epidemiology and biobanks.

Keywords: Panel data, Markov processes, bivariate Ornstein-Uhlenbeck processes, generalized estimating equations.

Session 62: Recent Advances on Multiple Fronts of Statistical Analysis for Genomics Data

A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer

Binghui Liu[1], *Xiaotong Shen*[2] *and* ♦*Wei Pan*[2]
[1]Northeast Normal University
[2]University of Minnesota
weip@biostat.umn.edu

Next-generation sequencing studies on cancer somatic mutations have discovered that driver mutations tend to appear in most tumor samples, but they barely overlap in any single tumor sample, presumably because a single driver mutation can perturb the whole pathway. Based on the corresponding new concepts of coverage and mutual exclusivity, new methods can be designed for de novo discovery of mutated driver pathways in cancer. Since the computational problem is a combinatorial optimization with an objective function involving a discontinuous indicator function in high dimension, many existing optimization algorithms, such as a brute force enumeration, gradient descent and Newton's methods, are not practically feasible or directly applicable. We develop a new algorithm based on a novel formulation of the problem as non-convex programming and non-convex regularization. The method is computationally more efficient, effective and scalable than existing Monte Carlo searching and several other algorithms, which have been applied to The Cancer Genome Atlas (TCGA) project. We demonstrate its promising performance with application to two cancer datasets to discover de novo mutated driver pathways.

A novel tail dependence measure to quantify the reproducibility

and quality of sequencing experiments

Tao Yang and ♦*Qunhua Li*
Penn State University
qunhua.li@psu.edu

The quality and reproducibility of sequencing experiments is essential to the reliability of downstream analysis and biological interpretation. Though Pearson and Spearman correlation coefficients are often used to assess the reproducibility of replicate sequencing experiments, they can be easily misled by highly repetitive regions or excessive amount of low count regions on the genome. Here we developed a novel reproducibility measure based on tail dependence. We evaluate our methods on different sequencing experiments. Our measure is robust and can effectively distinguish experiments with different levels of reproducibility and quality. It helps experimentalists identify suboptimal experiments failed due to different causes.

Large scale multiple testing for clustered signals

♦*Hongyuan Cao*[1] *and Wei Biao Wu* [2]
[1]University of Missouri-Columbia
[2]University of Chicago
caohong@missouri.edu

We propose a change point detection method for large scale multiple testing problems with clustered signals. Unlike the classic change point detection setup, the signals can vary in size and distribution within a cluster. The spatial structure on the signals enables us to accurately delineate the boundaries between null and alternative hypotheses. New test statistics are proposed for observations from one sequence and multiple sequences. Their asymptotic distributions are established with consistent estimators for unknown parameters. We allow the variances to be heteroscedastic in the multiple sequence case, which greatly expands the applicability of the proposed method. Simulation studies demonstrate that the large sample approximations are adequate for practical use and may yield favorable performance. Dataset from array CGH and DNA methylation are used to demonstrate the utility of the proposed methods.

A Cautionary Note on using Cross-validation for Molecular Classification

♦*Li-Xuan Qin and Huei-Chung Huang*
Memorial Sloan Kettering Cancer Center
qinl@mskcc.org

Cross validation is commonly used in molecular classification studies to derive an estimate of the classifier's error rate. However, it no longer works well when the data possess confounding variations due to experimental handling (that is, handling effects that induce confounding), regardless of the use of data normalization to remove bias. As a result, cross-validation of such data can lead to a spurious estimate of the error rate in the over-optimistic direction.

Methods We demonstrate this important yet over-looked complication of cross validation using a unique pair of datasets on the same set of tumor samples. One dataset was collected with uniform handling to prevent handling effects; the other dataset was collected without uniform handling and exhibited handling effects. The paired datasets were used to estimate the biological effects of the samples and the handling effects of the arrays in the latter dataset, which were then used to simulate data using re-sampling and virtual hybridization following various array-to-sample-group assignment schemes.

Results Our study showed that (1) cross-validation tended to underestimate the error rate when the data possessed confounding handling effects, (2) depending on the relative amount of handling effects, normalization may further worsen the under-estimation of the error rate, (3) balanced assignment of arrays to sample groups allowed cross-validation to provide an unbiased error estimate.

Conclusion Our study demonstrates the benefits of balanced array assignment for reproducible molecular classification and calls for caution on the routine use of data normalization and cross-validation in such analysis.

Session 63: Multi-regional Clinical Trials (MRCT): Statistical Challenges, Trial Design Approaches, and Other Aspects

A Few Considerations on Regional Differences in MRCT

♦*Bo Yang*[1] *and Yijie Zhou*[2]
[1]Vertex
[2]AbbVie
bo_yang@vrtx.com

While multi-regional clinical trials (MRCT) has gained extensive use in contemporary drug development with the intrinsic underlying assumption that drug effectiveness and safety are the same across regions, regional differences in other aspects can also make interpretation of MRCT results challenging. In this talk, we discuss four types of regional differences and associated statistical challenges: endpoint differences, background event rate differences control therapy differences, and geographic population shift.

Inconsistency and drop-minimum data analysis

Fei Chen[1], ♦*Gang Li*[1] *and Gordon Lan*
[1]J&J
gli@its.jnj.com

Even though consistency is an important issue in multi-regional clinical trials (MRCT) and inconsistency is often anticipated, solutions for handling inconsistency are rare. If a regions treatment effects are inconsistent with that of the other regions, pooling all the regions to estimate the overall treatment effect may not be reasonable. Unlike the multiple center clinical trials conducted in the US and Europe, in MRCT different regional regulatory agencies may have their own ways to interpret data and approve new drugs. It is therefore practical to consider the case in which the data from the region with the minimal observed treatment effect is excluded from the analysis in order to attain the regulatory approval of the study drug. Under such cases, what is the appropriate statistical approach for the remaining regions? We provide a solution first formulated within the fixed effects framework, and then extend it to discrete random effects models.

Regional Efficacy Assessment in Multi-Regional Clinical Development

Yijie Zhou
AbbVie
yijie.zhou@abbvie.com

Multi-regional clinical development is gaining more popularity in regions outside of the traditional market of US and EU. When a standard global trial cannot accommodate the sample size requirement for a particular target country, local trial is conducted and data from the global trial and the local trial together will be used to support local filing in the target country. This approach is considered efficient drug development both globally and in the target country. However it remains a challenge how to combine global trial data and local trial data toward local filing. To address this challenge, we propose an "interpretation-centric" evaluation criterion based on a weighted estimator that weights data from the target country and outside of the target country. This approach provides an unbiased

estimate of a global treatment effect with appropriate representation of the target country patient population, where the "appropriate representation" is the desired proportion of the target country participants in a global trial and is measured by the weight parameter. This natural interpretation can facilitate drug development discussion with local regulatory agencies. Sample size of the local trial can be determined using the proposed weighted estimator. Approaches for weight determination are also discussed.

New Method of Borrowing Information from Outside Regional Data for Analyzing the Local Regional Data

♦*Takahiro Hasegawa*[1]*, Lu Tian*[2]*, Brian Claggett*[3] *and Lee-Jen Wei*[4]
[1]Shionogi & Co., Ltd.
[2]Stanford University School of Medicine
[3]Brigham and Women's Hospital
[4]Harvard University
takahiro.hasegawa@shionogi.co.jp

To accelerate the drug development process and shorten approval time, multiregional clinical trials (MRCTs) have been designed to incorporate patients from many countries/regions around the world under the same protocol. After showing the overall efficacy of a drug in all regions, one can also simultaneously evaluate a treatment effect in each region to support drug registration in the corresponding region. However, in the case of small sample size of a local region, conventional treatment effect estimates may not be applicable due to a wide confidence interval. In addtion, an analysis population in the local region might differ from that in the outside region, which is defined as any region beyond the local region, possibly due to regional differences in patient characteristics. Therefore, finding a way to bridge the results of the MRCT to the local region is an important issue. In this talk, we focus on the specific region and establish statistical inference for the treatment effect in the local region by borrowing information from the outside regional data. More specifically, we construct an outcome regression model for covariates in each treatment group by using the outside regional data. Then the treatment effect inference is performed by plugging the covariates into the local region data into the established models. The proposed method is applied to real example data.

Session 64: High Dimensional Inference: Methods and Applications

Localized-Variate PCA for Multivariate Functional Data

Robert Krafty
University of Pittsburgh
rkrafty@pitt.edu

In this talk, we discuss localized-variate functional principal component analysis (LVFPCA) for finding basis functions that account for most of the variability in a random multivariate process. As opposed to traditional methods, the basis functions found by LVFPCA can be both sparse among variates (i.e. is zero across an entire functional variate) and localized within a variate (i.e. nonzero only within a subinterval of a variate). LVFPCA is formulated as a rank-one based convex optimization problem with matrix L1 and block Frobenius norm based penalties, which induce localization and variate sparsity, respectively. The approach not only provides more accurate estimates of PCs, but it also provides a tool for obtaining more interpretable PCs. An analysis of fMRI data reveals interpretable information that cannot be found by standard methods.

Nonparametric Screening under Conditional Strictly Convex Loss

Xu Han
Temple University
hanxu3@temple.edu

Ultrahigh-dimensional variable selection has received increasing attention in statistical learning due to the big data collection in a variety of scientific areas. The dimension p and the sample size n can satisfy the NP dimensionality (Fan & Lv 2008). In the current talk, we develop a general and unified framework for nonparametric sure screening methods from a loss function perspective. Consider a loss function to measure the divergence of the response variable and the underlying nonparametric function of covariates. We newly propose a class of loss functions called conditional strictly convex loss, which contains negative log-likelihood loss from exponential families, exponential loss for binary classification, quantile regression loss for robust estimation and beyond. We will establish our sure screening property and model selection size control within this class of loss functions. Our methods will be illustrated through simulation studies and real data analysis.

Dimension Reduction for Big Data Analysis

Dan Shen
University of South Florida
danshen@usf.edu

High dimensionality has become a common feature of "big data" encountered in many divergent fields, such as imaging and genetic analysis, which provides modern challenges for statistical analysis. To cope with the high dimensionality, dimension reduction becomes necessary.

I first introduce Multiscale Weighted PCA (MWPCA), a new variation of PCA, for imaging analysis. MWPCA introduces two sets of novel weights, including global and local spatial weights, to enable a selective treatment of individual features and incorporation of class label information as well as spatial pattern within imaging data. Simulation studies and real data analysis show that MWPCA outperforms several competing PCA methods.

Second we develop statistical methods for analyzing tree-structured data objects. This work is motivated by the statistical challenges of analyzing a set of blood artery trees, which is from a study of Magnetic Resonance Angiography (MRA) brain images of a set of 98 human subjects. We develop an entirely new approach that uses the Dyck path representation, which builds a bridge between the tree space (a non-Euclidean space) and curve space (standard Euclidean space). That bridge enables the exploitation of the power of functional data analysis to explore statistical properties of tree data sets.

Distance-Based Methods for Analyzing Data from 16S rRNA Microbiome Studies

Glen Satten
Centers for Disease Control and Prevention
gas0@cdc.gov

Appreciation of the importance of the microbiome is increasing as sequencing technology has made possible ascertaining the microbial content of a variety of samples. Studies that sequence the 16S rRNA gene, ubiquitous in and nearly exclusive to bacteria, have proliferated in the medical literature. Data from these studies are summarized in a data matrix with the observed counts from each operational taxonomic unit (OTU) or species for each sample. Analysis often reduces these data to a matrix of pairwise distances or dissimilarities; plotting the data using the first two or three principal components (PCs) of this distance matrix often reveals meaningful groupings in the data. However, once the distance matrix is calculated, it is no longer clear which OTUs or species are important

to the observed clustering; further, the PCs are hard to interpret and cannot be calculated for subsequent observations. The authors show how to construct approximate decompositions of the data matrix that pair PCs with linear combinations of OTU or species frequencies, and show how these decompositions can be used to construct biplots, select important OTUs and partition the variability in the data matrix into contributions corresponding to PCs of an arbitrary distance or dissimilarity matrix. An analysis of the bacteria found in 45 smokeless tobacco samples is used to illustrate the approach.

Bootstrap-Based Measures of Uncertainty for EEG Artifact Detection using ICA

♦*Rachel Nethery and Young Truong*
University of North Carolina at Chapel Hill
nethery@live.unc.edu

Independent component analysis (ICA) is a blind source separation technique which can be applied to multivariate data generated by the linear mixing of signals from unobserved sources of activity in order to recover these latent sources. Mathematically, the ICA model is written as X=AS, where X is the matrix of observed data, S is the matrix of independent latent sources, and A is a constant but unknown mixing matrix whose components represent the contribution of each latent source to each observation in X. Particularly in neuroscientific applications of ICA, the latent sources often arise from processes that contain temporal correlation, and ICA estimation procedures, such as the colored ICA (cICA) algorithm proposed by Lee et al. (2011), can exploit this temporal structure to estimate parameters of interest with greater precision than competing methods which assume independence within sources.

In the analysis of electroencephalogram (EEG) signals, which are often corrupted by artifacts such as eye and muscle movement, ICA has been identified as a useful tool for artifact detection and extraction. An analysis by Delorme et al. (2007) of several methods for detecting artifacts using ICA revealed that the most sensitive of these procedures is one which develops thresholds for the estimated spectral densities of the ICA estimated latent sources and identifies sources whose spectral densities exceed these thresholds as artifacts. This method, however, fails to take into account the statistical uncertainties in the ICA estimates, likely due to the difficulty in quantifying uncertainties under most popular ICA models. Here, we present a bootstrap algorithm for cICA which can be used to generate confidence bands for estimated source spectral densities and can be integrated into Delorme et al.'s spectral density thresholding method for EEG artifact detection.

The cICA algorithm, which assumes that each sources arises from an autoregressive moving average time series process, estimates A and S (in terms of the time series parameters for each source), which can be used to generate an estimate of the source spectral densities. Based on these estimates, we develop an algorithm which generates a pre-specified large number, B, of bootstrapped source matrices. We then obtain B bootstrap samples from X by multiplying each of these bootstrapped matrices by the cICA estimated mixing matrix. We again run cICA on each of these bootstrapped X matrices, resulting in B bootstrap estimates of all the cICA parameters and the spectral densities of the sources. In combination with the quantile method of forming bootstrap confidence intervals, these bootstrapped spectral density estimates can be used to create pointwise confidence bands for the true spectral density of each latent source signal.

In simulation studies, we report reliable point-wise coverage rates for the spectral density confidence bands, while acknowledging that multiple testing problems may arise if many points are utilized simultaneously in analyses. We also demonstrate the use of the spectral density confidence bands for artifact detection and removal in real EEG data.

Session 65: Modern Advancements in High-Dimensional Functional Data

Single-index Models for Function-on-Function Regression

♦*Guanqun Cao*[1] *and Lily Wang*[2]
[1]Auburn University
[2]Iowa State University
gzc0009@auburn.edu

The single-index model is a flexible and efficient tool to incorporate the effect of high dimensional covariates in a regression problem. We develop a functional single-index model for functional data analysis, where both the response and multiple predictors are functions, and we assume the response is related to a linear combination of the predictors via an unknown link function. The proposed method greatly enhances the flexibility of functional linear models and provides a useful tool for dimension reduction in regression with multiple functional predictors. Several numerical examples illustrate that the proposed model and estimation methodology are flexible and effective in practice.

Longitudinal Regression for Time-varying Functional Covariate

♦*Md Nazmul Islam and Ana-Maria Staicu*
North Carolina State University
mnislam@ncsu.edu

We propose a statistical framework to study the dynamic association between scalar outcomes and functional predictor which is observed longitudinally. The novelty of the proposed model is in the incorporation of the time-varying effect of the functional covariate and the proposal of a parsimonious approximation. We introduce an efficient estimation procedure that has excellent numerical properties and allows us to predict the full trajectory for outcome variable with high precision. The proposed method is illustrated with extensive simulation study and an application to the longitudinal sow data where the prediction of feeding behavior of sows is of primary concern and our method exhibits excellent numerical performance in terms of prediction efficiency and time.

A rotate-and-solve procedure for high dimensional classification

Ning Hao
The University of Arizona
nhao@math.arizona.edu

Many high dimensional classification techniques have been proposed in the literature based on sparse linear discriminant analysis. To efficiently use them, sparsity of linear classifiers is a prerequisite. However, this might not be readily available in many applications, and rotations of data are required to create the needed sparsity. In this talk, we consider a family of rotations to create the required sparsity. The basic idea is to use the principal components of the sample covariance matrix of the pooled samples and its variants to rotate the data first and to then apply an existing high dimensional classifier. This rotate-and-solve procedure can be combined with any existing classifiers, and is robust against the sparsity level of the true model. The effectiveness of the proposed method is demonstrated by a number of simulated and real data examples.

Multivariate Spatio-Temporal Models for High-Dimensional Areal Data

♦*Scott Holan, Jonathan Bradley and Christopher Wikle*

University of Missouri
holans@missouri.edu

Many data sources report related variables of interest that are also referenced over geographic regions and time; however, there are relatively few general statistical methods that one can readily use that incorporate these multivariate spatio-temporal dependencies. Additionally, many multivariate spatio-temporal areal data sets are extremely high dimensional, which leads to practical issues when formulating statistical models. For example, we analyze Quarterly Workforce Indicators (QWI) published by the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. QWIs are available by different variables, regions, and time points, resulting in millions of tabulations. Despite their already expansive coverage, by adopting a fully Bayesian framework, the scope of the QWIs can be extended to provide estimates of missing values along with associated measures of uncertainty. Motivated by the LEHD, and other applications in federal statistics, we introduce the multivariate spatio-temporal mixed effects model (MSTM), which can be used to efficiently model high-dimensional multivariate spatio- temporal areal data sets. The proposed MSTM extends the notion of Moran's I basis functions to the multivariate spatio-temporal setting. This extension leads to several methodological contributions, including extremely effective dimension reduction, a dynamic linear model for multivariate spatio-temporal areal processes, and the reduction of a high-dimensional parameter space using a novel parameter model.

Session 66: Bayesian Approaches for Medical Product Evaluation

BEANZ: A Web-Based Software for Bayesian Analysis of Heterogeneous Treatment Effect

Chenguang Wang
Johns Hopkins University
cwang68@jhmi.edu

When making health care decisions, it is vital to assess the heterogeneity of treatment effects (HTE). Nevertheless, it remains challenging to evaluate HTE based on information collected from clinical studies that are often designed and conducted to evaluate the efficacy of a treatment for the overall population. The Bayesian framework offers a principled and flexible approach to estimate and compare treatment effects across subgroups of patients defined by their characteristics. In this paper, we describe a web-based statistical software, BEANZ, that facilitates the conduct of Bayesian analysis of HTE in an interactive and user-friendly manner. The software allows users to upload clinical data, explore a wide range of Bayesian HTE analysis models, and produce posterior inferences about HTE.

Bayesian Models to Leverage Data from Early Visits in an Adaptive Design

Anna McGlothlin
Berry Consultants
anna@berryconsultants.com

Clinical trials to evaluate medical products often require long follow-up times to assess the primary endpoint. In this talk, we explore how modeling data from intermediate visits can facilitate a trial's adaptation decisions when the primary endpoint has a long time horizon. We give an example of a Goldilocks design that models the longitudinal data in order to select a sample size and to accelerate the decision of trial success. Methods are illustrated through simulated example trials.

Incorporation of stochastic engineering models as a prior in Bayesian medical device trials

Tarek Haddad[1], ♦*Adam Himes*[1], *Laura Thompson*[2], *Telba Irony*[2] *and Rajesh Nair*[2]
[1]Medtronic
[2]FDA
adam.k.himes@medtronic.com

Evaluation of medical devices via clinical trial is often a necessary step in the process of bringing a new product to market. In recent years, device manufacturers are increasingly using stochastic engineering models during the product development process. These models have the capability to simulate virtual patient outcomes.

In this work, we present a novel method based on the power prior for augmenting a clinical trial using virtual patient data. To properly inform clinical evaluation, the virtual patient model must simulate the clinical outcome of interest, incorporating patient variability, as well as the uncertainty in the engineering model and in its input parameters. The number of virtual patients is controlled by a loss function which uses the similarity between modeled and observed data.

This method is illustrated by a case study of cardiac lead fracture. Different loss functions are used to cover a range of scenarios in which the type I error rates and power vary for the same number of enrolled patients. We show how incorporation of engineering models as prior knowledge in a Bayesian clinical trial design can provide benefits of decreased sample size and trial length while still controlling type I error rate and power.

Evaluation of treatment efficacy using a Bayesian mixture piecewise linear model

♦*Lili Zhao*[1], *Dai Feng, Brian Neelon and Marc Buyse*
[1]Biostatistics, University of Michigan
zhaolili@umich.edu

Prostate-specific antigen (PSA) is a widely used marker in clinical trials for patients with prostate cancer. We develop a mixture model to estimate longitudinal PSA trajectory in response to treatment. The model accommodates subjects responding and not responding to therapy through a mixture of two functions. A responder is described by a piecewise linear function, represented by an intercept, a PSA decline rate, a period of PSA decline, and a PSA rising rate; a non-responder is described by an increasing linear function with an intercept and a PSA rising rate. Each trajectory is classified as a linear or a piecewise linear function with a certain probability, and the weighted average of these two functions sufficiently characterizes a variety of patterns of PSA trajectories. Furthermore, this mixture structure enables us to derive clinically useful endpoints such as a response rate and time-to-progression, as well as biologically meaningful endpoints such as a cancer cell killing fraction and tumor growth delay. We compare our model with the most commonly used dynamic model in the literature and show its advantages.

Session 67: New Methods for Complex Data

Jackknife Empirical Likelihood for the Gini Correlation

♦*YONGLI SANG*[1], *XIN DANG*[2] *and YICHUAN ZHAO* [3]
[1]UNIVERSITY OF MISSISSIPPI
[2]UNIVERSITY
[3]GEORGIA STATE UNIVERSITY
ysang@go.olemiss.edu

The main aim of this paper is to develop the Jackknife empirical likelihood for the Gini correlation between two variables. We construct confidence intervals for the Gini correlation without estimat-

ing the asymptotic variance. The two components of Gini correlation are equal only when the distributions of the two variables are exchangeable. We form a new way to test the exchangeability of two distributions via testing whether the difference of the two Gini correlations is 0 by applying the Jackknife empirical likelihood method. The simulation study confirms the advantage of the proposed method in applications.

Complex-valued wavelet lifting and applications

♦*Marina Knight*[1]*, Jean Sanderson*[2]*, Matt Nunes*[3] *and Piotr Fryzlewicz*[4]
[1]University of York, UK
[2]University of Sheffield, UK
[3]University of Lancaster, UK
[4]London School of Economics, UK
marina.knight@york.ac.uk

Signals with irregular sampling structures arise naturally in many fields. In applications such as nonparametric regression and spectral decomposition, classical methods often assume a regular sampling pattern, thus cannot be applied without prior data processing. This work proposes new complex-valued analysis techniques based on the wavelet lifting scheme that removes "one coefficient at a time". Our proposed lifting transform can be applied directly to irregularly sampled data and is able to adapt to the signal(s) characteristics. As our new lifting scheme produces complex-valued wavelet coefficients, it provides an alternative to the Fourier transform for irregular designs, allowing phase or directional information to be represented. We demonstrate the potential of this flexible methodology over real-valued analysis in the nonparametric regression context, where it outperforms its competitors. We also discuss applications in bivariate time series analysis, where the complex-valued lifting construction allows for coherence and phase quantification.

Neyman-Pearson (NP) Classification and NP-ROC

♦*Xin Tong* [1]*, Yang Feng*[2] *and Jingyi Li*[3]
[1]University of Southern California
[2]Columbia University
[3]University of California, Los Angeles
xint@marshall.usc.edu

In many binary classification applications such as disease diagnosis, type I errors are often more important than type II errors, and practitioners have the great need to control type I errors under a desired threshold α. However, common practices that tune empirical type I errors to α often lead to classifiers with type I errors much larger than α. In statistical learning theory, the Neyman-Pearson (NP) binary classification paradigm installs a type I error constraint under some user specified level α before it minimizes type II errors. Despite recent theoretical advances, the NP paradigm has not been implemented for many classification scenarios in practice. In this work, we propose an umbrella algorithm that adapts popular classification methods, including logistic regression, support vector machines, and random forests, to the NP paradigm. Powered by these NP classification methods, we propose the NP receiver operating characteristic (NP-ROC) curves, a variant of the receiver operating characteristic (ROC) curves, which have been widely used for evaluating the overall performance of binary classification methods at all possible type I error thresholds. Despite conceptual simplicity and wide applicability, ROC curves provide no reliable information on how to choose classifiers whose type I errors are under a desired threshold with high probability. In contrast, NP-ROC curves serve as effective tools to evaluate, compare and select binary classifiers with prioritized type I errors. We demonstrate the use and advantages of NP-ROC curves via simulation and real data case studies.

Reduced-Rank Linear Discriminant Analysis

♦*Yue Niu*[1]*, Ning Hao*[1] *and Bin Dong*[2]
[1]University of Arizona
[2]Peking University
yueniu@math.arizona.edu

Many high dimensional classification techniques have been developed recently. However, most works focus on only the binary classification problem. Available classification tools for the multi-class cases are either based on over-simplified covariance structure or computationally complicated. In this talk, following the idea of reduced ranked linear discriminant analysis, we introduce a new dimension reduction tool with the flavor of supervised principal component analysis. The proposed method is computationally efficient and can incorporate the correlation structure among the features. We illustrate our methods by simulated and real data examples.

Session 68: Recent Advances in Regression Analysis

Classification of Paper Citation Trajectories Through Functional Poisson Regression Model

♦*Ruizhi Zhang*[1]*, Jian Wang*[2] *and Yajun Mei*[1]
[1]Georgia Institute of Technology
[2]KU Leuven
rzhang320@gatech.edu

Citation-based metrics have been ubiquitously used to aid in science decision-making, but we still lack a comprehensive understanding of the citation trajectories of individual publications. Inspired by an empirical study of 30-year citation trajectories of 1699 papers published in 1980 in the American Physical Society journals, we propose a functional Poisson regression model for individual paper's citation trajectory, and fit it to the observed citations of individual papers by functional principal component analysis (FPCA) and maximum likelihood estimation. In addition, we apply the K-means clustering algorithm to individual-paper-specified coefficients of our proposed model to identify general patterns of citation trajectories. As compared with existing methods, our proposed approach yields a better fitting to the real dataset of observed 30- year citation trajectories of 1699 papers, and reveals interesting citation patterns. Specifically, it demonstrates the existence of evergreen cluster of papers that do not exhibit any declining in annual citations over 30 years.

Kernel Ridge Regression under Random Projection: Computational-and-Statistical Trade-off

♦*Meimei Liu*[1]*, Zuofeng Shang*[2] *and Guang Cheng*[1]
[1]Department of Statistics, Purdue University
[2]Department of Math Science, Binghamton University
liu1197@purdue.edu

Kernel Ridge Regression (KRR) is a broad class of nonparametric models including smoothing spline as a feature example. However, when sample size n is large, its applicability has been severely restricted due to the computational bottleneck. A common strategy to break this "curse of sample size" is through s-dimensional random projections of a large kernel matrix with $s \times n$. The main aim of this paper is to examine the effect of s (and also the projection manner) on the performances of local confidence interval and global likelihood ratio testing. In particular, we derive lower bounds on s under which the inferential accuracy of the proposed procedures is not sacrificed, i.e., statistical-and-computational trade-off. Concrete examples together with simulation study are given to support

our general theory. This type of computationally efficient inference (COFFEE) is particularly attractive in the era of big data.

Solving the Identifiability Problem with the Lasso Regularization in Age-period-cohort Analysis

♦*Beverly Fu*[1] *and Wenjiang Fu*[2]
[1]Okemos High School
[2]University of Houston
fubeverly99@gmail.com

Statistical age-period-cohort analysis has been studied extensively in the literature and has many applications to demography, public health, and sociology. However, due to the linear dependence of age, period and cohort, the regression model suffers from the identifiability problem, where multiple estimators fit the model equally well, making it difficult to determine which estimator yields the correct parameter estimation. In this work, we apply the Lasso shrinkage method, which not only helps to determine a unique estimate, but also yields consistent feature selection since the Lasso estimator yields consistent variable selection if the covariates of the model satisfy the irrepresentable condition. We apply the Lasso to the eigenvectors of the design matrix and it sets the coefficient for the null eigenvector to 0, leading to consistent estimation. We will compare the Lasso estimator with the intrinsic estimator in real data examples and illustrate that the Lasso method works well and yields sensible trend estimation in age, period and cohort.

Keywords: Eigenvector, Irrepresentable condition, Lasso, Linear dependence, multiple estimators, Singularity

A general framework for the regression analysis of pooled biomarker assessments

♦*Yan Liu, Christopher McMahan and Colin Gallagher*
Clemson University
yan5@g.clemson.edu

As a cost efficient data collection mechanism, the process of assaying pooled biospecimens is becoming increasingly common in epidemiological research; e.g. pooling has been proposed for the purpose of evaluating the diagnostic efficacy of biological markers (biomarkers). To this end, several authors have proposed techniques that allow for the analysis of continuous pooled biomarker assessments. Regretfully, most of these techniques proceed under restrictive assumptions, are unable to account for the effects of measurement error, and fail to control for confounding variables. These limitations are understandably attributable to the complex structure that is inherent to measurements taken on pooled specimens. Consequently, in order to provide practitioners with the tools necessary to accurately and efficiently analyze pooled biomarker assessments, herein a general Monte Carlo maximum likelihood based procedure is presented. The proposed approach allows for the regression analysis of pooled data under practically all parametric models and can be used to directly account for the effects of measurement error. Through simulation, it is shown that the proposed approach can accurately and efficiently estimate all unknown parameters and is more computational efficient than existing techniques. This new methodology is further illustrated using monocyte chemotactic protein-1 data collected by the Collaborative Perinatal Project in an effort to assess the relationship between this chemokine and the risk of miscarriage.

Permutation inference distribution for linear regression and related models

♦*Qiang Wu and Paul Vos*
East Carolina University
wuq@ecu.edu

For linear regression and related models, such as the two-sample means problem and the analysis of variance (ANOVA) model, the permutation inference distribution (PID) is introduced. The PID is sample-dependent and closely related to the unified Bayesian/Fiducial/Frequentist (BFF) inference framework. Like the confidence distribution in the BFF framework, the PID allows the construction of both confidence intervals and p-values. For two-sample problems and pairwise comparisons in the ANOVA model, a fast Fourier transformation (FFT) method can be used to find the exact PID for small to moderate samples. In general, however, random permutations are required except for small samples where all n! permutations can be generated. Simulation studies and real data applications are used to evaluate inferences obtained from the PID. PID methods are close to standard parametric methods when the errors are iid and normal. For skewed and heavy tailed errors, PID methods are superior to bootstrap and standard parametric methods. Confidence intervals are evaluated using coverage and mean length but also using sample-dependent criteria confidence error and confidence bias.

Comparison of Classical and Quantile Regression Methods for Modeling Childhood Obesity

♦*Gilson Honvoh*[1]*, Roger Zoh*[2]*, Hongwei Zhao*[2]*, Mark Benden*[2]*, Guoyao Wu*[3] *and Carmen Tekwe*[2]
[1]Texas A&M School of Public Health
[2]Texas A&M School of Public Health
[3]Texas A&M University
ghonvoh@sph.tamhsc.edu

The incidence of obesity among children continues to be a growing public health concern. Several approaches have been taken to reduce its prevalence, including targeted environmental interventions designed to increase physical activity. An approach to approximating each child's daily energy expenditure is through the use of body media devices. These devices are worn by the subjects and their caloric energy expenditures are measured every minute. The resulting data are often high dimensional functional data. Additionally, childhood obesity is defined based on levels of age- and sex- adjusted BMI percentiles. For example, a child is considered severely obese if his calculated BMI falls within the 99th percentile levels. In this study, we apply various statistical techniques to determine if energy expenditure at baseline can be used as a less invasive predictor of future obesity among 374 elementary school children exposed to stand-biased desks during two academic years in a Texas school district. Since targeted interventions for obesity are designed to assess how they affect individuals who are at higher risk for being overweight or obese, statistical approaches that are designed to address these questions will provide better answer when compared to approaches designed to answer questions related to how the interventions affect the average child. We compare statistical approaches that allow flexible modeling of the functional covariate, total daily energy expenditure, in both linear and quantile regression settings.

Key words: BMI, B-Splines, Childhood Obesity, Energy Expenditure, Functional Linear Regression, Quantile Regression

Session 69: Recent Development in Dose Finding Studies

A Simple and Efficient Statistical Approach for Designing an Early Phase II Clinical Trial

Yaohua Zhang[1]*, Qiqi Deng*[2]*, Susan Wang*[2] *and* ♦*Naitee Ting*[2]
[1]University of Connecticut
[2]Boehringer-Ingelheim Pharmaceuticals, Inc.

naitee.ting@boehringer-ingelheim.com

There are many challenges in designing early Phase II clinical trials. One reason is that there are many unknowns at this stage of the development, and the other is that the size of the trial at this stage is limited, even though results from such a clinical trial could impact many important decisions and there are high risks associated with each decision. In this manuscript, an ordinal linear contrast test (OLCT) is recommended to help design an efficient early Phase II trial. Performance of the proposed method is compared with MCP-Mod, ANOVA F, and Max T. Results indicate that the performance of ANOVA F and Max T approaches is sub-optimal. OLCT can be comparable with MCP-Mod. In practical applications, OLCT is simple to use, efficient, and robust. For practitioners with limited understanding of MCP-Mod, who have concerns of applying MCP-Mod to their studies, or who are not well versed with the complexity of MCP-Mod software, the OLCT is a useful alternative.

Sample Size Consideration based on Quantitative Decision Framework in Dose Finding Studies

♦*Huilin Hu and Dong Xi*
Novartis Pharmaceuticals Corporation
huilin.hu@novartis.com

The dose finding studies are very important in drug development and need to be planned carefully. The key objectives in dose finding studies are to detect the dose-response signal, to characterize its dose response relationship, and ultimately to determine an adequate dose level for a drug to carry forward to Ph3 for confirmatory testing. Traditional sample size justification for dose-finding studies is typically based on the statistical significance in certain forms (e.g. Dunnett test, trend test, etc). However, this does not fully align with what we wish to achieve at the end in terms of the study outcome, and may lead to insufficient evidence to support the Go and No-Go decision in drug development, such as whether or not there exists a dose for the new treatment that can reach the desired level of improvement compared to the current standard of care. We have investigated different Go/No-Go decision criteria based on what "success" and "failure" may look like in the data at the end for a dose-finding study. For example, it is natural to consider a "success" outcome as not only that a dose-response signal needs to be demonstrated, but also that the estimated dose-response curve should reach a clinically acceptable improvement over the placebo to certain extent, to enable identification of target dose(s). In this talk, we discuss how sample size can affect the operating characteristics of Go/No-Go criteria under different assumptions of true dose response curve and maximum effect size. We demonstrate via simulation how to choose sample sizes using the proposed decision criteria under the framework of the MCPMod methodology.

Improving dose finding studies with MCP-Mod design consideration

♦*Kuo-mei Chen and Jose Pinheiro*
Jassen Research & Development, LLC
kchen51@its.jnj.com

In a traditional Phase II dose-finding study, sample size calculation is commonly done by pairwise comparisons between placebo and several doses with type I error controlled and multiplicity adjusted. Traditional method focuses on assurance of adequate power to detect dose response and the sample size calculated is usually limited. However, the method is not aligned with other important Phase II study objectives: selection of doses or characterization of dose-response relationship. MCP-Mod was qualified by EMA in 2014 as an efficient methodology for design and analysis of dose-finding studies under model uncertainty. Initial sample size determination in MCP-Mod based on the power of the model-based trend test may not be sufficient either to provide precise estimation of the target dose, such as ED50. Modified approaches based on precision of target dose estimation and dose-response relationship will be presented and discussed. A case study and simulation results will be used to demonstrate benefits and limitations of the method.

Session 70: New Advances in High Dimensional and Complex Data Analysis

Probing the Pareto Frontier of Computational-Statistical Tradeoffs

Han Liu
Princeton University
hanliu@princeton.edu

In this talk, we discuss the fundamental tradeoffs between computational efficiency and statistical accuracy in big data. Based on an oracle computational model, we introduce a systematic hypothesis-free approach for developing minimax lower bounds under computational budget constraints. This approach mirrors the classical Le Cam's method, and draws explicit connections between algorithmic complexity and geometric structures of parameter spaces. Based on this approach, we sharply characterize the computational-statistical phase transitions that arise in structural normal mean detection, combinatorial detection in correlation graphs and Markov random fields, as well as sparse principal component analysis. Moreover, we resolve several open questions on the computational barriers arising in sparse mixture models, sparse phase retrieval, and tensor component analysis (Based on joint work with Zhaoran Wang, Quanquan Gu, and Zhaoran Yang).

Factor Adjusted Graphlet Screening

♦*Tracy Ke and Fan Yang*
University of Chicago
zke@galton.uchicago.edu

We consider the high-dimensional linear regression problem. Marginal Screening (MS, also known as sure screening) is an easy-to-implement variable selection method, but it is known to have the issue of "signal cancellation" when the design is not orthogonal. We propose a new screening method, Factor Adjusted Graphlet Screening (FA-GS), which improves MS for a large class of designs – the Gram matrix decomposes into the sum of a low rank matrix and a sparse matrix. FA-GS consists of two steps: the FA step applies PCA to construct a new linear model the Gram matrix of which is sparse, and the GS step is a non-trivial modification of the recent idea of Graphlet Screening.

Graphlet Screening generalizes MS from screening one variable at a time to screening a small number of variables at a time. It uses a graph to guide the screening so that the computational cost is only moderately larger than that of MS. Graphlet Screening is shown to successfully overcome "signal cancellation" when the Gram matrix is sparse. Compared to Graphlet Screening, first, FA-GS is able to handle non-sparse Gram matrices. Second, FA-GS produces a unique rank of all the variables, so that the users can conveniently select any target number of variables; the original Graphlet Screening does not have such a nice property.

We derive the convergence rates of both the number of false positives and the number of false negatives for FA-GS. We also show that, under mild conditions, not only the number of false negatives can be well-controlled, but also the set of remaining variables satis-

fies the "Separable After Screening" property. The theoretical analysis also includes a new result about the connection between PCA and factor analysis.

Truth, Knowledge, P-Values, Bayes, & Inductive Inference

Edel Pena
University of South Carolina
pena@stat.sc.edu

In the past few years the use of P-values in the context of scientific research has seen much, sometimes heated, discussions. The American Statistical Association was even compelled to release an official statement in early March 2016 regarding this issue, and a psychology journal has gone to the extreme of banning the use of P-values in articles appearing in its journal. This debate has also been in relation to important issues of reproducibility in scientific research. In fact, this debate goes to the core of inductive inference and the different schools of thought (significance testing approach, Neyman-Pearson paradigm, Bayesian approach, etc.) on how inductive inference should be done. In this talk I would like to delve into these issues and to offer some viewpoints on whether P-values should be relegated to the the dustbin of history or whether it will be there to stay as a tool of scientific investigations. In particular, I will touch on the representation of knowledge and its updating based on observations, and ask the question: "When given the p-value, what does it provide in the context of the updated knowledge of the phenomenon under consideration?",Edel,Pena,University of South Carolina

Regression in heterogeneous problems

Hanwen Huang
UGA
huanghw@uga.edu

We develop a new framework for modeling the impact of sub-cluster structure of data on regression. The proposed framework is specifically designed for handling situations where the sample is not homogeneous in the sense that the response variables in different regions of covariate space are generated through different mechanisms. In such situation, the sample can be viewed as a composition of multiple data sets each of which is homogeneous. The traditional linear and general nonlinear methods may not work very well because it is hard to find a model to fit multiple data sets simultaneously. The proposed method is flexible enough to ensure that the data generated from different regions can be modeled using different functions. The key step of our method incorporates the k-means clustering idea into the traditional regression framework so that the regression and clustering tasks can be performed simultaneously. The k-means clustering algorithm is extended to solve the optimization problem in our model that groups the samples with similar response-covariate relationship together. General conditions under which the estimation of the model parameters is consistent are investigated. By adding appropriate penalty terms, the proposed model can conduct variable selection to eliminate the uninformative variables. The conditions under which the proposed model can achieve asymptotic selection consistency are also studied. The effectiveness of the proposed method is demonstrated through simulations and real data analysis.

Session 71: Design of Experiments I

Minimax designs using clustering

♦*Simon Mak and V. Roshan Vengazhiyil*
Georgia Institute of Technology
smak6@gatech.edu

Minimax designs provide a uniform coverage of a design space $\mathcal{X} \subseteq \mathbb{R}^p$ by minimizing the maximum distance from any point in this space to its nearest design point. Although minimax designs have many useful applications, e.g., for optimal sensor allocation or as space-filling designs for computer experiments, there has been little work in developing algorithms for generating these designs. In this paper, a new clustering-based method is presented for computing minimax designs on any convex and bounded design region. The computation time of this algorithm scales linearly in dimensionality p, meaning our method can generate minimax designs efficiently for high-dimensional regions. Simulation studies and a real-world example show that the proposed algorithm provides improved minimax performance over existing methods on a variety of design regions. Finally, we introduce a new type of experimental design called a minimax projection design, and show that this proposed design provides better minimax performance on projected subspaces of $\mathcal{X}$ compared to existing designs.

Optimal Experimental Designs for Nonlinear Conjoint Analysis

Mercedes Esteban-Bravo[1], ♦*Agata Leszkiewicz*[2] *and Jose M. Vidal-Sanz*[1]
[1]Universidad Carlos III de Madrid
[2]Georgia State University
aleszkiewicz@gsu.edu

Estimators of choice-based multi-attribute preference models have a covariance matrix that depends on both the design matrix as well as the unknown parameters to be estimated from the data. As a consequence, researchers cannot optimally design the experiment (minimizing the variance). Several approaches have been considered in the literature, but they require prior assumptions about the values of the parameters that often are not available. Furthermore, the resulting design is neither optimal nor robust when the assumed values are far from the true parameters. In this paper, we develop efficient worst-case designs for the choice-based conjoint analysis which accounts for customer heterogeneity. The contributions of this method are manifold. First, we account for the uncertainty associated with ALL of the unknown parameters of the mixed logit model (both the mean and the elements in covariance matrix of the heterogeneity distribution). Second, we allow for the unknown parameters to be correlated. Third, this method is also computationally efficient, which in practical applications is an advantage over e.g. fully Bayesian designs. We conduct multiple simulations to evaluate the performance of this method. The worst case designs computed for the logit and mixed logit models are indeed more robust than the local and Bayesian benchmarks, when the prior guess about the parameters is far from their true values.

Obtaining locally D-optimal designs for binary response experiments via Particle Swarm Optimization

♦*Joshua Lukemire*[1], *Abhyuday Mandal*[2] *and Weng Kee Wong*[3]
[1]Emory University
[2]University of Georgia
[3]University of California at Los Angeles
joshlukemire@gmail.com

Obtaining optimal designs for experiments in which the outcome takes a binary response and is modeled by a generalized linear model is a difficult task due to the dependence of the optimal design on the model parameters. Theoretical results for these design problems are often unavailable, and instead computational methods must be used to obtain optimal designs. There are many popular such methods, however they generally require either an explicit ob-

jective function or a set of candidate design points in order to work. For experiments with mixed discrete and continuous factors this requirement is often problematic. It is often impossible to obtain an explicit objective function, and sets of candidate design points obtained by discretizing the continuous factor(s) will quickly grow prohibitively large. In this work, we demonstrate the use of a Particle Swarm Optimization algorithm for obtaining locally D-optimal designs that allows us to avoid both of these problems. Results are demonstrated for the redesign of an odor-removal experiment conducted at the University of Georgia.

Session 72: Recent Advancement about Adaptive Design in all Phases of Clinical Trial

Key Statistical Issues in Adaptive Design in Oncology Trials Application

♦*Qi Jiang and Chunlei Ke*

Amgen

qjiang@amgen.com

The draft adaptive design guidance released by FDA included a comprehensive overview of adaptive study designs. The flexibility of these designs in modifying trial procedures and potentially reducing trial size and duration could help to increase trial success rates. As a result, there has been increased emphasis placed on using adaptive designs. In this presentation we'll share some thoughts including some challenges and best practices regarding the key statistical and operational issues related to adaptive design in oncology trials in order to further encourage greater use of these designs.

Early phase trial designs

Sumithra Mandrekar

Mayo Clinic

mandrekar.sumithra@mayo.edu

In the current era of stratified medicine and targeted therapies, the focus has shifted from predictions based on the traditional anatomic staging systems to guide the choice of treatment for an individual patient to an integrated approach using the genetic makeup of the tumor and the genotype of the patient. In oncology, early phase trials with targeted therapeutics have mandated the development of novel study design strategies as the historical clinical trial design paradigm was no longer relevant. This talk will provide an overview of some of these designs while focusing on two specific design strategies: one in the phase I / dose-finding setting and the other in a phase II setting. In the dose finding setting, a continuous toxicity score utilizing information about multiple toxicity types and grades instead of the binary dose limiting toxicity endpoint has been developed as targeted therapies are often administered for multiple cycles and the toxicity profile for these agents are relatively mild. Dose finding designs utilizing the continuous score from just the first cycle of treatment versus multiple treatment cycles will be discussed. In the setting of phase II trials, biomarkers are increasingly utilized in the study design to help identify patients more likely to benefit from a treatment. One such design is the direct assignment design, which includes an option for direct assignment to the experimental treatment, when there is promising, but not definitive, evidence of a treatment benefit at the end of an initial randomized stage of the trial. This provides for an "extended confirmationphase" as an alternative to stopping the trial early for evidence of efficacy after the initial stage.

Session 73: Recent Advances on Statistical Analysis of Safety and/or Efficacy Endpoints in Clinical Trials

On the Restricted Mean Survival Time Curve in Survival Analysis

♦*Lihui Zhao*[1]*, Brian Claggett*[2]*, Lu Tian*[3]*, Hajime Uno*[4]*, Lorenzo Trippa*[5] *and Lee-Jen Wei*[5]

[1]Northwestern University

[2]Harvard Medical School

[3]Stanford University

[4]Dana-Farber Cancer Institute

[5]Harvard University

lihui.zhao@northwestern.edu

For a study with an event time as the endpoint, its survival function contains all the information regarding the temporal, stochastic profile of this outcome variable. The survival probability at a specific time point, say t, however, does not transparently capture the temporal profile of this endpoint up to t. An alternative is to use the restricted mean survival time (RMST) at time t to summarize the profile. The RMST is the mean survival time of all subjects in the study population followed up to t, and is simply the area under the survival curve up to t. The advantages of using such a quantification over the survival rate have been discussed in the setting of a fixed-time analysis. In this article, we generalize this approach by considering a curve based on the RMST over time as an alternative summary to the survival function. Inference, for instance, based on simultaneous confidence bands for a single RMST curve and also the difference between two RMST curves are proposed. The latter is informative for evaluating two groups under an equivalence or noninferiority setting, and quantifies the difference of two groups in a time scale. The proposal is illustrated with the data from two clinical trials, one from oncology and the other from cardiology.

Bayesian methods for meta-analysis combining randomized-controlled and single-arm studies

♦*Jing Zhang*[1]*, Chia-Wen Ko*[2]*, Lei Nie*[2]*, Yong Chen*[3] *and Ram Tiwari*[2]

[1]University of Maryland

[2]U.S. Food and Drug Administration

[3]University of Pennsylvania

jzhang86@umd.edu

Meta-analysis of interventions usually relies on randomized controlled trials (RCTs). However, when the dominant source of information comes from single-arm studies, or when the results from RCTs lack generalization due to strict inclusion and exclusion criteria, it is vital to synthesize both sources of evidence. One challenge of synthesizing both sources is that single-arm studies are usually less reliable than RCTs due to selection bias and confounding factors. In this paper, we propose a Bayesian hierarchical framework for the purpose of bias reduction and efficiency gain. Under this framework, six models are proposed: two of them treat single-arm studies equally with RCTs, two adjust for design difference and potential biases, and the rest two further downweight single-arm studies adaptively. Hierarchical power prior model and hierarchical commensurate prior model are recommended as primary methods for evidence synthesis. We illustrate our methods by applying all six models to two motivating datasets and evaluate their performance through simulation studies. We finish with a discussion of the advantages and limitations of our methods, as well as directions for future research in this area.

Quantifying treatment benefit in molecular subgroups to assess a predictive biomarker.

♦*Alexia Iasonos and Jaya Satagopan*
Memorial Sloan Kettering Cancer Center
iasonosa@mskcc.org

There is an increased interest in finding predictive biomarkers that can guide treatment options for both mutation carriers and non-carriers. The statistical assessment of variation in treatment benefit (TB) according to the biomarker carrier status plays an important role in evaluating predictive biomarkers. For time to event endpoints, the hazard ratio (HR) for interaction between treatment and a biomarker from a Proportional Hazards regression model is commonly used as a measure of variation in treatment benefit. While this can be easily obtained, the interpretation of HR is not straightforward. In this talk, we present two summary measures of differential TB on the scale of survival probabilities for evaluating a predictive biomarker. The proposed summary measures can be interpreted in terms of relative risk or excess absolute risk due to treatment in carriers versus non-carriers. We illustrate the use and interpretation of the proposed measures using data from completed clinical trials and evaluate their operating characteristics via simulations.

Bayesian proportional hazards model for interval censored data in cancer clinical trials

♦*Ai Ni, Zhigang Zhang and Mithat Gonen*
Memorial Sloan Kettering Cancer Center
nia@mskcc.org

In cancer clinical trials, progression free survival (PFS) is frequently used as an alternative endpoint to overall survival to accelerate the new drug development. The disease progression status is determined by clinical examinations such as CT scan, which usually occurs at scheduled clinical visits. Thus, PFS is naturally interval censored. Current practice typically ignores the interval censoring and treats the time when progression is first observed as the actual progression time. This can lead to biased estimation of treatment effect on PFS. In this study, we applied a recently developed Bayesian proportional hazards model for general interval censored data (Lin et al 2015) to a phase II randomized controlled trial to demonstrate its application. The trial sought to compare two embolization methods in patients with late stage liver cancer. One of the endpoint is PFS which is interval censored due to fixed clinical evaluation times. We analyzed the data using the Bayesian proportional hazards model and compared the results with naíve analysis that ignores interval censoring. We showed the importance of taking into account the interval censoring when analyzing PFS data. The Bayesian proportional hazards model for interval censored data is a flexible and efficient tool to analyze interval censored data.

Session 74: New Statistical Methods for Analysis of Large-Scale Genomic Data

Statistical Inference for Time Course RNA-Seq Data

Xiaoxiao Sun[1]*, Dalpiaz David*[2]*, Di Wu*[3]*, Jun Liu*[3] *and* ♦*Ping Ma*[1]
[1]University of Georgia
[2]University of Illinois at Urbana-Champaign
[3]Harvard University
pingma@uga.edu

Accurate identification of differentially expressed (DE) genes in time course RNA-Seq data is crucial for understanding the transcriptional regulatory network. Since gene expression profiles have many different trajectories, identification of DE genes is much more challenging in time course RNA-Seq data compared to that in static data. In this talk, I present a negative binomial mixed-effect model (NBMM) to identify DE genes in time course RNA-Seq data. In the NBMM, gene expression is characterized by a fixed effect, and time dependence is described by random effects. The NBMM method is very flexible and can be fitted to both unreplicated and replicated time course RNA-Seq data via penalized likelihood method. By comparing gene expression profiles over time, we further classify the DE genes into two subtypes to enhance the understanding of expression dynamics. A significance test for detecting DE genes is derived using a Kullback-Leibler distance ratio. Additionally, a significance test for gene sets is developed using a gene set score. Simulation analysis shows that the NBMM method outperforms currently available methods for detecting DE genes and gene sets. Moreover, our real data analysis of fruit fly developmental time course RNA-Seq data demonstrates our NBMM model identifies biologically relevant genes which are well justified by gene ontology analysis.

Nonparametric regularized regression for network construction and taxa selection

♦*Wenchuan Guo*[1]*, Zhenqiu Liu*[2] *and Shujie Ma*[1]
[1]University of California, Riverside
[2]Cedars-Sinai Medical Center
wguo007@ucr.edu

The network structure of taxa can be affected by the disease associated environmental conditions. In addition, taxa abundance is differentiated under conditions. Therefore, knowing how the correlation or relative abundance changes with these factors would be of great interest to researchers. We develop a non-parametric regularized regression method to construct taxa association networks under different clinical conditions. We let the coefficients be unknown functions of the environmental variable. The varying coefficients are estimated by using regression splines. The proposed method is regularized with concave penalties and an efficient group descent algorithm is used for computation. We also apply the varying coefficient model to estimate taxa abundance in order to see how it changes across different enviromental conditions. Moreover, for conducting inference, we propose a bootstrap method to construct the simultaneous confidence bands for the corresponding coefficients. We use different simulated designs and a real data set to demonstrate that our method can identify the network structures successfully under different environmental conditions. As such, the proposed method has potential applications for researchers to construct differential networks and identify taxa.

Statistical inference of allele-specific contacts from high-throughput chromosome conformation data

♦*Wenxiu Ma*[1]*, Xinxian Deng*[2]*, Vijay Ramani*[2]*, Zhijun Duan*[2]*, Jay Shendure*[2]*, William Noble*[2] *and Christine Disteche*[2]
[1]University of California Riverside
[2]University of Washington
wenxiu.ma@ucr.edu

High-throughput methods based on chromosome conformation capture have greatly advanced our understanding of the three-dimensional (3D) organization of genomes. However, little is known about the allele-specific chromatin conformation in diploid genomes and how the differences in homologous chromosome 3D foldings affect allelic gene expression. Here we have developed an Empirical Bayes hierarchical model to infer allele-specific chromatin contacts from DNase Hi-C data. DNase Hi-C produces higher-resolution and less biased chromatin maps than regular Hi-C. By applying this novel Hi-C method and new statistical and

computational methods to map allelic chromatin contacts in hybrid mouse systems, we discovered a specific bipartite organization of the mouse inactive X chromosome that probably plays an important role in maintenance of gene silencing on the inactive X.

Metagenomics Binning via sequential Monte Carlo (SMC) method

Xinping Cui, Chen Gao and ♦*Wei Cui*
University of California, Riverside
weicui@ucr.edu

Metagenomics is the study of DNA of microorganisms from environmental samples without cultivation and isolation. Recently high-throughput sequencing technology (HTS) efficiently sequence metagenomic DNA reads of multiple species. However, those mixed short reads make the separation of reads from different species more challenging. Existing binning algorithms fall into two main categories, supervised methods and unsupervised method. Supervised methods may leave a large fraction of reads unclassified due to low rate of related reference in database, while the performance of unsupervised methods rely heavily on the long length of reads and assumptions on the number of species. In this work, we present a novel algorithm based on sequential Monte Carlo (SMC) technique which incorporates Markovian structure of the nucleotide reads. The new algorithm has high binning accuracy and can determine the number of species automatically.

Session 75: New Development in Function Data Analysis

Nonlinear function on function regression model

♦*Xin Qi and Ruiyan Luo*
Georgia State University
xqi3@gsu.edu

We consider a nonlinear generalization of the linear function on function regression model. We replace the integral of the linear function of the predictive curve in the linear function on function model by the integral of a nonlinear function of the predictive curve. The form of the nonlinear function is unknown and we only assume that it is smooth and its value converges to zero when the argument of the function goes to infinity. Instead of estimating the nonlinear function, we focus on the predictive ability of the estimated model. We propose a signal approximation approach where the model is estimated through a penalized nonlinear optimization problem. We provide the upper bounds for the prediction error and estimation error of the signal function for the estimated model and propose algorithms to solve he optimization problem.

Bayesian Registration of Functions with a Gaussian Process Prior

♦*YI LU, Sebastian Kurtek and Radu Herbei*
The Ohio State University
morning820@gmail.com

We present a Bayesian method to register real-valued functional data by making inference on nonlinear time warping functions. We adopt transformations that are developed in a differential geometric framework to map the parameter space of warping functions into a linear space, which allows the definition of a Gaussian process prior distribution. Avoiding discretization until the last step, we keep the infinite dimensionality of the warping functions when we write down the model. Draws from the posterior distribution of those warping functions are then obtained by a novel MCMC algorithm that utilizes the Karhunen-Loeve expansion to approximate a continuous function with a finite number of basis functions. Furthermore, we extend our method to simultaneously register more than two functions, in which case we also make inference on the underlying template curve that all of the observed functions are matched to. We apply our method to real data sets including growth rate curves in the Berkeley growth study and kinematic measurements obtained in a gait cycle study.

Fast Bayesian inference for complex, ultra-high dimensional functional data

♦*Hongxiao Zhu*[1]*, Fengrong Wei*[2] *and Xiaowei Wu*[1]
[1]Virginia Tech
[2]University of West Georgia
hongxiaozhu@hotmail.com

Rapidly expanding modern technology enables automatic collection of high-dimensional data of functional form, such as neuroimages, genomic/epigenetic measurements, and sensor signals in engineering. While the data are functional in nature, existing approaches are mostly based on univariate/multivariate analysis, and many cutting-edge functional data analysis tools are not applicable due their computational limits. We propose a fast Bayesian inference framework that facilitates functional data regression for ultra-high dimensional functional measurements with dimension on the scale of O(1e6). This approach integrates compressive sensing with Bayesian approximate inference and parallel computing, achieving scalable, accurate, and robust inference. This novel framework is suitable for a large family of functional regression setups, and incorporates functional data with various complex structures, including hierarchical structures and spatial-temporal correlations. We demonstrate the performance of this framework through brain image and sonar sensing data.

Function on function regression with thousands of predictive curves

♦*Ruiyan Luo and Xin Qi*
Georgia State University
rluo@gsu.edu

Motivated by recent simultaneous EEG and fMRI data, we consider function-on-function linear regression models with thousands of predictive curves. We aim at finding an estimate of the model which has a good predictive ability. This goal is closely related to estimating the best random approximation to the signal in the response function. We establish the relationship between the best random approximation and the predictive and coefficient functions through a generalized functional eigenvalue problem. Then we propose a penalized generalized eigenvalue problem to estimate the best smooth random approximation and make prediction on the response function. We provide asymptotic upper bounds for the estimation error and prediction error when both the sample size and the number of predictive curves go to infinity. We apply the proposed method to a simultaneous EEG and fMRI data.

Session 76: Some Recent Developments in Robust High-dimensional Data Analysis

Fast Community Detection in Complex Networks with a K-Depths Classifier

Yahui Tian and ♦*Yulia Gel*
University of Texas at Dallas, USA
ygl@utdallas.edu

We introduce a notion of data depth for recovery of community structures in large complex networks. We propose a new data-driven algorithm, K-depths, for community detection using the L_1-depth

in an unsupervised setting. We evaluate finite sample properties of the K-depths method using synthetic networks and illustrate its performance for tracking communities in online social media platform Flickr. The new method significantly outperforms the classical K-means and yields comparable results to the regularized K-means. Being robust to low-degree vertices, the new K-depths method is computationally efficient, requiring up to 400 times less CPU time than the currently adopted regularization procedures based on optimizing the Davis-Kahan bound.

ROCKET: Robust Confidence Intervals via Kendall's Tau for Transelliptical Graphical Models

♦*Mladen Kolar*[1] *and Rina Foygel Barber*[2]
[1]The University of Chicago Booth School Of Business
[2]University of Chicago
mkolar@chicagobooth.edu

Undirected graphical models are used extensively in the biological and social sciences to encode a pattern of conditional independences between variables, where the absence of an edge between two nodes a and b indicates that the corresponding two variables are believed to be conditionally independent, after controlling for all other measured variables. In the Gaussian case, conditional independence corresponds to a zero entry in the precision matrix (the inverse of the covariance matrix). Real data often exhibits heavy tail dependence between variables, which cannot be captured by the commonly-used Gaussian or nonparanormal (Gaussian copula) graphical models. We study the transelliptical model, an elliptical copula model that generalizes Gaussian and nonparanormal models to a broader family of distributions. We propose the ROCKET method, which constructs an estimator of Ω_{ab} that we prove to be asymptotically normal under mild assumptions. Empirically, ROCKET outperforms the nonparanormal and Gaussian models in terms of achieving accurate inference on simulated data. We also compare the three methods on real data (daily stock returns), and find that the ROCKET estimator is the only method whose behavior across subsamples agrees with the distribution predicted by the theory.

Model diagnostics and robust estimation in low-rank models

Kun Chen
University of Connecticut
kun.chen@uconn.edu

Reduced-rank methods are popular in high-dimensional multivariate analysis for conducting simultaneous dimension reduction and model estimation. However, the commonly-used reduced-rank methods are not robust, as the underlying reduced-rank structure can be easily distorted by a few data outliers. Anomalies are bound to exist in big data problems, and yet in some applications they themselves could be of the primary interest. While naive residual analysis is often inadequate for outlier detection due to potential masking and swamping, robust reduced-rank estimation approaches could be computationally demanding. Under Stein's unbiased risk estimation framework, we propose a set of tools, including leverage score and generalized information score, to perform model diagnostics and outlier detection in large-scale reduced-rank estimation. The leverage scores give an exact decomposition of the so-called model degrees of freedom, which lead to exact decomposition of many commonly-used information criteria; the resulting quantities are thus named information scores. The proposed information score approach provides a principled way of combining the residuals and leverage scores for anomaly detection. Simulation studies, a pattern recognition example, and a time series modeling application using monthly U.S. macroeconomic data demonstrate the efficacy of the proposed approach. We also discuss some recent developments in joint robust reduced-rank estimation.

Variable selection for partially linear models via learning gradients

Lei Yang[1], ♦*Yixin Fang*[1], *Junhui Wang* [2] *and Yongzhao Shao*[1]
[1]New York University School of Medicine
[2]City University of Hong Kong
yf2113@gmail.com

Partially linear models, a compromise between parametric regression and non-parametric regression models, are very useful for analyzing high-dimensional data. Variable selection plays an important role in the use of partially linear models, which are of both linear and non-linear components. Variable selection for the linear component is well studied. However, variable selection for the non-linear component usually relies on some assumption imposed on the structure of the non-linear component. For example, variable selection methods have been developed for additive partially linear models and generalized additive partially linear models. In this manuscript, we propose a new variable selection method based on learning gradients for partially linear models without any assumption on the structure of the non-linear component. The proposed method utilizes the reproducing-kernel-Hilbert-space tool to learn the gradients and the group-lasso penalty to select variables. In addition, a block-coordinate descent algorithm is described and some theoretical properties are derived. The performance of the proposed method is evaluated via simulation studies and a real data application.

Session 77: Recent Advances in Statistical Methods for Handling Missing Data

Combining IRT with Multiple Imputation to Crosswalk between Health Assessment Questionnaires

Chenyang Gu and ♦*Roee Gutman*
Brown University
roee_gutman@brown.edu

The assessment of patients' functional status across the continuum of care requires a common patient assessment tool. Different health care settings rely on different assessment tools that cannot be easily contrasted. For example, the Functional Independence Measure (FIM) is used to evaluate the functional status of patients who stay in inpatient rehabilitation facilities (IRFs). After discharge from rehabilitation facilities, for patients that are transfer to a skilled nursing facilities (SNFs), the Minimum Data Set (MDS) is collected, while the Outcome and Assessment Information Set (OASIS) is collected for patients using home health care provided by home health agencies (HHAs). To compare patients that are discharged from rehabilitation facility to either SNFs or home health, a single measure of functionality is required. We assume that all patients have observed FIM measurements and treat the unmeasured MDS or OASIS items as missing. We propose a variant of the predictive mean matching method that relies on Item Response Theory (IRT) models to impute the missing measurement items. Using real data sets, we simulated missing measurements and compared our proposed approach to existing methods for missing data imputation. For all of the estimands that were examined, the proposed approach was generally valid and had the best operating characteristics.

A new framework for addressing selection bias due to missing

data in EHR-based research

♦*Sebastien Haneuse*[1] *and Michael Daniels*[2]
[1]Harvard T.H. Chan School of Public Health
[2]University of Texas - Austin
shaneuse@hsph.harvard.edu

Electronic health records (EHR) data are increasingly seen as a resource for cost-effective comparative effectiveness research (CER). Since EHR data are collected primarily for clinical and/or billing purposes, their use for CER requires consideration of numerous methodologic challenges including the potential for confounding bias, due to a lack of randomization, and for selection bias, due to missing data. In contrast to the recent literature on confounding bias in EHR-based CER, virtually no attention has been paid to selection bias possibly due to the belief that standard methods for missing data can be readily-applied. Such methods, however, hinge on an overly simplistic view of the available/missing EHR data, so that their application in the EHR setting will often fail to completely control selection bias. Motivated by challenges we face in an on-going EHR-based comparative effectiveness study of choice of antidepressant treatment and long-term weight change, we propose a new general framework for selection bias in EHR-based CER. Crucially, the framework provides structure within which researchers can consider the complex interplay between numerous decisions, made by patients and health care providers, which give rise to health-related information being recorded in the EHR system, as well as the wide variability across EHR systems themselves. This, in turn, provides structure within which: (i) the transparency of assumptions regarding missing data can be enhanced, (ii) factors relevant to each decision can be elicited, and (iii) statistical methods can be better aligned with the complexity of the data.

Imputing cost data that are missing not at random in SEER-Medicare linked data

Rebecca Andridge
The Ohio State University College of Public Health
andridge.1@osu.edu

The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program is one of the premier cancer surveillance programs in the world. The information collected on each and every cancer patient in SEER include demographics, a description of their cancer, limited initial treatment information, and patient follow-up including cause of death for deceased patients. SEER registries collect information on first-course treatment, but the SEER program does not release data on chemotherapy or hormone therapy due to uncertainties regarding data completeness. Recently, SEER chemotherapy data was augmented for older cancer patients (age 65+) with Medicare claims (e.g., cost data), but Medicare Part D claims were excluded. Since only approximately 50% of Medicare beneficiaries have Part D coverage, and those with coverage tend to be sicker and poorer, this is a complex missing data problem. In this talk we will review the scope of the missing data problem and discuss imputation strategies based on available patient and area-level socio-demographic information to inform usage about prescription drugs for the Medicare population without Part D coverage. In particular, we will consider sensitivity analyses to investigate the sensitivity of key analysis parameters to missing not at random mechanisms.

Nonparametric Imputation for non-ignorable missing data

Domonique Hodge[1]*, Chiu-Hsieh Hsu*[2] *and* ♦*Qi Long*[1]
[1]Emory University
[2]The University of Arizona
qlong@emory.edu

Imputation is widely used for handling missing data due to several attractive features including its ease of use. However, the vast majority of imputation techniques are designed for ignorable missing data. Under nonignorable missingness, the existing approaches typically require modeling the variables with missing values and the missing data mechanism jointly. As such, they are sensitive to the misspecification of the two models. We propose a more robust, nonparametric technique to impute missing data. Using the two models, we derive predictive scores to achieve dimension reduction and use the resulting scores coupled with a nearest neighbor hot deck to multiply impute the missing values. The nearest neighbor imputation step allows users to choose weights in defining distance between observations such that the resulting estimator can rely more heavily on the model that is more likely to be correctly specified. Our proposed approach is shown in simulations to outperform several existing multiple imputation methods for nonignorable missing data and is further illustrated using a real data example from the Georgia Coverdell Acute Stroke Registry.

Session 78: Statistical Research in Clinical Trials

A Statistical Approach for Clinical Operations in a Long-term Schizophrenia Study

Jun Zhao
AbbVie
zhao.jun@abbvie.com

Late stage confirmatory clinical trials are often large and well designed in order to confirm the efficacy and safety profile observed in the earlier stage trials. In order to move a product forward to registration and eventually market approval, clinical teams need to work together collaboratively and plan well in advance. The sample size of a study is one key design element, and it is linked to the power of the study to detect a predefined effect size by using planned statistical model and testing at a pre-defined significant level. The sample size also operationally links with the overall study cost, sites and regions selection and recruitment timeline of the study. During the conduct of a clinical trial, we may observe that the data from the trial itself might be away from our initial expectation. In this talk, we present our experience in a long-term schizophrenia study for maintenance of treatment effect, i.e., a randomized withdraw relapse prevention study, on how the clinical team worked collaboratively to assess the trial design assumptions, to evaluate the patient recruitment strategy, and to use statistical modeling in supporting the operational decision of the study.

Sensitivity Analyses for the Primary Efficacy Endpoint for Kalydeco R117H sNDA Submission

♦*Lan Lan*[1] *and Mei-Hsiu Ling*[2]
[1]Vertex
[2]Sr. Director
lan_lan@vrtx.com

In preparation for the Kalydeco R117H FDA Advisory Committee Meeting (ACM) for the sNDA submission, post-hoc sensitivity analyses were conducted to justify the trial data, considering the small sample size of the trial, the outlier and missing data. In this talk, the Wei-Lachin tests, the permutation tests and the interactions tree analyses will be briefly introduced, which provided insights to the justification of the trail data.

Estimating Optimal Treatment Regimes via Subgroup Identification in Randomized Control Trials and Observational Studies

Haoda Fu

Eli Lilly and Company
fuhaoda@gmail.com

With new treatments and novel technology available, personalized medicine has become an important piece in the new era of medical product development. Traditional statistics methods for personalized medicine and subgroup identification primarily focus on single treatment or two-arm randomized controlled trials. Motivated by the recent development of outcome weighted learning framework, we propose an alternative algorithm to search treatment assignments which has a connection with subgroup identification problems. Our method focuses on applications from clinical trials to generate easy-to-interpret results. This framework is able to handle two or more than two treatments from both randomized control trials and observational studies. We implement our algorithm in C++ and connect it with R. Its performance is evaluated by simulations, and we apply our method to a dataset from a diabetes study.

Session 79: Modeling and Analyzing Medical Device Data

A two-stage model for wearable device data

♦*Jiawei Bai, Yifei Sun, Ciprian Crainiceanu and Mei-Cheng Wang*
Johns Hopkins University
jiawei.bai@jhu.edu

Recent advances of wearable computing technology have allowed continuous health monitoring in large observational studies and clinical trials. Examples of data collected by wearable devices include minute-by-minute physical activity proxies measured by accelerometers or heart rate. The analysis of data generated by wearable devices has so far been quite limited to crude summaries, for example, the mean activity counts over the day. To better account for the temporal and population variability, we introduce a two-stage regression model for the minute-by-minute physical activity proxy data. Our model is designed to capture both the transition dynamics between active/inactive periods (stage 1) and activity intensity dynamics during active periods (stage 2). The approach is able to account for the high-dimensionality and time-dependence of the high density data generated by wearable devices. Methods are motivated by and applied to the Baltimore Longitudinal Study of Aging.

A functional data analysis framework for accelerometry data

♦*Chongzhi Di*[1]*, David Buchner*[2]*, Andrea LaCroix*[3] *and Ross Prentice*[1]
[1]Fred Hutchinson Cancer Research Center
[2]University of Illinois Urbana-Champaign
[3]Unviersity of California at San Diego
cdi@fredhutch.org

In large-scale epidemiological studies, it is increasing common to record physical activity objectively by wearable accelerometers. Accelerometry data are time series that allow more precise measurement of the intensity, frequency and duration of physical activity than self-reported questionnaires. However, standard analysis often reduce the high-resolution data into a few simple summary measures, which depends on choices of cut points and can be oversimplied. We develop a functional data framework for the analysis of accelerometry data. We first introduce functional indices to describe the profile of activity intensity, frequency and duration. These indices are then used as outcomes or predictors in functional regression analysis, which allows estimation of detailed dose-response relationship between activity patterns and health outcomes. These methods are motivated by and applied to the Objective Physical Activity and Cardiovascular Health Study among older women, where the aim is to study the association between objectively measured physical activity and cardiovascular diseases.

Variable selection in the concurrent functional linear model

♦*Jeff Goldsmith*[1] *and Joseph Schwartz*[2]
[1]Columbia University, Department of Biostatistics
[2]Columbia University Medical Center
ajg2202@cumc.columbia.edu

Methods for variable selection are developed for models of the association between a functional response and functional predictors that are observed on the same domain. This data structure, and the need for such methods, is exemplified by our motivating example: a study in which blood pressure values are observed throughout the day together with measurements of physical activity, heart rate, location, posture, attitude, and other quantities that may influence blood pressure. We estimate the coefficients of the concurrent functional linear model using variational Bayes and jointly model residual correlation using functional principal components analysis. Latent binary indicators partition coefficient functions into included and excluded sets, incorporating variable selection into the estimation framework. The proposed methods are evaluated in simulated- and real-data analyses.

Accelerometers, physical activity, and conditional random fields

♦ *John Staudenmayer and Evan Ray*
UMass-Amherst
jstauden@math.umass.edu

We consider the problem of recognizing aspects of physical activity from body worn accelerometers. The accelerometers provide vector valued observations at each time point, and the statistical problem is to classify the concurrent activity that the wearer is doing. We compare several approaches: hidden Markov models, conditional random fields, and static models. We apply our methods to simulated data and to several real datasets where what the wearers are actually doing is known from direct observation.

Session 80: Data Ming and Big Data Analysis

Wisdom of Crowds: Meta-analysis of Gene Set Enrichment Studies Utilizing Isoform Expression

♦*Xinlei Wang, Lie Li and Guanghua Xiao*
Southern Methodist University
swang@smu.edu

To understand molecular mechanisms underlying complex human diseases, one important task in transcriptome studies is to identify groups of related genes that are combinatorially involved in such biological processes, mainly through Gene Set Enrichment Analysis (GSEA). In the past, many biomedical studies have achieved spectacular successes with the aid of GSEA in the innovation of disease prevention and intervention strategies. However, in the dawn of a big data era, there is an increasingly urgent need to perform iGSEA (i.e., integrative analysis of multiple relevant GSEA studies), to avoid indecisive or potentially conflicting conclusions from individual data and to leverage "wisdom of crowds" for more effective and reliable scientific discoveries. Further, in the wake of next generation sequencing technologies, it has been made possible to measure genome-wide isoform-specific expression levels, calling for innovations that can utilize the unprecedented resolution. This research will be the first study to develop iGSEA methods that allow statistically efficient use of isoform-specific expression from multiple RNA-seq experiments. Unlike existing meta-analysis methods for GSEA that use ad-hoc summary statistics such as maximum,

minimum and sum, our approaches are model based, which integrate ideas from fixed-effect and random-effects methods that have been newly developed for GWAS meta-analysis. In addition, our methods are general-purpose: it can be applied to both discrete and continuous phenotypes. Simulation and real data analysis have been conducted to compare the statistical power of our methods with existing methods. The results have shown strong evidence favoring the proposed iGSEA methods.

Statistical Learning and Likelihood Methods as Educational Tools

Timothy OBrien
Loyola University Chicago
tobrie1@luc.edu

Beginning students with basic backgrounds in applied statistics may find the transition into courses such as regression analysis, simulation, and categorical data analysis challenging. Through a series of carefully chosen illustrations based on likelihood methods and statistical modelling, this talk demonstrates how students may best bridge this gap. Examples are also taken from courses in intermediate methods and statistical/machine learning. Computational methods will also be discussed, illustrated, and emphasized.

A multi-resolution functional ANOVA model for emulation of large-scale, high-dimensional simulations

Chih-Li Sung, Wenjia Wang and ♦*Benjamin Haaland*
Georgia Tech, ISyE
bhaaland3@gatech.edu

Simulation has become an increasingly widespread technique for studying real systems in science and engineering for which physical experimentation is difficult or expensive. As computational capabilities have continued to grow simulations have become both large-scale, in terms of number of runs, and high-dimensional, in terms of input size. Here, we propose a multi-resolution functional ANOVA model for emulating large-scale and high-dimensional simulations. The model is particularly well-suited to high-dimensional data with lower-dimensional structure. Techniques for tuning parameter selection and confidence interval construction are proposed.

Big Data: How Do We Stay Relevant?

David Morganstein
Westat, Inc.
DavidMorganstein@westat.com

Big Data has become a common term in our vocabulary. Yet, there is no single clear definition of nor common agreement on what it means. In spite of this lack of clarity, which may never be resolved, it is having an important impact on our profession. This talk will describe a perspective we need to be effective as statisticians in this new environment and to increase the chances that our unique contributions are valued and sought. It will give a current description of the membership of the American Statistical Association and how it is changing and review steps that the ASA is taking to assure the statisticians role at the Big Data table.

Session 81: Missing Data and Multiple Imputation

Quantitative assessment of exposure to fecal contamination for young children in Accra, Ghana

♦*Yuke Wang*[1]*, Peter Teunis*[2]*, Christine Moe*[1]*, Clair Null*[3]*, Suraja Raj*[1]*, Kelly Baker*[4] *and Habib Yakubu*[1]
[1]Emory University
[2]RIVM, Netherlands
[3]Mathematica Policy Research
[4]The University of Iowa
yuke.wang@emory.edu

Diarrheal diseases are a leading cause of death for children under five globally. In the developing world, lack of adequate sanitation results in fecal contamination of the environment and poses a risk of enteric disease transmission via multiple exposure pathways. To better understand how different sources and transmission routes contribute to overall exposure to fecal contamination, we identified eight different fecal exposure pathways for children under five years old in four high-density, low-income neighborhoods in Accra, Ghana, and quantified the contribution of each pathway to oral intake of fecal contamination. Data collection for the SaniPath study was from 2011 to 2012, and comprised 500 hours of structured observations for behaviors of 156 children, questionnaires from 800 households, and 1855 environmental samples for microbiological testing. Data were analyzed using Bayesian models, estimating the environmental and behavioral factors associated with exposure to fecal contamination. These estimates were applied in exposure models simulating sequences of behaviors and transfers of fecal indicators from the environment to oral ingestion. This approach allows us to identify the contribution of any sources of fecal contamination in the environment to child exposure and use dynamic fecal microbe transfer networks to track fecal bacteria from the environment to oral ingestion. Exposure pathways were categorized into four types (high/low by dose and frequency), as a basis for prioritizing pathways by their potential to reduce fecal exposure. Although we observed variation in exposure (magnitude ranged from 10^8 to 10^16 CFU/day for E. coli) between different age groups and neighborhoods, the greatest contribution consistently was through the food pathway (contributing ¿99.9% to total exposure) in Accra, Ghana. Hands played a pivotal role in fecal microbe transfer from the environment to ingestion. The fecal microbe transfer network provides a systematic approach to study the complex interaction of poor sanitation infrastructure and human behavior on exposure to fecal contamination.

Comparison of two approaches for imputing a composite categorical variable

♦*Yi Pan, Ruiguang Song, Yulei He, Qian An and Guoshen Wang*
Centers for Disease Control and Prevention
jnu5@cdc.gov

Missing data is a common problem in many data systems. Variables with missing values are often binary or categorical and the missing pattern can be arbitrary. For example, there is an increasing trend in missing transmission category among HIV cases reported to CDC through the National HIV Prevention Program Monitoring and Evaluation (NHM&E) system in the United States. Transmission category (categorical) summarizes the multiple risk factors (binary) that an individual may have had by hierarchically selecting the one through which HIV was most likely transmitted. Accurate estimation of the distribution of transmission category among persons with diagnosed HIV infection is critical for the adequate allocation of HIV prevention and care resources. Multiple imputation is a popular approach to analyses with missing data. There are two ways to impute transmission category. One is directly imputing missing values of the transmission category variable, and the other is imputing missing values of binary risk factors first and then computing transmission category based on the imputed risk factors. In this study, the performance of the two imputation methods is examined through simulation and application to NHM&E data.

Multiple Imputation of Missing Linkage to Care Data for CDC-

funded HIV Testing Program Data in 2015

♦*Guoshen Wang, Yi Pan, Puja Seth, Ruiguang Song and Lisa Belcher*
Centers for Disease Control and Prevention
IFD1@CDC.GOV

HIV testing and linkage of HIV-positive persons to HIV medical care are crucial steps in the HIV continuum of care. However, despite improvements in data quality, monitoring and evaluating HIV testing and prevention programs continue to be challenging because of missing data. For example, the percentage of missing data for the linkage to HIV medical care (LHMC) indicator (i.e., attendance at first medical appointment for HIV-positive persons) among newly diagnosed HIV-positive persons was 46% in 2011. While it decreased in 2012 (23.7%), it has remained stable at 23.8% in 2013 and 24.2% in 2014. However, significant improvements have been seen with the percentage of missing data for linkage to HIV medical care within 90 days – 56.2% in 2012, 43.2% in 2013, and 29.6% in 2014. Methods to address missing values are critical for data analysis and interpretation. Previously, a complete case analysis was used to address missing data. That is, only observations without any missing values were used to calculate the final indicator estimate. However, multiple imputation is also a popular approach to address this issue, whereby the variables that contribute to an indicator calculation are first imputed and then the indicator is calculated. We compared the results from the two methods using 2014 CDC-funded HIV testing program data. Before imputation, linkage to HIV medical care among newly diagnosed HIV-positive persons within 90 days in complete case analysis was 81.8%. After imputation, linkage to HIV medical care among newly diagnosed HIV-positive persons within 90 days was 83.6%, 95% CI (83.0%, 84.3%). We will compare the results from two methods using 2015 CDC-funded HIV testing data. Multiple imputation can give a more accurate point estimation for the percentage of persons linked within 90 days and is recommended for addressing missing linkage data in further analyses, as indicated.

Evaluation of imputation methods for Tdap vaccination among pregnant women, Internet panel survey

♦*Helen Ding*[1]*, Carla Black*[2]*, Srivastav Anup*[3] *and Greby Stacie*[2]
[1]CFD Research Corporation
[2]CDC
[3]LEIDO
hding@cdc.gov

Background

Accurately measuring tetanus toxoid, reduced diphtheria toxoid, and acellular pertussis (Tdap) vaccination coverage among pregnant women is important for immunization programs to ensure pregnant women and infants are protected from pertussis. Tdap vaccination coverage is estimated using self-reported Tdap vaccination status obtained from surveys. However, survey respondents may not know or report their Tdap vaccination status. This study evaluates methods for managing missing values for Tdap vaccination.

Methods

Data from an Internet panel survey of pregnant women conducted in April 2015 were analyzed. The study population was limited to the 686 women who had a live birth during August 2014-April 2015. Women who did not know whether they had ever received a Tdap vaccination or did not know if they received the vaccination during their most recent pregnancy were considered to have missing data for Tdap vaccination status. Patterns of missing data were evaluated overall and by demographic characteristics. Tdap vaccination coverage was calculated and compared using three methods for handling missing values: Method 1: exclude those women who had missing Tdap vaccination status from the analysis; Method 2: Impute missing Tdap vaccination status using a hot-deck approach, i.e., by randomly assigning a single value for the missing observation with a value from a donor pool matched by age, race/ethnicity, geographic region, and provider recommendation and offer of vaccine; and Method 3: Multiple imputation, which replaces each missing value with a set of plausible imputed values (10) that represent the uncertainty about the true value and then combines the 10 imputed values. Because the internet panel survey is not a probability based sampling survey, sampling error cannot be accurately approximated; therefore, 5 percentage points were used to indicate a notable difference between two coverage estimates.

Results

Overall, 15.5% ($n = 106$) of the selected sample had a missing value for Tdap vaccination status. Those who were below poverty and those who reported receiving no recommendation for Tdap from their provider had a higher proportion of missing data compared with their counterparts (21.8% of women below poverty vs. 13.2% of women at or above poverty; 32.5%, 15.0%, and 6.9% for those with no provider recommendation for vaccination, those with a recommendation but no offer, and those with an offer, respectively).

Overall Tdap vaccination coverage using method1 was 42.1%, approximately 3 percentage points higher than that using method 2 and method 3 (39.3% and 39.2 % respectively). Vaccination coverage estimates varied depending on the matching variables chosen for donors using the hot-deck imputation approach, or variables chosen for the multivariable model using multiple imputation approach (e.g.: 42.9% and 42.5% using age, race/ethnicity, and region for method 2 and method 3, respectively, and 39.3% and 39.2% using age, race/ethnicity, region, and provider recommendation and offer for method 2 and method 3, respectively).

Using method 1, vaccination coverage estimates were more than 5 percentage points higher for women aged 35-49 years and among those with 1-5 provider visits compared with method 2 and for women with a high school education or less, greater than college education, and from the northeast region compared with method 3. Using method 3, vaccination coverage estimates were more than 5 percentage points lower for women of non-Hispanic other race/ethnicity and more than 5 percentage points higher for women with 1-5 provider visits compared with method 2. By provider recommendation and offer, vaccination coverage was similar using methods 1, 2, and 3 (61.0% [received an offer], 19.7% [received a recommendation but no offer], and 2.3% [received no recommendation] for method 1, 61.0%, 20.9%, and 2.6%, respectively, for method 2, and 61.5%, 20.3% and 2.7%, respectively for method 3). For other subgroups, vaccination coverage using method 1 differed by less than 5 percentage points compared with estimates using method 2 and method 3.

Conclusions

Missing Tdap vaccination status was associated with certain respondent characteristics. Vaccination coverage might be biased upward by excluding respondents with missing Tdap status (method 1). Estimates based on the multiple imputation approach (method 3) were similar overall and within most subgroups compared with the hot deck imputation approach (method 2) if the variables included in the multivariable model in method 3 and the variables selected for donor matching in method 2 were the same.

Session 82: Statistical Innovations in the Analysis of Metagenomic Data

New development in alignment-free genome and metagenome comparison

Fengzhu Sun
University of Southern California
fsun@usc.edu

Next generation sequencing (NGS) technologies have generated enormous amount of shotgun read data and assembly of the reads is challenging, especially for organisms without reference sequences and metagenomes. We develop novel alignment-free and assembly-free statistics for genome and metagenome comparison. The key idea is to remove the background word counts from the observed counts when comparing genomes and metagenomes. Markov chains (MC) are usually used to model background molecular sequences and we develop a new statistical method to estimate the order of MCs based on short read data. The alignment-free sequence comparison statistics are used to study the relationships among species, to assign virus to their hosts, and to classify metagenomes and metatranscriptomes. In all applications, our novel methods yield results that are consistent with biological knowledge. Thus, our statistics provide powerful alternative approaches for genome and metagenome comparison based on NGS short reads.

Informative Approach in Differential Analysis on Time Course Microbial Studies

♦*Lingling An and Dan Luo*
University of Arizona
anling@email.arizona.edu

Recent and rapid advent of high-throughput sequencing technologies has greatly promoted the field of metagenomics, which studies the genetic materials of entire microbial communities. Detecting differentially abundant features (e.g., species or genes) plays a critical role in revealing the contributors (i.e., pathogens) to the status (e.g., disease) of microbial samples. However, currently available statistical methods lack power in detecting differentially abundant features across different conditions, in particular, for time series metagenomic data.

We have proposed a novel procedure to meet with the challenges in detecting differentially abundant features from metagenomic samples under different biological/medical conditions. The new approach takes advantage of dependence structure of time series data and the detection procedure relies on sound statistical support. Not only it can accurately identify the different features but also result in the information on the start and end time points. Compared with other existing methods the new approach shows the best performance in the comprehensive simulation studies. The new method is also applied to real metagenomic datasets and the new interesting findings may provide another angle of understanding the mechanism of the diseases.

Regression modeling of microbiome data integrating the phylogenetic tree

♦*Jun Chen*[1] *and Jian Xiao*[2]
[1]Mayo Clinic
[2]Mayo linic
chen.jun2@mayo.edu

The human microbiome, which has now been regarded as the "extended" human genome, has attracted considerable attention from both biomedical scientists and clinical investigators. Numerous studies have revealed a significant role of the human microbiome in disease development and prognosis. The human microbiome holds potential for precision medicine to improve patient care. Integrating the human microbiome data into medicine requires developing powerful predictive models while taking into the special characteristics of microbiome data. One popular type of microbiome data is generated by sequencing a region of the bacterial 16S rRNA gene. The output of the 16S targeted sequencing is an abundance table of the detected bacterial species, along with their phylogenetic relationship. Utilizing the phylogeny information in microbiome-based prediction is critical since the microbial taxa tend to be associated with the phenotype at various phylogenetic resolutions. However, efficient use of the phylogeny raises some challenges. Here I will present a framework for integrating the phylogenetic tree into predictive modeling of the microbiome data. I will introduce a new type of sparse regression model, inverse-correlation matrix regularized sparse regression (ICM-SR), where the correlation matrix is defined based on the phylogeny. Simulation as well as real 16S data will be used to illustrate the proposed method.

Regression analysis with compositional covariates

Hongmei Jiang
Northwestern University
hongmei@northwestern.edu

The abundance of an organism or a taxon is usually measured using relative proportion or percentage in sequencing-based metagenomics studies. Due to the constraint of the sum of the relative abundances being 1 or 100%, standard conventional statistical methods may not be suitable for the compositional data analysis. Various data transformations have been proposed so that standard statistical procedures can be applied. In this talk we will discuss and compare different approaches for regression analysis with compositional covariates. Current statistical and computational methods that are being developed to analyze the metagnoimcs data and the challenges will also be highlighted.

Session 83: Statistical Methods for Network Analysis

Co-clustering of nonsmooth graphons

David Choi
Carnegie Mellon University
davidch@andrew.cmu.edu

Theoretical results are becoming known for community detection and clustering of networks; however, these results assume an idealized generative model that is unlikely to hold in many settings. Here we consider exploratory co-clustering of a bipartite network, where the rows and columns of the adjacency matrix are assumed to be samples from an arbitrary population. This is equivalent to assuming that the data is generated from a nonparametric model known as a graphon. We show that co-clusters found by any method can be extended to the row and column populations, or equivalently that the estimated blockmodel approximates a blocked version of the generative graphon, with generalization error bounded by $n^{-1/2}$. Analogous results are also shown for degree-corrected co-blockmodels and random dot product bipartite graphs, with error rates depending on the dimensionality of the latent variable space.

Network Reconstruction From High Dimensional Ordinary Differential Equations

♦*Shizhe Chen, Ali Shojaie and Daniela Witten*
University of Washington
szchen@uw.edu

We consider the task of learning a dynamical system from high-dimensional time-course data. For instance, we might wish to es-

timate a gene regulatory network from gene expression data measured at discrete time points. We model the dynamical system nonparametrically as a system of additive ordinary differential equations. Most existing methods for parameter estimation in ordinary differential equations estimate the derivatives from noisy observations. This has been shown to be challenging and inefficient. We propose a novel approach that does not involve derivative estimation. We show that the proposed method can consistently recover the true network structure even in high dimensions, and we demonstrate empirical improvement over competing approaches.

Measuring Influence in Twitter Ecosystems Using a Counting Process Modeling Framework

Shawn Mankad
Cornell University
spm263@cornell.edu

Data extracted from social media platforms are both large in scale and complex in nature, since they contain both unstructured text, as well as structured data, such as time stamps and interactions between users. A key question for such platforms is to determine influential users, in the sense that they generate interactions between members of the platform. Common measures used both in the academic literature and by companies that provide analytics services are variants of the popular web-search PageRank algorithm applied to networks that capture connections between users. In this work, we develop a modeling framework using multivariate interacting counting processes to capture the detailed actions that users undertake on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Based on the proposed model, we also derive a novel influence measure. We discuss estimation of the model parameters through maximum likelihood and establish their asymptotic properties. The proposed model and the accompanying influence measure are illustrated on a data set covering a five year period of the Twitter actions of the members of the US Senate, as well as mainstream news organizations and media personalities.

Network Inference from Grouped Observations Using Star Models

♦*Charles Weko*[1] *and Yunpeng Zhao*[2]
[1]Department of Defense
[2]George Mason University
cweko@masonlive.gmu.edu

In medical research, economics, and the social sciences data frequently appears as subsets of a set of objects. Over the past century a number of descriptive statistics have been developed to construct network structure from such data. However, these measures lack a generating mechanism that links the inferred network structure to the observed groups. To address this issue, we propose a model-based approach called the Star Model which assumes that every observed group has a leader and that the leader has brought together the other members of the group. The performance of Star Models is demonstrated by simulation studies. We apply this model to infer the relationships among Senators serving in the 110th United States Congress, the characters in a famous 18th century Chinese novel, and the distribution of flora in North America.

Session 84: Robust EM Methods and Algorithms

Robust rank-based EM algorithm and trimmed BIC criterion

Xin Dang
University of Mississippi
xdang@olemiss.edu

The paper presents a new robust EM algorithm for the finite mixture learning procedures. The proposed Spatial-EM algorithm utilizes median-based location and rank-based scatter estimators to replace sample mean and sample covariance matrix in each M step, hence enhancing stability and robustness of the algorithm. It is robust to outliers and initial values. Compared with many robust mixture learning methods, the Spatial-EM has the advantages of simplicity in implementation and statistical efficiency. We apply Spatial-EM to supervised and unsupervised learning scenarios. More specifically, robust clustering and outlier detection methods based on Spatial-EM have been proposed. For clustering analysis, BIC and trimmed BIC can be used to select the number of components. Compared with the regular EM and many other existing methods such as K-median, X-EM and SVM, the rank based method demonstrates superior performance and high robustness.

Robust Expectation Maximization Algorithm for Mixture Models

♦*Yichen Qin*[1] *and Carey Priebe*[2]
[1]University of Cincinnati
[2]Johns Hopkins University
qinyn@ucmail.uc.edu

We introduce a robust estimation procedure for mixture models using a new expectation-maximization (EM) algorithm. Compared with the traditional EM algorithm, the proposed method provides a more robust estimation against outliers for small sample sizes. In particular, we study the performance of the proposed method in the context of the gross error model, where the true model of interest is a mixture of two normal distributions, and the contamination component is a third normal distribution with a large variance. A numerical comparison between the proposed method and the traditional method for this gross error model is presented in terms of Kullback Leibler (KL) distance and relative efficiency.

Robust estimation of clusters & mixture models based on trimming and constraints

Luis Angel García-Escudero[1]*, Alfonso Gordaliza*[1]*, Francesca Greselin*[2] *and* ♦*Agustin Mayo-Iscar*[1]
[1]Universidad de Valladolid
[2]University of Milano Bicocca
agustin@med.uva.es

Trimming procedures are commonly used in many statistical settings for getting robust estimators in the presence of contamination. When applying EM-algorithm, for achieve a robust proposal, both in model-based clustering and mixture modeling, trimming is not enough. It is also necessary to control the relative size of the components' scatter by applying some kind of constraints. The talk presents robust methodology based on the joint application of trimming and constraints for different clusters and mixture models settings. Proposals based in this methodology will be applied for estimating mixtures of regressions, mixtures of factor analyzers and mixtures of skew-symmetric models. The joint application of these tools produce additional input parameters. Data driven tools for helping to the user in choosing these additional input parameters will be presented.

Robust mixture regression by EM algorithm

Chun Yu[1]*,* ♦*Weixin Yao*[2] *and Kun Chen*[3]
[1]Jiangxi University of Finance and Economics
[2]University of California, Riverside
[3]University of Connecticut
weixin.yao@ucr.edu

Finite mixture regression models have been widely used for modelling mixed regression relationships arising from a clustered and thus heterogenous population. The classical normal mixture model, despite of its simplicity and wide applicability, may fail dramatically in the presence of severe outliers. We propose a robust mixture regression approach based on a sparse, case-specific, and scale-dependent mean-shift parameterization, for simultaneously conducting outlier detection and robust parameter estimation. A penalized likelihood approach is adopted to induce sparsity among the mean-shift parameters so that the outliers are distinguished from the good observations, and a thresholding-embedded Expectation-Maximization (EM) algorithm is developed to enable stable and efficient computation. The proposed penalized estimation approach is shown to have strong connections with other robust methods including the trimmed likelihood and the M-estimation methods. Comparing with several existing methods, the proposed methods show outstanding performance in numerical studies.

Session 85: Student Award Session

Incorporating Biological Information in Sparse PCA with Application to Genomic Data

♦*Ziyi Li, Sandra Safo and Qi Long*
Emory University
ziyi.li@emory.edu

Motivation: The advances in technology have lead to the collection of high dimensional data such as genomic data. Before applying the existing statistical methods on high dimensional data, principal component analysis (PCA) is often used to reduce dimensionality. Sparse PC loadings are usually desired in this situation for simplicity and better interpretation. Although PCA has been extended to produce sparse PC loadings, few methods take potential biological information into consideration. Method: In this article, we propose two novel structured sparse PCA methods which not only have sparse solutions but also incorporate available biological information. The proposed methods utilize group information of variables as well as the connections between variables within each group when calculating sparse PCA loadings.
Results: Our simulation study demonstrates incorporating known biological information improves the performance of sparse PCA methods, and the proposed methods are robust to potential misspecification of the biological information. We further illustrate the performance of our methods in a Glioblastoma genomic data set.

Semiparametric Estimation of the Accelerated Failure Time Model with Partly Interval-censored Data

♦*Fei Gao, Donglin Zeng and Danyu Lin*
University of North Carolina at Chapel Hill
fgao@live.unc.edu

Partly interval-censored (PIC) data arise when some failure times are exactly ob- served while others are only known to lie within certain intervals. In this paper, we consider efficient semiparametric estimation of the accelerated failure time (AFT) model with PIC data. We first generalize the Buckley-James estimator for right- censored data to PIC data. Then, we develop a one-step estimator by deriving and estimating the efficient score for the regression parameters. We show that, under mild regularity conditions, the generalized Buckley-James estimator is consistent and asymptotically normal, and the one-step estimator achieves the semiparametric efficiency bound. We conduct extensive simulation studies to examine the performance of the proposed estimators in finite samples and apply our methods to data derived from an AIDS study.

A Latent Class Modeling Approach for Predicting Kidney Obstruction in the Absence of a Gold Standard

♦*Lijia Wang*[1]*, Qi Long*[1]*, Andrew Taylor*[2] *and Amita Manatunga*[1]
[1]Biostatistics and Bioinformatics, Emory University
[2]Radiology and Imaging Sciences, Emory University
lwang87@emory.edu

Currently there is no gold standard for detection of kidney obstruction. A recent study investigated how to evaluate kidney obstruction and collected expert ratings and data from diuresis renography, a valuable clinical procedure for evaluating suspected kidney obstruction. Motivated by this study, we develop a latent class modeling approach for predicting kidney obstruction through integrative analysis of time-series renogram data and expert ratings. Conditioning on the latent disease status, a random effects model is used for time-series renogram data; a probit model is used for expert ratings, allowing for unstructured correlation among multiple experts. A Bayesian procedure is developed for obtaining parameters estimates and subsequently predicting kidney obstruction. Applying to the motivating study, our approach is compared with the approach of majority voting based on expert ratings only which has been used as a gold standard in practice. The patterns of the estimated renogram curves for the predicted "obstructed" and "normal" kidneys are consistent with the clinical interpretations, lending support to the usefulness of our method. Simulation studies are conducted to demonstrate superiority of the proposed method over majority voting and evaluate different prediction schemes.

Individualizing Drug Dosage with Longitudinal Data

♦*Xiaolu Zhu and Annie Qu*
University of Illinois, Urbana-Champaign
xzhu28@illinois.edu

We propose a two-step procedure to personalize drug dosage over time under the framework of a log-linear mixed-effect model. We model patients' heterogeneity using subject-specific random effects which are treated as the realizations of an unspecified stochastic process. We extend the conditional quadratic inference function to estimate both fixed-effect coefficients and individual random effects on a longitudinal training data sample in the first step, and propose an adaptive procedure to estimate new patients' random effects and provide dosage recommendations for new patients in the second step. An advantage of our approach is that we do not impose any distribution assumption on estimating random effects. Moreover, the new approach can accommodate more general time-varying covariates corresponding to random effects. We show in theory and numerical studies that the proposed method is more efficient compared to existing approaches, especially when covariates are time-varying. In addition, a real data example of a clozapine study confirms that our two-step procedure leads to more accurate drug dosage recommendations.

Session 86: Bayesian Approaches in Drug Development: How Many Things Can We Accomplish?

Bayesian approach in Proof-of-Concept in drug development: a case study

Michael Lee
Janssen
mlee60@its.jnj.com

Proof-of-Concept (PoC) is a key step in drug development. Conventionally, PoC is accomplished by showing efficacy in one dose of the drug being studied. Because of the primary objective of PoC studies, efficacy in other dose level is seldom studied. There are circumstances where knowledge of other dose level can help designing subsequent studies. For example, if we know a dose level that does not meet target effect size, this dose level can be the lower bound of the subsequent dose-finding study. An adaptive design can offer the possibility of studying more than 1 dose. In this presentation we will share an adaptive PoC study design that potentially allows studying more than one dose. A challenge in this design is that treatment period is long compared to enrollment period. To allow adaptation before all subjects are randomized, a prediction of outcome at the end of treatment based on early signal becomes necessary. The prediction is made by a statistical model and Bayesian approach so that data from historical studies as well as data accumulated from the current study thus far can be utilized. Characteristics of this study design will be presented.

A Bayesian approach to evaluate program success for programs with multiple trials

♦*Meihua Wang and Guanghan(Frank) Liu*

Merck & Co.

meihua.wang@merck.com

A late stage clinical development program typically contains multiple trials. Nowadays, interim analyses are often used to allow evaluation for early success and/or futility for each individual study. So it presents a good opportunity for us to estimate the probability of program success (POPS) for the entire clinical development earlier. The sponsor may abandon the program early if the estimated POPS is very low, and therefore permit the resource savings and reallocation to other products. A Bayesian approach is provided to calculate POPS for a clinical program with multiple trials in binary outcomes, as well as its confidence intervals. Information from prior early stage trials may be available to borrow. However, constructing the informative prior for borrowing is challenging. Simulations are conducted to evaluate the effect of prior information on POPS evaluation. The methods are illustrated with historical data retrospectively from a completed clinical program for depression, as well as several ways of borrowing including pooling, test-then-pool and power prior approaches.

Bayesian Benefit-Risk Assessments for Medical Products

Telba Irony

Center for Biologics Evaluation and Research - FDA

telba.irony@fda.hhs.gov

Benefit-risk assessments for medical products involve all available information about the benefits and risks of a treatment that is often obtained through clinical trials. A comprehensive benefit-risk determination should also take into account other factors reflecting the context in which these assessments are made. These factors include, among others, the uncertainty about the benefits and risks, the availability of alternative treatments, and the option of taking preliminary action and deferring a final benefit-risk determination to a later time, once more information is acquired. While the main challenge of benefit-risk determinations for medical product approval is to combine information with values, the essence of the Bayesian approach is to collect and update information, and merge it with values to make rational decisions. In this talk I will describe how the Bayesian approach can provide a framework for benefit-risk assessments of medical products.

Session 87: The Keys to Career Success as a Biostatistician

The Keys to Career Success as a Biostatistician

Lisa Lupinacci

Merck & Co., Inc.

lisa_lupinacci@merck.com

What is the biostatistician's role in the pharmaceutical industry? What differentiates an outstanding biostatistician from an average one? What skills do you need to acquire to get ahead in your career? How can you ensure that your efforts are recognized and appreciated? A successful career as a biostatistician in the pharmaceutical industry goes beyond "numbers". In addition to the strong technique skills, other skills and qualities such as effective communication, the "right" attitude and proactive mindsets are also the keys to prospering in this career! In this session, you will hear the inspiring stories, perspectives and guidance from leaders in prominent pharmaceutical companies and career consultants. You can ask questions and get personal advice and guidance for your career, and learn lessons about avoiding unnecessary diversions to your career path.

Making Sense of Data: Challenges of Being a Non-Clinical Biostatistician

♦*Binbing Yu and Harry Yang*

MedImmune

yub@medimmune.com

In addition to the majority of clinical statisticians in pharmaceutical industry, there is a group of elite data scientists, called nonclinical statisticians. The nonclinical statisticians provide support to the whole spectrum of pharmaceutical development, ranging from early-phase drug discovery to late-phase commercialization. Their work encompasses virtually all areas that is not involved in a clinic study. This implies that nonclinical statisticians are often faced with a wide variety of challenges.

The first challenge is to understand the study goal and data of the non-clinical research. For example, development and optimization of an assay requires both bioanalytical knowledge and quantitative skills. The second challenge is to use the statistical tools and models to address the issues wisely and appropriately. This often requires the usage of multiple data sources and state-of-the-art statistical methods. Especially for early-phase pharmaceutical development, sophisticated statistical methods do not necessary mean better. The third challenge is to convey statistical results in scientific or laymen's terms.

Overcoming these challenges will foster a collaborative relationship between statisticians and scientists. A successful non-clinical statistician must use scientifically sound methods, but a method is not useful until it is "suitable for its intended purpose."

You Are Statistically Significant–A career path for biostatisticians to make a difference

Helena Fan

The Lotus Group LLS

helena.fan@tlgcareers.com

This presentation provides useful information and insights gathered from the presenter's 12 years of experience in recruiting biostatisticians and from interviews with top statisticians in the pharmaceutical industry. Beginning with an introduction of the biostatistician's role and career path in the pharmaceutical industry, the discussion will focus on the traits of outstanding biostatisticians, strategies and tactics for gaining recognition in the workplace and suggestions for

the development of communication and leadership skills which are essential for biostatisticians to make a difference.

Session 88: Design of Experiments II

Partial Aliasing Relations in Mixed Two- and Three-Level Designs

Arman Sabbaghi
Purdue University Department of Statistics
sabbaghi@purdue.edu

The orthogonal polynomial parameterization of contrasts for mixed two- and three-level fractional factorial designs is important in practice because it yields partially aliased and interpretable interaction contrasts. However, its mathematics is not yet transparent, and this inhibits a simple understanding of its partial aliasing properties. A better understanding is achieved with indicator functions, and we develop the theory of indicator functions under this system for mixed two- and three-level designs. We prove that the algebra behind the calculation of indicator function coefficients is a product of individual algebraic operations for the different types of factors. This equivalence is then used to establish conditions for estimable interactions in mixed two- and three-level designs.

Designing covering arrays for testing statistical software

♦*Ryan Lekivetz and Joseph Morgan*
JMP Division of SAS
Ryan.Lekivetz@jmp.com

Testing software can be a time-consuming task. In validating statistical software, not only are there limitations on the number of tests that can be written, but there may also be limits on the number of tests that can be run for a certain time period. Covering arrays provide an efficient method to design test suites for testing software. In this talk, we introduce the concept of covering arrays and why they are useful for testing.

On Orthogonality through the block factor

Sunanda Bagchi
Indian Statistical Institute, Bangalore
sunanda_b@isibang.ac.in

The concept of main effect plans "orthogonal through the block factor" (POTB) has been introduced in Bagchi(2010). The main advantages of a POTB are that (a) it may exist in a set up where an "usual" orthogonal main effect plan (OMEP) cannot exist and (b) the data analysis is nearly as simple as that for an OMEP.

In the present paper this idea is extended to the set up where two-factor interactions among factors may also be present. We define the concept of orthogonality between a pair of factorial effects (main effects or interactions) "through the block facto" in the context of a symmetrical experiment.

Then we construct plans for symmetrical experiments in which interactions between three or more factors are absent. These plans have the following features.

(a) Blocks are of small size.

(b) Number of levels is two or three.

(c) In the plans for two-level factors orthogonality through the block factor is attained between every pair of factorial effects.

(d) In the plans for three-level factors every effect is orthogonal to all except one or two of the others through the block factor.

(e) The total number of runs is not more than that of any known plan for the same experiment.

Using fractional factorials to obtain efficient designs for certain generalized linear models

John Stufken
Arizona State University
jstufken@asu.edu

Fractional factorials are the workhorse in various areas of application of design of experiments, especially in industrial settings. In the context of determining efficient designs for certain generalized linear models, we discuss a problem in which fractional factorials facilitate the selection of optimal designs with a relatively small number of support points. Some of these optimality results can be explained through theoretical considerations only, while others are partly based on computational evidence. Both theoretical and computational tools that support these results will be discussed.

Session 89: Statistical Phylogenetics

Bayesian Analysis of Evolutionary Divergence with Genomic Data Under Diverse Demographic Models

♦*Yujin Chung and Jody Hey*
Temple University
ychung.wisc@gmail.com

We present a new Bayesian method for estimating demographic and phylogenetic history of population genomic samples in an Isolation with Migration framework. By generating Markov chain Monte Carlo (MCMC) samples from a simplified and efficient importance sampling distribution of coalescent trees that include neither demography nor migration, the new method provides for calculation of the joint posterior density for all model parameters. Moreover, using Markov chain representation of genealogy, the method analytically takes account of all possible migrations. Once sampled, the coalescent trees can be applied repeatedly to the inference of multiple diverse demographic models, providing for model comparison without resampling of trees through MCMC simulations. MIST implements the new method in parallel computing and scales to large numbers of loci without introducing the MCMC mixing problems typically associated with genealogy samplers. We demonstrate the method using simulated data and DNA sequences of two common chimpanzee subspecies: Pan troglodytes (P. t.) troglodytes and P. t. verus.

Displayed trees do not determine distinguishability under the network multispecies coalescent

♦*James Degnan*[1] *and Sha Zhu*[2]
[1]University of Canterbury
[2]Oxford University
jamdeg@gmail.com

Recent work in estimating species relationships from gene trees has included inferring networks assuming that past hybridization has occurred between species. Probabilistic models using the multispecies coalescent can be used in this framework for likelihood-based inference of both network topologies and parameters, including branch lengths and hybridization parameters. A difficulty for such methods is that it is not always clear whether, or to what extent, networks are identifiable — i.e., whether there could be two distinct networks that lead to the same distribution of gene trees. We present a new representation of the species network likelihood that represents the probability distribution of the gene tree topolgies as a linear combination of gene tree distributions given a set of species trees. This representation makes it clear that in some cases in which two distinct networks give the same distribution of gene trees when sampling one allele per species, the two networks can

be distinguished theoretically when multiple individuals are sampled per species. This result means that network identifiability is not only a function of the trees displayed by the networks but also depends on allele sampling within species. We additionally give an example in which two networks that display exactly the same trees can be distinguished from their gene trees even when there is only one lineage sampled per species.

Likelihood Estimation of Large Species Trees Using the Coalescent Process

Arindam RoyChoudhury
Columbia University
ar2946@columbia.edu

A species tree is a weighted tree-graph that represents the order and the magnitude of separation between a given set of species. Statistical estimation of trees is an integral part of studying species trees. Although various likelihood and Bayesian estimators of species trees are available, none of these methods are fast enough to estimate very large species trees under a certain commonly used model (the coalescent). This problem is especially relevant today because there has been a recent influx of large amount of genomic data. Here I will present an approach of fast likelihood estimation of species trees, exploiting a certain special structure of the tree space. Using this approach, one will be able to estimate larger species trees than previously possible in a reasonable time.

Mechanistic Models for the Retention of Duplicate Genes in Phylogenetic Birth-Death Processes

David Liberles
Temple University
daliberles@temple.edu

The protein coding content of genomes is shaped by the ongoing process of gene duplication and loss in addition to other processes. Thus, gene duplication is a fundamental process that enables rapid molecular diversification of protein sequences resulting in lineage-specific changes to gene content. Together with a discussion of analyses of the recently sequenced Atlantic salmon genome, new mechanistic models for duplicate gene retention will be presented. Models have been developed to characterize the probabilities of retention under six processes: non-functionalization, and then non-functionalization plus neo-functionalization, sub-functionalization, dosage balance, and lastly, with dosage balance as a transition state to subsequent sub-functionalization and dosage balance to neo-functionalization. Further, an addition to the model has been described that accounts for the probability of duplicated genes that are observed to segregate in a sampled genome but are destined to fail to fix. This modeling suite will be presented in the context of current bioinformatics pipelines for comparative genomics.

Session 90: Adaptive Methods and Regulation for Clinical Trials

Group sequential trial design under parametric survival model

♦*Jianrong Wu* [1] *and Xiaoping Xiong*[2]
[1]St. Jude Children's Research Hospital
[2]St. Jude Children's Research Hospital
jianrong.wu@stjude.org

In this talk, a sequential test is proposed under a parametric survival model. The proposed test is asymptotically normal distributed with an independent increments structure. The sample size for fixed sample test is derived for the purpose of group sequential trial design. In addition, a multi-stage group sequential procedure is given by applying the Brownian motion property of the test statistic and sequential conditional probability ratio test methodology.

A Robust Bayesian Dose-Finding Design for Phase I/II Clinical Trials

♦*Suyu Liu*[1] *and Valen Johnson*[2]
[1]The UT MD Anderson Cancer Center
[2]Texas A&M University
syliu@mdanderson.org

We propose a Bayesian phase I/II dose-finding trial design that simultaneously accounts for toxicity and efficacy. We model the toxicity and efficacy of investigational doses using a flexible Bayesian dynamic model, which borrows information across doses without imposing stringent parametric assumptions on the shape of the dose-toxicity and -efficacy curves. An intuitive utility function that reflects the desirability trade-offs between efficacy and toxicity is used to guide the dose assignment and selection. We also discuss the extension of this design to handle delayed toxicity and efficacy. We conduct extensive simulation studies to examine the operating characteristics of the proposed method under various practical scenarios. The results show that the proposed design possesses good operating characteristics and is robust to the shape of the dose-toxicity and dose-efficacy curves.

Continual reassessment method with multiple toxicity constraints

♦*Bin Cheng and Shing Lee*
Columbia University
bc2159@cumc.columbia.edu

Conventional dose finding methods require the dichotomization of toxicity outcome measures generally collected in an ordinal scale. To improve efficiency and include more information on the gradation of toxicities, a sequential likelihood procedure that accounts for multiple toxicity constraints is proposed to differentiate the tolerance for toxicity of various degrees of severity under a novel class of multiplicative models, and the asymptotic properties of the procedure under certain model misspecification are established.

The challenges and opportunities of adaptive designs in medical device studies

♦*Jie (Jack) Zhou, Hong (Laura) Lu and Xiting (Cindy) Yang*
FDA Center for Devices and Radiological Health
jack.zhou@fda.hhs.gov

Medical devices present unique challenges and opportunities in clinical trials. Nonrandomized, unblinded, and even single arm studies are frequently seen in medical device studies, presenting challenges to statistical inference and clinical trial conduct, especially for studies with adaptive designs. At the same time, the short development cycles, the similarity between each generation of medical devices, and the smaller sample sizes proposed by the smaller, more innovative companies often make adaptive designs attractive in these situations. We will present an overview of frequently used group sequential, as well as Bayesian and Frequentist adaptive designs in medical device trials, using real-life case studies as examples.

Session 91: Recent Developments of Nonparametric Methods for High-Dimensional Data

Variable Selection for Semiparametric Geospatial Models

♦*Guannan Wang*[1] *and Lily Wang*[2]
[1]College of William & Mary
[2]Iowa State University

gwang01@wm.edu

In this paper, we focus on the variable selection techniques for a class of semiparametric geospatial models which allow one to study the effect of the covariates in the presence of the spatial information. The smoothing problem in the nonparametric part is tackled by means of bivariate splines over triangulation, which is able to deal efficiently with data distributed over irregularly shaped regions. In addition, we develop a unified procedure for variable selection to identify significant linear covariates under a double penalization framework, and we show that the penalized estimators enjoy the oracle property. The proposed method can simultaneously identify non-zero covariates in the linear part and solve the problem of "leakage" across complex domains of the nonparametric part. To estimate the standard deviations of the proposed estimators for the coefficients, a sandwich formula is developed as well. In the end, several simulation examples and a real data analysis are studied to illustrate the proposed methods.

Optimal Prediction for Functional Linear Regression with A Functional Response

Xiaoxiao Sun[1]*, Xiao Wang*[2]*, Ping Ma*[1] *and* ♦*Pang Du*[3]
[1]University of Georgia
[2]Purdue University
[3]Virginia Tech
pangdu@vt.edu

This talk focuses on functional linear regresson model with a functional response and a functional predictor. A penalized likelihood approach is proposed to estimate the unknown intercept and coefficient functions in the model. Inference tools such as point-wise confidence intervals of the coefficient function and prediction intervals are derived. The minimax rate of convergence for the error in predicting the mean response is established. It is shown that the penalized likelihood estimator attains the optimal rate of convergence. Our simulations demonstrate a competitive performance against the existing approach. The method is further illustrated in a genetic study on the regulartory effect of histone acetylation on gene expression, both are measurements over time.

Ultra-high Dimensional Additive Partial Linear Models

♦*Xinyi Li, Li Wang and Dan Nettleton*
Iowa State University
lixinyi@iastate.edu

Motivated by the maize Shoot Apical Meristem (SAM) study, we consider variable selection in a flexible semiparametric additive partial linear regression model for analyzing ultra high dimensional data. We approximate the nonlinear additive components using B-spline basis functions. We apply the adaptive group Lasso to select nonzero components in high-dimensional settings where the dimension of covariates can be much larger than the sample size. The proposed method selects the correct model with probability approaching 1 under some regularity conditions. The estimators in both the linear part and nonlinear part are consistent, and their rates of convergence are also established. The proposed method enables us to select linear and nonparametric components and capture nonlinear patterns of some covariates simultaneously. The performance of the method is evaluated by some simulation studies. The proposed method is also applied to the genetic study for the SAM data.

Weighing Schemes for Functional Data

♦*Xiaoke Zhang*[1] *and Jane-Ling Wang*[2]
[1]University of Delaware
[2]University of California, Davis
xkzhang@udel.edu

Nonparametric estimation of mean function is important in functional data analysis. We investigate the performance of a local linear smoother with a general weighing scheme, which includes two commonly used schemes, equal weight per observation (OBS), and equal weight per subject (SUBJ), as two special cases. We provide a comprehensive analysis of their asymptotic properties on a unified platform for all types of sampling plan, be it dense, sparse, or neither. The asymptotic theories are unified on two aspects: (1) the weighing scheme is very general; (2) the magnitude of the number Ni of measurements for the ith subject relative to the sample size n can vary freely. Based on the relative order of Ni to n, functional data are partitioned into three types: non-dense, dense, and ultra-dense functional data for the OBS and SUBJ schemes. These two weighing schemes are compared both theoretically and numerically. We also propose a new class of weighing schemes in terms of a mixture of the OBS and SUBJ weights, of which theoretical and numerical performances are examined and compared.

Session 92: The Analysis of Complex Time-to-Event Data

The Analysis of Spontaneous Abortion with Left Truncation, Partly Interval Censoring and Cure Rate

♦*Yuan Wu*[1] *and Ronghui Xu*[2]
[1]Duke University
[2]UCSD
yuan.wu@duke.edu

Infections during pregnancy will increase women's risk of serious consequences. People have started to study the cohorts with safety data for vaccination during pregnancy. However, new advanced statistical methods are much needed to address the complicated data features of such cohorts including cure rate, partly interval censoring and left truncation. We propose to use semi-parametric sieve estimation method to deal with this complicated data structure and we assume the data follows non-mixture cure model with Cox proportion hazard regression. Simulation and real data studies are performed. We also provided asymptotic results for the proposed estimation method.

Residual-Based Model Diagnosis Methods for Mixture Cure Models

♦*Yingwei Peng*[1] *and Jeremy Taylor*[2]
[1]Queen's University
[2]University of Michigan
pengp@queensu.ca

Model diagnosis, an important issue in statistical modeling, has not yet been addressed adequately for cure models. We focus on mixture cure models in this work and propose some residual-based methods to examine the fit of the mixture cure model, particularly the fit of the latency part of the mixture cure model. The new methods extend the classical residual-based methods to the mixture cure model. Numerical work shows that the proposed methods are capable of detecting lack-of-fit of a mixture cure model, particularly in the latency part, such as outliers, improper covariate functional form, or nonproportionality in hazards if the proportional hazards assumption is employed in the latency part. The methods are illustrated with two real datasets that were previously analyzed with mixture cure models.

A New Coefficient of Determination for Regression Models

Chun Li
Case Western Reserve University
cxl791@case.edu

While the coefficient of determination, R^2, is a well known measure of fit for linear regression, measures of fit for other outcome types often are less satisfying. For regression models, we propose a new coefficient of determination as a correlation between the observed values and the fitted distributions, taking into account the variation in the latter. In linear regression, the measure is identical to R^2, thus giving R^2 a new interpretation as a correlation. In logistic regression, it is Tjur's coefficient of discrimination. For time-to-event outcomes, it is robust to the amount of censoring. Asymptotically, the measure reflects the fraction of total outcome variation explained by the fitted model, and it can be approximated using a Pythagorean property. We present a general definition of the measure, and evaluate it for continuous (including nonlinear and linear models fitted with ordinary or weighted least squares), count, binary, ordinal, and time-to-event outcomes. We give a list of desirable properties for a measure of fit, and show that our measure effectively satisfies all of them. We also discuss the use of correlation as a measure of fit, and describe the connections between correlation, measure of fit, and explained variation.

The Analysis and Desgin of Cancer Clinical Trials Based on Progression Free Survival

♦*Leilei Zeng*[1]*, Yidan Shi*[1]*, Lan Wen*[2] *and Richard Cook* [1]
[1]University of Waterloo
[2]MRC Biostatistics Units
lzeng@uwaterloo.ca

Cancer clinical trials are routinely designed on the basis of event-free survival time where the event of interest may represent a complication, metastasis, relapse, or progression. Given that such event is typically only assessed periodically, the event-free survival times are thus subject to dual censoring schemes. In this talk, we highlight statistical issues arising due to such censoring schemes, including the biases in the estimation of survival and loss of power in detecting its association with markers, and propose methods for sample size calculation and justification to address the loss of power.

Session 93: Recent Developments of Graphical Model for Big Data

Estimation and Inference for Dynamic Network Models for High-Dimensional Time Course Data

Hulin Wu
University of Texas Health Science Center at Houston
Hulin.Wu@uth.tmc.edu

It is very challenging to establish high-dimensional dynamic network models for high-dimensional time course data due to the curse of dimensionality and high computational cost. We have developed a novel method for parameter estimation and model selection for high-dimensional differential equation (ODE) models that may have more than 1 million parameters. The proposed methods can be applied to any application fields with high-dimensional time course data, including bioinformatics and stock market data modeling. The established models have dual-properties, dynamic system and network feature, which allow us to investigate the mechanisms and principles behind the observed data.

Nonparametric mixture of Gaussian graphical models, with applications to ADHD imaging data

♦*Kevin Lee and Lingzhou Xue*
Pennsylvania State University
khl119@psu.edu

Graphical model has been widely used to investigate complex dependence of various high-dimensional biomedical data. It is common to assume that observed data follow a homogeneous graphical model. However, in many real-world applications, observations usually come from different resources and have heterogeneous hidden commonality. Thus, it is important to account for heterogeneous dependencies and discover hidden commonality across the whole population. In this work, we introduce a novel regularized estimation scheme for learning nonparametric mixture of Gaussian graphical models, which extends the methodology and applicability of Gaussian graphical models and mixture models. We propose a unified penalized likelihood approach to effectively estimate nonparametric functional parameters and heterogeneous graphical parameters. We further design an efficient generalized effective EM algorithm to address three significant challenges: high-dimensionality, non-convexity, and label switching. We demonstrate our method in simulation studies and a real application to estimate human brain functional connectivity from ADHD imaging data, where two heterogeneous conditional dependencies are explained through profiling demographic variables and supported by existing scientific findings.

Consistent Estimation of Curved Exponential-Family Random Graph Models with Local Dependence and Growing Neighborhoods

Michael Schweinberger
Rice University
m.s@rice.edu

Statistical inference for discrete exponential-family random graph models given a single observation of a large random graph is in general problematic. We show that statistical inference is sensible as long as the model is endowed with additional structure. We consider a simple and common form of additional structure: multilevel structure in the form of neighborhood structure. Multilevel structure is observed in multilevel networks and is generated by either multilevel sampling designs or stochastic processes governing random graphs. We derive non-asymptotic concentration and consistency results concerning maximum likelihood estimators of full and non-full, curved exponential-family random graph models with weak and strong dependence within and between neighborhoods, where the neighborhoods and the natural parameter vectors of neighborhoods may grow with the number of neighborhoods. These consistency results are the first consistency results concerning a wide range of full and non-full, curved exponential-family random graph models under correct and incorrect model specifications and demonstrate that statistical inference is meaningful as long as models are endowed with additional structure. We discuss a local approach to computing maximum likelihood estimators that can be applied to large random graphs. Simulation results and an application are presented.

Composite Likelihood Inference on Stochastic Block Model for Big Networks

Ningtao Wang
University of Texas School of Public Health
Ningtao.Wang@uth.tmc.edu

Network data appears in diverse areas, such as brain functional connectivity in neuroscience, human communication network in social science, gene co-expression network and Hi-C data in genomics, to name a few. A fundamental step of analyzing networks is to detect and model the community structure within the network, where the stochastic block model is the most commonly used benchmark for

such a task. The first problem of fitting stochastic block model is the computation of optimal label assignments of communities, which is, in principle, NP-hard. We proposed to use composite likelihoods as an approximation of the infeasible full likelihood. The proposed method guarantees the unbiased estimating equations, and converts the NP-hard problem into a linear-hard one. A two-layer EM algorithms has been developed to obtain the parameter estimates. We proved the consistency of the latent class assignment function when the number of observations tends to infinity. The model was validated through simulation studies and by new discoveries from analyzing Hi-C data.

Session 94: Recent Developments in Statistics

Simulating Longer Vectors of Correlated Binary Random Variables via Multinomial Sampling

Justine Shults
University of Pennsylvania
jshults@mail.med.upenn.edu

The ability to simulate correlated binary data is important for sample size calculation and comparison of methods for analysis of clustered and longitudinal data with dichotomous outcomes. Sampling from the multinomial distribution of all possible length n permutations of zeros and ones is a straightforward simulation approach that was first proposed by Kang and Jung (Biom. J. 2001; 43 (3): 1521-4036). However, the multinomial sampling method has only been implemented in general form (without first making restrictive assumptions) for vectors of length 2 and 3. As noted by Haynes, Sabo, and Chaganty (2015), the CDF for establishing decision rules becomes complicated for cases of four or more repeated measures. While not impossible, constructing higher order joint probabilities can be computationally challenging. In this presentation I present an algorithm for simulating correlated binary data via multinomial sampling that can be applied to directly compute the joint distribution for any n. I demonstrate my algorithm to simulate vectors of length 4 and 8 in an assessment of power during the planning phases of a study and to assess the choice of working correlation structure in an analysis with GEE.

Stochastic Analysis of the Cost-Effectiveness Frontier

♦*Daniel Heitjan*[1] *and Yu Lan*[2]
[1]SMU/UTSW
[2]SMU
dheitjan@gmail.com

Suppose you wish to conduct a cost-effectiveness analysis for a set of potential treatments, for which you have estimates of costs and effectiveness from some combination of clinical trials, observational studies, or empirically-informed simulation exercises. The first step in the analysis is to identify all those treatments that lie on the CE frontier; the next is to select the most effective treatment whose incremental CE ratio (compared to the next less effective one on the frontier) is less than the current willingness-to-pay. The literature on incorporating uncertainty in a comparison of two treatments, as from a clinical trial or observational study, is voluminous, but the topic of comparing several treatments when there is uncertainty has received little study. We address the latter problem from a Bayesian perspective, which renders both the synthesis and analysis of the data conceptually straightforward. We propose to present posterior probabilities of two kinds of events: i) specific configurations of CE frontiers, and ii) indicators of whether specific treatments are on the frontier. One can readily extract these probabilities from any Monte Carlo posterior computation. We illustrate our ideas using a simulation model that projects lifetime survival and healthcare costs in a population of smokers attempting to quit.

On the uncertainty of data extrapolation in pediatric drug development

Alan Chiang
Eli Lilly and Company
chiangay@lilly.com

The conduct of clinical trials in children poses several problems. Methodological issues and ethical concerns represent the major obstacles that have traditionally limited research in pediatric drug development. By extending information and conclusions available from studies in one or more subgroups of the patient population, or in related conditions or with related medical products, to make inferences for another subgroup of the population can minimize the need to generate additional information to reach conclusions for pediatric patients. We propose a structured framework to enable more informed decisions on when it is appropriate for data extrapolation to be made. By integrating evidence via data from adults and other source, inferences can be made on the drug effects for children. We will evaluate whether the use of different methods leads to the evidence to demonstrate a positive benefit risk being altered.

The Role of Kernels in Data Analysis: A Statistical Perspective

Marianthi Markatou
University at Buffalo
markatou@buffalo.edu

Many problems in data integration can be posed by a single goodness-of-fit question, namely inquiring whether two or more data sets come from the same distribution. The same question is pertinent in comparative effectiveness research (CER), where "balancing" the groups to be compared is required. Motivated by these challenges in the analysis of biomedical data we discuss the inferential potential of kernels and their statistical properties. We show that many classical goodness of fit tests are functions of specific kernels. Given data and a question of interest, the choice of kernel is a matter of design with a number of design factors influencing its selection. We first introduce the concept of root kernel and discuss the considerations that enter the selection of the kernel. We then derive an easy to use normal approximation to the power of kernel based tests and base the construction of a noncentrality index, an analogue of the traditional noncentrality parameter, on it. This leads to a method akin to the Neyman-Pearson lemma for constructing optimal kernels for specific alternatives. We introduce a "mid-power" analysis as a device for choosing optimal degrees of freedom for a family of alternatives of interest. Simulation results illustrate the performance of the proposed methods.
This is joint work with B. G. Lindsay and other co-authors.

Session 95: Recent Advances in Biomarker Evaluation and Risk Prediction

Efficient Epidemiological Study Designs for Quantitative Longitudinal Data

Jonathan Schildcrout
Vanderbilt University
jonathan.schildcrout@vanderbilt.edu

The analysis of longitudinal trajectories usually focuses on evaluation of explanatory factors that are either associated with rates of change, or with overall mean levels of a quantitative outcome variable. We will discuss valid design and analysis methods that per-

mit outcome dependent sampling of longitudinal data for scenarios where all outcome data currently exist, but a targeted sub-study is being planned in order to collect additional key exposure information on a limited number of subjects. We propose a stratified sampling based on specific summaries of individual longitudinal trajectories, and we compare several analytical approaches. We demonstrate that the efficiency of an outcome-based sampling design relative to use of standard simple random sampling depends highly on the choice of outcome summary statistic used to direct sampling, and we show a natural link between the goals of the longitudinal regression model and corresponding desirable designs.

Least Squares Regression Methods for Clustered ROC Data with Discrete Covariates

♦*Liansheng Tang*[1]*, Wei Zhang*[2]*, Qizhai Li*[2]*, Xuan Ye*[3] *and Leighton Chan*[4]

[1]NIH and George Mason University

[2]Chinese Academy of Sciences

[3]George Mason University and FDA

[4]National Institute of Health

ltang1@gmu.edu

The receiver operating characteristic (ROC) curve is a popular tool to evaluate and compare the accuracy of diagnostic tests to distinguish the diseased group from the non-diseased group when test results from tests are continuous or ordinal. A complicated data setting occurs when multiple tests are measured on abnormal and normal locations from the same subject and the measurements are clustered within the subject. Although least squares regression methods can be used for the estimation of ROC curve from correlated data, how to develop the least squares methods to estimate the ROC curve from the clustered data has not been studied. Also, the statistical properties of the least squares methods under the clustering setting are unknown. We develop the least squares ROC methods to allow the baseline and link functions to differ, and more importantly, to accommodate clustered data with discrete covariates. We apply the methods to a real example in the detection of glaucomatous deterioration. We also derive the asymptotic properties of the proposed methods.

Genetic-based Prediction of Vaccine-derived Poliovirus Circulation

♦*Kun Zhao, Jaume Jorba, Jane Iber, Qi Chen, Kelley Bullard, Olen Kew and Cara Burns*

Centers for Disease Control and Prevention

vzt5@cdc.gov

Background. A key objective of the WHO 2013–2018 Polio Eradication Endgame and Strategic Plan is the detection and interruption of all poliovirus (PV) circulation. PV surveillance plays a critical role for detecting PV transmission. The Global Polio Laboratory Network (GPLN) routinely detects PV in clinical specimens from AFP cases, supplemented by environmental surveillance from sewage samples. In April 2016, WHO implemented a worldwide synchronized switch from trivalent oral poliovirus vaccine (tOPV; types 1, 2, and 3) to bivalent OPV (bOPV; types 1 and 3). The goal of the switch is to prevent the emergence of type 2 vaccine-derived polioviruses (VDPV2s) since more than 90% of the VDPV in the past two years has been type 2. VDPVs may emerge during person-to-person transmission (circulating VDPV [cVDPV]) or during prolonged replication in individuals with primary immunodeficiency (iVDPV). When planning the scope of an outbreak response, it would be helpful to know if the recently identified VDPV is likely to be cVDPV or iVDPV. If the source of the infection is unknown, such as from environmental surveillance, the VDPV is designated ambiguous VDPV [aVDPV]. Here we describe genetic distinctions between cVDPV and iVDPV capsid sequences that may help inform global poliovirus post-switch outbreak response, particularly for VDPV2. Method. We performed retrospective analysis of cVDPV and iVDPV nucleotide sequences encompassing the VP1 capsid region (900 nucleotides) from GPLN's routine acute flaccid paralysis (AFP) and environmental surveillance. The dataset contains 1,071 cVDPV2 and 53 iVDPV2 sequences from 36 countries during 1970-2015. A logistic regression model was developed to estimate the probability of VDPV being cVDPV, given the proportion of nucleotide (NT) and amino acid (AA) substitutions observed. The number of genetic NT and AA substitutions was determined by comparing each sequence to its Sabin 2 parental strain. Results. cVDPV2 and iVDPV2 sequences show similar profiles for NT and AA frequencies, codon usage, and AA conservation. Counts of NT, AA, and codon substitutions in comparative analyses between cVDPV2 and iVDPV2 were insufficient for a clear distinction between the two VDPV2 categories. However, the combined measures of NT and AA substitutions (from VP1 and neutralizing antigenic site sequences) can be more informative, with more non-synonymous (AA changes) in iVDPV2 than cVDPV2. Therefore, NT and AA substitutions were used to predict the response variable "circulation of VDPV" within a logistics regression framework. The two-way classification of the actual response levels and the predicted response levels suggests that the model's sensitivity and specificity can achieve 91% and 53% respectively, at a 0.9 cutoff level. Decreasing the cutoff level will increase the model sensitivity at a cost of decreasing its specificity. 53% specificity is somewhat low. We are in a process of expending to complete capsid sequence. Conclusion. Prediction of cVDPV2 can be achieved by applying a logistic regression model using simple genetic measurements. Our next steps include working on complete capsid sequences (2,600 nucleotides) with the aim of improving the specificity of the model. These efforts highlighted the feasibility of early detection of cVDPV2 outbreaks by utilizing genetic information as a powerful surveillance tool.

Novel diagnostic accuracy analysis for competing risks outcomes with ROC surface

♦*Song Zhang and Yu Cheng*

University of Pittsburgh

soz1@pitt.edu

Many medical conditions are marked by a sequence of events or statuses that are associated with continuous changes in some biomarkers. However, few works have been done to assess the overall accuracy of a biomarker on separating various competing events. Existing methods usually focus on a single cause and compare it with the event-free controls each time. In our study, the concept of ROC surface and the volume under the ROC surface (VUS) has been extended to competing risks outcomes, given its capability to handle ordinal multi-category outcomes. We propose two methods to estimate the VUS. The first method is based on the correct classification probabilities (CCPs) for the subjects who have experienced different cause-specific events given a pair of threshold points from the distribution of a diagnostic marker. The second method is to measure concordance between the marker and competing outcomes. Since the samples are often subject to independent censoring, inverse probability of censoring weigh is introduced to handle censored outcomes. We also extend our framework to unordered competing risk settings. Asymptotic results are derived using counting process techniques. The practical performances of the proposed es-

timators have been evaluated through numerical studies.

Session 96: Methods of Integrating Novel Functional Data with Multimodal Measurements

Three-Part Joint Modeling Methods for Complex Functional Data

♦*Haocheng Li*[1]*, John Staudenmayer*[2]*, Tianying Wang*[3] *and Carroll Raymond*[3]
[1]University of Calgary
[2]University of Massachusetts
[3]Texas A&M University
haocheng.li@ucalgary.ca

We take a functional data approach to longitudinal studies with complex bivariate outcomes. One response is obtained in continuous proportions with excess zeros and ones. The other outcome is a continuous variable featured by excess zeros and skewness. A three-part functional data joint modeling approach is introduced. The first part is a continuation-ratio model to postulate the ordinal features for proportional response to be 0, (0,1) or 1. The second part is to model the proportions when they are in interval (0,1). The last component specifies the skewed continuous measurements by Box-Cox transformations when the proportions are at (0,1) or 1. In this three-part model, the regression structures are specified as smooth curves measured at various time-points with random effects that have a correlation structure. The smoothed random curves for each variable are summarized using a few important principal components, and the association of the three longitudinal components is modeled through the association of the principal component scores. The difficulties in handling the ordinal and proportional variables are solved by using a quasilikelihood type approximation. We develop an efficient algorithm to fit the model, which involves the selection of the number of principal components. The method is applied to physical activity data, and is evaluated empirically by a simulation study.

Understanding the time-varying associations between two functional measurements

♦*Haochang Shou*[1]*, Simon Simon Vandekar*[1]*, Lihong Cui*[2]*, Vadim Zipunnikov*[3] *and Kathleen Merikangas*[2]
[1]University of Pennsylvania
[2]National Institutes of Mental Health
[3]Johns Hopkins University
haochang.shou@gmail.com

The availability of multiple domains of measurements in cohort studies poses challenges for statistical modeling. One motivating example is from the NIMH family study of spectrum disorders where daily physical activity profiles are continuously observed for two weeks using accelerometers, while in the mean time, ecological momentary assessments (EMA) are also surveyed four times a day for the same time period. Researchers are interested in understanding the time-varying associations between the two types of measurements, and after adjusting for time-varying confounding, understanding the link between disease-related deep phenotypes, such as physical activity, with mood disorders. In another example from Philadelphia Neurodevelopmental Cohort (PNC), one is interested in the time-updated relationship of brain perfusion and body growth through child development, and how they are associated with gender or social cognition. We propose functional data methods that estimate the time-dependent associations of the two types of functional data and test whether such associations vary over time.

Prediction Models of Dimentia Transition using longitudinal structural brain images

♦*Seonjoo Lee*[1]*, Liwen Wu*[2] *and Yaakov Stern*[2]
[1]NYSPI and Columbia University
[2]Columbia University
sl3670@cumc.columbia.edu

Endpoint prediction using (high-dimensional) multivariate longitudinal predictors is of interest. We model longitudinal trajectories of the predictors using mixed effect model, and build a function on scalar measurement model with the trajectories as predictors for the outcomes following exponential distribution. During the talk, extensive simulation studies will be presented to examine performance of the proposed method. Finally, Alzheimer's Disease Neuroimaging Initiative (ADNI) data analysis results will be presented.

Optimal Design for Sparse Functional Data

♦*So Young Park*[1]*, Luo Xiao*[1]*, Jayson Wilbur*[2] *and Ana-Maria Staicu*[1]
[1]North Carolina State University
[2]METRUM Research Group
spark13@ncsu.edu

We consider an optimal design problem for sparse functional data. The primary objective is to find optimal sampling points for future data collection such that response can be most accurately predicted with the observations collected at those points. We formulate the problem as an optimization problem, and provide a unifying formulation for two major functional model frameworks: functional principal component analysis (FPCA) and functional linear model (FLM). We also propose a method for selecting number of optimal sampling points. Performance of the proposed method is thoroughly investigated via simulation study and application to real data example.

Session 97: Recent Statistical Methodology Developments for Medical Diagnostics

Combining large number of weak biomarkers based on AUC

♦*Li Yan*[1]*, Lili Tian*[2] *and Song Liu*[1]
[1]RoswellPark Cancer Institute
[2]SUNY University at Buffalo
Li.Yan@roswellpark.org

Combining multiple biomarkers to improve diagnosis and/or prognosis accuracy is a common practice in clinical medicine. Both parametric and non-parametric methods have been developed for finding the optimal linear combination of biomarkers to maximize the area under the receiver operating characteristic curve (AUC), primarily focusing on the setting with a small number of well-defined biomarkers. This problem becomes more challenging when the number of observations is not order of magnitude greater than the number of variables, especially when the involved biomarkers are relatively weak. Such settings are not uncommon in certain applied fields, for example the development of innovative diagnostic tools based on new generation of biotechniques for personalized medicine. We empirically evaluated the performance of some existing linear combination methods under such settings, and proposed a new combination method, namely, the pairwise approach, to maximize AUC. Our simulation studies demonstrated that the performance of several existing methods can become unsatisfactory as the number of markers becomes large, while the newly proposed pairwise method performs reasonably well. Furthermore, we compared these methods with real datasets used for the development and validation of gene signatures.

Comparing two correlated diagnostic tests based on joint testing of the AUC and the Youden index

♦*Jingjing Yin*[1]*, Lili Tian*[2] *and Hani Samawi*[1]
[1]Biostatistics, Georgia Southern University
[2]Biostatistics, University at Buffalo
jyin@georgiasouthern.edu

In the ROC analysis, the area under the ROC curve (AUC) serves as an overall measure of a biomarker/diagnostic test's accuracy. Another popular index is Youden index (J), which corresponds to the maximum sum of sensitivity and specificity thus can be used for diagnostic threshold optimization. Although researchers mainly evaluate the diagnostic accuracy using the AUC, for the purpose of making diagnosis, Youden index provides a direct measure of the diagnostic accuracy at the optimal threshold and hence should be taken into consideration in addition to the AUC. Our previous research proposed the joint confidence region of the AUC and the Youden index for a single test. Furthermore, it is very common to compare the diagnostic accuracy of two tests in a paired design. To see if one biomarker is more preferable in terms of both AUC and Youden index, we can test Ho: AUC1-AUC2=0 and J1-J2=0 versus Ha: AUC1-AUC2¿0 and J1-J2¿0. The existing approach for testing such order restrictive hypothesis is the intersection-union test (IUT), which marginally test the AUC and the Youden index independently. We propose an alternative test procedure in both parametric and non-parametric settings, which is shown by simulations to be much more powerful than IUT test under the alternative and maintain the type I error rate under the null.

Empirical Likelihood Confidence Regions in the Evaluation of Medical Tests with Verification Bias

♦*Gengsheng Qin*[1] *and Binhuan Wang*[2]
[1]Georgia State University
[2]New York University
gqin@gsu.edu

In this paper, we propose various bias-corrected empirical likelihood confidence regions for any two of the three parameters, sensitivity, specificity and cutoff-value, with the remaining parameter fixed at a given value in the evaluation of a continuous-scale diagnostic test with verification bias. The proposed methods can be used to select a good and reliable cut-off value for a continuous-scale medical test. Simulation studies are conducted to evaluate the finite sample performance and robustness of the proposed empirical likelihood-based confidence regions in terms of coverage probabilities. A real case analysis is provided to illustrate the application of the proposed methods.

MLEs for diagnostic measures of accuracy by log-linear models for correction of workup bias.

♦*Haresh Rochani, Hani Samawi, Robert Vogel and Jingjing Yin*
Jian Ping Hsu College of Public Health
hrochani@georgiasouthern.edu

In diagnostic medicine, the test that determines the true disease status without an error is referred to as the gold standard. Even when a gold standard exists, it is extremely difficult to verify each patient due to the issues of costeffectiveness and invasive nature of the procedures. In practice some of the patients with test results are not selected for verification of the disease status which results in verification bias for diagnostic tests. The ability of the diagnostic test to correctly identify the patients with and without the disease can be evaluated by measures such as sensitivity, specificity and predictive values. However, these measures can give biased estimates if we only consider the patients with test results who also underwent the gold standard procedure. The emphasis of this paper is to apply the log-linear model approach to compute the maximum likelihood estimates for sensitivity, specificity and predictive values. We also compare the estimates with Zhou's results and apply this approach to analyze Hepatic Scintigraph data under the assumption of ignorable as well as non-ignorable missing data mechanisms. We demonstrated the efficiency of the estimators by using simulation studies.

Session 98: Combining Information in Biomedical Studies

Inference Using Combined Longitudinal and Survival Data in CKD

Wei Yang
University of Pennsylvania
weiyang@upenn.edu

In Chronic Kidney Disease (CKD) research, glomerular filtration rate (GFR) is an important marker of kidney function. End-stage renal diseases (ESRD), which include kidney transplantation and dialysis, and death are the clinical endpoints of CKD. Statistical inference is often made for either the longitudinal GFR measures or the survival outcomes separately. However, there are challenges in analyzing both types of outcomes. In the analysis of repeated GFR measures, there is a concern of informative missing, i.e., those who reached ESRD or death had low GFR levels on average than those who did not reach the clinical endpoints. Furthermore, the GFR level is not meaningful any more after patient's death. Informative censoring by death is also a concern in the analysis of ESRD, i.e., those who died would have higher risk of ESRD if they were alive. In this talk, we will review a few methods that extends the idea of principal stratification to handle the censoring by death issue. We will also talk about a specific causal modeling approach using a multi-state representation of both types of outcomes.

An adaptive Fisher's method for combining information across samples

♦*Xiaoyi Min*[1]*, Chi Song*[2] *and Heping Zhang*[3]
[1]Georgia State University
[2]Ohio State University
[3]Yale University
xmin@gsu.edu

Motivated by the problem of detecting DNA copy number variations in multiple samples, we propose an adaptive Fisher's method for combining information across samples. We prove that in the general problem of detecting heterogeneous and heteroscedastic Gaussian mixtures, this method could detect the non-null effects within the detectable region established by Cai, Jeng and Jin (2011, Journal of the Royal Statistical Society, Series B). The power of this method is shown to be quite robust to proportion of non-null component for most practical settings. Finally, the proposed method is to the CNV detection in a real dataset.

Bayesian Latent Hierarchical Model for Transcriptomic Meta-Analysis

Zhiguang Huo[1]*,* ♦*Chi Song*[2] *and George Tseng*[1]
[1]University of Pittsburgh
[2]The Ohio State University
song.1188@osu.edu

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have

been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. In this talk, I will introduce a Bayesian latent hierarchical model based on one-sided p-values from differential/association analysis to perform transcriptomic meta-analysis. The p-value based method is capable of combining data from different microarray and RNA-seq platforms, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and two examples using a microarray dataset from metabolism related knockout mice and an RNA-seq dataset from HIV transgenic rats are used to demonstrate performance of the proposed method.

Genetic Effect and Association Test for Covariance Heterogeneity in Multiple Trait Comorbidity

Yuan Jiang[1], ♦*Yaji Xu*[2] *and Heping Zhang*[3]
[1]Oregon State University
[2]Food and Drug Administration
[3]Yale University
yaji.xu@fda.hhs.gov

Genes and environmental factors may not only contribute to the prevalence of diseases, but can also create a "ripple effect" on multiple disorders. A concrete example includes the case that the population with a particular gene tend to develop multiple drug abuses simultaneously. To answer this question, we propose a new concept of genetic effects on multiple trait covariance, in contrast to the usual definition of genetic effects on disease prevalence. In addition to the new concept, we develop an association test for the covariance heterogeneity among multiple traits, unlike most existing methods that assume homogeneity of the covariance. Preliminary results show that ignoring the genetic effect on multiple trait covariance can result in loss of power when performing genetic association test for comorbidity. We also provide evidences for the importance of the new genetic effect from investigation of a real data set.

Session 99: New Frontiers in Genomics and Precision Medicine

Estimating interactions between a treatment and a large number of genomic features

James Dai
Fred Hutchinson Cancer Research Center, Seattle
jdai@fredhutch.org

We consider clinical trials with a large number of genomic features measured at baseline for discovering predictive markers of treatment effect. Modeling interactions between the treatment and the high-dimensional genomic features requires adequate modeling the main effects of the genomic features. We discuss two strategies that exploit the gene-treatment independence that eliminate the need to model the high-dimensional main effects. The first is the generalized case-only estimator, through which we are estimating the risk-ratio estimator of interaction. The rare-disease assumption for the usual case-only estimator is not required and the cost of genomic profiling is greatly reduced. The second strategy involves modified covariates for genomic features, as proposed in the recent literature. The gene-treatment interaction can be estimated in the risk difference scale without modeling the main effects. The two strategies were compared in simulations and data application.

New approaches for genetic association mapping with large-scale genetic data in diverse populations

♦*Timothy Thornton and Caitlin McHugh*
University of Washington
tathornt@uw.edu

GWAS and sequencing studies of complex traits in ancestrally diverse populations have recently become more common due to increased interest in both identifying novel, population specific variants that underlie phenotypic diversity and generalizing associations across populations. Admixed populations, such African Americans and U.S. Hispanics/Latinos, who have recent ancestry from multiple continents pose special challenges for genetic association mapping due to heterogeneous genetic backgrounds and allele frequency differentiation among populations that varies greatly across the genome; which can result in spurious associations if not properly accounted for in the association analysis. We consider the problem of genetic association testing in admixed population samples. New approaches for detecting genetic associations with large-scale genetic data in the presence of complex sample structure will be discussed. The utility of the methods will be demonstrated in applications to genome-wide data from large Hispanic cohort studies for complex trait mapping.

A Statistical Framework for Pathway and Gene Identification from Integrative Analysis

♦*Quefeng Li*[1], *Menggang Yu*[2] *and Sijian Wang*[2]
[1]University of North Carolina, Chapel Hill
[2]University of Wisconsin, Madison
liquefeng@gmail.com

In the era of big data, as the individual patient data (IPD) become more accessible, integrative analyses using IPD from multiple studies are now extensively conducted to identify prognostic genes. It has been recognized that genes do not work alone but through pathways. In this paper, we propose a general statistical framework for pathway and gene identification from integrative analysis. Our method employs a hierarchical decomposition on genes' effects followed by a proper regularization to identify important pathways and genes across multiple studies. Asymptotic theories are provided to show that our method is both pathway and gene selection consistent. We explicitly show that pathway selection consistency needs milder statistical conditions than gene selection consistency, as it would allow false positives/negatives at the gene selection level. Finite-sample performance of our method is shown to be superior than other ad hoc methods in various simulation studies. We further apply our method to analyze five cardiovascular disease studies.

Integrated analysis of DNA methylation and gene expression data in human aging

♦*Karen Conneely*[1], *Elizabeth Kennedy*[1], *Alicia Smith*[2], *Elisabeth Binder*[3] *and Kerry Ressler*[4]
[1]Department of Human Genetics, Emory University
[2]Department of Psychiatry, Emory University
[3]Department of Psychiatry, Max Planck Institute
[4]Department of Psychiatry, Harvard University
kconnee@emory.edu

Epigenome-wide association studies in humans have reported thousands of age-differentially-methylated CpG sites, and recent studies show that age can be predicted from DNA methylation data with great accuracy across a wide range of cell and tissue types. However, the role of these DNA methylation changes remains unelucidated. In whole blood, the profile of associations between age and gene expression has been reported to not align well with the

age-methylation association profile. There are several possible explanations for this phenomenon, including the possibility that many of the methylation changes observed in whole blood are not directly functional, but may be marks or side effects of another process. In this work, we use integrated data on methylation and expression in conjunction with the results of a large well-powered study of age and expression to further investigate whether age-related changes in methylation associate with changes in expression. Our ultimate goal is to explain the widely observed pattern of age-changes in methylation. Though this goal may be best achieved through analysis of longitudinal and/or familial data in multiple tissues, the aim of our current work is to see how far we can get with available cross-sectional microarray data, and to generate preliminary data and predictions for future studies.

Session 100: New Developments in BFF Inferences in the Era of Data Science

Fiducial Inference: Fisher's Big Blunder or Big Bang?

♦*Keli Liu*[1] *and Xiao-Li Meng*[2]
[1]Stanford University
[2]Harvard University
keliliu@stanford.edu

Estimating equations are a popular avenue for creating point estimates; in avoiding a full specification of the sampling distribution, they are often praised for their "objectivity" Fisher dared to also obtain a distributional inference by solving a system of equations; his Fiducial method avoids specification of the prior distribution but is generally viewed as a failed attempt at objective Bayes. However, the era of Big Data (and Efron's speculation that "Maybe Fisher's biggest blunder will become a big hit in the 21st century!" encourages us to look at Fiducial from a new perspective, as an algorithmic revolution: stochastic algebra replaces Markov Chain Monte Carlo in computing posterior inferences. The trick to achieving this algorithmic jujitsu lies in converting the usual problem of inferring signals into its dual form of predicting noise. Applying classical inference techniques, in particular partial likelihood, to the dual problem results in Fisher's Big Blunder (or Big Bang!): Fiducial inference.

Applications of the Poisson Dempster-Shafer Model (DSM) and the General Univariate DSM for Inference of Infection Time from Acute HIV-1 Genomes

Paul Edlefsen
Fred Hutchinson Cancer Research Center
pedlefse@fredhutch.org

HIV-1 is an RNA virus that diversifies rapidly after initial infection, and in most cases the Hamming distances among the aligned nucleotide RNA sequences of HIV-1 sampled from blood are well described by a Poisson distribution, at least for the most acute phase of the infection. Poisson goodness-of-fit tests are routinely employed to differentiate between samples conforming to a single-founder, purely drift process with low linkage (a "star-like phylogeny") versus those in violation of that model due to multiple initial "founders" of the infection or to varying selective pressures inducing viral lineages ("quasispeciation"). When there is a good fit, the Poisson model is then employed for estimation of HIV-1 infection times. Both the accurate estimation of HIV-1 infection times and the accurate assessment of the uncertainty of those estimates have important impact on the statistical properties of methods for evaluating correlates of vaccine efficacy in prophylactic HIV-1 clinical trials. Thus we sought to develop an approach to inference of HIV-1 infection times that yields uncertainty estimates of the Dempster-Shafer and Bayesian variety (that is, probabilistically-interpreted posterior assessments) rather than the typical confidence statements that are constructed by inverting small-sample-size-adjusted asymptotic hypothesis tests. Here we discuss the application of the classic Poisson DSM to the HIV-1 infection timing estimation problem, yielding posterior DSMs to address questions like "what is the evidence unambiguously against this infection having occurred over 30 days ago" or Bayesian posterior credible intervals. We also introduce the General Univariate DSM and its application to goodness-of-fit testing

Fusion Learning: combining inferences using data depth and confidence distribution

♦*Dungang Liu*[1]*, Regina Liu*[2] *and Min-ge Xie*[2]
[1]University of Cincinnati
[2]Rutgers University
dungang.liu@uc.edu

For the purpose of combining inferences from several nonparametric studies for a common hypothesis, we develop a new methodology using the concepts of data depth and confidence distribution. A confidence distribution (CD) is a sample-dependent distribution function that can be used to estimate parameters of interest. It is a purely frequentist concept yet can be viewed as a "distribution estimator" of the parameter of interest. In this project, we use the concept of CD, coupled with data depth, to develop a new approach for combining several nonparametric studies for a common multivariate parameter. This approach has several advantages. First, it allows us to resample directly from the empirical distribution, rather than from the estimated population distribution satisfying the null constraints. Second, it enables us to obtain test results directly without having to construct an explicit test statistic and then establish or approximate its sampling distribution. The proposed method provides a valid inference approach for a broad class of testing problems involving multiple studies where the parameters of interest can be either finite or infinite dimensional. The method will be illustrated using simulations and flight data from the Federal Aviation Administration (FAA).

Session 101: Topics in Statistics

Modelling cumulative effects of air pollution on respiratory illnesses

Xingfa Zhang
Guangzhou University China
xingfazhang@hotmail.com

Motivated by exploring and understanding the impact of air pollution on respiratory illnesses, this research developed a spline estimation for constraint additive single index model. It is widely recognized that the exposure in air pollution causes serious respiratory illnesses and the weather may also contribute to the seriousness. However, it is difficult to quantify the effects of pollutants and weather conditions due to the high unknown nonlinear relationship between illness and the environmental and climatic factors and incubative periods from the impact to the symptoms. The index model is proposed to capture the cumulative effect of the potential environmental and climatic factors, while the unknown nonlinearity is modelled by a semiparametric approach. Constraints are imposed on the index components based on a natural physical consideration. In order to derive a sensible estimation for this model with complicated structure and computing issues faced in the nonparametric functions, spline estimation is implemented due to its robustness

and ease of computing. The model is applied to the data collected from Hong Kong. Due to the SARS (Severe acute respiratory syndrome) epidemics in Hong Kong in 2003, extra care is needed in constructing a growth curve model to account for the impact by SARS. The results are comparable with previous researches and the length of incubative periods provide useful information on the impacts of pollutants on the respiratory illnesses over time and have potential applications in health monitoring program. The effects by weather factors are consistent with the findings in environmental research.

KEY WORDS: air pollution; cumulative effect; respiratory illness; severe acute respiratory syndrome (SARS); spline estimation.

Statistical Analysis of cochlear shape

♦*jean michel loubes, jose braga and laurent risser*
université de toulouse
loubes@math.univ-toulouse.fr

The site of Kromdraai B (KB) (Gauteng, South Africa) has yielded a minimum number of nine hominins including the type specimen of Paranthropus robustus (TM 1517), the only partial skeleton of this species known to date. Four of these individuals are juveniles, one is a subadult and four are young adults. They all occur with a macrofaunal assemblage spread across the succession of at least two time periods that occurred in South Africa approximately two million years ago. Here we report on an additional, newly discovered petrous temporal bone of a juvenile hominin, KB 6067. Following the description of KB 6067, we assess its affinities with Australopithecus africanus, P. robustus and early Homo. We discuss its developmental age and consider its association with other juvenile hominin specimens found at Kromdraai B. KB 6067 probably did not reach five years of age and in bony labyrinth morphology it is close to P. robustus, but also to StW 53, a specimen with uncertain affinities. However, its cochlear and oval window size are closer to some hominin specimens from Sterkfontein Member 4 and if KB 6067 is indeed P. robustus this may represent a condition that is evolutionarily less derived than that shown by TM 1517 and other conspecifics sampled so far. The ongoing fieldwork at KB, as well as the petrography and geochemistry of its deposits, will help to determine when the various KB breccias accumulated, and how time may be an important factor underlying the variation seen among KB 6067 and the rest of the fossil hominin sample from this site. We present the statistical study involving wasserstein distance on interpoint distributions in a deformation process framework.

Improving Online Education by Understanding Demographic Effects on Student Success and Retention

James Monroe
Kennesaw State University
kjamesmonroe@gmail.com

This study analyzes data from courses at Kennesaw State University to find trends related to student demographics and characteristics. I used multiple methods to test for significant predictors and interactions among the demographic categories of gender, race, class, age, and citizenship, as well as individual characteristics such as GPA, semesters of experience in college, and the level of the course being taken. With multiple regression I tested for significance in regards to relative success (grade in the course) and with logistic regression I tested for significance in regards to retention (pass or drop/withdraw/fail). Using a general linear model I tested for interaction between race and sex in relative success rates, both overall and in regards to class type. I used Chi-Square to examine trends in choice of class type by sex, by race, and by both sex and race. Another Chi-Square was performed to examine trends in class type by class level.

GPA is the strongest predictor of success and retention, included to control for its effects. Being female or being in a higher level course predicted success and retention in almost every case. However in opposition to the effects of course level, a student's semesters of experience were inversely related to their success in the course, as was their age. Having a more wealthy family increased success and retention, while being of a low enough SES to receive the Pell grant had a mixed impact. Unexpectedly, students do better in online and hybrid courses when they have not taken one before. Contrary to findings in studies which don't control for GPA and class, race has very mild effects on success and retention. Race has an insignificant effect in most cases, and interacts with gender only for Native American and white students. Trends in choice of course type were also examined and showed strong variance by race and gender. Overall, these findings illustrate the need for considering demographics while designing courses.

A Mixed Variance Component Model for Quantifying the Elasticity Modulus of Nanomaterials

♦*Angang Zhang*[1] *and Xinwei Deng*[2]
[1]Merck
[2]Virginia Tech
angang8@vt.edu

Nanomaterials possess great mechanical properties with wide applications in many areas. Experiments are often conducted for measuring certain mechanical properties of interest. How to accurate quantifying mechanical properties of nanomaterials is thus very important but challenge due to nanoscale manipulation and tactful measurement technique. Statistical modeling approach combined physical theories have been used for the quantification of nanomaterials. In this work, we propose a novel mixed variance component model to accommodate experimental variations and artifacts for analyzing the nanomaterial experiment data. The proposed method can automatically adjust systematic errors occurred in the experiments through a group adaptive forward backward selection (GFoBa). It thus leads to accurate estimation of mechanical properties with the ability to filter out various experimental errors. The performance of the proposed method is compared with other existing method through both simulation and a real data example.

The power of Rmarkdown and RStudio IDE: Lessons Learned Teaching R in Public Health and Nursing

Melinda Higgins
Emory University
melinda.higgins@emory.edu

All students struggle at first learning a new programming or any programming language. This is true for statistics students but is especially so for students who are in non-statistics majors such as public health and nursing, most of whom have never written a computer program before. On top of learning a programming language and environment such as those in R, SAS, SPSS, and Stata, these students are also often struggling simultaneously with understanding new statistical concepts while also figuring out how to apply these concepts to their research questions and field of study. In addition to the challenges of (a) applying statistical concepts to address research questions and (b) figuring out what programming is and how to do it, these students are also learning how to read their output from a statistical program and what to do with it. Statistical programs like SAS and SPSS provide very verbose output which is helpful since most procedures provide output related to statistical

assumptions of a given test or model, such as the test for homogeneity of variances needed for choosing the correct results from a 2-group independent t-test (pooled or unpooled). But this verbose output poses the challenge to the student of figuring out which part of the output is important and which numbers to use for interpreting and drawing conclusions. Programming languages like R, however, make getting statistical results even more challenging since the results have to be specifically requested (i.e. you have to know what you want and ask for it). In the long run, having to explicitly request the results you want and knowing why deepens the students understanding of the results and statistical analysis process. Helping to overcome these challenges is the R package, Rmarkdown, and the supporting interface for dynamic documentation within RStudio. After teaching R to a number of public health and nursing students at Emory and as well as public health researchers at the Centers for Disease Control (CDC), the consensus is that these "students" were thrilled with Rmarkdown and wanted to learn it first so it would be the conduit to learning R. Rmarkdown and the resulting document produced (HTML, PDF, or DOC) made the utility of learning R immediately clear (tangible reports and manuscript drafts). Rmarkdown also gave the students clear connectivity between (1) their stated research questions and objectives, (2) the statistical method chosen to address those questions, (3) the code to execute the statistical method and (4) the resulting documentation interpreting their results and presenting them.

Different estimations of time series models and application for foreign exchange in emerging markets

Jingjing Wang
Student
corrinewang01@gmail.com

Recent years,with the rapid development of emerging markets, the emerging markets have been enjoying a more and more significant role in the whole markets. After the financial crisis, the economic policy of developed markets(especially The United States) has a really big influence on the emerging market countries. So under such condition, we use different time series modesl to model the foreign exchange data of some emerging markets and make comparison with different models and talk about the volatility of these data sets.

Session 102: Topics in Biostatistics

A Maximum Likelihood Approach for Non-invasive Cancer Diagnosis Using Methylation Profiling of Blood

♦*Carol Sun*[1] *and Wenyuan Li*[2]
[1]Oak Park High School
[2]University of Southern California
fsunster@gmail.com

Cancer is the second most common killer for Americans and it will become the leading cause of death within the next decade. Fortunately cancer can be managed and even cured if diagnosed early. Thus early detection of cancer is essential for public health. Even at the early stage of cancer, some tumor cells can shed their DNA, tDNA, into the blood making it possible to use blood for early cancer diagnosis. Tumor DNAs differ from normal DNAs in their methylation patterns with a large fraction of CpG sites hyper- or hypo-methylated compared to normal cells Thus, it is possible to use methylation data of the blood samples for early cancer diagnosis. Using the methylation data for cancer and normal tissues from The Cancer Genome Atlas (TCGA), we design a maximum likelihood approach and a corresponding computational method to estimate the fraction of tumor DNAs in blood samples using next generation sequencing data. We model the DNAs from blood samples as a mixture of normal and tumor DNAs and assume the distributions of methylation levels for both normal and tumor DNAs to have different beta distributions. We study the effects of sequencing depth and fraction of tumor DNAs on the estimation accuracy using simulations. It is shown that the relative accuracy increases with sequence depth and the fraction of tumor DNAs. Applications to a set of normal individuals' and liver cancer patients' data, we show that our method can separate normal individuals from cancer patients well. We also show that the fraction of tDNAs in cancer patients is significantly decreased after surgery. Our method provides an effective approach for early cancer diagnosis using blood samples.

Parametric Bootstrap in Meta-analyses to Construct CIs for Event Rates and Differences in Event Rate

♦*Gaohong Dong*[1]*, Jennifer Ng*[2]*, Steffen Ballerstedt*[3] *and Marc Vandemeulebroecke*[3]
[1]iStats Inc. and Infotree Service Inc.
[2]Novartis Pharmaceuticals Corporation
[3]Novartis Pharma AG
Gaohong_Dong@istats.org

The pediatric investigational plan for Certican (everolimus) included pediatric studies in liver and kidney transplantation. However, the very slow enrolment made it impossible to recruit patients in a timely manner. Following the recent EMA concept paper on extrapolation (EMA, 2013) and with consultations with EMA, we developed an extrapolation methodology bridging adult and pediatric data via meta-analyses, which included 57 adult studies (¿19500 patients) and 7 pediatric studies (651 patients).

For the extrapolation, the meta-analytic model is parameterized such that the covariate effects are linear on the log odds scale. For the estimated event rates and differences in rates, initially we obtained confidence intervals (CIs) based on the delta method directly from PROC NLMIXED. However these intervals were deemed unsatisfactory, since the nonlinearity of the parameter transformation led to poor coverage properties and to nonsensical CI limits (i.e. negative lower limits). Therefore, we derived CIs via a parametric bootstrap approach.

In this talk, we will briefly introduce maximum likelihood and Bayesian meta-analysis models, then focus on the parametric bootstrap to construct CIs for event rates or differences in rates. We will also explain why this bootstrap approach is needed for maximum likelihood meta-analyses, but not for Bayesian meta-analyses.

Note: The first author carried out his work when he was employed by Novartis.

Estimation of Energy Expenditure

Shan Yang
Merck & Co., Inc.
shan.yang@merck.com

Abstract Glucagon is a peptide hormone that acutely stimulates energy expenditure and hepatic glucose production in humans. It is unclear whether these effects are sustained during prolonged hyperglucagonemia. To better understand the effects of glucagon on the amount of calories burned, a Phase 1, randomized, single blinded, placebo controlled, 3-treatment regimen crossover study was undertaken. Sleep energy expenditure and resting energy expenditure were compared between healthy volunteers who received an overnight glucagon infusion and those who received placebo infusions. Estimation results under different models, including a mixed

model and stochastic differential equations, are discussed using data simulated based on the results of the clinical study.

Meta-Analysis of Rare Binary Events in Treatment Groups with Unequal Variability

♦*Lie Li*[1]*, Ou Bai and Xinlei Wang*
[1]Southern Methodist University
liel@smu.edu

Meta-analysis has been widely used to synthesize information from related studies to achieve reliable findings. One of the most important applications is to evaluate treatment/intervention effects between two experimental conditions by combining data from multiple clinical trials. However, in studies of rare events, the event counts are often low or even zero, and standard meta-analysis methods based on fixed-effect models do not work well. Recently, to estimate the overall treatment effect, Bhaumik et al. (2012) developed a simple average (SA) estimator based on a random-effects model. They showed that the SA estimator with a continuity correction 0.5 is the least biased and has superior performance when compared with other commonly used estimators. However, the random-effects models used are restrictive because they assume that the variability in the treatment group is equal to or always greater than that in the control group. Under a general framework that allows treatment groups with unequal variability but assumes no direction, we consider the mean squared error (MSE) to assess the estimation efficiency and show the SA estimator cannot achieve the optimal MSE. We also adapt and extend two other methods to estimate the overall treatment effect, including median and shrinkage estimators. Under a new random-effects model that accommodates groups with unequal variability, we compare the performance of various methods for both large and small samples. Real data analysis will be conducted as well. Practical guidelines about when to use which estimator will be provided.

Combining Evidence of Regional Treatment Effects under Discrete Random-Effects Model (DREM) in MRCT

♦*Hsiao-Hui Tsou*[1]*, K. K. Gordon Lan*[2]*, Chi-Tian Chen*[1]*, H.M. James Hung*[3]*, Chin-Fu Hsiao*[1] *and Jung-Tzu Liu*[1]
[1]National Health Research Institutes
[2]Janssen Pharmaceutical Companies of Johnson & John
[3]US Food and Drug Administration
tsouhh@nhri.org.tw

In recent years, developing pharmaceutical products via a multiregional clinical trial (MRCT) has become standard. Traditionally, the treatment effect was assumed to have a fixed positive value over all regions in an MRCT. However, regional heterogeneity caused by differences in race, genetic fingerprints, diet, environment, culture, medical practice, and disease etiology among regions in an MRCT has been observed and may have an impact upon a medicine's treatment effect. In this presentation, we will discuss the issue of combining evidence of regional treatment effects in multiregional clinical trials. We will introduce a discrete random-effects model (DREM) to account for heterogeneous treatment effects across regions for the design and evaluation of MRCTs. We will illustrate determination of the overall sample size and address the issues of sample size re-calculation due to the uncertainty in the treatment effect or variability assumptions.

Meta-analysis with incomplete multinomial data: An application to tumor response in cancer patients

♦*Charity J. Morgan, Pooja Ghatalia and Guru Sonpavde*
University of Alabama at Birmingham
cjmorgan@uab.edu

Multinomial data may be "incomplete" if there exist observations that are only partially classified. That is, for some observations, while it is known that the response belongs to a given subset of the possible response categories, the true, precise classification is unknown. The incomplete multinomial distribution and incomplete multinomial regression have special relevance for the field of meta-analysis, in that studies may report multinomial outcomes using slightly different sets of response categories, complicating direct comparison of their results. For example, while many clinical trials for cancer patients use the Response Evaluation Criteria in Solid Tumors (RECIST) criteria to assess tumor response, investigators often choose to combine categories when reporting results. We discuss the extension of incomplete multinomial regression to the meta-analytic setting and demonstrate its use in a meta-analysis of tumor response in patients receiving either placebo or no anti-cancer therapy.

Session 103: New Development on Missing Data Problems

Multiply robust imputation procedures for the treatment of item nonresponse in surveys

♦*Sixia Chen*[1] *and David Haziza*[2]
[1]University of Oklahoma
[2]University of Montreal
Sixia-Chen@ouhsc.edu

Item nonresponse in surveys is often treated through some form of imputation. We introduce the concept of multiply robust imputation procedures in the context of finite population sampling, which is closely related to the concept of multiple robustness that can be viewed as an extension of the concept of double robustness. In practice, multiple nonresponse models and multiple imputation models may be fitted, each involving different subsets of covariates and possibly different link functions. An imputation procedure is said to be multiply robust if the resulting estimator is consistent if all but one model are misspecified. A jackknife variance estimator is proposed and shown to be consistent. Extension to random and fractional imputations is discussed. Finally, the results of a simulation study, assessing the performance of the proposed point and variance estimators in terms of bias and efficiency, are presented.

Empirical Likelihood Methods for Complex Surveys with Data Missing-by-Design

♦*Changbao Wu, Min Chen and Mary Thompson*
University of Waterloo
cbwu@uwaterloo.ca

We consider nonrandomized pretest-posttest designs with complex survey data for observational studies. We show that two-sample pseudo empirical likelihood methods provide efficient inferences on the treatment effect, with a missing-by-design feature used for forming the two samples and the baseline information incorporated through suitable constraints. The proposed maximum pseudo empirical likelihood estimators of the treatment effect are consistent and pseudo empirical likelihood ratio confidence intervals are constructed through bootstrap calibration methods. The proposed methods require estimation of propensity scores which depend on the underlying missing-by-design mechanism. A simulation study is conducted to examine finite sample performances of the proposed methods under different scenarios of ignorable and nonignorable missing patterns. An application to the International Tobacco Control Policy Evaluation Project Four Country Surveys (ITC 4-County) is also

presented.

Pseudo-population bootstrap methods for imputed survey data

♦*David Haziza*[1]*, Zeinab Mashreghi*[2] *and Christian Léger*[1]
[1]Université de Montréal
[2]University of Winninpeg
david.haziza@umontreal.ca

Item non-response in surveys is usually dealt with through single imputation. Treating the imputed values as if they were observed values may lead to serious underestimation of the variance of point estimators. Two approaches are used for deriving bootstrap variance estimators: the non-response model approach and the imputation model. We propose three pseudo-population bootstrap schemes: the first two lead to an approximately unbiased variance estimator with respect to the non-response model approach and the imputation model approach, respectively. The third scheme leads to a doubly robust bootstrap variance estimator that is approximately unbiased for the true variance if either the nonresponse model or the imputation model is correctly specified. Results from a simulation study suggest that the proposed methods perform well in terms of relative bias and coverage probability

Lack-of-fit testing of a regression model with response missing at random

Xiaoyu Li
Auburn University
xzl0037@auburn.edu

This paper proposes a class of lack-of-fit tests for fitting a linear regression model when the response observations are missing at random. These tests are based on a class of minimum integrated square distances between a kernel type estimator of a regression function and the parametric regression function being fitted. These tests are shown to be consistent against a large class of fixed alternatives. The corresponding test statistics are shown to have asymptotic normal distributions under null hypothesis and a class of nonparametric local alternatives.

Session 104: Advances in Ultra High Dimensional Data Analysis

Divergence based Inference for High-dimensional Regression Problems: Uncertainty Assessment, Robustn

Anand Vidyashankar
George Mason University
avidyash@gmu.edu

High dimensional data are ubiquitous in contemporary science and regression models are typically used to address related inferential questions. A strategy typically adopted by practicing statisticians is to first perform "exploratory analyses" and use model selection criteria such as BIC/GCV to select an appropriate model. The chosen model then gets treated as a "true model" and further inferences are performed. More recently, methods such as LASSO/ALASSO/MCP and their variants also get used towards simultaneous model selection and inference. It is now folklore that such a strategy towards inference may lead to inaccurate standard errors for the estimates of the regression parameters, potentially leading to erroneous decision making; also, additional complications arise if the underlying statistical model is misspecified. In this presentation, we provide a new framework, using divergences, to assess model selection variability and identify two important subcomponents: namely, intrinsic and extrinsic uncertainty. We evaluate the effect of these sub-components on the robustness and efficiency of inference. To achieve this, we establish the joint asymptotic distribution of the regression parameters accounting for model selection. To address the technical issues and make precise the notions of neighborhood arising in robust inference, we introduce a new key technical tool, *robust large deviations*-large deviations when the true probability distribution belongs to a neighborhood, which is determined by the divergence, of the postulated model. Our results can be applied to a variety of statistical models used in high-dimensional data analyses: for instance, generalized linear models, single index models, and partial linear models. We illustrate the ideas using examples from financial risk management and emerging drug resistance in biomedicine.

Tests for Nonparametric Interactions Using Random Forests

♦*Giles Hooker*[1] *and Lucas Mentch*[2]
[1]Cornell University
[2]SAMSI
gjh27@cornell.edu

While statistical learning methods have proved powerful tools for predictive modeling, the black-box nature of the models they produce can severely limit their interpretability and the ability to conduct formal inference. However, the natural structure of ensemble learners like bagged trees and random forests has been shown to admit desirable asymptotic properties when base learners are built with proper subsamples. We demonstrate that by defining an appropriate grid structure on the covariate space, we may carry out formal hypothesis tests for both variable importance and underlying additive model structure. To our knowledge, these tests represent the first statistical tools for investigating the underlying regression structure in a context such as random forests. We develop notions of total and partial additivity and further demonstrate that testing can be carried out at no additional computational cost by estimating the variance within the process of constructing the ensemble. Furthermore, we propose a novel extension of these testing procedures utilizing random projections in order to allow for computationally efficient testing procedures that retain high power even when the grid size is much larger than that of the training set.

High Dimensional Multivariate Testing with Applications in Neuroimaging

♦*Bret Hanlon*[1] *and Nagesh Adluru*[2]
[1]University of Wisconsin Statistics
[2]Waisman Laboratory for Brain Imaging and Behavior
hanlon@stat.wisc.edu

Abstract: Region of interest (ROI) testing is a commonly used tool in neuroimaging. The basic problem involves testing for a difference in the mean vectors of a collection of voxels between the healthy and disease groups. This problem typically occurs in a high-dimensional setting with the number of voxels p in the ROI, exceeding the number of subjects n. In this talk, we build on the work of Kuelbs and Vidyashankar (2010, Annals of Statistics) which developed an infinite-dimensional framework to study the comparisons of means when the number of parameters increase with the sample size. We describe an extension of their idea to more general statistical models, including regression models. Additionally, we provide an alternative computational method that does not require direct estimation of the covariance matrix. We illustrate the method on data from neuroimaging studies.

Conditional Variable Screening in High-Dimensional Binary Classification

♦*Guoqing Diao*[1] *and Jing Qin*[2]
[1]Department of Statistics, George Mason University

[2]National Institutes of Health, NIAID
gdiao@gmu.edu

Most existing variable screening methods for high-dimensional data rely on strong parametric modelling assumptions that are often violated in real applications. Recently, Mai and Zou (2013) proposed a Kolmogorov filter for high-dimensional binary classification based on the Kolmogorov-Smirnov statistic. This screening method, however, does not account for the effects of potential confounders. We propose a new nonparametric conditional screening method to assess the conditional contributions of the individual predictors in the presence of known confounders. A bootstrap method is proposed to assess the significance for each predictor. The proposed method retains the features of the Kolmogorov filter and is shown to enjoy the sure screening property under the much weakened model assumptions compared to the parametric conditional screening methods. We illustrate the proposed method through extensive simulation studies and real applications.

Session 105: Mixture Regression: New Methods and Applications

Order Dependence of Hypersphere Decomposition for Covariance Matrix

♦*Qingze Li and Jianxin Pan*
University of Manchester, UK
chinaliqingze@hotmail.com

In covariance matrix estimation and modelling, there are two obstacles, positive definiteness and order independence. Most existing methods guarantee the positive definiteness of the covariance matrix, but they suffer from the problem of order dependence, including Cholesky Decomposition and Modified Cholesky Decomposition (MCD), etc. A recently recongized method is the Hypersphere Decomposition (HPC), which is believed to be order-independent by many literature and is highly recommended in the areas of finance and risk management. In this paper, based on the HPC we first propose to jointly modelling of the mean and covariance structures using linear models strategy. We show that the estimators of the parameters in the mean and covariances are consistent and asymptotically normally distributed, which is true even in high-dimensional scenoaris where the dimension of covariance matrix p is much larger than the sample size n. Second, we show that the HPC is unfortunately order-dependent and the model parameters are lack of proper statistical interpretation. We then propose an improved Hypersphere Decomposition, which has clear statistical interpretation and is importantly order-independent. Theoretical properties and numerical evidences for the new HPC method are provided too.

Testing of multivariate spline growth model

♦*Jyrki Mottonen*[1] *and Tapio Nummi*[2]
[1]University of Helsinki
[2]University of Tampere
jyrki.mottonen@helsinki.fi

We present a new method for testing multivariate growth curves which is based on spline approximation and on F-test. We show how the basic spline regression model can easily be extended to the multiple response case. The method is illustrated by using a real data set.

Labor market attachment in early adulthood: A trajectory analysis approach

♦*Janne Salonen*[1]*, Tapio Nummi*[2]*, Antti Saloniemi*[2] *and Pekka Virtanen*[2]
[1]Finnish Centre for Pensions
[2]University of Tampere
janne.salonen@etk.fi

In this paper we investigate the labor market integration of young adults. The study is based on individual-level register data that contain information about working, studying, pensions and different types of social benefits in Finland. Using data from 2005 to 2013, we have studied the labour market attachment of the total birth year cohort of 1987. Our statistical method is a multivariate version of trajectory analysis that applies finite mixture modelling to longitudinal data. We have modelled the information about labour market outcomes as a binary dependent variable, and thus we apply a multivariate logistic regression model. The analysis is done for males and females separately, because their labour market situations are quite different. We find that there are ten main trajectories for males and females that lead to different labor market outcomes. Additionally some background information is provided on these groups at the end of the period of investigation. Information on, for example, family and working life support the findings of our trajectory analysis. Our results suggest that a trajectory analysis of a register-based population can reveal some new interesting information that may remain hidden in more formal census-based official statistics.

Keywords: trajectory analysis, education, labor markets, young people, working life

A semiparametric mixture regression model for longitudinal data

♦*Tapio Nummi*[1]*, Janne Salonen*[2]*, Lasse Koskinen*[1] *and Jianxin Pan*[3]
[1]University of Tampere, Finland
[2]The Finnish Centre of Pensions, Finland
[3]University of Manchester, UK
tan@uta.fi

A semiparametric normal mixture regression model for longitudinal data is proposed such that the model contains one smooth term and a set of possible linear predictors. Model terms are estimated using the penalized likelihood method with the EM-algorithm. A computationally appealing alternative that provides an approximate solution using an ordinary linear model methodology is also introduced. Simulation experiments and real data examples are used to illustrate the methods.

Session 106: Spatial and Spatio-temporal Statistical Modeling and their Applications

Disease Risk Estimation by Combining Case Control Data with Aggregated Information on Population at Risk

♦*Xiaohui Chang*[1]*, Rasmus Waagepetersen*[2]*, Herbert Yu*[3]*, Xiaomei Ma*[4]*, Theodore Holford*[4]*, Rong Wang*[4] *and Yongtao Guan*[5]
[1]College of Business, Oregon State University
[2]Dept of Mathematical Sciences, Aalborg University
[3]Epidemiology Program, University of Hawaii Cancer
[4]Yale School of Public Health
[5]Dept of Management Science, University of Miami
xiaohui.chang@oregonstate.edu

We propose a novel statistical framework by supplementing case-control data with summary statistics on the population at risk for a subset of risk factors. Our approach is to first form two unbiased estimating equations, one based on the case-control data and the other on both the case data and the summary statistics, and then optimally combine them to derive another estimating equation to be used for

the estimation. The proposed method is computationally simple and more efficient than standard approaches based on case-control data alone.We also establish asymptotic properties of the resulting estimator, and investigate its finite-sample performance through simulation. As a substantive application, we apply the proposed method to investigate risk factors for endometrial cancer, by using data from a recently completed population-based case-control study and summary statistics from the Behavioral Risk Factor Surveillance System, the Population Estimates Program of the US Census Bureau, and the Connecticut Department of Transportation.

Hierarchical Models for Spatial Data with Errors that are Correlated with the Latent Process

♦*Jonathan Bradley*[1]*, Christopher Wikle*[2] *and Scott Holan*[2]
[1]Florida State University
[2]University of Missouri
bradleyjr@missouri.edu

Prediction of a spatial Gaussian process using a "big dataset" has become a topical area of research over the last decade. The available solutions often involve placing strong assumptions on the error process associated with the data. Specifically, it has typically been assumed that the data is equal to the spatial process of principal interest plus a mutually independent error process. Furthermore, to obtain computationally efficient predictions, additional assumptions are placed on latent random processes and/or parameter models (e.g., low rank models, sparse precision matrices, etc.). We consider an alternative latent process modeling schematic, where it is assumed that the error process is spatially correlated and correlated with the spatial random process of principal interest. We show the counterintuitive result that error process dependencies allow one to remove assumptions on the spatial process of principal interest, and obtain computationally efficient predictions. At the core of this proposed methodology is the definition of a corrupted version of the latent process of interest, which we call the data specific latent process (DSLP). Demonstrations of the DSLP paradigm are provided through simulated examples and through an application using a large dataset consisting of US Census Bureau's American Community Survey estimates of median household income on census tracts.

Changes in Spatio-temporal Precipitation Patterns in Changing Climate Conditions

♦*Won Chang*[1]*, Michael Stein*[1]*, Jiali Wang*[2]*, Rao Kotamarthi*[2] *and Elisabeth Moyer*[1]
[1]University of Chicago
[2]Argonne National Laboratory
changwn@ucmail.uc.edu

Climate models robustly imply that some significant change in precipitation patterns will occur. Models consistently project that the intensity of individual precipitation events increases by approximately 6-7%/K, following the increase in atmospheric water content, but that total precipitation increases by a lesser amount (2-3%/K in the global average). Some other aspect of precipitation events must then change to compensate for this difference. We develop here a new methodology for identifying individual rainstorms and studying their physical characteristics - including starting location, intensity, spatial extent, duration, and trajectory - that allows identifying that compensating mechanism. We apply this technique to precipitation over the contiguous U.S. from both radar-based data products and high-resolution model runs simulating 100 years of business-as-usual warming. In model studies, we find that the dominant compensating mechanism is a reduction of storm size. In summer, rainstorms become more intense but smaller, in winter, rainstorm shrinkage still dominates, but storms also become less numerous and shorter duration. These results imply that flood impacts from climate change will be less severe than would be expected from changes in precipitation intensity alone. We show also that projected changes are smaller than model-observation biases, implying that the best means of incorporating them into impact assessments is via "data-driven simulations" that apply model-projected changes to observational data. We therefore develop a simulation algorithm that statistically describes model changes in precipitation characteristics and adjusts data accordingly, and show that, especially for summertime precipitation, it outperforms simulation approaches that do not include spatial information.

Estimating the Health Effects of Ambient Air Pollution Accounting for Spatial Exposure Uncertainty

♦*Howard Chang, Yang Liu and Stefanie Sarnat*
Emory University
howard.chang@emory.edu

Population studies of air pollution and health routinely assign exposures using measurements from outdoor monitoring networks that have limited spatial coverage. Moreover, ambient concentrations may not reflect human exposure to pollution from outdoor sources. As such, exposure uncertainty can arise in these studies due to unobserved spatial variation in ambient air pollution concentrations, as well as spatial variations in population or environmental characteristics that contribute to differential exposure. We will describe a daily time-series study of fine particulate matter and emergency department visits in Atlanta. To account for spatial exposure uncertainties, additional data sources are being incorporated to supplement ambient monitor measurements in a unified statistical modeling framework. Specifically, we first utilize remotely sensed data and simulations from numerical model to obtain spatially-resolved concentration predictions. These predictions are then combined with data from stochastic exposure simulators to obtain estimated personal exposures.

Session 107: Recent Development in Sufficient Dimension Reduction and Variable Selection

Variable selection via additive conditional independence

♦*Kuang-Yao Lee*[1]*, Bing Li*[2] *and Hongyu Zhao*[1]
[1]Yale University
[2]Pennsylvania State University
kuang-yao.lee@yale.edu

We introduce a nonparametric variable selection method which does not rely on any regression model or predictor distribution. The method is based on additive conditional independence (ACI), a newly proposed statistical relation for graphical models. Unlike most existing variable selection methods, which target the mean of the response, the proposed method targets a set of attributes of the response, such as its mean, variance, or entire distribution. In addition, the additive nature of this approach offers nonparametric flexibility without employing high-dimensional kernels. As a result it retains high accuracy for high-dimensional predictors. We establish estimation consistency, convergence rate, and variable-selection consistency of the proposed method. Through simulation comparisons we demonstrate that the proposed method performs better than existing methods when the predictor affects several attributes of the response, and performs competently in the classical setting where the predictors affect the mean only. We apply the new method to a

data set concerning how gene expression levels affect the weight of mice.

A BAYESIAN APPROACH FOR ENVELOPE MODELS

Kshitij Khare[1]*, Subhadip Pal*[2] *and* ♦*Zhihua Su*[1]
[1]University of Florida
[2]Emery University
zhihuasu@stat.ufl.edu

The envelope model is a new paradigm to address estimation and prediction in multivariate analysis. Using sufficient dimension reduction techniques, it has the potential to achieve substantial efficiency gains compared to standard models. This model was first introduced by Cook et al. (2010) for multivariate linear regression, and has since been adapted to many other contexts. However, a Bayesian approach for analyzing envelope models has not yet been investigated in the literature. In this paper, we develop a comprehensive Bayesian framework for estimation and model selection in envelope models in the context of multivariate linear regression. Our framework has the following attractive features. Firstly, we use the matrix Bingham distribution to construct a prior on the orthogonal basis matrix of the envelope subspace. This prior respects the manifold structure of the envelope model, and can directly incorporate prior information about the envelope subspace through the specification of hyperparamaters. This feature has potential applications in the broader Bayesian sufficient dimension reduction area. Secondly, sampling from the resulting posterior distribution can be achieved by using a block Gibbs sampler with standard associated conditionals. This in turn facilitates computationally efficient estimation and model selection. Thirdly, unlike the current frequentist approach, our approach can accommodate situations where the sample size is smaller than the number of responses. Lastly, the Bayesian approach inherently offers comprehensive uncertainty characterization through the posterior distribution. We illustrate the utility of our approach on simulated and real datasets.

Pseudo Estimation for High Dimensional Data

♦*Wenbo Wu*[1] *and Xiangrong Yin*[2]
[1]University of Oregon
[2]University of Kentucky
wenbow@uoregon.edu

Sufficient dimension reduction has achieved great success in recent years. When the sample covariance matrix of the predictors is not invertible, many sufficient dimension reduction methods take an ad hoc ridge regression approach. A question that has been raised for a long while is whether such an estimator is still in the central subspace. In this paper, we propose new concepts of pseudo sufficient dimension reduction and pseudo sufficient variable selection to answer this question. Based on an underlying relationship between the ridge regression and the measurement error regression, we propose a general procedure to obtain pseudo estimates. Using an ensemble idea, our proposed pseudo estimates are better than the traditional estimate or a ridge estimate for highly correlated predictors. Large p small n issue is discussed, while theoretical properties are obtained. Simulation studies and two real data analyses are used to demonstrate the advantage of our methods.

On the second-order inverse regression methods for a general type of elliptical predictors

Wei Luo
Baruch College
wei.luo@baruch.cuny.edu

In sufficient dimension reduction, the second-order inverse regression methods, such as the principal Hessian directions and directional regression, commonly require the predictor to be normally distributed. In this paper, we introduce a type of elliptical distributions called the quadratic variance ellipticity family, which covers and approximates a variety of commonly seen elliptical distributions, with the normal distribution as a special case. When the predictor belongs to this family, we study the properties of the second-order inverse regression methods and adjust them accordingly to preserve the consistency. When the dimension of the predictor is sufficiently large, we also show the consistency of the original methods, which strengthens a previous result in Li & Wang (2007). Simulation study is conducted to illustrate the effectiveness of the adjusted methods.

Session 108: New Approaches in Dimension Reduction for Modern Data Applications

Generalized Mahalanobis Depth in Point Process Data

Shuyi Liu and ♦*Wei Wu*
Florida State University
wwu@stat.fsu.edu

In this talk, we propose to generalize the notion of depth in temporal point process observations. The new depth is defined as product of two probability measurements: 1) the number of event in each process, and 2) the center-outward ranking on the event times conditioned on the number of events. In this study, we extend the Mahalanobis depth to define the center-outward ranking using a multivariate Gaussian model. We propose a bootstrapping method to estimate Gaussian parameters. In the case of Poisson process, the observed events are order statistics, and the parameters can be estimated robustly w.r.t. sample size (tri-diagonal precision matrix). We demonstrate the use of the new depth to rank data from a Poisson process. We also test the new method in classification problems for simulations and real data, and find that it provides more accurate and robust classification result as compared to commonly used likelihood methods.

On the Estimation of Ultra-High Dimensional Semiparametric Gaussian Copula Models

Qing Mai
Florida State University
mai@stat.fsu.edu

The semiparametric Gaussian copula model has wide applications in econometrics, finance and statistics. Recently, many have considered applications of semiparametric Gaussian copula model in several high-dimensional learning problems. In this talk we propose a slightly modified normal score estimator and a new Winsorized estimator for estimating both nonparametric transformation functions and the correlation matrix of the semiparametric Gaussian copula model. Two new concentration inequalities are derived, based on which we show that the normal score estimator and the new Winsorized estimator are consistent when the dimension grows at an exponential rate of the sample size. As demonstration, we apply our theory to two high-dimensional learning problems: semiparametric Gaussian graphical model and semiparametric discriminant analysis.

Supervised Integrative Principal Component Analysis

♦*Gen Li*[1] *and Sungkyu Jung*[2]
[1]Columbia University, Department of Biostatistics
[2]Pittsburgh University, Department of Statistics
gl2521@cumc.columbia.edu

It becomes increasingly common to have data from multiple sources for the same set of subjects in modern health science research. Integrative dimension reduction of multi-source data has the potential to leverage information in heterogeneous sources, and identify dominant patterns in data that facilitates interpretation and subsequent analyses. However, such methods are not well studied, and in particular, no existing method accounts for supervision effects from auxiliary covariates. In this talk, I will introduce a novel statistical framework for integrative dimension reduction of multi-source data with covariate adjustment. The method decomposes the total variation of multi-source data into the joint variation across sources and individual variation specific to each source. The framework is formulated as a hierarchical latent variable model where each latent variable easily incorporates covariate adjustment through a linear model or nonparametric models. We show that the model subsumes many recently developed dimension reduction methods as special cases. A computationally efficient algorithm is devised to fit the model. We apply the methods to two pediatric growth studies, where we discover interesting growth patterns and identify important covariates associated with growth.

A modern optimization perspective on iterative proportional scaling for large tables and count data

♦*Yiyuan She and Shao Tang*
Florida State University
yshe@stat.fsu.edu

In many modern applications people face the challenge of big multiway tables. Statistical inference and estimation of contingency tables heavily rely on the fitting of log affine models. Unfortunately, standard methods such as Newton's algorithm are prohibitive in computation due to the large problem size. Moreover, since the likelihood function does not have bounded Lipschitz gradient, there is no universal stepsize in applying gradient methods, and efficient line searches are lacking in the big data setting. We found that the traditional method of matrix raking or iterative proportional scaling (IPS), though perhaps only of historical interest in contemporary textbooks, is much more scalable. We give new interpretations of IPS from block coordinate descent and majorize-minimization. Our derivations recover several of its important variants and lead to faster algorithms for big count data.

Session 109: Deep Dive on Multiplicity Issues in Clinical Trials.

Use of intermediate endpoint in Phase II/III adaptive designs

♦*Xiaoyun (Nicole) Li and Cong Chen*
Merck& Co.
pkuxiaoyun02@gmail.com

In this presentation, we propose a study design for seamless phase II/III oncology studies, which uses intermediate endpoint at the interim analysis for study modification (i.e., dose selection or population selection) and the overall type I error of the phase II/III adaptive design is controlled at 2.5% (one-sided). Unlike some other designs, we mimic the reality where there is possibility that there may be treatment effect in the intermediate endpoint. We further ensure the overall type I error is controlled in any assumption of the intermediate treatment effect.

Keywords: intermediate endpoint, adaptive design, seamless Phase II/III, survival analysis, oncology clinical trials

Controlling Overall Type I Error with Hierarchical Testing Procedure: something not obvious

Li-an Xu
Bristol-Myers Squibb
li-an.xu@bms.com

Clinical trials often involve multiple hierarchically ordered hypotheses with logical restrictions, for example, multiple endpoints, multiple doses, interim/final analysis. One widely used method is to test these hypotheses using a hierarchical testing procedure to control the overall type I error so that all important claims may be made if they are statistically significant. When a hierarchical testing procedure is used with other complicated testing procedures such as Hochberg testing procedure, the controlling of overall type I error is not obvious. Often in these situations, the overall type I error is not controlled if the hierarchical testing procedure is applied naively. In this talk, I will review three examples arise from some real clinical trials. In these cases, careful consideration needs to be made in order to ensure overall type I error is controlled.

Power and sample size calculation in graphical approaches

♦*Dong Xi, Willi Maurer, Ekkehard Glimm and Frank Bretz*
Novartis
dong.xi@novartis.com

Power and sample size calculation is an essential part of clinical trial design. However, it could become complicated in trials with multiple objectives and there is a need for a flexible and intuitive way to carry out the calculation. Graphical approaches can be applied to common multiple test problems, such as comparing several treatments with a control, assessing the benefit of a new drug for more than one endpoint, and combined non-inferiority and superiority testing. Using graphical approaches, one can easily construct and explore different testing strategies and thus tailor the test procedure to the given study objectives. In this presentation, we illustrate power and sample size calculation to optimize a multiple test procedure for given study objectives. The calculation will be demonstrated using a clinical trial example. The presented methods will be illustrated using the graphical user interface from the gMCP package in R, which is freely available on CRAN.

Session 110: Adaptive Designs in Clinical Trials

Identifying Main Effects in Multi Factor Clinical Trials

♦*Abhishek Bhattacharjee*[1] *and Samuel Wu*[2]
[1]Department of Statistics, University of Florida
[2]Department of Biostatistics, University of Florida
a.bhattacharjee@ufl.edu

Factorial designs have been widely used in clinical research, especially for screening multi-factor interventions in early phases of trials. In this talk new tests for identifying main effects in factorial designs are proposed where the test statistics are weighted combination of simple effects. In the presence of interactions, the weighted tests are shown to be more powerful than traditional tests based on the average of simple effects. Furthermore, unequal sample size allocations are investigated to derive optimal design that achieve adequate power for a predetermined subset of tests for main effects. The computational burden is low because the optimal choice of the sample sizes under the power constraints can be formulated as a simple convex optimization problem. The new methods are illustrated by some real examples.

Interval Estimation in Multi-stage Drop-the-losers Designs

Xiaomin Lu[1], ♦*Ying He*[2] *and Samuel Wu*[1]
[1]University of Florida

[2]Clarkson University
yhe@clarkson.edu

Drop-the-losers designs have been discussed extensively in the past decades, mostly focusing on two-stage models. The designs with more than two stages have recently received increasing attention due to their improved efficiency over the corresponding two-stage designs. In this paper, we consider the problem of estimating and testing the effect of selected treatment under the setting of three-stage drop-the-losers designs. A conservative interval estimator is proposed, which is proved to have at least the specified coverage probability using a stochastic ordering approach. The proposed interval estimator is demonstrated numerically to have narrower interval width but higher coverage rate than the bootstrap method proposed by Bowden and Glimm (Biometrical Journal, vol. 56, pp. 332-349) in most cases. It is also a straightforward derivation from the stochastic ordering result that the family-wise error rate is strongly controlled with the maximum achieved at the global null hypothesis.

Graphical Approach to Multiple Test Procedures in 2×2 Factorial Designs

♦*Xiaomin Lu*[1]*, John Kairalla*[1]*, Hui Zeng*[2] *and Samuel Wu*[1]
[1]University of Florida
[2]Pharmaceutical Product Development (PPD)
xlu2@phhp.ufl.edu

For a study with multiple hypotheses, a multiple test procedure is needed in order to control the familywise error rate. The graphical approach is a novel and easy-to-understand way to express a multiple test procedure by a weighted directed graph, with each node representing an elementary hypothesis. Several commonly used multiple test procedures, such as Bonferroni-Holm procedure, fixed sequence procedure, and fallback procedure, can be viewed as special cases of the graphical approach. We discuss how the graphical approach can be applied to 2x2 factorial designs. Two test procedures using graphical approaches are proposed and compared with two Bonferroni-Holm procedures under various scenarios via numerical studies. The two proposed test procedures performed at least as well as the Bonferroni-Holm procedures in most cases in terms of number of correct rejections and the probability of at least one correct rejection.

Classification of Subtypes of Cancers Using Neural Networks and Gene Expression Data

Lan Gao
The University of Tennessee at Chattanooga
Cuilan-Gao@utc.edu

The major advantage of gene-expression data (measurement of the activity of tens of thousands of genes at once) by NGS technology is that the huge amount of molecular information can be extracted and integrated to find common patterns within a group of samples. Unfortunately, due to the very limited samples of human cancers, it is very challenging to build an effective predictive model directly from human gene expression data.The aim of this study is to develop a method of detecting the subtypes of cancers based on a artificial neural networks (ANN) classifier via analysis of the gene expression data of mouse models. Instead of building predictive model from the gene-expression data of human samples, first we will use gene expression data of mouse to calibrate ANN classifier to recognize cancers, then validate the model and finally apply the ANN classifier to human sample to diagnosis the subtypes of cancer. We will apply the proposed method to both simulated gene expression data and an example data.

Session 111: Shape Analysis in Applications

Non-Euclidean Classification of Medically Imaged Objects via SRNF

Xiaoming Dong
Florida State University
x.dong@stat.fsu.edu

In this talk, a new methodology for shape analysis of parameterized surfaces will be introduced, where the main issue is seeking a metric for shape comparison that is invariant to re-parameterization. We began by defining a general elastic metric on the space of parameterized surfaces, which provides a natural interpretation geometrically and is invariant under the action of re-parameterized group. Moreover, we introduce a novel representation of surfaces termed square root normal fields (SRNFs) and under this representation, a reduced general elastic metric becomes the simple L^2 metric, which will greatly simplify the implementation of our framework. We apply the new methodology to the hippocampi dataset and show the improved performance with proposed methodology.

A Novel Geometric Approach For Semiparametric Density Estimation

♦*Sutanoy Dasgupta*[1]*, Debdeep Pati and Anuj Srivastava*
[1]Florida State University
sutanoy26071991@gmail.com

We discuss the problem of estimating a probability density function given independent samples from the density. We propose a method which starts off with an initial estimate which might be far from the true density and then transform it with warping functions to get a better estimate. We exploit the nice geometric properties of warping functions to get an optimal warping function via maximum likelihood estimation.

An Analytical Approach for Computing Shape Geodesics

♦*Charles Hagwood*[1]*, Javier Bernal*[1] *and Gunay Dogan*[2]
[1]NIST
[2]Theiss Research and NIST
hagwood@nist.gov

In this talk, we will discuss a variational problem that occurs in shape analysis. Shape analysis is used to compare, classify and align D-dimensional objects and images. A distance function on shape space is involved, which is a Riemannian metric on shapes. This distance function is defined by the square root velocity function (SRVF), Srivastava et. al. Some analytical results are provided for this variational problem.

Shape metrology and its applications in Medicine and Forensics

Z.Q. John Lu
National Institute of Standards and Technology
john.lu@nist.gov

Both shape and size are important in biological and medical measurements. I will discuss how shape affects tumor volume measurements and how variability involving 3D volumes using thin-slice X-ray CT images based on semi-assisted CAD measurements should be analyzed using the data from RSNA QIBA studies. If time permit, I will discuss potential applications of shape metrology in forensics including collection of 3d impression evidence.

Session 112: Bias Reduction and Subgroup Identification in Observational Studies

Local Control Strategy: Can Current Indoor Radon Levels Cause Lung Cancer Mortality?

♦*Robert L. Obenchain*[1] *and S. Stanley Young*[2]
[1]Risk Benefit Statistics
[2]Omicsoft
wizbob@att.net

We present a case-study example of Local Control Strategy for analysis of observational data in which the "treatment" variable is not a binary choice. Rather, prevalent indoor Radon levels are continuous measures of exposure. We use an observational data set amassed from published government statistics for 2,881 US counties; it contains average measures of indoor Radon, lung cancer mortality and three X-confounder measures: % elderly (age ¿ 65), % current smokers and % obese. Surprisingly, perhaps, these data show that lung cancer mortality (deaths per 100,000 person-years) tends to be lower in counties where prevalent indoor Radon levels are higher ...with mortality heterogeneity directly predictable from (i.e. moderated by) the three X-confounding factors.

Time Series Smoother

♦*Cheng You*[1]*, Dennis Lin*[1] *and S. Stanley Young*[2]
[1]The Pennsylvania State University
[2]CGStat
czy112@psu.edu

In environmental epidemiology, how the air quality change would affect human health is of great interest. Due to its universality and controversy, the air quality change raised more and more concerns. United States Environmental Protection Agency claims that the small particulates can cause acute death and would enforce cleaner energy production by new protocols and facilities. Although their claim remains to be evaluated with caution, the air quality problem does not only affect long-term human welfare but also poses a major economical challenge to all the countries in the world.

While investigating the air quality and mortality data, we often encounter the dilemma of time series data with inner trend that cannot be accounted by the observed covariates. Inner trend is complicated to determine and separate from abnormal signals. We have found that the current spline smoothing or kernel smoothing methods can produce any result that one would prefer, by varying the smoothing parameter. Hence, we intend to address how to robustly and stably de-trend time series in order to recover the abnormal signals. Three classes of moving window smoothers to de-trend the time series data are being proposed. Their general properties are investigated and demonstrated. For real data analysis, the air quality and mortality data in Los Angeles has been studied, from data collection to statistical analysis. For simulation, synthetic datasets, almost indistinguishable from the real one, are generated to show the performance of different classes of time series smoothers. Finally, good practice and guidelines on how to recover informative signals of time series data with inner trend are provided upon users' preference.

Reliability of a Meta-analysis of Observational Studies

Kumer Das[1]*, Adam Fadhli-Theis*[2]*,* ♦*Allen Heller*[3] *and S. Stanley Young*[4]
[1]Lamar University
[2] Lamar University
[3]Pharma Study Design LLC
[4] CGStat LLC
ahheller2013@gmail.com

It is well known that claims coming from observational studies have failed to replicate in subsequent observational studies or when tested rigorously in randomized clinical trials. This problem may be related to the common practice of testing multiple questions and/or exploring multiple statistical models without adjusting the analysis for multiple testing. It is popular to gather observational studies and combine them into a "meta-analysis" in an effort to determine whether summary statistics from each of the base papers, when combined, yield a more convincing conclusion than the individual studies. There is a need to examine the reliability of meta-analyses of observational studies. Our idea is to select such a meta-analysis, gather the base papers selected for the meta-analysis, and estimate the size of the search space used in each base paper. We count the number of outcomes at issue, the number of predictors at issue and the number of adjusting covariates enumerated in each paper. If the search space of a base paper is large with no evidence of statistical adjustment for multiple testing, then the summary statistics from that paper could be biased. Unbiasedness is a requirement for a sound meta-analysis, and therefore the resulting meta-analysis may not be reliable. We present an evaluation of the base papers used in a meta-analysis of conditions triggering myocardial infarction. An evaluation of the reliability of the base papers speaks to the reliability of the meta-analysis.

Session 113: Advance in Statistical Method on Complex Data and Applications in Statistical Genomics

Joint Estimation of Multiple Dependent Gaussian Graphical Models with Applications to Mouse Genomics

♦*Yuying Xie*[1]*, Yufeng Liu*[2] *and William Valdar*[2]
[1]Michgan State University
[2]University of North Carolina at Chapel Hill
xyy@stt.msu.edu

Gaussian graphical models are widely used to represent conditional dependence among random variables. In this paper we propose a novel estimator for data arising from a group of Gaussian graphical models that are themselves dependent. A motivating example is that of modelling gene expression collected on multiple tissues from the same individual: a multivariate outcome that is affected by dependencies defined at the level of both the tissue and the whole body. Existing 20 methods that assume independence among graphs are not applicable in this setting. To estimate multiple dependent graphs, we decompose the problem into two graphical layers: the systemic layer, which is the network affecting all outcomes and thereby inducing cross-graph dependency, and the category-specific layer, which represents the graph-specific variation. We propose a new graphical EM technique that estimates the two layers jointly, establish the estimation consistency 25 and selection sparsistency of the proposed estimator, and confirm by simulation that the EM method is superior to a simple one-step method. Lastly, we apply our graphical EM technique to mouse genomics data and obtain biologically plausible results.

Identification of Pairwise Informative Features for Clustering Data with Growing Dimension

♦*Lulu Wang, Wen Zhou and Jennifer Hoeting*
Colorado State University
wanglulu@stat.colostate.edu

Identifying important features for separating unlabeled observations into homogeneous groups plays a critical role in dimension reduction and modeling data with complex structures. This problem is

directly related to selecting informative variables in cluster analysis, where a small fraction of features is identified for separating observed p-dimensional feature vectors into K possible classes. Utilizing the framework of model-based clustering, we introduce a PAirwise Reciprocal fuSE (PARSE) procedure based on a new class of penalization functions that imposes infinite penalties on features with small differences across clusters, which effectively avoids selecting overly dense number of features for separating observations in cluster analysis. We establish the consistency of the proposed procedure for identifying informative features for cluster analysis. The PARSE procedure is shown to enjoy certain optimality properties as well. We developed a backward selection algorithm, in conjunction with the EM algorithm, to implement PARSE. Simulation studies show that PARSE has competitive performance compared to other popular model-based clustering methods. PARSE is shown to select a sparse set of features and to produce accurate clustering results. We apply PARSE to a microarray experiment on human asthma and discuss the biological implications of the results.

Detect chromatin interaction from multiple Hi-C datasets by hierarchical hidden Markov random model

♦*Zheng Xu*[1]*, Ming Hu*[2] *and Yun Li*[1]
[1]University of North Carolina at Chapel Hill
[2]New York University School of Medicine
zheng.xu.isu@gmail.com

[Background] The constantly accumulating Hi-C data provide rich information for the detection of long range chromatin interactions (or peaks) across multiple experimental conditions and cell differentiation stages. However, statistical models and computational tools for characterizing the dynamics of chromatin interaction are still under development. Limited sequencing depth within each individual Hi-C experiment, as well as heterogeneity among different Hi-C experiments, pose great challenges for the downstream Hi-C data analysis.

[Method] We propose a peak caller based on a hierarchical hidden Markov random field (HHMRF) model to detect long range chromatin interactions from multiple Hi-C datasets. In addition to model the spatial dependency of chromatin interaction in the local neighborhood (Xu et al., Bioinformatics, 2016), HHMRF is also able to model the dependency across multiple Hi-C datasets, leading to further improved statistical power.

[Results] We conducted comprehensive simulation studies, and showed that HHMRF model outperforms other competing methods which ignore the dependency structure and call peaks separately in each individual Hi-C dataset. Next, we analyzed a real Hi-C dataset on human embryonic stem cells (h1ESC) and four H1 derived cells (Xie et al., Cell, 2013), and found that the cell-type-specific peaks identified by HHMRF show higher overlap with cell-type-specific epigenetic features and cell-type-specific gene expression, compared to those identified by competing methods. HHMRF model has the potential to unveil the structural basis of cell-type-specific transcription regulation mechanism.

Session 114: Statistical Issues in EHR Data

Comparative effectiveness of dynamic treatment strategies using electronic health records and the pa

Miguel Hernán
Harvard TH Chan School of Public Health
mhernan@hsph.harvard.edu

Comparative effectiveness of dynamic treatment strategies using electronic health records and the parametric g-formula

Causal questions about the comparative effectiveness and safety of health-related interventions are becoming increasingly complex. Decision makers are now often interested in the comparison of interventions that are sustained over time and that may be personalized according to the individuals' time-evolving characteristics. These dynamic treatment strategies cannot be adequately studied by using conventional analytic methods that were designed to compare "treatment"vs. "no treatment" The parametric g-formula was developed by Robins in 1986 with the explicit goal of comparing generalized treatment strategies sustained over time. However, despite its theoretical superiority over conventional methods, the parametric g-formula was rarely used for the next 25 years. Rather, the development of causal inference methods for longitudinal data with time-varying treatments focused on semiparametric approaches. In recent years, interest in the parametric g-formula is growing and the number of its applications increasing. This talk will review the parametric g-formula, the conditions for its applicability, its practical advantages and disadvantages compared with semiparametric methods, and several real world implementations for comparative effectiveness research.

Accounting for Error and Misclassification in Time to Event Analyses Using EHR-derived Endpoints

♦*Rebecca Hubbard*[1]*, Weiwei Zhu*[2]*, Le Wang*[1] *and Jessica Chubak*[2]
[1]University of Pennsylvania
[2]Group Health Research Institute
rhubb@upenn.edu

Estimates of the relationship between an outcome and an exposure are biased by imperfect ascertainment of the outcome of interest. In studies using data derived from electronic health records (EHRs), misclassification of outcomes is common and is often related to patient characteristics. For instance, patients with greater co-morbid disease burden may use the healthcare system more frequently making it more likely that the EHR will contain a record of their diagnosis, possibly resulting in poorer outcome classification for healthier patients. This is particularly problematic in studies of time to event outcomes in which both the occurrence of an event and the timing of the event may be captured with error in the EHR. Misclassification-adjusted estimators in the context of time to event outcomes are available using discrete time proportional hazards models but may be biased if operating characteristics of the EHR-derived endpoint vary across exposure categories. Motivated by an algorithm for identifying breast cancer recurrence using EHR data, we investigated the implications of using an imperfectly assessed outcome in time to event analyses. We used simulation studies to demonstrate the magnitude of bias induced by failure to account for error in the status and timing of recurrence and compared alternative methods for correcting this bias. We conclude with general guidance on accounting for outcome misclassification in time to event studies using EHR data.

Phenotype validation in Electronic Health Record based genetic association studies

Lu Wang, Jason Moore, Scott Damrauer and ♦*Jinbo Chen*
University of Pennsylvania
jinboche@mail.med.upenn.edu

The linkage between Electronic Health Records (EHRs) and genotype data makes it plausible to study genetic susceptibility of a wide range of disease phenotypes. While cost effective, a major challenge is that EHR-derived phenotype data is subjected to misclassification. It is well known that misclassification can lead to biased odds ratio (OR) estimation and decreased power for testing

association. The bias can be corrected when the phenotype data for a subset of patients can be validated. In this work, using a genotype-stratified case-control sampling strategy to select a subset of patients for validation, we derive the closed-form maximum likelihood estimator for the OR parameters and a two degree of freedom score statistic for testing genetic association. We assess the extent of power improvement through incorporation of validated phenotype data. Compared with standard case-control sampling, our genotype-stratified sampling strategy corrects bias in the OR estimates when the MAF is small, generally leads to smaller variance for the estimated ORs, and results in higher power for testing associations.

Robust methods for association studies in EHR data

Jing Huang, Rui Duan and ♦*Yong Chen*
University of Pennsylvania
ychen123@mail.med.upenn.edu

Over the last decade, Electronic Health Records (EHR) systems have been increasingly implemented at US hospitals. Despite their great potential, the complex and uneven nature of clinical documentation and data quality brings additional challenges for analyzing EHR data. A critical challenge is the information bias due to the measurement errors in outcome and covariates. Based on the characteristics of the Electronic Medical Records and Genomics (eMERGE) Network, we design and conduct simulation studies to quantified the loss of power due to misclassifications in case ascertainment and measurement errors in exposure status extraction. We also describe alternative methods for bias correction in EHR based association studies.

Session 115: Recent Advances in Integrative Analysis of Omics Data

Exploratory Factorization of Multi-Source Data

♦*Eric Lock and Michael OConnell*
University of Minnesota, Division of Biostatistics
elock@umn.edu

In biomedical research a growing number of platforms and technologies are used to measure diverse but related information. For a given study, the resulting data often constitute a collection of data arrays in which some dimensions are shared and some are not. For example, the NIEHS/NCATS/UNC 1000 Genomes toxicity screening project includes an array of toxicity measures for 884 genetically distinct human cell lines after exposure to 156 different chemicals (chemicals X cell lines), RNA-seq data giving gene expression profiles for each cell line (genes X cell lines), and an array with structural attributes for each chemical (chemicals X attributes). Factorization methods like Principal Components Analysis (PCA) are an invaluable tool for dimension reduction and exploratory analysis of a single data source, and the extension of such methods to the multi-source context is an active area of research that is relevant to the integrative analysis of omics data. The joint and individual variation explained (JIVE) method, and other approaches, were recently developed to identify shared and source-specific variation among multi-source data in which one dimension is common to all sources (e.g., multiple technologies measured on the same sample set); we compare these approaches and discuss new directions.

A Bayesian approach for the integrative analysis of omics data: A kidney cancer case study

♦*Thierry Chekouo*[1]*, Francesco Stingo*[1]*, James Doecke*[2] *and Kim-Anh Do*[1]
[1]U.T MD Anderson Cancer Center
[2]CSIRO Australia
tchekouo@mdanderson.org

Integration of genomic data from multiple platforms has the capability to increase precision, accuracy, and statistical power in the identification of prognostic biomarkers. A fundamental problem faced in many multi-platform studies is unbalanced sample sizes due to the inability to obtain measurements from all the platforms for all the patients in the study. We have developed a novel Bayesian approach that integrates multi-regression models to identify a small set of biomarkers that can accurately predict time-to-event outcomes. This method fully exploits the amount of available information across platforms and does not exclude any of the subjects from the analysis. Moreover, interactions between platforms can be incorporated through prior distributions to inform the selection of biomarkers and additionally improve biomarker selection accuracy. Through simulations, we demonstrate the utility of our method and compare its performance to that of methods that do not borrow information across regression models. Motivated by The Cancer Genome Atlas kidney renal cell carcinoma dataset, our methodology provides novel insights missed by non-integrative models.

Integrative analysis of multiple omics data with biological information.

♦*Sandra Safo, Shuzhao Li and Qi Long*
Emory University
ssafo@emory.edu

It is of increasing importance to integrate different types of omics data to examine biological mechanisms in disease etiology. Canonical Correlation Analysis (CCA) provides an attractive tool to investigate such mechanisms. Traditional CCA methods use all available variables and several sparse CCA methods have been proposed to constrain the size of the CCA vectors in order to yield interpretable results. It is well-known that variables in omics data are functionally structured in networks or pathways. We develop statistical methods for CCA that incorporate biological/structural information via undirected graphical networks. Our work is motivated by an in-vitro mouse toxicology study on the neurotoxicity of the combination of the herbicide paraquat and fungicide maneb in relation to Parkinson's disease (PD). We are interested in assessing association between transcriptomic and metabolomic data that may shed light on the etiology of PD. Our proposed methods use prior network structural information among genes and among metabolites to guide selection of relevant genes and metabolites in sparse CCA, providing insight on the molecular underpinning of PD. Our simulations demonstrate that the proposed CCA methods outperform several existing sparse CCA methods in selecting relevant genes and metabolites and when structural information is informative and are robust to mis-specified structural information. Our analysis of the PD toxicology data reveals that a number of genes and metabolic pathways including some known to be associated with PD are enriched in the subset of genes and metabolites selected by our proposed approach. With time permitting, I will also discuss ongoing work of integrating multiple big biomedical data types from different independent studies or subpopulations while accounting for data heterogeneity using factor analysis.

A Full Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-type Omics data

Qianxing Mo
Baylor College of Medicine
qmo@bcm.edu

Identification of clinically relevant tumor subtypes and genomic signatures is an important task in cancer translational research for pre-

cision medicine. Large-scale genomic profiling studies such as The Cancer Genome Atlas (TCGA) project have generated vast amounts of genomic, transcriptomic, epigenomic and proteomic data. While these studies have provided researchers great resources to discover clinically relevant tumor subtypes and driver genetic and genomic alterations, there are still lack of computationally efficient methods and tools for integrative analysis of these high-dimensional omics data. Therefore, we aim to develop a full Bayesian latent variable framework that can jointly model continuous and discrete omics data for identification of tumor subtype and relevant genomic feature. Specifically, this method uses a few latent variables to capture the inherent structure of multiple genomic data sets to achieve joint dimension reduction. As a result, the tumor samples can be clustered in the latent variable space. Simultaneously, relevant genomic features that contribute to clustering tumor samples are identified through Bayesian variable selection. By analyzing TCGA and simulated data sets, we demonstrate the excellent performance of the proposed method in revealing clinically meaningful tumor subtypes and their joint genomic patterns.

Session 116: Survival Analysis

A frailty model for recurrent events of the same type during alternating restraint and non-restraint

♦*Xiaoqi Li*[1]*, Yong Chen*[2] *and Ruosha Li*[3]
[1]Baylor College of Medicine
[2]University of Pennsylvania
[3]University of Texas School of Public Health
xiaoqil@bcm.edu

We consider recurrent events of the same type during alternating restraint and non-restraint time periods. This research is motivated by a study on juvenile recidivism, where the probationers were followed during alternating placement periods and free-time periods. During the placement periods,the probationers were under a restricted environment with direct observation by the probation officers. The probationers were released to home and not under direct observation during the free-time periods. Although re-offenses can occur during both types of periods, the intensities of the re-offenses are very different. Thus, these two types of periods should be modeled differently. In this paper, we propose a joint modeling framework that explicitly accounts for the different time periods, as well as the dependence between the two different intensities. The estimation procedure is implemented in SAS and is highly accessible to practical investigators. We evaluate the proposed method through simulation studies under both correctly specified and mis-specified models, and demonstrate the feasibility of the proposed method by applying it to the juvenile recidivism dataset. The proposed method is also applicable to studies in medicine and health care, such as tumor metastases during chemotherapy and chemo-free periods.

Variable Selection for Interval-Censored Survival Data Under the Proportional Hazards Model

♦*Tyler Cook*[1] *and Jianguo Sun*[2]
[1]University of Central Oklahoma
[2]University of Missouri
tcook14@uco.edu

When conducting a survival analysis, it is often of interest to identify a subset of variables significantly related with the event time under study. For example, a researcher measuring gene expression levels wants to determine which genes can predict tumor development time in cancer patients. This problem of variable selection has received a considerable amount of recent attention. Several methods using penalized likelihood procedures have been proposed to perform parameter estimation and variable selection simultaneously. A number of these techniques have been extended to the case of right-censored survival data, but little has been done in the context of interval-censoring. Here, we propose an approach for variable selection of interval-censored survival data under the proportional hazards model. This method uses imputation to create a new dataset of imputed exact failure times and right-censored observations. Variable selection can then be performed on the imputed dataset using any of the popular variable selection techniques created for right-censored data. Comprehensive simulation studies illustrate the effectiveness of this new approach. Moreover, this method is attractive due to how easy it is to implement, since it can take advantage of existing software for variable selection of right-censored data.

The Spike-and-Slab lasso Cox Model for Survival Prediction and Associated Genes Detection

♦*Zaixiang Tang*[1] *and Nengjun Yi*[2]
[1]Department of biostatistics, Soochow University
[2]Department of Biostatistics,UAB
tangzx81@gmail.com

Large-scale molecular profiling data generated by high-throughput technologies have offered extraordinary opportunities to improve survival prediction of cancers and other diseases and to detect disease associated genes. However, there are considerable challenges in analyzing large-scale molecular data, including large number of correlated predictors, limited number of samples, and small effect of each predictor. We propose new Bayesian hierarchical Cox proportional hazards models, called the spike-and-slab lasso Cox, for predicting survival outcomes and detecting associated genes. The proposed Cox models employ an indicator variable to specify a spike-and-slab mixture double-exponential prior with two shrinkage values for modeling irrelevant and large coefficients, respectively, and thus can adaptively shrink coefficients. We develop an efficient algorithm to fit the proposed models by incorporating EM steps (Expectation-Maximization) into the extremely fast cyclic coordinate descent algorithm. The proposed approach integrates two popular methods, i.e., penalized lasso and Bayesian spike-and-slab variable selection, into one unifying framework, and thus combines the nice features of the two popular methods while diminishing their shortcomings. The performance of the proposed method is assessed via extensive simulations and compared with the commonly used lasso Cox. The results show that the proposed approach can provide not only better prediction, but also more accurate estimates of the parameters. We demonstrate the proposed procedure on two cancer data sets with censored survival outcomes and thousands of molecular features: breast cancer and myelodysplastic syndromes. Our analyses suggest that the proposed procedure can generate powerful prognostic models for predicting cancer survival and detect survival associated genes. The methods have been implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

Variable Selection for Mixture and Promotion Time Cure Rate Models

♦*Abdullah Masud and Zhangsheng Yu*
Indiana University
amasud@iu.edu

Failure time data with cure fraction are commonly encountered in clinical studies. Variable selection methods for cure rate models are yet to be further studied. We propose to use the adaptive LASSO procedure for variable selection in both mixture cure rate

models and promotion time cure model. Both non parametric and parametric baseline hazard estimation methods are considered. An expectation-maximization (EM) algorithm is developed for estimation. Performance of selection using adaptive LASSO and LASSO procedures is evaluated using simulation studies. We finally apply the proposed method to evaluate risk factors on wheezing for a pediatric study.

Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach

♦*Xinyan Zhang*[1]*, Yan Li*[1]*, Tomi Akinyemiju*[1]*, Akinyemi I. Ojesina*[1]*, Phillip Buckhaults*[2]*, Bo Xu*[3] *and Nengjun Yi*[1]
[1]University of Alabama at Birmingham
[2]The University of South Carolina
[3]Southern Research Institute
abbyyan3@gmail.com

Heterogeneity of prognosis and prediction based on clinical factors or molecular signatures in cancer treatment across patients has been a persisted problem over decades. One of the main shortcomings of the previous studies is the failure to incorporate pathway-based genetic structure of cancer into the predictive model. To address this problem, we propose a two-stage procedure to incorporate pathway information into the predictive modeling using large-scale gene expression data. In the first stage, we fit all predictors within each pathway using penalized Cox model and Bayesian hierarchical Cox model. In the second stage, we combine the cross-validated prognostic scores of all pathways obtained in the first stage as new predictors to build an integrated prognostic model for prediction. We used the proposed method to analyze a breast cancer data set from The Cancer Genome Atlas (TCGA) project for predicting overall survival using gene expression profiling. The results show that the proposed approach not only improves survival prediction compared with the alternative analysis that ignores the pathway information, but also identifies significant biological pathways.

Application of Heavy-Tailed Probability Distributions on Financial Markets

Xing Yang
Jackson State University
xing.yang@jsums.edu

In this paper a comprehensive statistical simulation based on heavy-tailed distributions is performed on the US stock markets. The method modifies the Geometric Brownian Motion assumption of the Black-Scholes model with more realistic heavy-tailed distributions to reflect the higher probability occurrences of financial market collapses in the real world. The Standard & Poor's 500 stock index and other indices are used to examine the results of the simulation which are compared with that from the original Black-Scholes model.

Session 117: Advances in Hypothesis Testing

Methods for f lack of fit test without replicates

Tyler George
Central Michigan University
georg2t@cmich.edu

The lack of fit test is one of the most important diagnostics for assessing adequacies of linear regression models and the test has been commonly included in most regression textbooks. However, the test requires replicates on predictor variables, or in other words, repeated observations on predictor variables. Considering that many realistic data sets do not contain repeated predictor values; we need methods that do not require replicates. This made the lack of fit test well known but often unusable. There are already proposed, evaluated and published tests, Rainbow Test ([U]), Burn and Ryan ([BR]), and XLOF procedure ([M]), not being used that are more versatile, and can be used on any data set. Even with these methods, we find many insufficiencies with particular data structures. We aim to make these methods common practice; causing modern textbooks to include them and trigger future research. An introduction on each of these methods not requiring replicates is given. Recent results by Wang and Conerly ([WC],[WC2]) and Lawrance ([L]) show Minitab's XLOF procedure has the highest power in the majority of the data structures it studied. We pro- gram the XLOF in a new R program (verifying our programs validity) in an attempt to make these tests more widely available. It's then shown with a variety of examples that these current methods are not sufficient for some data structures. Finally we give recommended remedies; we added an option for flexibility in data grouping to improve the effectiveness in detecting lack of fit.

Note: this work has been done in collaboration with Daniel X. Wang.

Generalized Inference for the Reliability of of Stress-Strength Model for the Inverse Exponentials

Sumith Gunasekera
The University of Tennessee at Chattanooga
Sumith-Gunasekera@utc.edu

In this paper we consider the reliability $R_{s,k}$ = Pr(at least s of the $(X_1, X_2, ..., X_k)$ exceed Y) of an s-out-of-k : G system having several independent components with strengths $X_1, X_2, ..., X_k$ subjected to a common stress Y when X and Y are independent two-parameter general inverted exponentiated random variables. We are interested in inferences of the reliability of the multicomponent stress-strength s-out-of-k : G (or s-out-of-k : F) system. For this problem, there are no exact or approximate test and confidence intervals are available in the literature. Using the generalized p-value approach (also referred to as the generalized variable method), an exact test and generalized point and interval estimates are given. A simulation study is given to illustrate the proposed procedure. Based on the type I error rates, adjusted and unadjusted powers, coverage probabilities and expected lengths, and biases show that the generalized procedure outperforms the classical counterpart.

A New Approach to Multiple Testing of Grouped Hypotheses

Yanping Liu, Sanat Sarkar and ♦*Zhigen Zhao*
Temple University
zhaozhg@temple.edu

A new approach to multiple testing of grouped hypotheses controlling false discoveries is proposed. It decomposes a posterior measure of false discoveries across all hypotheses into within- and between-group components allowing a portion of the overall FDR level to be used to maintain control over within-group false discoveries. Numerical calculations performed under certain model assumption for the hidden states of the within-group hypotheses show its superior performance over its competitors that ignore the group structure, especially when only a few of the groups contain the signals, as expected in many modern applications. We offer data-driven version of our proposed method under our model by estimating the parameters using EM algorithms and provide simulation evidence of its favorable performance relative to these competitors. Real data applications have also produced encouraging results for the proposed approach.

Non-zero tests on heterogeneity variance in the quantitative syn-

thesis of cancer detection biomarker

Hanfei Xu and ♦*Kepher Makambi*
Georgetown University
khm33@georgetown.edu

It has recently been established that microRNA-155 is an important biomarker in the diagnosis and therapy of various cancers. Studies on the association between MicroRNA-155 and tumorigenesis are inconsistent. Therefore, the extent of heterogeneity in these studies is of great importance to cancer researchers. Unexplained heterogeneity can be evaluated through the heterogeneity variance, a parameter that can take on nonnegative values, especially when there is no covariate information. The non-zero hypothesis on the heterogeneity variance is a long outstanding problem in the one-way random effects ANOVA model that has found relevance in meta-analysis. We contribute to this problem by focusing on homoscedastic and heteroscedastic cases in balanced and unbalanced samples. Using simulations, we compare the performance of extended test procedures with respect to the actual attained type I error rate and power. The tests are shown to attain acceptable significance levels and power. A meta-analytic application of the tests is presented involving heterogeneity questions concerning microRNA-155 as a biomarker for cancer detection.

A Group of New F-Tests for Multiple Mean Comparisons and Their Applications in Medical Research

Jiajuan Liang
University of New Haven, U.S.A.
jliang@newhaven.edu

The problem of high-dimensional multiple mean comparisons belongs to multivariate analysis of variance (MANOVA for simplicity). It is well-known that the classical Wilks Lambda-statistic is for this purpose. A noticeable drawback of Wilks Lambda-statistic is its dependence on the sample covariance matrix that tends to singularity or becomes non-invertible when the total sample size is close to or less than the dimension of sample data. As a result, Wilks Lambda-statistic is losing power when the data dimension is increasing with a relatively stable sample size, and it finally becomes inapplicable if the data dimension is greater than the sample size. To overcome this weakness of Wilks Lambda-statistic, we develop a group of F-tests for high-dimensional multiple mean comparisons by employing the theory of spherical distributions and the idea of principal components. The new F-tests are applicable for both high and low dimensions without any restriction on the sample size. This makes it possible to solve one-way MANOVA problems under any sample sizes. Monte Carlo studies show that the new F-tests are comparable to Wilks Lambda-statistic in the case of a large sample size, but they outperform Wilks Lambda-statistic in the case of high dimension with a small sample size. Practical medical datasets are analyzed by the new F-tests to illustrate their potential applications in medical research.

Testing the Presence of Significant Covariates Through Conditional Marginal Regression

♦*Yanlin Tang, Huixia Wang and Emre Barut*
The George Washington University
tangyl@gwu.edu

In many statistical applications, researchers have *a priori* information on relative importance of the predictors; this information is then used to screen out a majority of the covariates. In these approaches, an important question is whether any of the discarded covariates have predictive powers when the "most relevant" predictors are included in the model. We consider the problem of testing whether none of the discarded covariates is significant conditional on some pre-chosen covariates. We propose a test statistic and show that the proposed statistic has a non-standard asymptotic distribution that can be estimated through bootstrap, giving rise to the conditional adaptive resampling test. We prove the consistency of the test procedure under very general assumptions. Moreover, we illustrate how the suggested test can be used as a stopping rule in forward regression. We show through simulation that the proposed test provides adequate control of family-wise error rate with competitive power even for high-dimensional cases even though the asymptotic theory used to calibrate the test assumes a fixed dimension, and the proposed forward regression has appealing advantage when covariates are correlated. Finally we demonstrate the effectiveness of the proposed approach through an analysis of an eQTL dataset.

Session 118: Statistical Analysis of Complex Data I

Spatial Data Fusion for large non-Gaussian Remote Sensing Datasets

♦*Hongxiang Shi and Emily Kang*
University of Cincinnati
shihn@mail.uc.edu

Remote sensing data are playing a vital role in understanding the pattern of the Earth's geophysical processes. In this paper, we proposed a spatial data fusion method that is able to take advantage of two (or potentially more) large remote sensing datasets with the exponential family of distributions. We take an empirical Hierarchical modeling (EHM) approach where any unknown parameters are estimated by Maximum Likelihood estimation via an efficient EM algorithm. Then through a MCMC algorithm the prediction is obtained by generating samples from the empirical predictive distribution where the unknown parameters are substituted by the estimates from the EM algorithm. Finally, the performance of our proposed method are investigated through simulation studies and real datasets. It shows that our data fusion method has capabilities to borrow strength across two complementary datasets and thus improving predictions reciprocally.

INCORPORATING GEOSTROPHIC WIND INFORMATION FOR IMPROVED SPACE-TIME SHORT-TERM WIND SPEED FORECASTING

♦*Xinxin Zhu*[1]*, Kenneth Bowman and Marc Genton*
[1]Merial
xxzhu2008@gmail.com

Accurate short-term wind speed forecasting is needed for the rapid development and efficient operation of wind energy resources. This is, however, a very challenging problem. Although on the large scale, the wind speed is related to atmospheric pressure, temperature, and other meteorological variables, no improvement in forecasting accuracy was found by incorporating air pressure and temperature directly into an advanced space-time statistical forecasting model, the trigonometric direction diurnal (TDD) model. This paper proposes to incorporate the geostrophic wind as a new predictor in the TDD model. The geostrophic wind captures the physical relationship between wind and pressure through the observed approximate balance between the pressure gradient force and the Coriolis acceleration due to the Earth's rotation. Based on our numerical experiments with data from West Texas, our new method produces more accurate forecasts than does the TDD model using air pressure and temperature for 1- to 6-hour-ahead forecasts based on three different evaluation criteria. Furthermore, forecasting errors can be further reduced by using moving average hourly wind speeds to fit the diurnal pattern. For example, our new method obtains between

13.9% and 22.4% overall mean absolute error reduction relative to persistence in 2-hour-ahead forecasts, and between 5.3% and 8.2% reduction relative to the best previous space-time methods in this setting

LESA: Longitudinal Elastic Shape Analysis of Brain Subcortical Structures

♦*Zhengwu Zhang*[1]*, Anuj Srivastava*[2]*, Jingwen Zhang*[3]*, Martin Styner*[3]*, Weili Lin*[3] *and Zhu Hongtu*[3]
[1]SAMSI
[2]Florida State University
[3]The University of North Carolina at Chapel Hill
zhengwustat@gmail.com

Longitudinal neuroimaging data plays a critical role in accurately mapping the neural developmental profile of major neuropsychiatric and neurodegenerative disorders, and in characterizing normal brain development. The aim of this paper is to develop a longitudinal elastic shape analysis (LESA) framework to accurately delineate the developmental trajectories of brain subcortical regions. This framework uniquely integrates ideas from an elastic shape representation for analyzing static cortical surfaces, a principal component analysis (PCA) model for capturing (temporal and population) variability, and a semiparametric mixed effect model for temporal evolution, all into a single framework. The two key novelties of LESA are: (i) it can efficiently capture complex subcortical structures using small number of basis, and (ii) it can accurately predict spatio-temporal changes in cortical shapes using statistical models.We illustrate LESA using longitudinal lateral ventricle surface data from a longitudinal study of early brain development.

Model based heritability scores for high-throughput sequencing data

Pratyaydipta Rudra[1]*,* ♦*W. Jenny Shi*[2]*, Brian Vestal*[1]*, Pamela Russel*[1]*, Laura Saba*[3] *and Katerina Kechris*[1]
[1]University of Colorado School of Public Health
[2]University of Colorado School of Medicine
[3]University of Colorado Skaggs School
wen.2.shi@ucdenver.edu

Heritability of a trait (either phenotypic or genetic) measures the proportion of variance that is attributable to genotypic variance. It is an important concept in breeding and genetics. Little work has been done for high-throughput sequenced genetic traits. We propose a statistical framework and different strategies to compute and test a heritability score for such data based on linear and generalized linear mixed effects models. The approaches have been compared for three types of simulations to test model sensitivity and performance. Our analyses show that the negative binomial mixed model (NBMM), compound Poisson mixed model (CPMM), and the variance stabilizing transformed linear mixed model (VST) well outperform the voom-transformed linear mixed model (VOOM). NBMM and VST appear to be more robust than CPMM, while NBMM is the most computational expensive. In addition, we apply the methods to the microRNA sequencing data from a recombinant inbred mouse panel. We show that the miRNA expression can be a highly heritable molecular trait in mouse.

Robust Functional Linear Models

♦*Melody Denhere*[1]*, Nedret Billor*[2] *and Huybrechts Bindele*[3]
[1]University of Mary Washington
[2]Auburn University
[3]University of South Alabama
mdenhere@umw.edu

The field of functional data analysis (FDA) is an increasingly active field of study in statistics. This has been attributed in part to the growing interest in the analysis of 'big data' and also the need to have robust techniques that can deal with data that is in the form of images. In this presentation, we discuss different estimation methods for functional regression models in the presence of outliers. The functional covariates and functional parameters of the models are approximated in a finite dimensional space generated by an appropriate basis. This approach reduces the functional model to a standard multiple model with highly collinear covariates and potentially high dimensionality issues. The proposed estimators tackles these issues and also minimizes the effect of functional outliers. Results from a simulation studies and real world examples are also presented to illustrate the performance of each proposed estimator.

Session 119: Nonparametric and Semiparametric Methods

Covariate Adjusted Cross Ratio Estimation

♦*Ran Liao*[1]*, Tianle Hu*[2] *and Sujuan Gao*[1]
[1]Indiana University
[2]Eli Lilly
ranliao@iu.edu

Cross ratio is formed as the ratio of two conditional hazard rates for one event given the other event, which is a local dependent measurement and also has inherited the nice interpretation of hazard rate. In this piece of work, we extended parametric estimation approach based on Clayton copula and Shih and Louis (1995)'s two stage semi-parametric estimation approach for constant cross ratio into covariate dependent cross ratio with multiplicative covariate effect set up. A comparison simulation study Hu et al. (2011)'s non-parametric estimator is the most robust approach compare to parametric estimator from copula approach and semi-parametric estimator from Shih and Louis (1995)'s two stage approach. To illustrate three estimation methodologies, we analyzed data from Indianapolis-Ibadan Dementia Project (IIDP) to investigate the gender effect between cardiac artery disease event and depression.

Predicting the Timing of the Final Event in a Clinical Trial using the Bayesian Bootstrap and beyond

♦*Marc Sobel*[1] *and Ibrahim Turkoz*[2]
[1]Temple University
[2]Janssen Research and Development
marc.sobel@temple.edu

Variable length time to event clinical trials are based on observing predetermined total numbers of events. Many of these trials involve following patients over a long span of time. The primary events of interest are e.g., death, relapse, adverse drug reaction, or new disease development. Research teams need to determine the conditions under which studies can be finalized in order to put together resources necessary to the dissemination of study results. We propose strategies for determining the timing of the final events in blinded settings. We employ Bayesian Bootstrap and nonparametric Bayesian survival models for this purpose; these are applied to a clinical trial example. Bayesian bootstrap and nonparametric Bayesian survival models are shown to be superior to their parametric counterparts; these counterparts (i) ignore issues involving staggered entry; (ii) do not take account of the evolution of hazard rates; and (iii) fail to provide model flexibility. Nonparametric Bayesian survival models and the Bayesian bootstrap address these concerns and are shown to be highly accurate predictors.

Modeling Hourly Electricity Demand Using Spline-Based Nonparametric Transfer Function Models

Jun Liu
Georgia Southern University
junliu6@gmail.com

In this paper a semi-parametric approach is developed to model nonlinear relationships in time series data. The nonlinearity is modeled using regression splines, which are computationally very efficient in comparison with many other nonparametric methods. This is especially relevant in short-term forecasting. With an explicit functional form, the results obtained with regression splines are also more interpretable. The serial correlation contained in the random error is captured using an ARMA model. The estimation procedure is developed, and the selection of smoothing parameters is discussed. In this paper this approach is used to forecast hourly electricity demand in a large residential area. The model considers the highly nonlinear effect of temperature, combined with those of time-of-day and type-of-day, and the seasonal correlation is modeled using an ARIMA model. Forecasting performance is evaluated by post-sample forecasting and comparative results are presented.

A Differences in Differences Approach for Bias Reduction in Semiparametric Models

♦*Chan Shen*[1] *and Roger Klein*[2]
[1]UT MD Anderson Cancer Center
[2]Rutgers University
cshen@mdanderson.org

There is a wide class of semiparametric models that include a nonparametric component of arbitrary dimension and a finite-dimension parameter vector. Reducing the order of the bias for the nonparametric component is central to both the theoretical and empirical properties of the semiparametric estimator. In this paper, we propose a "differences-in-differences" approach to reduce the bias to any desired order using regular kernels. We show that this approach results in estimators that have both desirable finite sample and asymptotic properties. A notable feature of our approach is that we use optimal windows instead of under-smoothing to reduce the bias to any desired order.

For every model in the class studied here, we provide theoretical results on the windows that ensure asymptotic normality. To examine the performance in finite samples, we ran Monte Carlo experiments for several models in which the conditional expectation of the dependent variable is an unknown function of three parametric indices. The proposed estimator does much better than higher order kernels and regular kernel methods that do not impose bias controls. The estimator is updated in a succession of stages, with each stage decreasing the order of the bias. In the simulation experiments reported here, the performance of the model improves with each stage, with the first stage involving no bias correction.

Exploiting Variance Reduction Potential in Local Gaussian Process Search

♦*Chih-Li Sung*[1]*, Robert Gramacy*[2] *and Benjamin Haaland*[1]
[1]Georgia Institute of Technology
[2]The University of Chicago
chihli_sung@gatech.edu

Gaussian process models are commonly used as emulators for computer experiments. However, developing a Gaussian process emulator can be computationally prohibitive when the number of experimental samples is even moderately large. Local Gaussian process approximation (Gramacy and Apley, 2015) has been proposed as an accurate and computationally feasible emulation alternative. Constructing local sub-designs specific to predictions at a particular location of interest remains a substantial computational bottleneck in that technique. In this paper, two computationally efficient neighborhood search limiting techniques are proposed, a maximum distance method and a feature approximation method. Two examples demonstrate that the proposed methods indeed save substantial computation while retaining emulation accuracy.

Semi-Parametric Estimation for Multivariate Skew-Elliptical Distributions

Jing Huang
European School of Management and Technology
jing.huang@esmt.org

We investigate a class of skew-elliptical distributions which generalize the skewed distributions proposed by Azzalini and Dalla Valle (1996). These distributions are generated by elliptical contours plus an unknown distribution function which assigns weights or densities to these contours. Our model can decompose such a skewed distribution into those two parts above and allow us to estimate them separately, with both parametric and Non-parametric method. Our model is analyzed explicitly with simulated data estimation in one and two dimensional case. In a parametric framework and under the assumption that the density assigning function has a closed form, our approach leads to a faster estimation process than traditional log likelihood estimation (MLE) through direct optimization or through expectation maximization (EM). At end we also show the results of a fitting a real data with our model which better captures the empirical tail distribution.

Session 120: Recent Advances in Network Data Inference

Network Inference From Time Varying Grouped Observations

Yunpeng Zhao
George Mason University
yzhao15@gmu.edu

In social network analysis, the observed data is usually some social behavior, such as the formation of groups, rather than explicit network structure. Zhao and Weko (2015) proposed a model-based approach called the star model to infer implicit networks from grouped observations. Star models assumed independence among groups, which sometimes is not valid for practical consideration. In this article, we generalize the idea of star models into the case of time varying grouped observations. Similarly to star models, we assume the group at each time point is gathered by one leader, but allow dependency among groups in the same time segment. We apply a variant of Expectation-Maximization algorithm – hard EM for identifying group leaders and apply a label switching technique to optimize Bayesian information criterion for identifying segments. The performance of the new model is evaluated under different simulation settings. We apply this model to a data set of Kibale chimpanzee project.

Studying the Communication Network on Facebook of French Election with Spectral Contextualization

♦*Yilin Zhang, Karl Rohe and Chris Wells*
University of Wisconsin-Madison
yzhang672@wisc.edu

We studied the Facebook discussion threads surrounding the 2012 French presidential election. The eight major candidates posted a total of 3k posts during the official campaign. In response, 90k individuals (fans) made 600k comments on the candidate posts. In this study, we aim to uncover the structures of the communication

network between these fans and the candidate posts. We search for the following two types of communication patterns: (1) Candidate-centered, where individuals primarily comment on the wall of one candidate and (2) Issue-centered, where, individuals' attention and expression is directed towards a specific set of issues (e.g. economics, immigration, etc).

By applying spectral clustering to the graph structure of the discussion threads (who comments where), we find the dominant structure of the communication is the candidate-centered structure.

To search for the Issue-centered communication, we develop a new spectral technique for clustering called Spectral Contextualization, which simultaneously analyzes the graph and the text in the comments. We identify four sub-populations which will be discussed in the talk. Statistical consistency is provided under the Node Contextualized Statistic co-Blockmodel.

Graph-limit Enabled Fast Computation for Fitting Exponential Random Graph Models to Large Networks

♦*Ran He*[1] *and Tian Zheng*[2]
[1]Columbia University, Bell Labs
[2]Columbia University
ryan.rhe@gmail.com

Large network, as a form of big data, has received increasing amount of attentions in data science, especially for large social network, which is reaching the size of hundreds of millions, with daily interactions on the scale of billions. Thus analyzing and modeling these data in order to understand the connectivities and dynamics of large networks is important in a wide range of scientific fields. Among popular models, exponential random graph models (ERGMs) have been developed to study these complex networks by directly modeling network structures and features. ERGMs, however, are hard to scale to large networks because maximum likelihood estimation of parameters in these models can be very difficult, due to the unknown normalizing constant. Alternative strategies based on Markov chain Monte Carlo (MCMC) draw samples to approximate the likelihood, which is then maximized to obtain the maximum likelihood estimators (MLE). These strategies have poor convergence due to model degeneracy issues and cannot be used on large networks. Chatterjee et al (2013) propose a new theoretical framework for estimating the parameters of ERGMs by approximating the normalizing constant using the emerging tools in graph theory—graph limits. In this paper, we construct a complete computational procedure built upon their results with practical innovations which is fast and is able to scale to large networks. More specifically, we evaluate the likelihood via simple function approximation of the corresponding ERGM's graph limit and iteratively maximize the likelihood to obtain the MLE. We also discuss the methods of conducting likelihood ratio test for ERGMs as well as related issues. Through simulation studies and real data analysis of two large social networks, we show that our new method outperforms the MCMC-based method, especially when the network size is large (more than 100 nodes). One limitation of our approach, inherited from the limitation of the result of Chatterjee et al (2013), is that it works only for sequences of graphs with a positive limiting density, i.e., dense graphs.

A Hypothesis Testing Framework for Modularity Based Network Community Detection

♦*Jingfei Zhang*[1] *and Yuguo Chen*[2]
[1]University of Miami
[2]University of Illinois at Urbana-Champaign
emmauiuc2009@gmail.com

One of the most relevant features of networks representing real systems is the community structure. Detecting communities is of great importance in understanding, analyzing, organizing networks, as well as in making informed decisions. In recent years, many approaches have been proposed for detecting the community structures in networks. However, few methods have been proposed for testing the statistical significance of detected community structures. In this talk, we describe a statistical framework for modularity based network community detection. Under the proposed framework, a hypothesis testing procedure is developed to determine the significance of an identified community structure. Moreover, the proposed modularity is shown to be consistent under a degree-corrected stochastic block model framework. Several synthetic and real networks are used to demonstrate the effectiveness of our method.

Session 121: Recent Developments in Design of Experiments

Optimal designs for the two-dimensional interference model

Wei Zheng and ♦*Heng Xu*
Indiana University Purdue University Indianapolis
hengxu@umail.iu.edu

In many applications of block designs, a treatment assigned to a plot could also affect the neighboring plots. This neighboring effect is typically accommodated by the inference model to avoid the bias. The optimal design of experiments for such models has been extensively studied in the recent two decades. However, all of them assumed that the plots a block is arranged in a one-dimensional layout, and a design is represented by a collection of treatment sequences. This paper focus on the two-dimensional case, where a design shall be a collection of arrays.

Maximum Empirical Likelihood Estimation in U-statistics-based General Estimating Equations

♦*Lingnan Li and Hanxiang Peng*
Indiana University-Purdue University, Indianapolis
lingli@umail.iu.edu

In this talk, we discuss maximum empirical likelihood estimates (MELE's) in U-statistics based general estimating equations (UGEE's) and their asymptotic behaviors. It is exhibited that the asymptotic variance of a MELE will not increase as the number of UGEE's increases. We give several important examples and report some simulation results.

Key words: Jackknife pseudo value, maximum empirical likelihood estimate, U-statistics, U-statistics based general estimating equations.

Bayesian D-Optimal Design of Experiments with Quantitative and Qualitative Responses

Lulu Kang[1], ♦*Xinwei Deng*[2] *and Ran Jin*[2]
[1]Illinois Institute of Technology
[2]Virginia Tech
xdeng@vt.edu

Systems with quantitative and qualitative (QQ) responses are widely encountered in many applications. Design of experiment methods are needed when experiments are conducted to study such systems. Classic experimental design methods are not suitable here because they often focus on one type of responses. In this paper, we develop a Bayesian D-optimal design method for experiments with one quantitative and one binary qualitative responses. Both non-informative and conjugate informative prior distributions on the unknown parameters are considered. The proposed design criterion

has meaningful interpretations in terms of D-optimality for both QQ response models. Efficient point-exchange search algorithm is developed to construct the local D-optimal designs for given parameter values. Global D-optimal designs are obtained by accumulating the frequencies of the design points in local D-optimal designs, where the parameters are sampled from the prior distributions. The performances of the proposed methods are evaluated through two examples.

BAYESIAN SEQUENTIAL DATA COLLECTION FOR CALIBRATING QUALITATIVE VARIABLES

♦*Qiong Zhang and yongjia song*
Virginia Commonwealth University
qzhang4@vcu.edu

Statistical calibration of computer experiments has drawn on much attention recently. Most related works focus on calibrating quantitative variables of computer experiments. In this talk, I will discuss a sequential data collection scheme to calibrate qualitative variables in computer experiments. In each step of the sequential procedure, new design points are selected according to the value of information, which can be computed very efficiently using a Bayesian model. Numerical results are provided to show the effectiveness of the proposed method.

Session 122: Design and Analysis of Traditional Therapy Studies

Assess the Accuracy of Diagnosis and Syndrome Differentiation Results Made by Traditional Medicine P

♦*Zheyu Wang*[1] *and Andrew Zhou*[2]
[1]Johns Hopkins University
[2]University of Washington
wangzy@jhu.edu

Complimentary and alternative medicine (CAM) has received a lot of attention in recent years due to their commendable or controversy performance. The attempt to understand CAM is often challenged by that fact that theories behind CAM practices often involve subjective evaluations or beliefs that are not well explained by modern science. In this talk, we use traditional Chinese medicine (TCM) as an example, and consider assessing the accuracy of different TCM doctors' diagnosis without relying on the controversy theory. We propose a nonparametric maximum likelihood method for estimating and comparing the accuracy of different doctors in detecting a particular symptom without a gold standard when the true symptom status had an ordered multiple class. In addition, we extend the concept of the area under the receiver operating characteristic curve to a hyper-dimensional overall accuracy for diagnostic accuracy and alternative graphs for displaying a visual result. Simulations results and real data application will also be discussed.

Evaluating Traditional Chinese Medicine using Modern Clinical Design and Statistical Methodology

Lixing Lao[1], ♦*Yi Huang*[2], *Chiguang Feng*[3], *Brian Berman*[3] *and Ming Tan*[4]
[1]The Univ. of Hong Kong, School of Chinese Medicine
[2]University of Maryland, Baltimore County
[3]University of Maryland, School of Medicine
[4]Georgetown University, Medical Center
yihuang@umbc.edu

Traditional Chinese medicine (TCM) has been used in China and other Asian counties for thousands of years and is increasingly utilized in Western countries. Due to inherent differences in how Western medicine and TCM are practiced, employing the so-called Western medicine-based gold standard research methodology to evaluate TCM, an ancient Chinese medicine treatment modality, is challenging. This article, a discussion of the obstacles inherent in the design and statistical analysis of clinical trials of TCM, is based on our experience in designing and conducting a randomized controlled clinical trial of acupuncture for post-operative dental pain control in which acupuncture was shown to be statistically and significantly better than placebo in lengthening the median survival time to rescue drug. Using that trial, we demonstrate that PH assumptions in the common Cox model do not hold, and more thoughtful modeling and more sophisticated models of statistical analysis are warranted in TCM trials. TCM study design entails all the challenges encountered in trials of drugs, devices, and surgical procedures in Western medicine. We present possible solutions to some but leave many issues unresolved.

Acupuncture and Prevention of Chronic Postsurgical Pain: From Experimental Study to Clinical Trial

Jiangang Song
Shanghai University of Traditional Chinese Medicine
songjg1993@126.com

Thoracic surgeries including thoracotomy and video-assisted thoracoscopic surgery are some of the highest risk procedures that often lead to chronic post-surgery pain (CPSP). The prevalence of CPSP varies from 14% to 83%. Acute pain management including multimodal analgesia techniques aspires to stop the transition from acute pain to chronic pain, yet a significant reduction in the occurrence rate and severity of chronic pain has not occurred. Acupuncture has been a widely used method in traditional medicine in East Asia for thousands of years. Now, many animal studies have shown electroacupuncture (EA) -induced analgesia on neuropathic pain but clinical studies have been sparse. As iatrogenic nerve injury-induced neuropathic pain is probably the most important cause of long-term postsurgical pain, we design this prospective, randomized, controlled trial to test the effects of perioperative application of electroacupuncture on the prevention of incidence or the degree of CPSP. The primary outcome was the change in chest wall pain intensity at rest between baseline and 3 months. More importantly, before the operation, all patients would be categorized into different types according to tongue manifestation instrument, inquiry scale, pulse-taking instrument, acoustic diagnostic information collection system, which is originated from traditional Chinese medicine (TCM). The methods of EA, such as acupoints choices, timing and duration of treatment, would depend on TCM Syndrome Types and theory of TCM. Whether there is a correlation among TCM Syndrome Types, severity of CPSP, and efficiency of EA would also be analysed. In a word, we attempted to provide an accurate EA treatment for CPSP based on TCM Syndrome Types, and to find a suitable clinical trial design and statistical methods to demonstrate efficiency of EA.

Session 123: Statistical Analysis of Structural Morphology and Functional Measures in Medical Studies

Statistical Shape Analysis of Anatomical Structures Using Square-Root Normal Fields

Sebastian Kurtek
The Ohio State University
kurtek.1@stat.osu.edu

We present a Riemannian framework for comprehensive statistical shape analysis of 3D objects, represented by their boundaries (parameterized surfaces). By comprehensive framework, we mean tools for registration, comparison, averaging, and modeling of observed surfaces. Registration is analogous to removing all shape preserving transformations, which include translation, scale, rotation and re-parameterization. This framework is based on a special representation of surfaces termed square-root normal fields and a closely related elastic metric. The main advantages of this method are: (1) the elastic metric provides a natural interpretation of shape deformations that are being quantified, (2) this metric is invariant to re-parameterizations of surfaces, and (3) under the square-root normal field transformation, the complicated elastic metric becomes the standard L2 metric, simplifying parts of the implementation. We present numerous examples of shape comparisons for various types of surfaces in different application areas. We also compute average shapes, covariances and perform principal component analysis to explore the variability in different shape classes. These quantities are used to define generative shape models and for random sampling. Specifically, we showcase the applicability of the proposed framework in shape analysis of anatomical structures in different medical applications including Attention Deficit Hyperactivity Disorder and endometriosis.

Statistical Shape Analysis of Biological Morphology

Shantanu Joshi
University of California Los Angeles
s.joshi@g.ucla.edu

We present a statistical framework for characterizing and comparing morphological variation in shapes derived from biological data. The statistical framework makes use of the tangent principal component approach to achieve dimension reduction on the space of infinite-dimensional, non-linear, quotient space of shapes and enables computation of shape averages and covariances on the shape space in an intrinsic manner (adapted to the shape space). We will present applications to biomedical imaging, paleontology, and brain morphometry.

Statistical Analysis and Simulations of Soft Tissue Shapes in Medical Studies

♦*Chafik Samir and Thomas Deregnaucourt*
University of Clermont Auvergne
chafik.samir@udamail.fr

Statistical shape analysis plays an important role in various medical imaging applications. In particular, such methods provide tools for registering, deforming, comparing, averaging, and modeling anatomical shapes. In this talk, we focus on a recent methods for statistical shape analysis of elastic parameterized data (surfaces, images, signals) for diseases characterization, curves-based multimodals registration, and to simulation of realistic samples from a given set of observation.

Recent advances in medical imaging offer increasingly detailed information on typical anatomical structures. However, there is a lack of validation techniques for automatic strategies, especially for ground truth, e.g. real data can only be extracted manually by an expert, and then used to validate numerical methods. Consequently, scarcity of data for evaluation results in restricted or incomplete studies. An interesting solution is to statistically analyze shapes of real clinical data and provide enough random or simulated samples to cover lack of recorded data and expertise.

We present different examples (curves and surfaces) where statistical analysis of soft tissue shapes helps to improve the accuracy of some medical imaging techniques and visualization of key information.

FUNCTIONAL CLUSTERING BASED ON PROJECTION ESTIMATION WITH APPLICATION TO BLOOD-SAMPLE SPECTRUM

♦*Anne-Françoise Yao*[1] *and Gerald GREGORI* [2]
[1]Lab. Mathematics Clermont-Ferrand Univ(France)
[2]MIO, Aix-Marseille University(France)
Anne-francoise.Yao@math.univ-bpclermont.fr

This work deals with the problem of clustering a functional dataset, $(X_i; 1 \leq i \leq n)$ from a finite-number basis function expansions, $V_1, .., V_k$. If classically, the V_i's are a priori fixed (Wavelet basis, Fourier basis, B-spline,...) or estimated from the X_i's, Functional Principal Component Analysis (FPCA), theses approaches cannot be used the problem we deal with. In fact, at the contrary to the previous cases, the number of k is known and the V_i's are unknown. In other words, one knows the number of sources of variability but not the sources. Let us consider the situation where the X_i's are the reactions of n lymphocytes to k fluorescent antibodies simultaneously administered to a blood sample. Then, for each j = 1, ..., k, V_j is the fluorescent signal which expresses the effect of the labelling T_j. So, the V_i's cannot be identified nor directly estimated from the rug dataset (X_i). This problem looks like an Independent Component Analysis (ICA) problem but unlike such a problem, the V_j's have specific features which have to be estimated. Each labeling T_j leads to a specific unknown signal V_j depending on the very fluorochrome coupled to the antibody fixed to its specific lymphocyte. As a lymphocyte can be labeled by several fluorescent antibodies, the corresponding spectrum, X_j, expresses simply a linear combination of all the label's spectra. In other word, if X represent the whole spectrum, and if F is an operator representing all the normalized spectra, then X=FA, where A is a vector of abundances (or concentrations) of all the fluorochromes brought by the different antibodies. Different clusters are then characterized by distinct combinations of fluorochrome abundances.

Session 124: Adaptive Randomization: Recent Advances in Theory and Practice

Utility of Outcome Adaptive Randomization for Multisite Comparative Trials

♦*Mi-Ok Kim, Chunyan Liu and Nusrat Harun*
Cincinnati Children's Hospital Medical Center
miok.kim@cchmc.org

Outcome adaptive randomization is ethical design that treats more study subjects with superior treatment arms. It has recently shown little benefit for two-armed comparative trials, even yielding the larger sample for the inferior treatment arm at a non-negligibly high probability. We aim to evaluate feasibility and utility of the randomization scheme for multi-site comparative trials. The Interventional Management of Stroke III (IMS-III) trial, a well publicised randomized controlled stroke trial, provided the clinical context. The trial involved fifty-eight sites and two strata defined by disease severity, and balanced treatment allocation within each stratum and across the sites. It adopted three interim analyses to allow early stopping for efficacy or futility, whereas the primary outcome was not immediately observable. We used a doubly adaptive biased coin design (DBCD) targeting randomization probabilities that minimize nonfavorable outcomes and modified it to accommodate the constraints of the trial. Monte Carlo computer simulation was used to compare

the design with fixed randomization designs including the original design for operating characteristics such as % of the study subjects treated with the superior treatment arm, % with non-favorable primary outcome, % stopped early correctly and % stopped wrongfully. Accrual and outcome data was simulated using the trial data for different accrual speed and treatment effects. We show that outcome adaptive randomization, if implemented efficiently, is applicable to multi-site comparative trials with modest yet reliable benefits.

OPTIMAL FLEXIBLE SAMPLE SIZE DESIGN WITH ROBUST POWER

♦*Lanju Zhang*[1]*, Lu Cui*[1] *and Bo Yang*[2]
[1]AbbVie Inc
[2]Vertex Pharmaceuticals
lanju.zhang@abbvie.com

Flexible sample size designs, as an alternative to traditional fixed sample size designs, have been used in industry sponsored clinical studies. Unlike a fixed sample size design with the power targeting at a single value of treatment difference with respect to the primary efficacy endpoint, a flexible sample size design is to achieve a robust power over a range of potential values of the treatment difference. This presentation illustrates that concept, and shows how to design a flexible sample size trial under the aforementioned objective. Performance of commonly used flexible sample size methods, including group sequential designs, is evaluated with respect to a chosen benchmark. Design optimization is achieved via global search and by varying multiple design parameters. In the presentation, clarifications of some common confusions and results of simulations will be provided.

Comparing efficiency, randomness of adaptive treatment allocation procedures in clinical trials

Li-Xin Zhang
Zhejiang University
stazlx@zju.edu.cn

The randomness, efficiency and desirable allocation proportion are important components for evaluating a response-adaptive design in clinical trials and conflicted demands in applications. Efron (1971) defined a selection bias as a measure of lack of randomness for an adaptive allocation procedure. We study the randomness of adaptive randomization procedures via this measure and the efficiency via the asymptotic variability. It is proved that each response-adaptive randomization procedure has a minimum value of the selection bias. A new family of efficient allocation procedures is proposed for targeting any allocation proportion, preserving randomization such that their selection bias is minimum, and attaining the lower bound of the allocation variances at the same time. A new measure of lack of randomness is also proposed for comparing the adaptive allocation procedures with the same selection bias. This talk is based on joint work with Weisum Chan, Siuhung Cheung and Feifang Hu.

Keywords Response-Adaptive Design; Efficiency; Selection Bias; Asymptotic Normality; Clinical Trial

References Efron, B. (1971). Forcing a sequential experiment to be balanced. Biometrika, 62, 347-352. Hu, F. and Zhang, L.-X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. Ann. Statist., 32, 268-301 Hu, F., Zhang, L.-X. and He, X. (2009), Efficient randomized adaptive designs, Ann. Statist., 37, 2543–560. Zhang, L.-X., Hu, F., Cheung, S. H., Chan, W. S. (2014). Multiple-treatment efficient randomized-adaptive design with minimum selection bias. Manuscript.

Covariate-adjusted response-adaptive deigns and their statistical inference

Wanying Zhao
George Washington University
zhaowanying90@gmail.com

Covariate-adjusted response-adaptive (CARA) designs can assign subjects based on all history information from the on-going trial and covariates of the current subject. Since the allocation scheme and the estimation of parameters are based on both responses and covariates, CARA designs are very difficult to formulate, and the theoretical properties of conventional hypotheses testing remain unknown. In the literature, most studies are only based on simulations. For a clinical trial with two treatments, we derive asymptotic distributions of test statistic for comparing treatment effects under both null and alternative hypotheses based on a simple linear model. Under a family of CARA designs, we find out that (i) the two-sample t-test is not always valid when omitting some covariates in inference procedures and its performance depends on the allocation function and unknown parameters in model settings; (ii) proper adjustment on the variance of test statistic is required to achieve correct Type I error; (iii) when including covariates in final inference, CARA designs can obtain comparable power with complete randomization, but assign fewer subjects to inferior treatment and have better overall mean responses. Numerical studies are performed to verify corresponding finite sample properties.

Session 125: ROC Analysis and Estimation of Large Covariance Matrices

Statistical Inferences for the Partial Youden Index

♦*Chenxue Li*[1]*, Jinyuan Chen*[2] *and Gengsheng Qin*[3]
[1]Georgia State Univerisity
[2]Lanzhou Univeristy
[3]Georgia State University
cli8@student.gsu.edu

In medical diagnostic studies, the receiver operating characteristic (ROC) curve is widely used in the evaluation of the accuracy of a diagnostic test. The sensitivity and specificity of a diagnostic test are important accuracy measures. The Youden index, defined as the maximum sum of sensitivity and specificity over all possible cut-off points, is a summary index for the ROC curve. The partial area under the ROC curve (AUC) is another important summary index for the ROC curve. Motivated by the definition of the partial AUC, a new summary index for the ROC curve, called "partial Youden index", is defined as the maximum sum of sensitivity and specificity on an interval of cut-off points of interest. The popular Youden index is a special case of the partial Youden index. In this article, we propose various parametric and non-parametric confidence intervals for the partial Youden index. Extensive simulation studies are conducted to evaluate the finite sample performances of the new intervals.

Jackknife empirical likelihood confidence regions for ROC curve with non-ignorable verification bias

♦*Haiqi Wang*[1]*, Gengsheng Qin*[1] *and Jinyuan Chen*[2]
[1]Georgia State University
[2]Lanzhou University
hwang41@student.gsu.edu

In a continuous-scale diagnostic test, when a cut-off level is given, the performance of the test in distinguishing diseased subjects from non-diseased subjects can be evaluated by its sensitivity and specificity. Joint inferences for sensitivity and specificity as well as cut-off level play an important role in the assessment of the diagnostic

accuracy of the test. Recently we propose various bias-corrected empirical likelihood confidence region for sensitivity and specificity under non-ignoble verification bias, and derived the asymptotic results. Simulation studies are conducted to evaluate the finite sample performance and robustness of the proposed jackknife empirical likelihood-based confidence regions in terms of coverage probabilities. Also we use NACC Minimum Data Set to examine the performance of our method.

Jackknife empirical likelihood inference for the pairs of sensitivity,specificity and cut-point

♦*Jinyuan Chen*[1]*, Haiqi Wang*[2] *and Gengsheng Qin*[2]
[1]Lanzhou University
[2]Georgia State University
jchen55@gsu.edu

In receiver operating characteristics (ROC) analysis, the confidence regions for the the pairs of sensitivity, specificity and cut-point are widely used tools for evaluating discriminative and diagnostic power of a biomarker. Recently, missing the biomarker values for some observations is a frequent phenomenon in some medical researches. In this article, jackknife empirical likelihood-based methods with missing biomarker values are proposed for any pairs of sensitivity, specificity and cut-point value with the remaining parameter fixed at a given value, and asymptotic results can be derived accordingly. Simulation studies are conducted to evaluate the finite sample performance and robustness of the proposed jackknife empirical likelihood-based confidence regions in terms of coverage probabilities. Finally, a real example is provided to illustrate the application of our methods.

Tuning parameter selection in regularized estimations of large covariance matrices

Yixin Fang[1]*,* ♦*Binhuan Wang*[1] *and Yang Feng*[2]
[1]NYU School of Medicine
[2]Columbia University
binhuan.wang@nyumc.org

Recently many regularized estimators of large covariance matrices have been proposed, and the tuning parameters in these estimators are usually selected via cross-validation. However, there is a lack of consensus on the number of folds for conducting cross-validation. One round of cross-validation involves partitioning a sample of data into two complementary subsets, a training set and a validation set. In this manuscript, we demonstrate that if the estimation accuracy is measured in the Frobenius norm, the training set should consist of majority of the data; whereas if the estimation accuracy is measured in the operator norm, the validation set should consist of majority of the data. We also develop methods for selecting tuning parameters based on the bootstrap and compare them with their cross-validation counterparts. We demonstrate that the cross-validation methods with optimal choices of folds are more appropriate than their bootstrap counterparts.

Session 126: Integrating Modeling and Simulation (M&S) in the Drug Development

Integrating pharmacometrics and statistics to accelerate early clinical development: a case study

♦*Bret Musser*[1]*, James Bolognese*[2]*, Ghassan Fayad*[1]*, Yue Shentu*[1]*, Lori Mixson*[1]*, Nitin Patel*[2] *and Jaydeep Bhattacharyya*[2]
[1]Merck & Co., Inc
[2]Cytel
bret_musser@merck.com

Dose choice for Phase 3 clinical trials is one of the most difficult but important factors for a successful drug development program. Often the information upon which Phase 3 dose choice is based includes early development biomarker trials and Phase 2 dose-finding trials using a clinically relevant endpoint. This paper investigates a framework that leverages the relationship between early biomarkers and the target clinical endpoint to optimize a Phase 1-2 development plan. The framework is used to assess different biomarker designs for PoC studies to ultimately improve Phase 3 dose choice. A case study using a Bayesian tri-variate normal distribution model illustrates the framework. Among the findings of the simulation study are: (1) at typical sizes of Ph1b and Ph2b trials, biomarkers appear useful for PoC, but not for clinical endpoint dose-finding; (2) even with near perfect correlation between biomarkers and large amounts of Ph1b prior information, improved Ph3 dose selection requires increased Ph2b sample sizes. Thus, the fastest development path may be to move into Ph2b and measure the clinical endpoint of interest soon after PoC.

Applications of Bayesian Modeling and Simulation Methodology in the Pediatric Drug Development

♦*Chyi-Hung Hsu and Steven Xu*
Janssen Research & Development
hsuch1us@gmail.com

Pediatric drug development is challenging in many aspects, particularly, slow and difficult recruitment may contribute to the failure of the pediatric program. To increase the efficiency of pediatric drug development, extrapolation approach, based on adult data and other data, was proposed and advocated by the US Food and Drug Administration (FDA) The impact of this proposed approach was evaluated by Dunne, et al, and they concluded that extrapolating streamlines pediatric drug development and helps to increase the number of approvals for pediatric use. On the other hand, the appropriate level of extrapolation, e.g. full or partial, is often determined by the similarity in disease progression, and in response to intervention. Furthermore, Iit has to be decided and agreed a prior, even before the collection of pediatric data. In this presentation, a Bayesian alternative method will be investigated of which the level of extrapolation and information borrowing based on the concordance between adult and pediatric data.

Application of population-based modeling and simulation for dose justification in clinical developme

♦*Emmanuel Chigutsa and Johan Wallin*
Global PKPD, Eli Lilly and Company
chigutsa_emmanuel@lilly.com

Application of population-based modeling and simulation for dose justification in clinical development of anticancer drugs

During early phases of drug development, data are often available from pre-clinical species only. Pharmacometric approaches for human dose selection using pre-clinical data are reviewed. The application of pharmacometric methods for dose justification and dose optimization using late phase data are also discussed.

The approach using pharmacokinetic (PK) and pharmacodynamic (PD) data from mouse xenograft experiments to describe the exposure-response relationship in pre-clinical species is described. Using PK data from preclinical species (mouse and/or monkey), interspecies scaling approaches are often used to predict human PK from preclinical data. Based on the expected human PK and the exposures required for optimal drug efficacy, a range of first human doses can be identified. During clinical studies, important patient characteristics that influence human PK or PD may be identified.

The clinical impact of these covariates in the context of the proposed dosing regimen needs to be determined, and an example using necitumumab for treatment of squamous non-small cell lung cancer is given. In this example, clinical trial data was comprised of tumor size and overall survival. A simultaneous approach to modeling the 2 metrics was applied to maximize the use of all the available data. Using the final exposure-response model for necitumumab, simulation based approaches were employed to evaluate the adequacy of the proposed dose and compare the impact of alternative dosing regimens.
In conclusion, modeling and simulation is a useful tool throughout drug development from identification of the first human dose through to justification of the registration dose.

PK/PD modeling of recurrent events and CTS in optimizing Phase 3 dose selection

♦*Zhaoling Meng, Tao Sheng, Lei Ma, Qiang Lu, Dimple Patel and Hui Quan*
sanofi
zhaoling.meng@sanofi.com

We explored the utilization of plasma concentration data for a more informative dose justification through the exposure-response relationship. A PK and dose model is first built to understand the PK profile of the drug. A negative binomial regression model is then applied to establish the relationship between PK and a recurrent event endpoint. Some baseline covariates which potentially can impact PK and the number of events are included in the regression model. Through both PK/dose model and PK/PD model, treatment effects of different doses can be predicted for dose justification. Data from an asthma program with the annualized exacerbation rate as the primary endpoint were fitted with the above models. The approach provided satisfactory results for dose justification. Also, trial simulation provided further useful information for the determination of Phase III sample size for the desired probability of success. The PK/PD modeling onf a recurrent event endpoint s approach combined with CTS can ensued in a more informative decision making for the Phase 3 dose selection.

Session 127: Recent Developments in Statistical Learning of Complex Data

Noncrossing Ordinal Classification of Complex Data

Xingye Qiao
Binghamton University
qiao@math.binghamton.edu

Ordinal data are often seen in real applications. Regular multicategory classification methods are not designed for this data type and a more proper treatment is needed. We consider a framework of ordinal classification which pools the results from binary classifiers together. An inherent difficulty of this framework is that the class prediction can be ambiguous due to boundary crossing. To fix this issue, we propose a noncrossing ordinal classification method which materializes the framework by imposing noncrossing constraints. An asymptotic study of the proposed method is conducted. We show by simulated and data examples that the proposed method can improve the classification performance for ordinal data without the ambiguity caused by boundary crossings.

A Conditional Dependence Measure with Applications in Undirected Graphical Models

Jianqing Fan[1]*, Yang Feng*[2] *and* ♦*Lucy Xia*[3]
[1]Princeton University
[2]Columbia University
[3]Stanford University
lucyxia@stanford.edu

Measuring conditional dependence is an important topic in statistics with broad applications including graphical models. Under a factor model setting, a new conditional dependence measure is proposed. The measure is derived by using distance covariance after adjusting the common observable factors or covariates. The corresponding conditional independence test is given with the asymptotic null distribu- tion unveiled. The latter gives a somewhat surprising result: the estimating errors in factor loading matrices, while of root-n order, do not have material impact on the asymptotic null distribution of the test statistic, which is also in the root-n domain. It is also shown that the new test has strict control over the asymptotic significance level and can be calculated efficiently. A generic method for building dependency graphs using the new test is elaborated. Numerical results and real data analysis show the superiority of the new method.

Provable Sparse Tensor Decomposition and Its Application to Personalized Recommendation

♦*Wei Sun*[1]*, Junwei Lu*[2]*, Han Liu*[2] *and Guang Cheng*[3]
[1]Yahoo
[2]Princeton
[3]Purdue
sunweisurrey8@gmail.com

Tensor as a multi-dimensional generalization of matrix has received increasing attention in industry due to its success in personalized recommendation systems. Traditional recommendation systems are mainly based on the user-item matrix, whose entry denotes each user's preference for a particular item. To incorporate additional information into the analysis, such as the temporal behavior of users, we encounter a user-item-time tensor. Existing tensor decomposition methods for personalized recommendation are mostly established in the non-sparse regime where the decomposition components include all features. For high dimensional tensor-valued data, many features in the components essentially contain no information about the tensor structure, and thus there is a great need for a more appropriate method that can simultaneously perform tensor decomposition and select informative features. In this talk, I will discuss a new sparse tensor decomposition method that incorporates the sparsity of each decomposition component to the CP tensor decomposition. Specifically, the sparsity is achieved via an efficient truncation procedure to directly solve an L0 sparsity constraint. In theory, in spite of the non-convexity of the optimization problem, it is proven that an alternating updating algorithm attains an estimator whose rate of convergence significantly improves those shown in non-sparse decomposition methods. As a by-product, our method is also widely applicable to solve a broad family of high dimensional latent variable models, including high dimensional Gaussian mixtures and mixtures of sparse regression. I will show the advantages of our method in two real applications, click-through rate prediction for online advertising and high dimensional gene clustering.

Joint Estimation of Multiple Undirected Graphs with Covariates

♦*Peng Wang*[1] *and Xiaotong Shen*[2]
[1]University of Cincinnati
[2]University of Minnesota
jwpeng@gmail.com

Graphical models are commonly used to analyze the conditional correlation structure among large number of variables. Although

joint estimation of multiple graphical models has been well studied, there has not been much literature that could also incorporate the influence of the covariates on the graphical models. In this project, we estimate multiple precision matrices and their relationship with corresponding covariates simultaneously. In order to achieve sparseness within each precision matrix and within the coefficients of the covariates, we use a nonconvex penalty in our objective function, and we apply a partition rule to divide nodes into smaller subgroups to reduce computational cost. Difference convex method, co-ordinate decent algorithm and augmented Lagrangian method are used to solve the nonconvex optimization problem. We perform simulation studies and real-data analysis to illustrate the numeric performance of the proposed approach. This is joint work with Professor Xiaotong Shen.

Session 128: Nonclinical Statistical Applications in Pharmaceutical Industry

Statistical Experiences on Qualification of a Screening Assay

Jorge Quiroz
Merck Research Laboratories
jorge.quiroz@merck.com

Before a screening assay can be used in the decision process, it is required to qualify the assay to ensure the integrity of the decision and compliance with internal and external regulatory requirements. In this presentation, we discuss some of our experience in interacting with scientist in the qualification of a screening assay used to assess monoclonality.

Frequentist and Bayesian Simulations Using Random Coefficients Model to Set Shelf-Life Specification

♦*Richard Montes*[1] *and David LeBlond*[2]
[1]Hospira, a Pfizer company
[2]CMC Statistics
richard.montes@pfizer.com

Setting proper specification limits is one component of the overall control strategy to ensure product quality and consistency. Specifications are numerical limits to which a product quality attribute must conform throughout the product shelf-life (shelf-life specification limits) to be considered acceptable. Ideally, specification limits are set based on a priori knowledge of the true ranges of quality attribute over which the drug product has acceptable safety and efficacy outcomes. If this information is not completely known and there are no applicable compendial standards, specification limits are established based on statistical analysis of available release and/or stability data. Although guidance documents (ICH Q6A and Q6B) describe the general considerations for setting specification limits, they do not prescribe specific statistical methodologies. Ad hoc methods based on a fixed effects (Analysis of Covariance) model, as adapted from methods described by Allen et al. (1991), have been applied in the pharmaceutical industry. These methods estimate future expected initial values from release data then the change over the shelf-life is estimated from stability data. Since such methods often employ worst-case assumptions about process and analytical uncertainties, the derived specification limits may be too conservative (wide) to provide an informative, risk-based, control strategy. Further, inferences from the fixed effects model are only applicable to the particular lots monitored in the stability studies and not to the future values of lots drawn from the population. This study aims to develop statistical tools that holistically analyze release and stability data to model random lot effects and analytical variability more probabilistically than current ad hoc methods. A mixed effects (hierarchical) model is used to account for random lot effects. Specifically, a random coefficients, linear regression model is applied to analyze available release and stability data. The estimated parameters of the random coefficients model are used to simulate future values at the product shelf-life expiry. Quantiles estimated from simulated future values, corresponding to a specified coverage, are used to derive the shelf-life specification limits which are then compared to the true population quantiles. Both Frequentist and Bayesian versions are explored. A range of realistic population parameters (intercept mean and variance, slope mean and variance, and analytical variance) as well as study design parameters (release and stability data sample sizes, quantile coverage, and prior assumption for the Bayesian approach) are examined. The confidence coefficient (i.e., the proportion of the repeated simulations wherein the estimated quantiles bracket the true population quantiles) is taken as a risk-based metric of performance. The results of this study inform the choice of analysis options against a range of realistic scenarios so that when applied in actual specification setting situations, acceptably conservative operating performance is assured.

A Bayesian Approach to Analytical Biosimilarity Assessment

♦*Yanbing Zheng and Lanju Zhang*
AbbVie Inc.
yanbing.zheng@abbvie.com

Demonstration of analytical biosimilarity is required for a biosimilar product development. Recently, FDA proposed and applied a tiered approach for demonstrating analytical biosimilarity, where tier-1 critical quality attributes are subjected to statistical equivalence testing within the frequentist paradigm. In this work, we propose an alternative method to assess analytical biosimilarity under a Bayesian framework with flexible prior distributions. Simulation studies and a real-life case study are conducted to evaluate the proposed method and comparisons are made with the FDA equivalence test approach.

Bioequivalence evaluation of sparse sampling data using bootstrap resampling method

Meiyu Shen
Food and Drug Administration
meiyu.shen@fda.hhs.gov

We develop a statistical method for bioequivalence evaluation of highly sparse pharmacokinetic data.

In the pharmacokinetic bioequivalence study design for ophthalmic products, each subject with bilateral cataracts is randomly assigned one of two treatments (the test and reference products) to one of two eyes (the left and right eyes). If the test product is randomly assigned to the left eye, then the reference product is assigned to the right eye; vice versa. A single sample of aqueous humor is collected from each eye at the same assigned sampling time. Hence each subject contributes a pair of one-time point aqueous humor drug concentrations. Since each subject contributes one concentration value per eye (if one eye receives the test product, the other eye receives the reference product, and vice versa.), we have a pair of one-point aqueous humor concentrations per subject. Assuming one eye of each subject contributes one replicated measurement at the assigned sampling time point for one product, we can calculate the sample mean of drug concentrations over many subjects for each of two products at each time point. From the mean drug concentration-time profile, one value of AUC (or Cmax) can be obtained for each product. The standard errors of AUC and Cmax can be obtained for each product by bootstrapping subjects at each time point with

replacement.

Session 129: Recent Advances in Analysis of Interval-Censored Failure Time data and Longitudinal Data

Maximum Likelihood Estimation for the Proportional Odds Model with Partly Interval-Censored Data

♦*Liang Zhu*[1]*, Dingjiao Cai*[2]*, Yimei Li*[3]*, Xingwei Tong*[2]*, Jianguo Sun*[4]*, Deo Kumar Srivastava*[3] *and Melissa M. Hudson*[3]
[1]University of Texas Health Science Center
[2]Beijing Normal University
[3]St. Jude Children's Research Hospital
[4]University of Missouri-Columbia
liang.zhu@uth.tmc.edu

This paper discusses regression analysis of partly interval-censored failure time data arising from the proportional odds model. Such data frequently occur in many areas, including clinical trials, epidemiology studies, and medical follow-up studies, and a great deal of literature on their analysis has been established. However, most of the existing methods do not separate the exact and interval-censored failure times; thus, they may not be efficient. Also it is well-known that the proportional hazards model may not be appropriate in some situations. Corresponding to these, we develop a maximum likelihood estimation for the proportional odds model and show that the resulting estimators are consistent and asymptotically Gaussian. An extensive simulation study is performed to assess the finite sample properties of the method and indicates that the proposed method works well for practical situations. The approach is then applied to a set of real data on the first occurrence of growth hormone deficiency, which motivated this study.

Case-cohort studies with interval-censored failure time data

♦*Qingning Zhou, Haibo Zhou and Jianwen Cai*
University of North Carolina at Chapel Hill
qz4z3@mail.missouri.edu

The case-cohort design has been widely used as a means of cost reduction in assembling or measuring expensive covariates in large cohort studies. The existing literature on the case-cohort design is mainly focused on right-censored data. In practice, however, the failure time is often subject to interval-censoring, that is, the failure time is never exactly observed but known only to fall within some random time interval. In this paper, we consider the case-cohort study design with interval-censored failure time and fit the proportional hazards model to data arising from this design. We employ the inverse probability weighted likelihood function and propose a sieve estimation approach via Bernstein polynomials. The consistency and asymptotic normality of the resulting regression parameter estimator are established and the weighted bootstrap procedure is considered for variance estimation. Simulation results show that the proposed method works well for practical situations, and an application to data from the Atherosclerosis Risk in Communities (ARIC) study is provided for illustration.

Bayesian Nonparametric Inference for Panel Count Data with Dependent Observation Times

♦*Ye Liang*[1] *and Yang Li*[2]
[1]Oklahoma State University
[2]University of North Carolina - Charlotte
ye.liang@okstate.edu

In this paper we consider a bivariate panel count data consisting of observation times and counts of some recurrent event for a number of subjects. We treat both observation times and counts as realizations of stochastic processes and use a Cox process model with bivariate Gaussian processes to model the dependence between observation times and counts. We use a Bayesian framework for the inference procedure and discuss Bayesian computations for our inference. We compare our modeling approach with other competitors in simulation studies and a real data analysis.

Semiparametric varying-coefficient regression analysis of recurrent events

♦*Yang Li*[1]*, Yanqing Sun*[1] *and Li Qi*
[1]University of North Carolina at Charlotte
y.li@uncc.edu

We investigate a generalized semiparametric varying-coefficient model for recurrent event data that can flexibly model three types of covariate effects: time-constant effects, time-varying effects, and covariate-varying effects. Different link functions can be selected to provide a rich family of models for recurrent event data. The model assumes that the time-varying effects are unspecified functions of time and the covariate-varying effects are parametric functions of an exposure variable specified up to a finite number of unknown parameters. The estimation procedure is developed using profile weighted least squares estimation techniques. The asymptotic distributions of the proposed estimators are established. Our simulation study shows that the proposed methods have satisfactory finite sample performance and the methods are applied to analyze data from an acyclovir study.

Session 130: Survival Analysis and its Applications

Estimating the optimal treatment regime based on restricted mean lifetime

♦*Min Zhang*[1] *and Baqun Zhang* [2]
[1]University of Michigan
[2]Renmin University
mzhangst@umich.edu

Individualizing treatment to account for patient heterogeneity in response to treatment has received much attention lately. A treatment regime is a rule that assigns a treatment, from among a set of possible treatments, to a patient based on his/her characteristics. The recent literature has seen much development in methodologies for estimating the optimal treatment regime. The majority of the methodological development has been focused on continuous responses. In this paper, we propose a method to estimate the optimal treatment regime for survival outcomes, where the optimal treatment regime is defined as the one that maximizes the restricted mean lifetime across all feasible regimes. We propose a direct optimization method that estimates the optimal treatment regime by optimizing the value of regimes. This direct optimization method is able to incorporate outcome-regression model (Cox model) to improve efficiency of estimation. However, as a direct optimization method as opposed to outcome-regression based method, the performance of the method is less sensitive to misspecification of the outcome-regression model and enjoys certain robustness property. The method will be evaluated by extensive simulation studies and illustrated by a real data application.

Parameter estimation in survival models with missing failure indicators in EMR by incorporating Expe

♦*Li Li and Tianle Hu*
Eli Lilly and Company
li_li_x1@lilly.com

Healthcare decisions can be improved by incorporating real-world evidence (RWE). Using claim/EMR data patients, healthcare providers, payers and pharmaceutical companies can better assess the value of treatments based on actual health outcomes. However, overall survival, the gold standard endpoint in cancer studies, is not directly observable in these databases, which severely limits the use of these databases. In the past, last appearance of patients was used as a proxy for survival and regular survival analysis was done based on this proxy. However, this approach underestimates survival and fails to propagate the uncertainty arising from treating the proxy as the observed. A solution to determine the mortality status is to link the claim/EMR database to another commercial database such as Experian data in which death dates can be reliably obtained. This solution is less than perfect, because only a portion of the patients can be linked. In this project, we utilized linked patient information to estimate the probability of last appearance being an event for patients using a parametric logistic regression. This probabilistic model will then be used to generate multiple imputed case status for unlinked patients. Survival analysis will be done among multiple imputed cases in combination with observed cases. We will show improved performance of the proposed approach over existing approaches via simulations and may apply the proposed approach to an RWE project estimating survival among patients with non-small cell lung cancer.

Survival trees for left-truncated / right-censored data, with application to time-varying covariates

Wei Fu and ♦*Jeffrey Simonoff*
New York University
jsimonof@stern.nyu.edu

Tree methods (recursive partitioning) are a popular class of nonparametric methods for analyzing data. One extension of the basic tree methodology is the survival tree, which applies recursive partitioning to censored survival data. There are several existing survival tree methods in the literature, which are mainly designed for right-censored data. We propose a new survival tree for left-truncated and right-censored (LTRC) data, which can be seen as a generalization of the traditional survival tree for right-censored data. Further, we show that such a tree can be used to analyze survival data with time-varying covariates, essentially building a time-varying covariates survival tree. Implementation of the method is easy, and simulations and real data analysis results show that the proposed methods work well for both LTRC data and survival data with time-varying covariates.

Computerized Multistage testing

Duanli Yan
Educational Testing Service
dyan@ets.org

Computerized Multistage Testing Duanli Yan Educational Testing Service

Recently, Multistage Testing (MST) has received much of attention. Similar to computerized adaptive testing (CAT), MST allows the adaptation of the difficulty of the test to the level of ability of a test taker. Specifically, in MST, items are interactively selected for each test taker, but rather than selecting individual items, groups of items are selected and the test is built in stages. Over the last decade, researchers have investigated ways for an MST to incorporate most of the advantages from CAT and linear testing, while minimize their disadvantages. These features include testing efficiency and accuracy, greater control of test content, more robust item review, as well as simplified test assembly and administration. Thus, MST becomes of more and more interest to researchers and practitioners as technology advances. At ETS, we are actively involved in theoretical and applied work in the area of CAT, MST and linear test. These research efforts in MST include theoretical research on statistical models, estimation algorithms, and related research, as well as the applications of MST with several operational test programs. MST is highly suitable for testing educational achievement because it adapted to educational surveys and student testing. An edited volume Computerized Multistage Testing: Theory and Applications (Yan, von Davier and Lewis, 2014), winner of 2016 AERA award for Significant Contributions to Educational Measurement and Research Methodology, covers the methodologies, underlying technology, and implementation aspects of this type of test design. This presentation provides a general overview of a multistage test (MST) design and its important concepts and processes. A multistage testing is described, as is why it is needed, and how it differs from other test designs, such as linear test and computer adaptive test (CAT) designs. It discusses the scientific perspectives and practical considerations for setting up an MST program, including a brief history of MST, test design and implementation for various purposes, item pool development and maintenance, IRT-based and classical test theory-based methodologies for test assembly, routing and scoring, and existing software. It also discusses the current research, operational programs, and innovative future assessments using MST.

Session 131: Recent Developments in Nonparametric Statistics and their Applications

Tensor sufficient dimension reduction

Shanshan Ding
University of Delaware
sding@udel.edu

Data with array (tensor)-valued predictors are commonly encountered in contemporary statistical applications, such as neuroimaging and social network areas. For these complex data, sufficient dimension reduction (SDR) is demanding to extract useful information from abundant measurements yet conventional SDR methods are limited to handle tensor data structures. We propose higher-order moment-based sufficient dimension reduction approaches with a focus on tensor sliced inverse regression (tensor SIR) for data with tensor-valued predictors. Tensor SIR is constructed based on tensor decompositions and can reduce a tensor-valued predictor's multiple dimensions simultaneously. The proposed method provides fast and efficient estimation. We further investigate asymptotic properties and show its advantages by simulation studies and a neuroimaging data application.

Sufficient Dimension Reduction via Distance Covariance

Wenhui Sheng
University of West Georgia
SHENGUGA@GMAIL.COM

We introduce a novel approach to sufficient dimension-reduction problems using distance covariance. Our method requires very mild conditions on the predictors. It estimates the central subspace effectively even when many predictors are categorical or discrete. Our method keeps the model-free advantage without estimating link function. Under regularity conditions, root-n consistency and asymptotic normality are established for our estimator.We compare the performance of our method with some existing dimension-reduction methods by simulations and find that our method is very

competitive and robust across a number of models. We also analyze a real data to demonstrate the efficacy of our method.

A subsampled double bootstrap for massive data

Srijan Sengupta[1]*, Stanislav Volgushev*[2] *and* ♦*Xiaofeng Shao*[1]
[1]University of Illinois at Urbana-Champaign
[2]Cornell University
xshao@illinois.edu

The bootstrap is a popular and powerful method for assessing precision of estimators and inferential methods. However, for massive datasets which are increasingly prevalent, the bootstrap becomes prohibitively costly in computation even with modern computing platforms. Building on Bag of Little Bootstraps or BLB (Kleiner et al, 2014) and the idea of fast double bootstrap, I will propose a fast resampling method, the subsampled double bootstrap (SDB), for both independent data and time series data. The SDB is consistent under mild conditions for both independent and dependent cases. Methodologically, SDB is superior to BLB in terms of speed, sample coverage and automatic implementation for a given time budget. Its advantage relative to BLB and bootstrap is also demonstrated in simulations and data illustration.

A new class of measures for independence test with its application in big data

♦*Xiangrong Yin and Qingcong Yuan*
University of Kentucky
yinxiangrong@uky.edu

We introduce a new class of measures for testing independence between two random vectors, using characteristic functions. In this talk, by choosing a particular weight function in the class, we study a new index for measuring independence and its property. Sample versions and their asymptotic properties using different estimations are developed. We demonstrate the advantage of our methods via simulations and real data. In particular, we illustrate the effective use of our methods in big data analysis.

Session 132: Statistical Methods for Medical Research

Variable screening for classification with errors in the class labels

Guanhua Chen
Vanderbilt University
g.chen@vanderbilt.edu

Variable screening is an important tool for handling ultra-high dimensional classification problem. When there are errors in the class labels, properly taking the errors into account is critical for correctly screening the variables. Given the auxiliary information about the tendency of the errors in the class labels for each record, we propose resampling and weighting based methods for variable screening. We show the competitive performance of proposed method via simulations. We demonstrate that the method is potentially useful for precision medicine.

Robust Tests for Genetic Association Analysis Incorporating Genotyping Uncertainty

♦*Juan Ding, Wenjun Xiong, Junjian Zhang and You Su*
Guangxi Normal University
dingjuan@gxnu.edu.cn

Genome-wide association studies have been widely used as an effective approach to uncover genetic susceptibility loci for complex diseases. In modern genetic studies, genotype imputation has become standard practice due to large amounts of genetic variation and budget limit. The uncertainty in the assignment of genotypes should be accounted for while analyzing the association with imputed genotypes. To address this issue, we develop a robust test to detect the association between a locus and a phenotype while the genotype is uncertain and the genetic model is unknown. We investigate the performance of our test when the genetic model is misspecified, comparing with existing methods such as the dosage test.

Performance Evaluation of Propensity Score Methods for Multi-level Treatments

♦*Hui Nian*[1]*, Juan Ding*[2]*, Chang Yu*[1]*, William Wu*[1]*, Richard Shelton*[2]*, William DUpont*[1] *and Pingsheng Wu*[2]
[1]Department of Biostatistics, Vanderbilt University
[2]Department of Medicine, Vanderbilt University
hui.nian@vanderbilt.edu

The propensity score method is widely used to estimate the average treatment effect in observational clinical studies, but it is generally confined to binary treatment assignment. In an extension to the settings of multilevel treatment, Imbens proposed generalized propensity score which is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables, and the average treatment effect can be estimated by conditioning solely on the generalized propensity score under the assumption of weak unconfoundedness. In the present work, we adopted this approach and conducted extensive simulations to evaluate the performance of several methods using the generalized propensity score, including subclassification, matching, inverse probability of treatment weighting and covariate adjustment. The results revealed that inverse probability of treatment weighting had the preferred overall performance. We also applied these methods to assess the impact of exposure to different types of antidepressant medications (no exposure, SSRI only, non-SSRI only, and both) during pregnancy in a retrospective cohort study of 228,876 pregnant women.

Batch Effects on Design and Analysis of Equivalence and Non-inferiority Studies

♦*Jason Liao*[1] *and Ziji Yu*[2]
[1]Merck
[2]University of Rochester
jason_liao@merck.com

Batch effects, which are not the biological differences, are the technical sources of variation that have been added to the drug substances and drug products during the process of handling. Although the implication of batch effect has been widely recognized in statistical literature, no batch variability information is considered in designing and analyzing clinical studies. In this paper, the impact of batch variability is systematically explored on both the design stage and data analysis stage of equivalence and non-inferiority studies. In the design stage, including more batches in the study can increase the probability of success for demonstrating equivalence or non-inferiority but maintain the control of type I error. In the data analysis stage, ignoring the batch effect may cause markedly underestimation of the variability and may lead to type I error inflation.

Session 133: Statistical Analysis of Complex Data II

A Mixed Effects Model for analyzing Complex AIDS Clinical Data

Tao Wang
School of Mathematics Yunnan Normal University CN
wtaokm@263.net

Chinese government started providing highly active antiretroviral therapy (HAART) free of charge in 2002 and has been developing China HAART Database since 2004 in which three kinds of correlative AIDS progression markers, i.e. unbalanced longitudinal CD4 , CD8, and viral RNA data are included. These data are typically intermittently missing and informatively left censored. Up to date, few model-based curative effect evaluation research on Chinese HAART based on this database have been published, and almost all of them only focused on modeling CD4 dynamic process without taking CD8, viral RNA data and the missing, censored mechanism into account, thus great bias and false result may be yielded. In this talk, we propose a parsimonious generalized linear mixed effects model to jointly inference the dynamic progression of CD4, CD8 and viral RNA data for AIDS clinical data sets in the database. We characterize the CD4 CD8 and viral RNA dynamic progress, by taking the correlation among such three AIDS progression markers into account .Simulation studies and real data analysis demonstrate that our model performs well and is appropriate for evaluating HAART in practice.

Simulated Data for SNP Set Association Tests in Family Samples

♦*Hung-Chih Ku*[1] *and Chao Xing*[2]
[1]DePaul University
[2]University of Texas Southwestern Medical Center
hku4@depaul.edu

The traditional statistical analysis of genome-wide association studies (GWAS) attempts to assess the association between a single nucleotide polymorphism (SNP) and a disease phenotype. Recently, kernel machine-based tests for association between a SNP set (e.g. SNPs in a gene) and the observed phenotype have been proposed to improve the power of capturing the association. We apply a generalized F-test on testing the genetic effects with familial data, including main and interaction effects, and compare with two existing R packages, famSKAT and KMFAM, for family-based sequence kernel association testing. We use simulation studies to evaluate the performance by type I error rate and power of the test from difference sample sizes. The results show that the type I error rate is controlled and the power is higher than competing methods in detecting association from correlated individuals. The key advantage is that it allows for an unspecified function in the linear mixed model. We hope this will facilitate data analysis to identify novel genes that are associated with diseases.

A Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations

♦*Chen Chen, Min Zhang and Dabao Zhang*
Purdue University
chen1167@purdue.edu

Linear systems of structural equations have been recently investigated to reveal the structures of genome-wide gene interactions in biological systems. However, building such a system usually involves a huge number of endogenous variables and even more exogenous variables, and hence demands a powerful statistical method which limits memory consumption and avoids intensive computation. We propose a two-stage penalized least squares method to build large systems of structural equations based on a new view of a classical method. Fitting one linear model for each endogenous variable at each stage, the method develops optimal prediction of a set of surrogate variables at the first stage, and consistent selection of regulatory endogenous variables from massive candidates at the second stage. While this method is computationally fast and allows for parallel implementation, it provides regulatory effect estimates which enjoy the oracle properties. We demonstrate the effectiveness of the method by conducting simulation studies, showing its improvements over other methods. Our method was applied to construct a yeast gene regulatory network with a genetical genomics data set.

Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams

♦*Kun Liu, Ruizhi Zhang and Yajun Mei*
Georgia Institute of Technology
liukun0924@gmail.com

In this article, motivated by biosurveillance and censoring sensor networks, we investigate the problem of distributed monitoring large-scale data streams where an undesired event may occur at some unknown time and affect only a few unknown data streams. We propose to develop scalable global monitoring schemes by parallel running local detection procedures and by combining these local procedures together to make a global decision based on SUM-shrinkage techniques. Our approach is illustrated in two concrete examples: one is the nonhomogeneous case when the pre-change and post-change local distributions are given, and the other is the homogeneous case of monitoring a large number of independent $N(0,1)$ data streams where the means of some data streams might shift to unknown positive or negative values. Numerical simulation studies demonstrate the usefulness of the proposed schemes.

Phase-amplitude functional framework for analyzing RNA sequencing data with point process filtering

♦*Sergiusz Wesolowski*[1]*, Daniel Vera*[2] *and Wei Wu*[3]
[1]Prefer not to mention
[2]Center of Genomics, Florida State Univ
[3]Department of Statistics, Florida State Univ
wesserg@gmail.com

The crucial component of the modern gene regulation analysis is a proper interpretation of the experimental results. The rapid development of the sequencing hardware left the statistical and mathematical tools lagging behind. Uncovering the critical pieces of information and the full potential of sequencing methods from the experimental outcomes is not a trivial task and requires more advanced models. In this work we aim to bridge that gap.

We propose, yet another, but revolutionary in design, framework to analyse the Next Generation Sequencing (NGS) experiments. Our approach is based on a functional interpretation of the NGS results through the point process filtering. The new model is capable of decoding new information from the NGS data. To construct the functional inference framework we utilize the Radon-Nikodym density representation of a point process. Variety of density representations allows the model to be easily adapted to various data questions over any genomic regions of interest. We evaluate the performance of the new framework in a simple, but novel way of analysing exon level differential expression. The RN density equips the analysis with a normalising procedure and a functional equivalent of a Fisher test to quantify differential patterns. The phase-amplitude separation gives an efficient noise reduction procedure improving the sensitivity and specificity of the method. A Heuristic justification of the test statistic distribution is provided. By comparison with Cufflinks, DESeq2, Limma-voom, we show that our new method uncovers, previously not available information about gene activity patterns.

Assessment of drug combination effects using mixed-effects model

Shouhao Zhou

UT - MD Anderson Cancer Center
szhou@mdanderson.org

We will discuss a novel approach to assess synergy of drug combination experiments. Starting from the Loewe additivity model and the marginal dose-effect curve for each drug involved in a combination, we first generalize the idea of interaction index generalized into a broad setting, and present a procedure to estimate the (generalized) interaction index and its associated confidence interval at a combination dose with observed effects. The approach is illustrated using data coming from a study in which the inhibition effect of a combination of two compounds is studied using 96-well plates and a fixed-ratio ray design.

Session 134: Statistical Genetics

Kernel-based Nonparametric Testing in High-dimensional Data with Applications to Gene Set Analysis

♦*Tao He*[1]*, Ping-Shou Zhong*[2]*, Yuehua Cui*[2] *and Vidyadhar Mandrekar*[2]
[1]San Francisco State University
[2]Michigan State University
hetao@sfsu.edu

We consider testing a nonparametric function of high-dimensional variates in a reproducing kernel Hilbert space, which is a function space generated by a positive definite kernel function. We propose a test statistic to test the nonparametric function under the high-dimensional setting. The asymptotic distributions of the test statistic are derived under the null hypothesis and a series of local alternative hypotheses, in the "large p, small n" setup. Extensive simulation studies and a real data analysis were conducted to evaluate the performance of the proposed method.

Optimal Filtering to Increase Detections of Differentially Expressed Genes in Microarray Data

♦*Zixin Nie and Kun Liang*
University of Waterloo
z5nie@uwaterloo.ca

Detecting differentially expressed genes in microarray datasets is difficult due to the large number of genes to be simultaneously tested, resulting in very low power for each test under multiple testing error rate controls. Controlling the false discovery rate (FDR) instead of the family-wise error rate (FWER) can increase the power of each test, however the number of genes detected is still limited compared to the expected number of differentially expressed genes. Methods that use empirical Bayesian estimation have been used to increase the number of detected genes while maintaining the FDR. In this paper we explore a different method by filtering out genes that are less likely to be differentially expressed. These filtering methods are compared to commonly used empirical Bayesian methods such as LIMMA and Empirical Bayesian Arrays through simulations and applications.

A unified X-chromosome genetic association test accounting for different XCI processes

♦*Jian Wang, Robert Yu and Sanjay Shete*
UT MD Anderson Cancer Center
jianwang@mdanderson.org

X-chromosome inactivation (XCI) on female X-chromosome loci states that in females during early embryonic development, 1 of the 2 copies of the X-chromosome present in each cell is randomly chosen to be inactivated to achieve dosage compensation of X-linked genes in males and females. That is, 50% of the cells have one allele inactive and the other 50% of the cells have the other allele inactive. As a result, the homozygous genotypes in females have the similar effects as the hemizygous allele types in males. In general, the XCI process is random; however, studies have suggested that skewed XCI is a biological plausibility in which more than 75% of cells have the same allele active. Furthermore, some of the X-chromosome genes escape XCI outside the pseudo-autosomal regions, i.e., both alleles are active in all cells. Current standard statistical tests for X-chromosome genetic association studies either account for random XCI only (e.g., Clayton's approach) or escaping of XCI only (e.g., PLINK software). Because for females, the true XCI process is unknown and differs across different regions on X-chromosome, we proposed a unified approach for analyzing X-chromosomal genetic data, which will account for all such biological possibilities: random XCI, skewed XCI and escaping of XCI. We conducted simulation studies to compare the performance of the proposed approach with existing approaches and applied the proposed and existing approaches to the X-chromosomal genetic association study of head and neck cancer.

Structured Sparse Co-Inertia Analysis with Applications to Genomics and Metabolomics Data

♦*Eun Jeong Min, Sandra Safo and Qi Long*
Emory University
ej.min@emory.edu

Rapid advances in technology have led to the explosion of omics data in biomedical research. As a result, there has been an increasing interest in methods for integrative analysis of multiple types of omics data. Co-inertia analysis (CIA) is one of the statistical tools for assessing relationships and trends in two data types. The well known multivariate analysis, canonical correlation analysis (CCA) can be regarded as a special case of CIA in terms of the choice of the distance measure between two data points. This generalization lets us enable to analyze a relationship between a paired data that exhibit high-dimensionality, which is impossible for CCA. Also CIA can be easily extended to the multiple data sets, the multiple co-inertia analysis (MCIA) stands for the extended CIA. Despite of these advantages that CIA have, it has been traditionally used in ecology area. More recently, it has been used for integrative analysis of omics data. We propose a structured sparse co-inertia analysis that incorporates biological information such as network information among genes or metabolites. Our proposed method is evaluated in simulations and illustrated in an application to integrative analysis of genomics and metabolomics data.

Testing for gene-gene interaction in case-control GWAS

Zhongxue Chen
Indiana University Bloomington
zc3@indiana.edu

Detecting gene-gene interaction is an important but challenging task in genome-wide association studies (GWASs). To this end, many statistical methods have been proposed in the literature. However, powerful yet robust approaches are yet to be developed. Herer we study the gene-gene interaction tests for case-control GWASs. A number of powerful tests can be constructed for given situations. We also discuss some tests for the main effects and the overall tests for association between genotype and phenotype. A simulation study is conducted to compare some of the proposed tests with existing methods. A real data application is also conducted to illustrate the use of the proposed tests.

A Set-Valued System Model for Secondary Trait Genetic Association Analysis in Case-Control Studies

Wenjian Bi

St. Jude Children's Research Hospital
wenjian.bi@stjude.org

In many case-control designs for genome-wide association studies (GWAS) or next generation sequencing studies (NGS), extensive data about correlated secondary continuous/binary traits that may share the common genetic variants with the primary disease status are available. Investigation of these secondary traits may provide critical insights about the disease etiology or pathology and enhance the primary GWAS or NGS results. As the case-control sample is not a random sample from the general population, the relationship between primary disease status, secondary trait, and genetic variations need to be considered carefully to avoid inflated type I error rates. Here we propose a set-valued model (SV) that efficiently models the dichotomizing process of binary trait with a latent continuous variable instead of traditional logistic regression-based model (LG). Across an extensive series of simulation studies, SV maintained type I error control across the spectrum of minor allele frequencies (MAFs) and LG method generated inflated type I error rates when SNPs are rare (MAF=0.005), especially at a stringent significant level of 10-5. Additionally, SV method has similar or greater power than LG method, especially when secondary phenotype is continuous and SNPs are rare (MAF=0.005). For instance, in a simulation that data was generated from logistic model with an odds ratio of 1.2 between disease status and secondary continuous phenotype, for a rare single nucleotide polymorphism with a MAF of 0.005 and a sample size of 4,000, the SV method had 80% power whereas the LG method had 40% power at the 10-5 level. We applied our method to a genome-wide association study on Benign Ethnic Neutropenia/Leukopenia, and found out more SNPs associated with secondary phenotypes compared with LG method.

Session 135: Topics in Statistics II

Semiparametric Inference via Sparsity-Induced Kriging for Massive Spatial Datasets

♦*Pulong Ma and Emily Kang*
University of Cincinnati
mapn@mail.uc.edu

With the development of new remote sensing technology, large or even massive spatial datasets covering the globe become available. Statistical analysis of such data is challenging. This manuscript proposes a semiparametric approach to modeling and inference for massive spatial datasets. In particular, a Gaussian process with additive components is considered, with its covariance structure coming from two components: one part is flexible without assuming a specific parametric covariance function but is able to achieve dimension reduction; the second part is parametric and simultaneously induces sparsity. The inference algorithm for parameter estimation and spatial prediction is devised. The resulting spatial prediction method that we call sparsity-induced kriging (SIK), is applied to simulated data and a massive satellite dataset. The results demonstrate the computational and inferential benefits of SIK over competing methods and show that SIK is more flexible and more robust against model misspecification.

Analyzing of mutation of TNBC cell cytoplasmic level using Bayesian partition methods

♦*Guanhao Wei*[1]*, Jing Zhang*[1]*, Remus Osan*[1] *and Ritu Aneja*[2]
[1]Dept. of Math and Stat,Georgia State University
[2]Dept. of Biology,Georgia State University
guanhao91@gmail.com

Patients suffering from triple-negative breast cancer (TNBC) have poor prognosis mainly because no standard treatment is currently available. And classical cyclophosphamide, methotrexate, 5-fluorouracil (CMF) regimens for TNBC patients may be more effective than any other therapies. However, it is still limited and may cause tumor cell resistance. Our objectives were to explore the characteristic diversity on cell cytoplasmic level patients with TNBC. We hope after using bayesian partition models, we can detect the probability distribution of multiple independent cell cytoplasmic mutation level, and infer interaction structures between these mutations based on different biomarkers. And we hope we can investigate the dependent effect from several different cytoplasmic levels and use the result to find out a new way to improve the curative effect of classical CMF method.

BAYESIAN METHOD FOR CAUSAL INFERENCE IN MULTIVARIATE TIME SERIES WITH APPLICATION ON SALES DATA

BO NING
North Carolina State University
ningbo1990118@gmail.com

Advertisement campaign often runs in multiple stores across different locations simultaneously during a given period, spatially correla- tions are possible exists between stores. We propose a novel Bayesian method for detecting causal impact in multivariate time-series with responses are spatially correlated. A G-Wishart prior with given graphical structure on the precision matrix is used to impose sparsity on the inverse of covariance matrices. We use Bayesian multivariate basic stricture model to carry out causal inference analysis. Stationary constraint is imposed on local approximation variable of linear trend to stabilize long-run prediction credit intervals. We measure the causal effect by comparing the posterior distribution of the trend given the entire data and that given a part of data without observations possible affected by the causal impact. Our method could dealing with more general datasets which contain missing values and post-campaign data. At last, the method is applied on a data set on sales to determine the effect of an advertising campaign.

PCAN: Probabilistic Correlation Analysis of Two Non-normal Data Sets

Roger Zoh
Texas A&M University School of Public Health
rszoh8@gmail.com

Most cancer research now involves one or more assays profiling various biological molecules, e.g., messenger RNA and micro RNA, in samples collected on the same individuals. The main interest with these genomic data sets lies in the identification of a subset of features that are active in explaining the dependence between platforms. To quantify the strength of the dependency between two variables, correlation is often preferred. However, expression data obtained from next-generation sequencing platforms are integer with very low counts for some important features. In this case, the sample Pearson correlation is not a valid estimate of the true correlation matrix, because the sample correlation estimate between two features/variables with low counts will often be close to zero, even when the natural parameters of the Poisson distribution are, in actuality, highly correlated. We propose a model-based approach to correlation estimation between two non-normal data sets, via a method we call Probabilistic Correlations ANalysis, or PCAN. PCAN takes into consideration the distributional assumption about both data sets and suggests that correlations estimated at the model natural pa-

rameter level are more appropriate than correlations estimated directly on the observed data. We demonstrate through a simulation study that PCAN outperforms other standard approaches in estimating the true correlation between the natural parameters. We then apply PCAN to the joint analysis of a microRNA (miRNA) and a messenger RNA (mRNA) expression data set from a squamous cell lung cancer study, finding a large number of negative correlation pairs when compared to the standard approaches.

Intensity Estimation in Poisson Process with Compositional Noise

♦*Glenna Gordon, Wei Wu and Anuj Srivastava*
Florida State University, Department of Statistics
ggordon@fsu.edu

Intensity estimation for Poisson processes is a classical problem and has been extensively studied over the past few decades. Practical observations, however, often contain compositional noise, i.e. a nonlinear shift along the time axis, which makes standard methods not directly applicable. The key challenge is that these observations are not "aligned", and registration procedures are required for a successful estimation. In this paper, we propose an alignment-based framework for positive intensity estimation. We first show that the intensity function is area-preserved with respect to compositional noise. Such property implies that the time warping is only encoded in the normalized intensity, or density, function. Then, we decompose the estimation of the intensity by the product of the estimated total intensity and estimated density. The estimation of the density relies on a metric which measures the phase difference between two density functions. An asymptotic study shows that the proposed estimation algorithm provides a consistent estimator for the normalized intensity. We then extend the framework to estimating non-negative intensity functions. The success of the proposed estimation algorithms is illustrated using two simulations. Finally, we apply the new framework in a real data set of neural spike trains, and find that the newly estimated intensities provide better classification accuracy than previous methods.

The application of association rules mining in vehicle crash study for Mississippi coastal areas

♦*Zhao Ma, Feng Wang and Ningning Wang*
Jackson State University
zhao.ma@students.jsums.edu

Vehicle crash is considered as one of the top 10 leading causes of death in the United States, and there have been considerable researches conducted on the analysis of vehicle crash data over the past 20 years. A large proportion of the attempts didn't end up reaching good results due to the high correlation among the crash characteristics. In this study, we have proposed an association analysis and discovered association rules among traffic events, environmental variables (e.g., weather, road surface, and light conditions), and transportation infrastructures in crash data from the coastal areas of Mississippi for the period of 2011 to 2013. Lift, the assessment index, is compared with 1 for each variables. Lift greater than 1 indicates positive interdependence between the antecedent and the consequent; the larger the lift, the stronger the interdependence. In another word, the results demonstrate the associations between vehicle crashes and crash characteristics. This methodology may be developed to be a decision assistance tool for the traffic safety administrators.

Index of Authors

See you in Chicago,
Illinois in 2017!

www.icsa.org