Acknowledgements

ENAR would like to acknowledge the generous support of the 2010 Local Arrangements Committee, chaired by Brian Marx, Louisiana State and Julia Voluafova, The Louisiana State University Health Sciences Center, and our student volunteers.

ENAR is grateful for the support of the National Institutes of Health (National Cancer Institute, National Institute of Allergy and Infectious) and of the ENAR Junior Researchers' Workshop Coalition (Emory University, Virginia Commonwealth University, Columbia University, Harvard University, The Johns Hopkins University, North Carolina State University, The Ohio State University, The University of Michigan, The University of North Carolina at Chapel Hill, and The University of Wisconsin).

We gratefully acknowledge the invaluable support and generosity of our Sponsors and Exhibitors.

SPONSORS

Abbott
Allergan
AMGEN
Cepahlon, Inc.
Cytel Inc.
GlaxoSmith Kline
Pfizer
PPD, Inc.
RPS, Inc.
Rho, Inc.
SAS
Statistics Collaborative, Inc.

EXHIBITORS

The Cambridge Group Ltd.
Cambridge University Press
CRC Press – Taylor & Francis
DHHS/FDA/Center for Biologics Evaluation and Research
Food and Drug Administration
Kforce Clinical Research
Oxford University Press
Penn State World Campus
RPS, Inc.
Salford Systems
SAS
SAS Institute
SAS Institute Inc. – JMP Division
SIAM
Springer
Wiley-Blackwell

Officers & Committees

EXECUTIVE COMMITTEE – OFFICERS

President	Sharon-Lise Normand
Past President	Lance Waller
President-Elect	Amy Herring
Secretary (2009-2010)	Maura Stokes
Treasurer (2010-2011)	Michael Daniels

REGIONAL COMMITTEE (RECOM)

President (Chair) Sharon-Lise Normand
Eight ordinary members (elected to 3-year terms):
Hormuzd Katki (RAB Chair)

2008-2010	2009-2011	2010-2012
Jianwen Cai	Daniel Heitjan	Scarlett Bellamy
Bradley Carlin	José Pinheiro	Vernon Chinchilli
Peter Macdonald	Joanna Shih	Brent Coull

REGIONAL MEMBERS OF THE COUNCIL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Timothy G. Gregoire, Xihong Lin, Jane Pendergast, José Pinheiro, and Jeremy Taylor

APPOINTED MEMBERS OF REGIONAL ADVISORY BOARD (3-year terms)
Chair: Hormuzd A. Katki

2008-2010	2009-2011	2010-2012
Karla V. Ballman	Dipankar Bandyopadhyay	Thomas Braun
Craig Borkowf	Andrew Finley	Jaroslaw Harezlak
Avital Cnaan	Haoda Fu	Yulei He
Kimberly Drews	Ronald Gangnon	Robert Krafty
Matthew Gurka	Eugene Huang	Sandra Lee
Monica Jackson	Renée Moore	Karen Lynn Price
Robert Johnson	Roger Peng	Bhramar Mukherjee
Robert Lyles	Jennifer Schumi	Hernando Ombao
Peter Song	Brian Smith	Juned Siddique
Ram Tawari		Rui Wang

Programs

2010 Joint Statistical Meeting
Yulei He

2011 Joint Statistical Meeting
Bhramar Mukherjee

2010 Spring Meeting – New Orleans, LA
Program Chair: Michael Daniels
Program Co-Chair: Jeffrey Morris
Local Arrangements Chairs: Brian Marx and Julia Volaufova

2011 Spring Meeting – Miami, Florida
Program Chair: Ciprian Crainiceanu
Program Co-Chair: Eugenio Andraca-Carrera
Local Arrangement Chairs: Tulay Koru-Sengul and John Kairalla

Biometrics Editor	Marie Davidian
Biometrics Co-Editors	Geert Molenberghs, Naisyin Wang, and David Zucker
Biometric Bulletin Editor	Urania Dafni
JABES Editor	Carl Schwarz
ENAR Correspondent for the Biometric Bulletin	Lillian Lin
ENAR Executive Director	Kathy Hoskins
International Biometric Society Business Manager	Dee Ann Walker

Representatives

COMMITTEE OF PRESIDENTS OF STATISTICAL SOCIETIES (COPSS)
ENAR Representatives
Sharon-Lise Normand (President) Lance Waller (Past-President) Amy Herring (President-Elect)

ENAR Standing and Continuing Committees

Nominations Committee
Eric Feuer, Chair
Karen Bandeen-Roche
Lisa LaVange
Tom Louis
Elizabeth Margosches
Linda Young

Sponsorship Committee
Christine Clark, Chair
Thomas Kelleher
IkSung Cho

ENAR Representative on the ASA Committee on Meetings
Maura Stokes, Committee Vice-Chair (January 2008-December 2010)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE (Joint with WNAR)
Terms were through February 22, 2011

Section E, Geology and Geography	Carol Gotway Crawford
Section N, Medical Sciences	Judy Bebchuk
Section G, Biological Sciences	Geof Givens
Section U, Statistics	Mary Foulkes
Section O, Agriculture	Mary Christman

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
(ENAR President is also an ex-officio member) Board of Trustees
Members: Sharon-Lise Normand, ENAR President
Donna Brogan

Workshop for Junior Researchers
Yi Li (Chair), Harvard University
DuBois Bowman, Emory University
Kimberly Drews, George Washington University
Amy Herring, The University of North Carolina at Chapel Hill
Bhramar Mukherjee, University of Michigan
Limin Peng, Emory University

2010 Fostering Diversity in Biostatistics Workshop
Reneé H. Moore (co-chair), University of Pennsylvania School of Medicine
Adriana Perez (co-chair), The University of Texas Health Science Center At Houston
Scarlett Bellamy, University of Pennsylvania School of Medicine
DuBois Bowman, Emory University
Kathryn Chaloner, University of Iowa
Tahera Darensburg, Emory University School of Public Health
Amita Manatunga, Emory University School of Public Health
Sastry Pantula, North Carolina State University
Dionne Price, Food and Drug Administration
DeJuran Richardson, Lake Forest College
Louise Ryan, Harvard University School of Public Health
Keith Soper, Merck Research Laboratories
Lance Waller, Emory University, Rollins School of Public Health

Distinguished Student Paper Awards Committee
Eric Feuer (Chair), National Cancer Institute
Christopher Bilder, University of Nebraska-Lincoln
Nilanjan Chatterjee, National Cancer Institute
Hongzhe Li, University of Pennsylvania
Yi Li, Harvard University
Robert Lyles, Emory University Rollins School of Public Health
Laura Meyerson, Biogen Idec
Peter Mueller, The University of Texas MD Anderson Cancer Center
Daniel Scharfstein, Johns Hopkins University
Jeremy Taylor, University of Michigan
Heping Zhang, Yale University
Haibo Zhou, University of North Carolina At Chapel Hill

Distinguished Student Paper Award Winners

Van Ryzin Award Winner
Brian Hobbs, University of Minnesota

Award Winners
Ying Ding, University of Michigan
Hedlin Haley, Johns Hopkins University
Yijuan Hu, University of Minnesota
Xin Huang, Georgia State University
Shahedul Khan, University of Waterloo
Se Hee Kim, University of North Carolina
Seonjoo Lee, University of North Carolina at Chapel Hill
Sang Mee Lee, University of Minnesota
Pei Li, University of Minnesota
Xiaoyun Li, Florida State University
Miguel Marino, Harvard University
Hui Nie, University of Pennsylvania
Seo Young Park, University of North Carolina at Chapel Hill
Debdeep Pati, Duke University
Xinyu Tang, University of Pittsburgh
Jincao Wu, University of Michigan
Hui Zhang, University of Rochester
Rongmei Zhang, University of Pennsylvania
Hong Zhu, Johns Hopkins University

Visit the ENAR website (www.enar.org) for the most up to date source of information on ENAR activities.

Special Thanks

2010 ENAR Program Committee
Michael Daniels (Chair), University of Florida
Jeffrey Morris (Co-Chair), University of Texas MD Anderson Cancer Center

At-Large Members
Kimberly Drews, George Washington University
Joseph Hogan, Brown University
Eugene Huang, Emory University
Hongtu Zhu, University of North Carolina at Chapel Hill

ASA Section Representatives
Alyson Wilson (Bayesian Statistical Sciences Section), Iowa State University
Liang Li (Biometrics Section), Cleveland Clinic
Dionne Price (Biopharmaceutical Section), Food and Drug Administration
Ji Zhu (Statistical Learning and Data Mining Section), University of Michigan
Recai Yucel (Health Policy Statistics Section), State University New York at Albany
Elizabeth Stuart (Social Statistics Section), Johns Hopkins University
Paul Albert (Statistics in Epidemiology Section), National Institute of Child and Human Development
Constantine Daskalakis (Teaching of Statistics in Health Sciences Section), Thomas Jefferson University

IMS Program Chair
Marie Davidian (Chair), North Carolina State University
Helen Zhang (Co-Chair), North Carolina State University

ENAR Educational Advisory Committee
Thomas Braun, University of Michigan
Xihong Lin, Harvard University
Paul Rathouz, University of Chicago

Local Arrangements Committee
Brian Marx (Chair), Louisiana State University
Julia Volaufova, Louisiana State University Health Science Center

ENAR Student Awards Committee
Eric (Rocky) Feuer (Chair), National Cancer Institute

ENAR Diversity Workshop Committee
Renee Moore (Co-Chair), University of Pennsylvania
Adriana Perez (Co-Chair), University of Texas at Houston Health Science Center

ENAR Workshop for Junior Researchers Committee
Yi Li (Chair), Harvard University

Welcome
to New Orleans!

There is no doubt that New Orleans is one of the most exciting and culturally diverse places to travel. New Orleans, also known as the Crescent City, is most well known for the recent catastrophe of Hurricane Katrina, the birthplace of jazz, and the celebration of Mardi Gras every spring. Normally when tourists or first-time residents come to New Orleans, they have a difficult time understanding the culture – strong French and Spanish influences have created a truly European city inside the United States. They quickly learn that bars have no closing hour, that the food is full of flavor, and that the music is pulsating almost everywhere – all these famous attributes of the city give New Orleans a powerful sense of identity.

As a major birthplace of musical and theatrical talent, New Orleans has set a standard throughout history for the showcasing of fine art. Head on down to Jackson Square to see artists at work in front of your very eyes. Check out the galleries in the French Quarter, the Arts District, the French Market, and on Magazine Street. Also in the French Quarter, the **House of Blues** offers live music every night of the week. **The Contemporary Arts Center** is a home to bold experiments and the celebration of art. Opera, symphony orchestras, ballets, and theatre can also be found in abundance throughout the city.

New Orleans is world-famous for its food. The indigenous cuisine is distinctive and influential. From centuries of amalgamation of the local Creole, haute Creole, and New Orleans French cuisines, New Orleans food has developed. Local ingredients, French, Spanish, Italian, African, Native American, Cajun, and a hint of Cuban traditions combine to produce a truly unique and easily recognizable Louisiana flavor. Stop by the **Bourbon House** in the French Quarter or any sub shop around town and pick up a famous New Orleans submarine sandwich, the po-boy, made with delicious Louisiana French bread. Another traditional New Orleans favorite is found at **Café Du Monde** on Decatur Street. Here you can enjoy original café au lait with beignets – square pieces of dough that are fried and covered with powdered sugar.

French Quarter | The French Quarter is the heart of New Orleans, the original city – the atmosphere and old-world charm is legendary, yet hard to describe. Window-shop on Royal or Magazine streets, rich in antiques, or for entertainment, take in Bourbon Street for intensive exposure to jam-packed bars, restaurants and music outposts. Street vendors hawking sweet pralines add a festive air to Jackson Square where sidewalk artists show off their skills. Browse the treasure trove at the French Market and munch on a muffellata at Central Grocery. The French Quarter is also home to some of New Orleans historic homes and buildings. Don't miss **Le Petit Theatre** – this 87-year-old theatre is has been recognized as one of the leading "little" or community theatres in the nation, and **St. Louis Cathedral**, which is the oldest continuously active Roman Catholic Cathedral in the United States.

The New Orleans Museum of Art | New Orleans received a gift of lasting culture in early 1910, when sugar broker Isaac Delgado offered the city $150,000 to build a "temple of art for rich and poor alike" in City Park. Today, Delgado's 25,000-square-foot "temple" is still at the center of the now much larger New Orleans Museum of Art and houses a $200 million collection in 46 galleries. The world-class collection of 50 modern and contemporary sculpture is presented in an incredible, five-acre natural setting with delights at every turn.

St. Charles Avenue | St. Charles Avenue has been described most aptly as "The Jewel of America's Grand Avenues." It is, indisputably, the most superb collection of great mansions of the South. The Avenue offers to all an open opportunity to enjoy the lofty magnificence of true, gracious living from l9th century New Orleans. A ride on the infamous Saint Charles streetcar provides a unique way to enjoy the splendor of the Avenue, from the statuesque monument at **Lee Circle** to its end point in the old town of Carrollton upriver.

The National World War II Museum | This is a must-see for history lovers and all patriots! Located in the **Arts/Warehouse District**, The National WWII's exhibits encompass the June 6, 1944 invasion of Normandy, the Home Front during WWII, and the D-Day Invasions in the Pacific. Exhibit galleries incorporate text panels, artifacts, and Personal Account stations in which visitors may listen to the stories of WWII veterans and others who supported the war effort.

The Botanical Gardens | The New Orleans Botanical Garden at City Park offers a serene retreat from the hustle and bustle of urban life. Surrounded by the nation's largest collection of mature live oaks, patrons enjoy a sensual walk past 2,000 varieties of plants, theme gardens such as butterfly walk, rose, tropical, Japanese and train garden. The recently renovated Conservatory features a simulated tropical rainforest complete with hanging vines and a roaring waterfall. Perhaps the most amazing thing about this wonderful Mid-City living museum is that it is always changing – you'll never make the same visit twice!

ENAR Recommends

Hot Spots in the Big Easy

Sharon-Lise Normand (2010 ENAR President)

Mid-City Lanes Rock 'N Bowl

Bowl while listening to live music! Regional music include Cajun, zydeco (OK, I admit I don't know what this is), and rockabilly. What could be more fun than sporting those fancy bowling shoes and shirts! Mid-City Lanes Rock 'N Bowl; 4133 S Carrolton Avenue, New Orleans; 504-482-3133. Open noon to midnight Sunday – Wednesday and noon to 2:00am Thursday – Friday.

Maura Stokes (ENAR Secretary)

New Orleans House of Blues

A ten minute walk to the heart of the French Quarter delivers you to live blues and dining like you've never experienced! Gospel Brunch, Voodoo Garden, 298 pieces of folk art, and a booth dedicated to blues legend Clarence "Gatemouth" Brown – what more can you possibly want? Well, visit the company story and buy a t-shirt just like my longtime favorite!

New Orleans House of Blues; 225 Decatur St., New Orleans; 504-310-4999. Open 11:30 am to 1am Wednesday through Sunday; call for Monday and Tuesday.

See the line-up and make reservations closer to fun time: www.houseofblues.com/venues/clubvenues/neworleans/

Lance Waller (2010 ENAR Past President)

Traditional New Orleans Music

Stroll over to the Louisiana Music Factory (www.louisianamusicfactory.com) at 210 Decatur St (not far from Secretary Stokes' pick). With any luck you'll catch an in-store performance in this tiny store, but if not, load up on traditional jazz, blues, zydeco (buy some for President Normand!!!), and everything else. Once you have your fill of recorded music, catch one, two, or more sets at Preservation Hall (www.preservationhall.com, open 8-11pm every night) at 726 St. Peter Street. No smoking, live music, unforgettable experience. Just remember, you can't have too many recordings by Duke Dejan, Dr. John, or Louis Armstrong.

More Hot Spots in the Big Easy

Amy Herring (2009 ENAR RAB Chair)

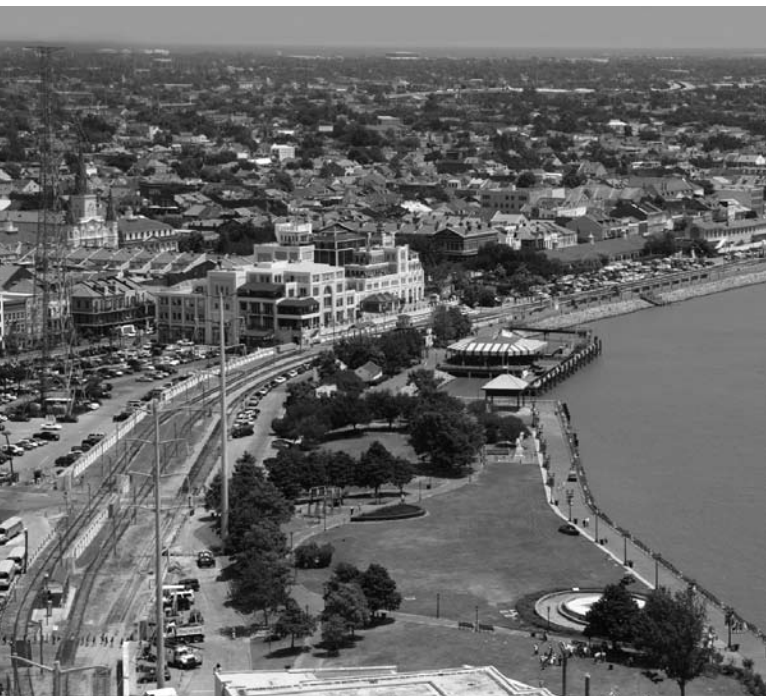
Audubon Aquarium of the Americas

My kid (and kid-at-heart!) friendly itinerary includes a trip to the Audubon Aquarium of the Americas (one block from the hotel) and the Audubon Zoo (especially the alligators and the Swamp Train). Taking a Mississippi River cruise on the Natchez steamboat is also sure to be a winner, though for those with more time I'd recommend renting a car and heading three hours north to its namesake town of Natchez, Mississippi for the famous Spring Pilgrimage. Natchez was once one of the wealthiest towns in the United States, and the spring pilgrimage includes tours of stunning antebellum homes, dozens of African-American heritage sites, and a variety of performances and pageants (www.natchezpilgrimage.com); the contrast between the splendors of Old South society and the social injustice at its core is somber and moving. Finally, only an hour and a half away from New Orleans is the Mississippi Gulf Coast, with beautiful beaches (Ocean Springs) and casino resorts (Biloxi).

Mike Daniels
(ENAR 2010 Spring Meeting Program Chair)

The Avenue Pub, 1732 St. Charles Avenue

Need to relax with a cold one after attending sessions and meetings all day? Then travel to The Avenue Pub in the Garden District for one of the best beer selections in New Orleans. From Abita Amber (brewed in LA) to Sierra Nevada to Harpoon Leviathan Imperial IPA, you will find a great selection both on tap and in the bottle. Don't want to walk? Take the Historic St Charles Avenue Street Car (a 5 minute walk from the hotel) and get off at stop #11. Cheers!



Jeffrey Morris
(2010 ENAR Spring Meeting Co-Chair)

New Orleans Ghost Tour and Mother's Po Boy Sandwich

New Orleans has a rich and storied history, and for those who like thrillers, you'll have a fun and unique experience going on a Ghost Walk. The Ghost Tour is offered by Haunted History Tours, and departs at 6pm and 8pm nightly at Rev. Zombie's Voodoo Shop at 723 St. Peter St. just a few blocks up from the Jackson Brewery (They suggest arriving at least 30 minutes before the tour). An entertaining tour guide will take you through the French Quarter and visit the sights of some of the documented hauntings and ghosts in New Orleans legend, and even visit a "haunted bar" along the way. The chilling yet fun tour lasts 2 hours. If you are in the mood for a quick sandwich before the tour, I highly recommend going to Mother's (401 Poydras Street near the Convention Center) and ordering "Famous Ferdi Special", an incredible sandwich consisting of Mother's world-famous baked ham, roast beef on a French roll, covered in what they call "debris and gravy." I am not even sure what "debris" is, but I can attest that it tastes incredible! The best sandwich I have ever had, and I make sure I get one every time I am in New Orleans. Even if you don't do the Ghost Tour, I highly recommend trying Mother's while you are in town!

Scarlett Bellamy (ENAR Treasurer)

Historic New Orleans Tours

Explore New Orleans' attractions and sights with the city's most qualified guides. Historic New Orleans Tours provides a variety of tours that will enhance your New Orleans vacation or visit, including walking tours and van tours alike! Try our van and cruise combo tour (the New Orleans Swamp Tour) and see a myriad of animals and creatures from alligators to snakes and crawfish on a cruise of Barataria-Terrebonne Estuary. See the devastation caused by hurricanes such as Hurricane Katrina and be inspired by New Orleans' rebirth and spirited recovery. Or take a general sightseeing tour of the city, and see neighborhoods from the famous French Quarter to the majestic St. Charles Avenue. Historic New Orleans Tours also offers walking tours of the French Quarter (Vieux Carre), the Garden District, St. Louis Cemetery #1, famous jazz locations, and even haunted sites! Learn about Voodoo culture, the history of music in the city, the architectural splendor of its neighborhoods, and movies that have been filmed here. These tours are some of the most fun things to do in New Orleans and are a great way to see and learn about Louisiana, whether you're a visitor or a resident.

I highly recommend the Garden District Tour. Highlights along the way include: Lafayette Cemetery #1; the homes of Anne Rice, Trent Reznor, Archie Manning (where Peyton and Eli grew up!), Nicolas Cage and John Goodman; the death site of Jefferson Davis; and the film site of Brad Pitt's "The Curious Case of Benjamin Button". For more info, visit: http://www.tourneworleans.com/index.html.

Hormuzd Katki
(Chair, ENAR Regional Advisory Board)
Mardi Gras World
Ever wonder where they make and keep all the Mardi Gras parade floats? Since 1947 they've been built and stored in Mardi Gras World. See a preview of the floats at http://www.flickr.com/photos/ekai/sets/72157621990139522/. Try on some Mardi Gras costumes and headdresses while you're there. Old floats get reused: extra points if you can spot the float of that was once of Mamie Eisenhower. Blaine Kern's Mardi Gras World; 1380 Port of New Orleans Place, New Orleans; 800/362-8213; Daily, 9:00 a.m.-4:30 p.m.; Adults, $17; children, $10; http://www.mardigrasworld.com

Mahlet Tadesse
(2009 ENAR Spring Meeting Program Co-Chair)
Garden District Walking Tour
Enjoy a walk through this historic and elegant neighborhood, known as the "American" section of New Orleans. The Garden District was created by Americans who moved to the area after the Louisiana Purchase of 1803. Stroll through the oak-tree lined streets and admire the varied architectural styles. If you are a fan of Anne Rice, you may want to view the writer's former residence, visit Lafayette Cemetery, which she frequently uses in her vampire books, and stop by the Garden District Book Shop, where she used to hold her book signings and signed editions can still be purchased. As an alternative to the many available escorted walking tours, the St. Charles Streetcar is an ideal way to take a self-guided tour of the Garden District and some of the other highlights of the city. If you plan to get on and off the streetcar, you may want to purchase a VisiTour pass ($1.25 for single fare, $5 for a one day pass).

Brian Marx (2010 Local Arrangements Chair)
Frenchman Street
Try something new and get away from the blaring music and tourists and explore the other side by taking a leisurely walk to the more local end of the French Quarter. Walk down either Chartres or Royal Streets, starting at Canal Street and continue clear to Esplanade Street. If you need a drink half way, you should stop by the historic Napoleon House at the corner of Chartes and Saint Louis, which offers an intimate and beautiful courtyard. Once you reach Esplande, cross it and continue outside the Quarter to the local neighborhood on Frenchman Street. Here you will find several local music venues, including the premier jazz club Snug Harbor, as well as several other local bars (with cheap or no cover), including The Spotted Cat and The Blue Nile. If you are waiting for a show to start or just passing time until dinner, shoot a simple game of pool at the R Bar (not named after the software) on the corner of Royal and Touro Streets- you are sure to feel local here. There is a little Thai restaurant a half block away. For dinner on the finer end, go a half block more to Marigny Brasserie at 640 Frenchman St. If you prefer the best burger in the world that is served with a baked potato, go to Port of Call bar a few steps away at 838 Esplanade Street. For Italian, just go back a few blocks into the quarter and go to Irene's Cuisine at the corner of Chartres and Saint Philip Streets (better to put your name on this list on your way to Frenchman). See you later in the evening at Café Du Monde for beignets and café au lait (Jackson Square at Decatur St.). Have some fun and enjoy New Orleans.

Kathy Hoskins (ENAR Executive Director)
Café Du Monde
Your visit to New Orleans (and the perfect ending to a night out in the Big Easy!) cannot be complete without a trip to Café Du Monde! While this famous New Orleans institution now has several locations – be sure to visit the original Café Du Monde in the French Market located at 800 Decatur Street – just a few blocks and an easy walk from the Hilton New Orleans Riverside Hotel. The Café is open 24 hours a day, seven days a week –so this makes the ideal ending to a night out in New Orleans. Great coffee and the best beignets in town! (If you are not familiar with beignets – they are square pieces of dough, fried and then covered in powdered sugar – so forget the calories and just enjoy!) If you have not visited this New Orleans landmark – it's time to stop by – grab a cup of coffee or café au lait, some warm beignets, and enjoy the flavor and history of New Orleans!

9

Presidential Invited Speaker

Bayes, BARS, and Brains: Statistics and Machine Learning in the Analysis of Neural Spike Train Data

Professor Robert Kass
Department of Statistics
Carnegie Mellon University

One of the most important techniques in learning about the functioning of the brain has involved examining neuronal activity in laboratory animals under varying experimental conditions. Neural information is represented and communicated through series of action potentials, or spike trains, and the central scientific issue in many studies concerns the physiological significance that should be attached to a particular neuron firing pattern in a particular part of the brain. In addition, a major comparatively new effort in neurophysiology involves the use of multielectrode recording, in which responses from dozens of neurons are recorded simultaneously. Among other things, this has made possible the construction of brain-controlled robotic devices, which could benefit people whose movement has been severely impaired.

In my talk I will briefly outline the progress made, by many people, over the past 10 years, highlighting some of the work my colleagues and I have contributed. The new methods of neural data analysis have resulted largely from adapting standard statistical approaches (additive models, state-space models, Bayesian inference, the bootstrap) to the context of neural spike trains, which may be considered point processes. Methods commonly associated with machine learning have also been applied. I will make some comments about the perspectives of statistics and machine learning, and will indicate current status and future challenges.

Biography
Rob Kass received his Ph.D. in Statistics from the University of Chicago in 1980. His early work formed the basis for his book Geometrical Foundations of Asymptotic Inference, co-authored with Paul Vos. His subsequent research has been in Bayesian inference and, most recently, in the application of statistics to neuroscience. Kass is known not only for his methodological contributions, but also for several major review articles, including one with Adrian Raftery on Bayes factors (JASA, 1995) one with Larry Wasserman on prior distributions (JASA, 1996), and a pair with Emery Brown on statistics in neuroscience (Nature Neuroscience, 2004, also with Partha Mitra; J. Neurophysiology, 2005, also with Valerie Ventura). Brown and Kass have recently attempted to stir debate about statistical education in an article entitled "What is Statistics?" (American Statistician, 2009)

Kass has served as Chair of the Section for Bayesian Statistical Science of the American Statistical Association, Chair of the Statistics Section of the American Association for the Advancement of Science, Executive Editor of the international review journal Statistical Science, and founding Editor-in-Chief of the journal Bayesian Analysis. 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 has been recognized by the Institute for Scientific Information as one of the 10 most highly cited researchers, 1995-2005, in the category of mathematics. In 1991 he began the series of workshops Case Studies in Bayesian Statistics, which are held at Carnegie Mellon every odd year, and was co-editor of the six proceedings volumes that were published by Springer. He is co-organizer of the workshop series Statistical Analysis of Neuronal Data, which began in 2002 and is held at Carnegie Mellon on even years. Kass has been been on the faculty of the Department of Statistics at Carnegie Mellon since 1981 and served as Department Head from 1995 to 2004; he joined the Center for the Neural Basis of Cognition in 1997, and the Machine Learning Department in 2007.

IMS Medallion Lecture

A Statistician's Adventures in Collaboration: Designing Better Treatment Strategies

Professor Marie Davidian
Department of Statistics
North Carolina State University

Over the past decade, there has been an increasing focus in a host of disease and disorder areas on not only traditional development and evaluation of new interventions but on elucidating more broadly the best ways to use both new and existing treatments. The recent emphasis by policymakers on comparative effectiveness research has only heightened this interest. Statisticians, mathematicians, and other quantitative scientists have developed various methodological approaches that can inform and guide this endeavor, particularly in the context of designing strategies for best using available treatment options over the entire course of a disease or disorder. Adopting the perspective that "treatment" of chronic diseases and disorders really involves a series of therapeutic decisions over time, each of which should ideally be made adaptively based on the evolving health status of the patient, holds the promise of developing treatment regimens that in this sense move toward "individualizing" treatment to the patient, thereby improving health outcomes. Multidisciplinary collaborations involving disease/disorder area specialists, statisticians, and other quantitative scientists to exploit these methodological advances are essential if this goal is to be achieved. I am fortunate to be involved in several collaborative research projects focused on development of such adaptive treatment strategies in areas ranging from HIV infection to alcohol abuse to organ transplantation. I will describe these exciting research projects and the statistical and mathematical methods that are being brought to bear to address this challenge.

Biography
Marie Davidian is William Neal Reynolds Professor of Statistics and Director of the Center for Quantitative Sciences in Biomedicine at North Carolina State University (NCSU). She received bachelors and master's degrees in applied mathematics in 1980 and 1981 from the University of Virginia and received a Ph.D. in statistics from the University of North Carolina at Chapel Hill in 1987 under the direction of Raymond J. Carroll. She joined the Department of Statistics at North Carolina State University in 1987 and returned in 1996 after serving on the faculty in the Department of Biostatistics at Harvard School of Public Health from 1994-1996. Dr. Davidian is an elected Fellow of the American Statistical Association (ASA), the Institute of Mathematical Statistics (IMS), and the American Association for the Advancement of Science (AAAS) and is an elected member of the International Statistical Institute (ISI). She has served as Coordinating and Executive Editor of Biometrics, as chair of the National Institutes of Health Biostatistical Methods and Research Design (BMRD) study section, on the International Biometric Society (IBS) and IMS Councils, and as ENAR president. She is the recipient of the G.W. Snedecor Award, the Janet L. Norwood Award for Outstanding Achievement by a Woman in the Statistical Sciences, the ASA Award for Outstanding Statistical Application, and several distinguished lectureships.

Short Courses

Date: Sunday, March 21, 2010

Full Day Fee:
Members $250 ($275 after 2/16)
Non-Members $300 ($320 after 2/16)

Half Day Fee:
Members $160 ($185 after 2/16)
Non-Members $200 ($225 after 2/16)

S1. A Practical Introduction to Bayesian Statistics

Room: Belle Chasse, 3rd Floor
Full Day: 8:00 am-5:00 pm
Instructor: Mark E. Glickman
Boston University School of Public Health

Description: Most statisticians' initial exposure to statistics is from the classical or "Frequentist" point of view. This comes as no surprise, as most statistical software encourages the nearly-exclusive use of classical statistical methods, and public health and medical journals tend to be much more accepting of the results of classical analyses. Exposure of practicing statisticians to Bayesian statistics has therefore been relatively limited. Few classically-trained statisticians appreciate, for example, that Bayesian statistics is not a collection of special tools or clever statistical models and procedures, but is actually a competing framework for statistical inference that is self-consistent, does not require inventing new methods when one encounters a difficult problem, and enables a statistician to prioritize thinking critically and scientifically about constructing appropriate data models rather than focus on the procedures to analyze the data. Recognizing some of the difficulties using classical procedures, a growing number of statisticians are beginning to understand the benefits of Bayesian statistics in part because, at long last, computational tools exist that makes Bayesian data analyses relatively straightforward to implement.

This short course is aimed at practicing statisticians who are comfortable using classical approaches to data modeling and analysis, but are curious to learn more about the details of Bayesian statistics. The course is equally appropriate as a refresher in Bayesian methods. No previous knowledge of Bayesian statistics is assumed. In the morning session, the fundamentals of the Bayesian framework for data analysis will be covered. This will include: the philosophical underpinnings of Bayesian statistics; Bayes rule; prior and posterior distributions; choice of a prior distribution; predictive distributions; summarizing inferences; sequential updating; model selection and Bayes factors; and basic applications. In the afternoon session, the course will concentrate on Monte Carlo simulation as a tool for summarizing the results of Bayesian analyses. Specifically, after briefly discussing basic Monte Carlo simulation, the bulk of the material will be on Markov chain Monte Carlo simulation as a standard method for fitting Bayesian models. The software package WinBUGS will be introduced and explained, and several example data analyses from health and medical applications will be performed. The examples in the afternoon session will involve more complex models than in the morning session, including hierarchical models. The course will emphasize interactive learning: In the afternoon session, in particular, participants will be asked to create and run their own WinBUGS programs with the instructor's guidance.

Prerequisites: Familiarity with classical statistics, regression models (least-squares and logistic regression), and basic probability. Participants are expected to bring their laptops with WinBUGS installed. Please visit the short course web site a few weeks prior to the course for instructions to download WinBUGS and for data files that will be used during the course.

NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website (www.enar.org) for instructions on download prior to the meeting.

S2. Statistical Methods for Analysis of High-Dimensional Data with Applications in Biosciences

Room: Versailles Ballroom, 3rd Floor
Full Day: 8:00 am-5:00 pm
Instructors: Tianxi Cai, Xihong Lin, and Armin Schwartzman, Department of Biostatistics, Harvard University

Description: High-dimensional data arise rapidly in bioscience research, especially in the current "omics" era. Analysis of such high-dimensional data presents daunting statistical challenges as well as exciting opportunities, as new statistical methods are needed to deal with high dimensional variables such as genes, proteins, images, medical records, and limited sample sizes. Analysis of high dimensional data often involves multiple stages: a discovery stage where a large number of variables are explored; a model building stage where a statistical model, such as a regression model, classification rule, or a prediction model, is constructed; a validation stage, where the findings are validated statistically or experimentally.

In this course, we will discuss recently developed statistical tools for analysis of high-dimensional bioscience data. We will start with basic ideas and methods, and move into more advanced methods. Comparisons of different methods will be discussed. Data examples from various bioscience applications will be used for illustration. Practical implementation of these methods using R will be discussed. The course will balance theory and method, applications and practice, and use of software.

Topics of this course include (i) testing a large number of hypotheses, family-wise error rate, false discovery rate control and estimation; (ii) variable selection and regularized regression methods, such as L0, LASSO, Adaptive LASSO, SCAD and Dantzig Selector; (iii) Classification methods, such as Bayes rule, discriminant analysis, support vector machines and kernel machines; (iv) Prediction and validation, e.g., prediction rules, prediction accuracy, ROC, AIC, BIC, CV.

Prerequisites: Familiarity with statistical linear models, multivariate regression and matrix algebra; familiarity with R. Please visit the short course website a few weeks prior to the course for R scripts and data that will be used for illustration.

NOTE: Wireless access will not be available in the classroom and participants should check the ENAR website (www.enar.org) for instructions on download prior to the meeting.

S3. Analyzing Complex Survey Data

Room: Elmwood, 3rd Floor
Full Day: 8:00 am-5:00 pm
Instructor: Michael Elliot, Department of Biostatistics,
University of Michigan School of Public Health

Description: Health surveys are now used to fit a large variety of statistical models to describe complex associations between health outcomes and a wide variety of clinical, behavioral, and social risk factors. This session will discuss the impact of survey sample design on the inferences obtained under these models, with a particular focus on the use of sample weights to account for model misspecification or non-ignorable sampling. Issue such as model fit and model checking will be addressed as well. Applications will focus on regression models - linear, generalized linear, and survival models in particular - although the general principles discussed can be applied in many modeling settings.

Prerequisites: Familiarity with linear and generalized linear models will be assumed. An introductory course in survey sampling will be very helpful, although this material will be briefly reviewed at the beginning of the course.

S4. Bioinformatics

Room: Oak Alley, 3rd Floor
Full Day: 8:00 am-5:00 pm
Instructor: Jun Liu, Department of Statistics
Harvard University

Description: The sheer amount and variety of the molecular biology data have already presented a major challenge to the scientific world. Analysis of these data using bioinformatics tools has played a key role in several recent advances and will play increasingly important roles in future biomedical researches. A distinctive feature of these data, be they microarray images, DNA sequences or protein structures, is that there is a large body of biological knowledge associated with them. This makes standard off-the-shelf data mining or statistical analysis tools less effective. Incorporating relevant scientific knowledge into the development of statistical or computational analysis tools is the key to success. This tutorial is intended to provide coverage of some key developments of bioinformatics in the past thirty years with an emphasis on topics of recent interest. Topics include: pair-wise sequence analysis, local alignment, dynamic programming, BLAST, multiple sequence alignment, Gibbs motif sampler, gene regulation, hidden Markov models, epigenetics, protein structure analysis, comparative genomics, model-based microarray analysis, clustering methods for microarrays, phylogenetic trees, etc.

Prerequisites: First-year graduate-level courses on probability and statistics. Course materials will be provided.

S5. Metabolomics

Room: Melrose, 3rd Floor
Half-Day: 8:00 am - noon
Instructor: David Banks
Department of Statistical Science, Duke University

Description: Metabolomics is a new area in bioinformatics that uses estimates of metabolite abundance to determine disease status. Although similar in many ways to proteomics, it provides both fresh statistical challenges and important scientific opportunities. The main challenges are (1) to make strong use of the additional information that is available from our greater knowledge of the human metabolome (as compared to the proteome); (2) to develop data mining procedures tuned to this kind of data; and (3) to create experimental designs that support cross-platform experiments. But the scientific opportunities are commensurate - in particular, the domain knowledge that exists on the chemical structure of metabolites and the pathways that produce them can eliminate much of the uncertainty that arises in proteomics. This half-day short course summarizes current work on all three challenges, as well as describing issues in quality control for metabolomics data. It includes detailed discussion of two case studies.

Prerequisites: This course assumes applied knowledge of the linear model and statistical inference (at about the level of a recent M.S. degree), and a degree of comfort with high school biochemistry.

S6. Adaptive Randomization for Phase III Studies

Room: Melrose, 3rd Floor
Half-Day: 1:00 pm-5:00pm
Instructor: Scott Emerson, University of Washington

Description: Recently much attention has been devoted to the re-design of ongoing randomized clinical trials based on interim results. One such area of interest is the use of interim trial results to determine the randomization of future subjects. In this short course we will review such methods as “covariate adaptive” randomization, “play-the-winner” randomization schemes, and adaptively dropping treatment arms or subgroups. We will focus on the ability of these methods to better address the efficiency and ethical issues inherent in randomized clinical trials. We will also discuss the special considerations that must be made at the time of study design, during conduct of the clinical trial, and when reporting results, and how those special considerations might impinge on gaining regulatory approval for a new drug, biologic, or device.

Prerequisites: A one year course in general biostatistical methods and familiarity with the common setting of randomized clinical trials.

Tutorials

Dates: Monday, March 22 and Tuesday, March 23, 2010

T1. Bayesian Computation in SAS

Room: Jefferson Ballroom, 3rd Floor
Monday, March 22, 8:30-10:15 am
Instructor: Joseph Ibrahim, University of Chapel Hill

Description: Survival analysis arises in many fields of study such as medicine, biology, engineering, public health, epidemiology, and economics. SAS has recently developed several procedures for fitting Bayesian models in SAS through their procedures LIFEREG, PHREG, GENMOD, and MCMC. This tutorial provides a comprehensive treatment of these four SAS procedures. Several topics are addressed, including generalized linear models, parametric and semiparametric survival models, and more general model structures such as random effects models, models for longitudinal data, missing data, and elicitation of informative prior distributions. Convergence diagnostics and model assessment tools will also be discussed, along with a general discussion of Markov chain Monte Carlo methods for Bayesian inference. Extensive examples and case studies will be presented, along with the SAS code and the output.

Prerequisites: Good working knowledge of SAS, a previous course in survival analysis, a previous course in Bayesian statistics.

T2. Comparative Effectiveness Research: An Introduction for Statisticians

Room: Jefferson Ballroom, 3rd Floor
Monday, March 22, 10:30 am-12:15 pm
Instructors: Constantine Gatsonis, Brown University

Description: Comparative Effectiveness Research (CER) is now a major initiative in the US, with wide ranging implications for both research and health care policy. The term CER is broadly used to refer to a body of research that generates and synthesizes evidence on the comparison of benefits and harms of alternative methods to prevent, diagnose, treat, and monitor clinical conditions, or to improve the delivery of care. The evidence from Comparative Effectiveness Research is intended to support clinical and policy decision making at both the individual and the population level. The mandate of CER places a premium on the study of outcomes that are of primary relevance to patients and on the derivation of conclusions that can inform individual patient choices.

The broad scope of CER requires a wide array of methodologic approaches. CER research may include both randomized and observational primary studies as well as research synthesis. In this tutorial we will provide an overview of the types of research questions addressed by CER and will survey the main areas of statistical methodology that are currently in use. We will also highlight limitations of current methods and discuss important open problems.

Prerequisites: This tutorial is intended for a broad audience of statisticians. Some familiarity with health services and outcomes research, and research synthesis studies is useful but not essential.

T3. SWEAVE

Room: Jefferson Ballroom, 3rd Floor
Monday, March 22, 1:45-3:30 pm
Instructor: Frank E. Harrell, Jr., Vanderbilt University

Description: Much of research that uses data analysis is not easily reproducible. This can be for a variety of reasons related to tweaking of instrumentation, the use of poorly studied high-dimensional feature selection algorithms, programming errors, lack of adequate documentation of what was done, too much copy and paste of results into manuscripts, and the use of spreadsheets and other interactive data manipulation and analysis tools that do not provide a usable audit trail of how results were obtained. Even when a research journal allows the authors the "luxury" of having space to describe their methods, such text can never be specific enough for readers to exactly reproduce what was done. All too often, the authors themselves are not able to reproduce their own results. Being able to reproduce an entire report or manuscript by issuing a single operating system command when any element of the data change, the statistical computing system is updated, graphics engines are improved, or the approach to analysis is improved, is also a major time saver.

It has been said that the analysis code provides the ultimate documentation of the "what, when, and how" for data analyses. Eminent computer scientist Donald Knuth invented literate programming in 1984 to provide programmers with the ability to mix code with documentation in the same file, with "pretty printing" customized to each. Lamport's LaTeX, an offshoot of Knuth's TeX typesetting system, became a prime tool for printing beautiful program documentation and manuals. When Friedrich Leisch developed Sweave in 2002, Knuth's literate programming model exploded onto the statistical computing scene with a highly functional and easy to use coding standard using R and LaTeX and for which the Emacs text editor has special dual editing modes using ESS. This approach has now been extended to other computing systems and to word processors. Using R with LaTeX to construct reproducible statistical reports remains the most flexible approach and yields the most beautiful reports, while using only free software. One of the advantages of this platform is that there are many high-level R functions for producing LaTeX markup code directly, and the output of these functions are easily directly to the LaTeX output stream created by Sweave.

This tutorial covers the basics of Sweave and shows how to enhance the default output in various ways by using: latex methods for converting R objects to LaTeX markup, your own floating figure environments, the LaTeX listings package to pretty-print R code and its output, and the R tikzDevice package of Cameron Bracken and colleagues to make full use of LaTeX fonts and typesetting inside R graphics objects. These methods apply to everyday statistical reports and to the production of 'live' journal articles and books.

Prerequisites: There are few prerequisites but those attendees who have used R, S-Plus, or LaTeX will digest the examples more quickly. See reproducibleresearch.net and biostat.mc.vanderbilt.edu/SweaveLatex for more information.

T4. Statistical Challenges in Genome-wide Association Studies

Room: Jefferson Ballroom, 3rd Floor
Tuesday, March 23, 8:30-10:15 am
Instructor: Sebastian Zöllner, University of Michigan

Description: In the last 4 years, genome-wide association studies have been widely successful, identifying >800 loci for complex phenotypes. Further studies are ongoing and promise to elucidate the molecular basis of many diseases. In this tutorial, we will review statistical challenges of these study designs such as developing quality-control filters for the genotype data, controlling for confounding due to population stratification, and the multiple comparisons problems of performing up to 3 million association tests. Designs have to account for the generally low effect sizes of common genetic variants. Hence enormous sample sizes are required to provide adequate power; usually such sample sizes are acquired by combining multiple large studies. We will describe methods imputation methods that enable such combined analyses and commonly used approaches for meta-analyses.

T5. Likelihood Methods for Measuring Statistical Evidence

Room: Jefferson Ballroom, 3rd Floor
Tuesday, March 23, 1:15-3:30 pm
Instructors: Jeffrey Blume and Gregory D. (Dan) Ayers
Vanderbilt University

Description: Likelihood methods for measuring statistical evidence have evolved slowly since they were first introduced by R.A. Fisher in the early 1920's. Nearly a century later the Likelihood paradigm has matured enough to warrant careful consideration. Likelihood methods are the natural compromise between Bayesian and frequentists approaches; they retain desirable properties of both paradigms (irrelevance of sample space, good performance probabilities) while shedding undesirable ones (dependence on prior distributions, ad-hoc adjustments to control error probabilities).

This tutorial will present the fundamentals of likelihood methods for measuring statistical evidence. We will only briefly touch upon its lively history and philosophical components. Instead, the focus will be on illustrating these methods in real-world examples, understanding their operational characteristics, and discussing recent advancements, e.g., robustness. The goal is to generate familiarity with how these methods work and when they might be used. We will discuss the likelihood approach to representing evidence, designing studies, re-examining accumulating evidence during the course of a clinical trial (multiple looks problem), dealing with multiple endpoints (multiple comparisons problem), non-inferiority trials and small scale Phase I dose escalation studies. Throughout these examples we will highlight the Likelihood paradigm's exceptional flexibility, efficiency, and accuracy.

Prerequisites: General understanding of statistical inference and clinical trials. Knowledge of R or S-plus helpful but not required.

Roundtables

Monday, March 22
12:15 - 1:30 pm
Napolean Ballroom

R1. Exploring Roads to Successful Publishing

Discussion Leader: Joel Greenhouse*
(*Statistics in Medicine Editor), Carnegie Mellon University

Description: Dissemination of research results through publication remains the primary venue for communication in the statistical sciences. In this roundtable, Joel Greenhouse, a co-editor of Statistics in Medicine and past editor of the IMS Lecture Note and Monograph Series, will describe the review and publication process from the editor side, and will facilitate a discussion on how to improve your chances of a successful submission. This roundtable will focus on issues of most concern to new researchers.

R2. How To Publish and Flourish

Discussion Leader: Anastasios (Butch) Tsiatis, North Carolina State University

Description: Publishing in the peer reviewed literature is necessary for academic advancement and an obligation to society. It is challenging for the experienced researcher; daunting for those early in their career. To shed some light and possibly generate a little heat, we will discuss a wide range of topics related to scientific publishing. Topics include identifying your audience, choosing a target journal, aiming your manuscript at your target, dealing with journal decisions, and components of effective writing. We'll discuss the benefits and possible drawbacks of "aiming high" and outline the review process at Biometrics. Participants should be ready to pose questions and to discuss their experiences, both positive and negative. All should leave the roundtable well nourished with improved understanding of the publication process, of how to maximize publication success and how to build on rejections to produce subsequent successes.

R3. Strategies for Publishing a Book

Discussion Leader: John Kimmel, Springer Science+Business Media, LLC

Description: The discussion will concentrate on publishing issues of interest to the participants. Possible topics include when in your career you should write a book and what are the benefits; authored vs. edited books and monographs vs. textbooks; how long writing a book takes, how book publishers differ; what publishers want to see in a proposal; what you should ask for from a publisher; what contractual clauses are important and which ones you should beware of; how types of royalties are computed; and changes coming soon such as print on demand and e-books.

R4. Innovative Statistical Methods for Clinical Trials

Discussion Leader: Marie Davidian, North Carolina State University

Description: Randomized clinical trials are the gold standard for evaluation of the efficacy and effectiveness of new and existing therapeutic interventions, and statistical principles are fundamental to their design and analysis. New methodological advances and new perspectives on how treatments should be evaluated offer opportunities to enhance the quality of inferences and what can be learned from clinical trials. Some examples include new methods for covariate adjustment that can improve precision, the use of mathematical and statistical modeling and simulation for trial design, adaptive designs, and sequential multiple assignment randomized trials (SMARTs) for making inference on time-dependent treatment strategies (dynamic treatment regimes). This roundtable will provide a forum for discussing these and other innovative methods for trial design and analysis and for brainstorming about approaches for integrating new advances into the clinical trial paradigm and for promoting their acceptance by trialists, clinicians, and regulatory authorities.

R5. Statistical Challenges in Genomics Data Analysis

Discussion Leader: Hongyu Zhao, Yale University

Description: Much progress has been made in the past 10 years on numerous genomics technologies so that researchers now can study biological systems at the genome level from different perspectives. The large and diverse types of genomics data pose many statistical and computational challenges, and have stimulated vigorous methodology and theory developments in the statistics community. Genomics research has proved to be a fertile field for statisticians reflected by many papers published recently in top statistics journals that were motivated from the analysis and interpretation of genomics data. In this roundtable, we will discuss what has been accomplished, the many gaps remaining between biological questions to be addressed and the available statistical tools, new technologies that demand even greater statistical efforts, and collaborative and grant opportunities in genomics studies.

R6. FDA Statistics: Innovations and Opportunities for Biostatisticians

Discussion Leader: Greg Campbell, Food and Drug Administration

Description: The U.S. Food and Drug Administration employs a large number of statisticians in a variety of capacities and there will likely continue to be numerous opportunities for future employment as well. One of the main responsibilities of many FDA statisticians involves the review of study plans before trials begin and then the subsequent results from clinical studies that companies perform to demonstrate that their therapeutic or diagnostic medical products are safe and effective prior to marketing in the US. FDA statisticians are often involved in cutting-edge and innovative statistical techniques. Examples that will be discussed include Bayesian statistics, adaptive designs, microarrays, personalized medicine, evaluation of diagnostic medical tests and imaging systems, surrogate endpoints and post-market issues.

R7. Dynamic Treatment Regimes: Problems and Challenges

Discussion Leader: Susan Murphy, University of Michigan

Description: In many areas of health (mental health, cancer, HIV infection, obesity, substance abuse, etc.), clinicians must consider treatment strategies that require answers to questions such as which treatment to provide initially, when to stop the initial treatment, which treatment to provide second and so on. Statisticians are playing an ever more important role in how to design clinical trials to collect high quality data and in how to use both clinical trial data and longitudinal studies to address these sequencing and timing questions. In this roundtable discussion we will discuss these issues and how these issues raise exciting challenges and provide open problems for our field.

R8. Opportunities for Funding of Statistical Methodology and Software at NIH

Discussion Leader: Eric J. (Rocky) Feuer, National Cancer Institute, and Michelle Dunn, National Cancer Institute

Description: At this roundtable the grant review process at NIH will be discussed, especially mechanisms that are appropriate for newer investigators. In addition, the facilitator will share the types of grants that are currently funded. Participants will be encouraged to share experiences, both positive and negative, in the application, review, and conduct of NIH grants.

R9. Quantitative Methods for HIV Prevention/ Control

Discussion Leader: Victor De Gruttola, Harvard School of Public Health

Description: Treatment for HIV has attained a high level of sophistication, including the development of over 20 approved drugs in drug classes as well as methods for monitoring treatment and for selecting salvage regimens. However prevention of HIV has lagged behind, and the incidence remains stubbornly high in risk groups throughout the world. Current focus for prevention is on combining different modalities, including treatment, prophyaxis (for those at high risk), circumcision and behavioral modifications. Developing an appropriate package of interventions for investigation for some target population, and determining the best design for such a study is a considerable challenge; to address it requires integration of a variety of disciplines, especially in those involving quantitative methods. These include network and epidemic modeling as well as epidemiology and biostatistics. The topic of this roundtable will be how to identify the most important problems that experts in these fields might work on, and how to facilitate communication among such experts to facilitate progress in this cross-disciplinary research.

R10. Opportunities for Biostatisticians in Pharmaceutical Companies

Discussion Leader: Stacy Lindborg, Eli Lilly

Description: The pharmaceutical industry will experience unprecedented patent loss over the next decade. A recent report estimated that more than $209 billion in annual sales were at risk to low cost generic substitution across the pharmaceutical industry between 2010 - 2014, which realistically will result in a lower investment in R&D for future drugs. At the same time for a host of reasons, the number of truly innovative new medicines approved by the FDA and other major regulatory bodies around the world has continued to decrease (50% fewer NME's approved compared to the previous 5 years). I believe that statisticians in the pharmaceutical industry have an increasingly important and strategic role to play in determining the future of individual companies and more broadly across the industry. During this discussion we will focus on topics such as statistical contributions to the goal of better selecting molecules early in development, advancing innovative trial designs, and understanding tailored therapies. The ability to live up to the potential our discipline provides will be influenced by technical skills in addition to leadership and influence. We will use this lunch to discuss opportunities, challenges, and areas of richest focus.

R11. Opportunities in Environmental and Climate Change Research

Discussion Leader: Andrew Finley, Michigan State University

Description: With an astonishing proliferation of spatial-temporal datasets and need for analysis, there are unprecedented opportunities for the statistical community to contribute to the interrelated fields of climate change, environmental inventory and monitoring, and natural resources protection, utilization, and management. This round table is being proposed to provide a venue to identify and discuss statistical challenges within, and at the interface of, these and related fields. Within these fields, we encounter a host of exciting theoretical and applied modeling and computing challenges including accommodation of space-time structures, nonstationarity in space and time, and multivariate structures that are not temporally separable. Further, the datasets used in these fields are often massive, with missingness in response and predictor variables, and variable misalignment and change of support. The round table is open to participants currently working in these and related fields, those who are interested in contributing their expertise from other fields, and junior investigators seeking collaborative opportunities for interdisciplinary, methodological, and theoretical research.

Program Summary

SATURDAY, MARCH 20

9:00 a.m.—9:00 p.m. **Workshop for Junior Researchers** *Sterling Room (Riverside Building)*

3:30 p.m. - 5:30 p.m. **Conference Registration** *3rd Floor Foyer*

SUNDAY, MARCH 21

7:30 a.m.—6:30 p.m. **Conference Registration** *3rd Floor Foyer*

8:00 a.m.—12:00 p.m. **Short Courses**
SC5: Metabolomics *Melrose Room (3rd Floor)*

8:00 a.m.—5:00 p.m. **Short Courses**
SC1: A Practical Introduction to Bayesian Statistics *Belle Chasse Room (3rd Floor)*
SC2: Statistical Methods for Analysis of High-Dimensional Data with Applications in Biosciences *Versailles Ballroom (3rd Floor)*
SC3: Analyzing Complex Survey Data *Elmwood Room (3rd Floor)*
SC4: Bioinformatics *Oak Alley Room (3rd Floor)*

12:30 p.m. – 5:00 p.m. **Diversity Workshop** *Rosedown Room (3rd Floor)*

1:00 p.m.—5:00 p.m. **Short Course**
SC6: Adaptive Randomization for Phase III Studies *Melrose Room (3rd Floor)*

3:00 p.m.—5:00 p.m. **Exhibits Open** *Court Assembly Foyer (3rd Floor)*

4:00 p.m.—7:00 p.m. **ENAR Executive Committee (By Invitation Only)** *Pelican Room (Riverside Building)*

4:30 p.m.—6:30 p.m. **Placement Service** *Windsor/Ascot Rooms (3rd Floor)*

7:30 p.m. – 8:00 p.m. **New Member Reception** *Napoleon Ballroom (3rd Floor)*

8:00 p.m.—11:00 p.m. **Social Mixer and Poster Session** *Napoleon Ballroom (3rd Floor)*
1. Posters: Clinical Trials
2. Posters: Survival Analysis I : Methodolgy
3. Posters: Survival Analysis II: Applications and Methodology
4. Posters: Statistical Genetics I
5. Posters: Statistical Genetics II
6. Posters: Imaging and Spatial Modeling
7. Posters: Genomics and Biomarkers
8. Posters: Longitudinal Data Analysis
9. Posters: Spatial/Temporal Modeling and Environmental/Ecological Applications
10. Posters: Applications and Case Studies I
11. Posters: Power/Sample Size
12. Posters: Nonparametric Methods
13. Posters: Missing Data and Measurement Error
14. Posters: Statistical Methods
15. Posters: Survey Research and Categorical Data Analysis
16. Posters: Biologics, Pharmaceuticals and Medical Devices

MONDAY, MARCH 22

7:15 a.m.—8:15 a.m. **Open Forum – Revitalization of Biostatistical Grant Applications at NIH**
Melrose Room (3rd Floor)

7:30 a.m.—8:30 a.m. **Student Breakfast** *Napoleon Ballroom (3rd Floor)*

7:30 a.m.—5:00 p.m. **Conference Registration** *3rd Floor Foyer*

8:30 a.m.—5:00 p.m. **Exhibits Open** *Court Assembly Foyer (3rd Floor)*

8:30 a.m.—10:15 a.m. **Tutorial 1: Bayesian Computation in SAS** *Jefferson Ballroom (3rd Floor)*

8:30 a.m.—10:15 a.m. **Scientific Program**

17. Statistical Methods for Estimating the Public Health Impact of Policy Interventions *Melrose Room (3rd Floor)*
18. Advances in Time Series Analysis of Neurological Studies *Rosedown Room (3rd Floor)*
19. Inference and Prediction for Sparse Networks *Magnolia Room (3rd Floor)*
20. Recent Developments in High Dimensional Inference *Jasperwood Room (3rd Floor)*
21. Contributed Papers: Image Analysis *Chequers Room (2nd Floor)*
22. Contributed Papers: Missing Data in Longitudinal Studies *Oak Alley Room (3rd Floor)*
23. Contributed Papers: Survival Analysis *Eglinton Winton Room (2nd Floor)*
24. Contributed Papers: Microarrays: Differential Expression and Reproductibility *Elwood Room (3rd Floor)*
25. Contributed Papers: Stephen Lagakos: Views from Former Students Spanning 2.5 Decades *Belle Chasse Room (3rd Floor)*
26. Contributed Papers: Adaptive Clinical Trial Designs for Dose Finding *Cambridge Room (2nd Floor)*

9:30 a.m.—5:00 p.m. **Placement Service** *Windsor/Ascot Rooms (3rd Floor)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *Court Assembly Foyer (3rd Floor)*

10:30 a.m.—12:15 p.m. **Tutorial 2: Comparative Effectiveness Research: An Introduction for Statisticians** *Jefferson Ballroom (3rd Floor)*

Scientific Program

27. Bayesian Methods for Combining Data from Multiple Sources for Clinical Trials *Melrose Room (3rd Floor)*
28. Competing Risks in Action *Rosedown Room (3rd Floor)*
29. Studying Genetic and Environmental Risk Factors of Complex Human Disorders and Their Interactions *Magnolia Room (3rd Floor)*
30. Contributed Papers: Spatial/temporal: Modeling and Methodology *Jasperwood Room (3rd Floor)*
31. Contributed Papers: Pathway and Network-Based Genomic Analysis *Oak Alley Room (3rd Floor)*
32. Contributed Papers: Causal Inference: Methodology *Elmwood Room (2nd Floor)*
33. Contributed Papers: Epidemiological Methods *Chequers Room (2nd Floor)*
34. Contributed Papers: Categorical Data Analysis *Eglinton Winton Room (2nd Floor)*
35. Contributed Papers: Adaptive Clinical Trial Design *Belle Chasse Room (3rd Floor)*
36. Contributed Papers: ROC Analysis *Cambridge Room (2nd Floor)*

12:15 p.m.—1:30 p.m. **Roundtable Luncheons** *Napoleon Ballroom (3rd Floor)*

12:30 p.m.—4:30 p.m. **Regional Advisory Board (RAB) Luncheon Meeting (By Invitation Only)** *Prince of Wales Room (2nd Floor)*

1:45 p.m.—3:30 p.m. **Tutorial 3: SWEAVE** *Jefferson Ballroom (3rd Floor)*

Scientific Program

37. Comparative Effectiveness Research: The Methodologic Challenges *Melrose Room (3rd Floor)*
38. Advances/Challenges in Jointly Modeling Multivariate Longitudinal Measurements and Time-to-Event Data *Rosedown Room (3rd Floor)*
39. Statistical Models and Practice that Improve Reproducibility in Genomics Research *Magnolia Room (3rd Floor)*
40. Dynamic Treatment Regimes and Reinforcement Learning in Clinical Trials *Jasperwood Room (3rd Floor)*
41. Contributed Papers: Multivariate Analysis *Chequers Room (2nd Floor)*
42. Contributed Papers: Biopharmaceutical Methods *Oak Alley Room (3rd Floor)*
43. Contributed Papers: Clinical Trials with Adaptive Sample Sizes *Newberry Room (3rd Floor)*
44. Contributed Papers: Causal Inference: Methodology and Applications *Eglinton Winton Room (2nd Floor)*
45. Contributed Papers: Survival Analysis Methodology *Cambridge Room (2nd Floor)*
46. Contributed Papers: Statistical Genetics: Complex Diseases and Linkage Analysis *Elmwood Room (3rd Floor)*
47. Contributed Papers: Functional Data Analysis *Belle Chasse Room (3rd Floor)*

3:30 p.m.—3:45 p.m. **Refreshment Break and Visit the Exhibitors** *Court Assembly Foyer (3rd Floor)*

3:45 p.m.—5:30 p.m. **Scientific Program**

48. Celebrating 70: The Contributions of Donald A. Berry to Statistical Science and Statistical Practice in Academics, Industry and Government *Versailles Ballroom (3rd Floor)*
49. Current Issues in Statistical Proteomics *Melrose Room (3rd Floor)*

50. Survey of Methodologies for Population Analysis in Public Health *Rosedown Room (3rd Floor)*
51. Prediction and Model Selection with Applications in Biomedicine *Magnolia Room (3rd Floor)*
52. Contributed Papers: Missing Data *Jasperwood Room (3rd Floor)*
53. Contributed Papers: Inference for Clinical Trials *Chequers Room (2nd Floor)*
54. Contributed Papers: Methods for Image Data and Time Series *Eglinton Winton Room (2nd Floor)*
55. Contributed Papers: Survival Analysis: Cure Models and Competing Risks *Oak Alley Room (3rd Floor)*
56. Contributed Papers: Statistical Genetics: Quantitative Trait Loci *Elmwood Room (3rd Floor)*
57. Contributed Papers: Microarrays: Time Course and Gene Sets *Belle Chasse Room (3rd Floor)*
58. Contributed Papers: Experimental Design, Power/Sample Size, and Survey Research *Cambridge Room (2nd Floor)*

6:00 p.m.—7:30 p.m. **President's Reception (By Invitation Only)** *River Room (Riverside Building)*

TUESDAY, MARCH 23

7:30 a.m.—5:00 p.m. **Conference Registration** *3rd Floor Foyer*

8:30 a.m.—10:15 a.m. **Tutorial 4: Statistical Challenges in Genome-wide Association Studies** *Jefferson Ballroom (3rd Floor)*

8:30 a.m.—5:00 p.m. **Exhibits Open** *Court Assembly Foyer (3rd Floor)*

9:30 a.m.—3:30 p.m. **Placement Service** *Windsor/Ascot Rooms (3rd Floor)*

8:30 a.m.—10:15 a.m. **Scientific Program**

59. Combining Multiple Sources of Exposure Data in Health Environment Studies *Melrose Room (3rd Floor)*
60. Imagining 2025: The Hopes for Clinical Trials *Versailles Ballroom (3rd Floor)*
61. Innovative Methods in Biosurveillance: Accurate Methods for Rapid Detection of Events and Patterns *Rosedown Room (3rd Floor)*
62. Innovative Statistical Methods for Functional and Image Data *Magnolia Room (3rd Floor)*
63. Contributed Papers: Joint Models for Longitudinal and Survival Data *Cambridge Room (2nd Floor)*
64. Contributed Papers: Health Services Research *Eglinton Winton Room (2nd Floor)*
65. Contributed Papers: Models Involving Latent Variables *Jasperwood Room (3rd Floor)*
66. Contributed Papers: Next Generation Sequencing and Transcription Factor Binding Sites *Oak Alley Room (3rd Floor)*
67. Contributed Papers: Variable Selection for High-Dimensional Data *Elmwood Room (2nd Floor)*
68. Contributed Papers: Quantile Regression and Symbolic Regression *Chequers Room (2nd Floor)*
69. Contributed Papers: Personalized Therapy and Variable Selection in Clinical Applications *Belle Chasse Room (3rd Floor)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *Court Assembly Foyer (3rd Floor)*

10:30 a.m.—12:15 p.m. **70. Presidential Invited Address** *Napoleon Ballroom (3rd Floor)*

12:30 p.m.—4:30 p.m. **Regional Committee Meeting (RECOM) (By Invitation Only)** *Prince of Wales Room (2nd Floor)*

1:15 p.m.—3:30 p.m. **Tutorial 5: Likelihood Methods for Measuring Statistical Evidence** *Jefferson Ballroom (3rd Floor)*

1:45 p.m.—3:30 p.m. **Scientific Program**

71. Bayesian Methods in Genomic Research *Melrose Room (3rd Floor)*
72. Interference and Spillover Effects in Causal Inference *Rosedown Room (3rd Floor)*
73. Statistical Methods in Neuroscience *Magnolia Room (3rd Floor)*
74. Recent Advances in Variable Selection Methodology *Versailles Ballroom (3rd Floor)*
75. Contributed Papers: Biomarkers and Diagnostic Tests *Jasperwood Room (3rd Floor)*
76. Contributed Papers: Survival Analysis in Clinical Trials *Cambridge Room (2nd Floor)*
77. Contributed Papers: High Dimensional Modeling: Segment Detection and Clustering *Eglinton Wintom Room (2nd Floor)*
78. Contributed Papers: Longitudinal Data Analysis *Oak Alley Room (3rd Floor)*
79. Contributed Papers: Statistical Genetics: Epistasis, Gene-Gene and Gene-Environment Interactions *Elmwood Room (2nd Floor)*
80. Contributed Papers: Bayesian Methods and Applications *Belle Chasse Room (3rd Floor)*
81. Contributed Papers: Spatial/temporal Applications, and Infectious Disease Modeling *Chequers Room (2nd Floor)*

Program Summary

3:30 p.m.—3:45 p.m. **Refreshment Break and Visit the Exhibitors** *Court Assembly Foyer (3rd Floor)*

3:45 p.m.—5:30 p.m. **Scientific Program**

82. IMS Medallion Lecture *Versailles Ballroom (3rd Floor)*
83. Shrinkage Estimation in Microarray Data Analysis *Melrose Room (3rd Floor)*
84. Opportunities for Biostatisticians Inside (research) and Outside (funding) of NIH *Rosedown Room (3rd Floor)*
85. ROC Methods: Experiments with Time-dependent and Clustered Data *Magnolia Room (3rd Floor)*
86. Contributed Papers: Analysis of Clinical Trials and Biopharmaceutical Studies *Chequers Room (2nd Floor)*
87. Contributed Papers: Clustered Data Methods *Jasperwood Room (3rd Floor)*
88. Contributed Papers: Multivariate Survival *Cambridge Room (2nd Floor)*
89. Contributed Papers: Genome-Wide Association Studies *Oak Alley Room (3rd Floor)*
90. Contributed Papers: New Methodology for Linear Mixed Model Framework *Eglinton Wintom Room (2nd Floor)*
91. Contributed Papers: Environmental and Ecological Applications *Elmwood Room (2nd Floor)*
92. Contributed Papers: Variable Selection and Penalized Regression Models *Belle Chasse Room (3rd Floor)*

5:30 p.m.—6:30 p.m. **ENAR Business Meeting (Open to all ENAR Members)** *Rosedown Room (3rd Floor)*

WEDNESDAY, MARCH 24

7:30 a.m.—9:00 a.m. **Planning Committee Breakfast Meeting (By Invitation Only)** *Warwick Room (3rd Floor)*

8:00 a.m.—12:30 p.m. **Conference Registration** *3rd Floor Foyer*

8:00 a.m. **Exhibits Open** *Court Assembly Foyer (3rd Floor)*

8:30 a.m.—10:15 a.m. **Scientific Program**

93. Statistical Analysis of Brain Imaging Data *Rosedown Room (3rd Floor)*
94. Regression Models with Complex Covariate Inputs *Magnolia Room (3rd Floor)*
95. Methods for Combining Matched and Unmatched Case-Control Studies *Jasperwood Room (3rd Floor)*
96. High Dimensional Data Analysis: Dimension Reduction and Variable Selection *Melrose Room (3rd Floor)*
97. Contributed Papers: Nonparametric Methods *Ascot Room (3rd Floor)*
98. Contributed Papers: Nonparametric and Semiparametric Survival Models *Chequers Room (2nd Floor)*
99. Contributed Papers: Rater Agreement and Screening Tests *Cambridge Room (2nd Floor)*
100. Contributed Papers: Bayesian Methods: Joint Longitudinal/Survival Modeling and Disease Modeling *Eglinton Winton Room (2nd Floor)*
101. Contributed Papers: Integration of Information across Multiple Studies or Multiple -omics Platforms *Belle Chasse Room (3rd Floor)*
102. Contributed Papers: Estimating Equations *Newberry Room (3rd Floor)*

10:15 a.m.—10:30 a.m. **Refreshment Break and Visit the Exhibitors** *Court Assembly Foyer (3rd Floor)*

10:30 a.m.—12:15 p.m. **Scientific Program**

103. Recent Advances in Modeling Nonlinear Measurement Error *Rosedown Room (3rd Floor)*
104. Use of Biomarkers in Personalized Medicine *Melrose Room (3rd Floor)*
105. Analysis of Recurrent Events Data in the Presence of a Terminal Event *Magnolia Room (3rd Floor)*
106. Contributed Papers: Genetic Studies with Related Individuals *Newberry Room (3rd Floor)*
107. Contributed Papers: Hypothesis Testing, Multiple Testing, and Computational Methods *Cambridge Room (2nd Floor)*
108. Contributed Papers: Genomics and Proteomics *Jasperwood Room (3rd Floor)*
109. Contributed Papers: Infectious Disease and Medical Case Studies *Eglinton Winton Room (2nd Floor)*
110. Contributed Papers: Variable Selection Methods *Ascot Room (3rd Floor)*
111. Contributed Papers: Generalized Linear Models *Chequers Room (2nd Floor)*

Poster Presentations

SUNDAY, MARCH 21

7:30—8:00 p.m. New Member Reception
Napolean Ballroom, 3rd Floor

8:00—11:00 p.m. Opening Mixer and Poster Presentations
Reception Napolean Ballroom, 3rd Floor

1. POSTERS: CLINICAL TRIALS
Sponsor: ASA Biopharmaceutical Section

1a. A Predictive Model for Imbalance in Stratified Permuted Block Designs
Joseph W. Adair and Aaron C. Camp*, PPD

1b. Optical Properties of Human Skin as a Criterion for Non-melanoma Skin Cancer Detection
Vadim S. Tyuryaev* and Clyde F. Martin, Texas Tech University

1c. Bayesian Adaptively Randomized Clinical Trial of End-stage Non-small Cell Lung Cancer
Valen E. Johnson, University of Texas M.D. Anderson Cancer Center and Chunyan Cai*, University of Texas Health Science Center at Houston and University of Texas M.D. Anderson Cancer Center

1d. A Distribution-Free Bayesian Method for Estimating the Probability of Response in Combination Drug Tests
John W. Seaman III*, John W. Seaman II and James D. Stamey, Baylor University

1e. A Two-stage Design for Randomized Phase II Clinical Trials with Bivariate Binary Outcome
Rui Qin* and Qian Shi, Mayo Clinic

1f. Improving Small-Sample Inference in Group Randomized Trials with Binary Outcomes
Philip Westgate* and Thomas M. Braun, University of Michigan

1g. A Comparison of Two Simulation Models of Clinical Trials
Maryna Ptukhina* and Clyde Martin, Texas Tech University

1h. Finding and Validating Subgroups in Clinical Trials
Jared C. Foster*, Jeremy M.G. Taylor, University of Michigan and Stephen J. Ruberg, Eli Lilly and Company

2. POSTERS: SURVIVAL ANALYSIS I: METHODOLGY
Sponsor: ASA Biometrics Section

2a. Bayesian Influence Methods with Missing Covariates in Survival Analysis
Diana Lam*, Joseph Ibrahim and Hongtu Zhu, University of North Carolina-Chapel Hill

2b. Bayesian Predictive Distributions under Cox's Proportional Hazard Model
Yijie Liao* and Ronald Butler, Southern Methodist University

2c. Bayesian Joint Model for Longitudinal and Competing Risks with Copula
Yi Qian*, Amgen, Deukwoo Kwon, National Cancer Institute and Jeesun Jung, Indiana University School of Medicine

2d. Inference for Accelerated Failure Time Models for Clustered Survival Data with Potentially Informative Cluster Size
Jie Fan*and Somnath Datta, University of Louisville

2e. Time Dependent Cross-Ratio Estimation
Tianle Hu* and Bin Nan, University of Michigan, Xihong Lin and James Robins, Harvard University

2f. Nonparametric Survival Analysis on Time-Dependent Covariate Effects
Chan-Hee Jo*, University of Arkansas for Medical Sciences, Chunfeng Huang, Indiana University and Haimeng Zhang, Mississippi State University

2g. Hazard-type Empirical Likelihood and General Estimating Equations for Censored Data
Yanling Hu* and Mai Zhou, University of Kentucky

2h. Nonparametric Regression Models for Right-censored Data Using Bernstein Polynomials
Muhtarjan Osman* and Sujit K. Ghosh, North Carolina State University

2i. Practical Mixed Effects Cox Models
Terry M. Therneau*, Mayo Clinic

2j. Model Checking Techniques for Censored Linear Regression Models
Larry F. Leon*, Bristol-Myers Squibb, Tianxi Cai and Lee Jen Wei, Harvard School of Public Health

3. POSTERS: SURVIVAL ANALYSIS II: APPLICATIONS AND METHODOLOGY
Sponsor: ASA Biometrics Section

3a. Evaluation of Risk Factors for Left Atrio-ventricular Valve Stenosis after Atrio-ventricular Septal Defect Repair: A Competing Risks Framework
Adriana C. Dornelles*, Vitor Guerra and Leann Myers, Tulane University

3b. Examine the Dynamic Association between BMI and All-Cause Mortality
Jianghua He*, University of Kansas Medical Center

3c. Estimating Colorectal Cancer Screening in the Presence of Missing Data in a Population with a Resistant Subset and Multiple Observations
Yolanda Hagar*, Laurel Beckett and Joshua Fenton, University of California-Davis Medical Center

3d. Using Q Learning to Construct Dynamic Treatment Regimes with Time-to-Event Outcomes
Zhiguo Li* and Susan Murphy, University of Michigan

3e. Modifications and Alternatives to the SMR in Evaluating Center-Specific Mortality
Kevin He* and Douglas E. Schaubel, University of Michigan

3f. Time Scales In Epidemiological Analysis
Prabhakar Chalise*, Mayo Clinic

3g. Some Graphical Approaches to Monitoring the Outcomes of Liver Transplant Centers
Jie (Rena) Sun* and John D. Kalbfleisch, University of Michigan

4. POSTERS: STATISTICAL GENETICS I
Sponsor: ASA Biometrics Section

4a. Composite Likelihood in Long Sequence Data
Bruce G. Lindsay and Jianping Sun*, Penn State University

4b. A Generalized Linear Model for Peak Calling in ChIP-Seq Data
Jialin Xu* and Yu Zhang, Penn State University

4c. Genome Wise Association Study with Longitudinally Observed Data: QT Interval
JungBok Lee*, Seung Ku, Soriul Kim and Chol Shin, Korea University and Byoung Cheol Jung, University of Seoul

4d. A New Variable Selection Method for Genome-wide Association Studies
Qianchuan He* and Danyu Lin, University of North Carolina-Chapel Hill

4e. Is it Rare or Common? A Coalescent Tree Approach to Identify the Genetic Types of Variants Underlying Complex Diseases
Kaustubh Adhikari*, Harvard University

4f. Effects of Population Stratification in Logistic Regression and an Alternative to Achieving Greater Power
Adrian Tan* and Saonli Basu, University of Minnesota

4g. The Effect of Retrospective Sampling on Estimates of Prediction Error for Multifactor Dimensionality Reduction
Stacey J. Winham* and Alison A. Motsinger-Reif, North Carolina State University

4h. Statistical Models for Detecting Rare Variants Associated with Disease
Jeesun Jung*, Indiana University School of Medicine and Deukwoo Kwon, National Cancer Institute

5. POSTERS: STATISTICAL GENETICS II
Sponsor: ASA Biometrics Section

5a. Comparison of Conditional and Unconditional Analysis of Left-Truncated Data: Simulation Study and Application
Lydia C. Kwee* and Silke Schmidt, Duke University Medical Center

5b. Detecting Copy Number Variation via Mixture-model and ROC Curve
Hui-Min Lin*, Ching-Wei Chang and James Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

5c. Utilizing Genotype Imputation for the Augmentation of Sequence Data
Brooke L. Fridley*, Gregory Jenkins, Matthew Deyo-Svendsen and Scott Hebbring, Mayo Clinic

5d. Trees Assembling Based Mann-Whitney Test for Large-scale Genetic Association Study
Changshuai Wei* and Qing Lu, Michigan State University

5e. Artifact due to Differential Genotyping Error when Cases and Controls are Genotyped using Different Platforms
Jennifer A. Sinnott* and Peter Kraft, Harvard University

5f. A Cross-validated Bagging ROC Method for Predictive Genetic Tests
Chengyin Ye*, Michigan State University and Zhejiang University, Yuehua Cui, Michigan State University, Robert C. Elston, Case Western Reserve University, Jun Zhu, Zhejiang University and Qing Lu, Michigan State University

6. POSTERS: IMAGING AND SPATIAL MODELING
Sponsor: ENAR

6a. Multiscale Adaptive Smoothing Models for Functional Imaging Construction, Segmentation and Classification
Jiaping Wang*, Hongtu Zhu and Weili Lin, University of North Carolina-Chapel Hill

6b. Predicting Post-treatment Neural Activity Based on Pre-treatment Functional Neuroimaging Data
Gordana Derado* and F.D. Bowman, Emory University

6c. Enhanced Global Error Rate Control in Neuroimaging Data
Shuzhen Li* and Lynn Eberly, University of Minnesota

6d. Motion Correction for Two-Photon Laser-Scanning Microscopy
Mihaela Obreja*, University of Pittsburgh and William Eddy, Carnegie Mellon University

6e. Testing Similarity of 2D Electrophoresis Gels Across Groups Based on Independent Component Analysis
Hui Huang*, Anindya Roy, Nicolle Correa and Tulay Adali, University of Maryland Baltimore County

7. POSTERS: GENOMICS AND BIOMARKERS
Sponsor: ENAR

7a. Regularized Gaussian Mixture Modeling with Covariance Shrinkage
Hyang Min Lee* and Jia Li, Penn State University

7b. Effect of Gene Reference Selection on Enrichment Analysis of Gene Lists
Laura Kelly Vaughan*, University of Alabama-Birmingham

7c. Incorporation of Prior Information into Lasso via Linear Constraints
Tianhong He* and Michael Zhu, Purdue University

7d. Multiple Imputation for Missing Values in Microarray Data Analysis
Richard E. Kennedy* and Hemant K. Tiwari, University of Alabama-Birmingham

7e. Gene Clustering and Identification using Composite Likelihood
Ran Li* and Baolin Wu, University of Minnesota

7f. Hybrid Pooled-Unpooled Design for Cost-Efficient Measurement of Biomarkers
Enrique F. Schisterman and Sunni Mumford, National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Albert Vexler, SUNY, Albany and Neil J. Perkins*, National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development

7g. Weakest-link Models for Joint Effects in Cell-based Data
The Minh Luong* and Roger Day, University of Pittsburgh

8. POSTERS: LONGITUDINAL DATA ANALYSIS
Sponsor: ASA Biometrics Section

8a. Covariate Adjustment in Latent Class Models for Joint Analysis of Longitudinal and Time-to-Event Outcomes
Benjamin E. Leiby*, Thomas Jefferson University, Mary D. Sammel, University of Pennsylvania and Terry Hyslop, Thomas Jefferson University

8b. Joint Modeling of Primary Outcome and Longitudinal Data Measured at Informative Observation Times
Song Yan*, Daowen Zhang and Wenbin Lu, North Carolina State University

8c. The Two-group Latent Growth Modeling in Studying of Change over Time
Yingchun Zhan*, Du Feng and Clyde Martin, Texas Tech University

8d. A Wilcoxon-type Statistic for Repeated Binary Measures with Multiple Outcomes
Okan U. Elci* and Howard E. Rockette, University of Pittsburgh

8e. Modeling for Longitudinal data with High Dropout Rates
Sunkyung Yu*, James Dziura, Melissa M. Shaw, and Mary Savoye, Yale University

8f. A Permutation Test for Random Effects in Linear Mixed Models
Oliver Lee* and Thomas M. Braun, University of Michigan

9. POSTERS: SPATIAL/TEMPORAL MODELING AND ENVIRONMENTAL/ECOLOGICAL APPLICATIONS
Sponsor: ENAR

9a. Estimation of Br Concentration Distribution in Groundwater
Yunjin Park* and Yongsung Joo, Dongguk University, Korea

9b. Cumulative Dietary Exposure to Malathion and Chloypyrifos in the NHEXAS-Maryland Investigation
Anne M. Riederer, Emory University, Ayona Chatterjee*, University of West Georgia, Scott M. Bartell, University of California and Barry P.Ryan, Emory University

9c. Adjusting for Measurement Error in Maternal PM Exposure and Birth Weight Models
Simone Gray*, Alan Gelfand and Marie Lynn Miranda, Duke University

9d. Product Partition Models with Correlated Parameters
Joao VD Monteiro*, University of Minnesota, Rosangela H. Loschi and Renato M. Assuncao, Universidade Federal de Minas Gerais

10. POSTERS: APPLICATIONS AND CASE STUDIES I
Sponsor: ENAR

10a. Psychological Correlates of Placebo Response
Hamdan Azhar*, University of Michigan, Christian S. Stohler, University of Maryland and Jon-Kar Zubieta, University of Michigan

10b. Assessment of Reliability, Validity and Effects of Missing Data of the FACT-M Questionnaire
J. Lynn Palmer*, University of Texas M.D. Anderson Cancer Center

10c. Estimating Disease Prevalence using Inverse Binomial Pooled Screening
Joshua M. Tebbs*, University of South Carolina and Nicholas A. Pritchard, Coastal Carolina University

10d. Generating a Simulated Kidney Paired Donation Program
Yan Zhou*, Yijiang (John) Li, Jack D. Kalbfleisch and Peter X. K. Song, University of Michigan

10e. Factors Associated with Difficulty in using Community Based Services Among Children with Special Health Care Needs
Sreelakshmi Talasila*, Kimberly Fulda, Sejong Bae and Karan Singh, University of North Texas Health Science Center

10f. Use of Multiple Singular Value Decompositions to Analyze Complex Calcium Ion Signals
Josue G. Martinez*, Jianhua Z. Huang, Robert C. Burghardt, Rola Barhoumi and Raymond J. Carroll, Texas A&M University

10g. Psychosocial Stress and Health Disparities
Brisa N. Sanchez*, Trivellore E. Raghunathan, Meihua Wu and Ana V. Diez-Roux, University of Michigan

10h. Selection of a Working Correlation Structure in Pediatric Studies of Renal and Crohn's Disease
Matthew White*, Justine Shults, Meena Thayu, Michelle Denburg and Mary Leonard, University of Pennsylvania School of Medicine

11. POSTERS: POWER/SAMPLE SIZE
Sponsor: ASA Biopharmaceutical Section

11a. Bayesian Sample Size Determination for Studies Designed to Evaluate Continuous Medical Tests
Adam J. Branscum*, University of Kentucky, Dunlei Cheng, Baylor Health Care System and James D. Stamey, Baylor University

11b. Bayesian Sample Size Determination for Two Independent Poisson Rates
Austin L. Hand*, James Stamey and Dean Young, Baylor University

11c. Comparison of Sample Size Requirements in 3-way Comparisons for Fixed-dose Combination Drug Efficacy Studies
Linlin Luo* and Julia N. Soulakova, University of Nebraska

11d. A Simple Method for Approximating Effect on Power of Linear Model Misspecification
T. Robert Harris*, University of Texas

12. POSTERS: NONPARAMETRIC METHODS
Sponsor: ENAR

12a. A Bayesian Nonparametric Goodness of Fit Test for Logistic Regression with Continuous Response Data
Angela Schörgendorfer* and Adam J. Branscum, University of Kentucky and Timothy E. Hanson, University of Minnesota

12b. A Study of Bayesian Density Estimation using Bernstein Polynomials
Charlotte C. Gard* and Elizabeth R. Brown, University of Washington

12c. Non Parametric Analysis of Multidimensional Profiles
Margo A. Sidell* and Leann Myers, Tulane University

12d. A Functional Data Analysis Method for Evaluation of Inherence of Medical Guideline
Lingsong Zhang*, Purdue University

12e. Robust Lower-dimensional Approximation for Sparse Functional Data with its Application to Screening Young Children's Growth Paths
Wei Ying and Wenfei Zhang*, Columbia University

12f. A Copula Approach for Estimating Correlation of Shared Couple Behaviors
Seunghee Baek*, Scarlett L. Bellamy, Andrea B. Troxel, Thomas R. Ten Have and John B. Jemmott III, University of Pennsylvania

12g. A Comparative Study of Nonparametric Estimation in Weibull Regression: A Penalized Likelihood Approach
Young-Ju Kim*, Kangwon National University

12h. Multivariate Shape Restriction Regression with Bernstein Polynomial
Jiangdian Wang* and Sujit Ghosh, North Carolina State University

13. POSTERS: MISSING DATA AND MEASUREMENT ERROR
Sponsors: ASA Biometrics Section/ASA Social Statistics Section/ ASA Section on Statistics in Epidemiology

13a. Combining Disparate Measures of Metabolic Rate During Simulated Spacewalks
Robert J. Ploutz-Snyder*, Alan H. Feiveson, Dan Nguyen and Lawrence Kuznetz, NASA Johnson Space Center

13b. A Bayesian Approach to Multilevel Poisson Regression with Misclassification
Monica M. Bennett*, John W. Seaman, Jr. and James D. Stamey, Baylor University

13c. Misrecording in the Negative Binomial Regression Model
Mavis Pararai*, Indiana University of Pennsylvania

13d. A Non-parametric Method for Estimating a Heaping Mechanism from Precise and Heaped Self-report Data
Sandra D. Griffith*, University of Pennsylvania, Saul Shiffman, University of Pittsburgh and Daniel F. Heitjan, University of Pennsylvania

13e. Multiple Imputation in Group Randomized Trials: The Impact of Misspecifying Clustering in the Imputation Model
Rebecca R. Andridge*, The Ohio State University

13f. Estimation in Hierarchical Models with Incomplete Binary Response and Binary Covariates
Yong Zhang* and Trivellore Raghunathan, University of Michigan

13g. Bayesian Multilevel Regression Models with Spatial Variability and Errors in Covariates
Theodore J. Thompson* and James P. Boyle, Centers for Disease Control and Prevention

14. POSTERS: STATISTICAL METHODS
Sponsor: ENAR

14a. Bayesian Inference for Censored Binomial Sampling
Jessica Pruszynski* and John W. Seaman, Jr., Baylor University

14b. A Comparison of Model-based versus Moment-based Estimator of Complier Average Causal Effect (CACE)
Nanhua Zhang* and Roderick J. A. Little, University of Michigan

14c. Assessing the Effect of a Treatment with a Clump of Observations at Zero
Jing Cheng*, University of Florida and Dylan Small, University of Pennsylvania

14d. Nonconvergence in Logistic and Poisson Models for Neural Spiking
Mengyuan Zhao* and Satish Iyengar, University of Pittsburgh

14e. Approximate Inferences for Nonlinear Mixed-effects Models with Skew-normal Independent Distributions
Victor H. Lachos Davila*, Campinas State University, Brazil and Dipak K. Dey, University of Connecticut

15. POSTERS: SURVEY RESEARCH AND CATEGORICAL DATA ANALYSIS
Sponsor: ASA Social Statistics Section

15a. The Intracluster Correlation as a Function of Inherent and Design Induced Covariances
Robert E. Johnson* and Tina D. Cunningham, Virginia Commonwealth University

15b. Identifiability of A Restricted Finite Mixture Model for Few Binary Responses Allowing Covariate Dependence in Mixing Distribution
Yi Huang*, University of Maryland Baltimore County

15c. Estimating Disease Prevalence when Testing Consent Rates Are Low: A Pooled Testing Approach
Lauren Hund* and Marcello Pagano, Harvard University

15d. To Weight or Not to Weight: A Survey Sampling Simulation Study
Marnie Bertolet*, University of Pittsburgh

16. POSTERS: BIOLOGICS, PHARMACEUTICALS AND MEDICAL DEVICES
Sponsor: ASA Biopharmaceutical Section

16a. Hierarchical Modeling to Assess Precision and Treatment Effects in an Interlaboratory Study
Jason Schroeder*, Center for Devices and Radiological Health, Food and Drug Administration

16b. A Statistical Test for Evaluation of Biosimilarity in Variability of Biologic Products
Tsung-Cheng Hsieh, National Taiwan University, Taiwan, Shein-Chung Chow, Duke University School of Medicine, Jen-Pei Liu, National Taiwan University, Taiwan, Chin-Fu Hsiao, National Health Research Institutes, Taiwan and Eric M. Chi*, Amgen Inc.

16c. Bayesian Hierarchical Monotone Regression Splines for Dose-Response Assessment and Drug-Drug Interaction Analysis
Violeta G. Hennessey*, Veerabhadran Baladandayuthapani and Gary L. Rosner, University of Texas M.D. Anderson Cancer Center

MONDAY, MARCH 22

7:15 a.m.—8:15 a.m. Revitalization of Biostatistical Grant Applications at NIH – An Open Forum *Melrose Room (3rd Floor)*

8:30 a.m.—10:15 a.m.

17. STATISTICAL METHODS FOR ESTIMATING THE PUBLIC HEALTH IMPACT OF POLICY INTERVENTIONS
Melrose Room (3rd Floor)
Sponsors: ASA Health Policy Statistics Section, ASA Social Statistics Section
Organizer: Francesca Dominici, Harvard University
Chair: Francesca Dominici, Harvard University

8:30 **Analysis of Longitudinal Data to Evaluate a Policy Change**
Benjamin French*, University of Pennsylvania and Patrick J. Heagerty, University of Washington

9:00 **A Statistical Approach for Assessing the Public Health Impact of Smoking Bans**
Christopher D. Barr*, Harvard University

9:30 **Estimating Longitudinal Effects using Propensity Scores as Regressors**
Aristide C. Achy-Brou, JP Morgan, Constantine E. Frangakis*, Johns Hopkins University and Michael Griswold. University of Mississippi

10:00 **Floor Discussion**

18. ADVANCES IN TIME SERIES ANALYSIS OF NEUROLOGICAL STUDIES
Rosedown Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Robert Krafty, University of Pittsburgh
Chair: Yu Cheng, University of Pittsburgh

8:30 **Nonparametric Spectral Analysis with Applications to Seizure Characterization Using EEG Time Series**
Li Qin*, Fred Hutchinson Cancer Research Center and Yuedong Wang, University of California-Santa Barbara

8:55 **Classification of Families of Locally Stationary Time Series**
Robert T. Krafty*, University of Pittsburgh and Wensheng Guo, University of Pennsylvania

9:20 **Evolutionary Factor Analysis of EEG data**
Giovanni Motta, University of Maastricht and Hernando Ombao*, Brown University

9:45 **Stimulus-Locked VAR Models for Event-Related fMRI**
Wesley K. Thompson*, University of California-San Diego

10:10 **Floor Discussion**

19. INFERENCE AND PREDICTION FOR SPARSE NETWORKS
Magnolia Room (3rd Floor)
Sponsor: ASA Statistical Learning and Data Mining Section
Organizer: Annie Qu, University of Illinois at Urbana Champaign
Chair: Annie Qu, University of Illinois at Urbana Champaign

8:30 **Variational EM Algorithms for a Class of Network Mixture Models**
Duy Vu and David R. Hunter*, Penn State University

8:55 **Sparse Regression Models for Constructing Genetic Regulatory Networks**
Jie Peng*, University of California-Davis and Pei Wang, Fred Hutchinson Cancer Research Center

9:20 **Time Varying Networks: Reverse Engineering and Analyzing Rewiring Social and Genetic Interactions**
Eric P. Xing, Carnegie Mellon University

9:45 **Penalized Regression with Networked Predictors and Its Application to eQTL Analysis**
Wei Pan*, University of Minnesota, Benhua Xie, Takeda Global Research and Development, Xiaotong Shen, University of Minnesota

10:10 **Floor Discussion**

20. RECENT DEVELOPMENTS IN HIGH DIMENSIONAL INFERENCE
Jasperwood Room (3rd Floor)
Sponsor: IMS
Organizer: Jiashun Jin, Carnegie Mellon University
Chair: Ji Zhu, University of Michigan

8:30 **Optimal Screening for Sparse Signals**
Tony Cai, University of Pennsylvania and Wenguang Sun*, North Carolina State University

8:55 **Revisiting Marginal Regression**
Jiashun Jin*, Christopher Genovese and Larry Wasserman, Carnegie Mellon University

9:20 **Theoretical Support for High Dimensional Data Analysis based on Student's t Statistic**
Aurore Delaigle and Peter Hall*, University of Melbourne and Jiashun Jin, Carnegie Mellon University

9:45 **Risk Predictions from Genome Wide Association Data**
Hongyu Zhao*, Jia Kang, Ruiyan Luo and Judy Cho, Yale University

10:10 **Floor Discussion**

21. CONTRIBUTED PAPERS: IMAGE ANALYSIS
Chequers Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Chongzhi Di, Fred Hutchinson Cancer Research Center

● **8:30** **Predicting Treatment Efficacy via Quantitative MRI: A Bayesian Joint Model**
Jincao Wu* and Timothy D. Johnson, University of Michigan

8:45 **A Bayesian Hierarchical Framework for Modeling of Resting-state fMRI Data**
Shuo Chen* and DuBois F. Bowman, Emory University

● **9:00** **Covariate-adjusted Nonparametric Analysis of Magnetic Resonance Images using Markov Chain Monte Carlo**
Haley Hedlin* and Brian Caffo, Johns Hopkins University, Ziyad Mahfoud, American University of Beirut and Susan S. Bassett, Johns Hopkins University School of Medicine

9:15 **Meta Analysis of Functional Neuroimaging Data via Bayesian Spatial Point Processes**
Jian Kang* and Timothy D. Johnson, University of Michigan, Thomas E. Nichols, University of Warwick and Tor D. Wager, Columbia University

9:30 **Multiscale Adaptive Supervised Feature Selction for Image Data**
Ruixin Guo* and Hongtu Zhu, University of North Carolina-Chapel Hill

9:45 **A Multiresolution Analysis of Environmental Lead Exposure's Impact on Brain Structure**
Shu-Chih Su*, Merck & Co. and Brian Caffo, Johns Hopkins University

10:00 **Semiparametric Approaches to Separation of Sources Using Independent Component Analysis**
Ani Eloyan* and Sujit K. Ghosh, North Carolina State University

22. CONTRIBUTED PAPERS: MISSING DATA IN LONGITUDINAL STUDIES
Oak Alley Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section
Chair: Chunling Catherine Liu, National Institutes of Health

8:30 **Pseudo-likelihood Estimation for Incomplete Data**
Geert Molenberghs*, Universiteit Hasselt & Katholieke Universiteit Leuven, Belgium, Michael G. Kenward, London School of Hygiene and Tropical Medicine, Geert Verbeke, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium and Teshome Birhanu, Universiteit Hasselt, Belgium

8:45 **A Bayesian Shrinkage Model for Longitudinal Binary Data with Intermittent Missingness and Dropout with Application to the Breast Cancer Prevention Trial**
Chenguang Wang * and Michael J. Daniels, University of Florida, Daniel O. Scharfstein, Johns Hopkins University and Stephanie Land, University of Pittsburgh

9:00 **A Test of Missing Completely at Random for Regression Data with Nonresponse**
Gong Tang*, University of Pittsburgh

9:15 **Evaluating Statistical Hypotheses for Non-identifiable Models using General Estimating Functions**
Guanqun Cao*, David Todem, Lijian Yang, Michigan State University and Jason P. Fine, University of North Carolina-Chapel Hill

● **9:30** **Generalized ANOVA for Concurrently Modeling Mean and Variance within a Longitudinal Data Setting**
Hui Zhang* and Xin M. Tu, University of Rochester

9:45 **Semiparametric Regression Models for Repeated Measures of Mortal Cohorts with Non-Monotone Missing Outcomes and Time-Dependent Covariates**
Michelle Shardell*, University of Maryland School of Medicine, Gregory E. Hicks, University of Delaware, Ram R. Miller and Jay Magaziner, University of Maryland School of Medicine

10:00 **Multiple Imputation Methods for Prediction with Multiple Left-censored Biomarkers Due to Detection Limits**
Minjae Lee* and Lan Kong, University of Pittsburgh

23. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS
Eglinton Winton Room (2nd Floor)
Sponsor: ENAR
Chair: Layla Parast, Harvard University

8:30 **Risk-Adjusted Monitoring of Time to Event**
Axel Gandy, Imperial College, London, Jan Terje*, Kvaløy University of Stavanger, Norway, Alex Bottle and Fanyin Zhou, Imperial College, London

8:45 **Bayesian Inference for Cumulative Incidence Function Under Additive Risks Model**
Junhee Han*, University of Arkansas and Minjung Lee, National Cancer Institute

9:00 **Semiparametric Hybrid Empirical Likelihood Inference for Two-sample Comparison with Censored Data**
Mai Zhou, University of Kentucky, Haiyan Su*, Montclair State University and Hua Liang, University of Rochester

9:15 **Semiparametric Transformation Models Based on Degradation Processes**
Sangbum Choi* and Kjell A. Doksum, University of Wisconsin-Madison

9:30 **Multiple Imputation Methods for Inference on Cumulative Incidence with Missing Cause of Failure**
Minjung Lee*, Kathleen A. Cronin, Mitchell H. Gail and Eric J. Feuer, National Cancer Institute

9:45 **Constrained Nonparametric Maximum Likelihood Estimation of Survivor Functions under Stochastic Ordering in One-sample and Two-sample Cases**
Yong Seok Park*, John D. Kalbfleisch and Jeremy MG Taylor, University of Michigan

24. CONTRIBUTED PAPERS: MICROARRAYS: DIFFERENTIAL EXPRESSION AND REPRODUCIBILITY
Elmwood Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Mehmet Kocak, St. Jude Children's Research Hospital

8:30 **A Bayesian Model Averaging Approach to Differentially Expressed Gene Detection in Observational Microarray Studies**
Xi K. Zhou*, Weill Medical College of Cornell University, Fei Liu, University of Missouri-Columbia and Andrew J. Dannenberg, Weill Medical College of Cornell University

8:45 **Gene Expression Barcodes Based on Data from 8,277 Microarrays**
Matthew N. McCall*, Johns Hopkins University, Michael J. Zilliox, Emory University School of Medicine and Rafael A. Irizarry, Johns Hopkins University

9:00 **A Multivariate Empirical Bayes Modeling Approach to Simultaneous Inference of Multi-class Comparison Problems**
Xiting Cao* and Baolin Wu, University of Minnesota

9:15 **Testing Multiple Hypotheses Using Population Information of Samples**
Mingqi Wu* and Faming Liang, Texas A&M University

9:30 **Estimate of Transcript Absolute Concentration from DNA Micorarrays**
Yunxia Sui* and Zhijin Wu, Brown University

9:45 **Statistical Practice in High-Throughput siRNA Screens Identifying Genes Mediating Sensitivity to Chemotherapeutic Drugs**
Fei Ye*, Joshua A. Bauer, Huiyun Wu, Jennifer A. Pietenpol and Yu Shyr, Vanderbilt University

10:00 **Bayesian Hierarchical Models for Correlating Expression Data across Chips**
Bernard Omolo*, University of North Carolina-Chapel Hill

25. CONTRIBUTED PAPERS: STEPHEN LAGAKOS: VIEWS FROM FORMER STUDENTS SPANNING 2.5 DECADES
Belle Chasse Room (3rd Floor)
Sponsor: ENAR
Chair: Victor Degruttola, Harvard University

8:30 **Steve Lagakos, a Legacy of Interactions**
Roger S. Day*, University of Pittsburgh

9:00 **Biostatistics in the Era of Interdisciplinary Science**
Melissa D. Begg* and Roger D. Vaughan, Columbia University

9:30 **Cross-Sectional Prevalence Testing for Estimating HIV Incidence**
Rui Wang* and Stephen W. Lagakos, Harvard University

10:00 **Trends and Challenges in Research Involving Elderly and Impaired Drivers**
Jeffrey D. Dawson*, Elizabeth Dastrup, Amy M. Johnson, Ergun Y. Uc and Matthew Rizzo, University of Iowa

26. CONTRIBUTED PAPERS: ADAPTIVE CLINICAL TRIAL DESIGNS FOR DOSE FINDING
Cambridge Room (2nd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Yang Xie, University of Texas Southwestern Medical Center

8:30 **Bayesian Adaptive Dose-finding Studies with Delayed Responses**
Haoda Fu* and David Manner, Eli Lilly and Company

8:45 **Bayesian Phase I Dose-finding Design Modeling Discrete-time Toxicity Grades**
Lin Yang*, Nebiyou B. Bekele and Donald A. Berry, University of Texas M.D. Anderson Cancer Center

9:00 **Bayesian Phase I/II Drug-combination Trial Design in Oncology**
Ying Yuan, University of Texas M.D. Anderson Cancer Center and Guosheng Yin*, University of Hong Kong

9:15 **Bayesian Model Averaging Continual Reassessment Method in Phase I Clinical Trials**
Guosheng Yin, University of Hong Kong and Ying Yuan*, University of Texas M.D. Anderson Cancer Center

9:30 **Dose Finding in Drug Combinations with Discrete and Continuous Doses**
Lin Huo* and Ying Yuan, University of Texas M.D. Anderson Cancer Center and Guosheng Yin, University of Hong Kong

9:45 **A Two-stage Dose-response Adaptive Design Method for Establishing Proof of Concept**
Yoko Tanaka*, Stewart Anderson and Allan R. Sampson, University of Pittsburgh

10:00 **Proportional Odds Model for Dose Finding Clinical Trial Designs with Ordinal Toxicity Grading**
Emily M. Van Meter*, Elizabeth Garrett-Mayer and Dipankar Bandyopadhyay, Medical University of South Carolina

MONDAY, MARCH 22

10:15—10:30 a.m.
Refreshment Break and Visit the Exhibitors
Court Assembly Foyer (3rd Floor)

10:30 a.m.—12:15 p.m.

27. BAYESIAN METHODS FOR COMBINING DATA FROM MULTIPLE SOURCES FOR CLINICAL TRIALS
Melrose Room (3rd Floor)
Sponsors: ASA Section on Bayesian Statistical Sciences, ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Sourish Das, Duke University
Chair: David Banks, Duke University

10:30 **Synthetic Priors from Analysis of Multiple Experts' Opinions**
Sourish Das*, Hongxia Yang and David Banks, Duke University

10:55 **A Novel Approach for Eliciting an Odds Ratio in the Setting of Incomplete Longitudinal Data**
Michael Daniels* and Chenguang Wang, University of Florida and Daniel Scharfstein, Johns Hopkins University

11:20 **Using Prior Information and Elicited Utilities for Adaptive Decision Making in Phase I/II Trials**
Peter F. Thall*, University of Texas M.D. Anderson Cancer Center

11:45 **Borrowing Strength with Non-Exchangeable Priors over Subpopulations**
Luis Gonzalo Leon-Novelo, Univ of Florida B. Nebiyou Bekele, M.D, Anderson Cancer Center, Peter Mueller, M.D Anderson Cancer Center, Fernando Quintana, Pontificia Universidad Catolica de Chile, Kyle Wathen, M.D Anderson Cancer Center

12:10 **Floor Discussion**

28. COMPETING RISKS IN ACTION
Rosedown Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Jason Fine, University of North Carolina-Chapel Hill
Chair: Limin Peng, Emory University

10:30 **Regression Strategy for the Conditional Probability of a Competing Risk**
Aurelien Latouche*, University of Versailles and European Group for Blood and Marrow Transplantation

11:00 **Summarizing Differences in Cumulative Incidence Function**
Mei-Jie Zhang*, Medical College of Wisconsin and Jason Fine, University of North Carolina-Chapel Hill

11:30 **Inference on Quantile Residual Life under Competing Risks**
Jong-Hyeon Jeong*, University of Pittsburgh

12:00 **Floor Discussion**

29. STUDYING GENETIC AND ENVIRONMENTAL RISK FACTORS OF COMPLEX HUMAN DISORDERS AND THEIR INTERACTIONS
Magnolia Room (3rd Floor)
Sponsor: IMS
Organizer: Tian Zheng, Columbia University
Chair: Tian Zheng, Columbia University

3:45 **Discovering Influential Variables: A Method of Partitions**
Shaw-Hwa Lo*, Columbia University, Herman Chernoff, Harvard University and Tian Zheng, Columbia University

4:10 **Combining Disease Models to Test for Gene-Environment Interaction in Nuclear Families**
Thomas J. Hoffmann* University of California-SanFrancisco and Nan M. Laird, Harvard University

4:35 **The Value of SNPs for Projecting Breast Cancer Risk**
Mitchell H. Gail*, National Cancer Institute

5:00 **Testing for the Effect of Rare Variants in Complex Traits: A Novel Approach**
Iuliana Ionita-Laza*, Columbia University, Christoph Lange and Nan M. Laird, Harvard University

5:25 **Floor Discussion**

30. CONTRIBUTED PAPERS: SPATIAL/TEMPORAL: MODELING AND METHODOLOGY
Jasperwood Room (3rd Floor)
Sponsor: ENAR
Chair: Andrew O. Finley, Michigan State University

10:30 **Variational Bayesian Method for Spatial Data Analysis**
Qian Ren* and Sudipto Banerjee, University of Minnesota

10:45 **Gaussian Predictive Process Model for Random Knots**
Rajarshi Guhaniyogi*, University of Minnesota, Andrew O. Finley, Michigan State University, Sudipto Banerjee, University of Minnesota and Alan Gelfand, Duke University

11:00 **Additive Models with Spatio-temporal Data**
Xiangming Fang*, East Carolina University and Kung-Sik Chan, University of Iowa

11:15 **Composite Quadratic Inference Functions and Applications in Spatio-temporal Models**
Yun Bai*, Peter X.K. Song and Trivellore Raghunathan, University of Michigan

● **11:30** **Nonparametric Hierarchical Modeling for Detecting Boundaries in Areally Referenced Spatial Datasets**
Pei Li*, Sudipto Banerjee, Timothy E. Hanson and Alexander M. McBean, University of Minnesota

11:45 **Bayesian Analysis of High-throughput Data via Regression Models with Spatially Varying Coefficients**
Xinlei Wang*, Southern Methodist University and Guanghua Xiao, University of Texas Southwestern Medical Center

12:00 **Modeling Time Series Data with Semi-reflective Boundaries with Application to Lateral Control of Motor Vehicles**
Amy M. Johnson* and Jeffrey D. Dawson, University of Iowa

31. CONTRIBUTED PAPERS: PATHWAY AND NETWORK-BASED GENOMIC ANALYSIS
Oak Alley Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Bernard Omolo, University of North Carolina-Chapel Hill

10:30 **Genetic Network Learning in Genetical Genomics Experiments**
Jianxin Yin* and Hongzhe Li, University of Pennsylvania

10:45 **Robust Gene Pathway Testing**
Hongyuan Cao*, Fred Wright and Michael Kosorok, University of North Carolina-Chapel Hill

11:00 **Structured Varying-Coefficient Model for High-Dimensional Feature Discovery with Applications in Genomic Analysis**
Zhongyin J. Daye* and Hongzhe Li, University of Pennsylvania

11:15 **An Integrative Pathway-based Clinical-genomic Model for Cancer Survival Prediction**
Xi Chen* and Lily Wang, Vanderbilt University and Hemant Ishwaran, Cleveland Clinic

11:30 **Analysis of Biological Pathways Using Laplacian Eigenmaps and Penalized Principal Component Regression on Graphs**
Ali Shojaie* and George Michailidis, University of Michigan

11:45 **Mixed Effects Cox Models for Gene Set Analysis**
Marianne Huebner* and Terry Therneau, Mayo Clinic

12:00 **Network-based Empirical Bayes Methods for Linear Models with Applications to Genomic Data**
Caiyan Li*, The U.S. Food and Drug Administration, Hongzhe Li, University of Pennsylvania and Zhi Wei, New Jersey Institute of Technology

32. CONTRIBUTED PAPERS: CAUSAL INFERENCE: METHODOLOGY
Elmwood Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Health Policy Statistics Section, ASA Social Statistics Section, ASA Section on Statistics in Epidemiology
Chair: Frank B. Yoon, Harvard University

10:30 **Sensitivity Analysis for Unmeasured Confounding in Principal Stratification**
Scott Schwartz*, Fan Li and Jerry Reiter, Duke University

10:45 **Inference for the Effect of Treatment on Survival Probability in Randomized Trials with Noncompliance and Administrative Censoring**
Hui Nie*, University of Pennsylvania, Jing Cheng, University of Florida, College of Medicine and Dylan S. Small, University of Pennsylvania

11:00 **Bias Associated with Using Propensity Score as a Regression Predictor**
Bo Lu* and Erinn Hade, The Ohio State University

11:15 **Doubly Robust Instrumental Variable Estimation of LATE(x)**
Elizabeth L. Ogburn*, Andrea Rotnitzky and James Robins, Harvard University

11:30 **A Powerful and Robust Test Statistic for Randomization Inference in Group-Randomized Trials**
Kai Zhang*, Mikhail Traskin and Dylan Small, University of Pennsylvania

11:45 **A Markov Compliance Classes and Outcomes Model for Causal Analysis in the Longitudinal Studies**
Xin Gao* and Michael R. Elliott, University of Michigan

12:00 **Semiparametric Estimation of Causal Mediation Effects in Randomized Trials**
Jing Zhang* and Joseph Hogan, Brown University

33. CONTRIBUTED PAPERS: EPIDEMIOLOGICAL METHODS
Chequers Room (2nd Floor)
Sponsor: ASA Section on Statistics in Epidemiology
Chair: Howard Chang, Statistical and Applied Mathematical Sciences Institute

10:30 **Estimation and Testing of the Relative Risk of Disease in Case Control Studies with a Set of k Matched Controls Per Case**
Barry K. Moser* and Susan Halabi, Duke University Medical Center

10:45 **Nested Case-control Analysis for Observational Data in Cardiovascular Disease**
Zugui Zhang*, Edward F. Ewen and Paul Kolm, Christiana Care Health System

11:00 **Simple Adjustments to Reduce Bias and Mean Squared Error Associated with Regression-Based Odds Ratio and Relative Risk Estimators**
Robert H. Lyles* and Ying Guo, Emory University

11:15 **Mortality Model for Prostate Cancer**
Shih-Yuan Lee* and Alexander Tsodikov, University of Michigan

11:30 **Binary Regression Analysis with Pooled Exposure Measurements**
Zhiwei Zhang* and Paul S. Albert, Eunice Kennedy Shriver National Institute of Childe Health of Human Development

11:45 **Attributable Fraction Functions for Censored Event Times**
Li Chen*, Danyu Lin and Donglin Zeng, University of North Carolina-Chapel Hill

12:00 **A Multivariate Nonlinear Measurement Error Model for Episodically Consumed Foods**
Saijuan Zhang*, Texas A&M University, Adriana Pérez, University of Texas Health Science Center at Houston, Victor Kipnis, National Cancer Institute, Laurence Freedman, Sheba Medical Center, Israel, Kevin Dodd, National Cancer Institute, Raymond J. Carroll, Texas A&M University and Douglas Midthune, National Cancer Institute

34. CONTRIBUTED PAPERS: CATEGORICAL DATA ANALYSIS
Eglinton Winton Room (2nd Floor)
Sponsor: ENAR
Chair: Inyoung Kim, Virginia Tech University

10:30 **Kullback Leibler Risk of Estimators for Univariate Discrete Exponential Family Distributions**
Qiang Wu* and Paul Vos, East Carolina University

10:45 **Correlated Ordinal Categorical Data Analysis: Comparing Braun-Blanquet Sea Grass Coverage Abundance Scores**
Nate Holt* and Mary Christman, University of Florida

11:00 **A Bayesian Approach for Correcting Misclassification in both Outcome Variable and Covariate**
Sheng Luo* and Wenyaw Chan, University of Texas Health Science Center and Michelle Detry, University of Wisconsin-Madison

11:15 **Analysis of Zero-Inflated Clustered Count Data Using Marginalized Model Approach**
Keunbaik Lee*, Louisiana State University-New Orleans, Yongsuung Joo, Dongguk University, Korea and Joon Jin Song, University of Arkansas

11:30 **Confidence Intervals that Match Fisher's Exact or Blaker's Exact Tests**
Michael P. Fay*, National Institute of Allergy and Infectious Diseases

11:45 **On Finding the Upper Confidence Limit for a Binomial Proportion When Zero Successes Are Observed**
Courtney Wimmer*, Medical College of Georgia

12:00 **Floor Discussion**

35. CONTRIBUTED PAPERS: ADAPTIVE CLINICAL TRIAL DESIGN
Belle Chasse Room (3rd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Michelle Detry, University of Wisconsin-Madison

● **10:30** **Hierarchical Gaussian Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials**
Brian P. Hobbs* and Bradley P. Carlin, University of Minnesota, Daniel Sargent and Sumithra Mandrekar, Mayo Clinic

10:45 **Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer**
Lu Wang*, University of Michigan, Peter Thall, University of Texas M.D. Anderson Cancer Center, Andrea Rotnitzky, Universidad Torcuato Di Tella, Xihong Lin, Harvard University and Randall Millikan, University of Texas M.D. Anderson Cancer Center

11:00 **Median Residual Life Time Estimation in Sequentially Randomized Trials**
Jin H. Ko* and Abdus S. Wahed, University of Pittsburgh

11:15 **Design of Dose-finding Experiments with Correlated Responses of Different Types**
Valerii V. Fedorov and Yuehui Wu, GlaxoSmithKline and Rongmei Zhang*, University of Pennsylvania

11:30 **Issues to Consider in Selecting a Response-Adaptive Design for Dose-finding Experiments**
Nancy Flournoy*, University of Missouri-Columbia

11:45 **Estimating the Dose-toxicity Curve in Completed Phase I Studies**
Irina Ostrovnaya* and Alexia Iasonos, Memorial Sloan-Kettering Cancer Center

12:00 **Information in a Simple Adaptive Optimal Design**
Ping Yao*, Northern Illinois University and Nancy Flournoy, University of Missouri-Columbia

36. CONTRIBUTED PAPERS: ROC ANALYSIS
Cambridge Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Sergei Leonov, GlaxoSmithKline

10:30 **Biomarker Validation with an Imperfect Reference: Bounds and Issues**
Sarah C. Emerson* and Rebecca A. Betensky, Harvard University

10:45 **Logistic Regression-based Approach to Borrowing Information Across Common-ROC Populations in Risk Prediction**
Ying Huang*, Columbia University and Ziding Feng, Fred Hutchinson Cancer Research Center

11:00 **Subject-specific Type of Approach in FROC Analysis**
Andriy Bandos*, Howard E. Rockette and David Gur, University of Pittsburgh

11:15 **Classification of Binormal ROC Curves with Respect to Improperness**
Stephen L. Hillis*, Iowa City VA Medical Center and Kevin S. Berbaum, University of Iowa

11:30 **Nonparametric Estimation of Time-Dependent Predictive Accuracy Curve**
Paramita Saha*, National Institute of Environmental Health Sciences and Patrick J. Heagerty, University of Washington

● **11:45** **Optimal Combinations of Diagnostic Tests Based on AUC**
Xin Huang*, Yixin Fang and Gengsheng Qin, Georgia State University

12:00 **Floor Discussion**

MONDAY, MARCH 22

12:15—1:30 p.m.
ROUNDTABLE LUNCHEONS
Napoleon Ballroom (3rd Floor)

Monday, March 22
1:45—3:30 p.m.

37. COMPARATIVE EFFECTIVENESS RESEARCH: THE METHODOLOGIC CHALLENGES
Melrose Room (3rd Floor)
Sponsors: ASA Health Policy Statistics Section, ASA Social Statistics Section
Organizer: Constantine Gatsonis, Brown University
Chair: Constantine Gatsonis, Brown University

1:45 **Comparative Effectiveness Research: Promises and Challenges**
Kathleen N. Lohr*, RTI International
2:10 **Comparative Effectiveness Research: The Role of Research Synthesis**
Joel B. Greenhouse*, Carnegie Mellon University
2:35 **Comparative Effectiveness of Hip Replacement Systems**
Sharon-Lise T. Normand*, Harvard Medical School and Harvard School of Public Health, Danica Marinac-Dabic and Art Sedrakyan, Center for Devices and Radiological Health, U.S. Food and Drug Administration
3:00 **How to Compare the Effectiveness of Hypothetical Interventions (Hint: First Specify the Interventions)**
Miguel A. Hernan*, Harvard University
3:25 **Floor Discussion**

38. ADVANCES/CHALLENGES IN JOINTLY MODELING MULTIVARIATE LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA
Rosedown Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Joanna H. Shih, National Cancer Institute
Chair: Joanna H. Shih, National Cancer Institute

1:45 **Sample Size and Power Determination in Joint Modeling of Longitudinal and Survival Data**
Joseph G. Ibrahim*, Liddy Chen and Haitao Chu, University of North Carolina-Chapel Hill
2:15 **Predicting Renal Graft Failure using Multivariate Longitudinal Profiles**
Geert Verbeke* and Steffen Fieuws, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium
2:45 **An Analysis for Jointly Modeling Mulitvariate Longitudinal Measurements and Time-to-Event Data**
Paul S. Albert*, Eunice Kennedy Shriver National Institute of Child Health and Human Development and Joanna H. Shih, National Cancer Institute
3:15 **Discussant**
Geert Molenberghs, Universiteit Hasselt & Katholieke Universiteit Leuven, Belgium

39. STATISTICAL MODELS AND PRACTICE THAT IMPROVE REPRODUCIBILITY IN GENOMICS RESEARCH
Magnolia Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Rob Scharpf, Johns Hopkins University
Chair: Rob Scharpf, Johns Hopkins University

1:45 **The Importance of Reproducibility in High-Throughput Biology: Some Case Studies**
Keith A. Baggerly* and Kevin R. Coombes, University of Texas M.D. Anderson Cancer Center
2:15 **Normalization using Negative Control Features**
Zhijin Wu* and Yunxia Sui, Brown University
2:45 **Statistical Reproducibility in Clinical Genomics**
Jeffrey T. Leek*, Johns Hopkins University and John D. Storey, Princeton University
3:15 **Floor Discusssion**

40. DYNAMIC TREATMENT REGIMES AND REINFORCEMENT LEARNING IN CLINICAL TRIALS
Jasperwood Room (3rd Floor)
Sponsor: IMS
Organizer: Michael Kosorok, University of North Carolina-Chapel Hill
Chair: Michael Kosorok, University of North Carolina-Chapel Hill

1:45 **Model-checking for Semiparametric Estimation of Optimal Dynamic Treatment Regimes**
Erica Moodie*, Benjamin Rich and David A. Stephens, McGill University
2:15 **Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes**
Bibhas Chakraborty*, Columbia University, Susan A. Murphy and Victor J. Strecher, University of Michigan
2:45 **Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer**
Yufan Zhao*, Amgen Inc., Michael R. Kosorok, Donglin Zeng and Mark A. Socinski, University of North Carolina-Chapel Hill
3:15 **Floor Discussion**

41. CONTRIBUTED PAPERS: MULTIVARIATE ANALYSIS
Chequers Room (2nd Floor)
Sponsor: ENAR
Chair: Ali Shojaie, University of Michigan

1:45 **Continuity and Analysis of Principal Components**
Ahmad Reza Soltani*, Fatemah Alqallaf and Noriah Alkandari, Kuwait University
2:00 **Convergence and Prediction of Principal Component Scores in High-Dimensional Settings**
Seunggeun Lee*, Fei Zou and Fred A. Wright, University of North Carolina-Chapel Hill
2:15 **Sufficient Dimension Reduction in Regression and Applications to SNP datasets**
Kofi P. Adragni*, University of Alabama-Birmingham
2:30 **Between Estimator in the Intraclass Correlation Model with Missing Data**
Mixia Wu, Beijing University of Technology and Kai F. Yu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development

2:45 **Distribution-free Tests of Mean Vectors and Covariance Matrices for Multivariate Paired Data**
Erning Li*, Texas A&M University, Johan Lim, Kyunga Kim, and Shin-Jae Lee, Seoul National University, Korea

● **3:00** **Inference for Factor Analysis with Large p and Small n: Understanding the Change Patterns of the US Cancer Mortality Rates**
Miguel Marino* and Yi Li, Harvard University

3:15 **Use of Factor Analysis in Medical Education**
Jay Mandrekar*, Mayo Clinic

42. CONTRIBUTED PAPERS: BIOPHARMACEUTICAL METHODS

Oak Alley Room (3rd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Byron J Gajewski, University of Kansas

1:45 **An Extended F-test for Biosimilarity of Variability to Assess Follow-on Biologics**
Jun Yang* and Nan Zhang, Amgen, Inc., Shein-Chung Chow, Duke University and Eric Chi, Amgen, Inc.

2:00 **Bioequivalence Analyses for Replicated Crossovers: Structured Covariance?**
Donna L. Kowalski*, Astellas Pharma and Devan V. Mehrotra, Merck & Co., Inc.

2:15 **Stochastic Differential Equations with Positive Solutions in Modeling and Design of Pharmacokinetic Studies**
Valerii Fedorov and Sergei Leonov*, GlaxoSmithKline

2:30 **A Likelihood Based Method for Signal Detection in Safety Surveillance with Application to FDA's Drug Safety Data**
Lan Huang*, Jyoti Zalkikar and Ram Tiwari, U.S. Food and Drug Administration

2:45 **Group Sequential Methods for Observational data Incorporating Confounding through Estimating Equations with Application in Post-Marketing Vaccine/Drug Surveillance**
Andrea J. Cook* and Jennifer C. Nelson, University of Washington and Ram C. Tiwari, U.S.Food and Drug Administration

3:00 **Improved Analysis of 2x2 Crossover Trials with Potentially Missing Data**
Devan V. Mehrotra*, Merck Research Laboratories, Yu Ding, Yang Liu, John Palcza

3:15 **On the Relationship Between the Distribution of Batch Means and the Distribution of Batch Shelf Lives in Estimating Product Shelf Life**
Michelle Quinlan* and Walt Stroup, University of Nebraska-Lincoln, Dave Christopher, Schering-Plough Corporation and James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc.

43. CONTRIBUTED PAPERS: CLINICAL TRIALS WITH ADAPTIVE SAMPLE SIZES

Newberry Room (3rd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Haoda Fu, Eli Lilly and Company

1:45 **Application of Group Sequential Methods with Response Adaptive Randomization Design for Comparing the Treatment Effect with Binary Outcomes: An Evaluation of Bayesian Decision Theory Approach**
Fei Jiang*, J. Jack Lee and Peter Mueller, University of Texas M.D. Anderson Cancer Center

2:00 **Adjustment of Patient Recruitment in the Bayesian Setting**
Frank V. Mannino*, Valerii Fedorov and Darryl Downing, GlaxoSmithKline

2:15 **Bayesian Adaptive Designs for Phase III Cardiovascular Safety**
Jason T. Connor* and Scott M. Berry, Berry Consultants

2:30 **Adaptive Increase in Sample Size When Interim Results are Promising**
Cyrus R. Mehta*, President, Cytel Inc. and Stuart J. Pocock, London School of Hygiene and Tropical Medicine

2:45 **Design of Sequential Probability Likelihood Ratio Test Methodology for Poisson GLMM with Applications to Multicenter Randomized Clinical Trials**
Judy X. Li* and Daniel R. Jeske, University of California-Riverside and Jeffrey A. Klein, University of California-Irvine

3:00 **Predicting Number of Events in Clinical Trials When Treatment Arms are Masked**
Yu Gu*, Florida State University, Liqiang Yang and Ke Zhang, Pfizer Inc., Debajyoti Sinha, Florida State University

3:15 **Open-Source Simulation Experiment Platform for Evaluating Clinical Trial Designs**
Yuanyuan Wang* and Roger S. Day, University of Pittsburgh

44. CONTRIBUTED PAPERS: CAUSAL INFERENCE: METHODOLOGY AND APPLICATIONS

Eglinton Winton Room (2nd Floor)
Sponsors: ASA Biometrics Section/ASA Health Policy Statistics Section/ASA Social Statistics Section/ASA Section on Statistics in Epidemiology
Chair: Hui Nie, University of Pennsylvania

1:45 **Validation of Surrogate Outcomes using a Causal Inference Framework**
Andreas G. Klein*, University of Western Ontario

2:00 **Entire Matching with Fine Balance and its Application in Psychiatry**
Frank B. Yoon*, Harvard Medical School

2:15 **Causal Surrogacy Assessment in a Meta Analysis of Colorectal Clinical Trials**
Yun Li*, Jeremy MG Taylor, Michael R. Elliott and Bhramar Mukherjee, University of Michigan

2:30 **Challenges in Evaluating the Efficacy of a Malaria Vaccine**
Dylan S. Small*, University of Pennsylvania, Jing Cheng, University of Florida and Thomas R. Ten Have, University of Pennsylvania

2:45 **Cross-time Matching in an Observational Study of Smoking Cessation**
Bo Lu and Chih-Lin Li*, The Ohio State University

3:00 **G-Computation Algorithm for Smoking Cessation Data**
Shira I. Dunsiger*, Centers for Behavioral and Preventative Medicine, The Miriam Hospital, Joseph W. Hogan and Bess H. Marcus, Brown University

3:15 **Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants**
Mike Baiocchi*, Dylan Small, Scott Lorch and Paul Rosenbaum, University of Pennsylvania

45. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS METHODOLOGY
Cambridge Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Junhee Han, University of Arkansas

1:45 **Landmark Prediction of Survival**
Layla Parast* and Tianxi Cai, Harvard University

2:00 **Proportional Hazards and Threshold Regression: Their Theoretical and Practical Connections**
George A. Whitmore, McGill University, Montreal

2:15 **Parameter Estimations for Generalized Exponential Distribution under Progressive Type-I Interval Censoring**
Din Chen*, Georgia Southern University and Yuhlong Lio, University of South Dakota

2:30 **Partly Proportional Single-Index Model For Censored Survival Data**
Kai Ding*, Michael R. Kosorok, Donglin Zeng and David B. Richardson, University of North Carolina-Chapel Hill

2:45 **Stochastic Frailty Model Induced by Time Dependent Covariates**
Lyrica Xiaohong Liu*, Alex Tsodikov and Susan Murray, University of Michigan

3:00 **Floor Discussion**

46. CONTRIBUTED PAPERS: STATISTICAL GENETICS: COMPLEX DISEASES AND LINKAGE ANALYSIS
Elmwood Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Michael C. Wu, University of North Carolina-Chapel Hill

1:45 **Random Forests and MARS for Detecting SNP-SNP Interactions in Complex Diseases**
Lin Hui-Yi*, Moffitt Cancer Center & Research Institute

2:00 **Theoretical Basis for Haplotyping Complex Diseases**
Li Zhang*, Cleveland Clinic, Jiangtao Luo, University of Florida and Rongling Wu, Penn State University

2:15 **Bayesian Variable Selection with Biological Prior Information**
Deukwoo Kwon*, National Cancer Institute

2:30 **Assessing Statistical Significance in Genetic Linkage Analysis with the Variance Components Model**
Gengxin Li* and Yuehua Cui, Michigan State University

2:45 **Assessing Genetic Association in Case-Control Studies with Unmeasured Population Structure**
Yong Chen*, Kung-Yee Liang, Terri H. Beaty and Kathleen C. Barnes, Johns Hopkins University

3:00 **A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants**
Fang Han* and Wei Pan, University of Minnesota

3:15 **Epistatic Interactions**
Tyler J. VanderWeele*, Harvard University

3:30 **Floor Discussion**

47. CONTRIBUTED PAPERS: FUNCTIONAL DATA ANALYSIS
Belle Chasse Room (3rd Floor)
Sponsor: ASA Statistical Learning and Data Mining Section
Chair: Nan Chen, George Mason University

1:45 **Robust Functional Mixed Models**
Hongxiao Zhu*, University of Texas M.D. Anderson Cancer Center, Philip J. Brown, University of Kent, Canterbury, UK. and Jeffrey S. Morris, University of Texas M.D. Anderson Cancer Center

2:00 **Multilevel Functional Principal Component Analysis for Sparsely Sampled Hierarchical Curves**
Chongzhi Di*, Fred Hutchinson Cancer Research Center and Ciprian M. Crainiceanu, Johns Hopkins University

2:15 **Functional Data Analysis via Multiple Principle Components Variables**
Andrew Redd*, Texas A&M University

2:30 **Semiparametric Bayes Multivariate Functional Data Clustering with Variable Selection**
Yeonseung Chung* and Brent Coull, Harvard University

2:45 **Clustering Analysis of fMRI Time Series using Wavelets**
Cheolwoo Park*, Jinae Lee, Benjamin Austin, Kara Dyckman, Qingyang Li, Jennifer McDowell and Nicole A. Lazar, University of Georgia

3:00 **A Bayesian Approach on Smoothing and Mapping Functional Connectivity for Event-Related fMRI Time Series**
Dongli Zhou*, University of Pittsburgh and Wesley Thompson, University of California-San Diego

3:15 **Cross-Correlation Analysis of Spatio-Temporal Processes**
Huijing Jiang* and Nicoleta Serban, Georgia Institute of Technology

MONDAY, MARCH 22

3:30—3:45 p.m.
Refreshment Break and Visit the Exhbitors
Court Assembly Foyer (3rd Floor)

Monday, March 22
3:45—5:30 p.m.

48. CELEBRATING 70: THE CONTRIBUTIONS OF DONALD A. BERRY TO STATISTICAL SCIENCE AND STATISTICAL PRACTICE IN ACADEMICS, INDUSTRY AND GOVERNMENT
Versailles Ballroom (3rd Floor)
Sponsor: ENAR
Organizers: Lurdes Inoue, University of Washington and Dalene Stangl, Duke University
Chair: Dalene Stangl, Duke University

3:45 **Don Berry, Statistics, and Other Dangerous Things**
Michael Krams*, Pfizer Inc.
4:10 **Don Berry's Impact in the Design and Analysis of Medical Device Clinical Trials in the Regulatory Setting**
Telba Irony*, Center for Devices and Radiological Health, U.S. Food and Drug Administration
4:35 **Using Statistics to Fight Cancer: Examples from Don Berry's Career.**
Giovanni Parmigiani*, Dana-Farber Cancer Institute
5:00 **Bandits, Stopping Rules and Multiplicity**
James Berger*, Duke University
5:25 **Floor Discussion**

49. CURRENT ISSUES IN STATISTICAL PROTEOMICS
Melrose Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section
Organizer: Somnath Datta, University of Louisville
Chair: Maiying Kong, University of Louisville

3:45 **International Competition on Proteomic Diagnosis**
Bart Mertens*, Leiden University Medical Center
4:15 **Monoisotopic Peak Detection and Disease Classification for Mass Spectrometry Data**
Susmita Datta*, University of Louisville and Mourad Atlas, U.S. Food and Drug Administration
4:45 **Multiple Testing Issues and Dimension Reduction in Proteomics**
Francoise Seillier-Moiseiwitsch*, Georgetown University Medical Center
5:15 **Discussant:** Somnath Datta, University of Louisville

50. SURVEY OF METHODOLOGIES FOR POPULATION ANALYSIS IN PUBLIC HEALTH
Rosedown Room (3rd Floor)
Sponsor: ASA Social Statistics Section
Organizer: Andrew Gelman, Columbia University
Chair: Martin Lindquist, Columbia University

3:45 **Statistics Can Lie But Can Also Correct for Lies: Reducing Response Bias in NLAAS via Bayesian Imputation**
Jingchen Liu*, Columbia University, Xiao-Li Meng, Harvard University, Margarita Alegria, Cambridge Health Alliance and Harvard University and Chih-Nan Chen, Cambridge Health Alliance
4:15 **Latent Space Models for Aggregated Relation Data for the Study of High Risk Populations of HIV+/AIDS**
Tian Zheng* and Tyler H. McCormick, Columbia University
4:45 **Sampling Unsettled Populations**
David Banks, Duke University
5:15 **Floor Discussion**

51. PREDICTION AND MODEL SELECTION WITH APPLICATIONS IN BIOMEDICINE
Magnolia Room (3rd Floor)
Sponsor: IMS
Organizer: Yi Li, Harvard University
Chair: Miguel Marino, Harvard University

10:30 **Feature Selection with in Ultradimensional Statistical Problems**
Jianqing Fan*, Princeton University
11:00 **Adaptive Index Models**
Lu Tian* and Robert Tibshirani, Stanford University
11:30 **Variable Selection in Censored Quantile Regression**
Huixia Judy Wang*, North Carolina State University
12:00 **Floor Discussion**

52. CONTRIBUTED PAPERS: MISSING DATA
Jasperwood Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section, ASA Social Statistics Section
Chair: Geert Molenberghs, Universiteit Hasselt & Katholieke Universiteit, Belgium

3:45 **Binary Regression Analysis with Covariate Subject to Detection Limit**
Chunling Liu*, Aiyi Liu and Paul Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development
4:00 **Parametric Fractional Imputation for Missing Data Analysis**
Jae-kwang Kim*, Iowa State University
4:15 **The Convergence of Multiple Imputation Algorithms Using a Sequence of Regression Models**
Jian Zhu* and Trivellore E. Raghunathan, University of Michigan
4:30 **Logistic Regression Models with Monotone Missing Covariates**
Qixuan Chen* and Myunghee C. Paik, Columbia University
4:45 **Subsample Ignorable Maximum Likelihood for Regression with Missing Data**
Nanhua Zhang and Roderick J. Little, University of Michigan
5:00 **Multiple Imputation for Regression Analysis with Measurement Error in a Covariate**
Ying Guo* and Roderick Little, University of Michigan

5:15 **Informative Model Specification Test using Coarsened Data**
Xianzheng Huang*, University of South Carolina

53. CONTRIBUTED PAPERS: INFERENCE FOR CLINICAL TRIALS
Chequers Room (2nd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Irina Ostrovnaya, Memorial Sloan-Kettering Cancer Center

3:45 **Methods to Test Mediated Moderation in Logistic Regression: An Application to the TORDIA Clinical Trial**
Kaleab Z. Abebe*, Satish Iyengar and David A. Brent, University of Pittsburgh

4:00 **Exact Tests Using Two Correlated Binomial Variables in Contemporary Cancer Clinical Trials**
Jihnhee Yu*, University at Buffalo, James Kepner, American Cancer Society and Renuka Iyer, Roswell Park Cancer Institute

4:15 **Hypothesis Testing Problem for Three-arm Non-inferiority Trials**
Xiaochun Li*, New York University, Yongchao Ge, Mount Sinai School of Medicine and Judith D. Goldberg, New York University

4:30 **An Empirical Likelihood Approach to Nonparametric Covariate Adjustment in Randomized Clinical Trials**
Xiaoru Wu* and Zhiliang Ying, Columbia University

4:45 **Incorporating the Risk Difference into Non-Inferiority Trials of Safety**
Kristine R. Broglio*, Berry Consultants and Texas A&M University, Jason T. Connor and Scott M. Berry, Berry Consultants

● **5:00** **Post-randomization Interaction Analyses in Clinical Trials with Standard Regression**
Rongmei Zhang*, Jennifer Faerber, Marshall Joffe and Tom Ten Have, University of Pennsylvania

5:15 **Estimating Treatment Effects in Randomized Clinical Trials with Non-compliance and Missing Outcomes**
Yan Zhou*, Food and Drug Administration, Jack Kalbfleisch and Rod Little, University of Michigan

54. CONTRIBUTED PAPERS: METHODS FOR IMAGE DATA AND TIME SERIES
Eglinton Winton Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Jian Kang, University of Michigan

3:45 **fMRI Analysis via Bayesian Variable Selection with a Spatial Prior**
Jing Xia*, Feng Liang and Yongmei M. Wang, University of Illinois at Urbana-Champaign

4:00 **Wavelet Thresholding Using Oracle False Discovery Rate with Application to Functional Magnetic Resonance Imaging**
Nan Chen* and Edward J. Wegman, George Mason University

4:15 **A Semiparametric Heterogeneous-Mixture Model for Certain Quantum-Dot Images, and Likelihood Inference for Dot Count and Location**
John Hughes* and John Fricks, Penn State University

4:30 **The Generalized Shrinkage Estimator for Partial Coherence Estimation in Multivariate Time Series**
Mark Fiecas* and Hernando Ombao, Brown University

4:45 **Elastic-nt Based Model for Imaging MS Proteomic Data Processing**
Fengqing Zhang* and Don Hong, Middle Tennessee State University

5:00 **Subdiffusion Detection in Microrheological Experiments**
Gustavo Didier*, Tulane University and John Fricks, Penn State University

5:15 **Floor Discussion**

55. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS: CURE MODELS AND COMPETING RISKS
Oak Alley Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Ori Stitelman, University of California-Berkeley

3:45 **Semiparametric Regression Cure Models for Interval-Censored Data**
Hao Liu*, Baylor College of Medicine and Yu Shen, University of Texas M. D. Anderson Cancer Center

4:00 **Incorporating Short-term Effects in Cure Models with Bounded Cumulative Hazards**
Xiang Liu* and Li-Shan Huang, University of Rochester Medical Center

4:15 **Cure Rate Model with Nonparametric Spline Estimated Components**
Lu Wang and Pang Du*, Virginia Tech University

4:30 **Cause-specific Association Measures for Multivariate Competing Risks Data and Their Nonparametric Estimators**
Yu Cheng and Hao Wang*, University of Pittsburgh

4:45 **Semiparametric Analysis of Competing Risks Model with a Misattribution of Cause of Death**
Jinkyung Ha* and Alex Tsodikov, University of Michigan

5:00 **Regression Strategy for the Conditional Probability of a Competing Risk**
Aurelien Latouche*, University of Versailles and European Group for Blood and Marrow Transplantation

5:15 **Number Needed to Treat for Time to Event Data with Competing Risks**
Suprateek Kundu* and Jason P. Fine, University of North Carolina-Chapel Hill

56. CONTRIBUTED PAPERS: STATISTICAL GENETICS: QUANTITATIVE TRAIT LOCI
Elmwood Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Ruzong Fan, Texas A&M University

3:45 **Bayesian Nonparametric Multivariate Statistical Models for Quantitative Traits and Candidate Genes Association Tests in Structured Populations**
Meijuan Li*, U.S. Food and Drug Administration and Tim Hanson, University of Minnesota

4:00 **Robust Score Statistics for QTL Linkage Analysis using Extended Pedigrees**
Chia-Ling Kuo* and Eleanor Feingold, University of Pittsburgh

4:15 **Identifying QTLs in Crop Breeding Populations Using Adaptive Mixed LASSO**
Dong Wang* and Kent M. Eskridge, University of Nebraska and Jose Crossa, International Maize and Wheat Improvement Center

4:30 **An Approach to Testing Pleiotropy with Quantitative Traits in Genome-wide Association Studies**
Emily Kistner-Griffin*, Medical University of South Carolina, Nancy J. Cox and Dan L. Nicolae, University of Chicago

4:45 **Enriching our Knowledge in Gene Regulation via eQTL Mapping: A Combined p-value Approach**
Shaoyu Li* and Yuehua Cui, Michigan State University

5:00 **Comparison of Methods for gXg Interaction for Quantitative Traits in Case-Control Association Studies**
Raymond G. Hoffmann* and Soumitra Ghosh, Medical College of Wisconsin, Thomas J. Hoffmann, University of California-San Francisco and Pippa M. Simpson, Medical College of Wisconsin

5:15 **Mapping Quantitative Trait Loci for Time-to-Event Phenotype with Cured Individuals**
Chi Wang*, University of California-Riverside, Zhiqiang Tan, Rutgers University and Thomas A. Louis, Johns Hopkins University

57. CONTRIBUTED PAPERS: MICROARRAYS: TIME COURSE AND GENE SETS
Belle Chasse Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Xi Kathy Zhou, Cornell University

3:45 **A Novel Approach in Testing for Periodicity in Cell-Cycle Gene Expression Profiles**
Mehmet Kocak*, St. Jude Children's Research Hospital, E. Olusegun George, University of Memphis and Saumyadipta Pyne, MIT and Harvard University

4:00 **A Unified Mixed Effects Model for Gene Set Analysis of Time Course Microarray Experiments**
Lily Wang* and Xi Chen, Vanderbilt University and Russell D. Wolfinger, SAS Institute Inc.

4:15 **Rank Based Gene Selection for Classification**
Shuxin Yin* and Asheber Abebe, Auburn University

4:30 **Adaptive Prediction in Genomic Signatures-Based Clinical Trials**
Yang Xie*, Guanghua Xiao, Chul Ahn, Luc Girard and John Minna, University of Texas Southwestern Medical Center

4:45 **Stereotype Logit Models for High-dimensional Data**
Andre A.A. Williams* and Kellie J. Archer, Virginia Commonwealth University

● **5:00** **Likelihood Based Approach to Gene Set Enrichment Analysis with a Finite Mixture Model**
Sang Mee Lee* and Baolin Wu, University of Minnesota

58. CONTRIBUTED PAPERS: EXPERIMENTAL DESIGN, POWER/SAMPLE SIZE AND SURVEY RESEARCH
Cambridge Room (2nd Floor)
Sponsors: ASA Biopharmaceutical Section, ASA Social Statistics Section
Chair: Xiangqin Cui, University of Alabama-Birmingham

3:45 **Sequential Design for Microarray Studies**
Laurent Briollais*, Mount Sinai Hospital, Toronto and Gilles Durrieu, University of Bordeaux, France

4:15 **Tests for Unequal Treatment Variances in Crossover Designs**
Yoon-Sung Jung*, Alcorn State University and Dallas E. Johnson, Kansas State University

4:30 **Optimal Designs for Response Functions with a Downturn**
Seung Won Hyun*, Min Yang and Nancy Flournoy, University of Missouri

4:45 **Power Analysis for Longitudinal Studies with Time Dependent Covariate**
Cuiling Wang*, Albert Einstein College of Medicine

5:00 **The Use of Percentiles for Estimating Variability of Normal and Uniform Distributions**
Chand K. Chauhan* and Yvonne M. Zubovic, Indiana University-Purdue University, Fort Wayne

5:15 **The Impact of Survey Order on Identifiability of Response Fatigue and Estimation of Treatment Effects**
Brian L. Egleston*, Fox Chase Cancer Center

TUESDAY, MARCH 23

8:30—10:15 a.m.

59. COMBINING MULTIPLE SOURCES OF EXPOSURE DATA IN HEALTH ENVIRONMENT STUDIES
Melrose Room (3rd Floor)
Sponsor: ENAR
Organizers: Brent Coull, Harvard University and Sylvia Richardson, Imperial College London, UK
Chair: Brent Coull, Harvard University

8:30 **Bayesian Graphical Models for Combining Multiple Data Sources, with Applications in Environmental Epidemiology**
Sylvia Richardson* and Alexina Mason, Imperial College London, UK, Lawrence McCandless, Simon Fraser University, Canada and Nicky Best, Imperial College London, UK

9:00 **Nonlinear Latent Process Models for Addressing Temporal Change of Support in Spatio-temporal Studies of Environmental Exposures**
Nikolay Bliznyuk*, Texas A&M University, Christopher Paciorek, University of California-Berkeley and Brent Coull, Harvard University

9:30 **Mortality Risks of Short and Long-term Exposure to Chemical Composition of Fine Particulate Air Pollution (2000-2007): Statistical Challenges**
Francesca Dominici*, Harvard University

10:00 **Floor Discussion**

60. IMAGINING 2025: THE HOPES FOR CLINICAL TRIALS
Versailles Balloom (3rd Floor)
Sponsor: ASA Biopharmaceutical Section
Organizers: Dalene Stangl, Duke University, Lurdes Inoue, University of Washington and Telba Irony, U.S. Food and Drug Administration
Chair: Joseph (Jay) Kadane, Carnegie Mellon University

8:30 **Predicting the Predictable? Clinical trials in 2025**
Steven N. Goodman*, Johns Hopkins University
9:00 **Peering into the Hopeful Crystal Ball: Clinical Trials in 2025**
Janet Wittes*, Statistics Collaborative
9:30 **A Bayesian 21st Century?**
Donald A. Berry*, University of Texas M D Anderson Cancer Center
10:00 **Discussant:** Dalene Stangl, Duke University

61. INNOVATIVE METHODS IN BIOSURVEILLANCE: ACCURATE METHODS FOR RAPID DETECTION OF EVENTS AND PATTERNS
Rosedown Room (3rd Floor)
Sponsor: ENAR
Organizer: Amy Herring, University of North Carolina-Chapel Hill
Chair: Yingqi Zhao, University of North Carolina-Chapel Hill

8:30 **Using Spatiotemporal Regression Methods To Identify Causes of Disease Outbreaks**
Michael R. Kosorok*, Yingqi Zhao, Donglin Zeng, Amy H. Herring and David Richardson, University of North Carolina-Chapel Hill
9:00 **Fast Subset Scanning for Multivariate Event Detection**
Daniel B. Neill*, Carnegie Mellon University
9:30 **Adjustments for Secondary Multiplicity Problems with Scan Statistics**
Ronald Gangnon*, University of Wisconsin-Madison
10:00 **Floor Discussion**

62. INNOVATIVE STATISTISTICAL METHODS FOR FUNCTIONAL AND IMAGE DATA
Magnolia Room (3rd Floor)
Sponsor: IMS
Organizer: Jeffrey Morris, University of Texas M.D. Anderson Cancer Center
Chair: Hongxiao Zhu, University of Texas M.D. Anderson Cancer Center

8:30 **Bayesian Sparse Factor Models for Multivariate Functional Data**
David B. Dunson*, Duke University
9:00 **Automated, Robust Analysis of Functional and Quantitative Image Data using Functional Mixed Models and Isomorphic Basis-Space Modeling**
Jeffrey S. Morris*, Veerabhadran Baladandayuthapani and Hongxiao Zhu, University of Texas M.D. Anderson Cancer Center
9:30 **Reduced-rank t Models for Robust Functional Data Analysis**
Daniel Gervini*, University of Wisconsin-Milwaukee
10:10 **Floor Discussion**

63. CONTRIBUTED PAPERS: JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA
Cambridge Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Alex McLain, National Institutes of Health

8:45 **An Estimation Method of Marginal Treatment Effects on Correlated Longitudinal and Survival Outcomes**
Qing Pan*, George Washington University and Grace Y. Yi, University of Waterloo
9:00 **Joint Modeling of Longitudinal and Survival Data Subject to Non-ignorable Missing and Left-censoring in Repeated Measurements**
Abdus Sattar*, University of Pittsburgh
9:15 **Joint Modeling of Longitudinal and Time to Event Data with Random Changepoints**
Chengjie Xiong* and Yuan Xu, Washington University
9:30 **Joint Modeling of the Relationship between Longitudinal and Survival Data Subject to Both Left Truncation and Right Censoring with Applications to Cystic Fibrosis**
Mark D. Schluchter and Annalisa VanderWyden Piccorelli*, Case Western Reserve University
9:45 **Semiparametric Estimation of Treatment-Free Restricted Mean Lifetime using Landmark Analysis with a Partly Conditional Model**
Qi Gong* and Douglas E. Schaubel, University of Michigan
10:00 **Joint Models of Longitudinal Data and Recurrent Events with Informative Terminal Event**
Se Hee Kim*, Donglin Zeng and Lloyd Chambless, University of North Carolina-Chapel Hill

64. CONTRIBUTED PAPERS: HEALTH SERVICES RESEARCH
Eglinton Winton Room (2nd Floor)
Sponsor: ASA Health Policy Statistics Section
Chair: Jing Zhang, Miami University

8:30 **A Logistic Regression Model of Obesity in Pre-school Children**
MaryAnn Morgan-Cox*, Baylor University, Veronica Piziak, M.D., Scott & White Medical Center, Jack D. Tubbs, James D. Stamey and John W. Seaman, II, Baylor University
8:45 **Assessing Content Validity through Correlation and Relevance Tools: A Bayesian Randomized Equivalency Experiment**
Byron J. Gajewski*, Valorie Coffland, Diane K. Boyle, Marjorie Bott, Jamie Leopold and Nancy Dunton, University of Kansas

9:00 **Hierarchical Bayesian Models to Quantify Hospital Performance**
Yulei He*, Sharon-Lise T. Normand and Robert Wolf, Harvard Medical School

9:15 **Analysis of Interval-grouped Recurrent Event Data with Application to National Hospitalization Data**
Dandan Liu*, Jack D. Kalbfleisch and Douglas E. Schaubel, University of Michigan

9:30 **Implementation of a Kronecker Product Correlation Structure for the Analysis of Unbalanced Data**
Arwin M. Thomasson*, University of Pennsylvania, Hanjoo Kim, Forest Labs, Justine Shults, Russell Localio, Harold I. Feldman and Peter P. Reese, University of Pennsylvania

9:45 **Power in the Design of Two-Level Randomized Trials**
Yongyun Shin*, Virginia Commonwealth University

10:00 **Floor Discussion**

65. CONTRIBUTED PAPERS: MODELS INVOLVING LATENT VARIABLES
Jasperwood Room (3rd Floor)
Sponsor: ASA Social Statistics Section
Chair: Lee Hye-Seung, University of South Florida

8:30 **A Penalized Maximum Likelihood Approach to Sparse Factor Analysis**
Jang Choi*, Hui Zou and Gary Oehlert, University of Minnesota

8:45 **Modern Cluster Methods for Incomplete Longitudinal Data**
Liesbeth Bruckers* and Geert Molenberghs, Hasselt University

9:00 **Latent Variable Models for Development of Composite Indices**
Xuefeng Liu*, Meng Liu and Kesheng Wang, East Tennessee State University and Jeffray Roth, University of Florida

9:15 **Efficiency of Likelihood Based Estimators in the Analysis of Familial Binary Variables**
Yihao Deng*, Indiana University Purdue University-Fort Wayne, Roy T. Sabo, Virginia Commonwealth University and N. Rao Chaganty, Old Dominion University

9:30 **A Structured Latent Class Model Versus a Factor Mixture Model for Exploring Clusters of Obesogenic Behaviors and Environments in Adolescents**
Melanie M. Wall*, University of Minnesota

9:45 **Confidence Intervals for the Difference in TPRs and FPRs of Two Diagnostic Tests with Unverified Negatives**
Eileen M. Stock*, Baylor University

10:00 **Sparse Bayesian Infinite Factor Models**
Anirban Bhattacharya* and David B. Dunson, Duke University

66. CONTRIBUTED PAPERS: NEXT GENERATION SEQUENCING AND TRANSCRIPTION FACTOR BINDING SITES
Oak Alley Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Peng Wei, University of Texas

8:30 **Identification of miRNAs in Next-generation Sequencing Data**
W. Evan Johnson*, Brigham Young University

8:45 **Bayesian Hierarchical Models for Quantifying Methylation Levels by Next-generation Sequencing**
Guodong Wu and Nengjun Yi, University of Alabama-Birmingham, Devin Absher, HudsonAlpha Institute for Biotechnology and Degui Zhi*, University of Alabama-Birmingham

9:00 **Gene Class Enrichment Analysis for RNA-sequencing**
Liyan Gao, Degui Zhi, Kui Zhang and Xiangqin Cui*, University of Alabama-Birmingham

9:15 **A Statistical Framework for the Analysis of ChIP-Seq Data**
Pei Fen Kuan*, University of North Carolina-Chapel Hill, Guangjin Pan, Genome Center of Wisconsin, James A. Thomson, Ron Stewart and Sunduz Keles, University of Wisconsin-Madison

9:30 **A Comparison of Monte-Carlo Logic and LogicFS Regression Methods for Identifying Important Co-Regulators of Gene Expression with Application to a Study of Human Heart Failure**
Yun Lu*, Sridhar Hannenhalli, Thomas Cappola and Mary Putt, University of Pennsylvania School of Medicine

9:45 **Detection and Refinement of Transcription Factor Binding Sites Using Hybrid Monte Carlo Method**
Ming Hu*, University of Michigan, Jindan Yu, University of Michigan and Northwestern University, Jeremy Taylor, Arul Chinnaiyan and Zhaohui Qin, University of Michigan

10:00 **Floor Discussion**

67. CONTRIBUTED PAPERS: VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA
Elmwood Room (3rd Floor)
Sponsor: ASA Statistical Learning and Data Mining Section
Chair: Peng Zeng, Auburn University

8:30 **Random Forests: A Summary of the Algorithm and a Description of the Features that are Available, Including a Presentation of Recent Work on Variable Importance and Proximities**
Adele Cutler*, Utah State University

8:45 **Nonparametric Independence Screening in Ultra-high Dimensional Additive Models**
Jianqing Fan and Yang Feng*, Princeton University and Rui Song, Colorado State University

9:00 **Grouped Variable Selection in High-Dimensional Partially Linear Additive Cox Model**
Li Liu* and Jian Huang, University of Iowa

9:15 **Model-free Feature Selection**
Liping Zhu*, Penn State University and East China Normal University and Runze Li, Penn State University

9:30 **Generalized Forward Selection: Subset Selection in High Dimensions**
Alexander T. Pearson and Derick R. Peterson*, University of Rochester

9:45 **Feature Selection in Microarray Data using Tight Clustering**
Ami Yu* and Jae Won Lee, Korea University

68. CONTRIBUTED PAPERS: QUANTILE REGRESSION AND SYMBOLIC REGRESSION
Chequers Room (2nd Floor)
Sponsor: ENAR
Chair: Scott S. Emerson, University of Washington

8:30 **Spatial Quantile Regression**
Kristian Lum* and Alan Gelfand, Duke University

8:45 **On Rank Score Test for Longitudinal Best Line Quantile Model**
Nanshi Sha* and Ying Wei, Columbia University

9:00 **An Exact Bootstrap Approach Towards Modification of the Harrell-Davis Quantile Function Estimator for Censored Data**
Dongliang Wang*, Alan D. Hutson and Daniel P. Gaile, University of Buffalo

9:15 **Bent Line Quantile Regression with Application to an Allometric Study of Land Mammals' Speed and Mass**
Chenxi Li*, University of Wisconsin-Madison, Ying Wei, Columbia University, Rick Chappell, University of Wisconsin-Madison and Xuming He, University of Illinois at Urbana-Champaign

9:30 **Using Logistic Regression to Construct Confidence Intervals for Quantile Regression Coefficients**
Junlong Wu* and Matteo Bottai, University of South Carolina

9:45 **Detection of Treatment Effects by Covariate-adjusted Expected Shortfall and Trimmed Rankscore**
Ya-Hui Hsu*, University of Illinois at Urbana-Champaign

10:00 **Regression for Interval-valued Symbolic Data**
Wei Xu* and Lynne Billard, University of Georgia

69. CONTRIBUTED PAPERS: PERSONALIZED THERAPY AND VARIABLE SELECTION IN CLINICAL APPLICATIONS
Belle Chasse Room (3rd Floor)
Sponsors: ASA Biometrics Section/ASA Biopharmaceutical Section/ ASA Health Policy Statistics Section
Chair: Ying Yuan, University of Texas M.D. Anderson Cancer Center

8:30 **Developing Adaptive Personalized Therapy for Cystic Fibrosis by Reinforcement Learning**
Yiyun Tang* and Michael R. Kosorok, University of North Carolina-Chapel Hill

8:45 **Penalized Models for Ordinal Response Prediction: Application Discriminating Patients with Early-stage Parkinson's Disease**
Kellie J. Archer* and Andre A.A. Williams, Virginia Commonwealth University

9:00 **Screening Suboptimal Treatments using the Fused Lasso**
Eric B. Laber*, University of Michigan, Mahdi Fard and Joelle Pineau, McGill University and Susan A. Murphy, University of Michigan

9:15 **Identify Interesting Interactions in Decision Making**
Peng Zhang* and Susan A. Murphy, University of Michigan

9:30 **Examples in Epidemiology Using Advanced Data Mining Techniques: CART, MARS and TreeNet/MART**
Shenghan Lai*, Johns Hopkins University and Mikhail Golovnya, Salford Systems

9:45 **Change-Line Classification and Regression for Chemical Toxicity Analysis**
Chaeryon Kang*, Fei Zou, Hao Zhu and Michael R. Kosorok, University of North Carolina-Chapel Hill

10:00 **Floor Discussion**

TUESDAY, MARCH 23

10:15—10:30 a.m. Refreshment Break and Visit the Exhibitors
Court Assembly Foyer (3rd Floor)

Tuesday, March 23
10:30 a.m.—12:15 p.m.

70. PRESIDENTIAL INVITED ADDRESS
Napoleon Ballroom (3rd Floor)
Sponsor: Sharon-Lise Normand, Harvard University
Organizer/Chair: Sharon-Lise Normand, Harvard University

10:30 **Introduction: Sharon-Lise Normand, Harvard University**

10:35 **Distinguished Student Paper Awards**

10:45 **Bayes, BARS, and Brains: Statistics and Machine Learning in the Analysis of Neural Spike Train Data**
Robert E. Kass, Department of Statistics, Center for the Neural Basis of Cognition, Machine Learning Department, Carnegie Mellon University

Tuesday, March 23
1:45—3:30 p.m.

71. BAYESIAN METHODS IN GENOMIC RESEARCH
Melrose Room (3rd Floor)
Sponsor: ASA Section on Bayesian Statistical Sciences
Organizer: Gary L. Rosner, University of Texas M.D. Anderson Cancer Center
Chair: Gary L. Rosner, University of Texas M.D. Anderson Cancer Center

1:45 **Bayesian Approaches for Incorporating Intermediate Biomarkers in Genetic Association Studies**
David V. Conti* and Wonho Lee, University of Southern California, Rachel Tyndale, University of Toronto, Andrew Bergen and Gary Swan, SRI International

2:10 **PICS: Probabilistic Inference for ChIP-Seq**
Raphael Gottardo*, Clinical Research Institute of Montreal and University of British Columbia

2:35 **Bayesian Model-based Methods for Analyzing ChIP Sequencing Data**
Ming Hu, University of Michigan, Jindan Yu, Northwestern University Feinberg Medical School, Jeremy MG Taylor, Arul M. Chinnaiyan and Zhaohui S. Qin*, University of Michigan

3:00 **Modeling Population Haplotype Variation**
Paul Scheet*, University of Texas M. D. Anderson Cancer Center

3:25 **Floor Discussion**

72. INTERFERENCE AND SPILLOVER EFFECTS IN CAUSAL INFERENCE
Rosewood Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Social Statistics Section
Organizer: Tyler VanderWeele, University of Chicago
Chair: Tyler VanderWeele, University of Chicago

1:45 **On Interference in Inference for Causal Effects and Extensions with Application to Infectious Diseases**
Eric J. Tchetgen* and Tyler VanderWeele, Harvard University

2:15 **Strategies for Modeling Inference Between Units in Multi-site Trials**
Stephen Raudenbush*, University of Chicago

2:45 **Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers**
Bryan S. Graham*, New York University, Guido W. Imbens, Harvard University and Geert Ridder, University of Southern California

3:15 **Floor Discussion**

73. STATISTICAL METHODS IN NEUROSCIENCE
Magnolia Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Wei Wu, Florida State University
Chair: Robert Kass, Carnegie Mellon University

1:45 **A New Look at State-space Models in Neuroscience**
Liam Paninski*, Columbia University

2:10 **Multi-scale Multiple Hypothesis Testing for Spike Trains**
Matthew T. Harrison*, Brown University and Asohan Amarasingham, Rutgers University

2:35 **Tests for Differential Spiking Activity based on Point Process Models**
Uri Eden*, Boston University

3:00 **Motor Cortical Decoding using Hidden State Models**
Vernon Lawhern and Wei Wu*, Florida State University, Nicholas Hatsopoulos, University of Chicago and Liam Paninski, Columbia University

3:25 **Floor Discussion**

74. RECENT ADVANCES IN VARIABLE SELECTION METHODOLOGY
Versailles Ballroom (3rd Floor)
Sponsor: IMS
Organizer: Howard Bondell, North Carolina State University
Chair: Brian Reich, North Carolina State University

1:45 **Penalized Regression Methods for Ranking Variables by Effect Size, with Applications to Genetic Mapping Studies**
Nam-Hee Choi, Kerby Shedden and Ji Zhu*, University of Michigan

2:10 **Variable Selection and Tuning via Confidence Regions**
Howard D. Bondell* and Funda Gunes, North Carolina State University

2:35 **Computing the Solution Path of Penalized Cox Regression**
Hui Zou*, University of Minnesota

3:00 **Coordinate Descent Algorithms for Variable Selection**
Trevor J. Hastie*, Rahul Mazumder and Jerome Friedman, Stanford University

3:25 **Floor Discussion**

75. CONTRIBUTED PAPERS: BIOMARKERS AND DIAGNOSTIC TESTS
Jasperwood Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Biopharmaceutical Section
Chair: Andreas Klein, University of Western Ontario

1:45 **Using Tumor Response in Designing Efficient Cancer Clinical Trials with Overall Survival as Primary Endpoint**
Donald A. Berry and Haiying Pang*, University of Texas M. D. Anderson Cancer Center

2:00 **A Strategy to Identify Patients Sensitive to Drug-induced Adverse Events**
Wei-Jiun Lin* and James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

2:15 **Early Detection of Alzheimer's Disease using Partially Ordered Classification Models**
Curtis Tatsuoka*, Case Western Reserve University, Huiyun Tseng, Columbia University, Judith Jaeger, AstraZeneca Pharmaceuticals and Alan Lerner, Case Western Reserve University

2:30 **Challenges and Technical Issues of Assessing Trial-level Surrogacy of Putative Surrogate Endpoints in the Meta-analytic Framework for Clinical Trials**
Qian Shi*, Mayo Clinic, Lindsay A. Renfro, Baylor University, Brian M. Bot and Daniel J. Sargent, Mayo Clinic

2:45 **Probit Latent Class Models for Evaluating Accuracy of Diagnostic Tests with Indeterminate Results**
Huiping Xu*, Mississippi State University and Bruce A. Craig, Purdue University

3:00 **Bayesian Analysis and Classification of Two Quantitative Diagnostic Tests for Occaisionally Absent Markers and No Gold Standard**
Jingyang Zhang*, Kathryn Chaloner and Jack T. Stapleton, University of Iowa

76. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS IN CLINICAL TRIALS
Cambridge Room (2nd Floor)
Sponsor: ASA Biopharmaceutical Section
Chair: Din Chen, Georgia Southern University

1:45 **A Test for Equivalence of Two survival Functions in Proportional Odds Model**
Wenting Wang* and Debajyoti Sinha, Florida State University, Stuart R. Lipsitz, Harvard Medical School and Richard J. Chappell, University of Wisconsin-Madison

2:00 **A Nonparametric Test for Equality of Survival Medians**
Mohammad H. Rahbar*, University of Texas Health Science Center at Houston and University of Texas School of Public Health, Zhongxue Chen, Florida International University, Sangchoon Jeon, Yale University and Joseph C. Gardiner, Michigan State University

2:15 **Evaluation of Treatment-Effect Heterogeneity in the Age of Biomarkers**
Ann A. Lazar*, Harvard University and Dana-Farber Cancer Institute, Bernard F. Cole, University of Vermont and Dana-Farber Cancer Institute, Marco Bonetti, Bocconi University and Richard D. Gelber, Harvard School of Public Health, Harvard Medical School and Dana-Farber Cancer Institute

2:30 **Stratified and Unstratified Log-rank Tests in Multicenter Clinical Trial**
Changyong Feng*, University of Rochester

2:45 **Hypothesis Testing in Randomized Trials for Survival Data with Misspecified Regression Models**
Jane Paik*, Stanford University

3:00 **Semiparametric Estimation of Treatment Effect with Time-Lagged Response in the Presence of Informative Censoring**
Xiaomin Lu*, University of Florida and Anastasios Tsiatis, North Carolina State University

● **3:15** **A Generalized Cox Proportional Hazard Model for Comparing Dynamic Treatment Regimes**
Xinyu Tang* and Abdus S. Wahed, University of Pittsburgh

77. CONTRIBUTED PAPERS: HIGH DIMENSIONAL MODELING: SEGMENT DETECTION AND CLUSTERING
Eglinton Winton Room (2nd Floor)
Sponsor: ASA Statistical Learning and Data Mining Section
Chair: Hongyuan Cao, University of North Carolina-Chapel Hill

1:45 **Optimal Sparse Segment Identification**
Jessie Jeng*, Tony Cai and Hongzhe Li, University of Pennsylvania

2:00 **Using Scan Statistics on Dependent Signals and Assessing its Distribution, with Application to Searching Sequences of Interest along the Genome**
Anat Reiner-Benaim*, Haifa University

2:15 **A Framework for Density Estimation for Binary Sequences**
Xianyun Mao* and Bruce Lindsay, Penn State University

2:30 **Bayesian Species Clustering via DP and Sampling Designs via Monte Carlo**
Hongmei Zhang*, University of South Carolina, Kaushik Ghosh, University of Nevada-Las Vegas and Pulak Ghosh, Novartis Pharmaceuticals

2:45 **Feature-Subset-Specific Clustering Using Stochastic Search**
Younghwan Namkoong, University of Florida, Yongsung Joo*, Dongguk University, Korea and Douglas D. Dankel, University of Florida

3:00 **Floor Discussion**

78. CONTRIBUTED PAPERS: LONGITUDINAL DATA ANALYSIS
Oak Alley Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Matthew Guerra, University of Pennsylvania

1:45 **Efficient Semiparametric Regression for Longitudinal Data with Nonparametric Covariance Estimation**
Yehua Li*, University of Georgia

2:00 **A Distribution-free Association Measure for Longitudinal Data with Applications to HIV/AIDS Research**
Sujay Datta*, Li Qin and Stephen G. Self, Fred Hutchinson Cancer Research Center and the University of Washington

● **2:15** **Flexible Bent-Cable Models for Mixture Longitudinal Data**
Shahedul A. Khan*, Grace Chiu and Joel A. Dubin, University of Waterloo

2:30 **A Circular LEAR Correlation Structure for Cyclical Longitudinal Data**
Sean L. Simpson*, Wake Forest University School of Medicine and Lloyd J. Edwards, University of North Carolina-Chapel Hill

2:45 **Discrete Time-Transformation Model with Random Effects and Sterile Fraction: An Application to Time to Pregnancy**
Alexander McLain* and Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development

3:00 **A Correlated Bayesian Human Fecundability Model with Missing Covariates**
Sungduk Kim*, Rajeshwari Sundaram and Germaine B. Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development

3:15 **Bayesian Changepoint Models in Detecting Women's Menopausal Transition**
Xiaobi Huang*, Siob'an D. Harlow and Michael R. Elliott, University of Michigan

79. CONTRIBUTED PAPERS: STATISTICAL GENETICS: EPISTASIS, GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS
Elmwood Room (3rd Floor)
Sponsors: ASA Biometrics Section, ASA Section on Statistics in Epidemiology
Chair: Meijuan Li, U.S. Food and Drug Administration

1:45 **Epistasis in Genome-wide Association Studies**
Huei-Wen Teng* and Yu Zhang, Penn State University

2:00 **Entropy-based Tests for Genetic Epistasis in Genome Wide Association Studies**
Xin Wang* and Mariza de Andrade, Mayo Clinic College of Medicine

2:15 **A Dimension Reduction Approach to Detect Multilocus Interaction in a Case Control Study**
Saonli Basu*, University of Minnesota

2:30 **Estimating Gene-Environment Interaction by Pooling Biomarkers**
Michelle R. Danaher*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Anindya Roy, University of Maryland, Baltimore County, Paul Albert and Enrique Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development

2:45 **Gene-environment Interaction Testing in Family-based Association Studies with Phenotypically Ascertained Samples: A Causal Inference Approach**
David Fardo* and Yan Huang, University of Kentucky and Stijn Vansteelandt, Ghent University

● **3:00** **A General Framework for Studying Genetic Effects and Gene-Environment Interactions With Missing Data**
Yijuan Hu*, Danyu Lin and Donglin Zeng, University of North Carolina-Chapel Hill

80. CONTRIBUTED PAPERS: BAYESIAN METHODS AND APPLICATIONS
Belle Chasse Room (3rd Floor)
Sponsor: ASA Section on Bayesian Statistical Sciences
Chair: Eleanor M Pullenayegum, McMaster University

1:45 **Robustness of Nonparametric Bayesian Methods**
Steven N. MacEachern, The Ohio State University

2:00 **Modeling Relational Data Using Nested Partition Models**
Abel Rodriguez, University of California-Santa Cruz and Kaushik Ghosh, University of Nevada Las Vegas

2:15 **Bayesian Inference on Parameters of a Zero Inflated Negative Multinomial Distribution**
Santanu Chakraborty*, University of Texas Pan American

2:30 **Performance of Bayesian Ranking Methods for Identifying the Extreme Parameter**
Yi-Ting Chang* and Thomas A. Louis, Johns Hopkins University

2:45 **An Efficient Markov Chain Monte Carlo Method for Mixture Models by Neighborhood Pruning**
Youyi Fong*, Jon Wakefield and Ken Rice, University of Washington

3:00 **Nonparametric Bayes Stochastically Ordered Latent Class Models**
Hongxia Yang*, David Dunson and Sean O'brien, Duke University

3:15 **Floor Discussions**

81. CONTRIBUTED PAPERS: SPATIAL/TEMPORAL APPLICATIONS, AND INFECTIOUS DISEASE MODELING
Chequers Room (2nd Floor)
Sponsor: ENAR
Chair: Xiangming Fang, East Carolina University

● **1:45** **Bayesian Geostatistical Modeling with Informative Sampling Locations**
Debdeep Pati*, Duke University, Brian J. Reich, North Carolina State University and David B. Dunson, Duke University

● **2:00** **Independent Component Analysis for Colored Sources with Application to Functional Magnetic Resonance Imaging**
Seonjoo Lee*, Haipeng Shen and Young Truong, University of North Carolina-Chapel Hill

2:15 **A Hierarchical Model for Predicting Forest Variables over Large Heterogeneous Domains**
Andrew O. Finley*, Michigan State University and Sudipto Banerjee, University of Minnesota

2:30 **Estimating Case Fatality Ratios from Infectious Disease Surveillance Data**
Nicholas G. Reich*, Justin Lessler and Ron Brookmeyer, Johns Hopkins University

2:45 **Modeling the Spatio-temporal Transmission of Infectious Diseases in Animals**
Christel Faes*, Marc Aerts and Niel Hens, Hasselt University

3:00 **Assessing the Spatial Variability of Syphilis in Baltimore in the 1990s using Geographically Weighted Regression**
Jeffrey M. Switchenko* and Lance A. Waller, Emory University

3:15 **Detecting Disease Outbreaks Using Local Spatiotemporal Methods**
Yingqi Zhao*, Donglin Zeng, Amy H. Herring, David Richardson and Michael R. Kosorok, University of North Carolina-Chapel Hill

TUESDAY, MARCH 23

3:30—3:45 p.m.
Refreshment Break and Visit the Exhibitors
Court Assembly Room (3rd Floor)

Tuesday, March 23
3:45—5:30 p.m

82. IMS MEDALLION LECTURE
Versailles Ballroom (3rd Floor)
Sponsor: IMS
Organizer: Hao Helen Zhang, North Carolina State University
Chair: Hao Helen Zhang, North Carolina State University

3:45 **A Statistician's Adventures in Collaboration: Designing Better Treatment Strategies**
Marie Davidian, North Carolina State University

83. SHRINKAGE ESTIMATION IN MICROARRAY DATA ANALYSIS
Melrose Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Samiran Sinha, Texas A&M University
Chair: Samiran Sinha, Texas A&M University

3:45 **LEMMA: Laplace Approximated EM Microarray Analysis**
Bar Haim, James G. Booth*, Elizabeth Schifano and Martin T. Wells, Cornell University

4:10 **Borrowing Information across Genes and across Experiments for Improved Residual Variance Estimation in Microarray Data Analysis**
Tieming Ji, Peng Liu and Dan Nettleton*, Iowa State University

4:35 **Mixture Priors for Variable Selection with Application in Genomics**
Marina Vannucci*, Rice University

5:00 **Assessing Differential Gene Expression Using a Nonparametric Mean-Variance Smoothing: Application to Arabidopsis Thaliana Abiotic Stress Microarray Experiments**
Taps Maiti* and Pingsha Hu, Michigan State University

5:25 **Floor Discussion**

84. OPPORTUNITIES FOR BIOSTATISTICIANS INSIDE (RESEARCH) AND OUTSIDE (FUNDING) OF NIH
Rosedown Room (3rd Floor)
Sponsor: ENAR
Organizer: Gang Zheng, National Heart, Lung and Blood Institute
Chair: Gang Zheng, National Heart, Lung and Blood Institute

3:45 **Biostatistics at the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH)**
Nancy L. Geller*, National Heart, Lung, and Blood Institute

4:10 **My Research Experience at the NIH and UTSPH**
Sheng Luo*, University of Texas School of Public Health

4:35 **NIH Funding Opportunities for Biostatisticians**
Michelle C. Dunn*, National Cancer Institute

5:00 **Applying for Biostatistical Research Grants from NIH: Why? What? And How?**
Hulin Wu*, University of Rochester

5:25 **Floor Discussion**

85. ROC METHODS: EXPERIMENTS WITH TIME-DEPENDENT AND CLUSTERED DATA
Magnolia Room (3rd Floor)
Sponsor ASA Biometrics Section
Organizer: Abdus Wahed, University of Pittsburgh
Chair: Andriy Bandos, University of Pittsburgh

3:45 **Stratified and Clustered Receiver Operating Characteristic Analysis**
Kelly H. Zou*, Pfizer Inc. and Simon K. Warfield, Children's Hospital Boston and Harvard Medical School

4:10 **Evaluation of Biomarker Accuracy under Nested Case-control Studies**
Tianxi Cai*, Harvard University and Yingye Zheng, Fred Hutchinson Cancer Research Institute

4:35 **Summarizing Performance in FROC Experiments**
Howard E. Rockette* and Andriy Bandos, University of Pittsburgh

5:00 **Sample Size Considerations for Time-Dependent ROC Estimation**
Hong Li, Harvard University and Constantine Gatsonis*, Brown University

5:25 **Floor Discussion**

86. CONTRIBUTED PAPERS: ANALYSIS OF CLINICAL TRIALS AND BIOPHARMACEUTICAL STUDIES
Chequers Room (2nd Floor)
Sponsor ASA Biopharmaceutical Section
Organizer: Andrea J. Cook, University of Washington
Chair: Andrea J. Cook, University of Washington

3:45 **Comparison of Estimates for the Common Correlation Coefficient in a Stratified Experiment**
Yougui Wu, University of South Florida and Jason Liao*, Merck Research Laboratories

4:00 **A Consistency-adjusted Strategy for Testing Alternative Endpoints in a Clinical Trial**
Mohamed Alosh,* U.S. Food and Drug Administration

4:15 **Adaptive Confidence Intervals for Non-regular Parameters in Dynamic Treatment Regimes**
Min Qian*, Eric B. Laber and Susan A. Murphy, University of Michigan

4:30 **Prediction of the Biggest Loser: A Bridge Connecting Obesity Clinical Outcomes and Animal Models**
Yuefeng Lu*, Eli Lilly and Company

4:45 **Semiparametric Causal Inference for Randomized Clinical Trials with a Time-to-Event Outcome and All-or-None Treatment-Noncompliance**
Ying Zhou* and Gang Li, University of California-Los Angeles and Huazhen Lin, Sichuan University, China

5:00 **Pooling Strategies for Outcome Under a Gaussian Random Effects Model**
Yaakov Malinovsky*, Paul S. Albert and Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development and Vyacheslav Vasiliev, Tomsk State University

5:15 **M-estimation Procedures in Heteroscedastic Nonlinear Regression Models with Parametric Variance Model**
Changwon Lim*, National Institutes of Health, Pranab K. Sen, University of North Carolina-Chapel Hill and Shyamal D. Peddada, National Institutes of Health

87. CONTRIBUTED PAPERS: CLUSTERED DATA METHODS
Jasperwood Room (3rd Floor)
Sponsor: ENAR
Chair: Inyoung Kim, Virginia Tech University

3:45 **Analysis of Unbalanced and Unequally-spaced Familial Data using Generalized Markov Dependence**
Roy T. Sabo*, Virginia Commonwealth University

4:00 **Identifying Distinct Subtypes in Acute Myeloid Leukemia: A Model Based Clustering Approach**
Matthias Kormaksson*, Cornell University

4:15 **Weighted Scores Method to Analyze Multivariate Overdispersed Count Data**
Aristidis K. Nikoloulopoulos, Athens University of Economics and Business, Harry Joe, University of British Columbia and N. Rao Chaganty*, Old Dominion University

4:30 **Re-sampling based Method to Estimate Intra-cluster Correlation for Clustered Binary Data**
Hrishikesh Chakraborty*, RTI International and Pranab K. Sen, University of North Carolina-Chapel Hill

4:45 **A Bi-directional Random Effects Specification for Heterogeneity in Mixed Effects Models**
Zhen Chen*, National Institutes of Health

5:00 **Regression Analysis of Clustered Interval-Censored Failure Time Data with Informative Cluster Size**
Xinyan Zhang* and Jianguo Sun, University of Missouri

5:15 **Likelihood Methods for Binary Responses of Present Components in a Cluster**
Xiaoyun Li*, Florida State University, Dipankar Bandyopadhyay, Medical University of South Carolina, Stuart Lipsitz, Harvard Medical School and Debajyoti Sinha, Florida State University

88. CONTRIBUTED PAPERS: MULTIVARIATE SURVIVAL
Cambridge Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Jane Paik, Stanford University

3:45 **Incorporating Sampling Bias in Analyzing Bivariate Survival Data with Interval Sampling and Application to HIV Research**
Hong Zhu* and Mei-Cheng Wang, Johns Hopkins University

4:00 **Estimation of the Gap Time Distributions for Ordered Multivariate Failure Time Data: A Sieve Likelihood Approach**
Chia-Ning Wang*, Bin Nan, Roderick Little and Harlow Sioban, University of Michigan

4:15 **Semiparametric Inference for Successive Durations**
Jing Qian*, Harvard University and Yijian Huang, Emory University

4:30 **A Joint Modeling of Correlated Recurrent and Terminal Events with Multivariate Frailty in the Analysis of Driving Safety Data in Old Drivers with Neurological Diseases**
Dawei Liu*, Ergun Uc, Elizabeth Dastrup, Aaron Porter, Jeff Dawson and Matt Rizzo, University of Iowa

4:45 **Exponential Gap-Time Estimation for Correlated Recurrent Event data under Informative Monitoring**
Akim Adekpedjou*, Missouri University of Science and Technology and Gideon Zamba, University of Iowa

5:00 **On Several Test Statistics for Paired Censored Data**
Liang Li*, Cleveland Clinic

5:15 **Floor Discussion**

89. CONTRIBUTED PAPERS: GENOME-WIDE ASSOCIATION STUDIES
Oak Alley Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: John Barnard, Cleveland Clinic

3:45 **Bayesian Variable Selection in a Genetic Association Study**
Mengye Guo*, Dana-Farber Cancer Institute, Edward George, Nandita Mitra and Daniel Heitjan, University of Pennsylvania

4:00 **Survival Kernel Machine SNP-set Analysis for Genome-wide Association Studies**
Xinyi Lin*, Tianxi Cai and Xihong Lin, Harvard University

4:15 **A Genome-wide Association Approach on Detecting CNVs for SNP Genotyping Data**
Yaji Xu*, Bo Peng and Christopher I. Amos, University of Texas M.D. Anderson Cancer Center

4:30 **Kernel Machine Methods for the Analysis of Large Scale Genetic Association Studies**
Michael C. Wu*, University of North Carolina-Chapel Hill

4:45 **A Genome Imprinting Test with Application to Whole-Genome Scans of Insulin Resistance and Glucose**
Rui Feng*, University of Pennsylvania and Donna Arnett, University of Alabama-Birmingham

5:00 **A Cross-Population Comparison Score Test to Detect Positive Selection in Genome-wide Scans**
Ming Zhong*, Texas A&M University, Kenneth Lange, University of California-Los Angeles and Ruzong Fan, Texas A&M University

5:15 **Optimal Choice of Latent Ancestry Variables for Adjustment in Genome-wide Association Studies**
Gina M. Peloso*, L Adrienne Cupples and Kathryn L. Lunetta, Boston University

90. CONTRIBUTED PAPERS: NEW METHODOLOGY FOR LINEAR MIXED MODEL FRAMEWORK
Eglinton Winton Room (2nd Floor)
Sponsor: ENAR
Chair: Ni Li, University of Missouri

3:45 **Goodness of Fit Tests in Linear Mixed Models**
Min Tang* and Eric V. Slud, University of Maryland-College Park

4:00 **Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models**
Sijian Wang*, University of Wisconsin-Madison, Peter XK Song and Ji Zhu, University of Michigan

4:15 **Likelihood Reformulation Method in Non-normal Random Effects Models**
Lei Liu*, University of Virginia and Zhangsheng Yu, Indiana University School of Medicine

4:30 **A Goodness-of-fit Test for the Random-effects Distribution in Mixed Models**
Roula Tsonaka*, Leiden University Medical Center, The Netherlands, Dimitris Rizopoulos, Erasmus University Medical Center, The Netherlands, Geert Verbeke and Geert Molenberghs, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium

4:45 **Normalization and Analysis of Longitudinal Quantitative PCR Data by Linear Mixed Models**
Xuelin Huang*, University of Texas, M.D. Anderson Cancer Center

5:00 **Asymptotic Equivalence Between Cross-validations and Akaike Information Criteria in Mixed-effects Models**
Yixin Fang*, Georgia State University

5:15 **Variable Selection in Linear Mixed Models for Longitudinal Data**
Rain Cui*, Xihong Lin and Victor DeGruttola, Harvard University

91. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS
Elmwood Room (3rd Floor)
Sponsor: ENAR
Chair: Pei Li, University of Minnesota

3:45 **Estimation of Population Abundance Using a Hierarchical Depletion Model**
Thomas F. Bohrmann*, Mary C. Christman and Xiaobo Li, University of Florida

4:00 **Assessing the Efficacy of Slow Speed Zones in Florida's Waterways**
Kenneth K. Lopiano* and Linda J. Young, University of Florida

4:15 **A Single State Super Population Capture-Recapture Model Augmented with Information on Population of Origin**
Zhi Wen* and Kenneth Pollock, North Carolina State University, James Nichols, U.S.Geological Survey Patuxent Research Center and Peter Waser, Purdue University

4:30 **Modification by Frailty Status of the Association Between Air Pollution and Lung Function in Older Adults**
Sandrah P. Eckel*, University of Southern California, Thomas A. Louis, Karen Bandeen-Roche and Paulo H. Chaves, Johns Hopkins University, Linda P. Fried, Columbia University and Helene Margolis, University of California-Davis

4:45 **A New Stochastic Model of Carcinogenesis for Initiation-Promotion Bioassay**
Wai-Yuan Tan and Xiaowei (Sherry) Yan*, University of Memphis

5:00 **Estimating the Acute Health Effects of Coarse Particulate Matter Accounting for Exposure Measurement Error**
Howard Chang*, Statistical and Applied Mathematical Sciences Institute, Roger D. Peng, Johns Hopkins University and Francesca Dominici, Harvard University

5:15 **Rank Tests for Selective Predation**
Yankun Gong*, Shuxin Yin and Asheber Abebe, Auburn University

92. CONTRIBUTED PAPERS: VARIABLE SELECTION AND PENALIZED REGRESSION MODELS
Belle Chasse Room (3rd Floor)
Sponsor: ENAR
Chair: Yun Lu, University of Pennsylvania

3:45 **A Perturbation Method for Inference on Adaptive LASSO Regression Estimates**
Jessica Minnier* and Tianxi Cai, Harvard University

4:00 **Bootstrap Inconsistency and an Oracle Bootstrap**
Mihai C. Giurcanu*, University of Louisiana-Lafayette and Brett D. Presnell, University of Florida

4:15 **A Lasso-Type Approach for Estimation and Variable Selection in Single Index Models**
Peng Zeng*, Auburn University, Tianhong He and Yu Zhu, Purdue University

4:30 **Coordinate Descent Algorithms for Nonconvex Penalized Regression Methods**
Patrick Breheny*, University of Kentucky and Jian Huang, University of Iowa

4:45 **Variable Selection for Panel Count Data via Non-concave Penalized Estimating Function**
Xingwei Tong, Beijing Normal University, Xin He*, University of Maryland-College Park, Liuquan Sun, Chinese Academy of Sciences and Jianguo Sun, University of Missouri-Columbia

5:00 **Variable Selection with the Seamless-L0 Penalty**
Lee Dicker*, Baosheng Huang and Xihong Lin, Harvard University

5:15 **EEBoost: A General Framework for High-dimensional Variable Selection Based on Estimating Equations**
Julian Wolfson*, University of Minnesota

WEDNESDAY, MARCH 24

8:30—10:15 a.m.

93. STATISTICAL ANALYSIS OF BRAIN IMAGING DATA
Rosedown Room (3rd Floor)
Sponsor: ENAR
Organizer: Hongtu Zhu, University of North Carolina-Chapel Hill
Chair: Lynn Eberly, University of Minnesota

8:30 **Determining Differences in Resting-State Brain Connectivity between Patients with Depression and Healthy Controls: A Combined fMRI/DTI Analysis**
DuBois Bowman*, Gordana Derado and Shuo Chen, Emory University

8:55 **Statistical Methods for Evaluating Connectivity in the Human Brain**
Brian S. Caffo* and Cirpian M. Crainiceanu, Johns Hopkins University

9:20 **FRATS: Functional Regression Analysis of DTI Tract Statistics**
Hongtu Zhu*, Martin G. Styner, Weili Lin, and Zhexing Liu of University of North Carolina-Chapel Hill, Niansheng Tang, Yunnan University and John H. Gilmore, University of North Carolina-Chapel Hill

9:45 **Over-connectivity of 3D Brain Network using Diffusion Tensor Imaging**
Moo K. Chung*, University of Wisconsin-Madison

10:10 **Floor Discussion**

94. REGRESSION MODELS WITH COMPLEX COVARIATE INPUTS
Magnolia Room (3rd Floor)
Sponsor: ENAR
Organizer: Naisyin Wang,Texas A&M University
Chair: Naisyin Wang, Texas A&M University

8:30 **Regression Analysis of Graph-Structured Data With Genomic Applications**
Hongzhe Li*, University of Pennsylvania

8:55 **Covariate Adjusted Association Tests for Ordinal Traits**
Wensheng Zhu, Yuan Jiang and Heping Zhang*, Yale University

9:20 **Adaptive Functional Linear Mixed Models**
Veera Baladandayuthapani* and Jeffrey S. Morris, University of Texas M.D. Anderson Cancer Center

9:45 **Functional Single Index Models for Functional Covariate**
Ciren Jiang, University of California-Berkeley and Jane-Ling Wang*, University of California-Davis

10:10 **Floor Discussion**

95. METHODS FOR COMBINING MATCHED AND UNMATCHED CASE-CONTROL STUDIES

Jasperwood Room (3rd Floor)
Sponsor: ASA Section on Statistics in Epidemiology
Organizer: Mulugeta Gebregziabher, Medical University of South Carolina
Chair: Paulo Gumareas, University of South Carolina

8:30 **Combining Matched and Unmatched Case-Control Studies using Standard Conditional Logistic Regression Software**
Bryan Langholz*, University of Southern California

8:55 **Combining Matched and Unmatched Control Groups in Case-Control Studies, using Sandwich or Bootstrapping Methods**
Saskia le Cessie*, Leiden University Medical Center

9:20 **On Combining Related and Unrelated Controls**
Nilanjan Chatterjee*, National Cancer Institute and Bhramar Mukherjee, University of Michigan

9:45 **A Polytomous Conditional Likelihood Approach for Combining Matched and Unmatched Case-Control Studies**
Mulugeta Gebregziabher*, Medical University of South Carolina, Paulo Guimaraes, Wendy Cozen and David Conti, University of South Carolina

10:10 **Floor Discussion**

96. HIGH DIMENSIONAL DATA ANALYSIS: DIMENSION REDUCTION AND VARIABLE SELECTION

Melrose Room (3rd Floor)
Sponsor: IMS
Organizer: Junhui Wang, University of Illinois at Chicago
Chair: Junhui Wang, University of Illinois at Chicago

8:30 **Groupwise Dimension Reduction**
Lexin Li*, North Carolina State University, Bing Li, Penn State University and Li-Xing Zhu, Hong Kong Baptist University

8:55 **Non-Euclidean Dimension Reduction Via Graph Embedding**
Michael W. Trosset* and Minh Tang. Indiana University

9:20 **Non-Linear Penalized Variable Selection**
Gareth James* and Peter Radchenko, University of Southern California

9:45 **Boosting for High-Dimensional Linear Models with Grouped Variables**
Lifeng Wang*, Michigan State University, Yihui Luan, Shandong University and Hongzhe Li, University of Pennsylvania

10:10 **Floor Discussion**

97. CONTRIBUTED PAPERS: NONPARAMETRIC METHODS

Ascot Room (3rd Floor)
Sponsor: ENAR
Chair: Hani Samawi, Georgia Southern University

8:30 **Semi-parametric Measurement Error Modeling in Logistic Regression**
Jianjun Gan* and Hongmei Zhang, University of South Carolina

8:45 **Estimation of the Probability that Treatment is Better than Control in Clinical Trials with Continuous Outcomes**
Suporn Sukpraprut* and Michael P. LaValley, Boston University

9:00 **Semiparametric Spline Regression for Longitudinal/ Clustered Data**
Arnab Maity* and Xihong Lin, Harvard University and Raymond J. Carroll, Texas A&M University

9:15 **Dimension Reduction for the Conditional Kth Moment via Central Solution Space**
Yuexiao Dong*, Temple University and Zhou Yu, East China Normal University

9:30 **Testing for Constant Nonparametric Effects in General Semiparametric Regression Models with Interactions**
Jiawei Wei*, Texas A&M University and Arnab Maity, Harvard University

9:45 **Nonparametric Additivity Test under Random Design**
Zhi He* and Douglas G. Simpson, University of Illinois at Urbana-Champaign

10:00 **Balancing the Optimal Match: A Clever Swap**
Shoshana R. Daniel. Covance, Inc.

98. CONTRIBUTED PAPERS: NONPARAMETRIC AND SEMIPARAMETRIC SURVIVAL MODELS

Chequers Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Dawei Liu, University of Iowa

● **8:30** **Sieve Maximum Likelihood Estimation Using B-Splines for the Accelerated Failure Time Model**
Ying Ding* and Bin Nan, University of Michigan

8:45 **Efficient Computation of Nonparametric Survival Functions via a Hierarchical Mixture Formulation**
Yong Wang*, University of Auckland, New Zealand and Stephen M. Taylor, Auckland University of Technology, New Zealand

9:00 **Scientific Implications of Parametric, Semiparametric, and Nonparametric Statistical Models**
Scott S. Emerson*, University of Washington

9:15 **Targeted Maximum Likelihood For Time To Event Data**
Ori M. Stitelman* and Mark J. van der Laan, University of California-Berkeley

9:30 **Partially Monotone Tensor Spline Estimation of Joint Distribution Function with Bivariate Current Status Data**
Yuan Wu* and Ying Zhang, University of Iowa

9:45 **A Non-parametric Maximum Likelihood Estimation Approach to Frailty Model**
Zhenzhen Xu* and John D. Kalbfleisch, University of Michigan

10:00 **A Novel Semiparametric Method for Modeling Interval-Censored Data**
Seungbong Han*, Adin-Cristian Andrei and Kam-Wah Tsui, University of Wisconsin-Madison

99. CONTRIBUTED PAPERS: RATER AGREEMENT AND SCREENING TESTS
Cambridge Room (2nd Floor)
Sponsor: ASA Biometrics Section
Chair: Stephen L. Hillis, University of Iowa

8:30 **Assessing the "Broad Sense Agreement" between Ordinal and Continuous Measurements**
Limin Peng*, Ruosha Li, Ying Guo and Amita Manatunga, Emory University

8:45 **Estimating the Agreement and Diagnostic Accuracy of Two Diagnostic Tests When One Test is Conducted on only a Subsample of Specimens**
Hormuzd A. Katki*, National Cancer Institute, Yan Li, University of Texas at Arlington and Philip E. Castle, National Cancer Institute

9:00 **Testing the Significance of Overlapping Sets in a Venn Diagram**
Aixiang Jiang*, Vanderbilt University

9:15 **Estimation of Cut-Points on a Continuous Scale According to a Categorical Scale**
Ming Wang*, Amita Manatunga, Ying Guo and Limin Peng, Emory University

9:30 **Comparing the Cumulative False-Positive Risk of Screening Mammography Programs Using a Discrete Time Survival Model Allowing for Non-Ignorable Drop-Out**
Rebecca A. Hubbard* and Diana L. Miglioretti, Group Health Research Institute

9:45 **Logic Forest: An Ensemble Classifier for Discovering Logical Combinations of Binary Markers**
Bethany J. Wolf*, Elizabeth H. Slate and Elizabeth G. Hill, Medical University of South Carolina

10:00 **Floor Discussion**

100. CONTRIBUTED PAPERS: BAYESIAN METHODS: JOINT LONGITUDINAL/SURVIVAL MODELING AND DISEASE MODELING
Eglinton Winton Room (2nd Floor)
Sponsors: ASA Section on Bayesian Statistical Sciences, ASA Biometrics Section
Chair: Steven MacEachern, The Ohio State University

8:30 **Bayesian Joint Models of Zero-Inflated Longitudinal Patient-Reported Outcomes and Progression-Free Survival Times in Mesothelioma**
Laura A. Hatfield*, University of Minnesota, Mark E. Boye and Michelle D. Hackshaw, Eli Lilly and Company and Bradley P. Carlin, University of Minnesota

8:45 **Bayesian Semiparametric Multivariate Joint Models**
Dimitris Rizopoulos*, Erasmus University Medical Center, The Netherlands and Pulak Ghosh, Novartis Pharmaceuticals

9:00 **Semiparametric Bayesian Joint Model with Variable Selection**
Haikun Bao* and Bo Cai, University of South Carolina, Pulak Ghosh, Novartis Pharmaceuticals and Nicole Lazar, University of Georgia

9:15 **A Dynamic Projection Model of the Burden of Diabetes in the U.S. Adult Population**
James P. Boyle*, Theodore J. Thompson and Lawrence Barker, Centers for Disease Control

9:30 **An Empirical, Informed Prior for the Between-Study Heterogeneity in Meta-Analyses**
Eleanor M. Pullenayegum*, McMaster University

9:45 **Bayesian Hierarchical Modeling of Host Genetic Correlates of Immune Response to Anthrax Vaccine**
Nicholas M. Pajewski*, University of Alabama Birmingham, Purushottam W. Laud, Medical College of Wisconsin, Scott D. Parker, Robert P. Kimberly and Richard A. Kaslow, University of Alabama-Birmingham

10:00 **Relative Breadth of Mosaic and CON-S HIV-1 Vaccine Design Strategies**
Sydeaka P. Watson*, Baylor University and Los Alamos National Laboratory, Bette T. Korber, Los Alamos National Laboratory, Mark R. Muldoon, University of Manchester, John W. Seaman and James Stamey, Baylor University

101. CONTRIBUTED PAPERS: INTEGRATION OF INFORMATION ACROSS MULTIPLE STUDIES OR MULTIPLE -OMICS PLATFORMS
Belle Chasse Room (3rd Floor)
Sponsors: ASA Biometrics Section/ASA Biopharmaceutical Section
Chair: W. Evan Johnson, Brigham Young University

8:30 **Integrating Diverse Genomic Data Using Gene Sets**
Svitlana Tyekucheva*, Dana-Farber Cancer Institute, Rachel Karchin, Johns Hopkins University and Giovanni Parmigiani, Dana-Farber Cancer Institute

8:45 **Bayesian Joint Modeling of Multiple Gene Networks and Diverse Genomic Data to Identify Target Genes of a Transcription Factor**
Peng Wei*, University of Texas and Wei Pan, University of Minnesota

9:00 **A Latent Mixture Model for Analyzing Multiple Gene Expression and ChIP-chip Data Sets**
Hongkai Ji*, Johns Hopkins University

9:15 **Effect of Combining Statistical Tests and Fold-Change Criteria**
Doug Landsittel*, University of Pittsburgh and Nathan Donohue-Babiak, Duquesne University

9:30 **A Comparison of Methods for Integrated Omics Analysis**
John Barnard*, Cleveland Clinic

9:45 **Pathway-directed Weighted Testing Procedures for the Integrative Analysis of Gene Expression and Metabolomic Data**
Laila M. Poisson*, University of Michigan and Debashis Ghosh, Penn State University

10:00 **An Empirical Bayes Approach to Joint Analysis of Multiple Microarray Gene Expression Studies**
Lingyan Ruan* and Ming Yuan, Georgia Institute of Technology

102. CONTRIBUTED PAPERS: ESTIMATING EQUATIONS
Newberry Room (3rd Floor)
Sponsor: ENAR
Chair: Xin He, University of Maryland-College Park

8:30 **Estimating Equations in Biased Sampling Problems**
Bin Zhang*, University of Alabama-Birmingham and Jing Qin, National Institute of Allergy and Infectious Diseases

8:45 **Bias Sampling, Nuisance Parameters, and Estimating Equations**
Kunthel By*, University of North Carolina-Chapel Hill

9:00 **Estimation in Logistic Regression Models for Clustered/Longitudinal Data with Covariate Measurement Error**
Jeff Buzas*, University of Vermont

9:15 **Semiparametric Transformation Models for Panel Count Data with Dependent Observation Process**
Ni Li*, University of Missouri, Liuquan Sun, Chinese Academy of Sciences and Jianguo Sun, University of Missouri

9:30 **A Comparison of Several Approaches for Analysis of Longitudinal Binary Data**
Matthew Guerra*, Justine Shults and Thomas Ten Have, University of Pennsylvania

9:45 **Augmented Estimating Equations for Semiparametric Panel Count Regression with Informative Observation Times and Censoring Time**
Xiaojing Wang* and Jun Yan, University of Connecticut

10:00 **Analyzing Length-biased Data with Accelerated Failure Time Models**
Jing Ning*, University of Texas Health Science Center at Houston, Jing Qin, National Institute of Allergy and Infectious Diseases and Yu Shen, University of Texas M. D. Anderson Cancer Center

WEDNESDAY, MARCH 24

10:15—10:30 a.m. Refreshment Break and Visit the Exhibitors
Court Assembly Foyer (3rd Floor)

Wednesday, March 24
10:30 a.m.—12:15 p.m.

103. RECENT ADVANCES IN MODELING NONLINEAR MEASUREMENT ERROR
Rosedown Room (3rd Floor)
Sponsor: ASA Section on Statistics in Epidemiology
Organizer: Victor Kipnis, National Cancer Institute
Chair: Douglas Midthune, NCI

10:30 **Correction for Measurement Error in Covariates for Interaction Models**
Havi Murad, Gertner Institute for Epidemiology, Israel, Victor Kipnis, National Cancer Institute and Laurence S. Freedman*, Gertner Institute for Epidemiology, Israel

11:00 **Correction for Measurement Error in Nutritional Epidemiology: Allowing for Never and Episodic-Consumers in Measurement Error Models for Dietary Assessment Instruments**
Ruth Keogh*, MRC Centre for Nutritional Epidemiology in Cancer Prevention and Survival and MRC Biostatistics Unit, University of Cambridge
Ian White, MRC Biostatistics Unit, Cambridge, UK

11:30 **Simultaneous Modeling of Multivariate Data with Excess Zeros and Measurement Error with Application to Dietary Surveys**
Victor Kipnis*, National Cancer Institute, Raymond J. Carroll, Texas A&M University, Laurence S. Freedman, Gertner Institute for Epidemiology and Public Health Policy, Israel and Douglas Midthune, National Cancer Institute

12:00 **Floor Discussion**

104. USE OF BIOMARKERS IN PERSONALIZED MEDICINE
Melrose Room (3rd Floor)
Sponsor: ASA Health Policy Statistics Section
Organizer: Liansheng Larry Tang, George Mason University
Chair: Pang Du, Virginia Tech University

10:30 **A Procedure for Evaluating Predictive Accuracy of Biomarkers for Selecting Optimal Treatments**
Xiao-Hua A. Zhou* and Yunbei Ma, University of Washington

10:55 **Clinical Trial Designs for Predictive Biomarker Validation: Theoretical Considerations and Practical Challenges**
Sumithra J. Mandrekar* and Daniel J. Sargent, Mayo Clinic

11:20 **Adaptive Designs to Validate Cancer Biomarkers**
Liansheng Tang*, George Mason University

11:45 A Threshold Sample-Enrichment Approach in a Clinical Trial with Heterogeneous Subpopulations
Aiyi Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Qizhai Li, Chinese Academy of Sciences, Chunling Liu and Kai F. Yu, Eunice Kennedy Shriver National Institute of Child Health and Human Development and Vivian Yuan, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

12:10 Floor Discussion

105. ANALYSIS OF RECURRENT EVENTS DATA IN THE PRESENCE OF A TERMINAL EVENT
Magnolia Room (3rd Floor)
Sponsor: ASA Biometrics Section
Organizer: Lei Liu, University of Virginia
Chair: Terry Therneau, Mayo Clinic

10:30 Semiparametric Additive Rate Model for Recurrent Event with Informative Terminal Event
Jianwen Cai* and Donglin Zeng, University of North Carolina-Chapel Hill

10:55 Models for Joint Longitudinal and Event-Time Outcomes
Elizabeth H. Slate*, Medical University of South Carolina

11:20 Robust Estimation of Mean Functions and Treatment Effects for Recurrent Events Under Event-Dependent Censoring and Termination
Richard J. Cook* and Jerald F. Lawless, University of Waterloo, Lajmi Lakhal-Chaieb, Université Laval, Québec and Ker-Ai Lee, University of Waterloo

11:45 Analyzing Recurrent Events Data: A Bayesian Perspective
Debajyoti Sinha*, Florida State University, Bichun Ouyang, Rush Medical Center, Elizabeth Slate, Medical University of South Carolina and Yu Gu, Florida State University

12:10 Floor Discussion

106. CONTRIBUTED PAPERS: GENETIC STUDIES WITH RELATED INDIVIDUALS
Newberry Room (3rd Floor)
Sponsor: ENAR
Chair: Dong Wang, University of Nebraska

10:30 Heritability Estimation using Regression Models for Correlation: Quantitative Traits from Extended Families
Hye-Seung Lee*, University of South Florida, Myunghee C. Paik, Columbia University and Jefferey P. Krischer, University of South Florida

10:45 Association Analysis of Ordinal Traits on Related Individuals
Zuoheng Wang*, Yale University

11:00 Penalized Estimation of Haplotype Frequencies from General Pedigrees
Kui Zhang*, University of Alabama-Birmingham

11:15 An EM Composite Likelihood Approach for Multistage Sampling of Family Data
Yun-Hee Choi*, University of Western Ontario and Laurent Briollais, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto

11:30 A Likelihood Approach for Detection of Imprinting and Maternal Effects Using General Pedigree Data
Jingyuan Yang* and Shili Lin, The Ohio State University

11:45 Functional Mapping in Human Population with Genetic Data Structure of Parents and Children
Jiangtao Luo*, Penn State College of Medicine, William W. Hager, University of Florida and Rongling Wu, Penn State College of Medicine

12:00 Floor Discussion

107. CONTRIBUTED PAPERS: HYPOTHESIS TESTING, MULTIPLE TESTING, AND COMPUTATIONAL METHODS
Cambridge Room (2nd Floor)
Sponsor: ENAR
Chair: Eric Laber, University of Michigan

10:30 False Discovery Rate Control For High Dimensional Multivariate Data
Jichun Xie*, Tianwen Tony Cai and Hongzhe Li, University of Pennsylvania

10:45 Adaptive Multiple Testing Procedures under Dependence
Wenge Guo*, New Jersey Institute of Technology

11:00 A Univariate Approach to Repeated Measures and MANOVA for High Dimension, Low Sample Size
Yueh-Yun Chi* and Keith E. Muller, University of Florida

11:15 Application of Anbar's Approach to Hypothesis Testing to Detect the Difference between Two Proportions
Julia Soulakova* and Ananya Roy, University of Nebraska

11:30 Nonparametric Test of Symmetry Based on Overlapping Coefficient
Hani M. Samawi*, Georgia Southern University, Amal Helu, University of Jordan and Robert Vegal, Georgia Southern University

11:45 RapidStat: A Hybrid of Excel and Graphical Language to Expedite User Interface Creation
Pallabi Saboo and Marc Abrams*, Harmonia Inc.

12:00 A Summary of Graphic Approaches to Monitor Performance of Liver Transplant Centers
Jie (Rena) Sun* and John D. Kalbfleisch, University of Michigan

108. CONTRIBUTED PAPERS: GENOMICS AND PROTEOMICS
Jasperwood Room (3rd Floor)
Sponsor: ASA Biometrics Section
Chair: Hongkai Ji, Johns Hopkins University

10:30 Multi-gene Domain Clusters Found Throughout the Mouse Genome via Hidden Markov Models
Jessica L. Larson* and Guocheng Yuan, Harvard University

10:45 Dissection of Allele Specific Copy Number Changes and its Applications
Wei Sun*, University of North Carolina-Chapel Hill

11:00 **Bayesian Modeling of ChIP-Chip Data through a High-order Ising Model**
Qianxing Mo*, Memorial Sloan-Kettering Cancer Center and Faming Liang, Texas A&M University

11:15 **Signal Extraction and Breakpoint Identification for Array CGH Data using State Space Model**
Bin Zhu*, Peter Song and Jeremy Taylor, University of Michigan

11:30 **A Bayesian Model for Analysis of Copy Number Variants in Genetic Studies**
Juan R. Gonzalez*, Juan J. Abellan and Carlos Abellan, CIBER Epidemiology and Public Health (CIBERESP), Spain Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, Spain

11:45 **Sequential Sampling Designs for Small-Scale Protein Interaction Experiments**
Denise Scholtens* and Bruce Spencer, Northwestern University

12:00 **A Multi-step Protein Lysate Array Quantification Method and its Statistical Properties**
Ji-Yeon Yang* and Xuming He, University of Illinois at Urbana-Champaign

109. CONTRIBUTED PAPERS: INFECTIOUS DISEASE AND MEDICAL CASE STUDIES

Eglinton Winton Room (2nd Floor)
Sponsors: ASA Biopharmaceutical Section, ASA Section on Statistics in Epidemiology
Chair: Nicholas Reich, Johns Hopkins University

10:30 **Modeling Infectivity Rates and Attack Windows for Two Viruses**
Jing Zhang*, Douglas Noe, Jian Wu, A. John Bailer and Stephen Wright, Miami University

10:45 **Bayesian Inference for Contact Networks Given Epidemic Data**
Chris Groendyke*, David R. Hunter and David Welch, Penn State University

11:00 **Optimizing Exchanges in a Kidney Paired Donation (KPD) Program**
Yijiang (John) Li*, Yan Zhou, John D. Kalbfleisch and Peter X.-K. Song, University of Michigan

11:15 **Estimation in Type I Censored Viral Load Assays Under Non-Normality**
Evrim Oral, Robbie A. Beyl* and William T. Robinson, Louisiana State University

11:30 **Improved Meta Analysis of Randomized Controlled Trials on the Comparative Efficacy of Daily Low-intake of Dark Chocolate Among Middle-aged Hypertensive Patients as Compared to Placebo**
Martin Dunbar*, Georgia Southern University

11:45 **Informative Dorfman Screening with Risk Thresholds**
Christopher S. McMahan* and Joshua M. Tebbs, University of South Carolina and Christopher R. Bilder, University of Nebraska

12:00 **Floor Discussion**

110. CONTRIBUTED PAPERS: VARIABLE SELECTION METHODS

Ascot Room (3rd Floor)
Sponsor: ENAR
Chair: Patrick Breheny, University of Kentucky

10:30 **A Path Following Algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE)**
Guilherme V. Rocha*, Indiana University, Peng Zhao, Citadel Investment Group and Bin Yu, University of California-Berkeley

10:45 **Multicategory Vertex Discriminant Analysis for High-Dimensional Data**
Togntong Wu*, University of Maryland and Kenneth Lange, University of California-Los Angeles

11:00 **Order Thresholding**
Min HeeKim* and Michael G. Akritas, Penn State University

11:15 **Variable Selection in Partial Linear Additive Model**
Fengrong Wei*, University of West Georgia

11:30 **Condence Region Based Tuning for Forward and Backward Selection**
Funda Gunes* and Howard Bondell, North Carolina State University

● **11:45** **Robust Penalized Logistic Regression with Truncated Loss Functions**
Seo Young Park* and Yufeng Liu, University of North Carolina-Chapel Hill

12:00 **Adaptive Model Selection in Linear Mixed Models**
Bo Zhang*, National Institute of Child Health and Human Development, Xiaotong Shen, University of Minnesota and Zhen Chen, National Institute of Child Health and Human Development

111. CONTRIBUTED PAPERS: GENERALIZED LINEAR MODELS

Chequers Room (2nd Floor)
Sponsor: ENAR
Chair: Yehua Li, University of Georgia

10:30 **Conditional Logistic Mixed Effects Model for Unbalanced Matched Case-Control Studies**
Inyoung Kim* and Feng Guo, Virginia Tech University

10:45 **Accuracy and Precision of Estimates in Covariate-Adjusted Generalized Linear Regression Models with or without Treatment and Covariate Interaction**
Junyi Lin*, Penn State University, Lei Nie, U.S. Food and Drug Administration and Runze Li, Penn State University

11:00 **MCEM-SR and EM-LA2 for fitting Generalized Linear Mixed Models**
Vadim V. Zipunnikov*, Johns Hopkins University and James G. Booth, Cornell University

11:15 **Rationale for Choosing Explicit Correlation Structure in a Multivariate**
Folefac D. Atem* and Stewart J. Anderson, University of Pittsburgh

11:30 **Estimation of the Standard Deviation for an Exponential Distribution from Limited Data**
Yvonne M. Zubovic* and Chand K. Chauhan, Indiana University Purdue University-Fort Wayne

11:45 **Performance of Beta Regression in Detecting Independent Group Differences**
Christopher J. Swearingen*, University of Arkansas for Medical Sciences, Dipankar Bandyopadhyay and Robert F. Woolson, Medical University of South Carolina and Barbara C. Tilley, University of Texas Health Science Center

12:00 **Floor Discussion**

Abstracts

1. POSTERS: CLINICAL TRIALS

1a. A PREDICTIVE MODEL FOR IMBALANCE IN STRATIFIED PERMUTED BLOCK DESIGNS

Joseph W. Adair, PPD
Aaron C. Camp*, PPD

Quantifying the probability of imbalance in multiple stratified permuted block designs at a given level of enrollment may be useful in the decision-making process when deciding among randomization schemes for clinical trials (e.g., Pocock and Simon (1975) adaptive design vs. stratified permuted block design). A closed form is found for the probability density function (PDF) of the imbalance in such designs. For a small number of strata the closed form may be computed easily, but for more unwieldy designs a simulation is presented which allows quantification of the probability of imbalance for a given randomization. A permuted block design with a 2:1 randomization between two treatment arms, block size six, and 18 subjects classified into four strata level combinations is considered. The PDF of imbalance calculated using the closed form solution and the empirical density function resulting from the simulation are found to be congruent. The simulation is then generalized and implemented in order to obtain empirical density functions for several sample permuted block designs and the results provide a standard measure to compare permuted block designs and adaptive randomization designs using Pocock and Simon's method.

email: aaron.camp@ppdi.com

1b. OPTICAL PROPERTIES OF HUMAN SKIN AS A CRITERION FOR NONMELANOMA SKIN CANCER DETECTION

Vadim S. Tyuryaev*, Texas Tech University
Clyde F. Martin, Texas Tech University

Estimated new cases from nonmelanoma skin cancer in the United States in 2009, is 1,000,000. Early detection of skin cancer is a key to its successful treatment. One of the ways to solve it is implementation of IT technologies for supplementary diagnostics, based on difference in optical properties of cancerous and healthy skin. Non-invasive means of diagnostics have a potential to detect slight pathological changes in the human skin on early stages of cancer, which will allow its on-time detection. Current paper investigates the possibility of usage of absorption and scattering coefficients in order to differentiate between normal and healthy skin. We would like to thank Dr. Yaroslavsky, Harvard Medical School, Massachusetts General Hospital for the data. The therapeutic window, 600-1200 nm, where the difference between scattering and absorption is much more pronounced, is used. Results, based on ANOVA and multiple mean comparisons, indicate that Nodular Basal Cell Carcinoma (NBCC) can be distinguished from other nonmelanoma skin cancer and healthy skin on the base of absorption coefficient, NBCC and Squamous Cell Carcinoma (SCC) can be distinguished from Infiltrative Basal Cell Carcinoma (IBCC) and healthy skin on the base of scattering coefficient.

email: vadim.tyuryaev@ttu.edu

1c. BAYESIAN ADAPTIVELY RANDOMIZED CLINICAL TRIAL OF ENDSTAGE NON-SMALL CELL LUNG CANCER

Valen E. Johnson, University of Texas M.D. Anderson Cancer Center
Chunyan Cai*, University of Texas Health Science Center at Houston and University of Texas M.D. Anderson Cancer Center

Bayesian adaptive randomization clinical trials assign patients to treatments with probabilities that are calculated using the outcomes of previous patients. Recently, the use of Bayesian adaptive trials has increased due to their potential for increasing the number of patients assigned to efficacious treatments. In this article, we develop a Bayesian adaptive randomization design to test the efficacy of 15 combinations of 4 trial agents for reducing symptoms of end-stage non-small cell lung cancer patients. To obtain initial estimates of treatments effects, we assign the first 32 patients to treatments following a randomized factorial design; subsequent patients are assigned to treatments according to the posterior probability that each treatment combination is most efficacious. Bayes factors used in the computation of posterior probabilities are based on non-local (MOM and iMOM) prior densities on treatment effects, which we show increases the rate at which evidence is accumulated in favor of effective treatments. Compared to Bayes factors based on standard objective prior densities, we show that our method provides better operating characteristics and assigns more patients to efficacious treatments.

email: cyretni@gmail.com

1d. A DISTRIBUTION-FREE BAYESIAN METHOD FOR ESTIMATING THE PROBABILITY OF RESPONSE IN COMBINATION DRUG TESTS

John W. Seaman III*, Baylor University
John W. Seaman II, Baylor University
James D. Stamey, Baylor University

There is often interest in combining two drugs for treatment of some disease. Concerns over safety may present ethical problems in testing both drugs simultaneously. Information on the component drugs is often available and can be used to estimate the probability

of an adverse event via a method called proof-loading. We consider a distribution-free Bayesian approach to proof-loading, investigate some of its properties, and propose a novel application to drug safety.

email: john_w_seaman@baylor.edu

1e. A TWO-STAGE DESIGN FOR RANDOMIZED PHASE II CLINICAL TRIALS WITH BIVARIATE BINARY OUTCOME

Rui Qin*, Mayo Clinic
Qian Shi, Mayo Clinic

Randomized phase II clinical trials for screening preliminary efficacy of a new regimen in oncology often use un-definitive endpoints, such as response rate (RR). However, RR does not always predict treatment effect on survival. Incorporating additional endpoints which may be observed later but more close to survival endpoint than RR will increase the likelihood of detecting promising regimens for further validation in large scale phase III studies. In current research, progression-free survival (PFS) rate has been considered in conjunction with RR to develop an innovative randomized phase II design for cancer clinical trials. A Bayesian Dirichlet-multinomial model is adopted for determining the two-stage sequential monitoring and decision-making rules. Sample sizes at both stages are optimized according to desired frequentist performance criteria, i.e. significance level and power. Simulation studies are conducted to evaluate the operating characteristics of this novel randomized phase II design with a bivariate endpoint. We have illustrated its application through a phase II trial of temsirolimus and bevacizumab in patients with ovarian cancer.

email: qin.rui@mayo.edu

1f. IMPROVING SMALL-SAMPLE INFERENCE IN GROUP RANDOMIZED TRIALS WITH BINARY OUTCOMES

Philip Westgate*, University of Michigan
Thomas M. Braun, University of Michigan

Group Randomized Trials (GRTs) randomize groups/clusters of people to treatment or control arms instead of individually randomizing subjects. Typically, GRTs have a small number, n, of independent clusters, each of which can be quite large. When each subject has a binary outcome, over-dispersed binomial data may result, quantified using the intra-cluster correlation coefficient (ICC). Treating the ICC as a nuisance parameter, inference for a treatment effect can be done using quasi-likelihood with a logistic link. A Wald statistic, which asymptotically has a standard normal distribution, can be used to test for a marginal treatment effect. However, we have found in our setting that the Wald statistic may have a variance less than one, resulting in a test size smaller

than its nominal value. When the ICC is known, we develop a method for adjusting the estimated standard error appropriately such that the Wald statistic will approximately have a standard normal distribution. We also propose a way to handle non-nominal test sizes when the ICC is estimated. Through simulation results covering a variety of realistic settings for GRTs, we examine the performance of our methods.

email: pwestgat@umich.edu

1g. A COMPARISON OF TWO SIMULATION MODELS OF CLINICAL TRIALS

Maryna Ptukhina*, Texas Tech University
Clyde Martin, Texas Tech University

Many different models can be developed to describe the treatment effect on patients in clinical trials. We develop two simple simulation models of clinical trials and compare them. Mathematically both models are based on random walk process. We introduce these models as rival simulation models, each of which describe the treatment procedure for the patient with different dosage of the drug treatments and takes into account the individual reaction of the patient to this specific dosage. The models are developed in the following manner. We assume the clinical trial with 500 patients involved in the study for the 300 time points. The initial stage of the patients entering the study is represented by vector Xo. The boundary values 0 and1 represent correspondingly death and remission. We assume two models of the form: Xn+1=(lambda_i)*Xn+e_i, where the parameter lambda represents the dosage of the treatment and e_i are independent and identically distributed random perturbation terms, which represent the individual reaction of the patient. Model one assumes that the distribution of perturbation terms is normal. Model two assumes uniform [0; 1] distribution of perturbation terms. The main result of this work is to show that the model that has exible assumptions about the patient's reaction to the treatment is more realistic and therefore applicable in practice.

email: maryna.ptukhina@ttu.edu

1h. FINDING AND VALIDATING SUBGROUPS IN CLINICAL TRIALS

Jared C. Foster*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Stephen J. Ruberg, Eli Lilly and Company

We consider the problem of subgroups of patients who may have an enhanced treatment effect in a randomized clinical trial, and it is desirable that the subgroup be defined by a limited number of covariates. The development of a standard, pre-determined strategy may help to avoid the well-known dangers of subset analysis. We present two methods developed to find subgroups of enhanced treatment effect. The first method involves the use of logistic regression and forward selection, with the largest possible model being that with all main effects, one and two-way interaction terms of the covariates and the treatment group indicator. The second method, referred to as 'Virtual Twins', involves predicting response probabilities for treatment and control 'twins' for each subject. The difference in these probabilities is then used as the outcome in a regression tree, which can potentially include any set of the covariates. The estimated tree then defines the subgroup of enhanced treatment effect. Simulation studies are presented for situations in which there are and are not true subgroups of enhanced treatment effect, and the methods are compared using a variety of metrics, including area under the curve, sensitivity, specificity, positive and negative predicted values, and a cross-validation-based estimate of treatment effect.

email: jaredcf@umich.edu

2. POSTERS: SURVIVAL ANALYSIS I: METHODOLGY

2a. BAYESIAN INFLUENCE METHODS WITH MISSING COVARIATES IN SURVIVAL ANALYSIS

Diana Lam*, University of North Carolina-Chapel Hill
Joseph Ibrahim, University of North Carolina-Chapel Hill
Hongtu Zhu, University of North Carolina-Chapel Hill

In this talk we formally develop general Bayesian local and global influence methods to carry out sensitivity analyses of perturbations to survival models in the presence of missing covariate data. We examine several types of perturbation schemes for perturbing various assumptions in this setting. In doing so, we show that the metric tensor of a Bayesian perturbation manifold provides useful information for selecting an appropriate perturbation. We also develop several Bayesian local influence measures to identify influential points, assess model assumptions and examine robustness of the proposed model. Simulation studies are conducted to evaluate our methods, and real data sets are analyzed to illustrate the use of our influence measures.

email: dlam@bios.unc.edu

2b. BAYESIAN PREDICTIVE DISTRIBUTIONS UNDER COX'S PROPORTIONAL HAZARD MODEL

Yijie Liao*, Southern Methodist University
Ronald Butler, Southern Methodist University

Bayesian posterior predictive distributions are derived and computed for Cox's proportional hazard model. The priors used on model parameters are semi-parametric; gamma and Dirichlet process priors are used on the baseline survival function while a parametric prior is used on regression parameters. New insights are gained by using a more acceptable form of the likelihood function than has previously been considered. In particular, the posterior

on regression parameters using a gamma process prior agrees with and extends an approximate likelihood function originally derived by Kalbfleisch (1978). A new posterior results for such regression parameters when a Dirichlet process prior is used. Also, contrary to existing literature, it is shown that Cox's partial likelihood cannot be justified as Bayesian in any asymptotic sense and furthermore cannot be construed as the limit of a proper Bayesian posterior distribution.

email: yijiel@smu.edu

2c. BAYESIAN JOINT MODEL FOR LONGITUDINAL AND COMPETING RISKS WITH COPULA

Yi Qian*, Amgen
Deukwoo Kwon, National Cancer Institute
Jeesun Jung, Indiana University School of Medicine

Joint modeling of longitudinal outcome and time to an event of interest is of increasing interest in clinical studies. We propose a Bayesian approach for the joint analysis of longitudinal outcome and competing risks failure time data, where copulas are incorporated to allow for the flexibility of the dependence structure of two random processes in the joint model. Simulation study is conducted to evaluate robustness of the estimates when normality assumption of the random effects is not satisfied.

email: yiq@amgen.com

2d. INFERENCE FOR ACCELERATED FAILURE TIME MODELS FOR CLUSTERED SURVIVAL DATA WITH POTENTIALLY INFORMATIVE CLUSTER SIZE

Jie Fan*, University of Louisville
Somnath Datta, University of Louisville

Accelerated failure time model (AFT) is an important alternative method to the Cox proportional hazards model to analyze time to event data with censored observations. In this work, we consider marginal AFT models for correlated survival data with potentially informative cluster size, which means that the size of the correlated groups may be predictive of their survival characteristics. Two competing proposals, cluster-weighted AFT (CWAFT) marginal model and non-cluster-weighted AFT (NCWAFT) marginal model, are investigated. Simulation and theoretical results show that the CWAFT approach produces unbiased parameter estimation, but that the NCWAFT model does not when the cluster size is informative. We use probability-probability plots to investigate statistical properties of confidence intervals, and adopt Wald tests to examine power properties for the CWAFT model. To illustrate our analysis, we apply the CWAFT model to a dental study.

email: j0fan001@louisville.edu

2e. TIME DEPENDENT CROSS-RATIO ESTIMATION

Tianle Hu*, University of Michigan
Bin Nan, University of Michigan
Xihong Lin, Harvard University
James Robins, Harvard University

In the analysis of bivariate correlated failure time data, it is important to measure the strength of association among the correlated failure times. One commonly used such measure is the cross-ratio. Motivated by the Cox's partial likelihood idea, we propose a novel parametric estimator for the cross-ratio as a continuous function of both components of the bivariate survival times. We show that the proposed parameter estimator is consistent and asymptotically normal. The performance of the proposed technique in finite samples is examined using simulation studies. In addition, the proposed method is applied to the Australian twin data for the estimation of dependence of age at appendectomy between members in the monozygotic and dizygotic twin pairs.

email: hutianle@umich.edu

2f. NONPARAMETRIC SURVIVAL ANALYSIS ON TIME-DEPENDENT COVARIATE EFFECTS

Chan-Hee Jo*, University of Arkansas for Medical Sciences
Chunfeng Huang, Indiana University
Haimeng Zhang, Mississippi State University

Cox's regression model is widely used to assess the influence of exposure with other covariates on mortality or morbidity. In this project, a nonparametric smoothing spline estimator is proposed to study the covariate effects on the survival time in Cox's regression model. We design an efficient algorithm to compute the resultant estimator through the Kalman filter. A data driven procedure is used to choose the smoothing parameter. A simulation study is also presented to demonstrate our method.

email: jochanhee@uams.edu

2g. HAZARD-TYPE EMPIRICAL LIKELIHOOD AND GENERAL ESTIMATING EQUATIONS FOR CENSORED DATA

Yanling Hu*, University of Kentucky
Mai Zhou, University of Kentucky

Qin and Lawless (1994) have shown that (with complete data) the empirical likelihood ratios in terms of distribution function with over-determined estimating equation constraints have very nice asymptotic properties. They may be used to obtain tests or confidence intervals in a way that is analogous to that used with parametric likelihoods. To incorporate censored data, we study here a parallel construct to use a hazard-type empirical likelihood

function and over-determined hazard-type constraints. Over-determined constraint is often used in econometrics, where the number of constraints is larger than the number of parameters. Martingale techniques make the asymptotic analysis easier. Similar asymptotic results of the maximum empirical likelihood estimator and statistics are obtained. Several examples are provided to illustrate the application of this method.

email: vivianhyl@hotmail.com

2h. NONPARAMETRIC REGRESSION MODELS FOR RIGHT-CENSORED DATA USING BERNSTEIN POLYNOMIALS

Muhtarjan Osman*, North Carolina State University
Sujit K. Ghosh, North Carolina State University

In some applications of survival analysis with covariates, the commonly used assumptions (e.g., PH, AFT etc.) may turn out to be stringent and unrealistic, particularly when there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. For instance, in gastric cancer clinical trails, patients receiving only chemotherapy may have a higher survival rate initially but such rates decay much faster compared to a group of patients receiving chemotherapy and radiotherapy. A new nonparametric regression model is developed for conditional hazard rate using a suitable sieve of Bernstein polynomials. The resulting model is shown to nest PH model as a special case. Sieve maximum likelihood estimator is used to obtain the smooth estimators of the conditional survival rate. Large sample properties including semi-parametric consistency, efficiency, and asymptotic normality of the estimator are established under some regularity conditions. Results of simulation studies indicate that the proposed model has reasonably robust performance compared to other semi-parametric models particularly when the semi-parametric assumptions are violated.

email: mosman@ncsu.edu

2i. PRACTICAL MIXED EFFECTS COX MODELS

Terry M. Therneau*, Mayo Clinic

The new R library and function coxme fits general mixed-effects Cox models of the form h(t) = h0(t) exp(X beta + Z b) b ~ N(0, A) where h0 is the baseline hazard, beta is the vector of fixed effects, b the vector of random effects, and X and Z are the covariate matrices. The variance matrix A can be any of the several forms, including user-written specifications. The likelihood is solved using a Laplace approximation, which turns out to be extremely accurate in this setting. I will present examples, discuss several statistical issues that have arisen, and point out areas for future research.

email: therneau@mayo.edu

2j. MODEL CHECKING TECHNIQUES FOR CENSORED LINEAR REGRESSION MODELS

Larry F. Leon*, Bristol-Myers Squibb
Tianxi Cai, Harvard School of Public Health
Lee Jen Wei, Harvard School of Public Health

We present model checking techniques for assessing functional form specifications of covariates in censored linear regression models. The procedures are based on a censored data analog to taking cumulative sums of 'robust' residuals over the space of the covariate under investigation. These cumulative sums are formed by integrating certain Kaplan-Meier estimators and may be viewed as 'robust' censored data analogs to those developed by Lin, Wei, and Ying ('Model-checking techniques based on cumulative residuals', Biometrics, 2002). The null distributions of these stochastic processes can be approximated by computer simulation of certain zero-mean Gaussian processes. Each observed process can be graphically compared with a few realizations from the Gaussian process. We also develop formal test statistics for numerical comparison. Such comparisons enable one to assess objectively whether an apparent trend seen in a residual plot reflects model misspecification or natural variation. We illustrate the methods with a well known dataset and examine the finite sample performance of the test statistics in simulation experiments. In our simulation experiments, the proposed test statistics have good power of detecting misspecification while at the same time controlling the size of the test.

email: larry.leon@bms.com

2k. METHODS FOR MULTIPLY TRUNCATED SURVIVAL DATA: APPLICATION TO AGE OF ONSET OF ALS

Matthew D. Austin, Harvard University
Rebecca A. Betensky, Harvard University

The causes of sporadic amyotrophic lateral sclerosis (ALS) are unknown. Several risk factors have been implicated, including a positive family history and increasing age. A case-control study was performed at Massachusetts General Hospital to examine the role of hypothesized risk factors (i.e. early head trauma) and to identify additional risk factors. Sampling into the study required that subjects had experienced onset and diagnosis of ALS prior to study entry, and that they were alive and being followed at study entry. We propose two models for sequential truncation that cover the scenarios of interest. Within this framework, we propose nonparametric and semi-parametric estimators for the distribution of age of onset that are consistent. The semi-parametric estimators achieve improved efficiency through flexible parametric modeling of age at death (or end of follow-up) and/or age at study entry. We obtain estimates of the median age of onset of ALS of 63.2 (95% CI: 55.7,70,.4) among ALS patients with prior trauma and 66.3 (95% CI: 55.8,71.2) among ALS subjects with no prior trauma, suggesting a possible relationship between early trauma and ALS

onset. We derive asymptotic variance formulas and proofs of consistency, and we validate our results in simulation studies.

email: maustin@hsph.harvard.edu

3. POSTERS: SURVIVAL ANALYSIS II: APPLICATIONS AND METHODOLOGY

3a. EVALUATION OF RISK FACTORS FOR LEFT ATRIO-VENTRICULAR VALVE STENOSIS AFTER ATRIO-VENTRICULAR SEPTAL DEFECT REPAIR: A COMPETING RISKS FRAMEWORK

Adriana C. Dornelles*, Tulane University
Vitor Guerra, Tulane University
Leann Myers, Tulane University

Residual Left-sided atrioventricular valve insufficiency (LAVV) is the main cause for reoperation in patients after repair of atrioventricular septal defect (ASVD). However, LAVV stenosis after repair has not been investigated. The main purpose of this paper is to determine the risk factors and outcome of patients with residual stenosis of LAVV. In a retrospective study from 2001 to 2007, the status of all patients who underwent surgery for atrio-ventricular septal defect was followed up. Among them, it was found three competing outcomes. After ASVD surgery, a patient may survive without further complications; may develop LAVV stenosis (from mild to severe) and may undergo to a reoperation or die. These mutually exclusive end points were analyzed in a competing risk analysis framework. Classical survival analysis as Kaplan-Meier method and Cox regression were also performed.

email: adridornelles@gmail.com

3b. EXAMINE THE DYNAMIC ASSOCIATION BETWEEN BMI AND ALL- CAUSE MORTALITY

Jianghua He*, University of Kansas Medical Center

The association between BMI and mortality has been reported as no association, U-shaped, J-shaped, direct, or even inverse in the epidemiological research. Traditional analysis methods can only model the association between BMI and morality as fixed with the follow-up time. Previous research based on the Framingham Heart Study suggested that there is a dynamic association between BMI and mortality, especially for men. Due to the limitation of the data and analysis methods, only the linear association between BMI and mortality was examined. In this paper, the author examined how the nonlinear association between BMI and mortality changes with the follow-up time. Both time-varying covariate Cox models and time-varying coefficient survival models using nonparametric smoothing are applied. BMI is also transformed to improve the analysis.

email: jhe@kumc.edu

3c. ESTIMATING COLORECTAL CANCER SCREENING IN THE PRESENCE OF MISSING DATA IN A POPULATION WITH A RESISTANT SUBSET AND MULTIPLE OBSERVATIONS

Yolanda Hagar*, University of California-Davis Medical Center
Laurel Beckett, University of California-Davis Medical Center
Joshua Fenton, University of California-Davis Medical Center

Purpose: We use SEER/Medicare data to estimate colorectal cancer screening adherence rates to guidelines set forth by the U.S. government. Analysis of SEER/Medicare data poses substantial challenges. The data are both left truncated and right censored and some individuals may have multiple screening observations while a resistant subset will never be screened. Methods: We propose a Bayesian multivariate parametric model for estimating time to screening. We assume that the number of screenings an individual will receive is a latent random variable, and we calculate likelihoods based on observed screening, length of observation, and whether truncation and/or censoring is present. The parameters of these probabilities are estimated through Gibbs sampler. Results: Simulations have shown acceptable performance for realistic sample sizes. Preliminary results from the data estimate that half of individuals will not be screened. Age and gender significantly affect likelihood of screening, with older people and women having lower adherence. Results to date do not differ by race/ethnicity. Conclusions: Bayesian estimation of a parametric model allows us to characterize adherence to screening guidelines and effects of demographics and policy shifts, despite the challenges in SEER/Medicare data.

email: ychagar@ucdavis.edu

3d. USING Q LEARNING TO CONSTRUCT DYNAMIC TREATMENT REGIMES WITH TIME-TO-EVENT OUTCOMES

Zhiguo Li*, University of Michigan
Susan Murphy, University of Michigan

We propose using the Q learning algorithm to construct dynamic treatment regimes from data in a SMART study (especially a two-stage randomized trial) when the outcome is time to event, which can be censored. The reward functions are chosen such that we optimize the area under the survival curve before the end of follow up (or the restricted mean survival time). In this approach, it is necessary to choose a model for the mean restricted survival time E[min(T,t)], where T is the time to event and t is the follow up time. A linear model may not be adequate, and we propose fitting a varying-coefficient model and check for evidence of nonlinearity of the coefficients. At first, we use local polynomial regression techniques to estimate the varying coefficients. To test if the coefficients are linear or not, we extend two methods in the literature to our setting. One is an F-type test for testing if the nonlinear part of a partly nonlinear model is actually linear, and another one is a generalized quasi likelihood ratio test for testing

whether some of the coefficients are identically 0 in a varying-coefficient model. We explore both the asymptotic properties of the extended tests and their small sample performances. Also, we compare the inverse probability weighting method and the multiple imputation method to account for censoring.

email: zhiguo@umich.edu

3e. MODIFICATIONS AND ALTERNATIVES TO THE SMR IN EVALUATING CENTER-SPECIFIC MORTALITY

Kevin He*, University of Michigan
Douglas E. Schaubel, University of Michigan

Post-transplant mortality may differ significantly across centers. Estimating center effects through fixed or random effects typically involves the unrealistic assumption of proportionality among center-specific hazards. As a solution, the standardized mortality ratio (SMR) has been used to evaluate the center-specific mortality. However, existing asymptotic properties of SMR are based on person-year methods or parametric models with simple intensity functions. The behavior of the SMR under a Cox model is not well-studied. In this study, we first provide a rigorous examination of the SMR under the null and alternative hypotheses. We then develop an alternative method to compute the SMR, which is based on the stratified Cox model and hence remedies some strong limitations of the usual measure. We also propose a class of kernel-smoothed estimators. The measures are process-based and allow one to not only identify if a center's mortality is outlying, but also when during the follow-up the excess mortality is tending to occur. A sup test is developed to evaluate whether the center effects are constant over time. Asymptotic properties of proposed estimators are derived. The finite-sample properties are examined and compared to SMR in simulation studies. The proposed methods are applied to national kidney transplant data.

email: kevinhe@umich.edu

3f. TIME SCALES IN EPIDEMIOLOGICAL ANALYSIS

Prabhakar Chalise*, Mayo Clinic

The Cox proportional hazards model is routinely used to determine the time until an event of interest. Two time scales are used in practice: time-on-study and chronological age. The former is the most frequently used time scale both in clinical studies and longitudinal observational studies. However, there is no general consensus about which time scale is the best. In recent years, papers have appeared arguing for using chronological age as the time scale either with or without adjusting the entry-age. Also, it has been asserted that if the cumulative baseline hazard is exponential or if the age-at-entry is independent of covariate, the two models are equivalent. Our studies do not satisfy these two conditions in general. We found that the true factor that makes the models perform significantly different is the variability in the age-at-entry. Both of our empirical and simulation studies show that time-on-study time scale model using age at entry as a covariate is better than the chronological age. This finding is illustrated with two examples with data from Diverse Population Collaboration group. Based on our findings, we recommend using time-on-study time as a time scale for epidemiological analysis.

email: Chalise.Prabhakar@mayo.edu

3g. SOME GRAPHICAL APPROACHES TO MONITORING THE OUTCOMES OF LIVER TRANSPLANT CENTERS

Jie (Rena) Sun*, University of Michigan
John D. Kalbfleisch, University of Michigan

We present various graphical approaches to monitoring survival outcomes in medical centers over time using nationwide liver transplant centers as an example. A one-sided risk adjusted CUSUM with a constant control limit and an O-E risk-adjusted CUSUM with a V-mask as a control mechanism are introduced and evaluated theoretically and through simulation. We discuss processes associated with reviewing and reacting to signals and of restarting a CUSUM following such review. We also study the performance of both CUSUMs under different departures from the null distribution, and compare the methods through simulation with more traditional approaches to monitoring survival outcomes. Finally, the use of such charts in a national quality improvement program is discussed.

email: renajsun@umich.edu

4. POSTERS: STATISTICAL GENETICS I

4a. COMPOSITE LIKELIHOOD IN LONG SEQUENCE DATA

Bruce G. Lindsay, Penn State University
Jianping Sun*, Penn State University

The primary goal of this talk is the analysis of long sequence data generated in biology, such as SNP data. Suppose we have observed n current descendant sequences of length L, one interesting question is that how to estimate the unknown ancestral distribution from the observed descendants, considering realistic biology complexities such as mutation and recombination. We have developed a statistical model by extending the ancestor mixture model (Chen and Lindsay (2006)) with both mutation and recombination to estimate the ancestral distribution. However, though we can write out the full likelihood for ancestral distribution explicitly, there is an enormous computation challenge when applying it on data due to an enormous number of recombination possibilities, which grows exponentially in sequence length. Therefore, we apply composite likelihood as an

approximation to solve the problem. In this talk, we first introduce our developed statistical model and composite likelihood method. Then, some simulation results are shown to investigate the performance of composite likelihood in long sequence data.

email: jxs1021@psu.edu

4b. A GENERALIZED LINEAR MODEL FOR PEAK CALLING IN CHIP-SEQ DATA

Jialin Xu*, Penn State University
Yu Zhang, Penn State University

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) is a new powerful tool for detecting protein-DNA interactions. Compared with traditional technologies, ChIP-Seq has been shown to produce data in much higher resolution with stronger signal to noise ratios. ChIP-Seq data are discrete and consist of read counts from both forward and backward strands of the chromosomes. Signals from protein-DNA interaction loci typically appear in pairs from both strands. Powerful statistical methods that can most effectively analyze such data, with improved sensitivity and specificity, are therefore urgently needed. In particular, modeling short read counts from the two strands is the first main step towards peak calling of protein-DNA interactions in the genome. We will present a generalized linear model that aims to capture the specific characteristics of real interaction signals in ChIP-Seq data. In our model, we not only consider the signal profile of both strands, but also incorporate a varying parameter for the distance between interaction signals on the two strands of chromosomes. Our first goal is to develop a powerful test statistic to call peaks with sufficient sensitivity and specificity. We then generalize the model to involve data from additional tracks of information to further improve the peak calling accuracy.

email: jxx120@psu.edu

4c. GENOME WISE ASSOCIATION STUDY WITH LONGITUDINALLY OBSERVED DATA: QT INTERVAL

JungBok Lee*, Korea University
Seung Ku, Korea University
Soriul Kim, Korea University
Chol Shin, Korea University
Byoung Cheol Jung, University of Seoul

The rapid increase of GWAS provides an opportunity to examine the potential impact of common genetic variation on complex diseases by systematically cataloging and summarizing key characteristics of the observed associations and the trait/disease. However, most statistical analysis for GWAS was limited to univariate phenotype measurement, which can not reflect complex traits and characteristics of disease in some cases. Recently KARE (Korea Association Resource) project sponsored by Korea CDC has conducted genomewide scanning with KoGes (Korean Genomic Epidemiology Study) which is an ongoing 10 years follow-up study. Based on the data, we perform GWAS with longitudinally observed QT interval data and compare the results between univariate trait and longitudinally observed phenotypes.

email: jungboklee@korea.ac.kr

4d. A NEW VARIABLE SELECTION METHOD FOR GENOME-WIDE ASSOCIATION STUDIES

Qianchuan He*, University of North Carolina-Chapel Hill
Danyu Lin, University of North Carolina-Chapel Hill

Variable selection for genome-wide association studies (GWAS) is very challenging because of the extremely high dimension and strong correlation among single nucleotide polymorphisms (SNPs). We introduce GWASelect, a method designed for variable selection at the genome-wide level. This approach takes advantage of several recent advances in variable selection and performs impressively well under a series of scenarios. We show through simulation studies that our approach is capable of capturing important SNPs that are either marginally correlated or marginally uncorrelated with the disease status. Moreover, it has markedly reduced false discovery rates compared to other variable selection methods. Applying our method to the Wellcome Trust Case-Control Consortium (WTCCC) data leads to a number of novel discoveries. Our method provides a powerful new tool for variable selection at the scale of half million SNPs.

email: heqianch@email.unc.edu

4e. IS IT RARE OR COMMON? A COALESCENT TREE APPROACH TO IDENTIFY THE GENETIC TYPES OF VARIANTS UNDERLYING COMPLEX DISEASES

Kaustubh Adhikari*, Harvard University

An important problem in modern genetic studies is whether a given disease is caused by a few underlying common variants or by several rare variants in a gene. This paper presents a unique method of identifying the type of variants in a gene segment (marked for detailed analysis through candidate-gene, functional pathway or GWAS study) responsible for a disease through SNP genotyping in that segment. Based on case-control data of SNP genotypes in a gene segment, this method obtains joint posterior distribution of the number of common and rare variants present in the segment (phenocopy is allowed). We do a Bayesian modeling using coalescent genealogical trees to model the unknown ancestral history of both the cases and the controls jointly. A particular type on Ancestral Recombination Graphs - minARG - is used for the modeling. Such trees have been used frequently in the past to identify SNPs associated with causal variants, but the complexity grows out of hand with thousands of genome-wide SNPs. This

method, applied to a candidate-gene segment, utilizes the ability of ARGs to model the ancestral structural history (with mutation and recombination) by employing the information present in the SNP markers.

email: kadhikar@fas.harvard.edu

4f. EFFECTS OF POPULATION STRATIFICATION IN LOGISTIC REGRESSION AND AN ALTERNATIVE TO ACHIEVING GREATER POWER

Adrian Tan*, University of Minnesota
Saonli Basu, University of Minnesota

Population stratification has been shown to be a cause of spurious associations in whole genome association case-control studies. To a significant extent, this problem has been resolved by recent methods that corrects for population stratification by incorporating measures of ancestry via population membership coefficients from STRUCTURE or ancestral principal components from PCA in methods such as Structured Association, Eigenstrat and Stratscore. In this study, we investigate the effects of stratification in Logistic Regression applied to Case Control studies in 2 common scenarios: 1) Difference in allele frequencies between multiple discrete populations, and 2) Difference in disease prevalence between multiple discrete populations. We derive the closed forms for the bias and variance distortion of the estimated log odds under recessive, additive and dominant genotype models when stratification is unaccounted for and propose an alternative to achieving greater power while maintaining type I error.

email: tanxx201@umn.edu

4g. THE EFFECT OF RETROSPECTIVE SAMPLING ON ESTIMATES OF PREDICTION ERROR FOR MULTIFACTOR DIMENSIONALITY REDUCTION

Stacey J. Winham*, North Carolina State University
Alison A. Motsinger-Reif, North Carolina State University

Recently, a number of novel analytical approaches have been developed for genetic epidemiological studies to identify predictive models that account for complex etiologies. Multifactor Dimensionality Reduction (MDR) is a highly successful data-mining method designed to generate testable hypotheses about possible gene-gene interactions to be further investigated by geneticists, and relies on classification error in conjunction with cross-validation from retrospective case-control data to rank and test potential models. Previous work has focused on power to detect functional loci, but has not considered bias and variance of prediction error estimates. These error estimates are frequently reported, particularly for prediction, and accuracy is critical in terms of proper prioritization of identified models for follow-up study. We evaluate the bias and variance of the MDR error estimate and show that MDR can both underestimate and overestimate error, in part because of retrospective sampling in case-control studies. We argue that a prospective error estimate is necessary if the model is to be used for prediction and prioritization, and propose and demonstrate the use of an estimate constructed with bootstrap resampling to accurately estimate prospective error. The proposed estimation is potentially applicable to all data-mining methods that estimate classification and prediction errors.

email: staceyjeanwood@gmail.com

4h. STATISTICAL MODELS FOR DETECTING RARE VARIANTS ASSOCIATED WITH DISEASE

Jeesun Jung*, Indiana University School of Medicine
Deukwoo Kwon, National Cancer Institute

Detecting common variants associated with common disease has been successful in a framework of genome-wide association studies (GWAS) where rare variants have not been focused to study. Due to recent advance in sequencing technologies, deep re-sequencing data becomes available to discover the rare variants influencing a disease. In this study, we propose a novel statistical method for identifying disease associated rare variants with two scenarios: (1) 1-5% of rare variants, and (2) less than 1% of rare variants. We assume the variants have poisson or zero-inflated poisson depending on rate of variants and test for association of a disease using score test statistics. Based on simulation studies, we performed power and type I error rate studies and we have demonstrated that our proposed method is statistically robust and achieves a good power to detect differences between case and control subjects.

email: jeejung@iupui.edu

5. POSTERS: STATISTICAL GENETICS II

5a. COMPARISON OF CONDITIONAL AND UNCONDITIONAL ANALYSIS OF LEFT-TRUNCATED DATA: SIMULATION STUDY AND APPLICATION

Lydia C. Kwee*, Duke University Medical Center
Silke Schmidt, Duke University Medical Center

Observational studies of rare diseases frequently utilize a case-control design, with both incident and prevalent cases sampled retrospectively. For progressive diseases, subjects may be followed longitudinally in order to simultaneously study factors which affect disease incidence and/or survival. Data from such 'prevalent cohort' studies are left-truncated since subjects enter the study at variable times after diagnosis and hence, only subjects who survive long enough to be sampled are included. Without proper adjustment, a survival analysis using the Cox model will yield biased hazard ratio estimates. As a special case of left-truncated data, length-biased data

are observed when the underlying disease incidence is a stationary Poisson process. Here, we use a simulation study to compare the traditional conditional survival analysis for length-biased data with a recently proposed unconditional method with greater efficiency (Qin & Shen 2009). We simulate covariate influences on incidence only, survival only, neither, or both. We also apply the conditional and unconditional analysis to data from the National Registry of Veterans with ALS in order to evaluate whether coding changes in HFE, a previously implicated candidate gene, are associated with the incidence of, or survival with, ALS.

email: lydia.kwee@duke.edu

5b. DETECTING COPY NUMBER VARIATION VIA MIXTURE-MODEL AND ROC CURVE

Hui-Min Lin*, National Center for Toxicological Research, U.S. Food and Drug Administration
Ching-Wei Chang, National Center for Toxicological Research, U.S. Food and Drug Administration
James Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

DNA copy number variation (CNV) is a segment of DNA in which copy number amplifications or deletions have been found by comparing with a normal reference genomes. CNV has long been known as an important role in complex diseases and correlated with the degree of disease predisposition. Therefore, to identify the exact copy number is important to understand genesis and progression of human diseases. Recently, high-throughput technology has been developed to detect changes in chromosomal copy number. However, there are some technical problems arisen from using the Affymetrix SNP chips. Firstly, intensities may vary among array elements even if there are no copy number changes. Secondly, the sample size is not large enough to perform powerful statistical testing or modeling. Most of current methods are only used for detecting gain or lost, could not be applied to specify the exact copy number. In this study, we develop a procedure based on mixture-model for automatically detecting the number of copy number distribution. With combining the concept of receiver operating characteristic (ROC) curve, we can not only detect gain or lost but also provide cut-off estimation to recognize the exact copy number. A simulation study and a real data analysis will be used to illustrate our procedure.

email: HuiMin.Lin@fda.hhs.gov

5c. UTILIZING GENOTYPE IMPUTATION FOR THE AUGMENTATION OF SEQUENCE DATA

Brooke L. Fridley*, Mayo Clinic
Gregory Jenkins, Mayo Clinic
Matthew Deyo-Svendsen, Mayo Clinic
Scott Hebbring, Mayo Clinic

The advancement in genotyping technology has led to an increase in the number of genome-wide association studies (GWAS). However, the variants identified from these GWAS are not necessarily the functional variants, with the next phase in GWAS involving the refining of these putative loci. One possible approach for refining the locus would be to catalog all variants via sequencing, followed by association analysis. However, sequencing a locus in a large number of subjects is still relatively expensive. By sequencing only a portion of the samples, followed by imputation in the remaining samples, one can significantly reduced the cost to localize the punitive variant. A potentially attractive alternative option would be imputation based only on the 1000 Genomes Project; however, this has the drawbacks of using a reference population that does not necessarily match with respect to disease status and LD pattern. We conducted a study to investigate the use of genotype imputation using sequencing data for a fraction of the study participants. Various approaches were implemented using data from both a sequencing study conducted at the Mayo Clinic and the 1000 Genomes Project. Our results show that imputation based on sequencing a portion of the study participants is a reasonable, cost-saving, approach for disease mapping and the refinement of putative loci detected from GWAS.

email: fridley.brooke@mayo.edu

5d. TREES ASSEMBLING BASED MANN-WHITNEY TEST FOR LARGESCALE GENETIC ASSOCIATION STUDY

Changshuai Wei*, Michigan State University
Qing Lu, Michigan State University

The complex genetics architecture of common disease requires more powerful and computationally efficient statistical tools. While new methods have been proposed for the gene-gene interaction association analysis, many of them lack the ability to tack complex interaction among a large number of genetic variants. Here, based on Mann-Whitney test, we propose a novel non-parametric method to assess gene-gene interaction on high dimension data. It adopts a randomization mechanism to strengthen the method's reliability and a tree-assembling procedure to improve the method's statistical power. Through simulation, we showed the new method was computationally more efficient and more powerful over the existing methods, such as the Multifactor Dimensionality Reduction method. In particular, we found the new method had much more advantage over the existing methods when a large number of loci, mostly associated with small effect size, influence the disease. We believe such disease model is more reasonable for complex diseases. Previously, 29 loci were found to be associated with Crohn's Disease. We evaluated their joint effect by using the Wellcome Trust Genome-Wide Crohn's Disease dataset. The new method identified a strong association of 29 loci with Crohn's Disease (P-value < 1e-14), while MDR obtains a less significant association (P-value = 5.81e-5).

email: changs18@msu.edu

5e. ARTIFACT DUE TO DIFFERENTIAL GENOTYPING ERROR WHEN CASES AND CONTROLS ARE GENOTYPED USING DIFFERENT PLATFORMS

Jennifer A. Sinnott*, Harvard University
Peter Kraft, Harvard University

When cases and controls are genotyped on different platforms in a GWAS, the platforms produce different collections of SNPs, so researchers impute SNPs to get a data set suitable for standard analysis. This imputation introduces measurement error into the analysis, and we investigated the effect of this error. We compared genotype frequencies of two groups of healthy controls from the Nurses' Health Study -- 1370 controls genotyped on Affymetrix and 1038 controls genotyped on Illumina. We observed many more statistically significant SNPs than expected: 6247 SNPs out of 825956 (0.8%) were significant at the 5e-8 level, and the genomic control lambda was 1.47. We explored three methods for controlling for this problem. One method was to restrict to SNPs of highest quality imputation; another was to remove platform effects using Eigenstrat; and a third was to genotype some controls alongside the cases and use them to isolate SNPs that are statistical artifact before proceeding with analysis. We evaluate these methods and compare their effectiveness on our data. Researchers using this type of data need to be aware of the inflation in error rate, and should consider controlling for it at the design or analysis stage with one of our methods.

email: jsinnott@hsph.harvard.edu

5f. A CROSS-VALIDATED BAGGING ROC METHOD FOR PREDICTIVE GENETIC TESTS

Chengyin Ye*, Michigan State University and Zhejiang University
Yuehua Cui, Michigan State University
Robert C. Elston, Case Western Reserve University
Jun Zhu, Zhejiang University
Qing Lu, Michigan State University

Recent Investigations in genome-wide association studies have identified numerous genetic variants predisposing to common complex diseases. These genetic variants, combined with the existing clinical/genetic risk factors, bring new opportunities for early disease prediction. Meanwhile, it also raises a statistical challenge of combining a large number of variants and their possible interactions for disease prediction. To fulfill this need, we propose a novel nonparametric algorithm, called the cross-validated bagging ROC method. The new method is based on the concept of the optimal ROC curve, thus, theoretically, it can build a test with many ideal properties (e.g., having the highest discriminative ability). By adopting both a forward selection algorithm and a cross-validated bagging procedure, the new method is able to handle a large number of genetic and environmental risk predictors, taking their possible interactions into consideration, while maintains robust performance. In addition, we also introduce into the method an efficient procedure to handle missing data. Through simulations and real data application, we compared the new method with classification and regression tree and the allele counting methods, and found that the new method performed better than the other two methods.

email: cye@epi.msu.edu

6. POSTERS: IMAGING AND SPATIAL MODELING

6a. MULTISCALE ADAPTIVE SMOOTHING MODELS FOR FUNCTIONAL IMAGING CONSTRUCTION, SEGMENTATION AND CLASSIFICATION

Jiaping Wang*, University of North Carolina-Chapel Hill
Hongtu Zhu, University of North Carolina-Chapel Hill
Weili Lin, University of North Carolina-Chapel Hill

Functional imaging studies analyze imaging data with complex spatial and temporal correlation structures and varied activation patterns on a 2D surface or in a 3D volume. This paper develops a multiscale adaptive smoothing model (MASM) for spatial and adaptive analysis of functional imaging data. Compared with the existing smoothing approaches, MASM has four unique features: spatial, connected, hierarchical and adaptive. MASM not only creates adaptive ellipsoid at each location (called voxel) but also groups them into homogeneous clusters. MASM analyzes all observations in the ellipsoid of each voxel and its homogeneous cluster. These consecutively connected ellipsoids across all voxels can capture spatial dependence among imaging observations while these homogeneous clusters allow combining spatial disconnected regions. Finally, MASM combines imaging observations with adaptive weights in the voxels within the ellipsoid of the current voxel to adaptively, spatially smooth functional images. Theoretically, we establish consistency of the adaptive estimates under some mild conditions. Three sets of simulation studies demonstrate the methodology and examine its finite sample performance in imaging construction, segmentation and classification. Our simulation studies and real data analysis confirm that MASM significantly outperforms the existing methods.

email: jwang@bios.unc.edu

6b. PREDICTING POST-TREATMENT NEURAL ACTIVITY BASED ON PRETREATMENT FUNCTIONAL NEUROIMAGING DATA

Gordana Derado*, Emory University
F.D. Bowman, Emory University

There is growing interest in increasing the clinical applicability of functional neuroimaging data, for example for diagnostic purposes and for predicting patients' future health outcomes. Researchers

have long sought to predict response to antidepressant treatments and a number of key predictors have been identified and replicated in previous studies (Kemp et al., 2008). However, a number of methodological issues, including small sample sizes, heterogeneity, and subtypes of depression have hindered previous research. In context of resting-state neuroimaging data, these issues may lead to variability in the model parameter estimators and limited sensitivity and specificity of the predictors. We propose a novel Bayesian hierarchical framework for predicting post-treatment neural activity based on the pre-treatment functional neuroimaging data that attempts to overcome the aforementioned shortcomings by borrowing strength from the spatial correlations present in the data. We apply our proposed methodology to the data from a study on depression.

email: gderado@emory.edu

6c. ENHANCED GLOBAL ERROR RATE CONTROL IN NEUROIMAGING DATA

Shuzhen Li*, University of Minnesota
Lynn Eberly, University of Minnesota

When we threshold a statistical map of neuroimaging data, it is vital for us to account for the problem of multiple testing. False Discovery Rate (FDR) has been commonly used in the context of large scale analyses. Many methods have been proposed for this issue. Storey (2002) introduced the pFDR error measure and also showed its better performance both on sensitivity and specificity than the standard FDR measure under an independence assumption. Langers et al. (2007) provided an approach of regional control of the global false discovery rate taking into consideration the natural clustered nature of neuroimaging data. In this paper, we consider several modifications of Storey and Langers to estimate pFDR and control the global error rate for neuroimaging data. We use simulation study to examine whether the proposed methods performs significantly better.

email: lixxx466@gmail.com

6d. MOTION CORRECTION FOR TWO-PHOTON LASER-SCANNING MICROSCOPY

Mihaela Obreja*, University of Pittsburgh
William Eddy, Carnegie Mellon University

Two-photon laser-scanning microscopy (TPLSM) can be used for in vivo neuroimaging of small animals. Due to the very high resolution of the images, brain motion is a source of large artifacts; tissue may be displaced by 10 or more pixels from its rest position. Thus, because the scanning rate is relatively slow comparing with the cardiac and respiratory cycles, some tissue pixels are scanned several times while others are never scanned. Consequently, although the images superficially appear reasonable, they can lead to incorrect conclusions with respect to brain structure and function. As a line is scanned almost instantaneous (~1ms), our problem is reduced to relocating each of the lines in a three-dimensional stack of images to its 'correct' location. Addressing the motion effects, we describe a Hidden Markov Model to estimate the sequence of hidden states most likely to have generated the sequence of observations. Our algorithm assigns probabilities for the states based on concomitant physiological measurements and estimates the most likely path of observed lines from the areas which move the least. Because there is no gold standard for comparison we compare our result with an image collected after the animal is sacrificed.

email: mio8@pitt.edu

6e. TESTING SIMILARITY OF 2D ELECTROPHORESIS GELS ACROSS GROUPS BASED ON INDEPENDENT COMPONENT ANALYSIS

Hui Huang*, University of Maryland Baltimore County
Anindya Roy, University of Maryland Baltimore County
Nicolle Correa, University of Maryland Baltimore County
Tulay Adali, University of Maryland Baltimore County

Statistical procedures are often used to pre-screen spots that may be differentially expressed across groups of two dimensional electrophoresis gels. Most of the commonly used inference procedures are univariate and have poor detection capabilities due to their reliance on unrealistic assumptions and result in a large number of false positives, thereby reducing the efficiency and costing effectiveness of the statistical pre-screening procedures. We develop a completely data-driven statistical approach that provides accurate identification of statistically significant differences in protein expression profiles across groups. Our methodology relies on two data-driven techniques. First, features are extracted from each group of gels using a data-driven approach for feature extraction -independent component analysis (ICA). Second, based on bootstrap resampling technique, we develop a novel data-driven testing procedure for detecting group differences in the features across groups. The procedures are based on Kolmogorov-Smirnov type statistics that are appropriate for spatial dependence in data such as in gel images. Simulation study based on synthetic gels shows that the testing procedure has high efficiency in detecting group differences. The methodology is also illustrated via a set of real2DE gels.

email: hh2@umbc.edu

7. POSTERS: GENOMICS AND BIOMARKERS

7a. REGULARIZED GAUSSIAN MIXTURE MODELING WITH COVARIANCE SHRINKAGE

Hyang Min Lee*, Penn State University
Jia Li, Penn State University

We introduce a covariance shrinkage method for Gaussian mixture models that allow different components to have different levels of complexity. A complexity parameter is assigned to each component to determine the extent of shrinkage towards a diagonal or common covariance matrix. A BIC-type penalized log-likelihood is proposed to estimate the model parameters and the complexity parameters. A generalized EM algorithm is developed for model estimation. Based on both simulated and real data sets, we will compare the proposed covariance shrinkage method with covariance shrinkage using a single complexity parameters and estimation without shrinkage.

email: hul145@psu.edu

7b. EFFECT OF GENE REFERENCE SELECTION ON ENRICHMENT ANALYSIS OF GENE LISTS

Laura Kelly Vaughan*, University of Alabama-Birmingham

Gene functional enrichment analysis has become a powerful tool in the analysis of lists of genes obtained from high dimensional genomics experiments such as gene expression microarrays and genotyping chips. There are numerous methods which have been designed to aid interpretation and identify interesting genes for further analysis by identifying groups of genes which occur more likely than what would be expected by chance alone. In addition to the statistical methodology employed, this enrichment analysis is highly dependent on the selection of the background or reference gene list. Although the importance of the reference list has been discussed, there have been no systematic studies illustrating the effect of list selection. Here we present the results of the selection of reference background on enrichment analysis of a gene list derived from a genome wide association study. We employed several popular tools, with background defined as all genes 1) in the human genome 2) on the analysis platform 3) that have annotations and 4) on the platform which are annotated. Our results indicate that there is no single "gold standard" background and that researchers should specify a reference list based on each study and analysis method.

email: lkvaughan@uab.edu

7c. INCORPORATION OF PRIOR INFORMATION INTO LASSO VIA LINEAR CONSTRAINTS

Tianhong He*, Purdue University
Michael Zhu, Purdue University

The Lasso proposed by Tibshirani (1996) has become a popular variable selection method for high dimensional data analysis. Much effort has been dedicated to the further improvement of Lasso in recent statistical literature. It is well-known that incorporation of prior information regarding predictor variables can lead to more accurate estimates of the regression coefficients in linear regression. In this article, we propose a systematic approach to incorporating prior information into the Lasso via linear constraints. An efficient algorithm has been developed to compute the lasso solution under linear constraints and the theoretical properties of the resulting estimates including estimation and variable selection consistencies have been established. We use the proposed method to incorporate prior knowledge from regulatory networks or metabolic pathways study into genomic data analysis. In contrast with some existing methods that represent prior knowledge by quadratic penalty functions, we use linear constraints instead, which lead to more interpretable results.

email: het@purdue.edu

7d. MULTIPLE IMPUTATION FOR MISSING VALUES IN MICROARRAY DATA ANALYSIS

Richard E. Kennedy*, University of Alabama-Birmingham
Hemant K. Tiwari, University of Alabama-Birmingham

Several imputation algorithms have been proposed to estimate the missing values that are often encountered when analyzing microarray gene expression data. All of these methods utilize single imputation, which provides estimates of the missing gene expression values but does not provide measures of uncertainty associated with the estimates. Furthermore, these methods have been evaluated with simulations that assume that may not accurately reflect the patterns seen in real microarray datasets. We present an application of multiple imputation (MI) as an alternative to impute probable expression values and associated measures of uncertainty. We validate the MI process with a missing at random (MAR) simulation using other covariate information in the linear model context, as well as the missing completely at random (MCAR) and not missing at random (NMAR) deletion of entire genes that have been used in previous studies, across a range of percentages for missingness. We investigate bias and root mean square error (RMSE) of the estimates, as well as the effects of MI on the declaration of differential gene expression.

email: rkennedy@ms.soph.uab.edu

7e. GENE CLUSTERING AND IDENTIFICATION USING COMPOSITE LIKELIHOOD

Ran Li*, University of Minnesota
Baolin Wu, University of Minnesota

In this paper, we propose a two-step procedure that incorporates the correlation between genes into the process of gene identification. Our approach uses the correlation information to cluster genes into groups and then apply lasso on genes groups. We define supergenes as the representative for a group of genes and use the fitted value from a linear regression (where the gene measurement in the current group is used as the covariate and the clinical phenotype is used as the response) as the pseudo measurement for the supergenes. Penalized likelihood approach was

adopted to select these supergenes. Our approach utilized not only the positive correlations between different genes within the same group but also negative correlations. It is more richly parameterized and outperforms Lasso on both the simulated data and real data.

email: liran1061@hotmail.com

7f. HYBRID POOLED-UNPOOLED DESIGN FOR COST-EFFICIENT MEASUREMENT OF BIOMARKERS

Enrique F. Schisterman, National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Sunni Mumford, National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Albert Vexler, SUNY, Albany
Neil J. Perkins*, National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Evaluating biomarkers in epidemiological studies can be expensive and time consuming. Investigators are often forced to use techniques such as random sampling or pooling biospecimens in order to cut costs and save time. Random sampling provides data that can be easily analyzed. However, random sampling methods are not optimal cost-efficient designs for estimating means. Pooling can be much more efficient but pooled data are strongly restricted by distributional assumptions which are challenging to validate. We propose and examine a cost-efficient hybrid design that involves taking a sample of both pooled and unpooled data in an optimal proportion in order to efficiently estimate the unknown parameters of the biomarker distribution. In addition, we find that this design can be utilized to estimate and account for different types of measurement and pooling error, without the need to collect validation data or repeated measurements. The hybrid design, when applied, leads to minimization of a given loss function based on variances of the estimators of the unknown parameters. This optimization with respect to a quantity of interest is shown via Monte Carlo simulation and exemplified using biomarker data from a study on coronary heart disease.

email: perkinsn@mail.nih.gov

7g. WEAKEST-LINK MODELS FOR JOINT EFFECTS IN CELL-BASED DATA

The Minh Luong*, University of Pittsburgh
Roger Day, University of Pittsburgh

The joint effect of multiple biomarkers strongly associated with outcomes may point to an important molecular mechanism in cancer. Some methods for detecting interesting variable combinations require categorization and lose information, while others do not plausibly model the underlying biology. The weakest-link paradigm states that, for almost all covariate space points, the mechanism's level of activity is insensitive to changes from all covariates except one, labeled the "weakest-link" covariate for that point. However, the identity of this weakest-link varies across the covariate space, and is determined by projecting onto the curve of optimal use (COU). The COU is an intersection of (p-1)-dimensional surfaces in p-dimensional covariate space that sweeps out optimal combinations of ingredient quantities. We began with multi-parameter cytometry lung cancer data, consisting of thousands of cells per patient. We used a weakest-link model for the joint effect of four biomarkers within individual cells, in light of a previous breast cancer study suggesting that accounting for cells where simultaneous abnormalities occur could provide additional prognostic information. This weakest-link model, compared to logic regression and linear regression, performed the best in predicting recurrence-free survival, according to cross-validation criteria.

email: thl13@pitt.edu

8. POSTERS: LONGITUDINAL DATA ANALYSIS

8a. COVARIATE ADJUSTMENT IN LATENT CLASS MODELS FOR JOINT ANALYSIS OF LONGITUDINAL AND TIME-TO-EVENT OUTCOMES

Benjamin E. Leiby*, Thomas Jefferson University
Mary D. Sammel, University of Pennsylvania
Terry Hyslop, Thomas Jefferson University

Joint analysis of longitudinal and time-to-event outcomes is increasingly common in the medical literature. Among the approaches to joint analysis are latent class models which assume subjects belong to a latent class with some probability. The trajectory of the longitudinal outcome and the distribution of the time-to-event outcome differ by latent class. These models allow for adjustment for covariates in multiple places: the class-specific models for the time-to-event outcome, the class-specific trajectory model for the longitudinal outcome, and the class probability model. Different placement of covariates yields different interpretation of the covariate effect. Neither the appropriate place for covariates nor the effect of misplacement of the covariates in modeling is well-understood. We explore issues in covariate adjustment in latent class models through simulation studies and a case study where we assess the effect of treatment in a joint analysis of repeated biomarker measurements and time to recurrence of colon cancer.

email: bleiby@mail.jci.tju.edu

8b. JOINT MODELING OF PRIMARY OUTCOME AND LONGITUDINAL DATA MEASURED AT INFORMATIVE OBSERVATION TIMES

Song Yan*, North Carolina State University
Daowen Zhang, North Carolina State University
Wenbin Lu, North Carolina State University

In some biomedical research, we are interested in the relationship between a primary outcome and longitudinal data profiles which are taken at informative observation times. We propose a joint model which consists of (1) the frailty cox model for informative observation times, (2) longitudinal semiparametric mixed model, (3) logistic model for primary binary outcome. These three submodels are linked by subject-specific random effects. The estimation can be conveniently accomplished by Gaussian quadrature techniques, e.g., SAS Proc NLMIXED or by Monte Carlo EM algorithm. The proposed joint model is evaluated by simulation and is applied to a study that investigates the relationship between pregnancy outcome and early beta-human chorionic gonadotrophin (beta-HCG) among patients.

email: yansong@gmail.com

8c. THE TWO-GROUP LATENT GROWTH MODELING IN STUDYING OF CHANGE OVER TIME

Yingchun Zhan*, Texas Tech University
Du Feng, Texas Tech University
Clyde Martin, Texas Tech University

The latent growth model (LG) is one of the main methodologies applied to study the change of repeated measures. We introduce a two-group latent growth (LG) model by Muthen and Curran (1997; see also Curran and Muthen, 1998) and compare it with the classical and traditional repeated measure ANOVA. Briefly, the two-group LG model estimations involve two groups: a control group in the context providing a normative growth trajectory from which the intervention group is estimated by adding a treatment factor. Thus, the added treatment factor in the intervention group captures the incremental or decremental effect. The individual growth models in the comparison and the intervention group are the following: y_ti=•_0i+•_1i x_t+μ_ti (Comparison group) and y_ti=•_0i+•_1i x_t+•_add a_(add)+μ_ti (Intervention group). An example of a longitudinal data in children obesity intervention study is used on these two methods. The outcome variables include subjects' BMI-for-age-and-gender (BMI percentile), waist circumference and body composition. A computer program Mplus is applied to study the changes of the outcome variables.

email: yingchun.zhan@ttuhsc.edu

8d. A WILCOXON-TYPE STATISTIC FOR REPEATED BINARY MEASURES WITH MULTIPLE OUTCOMES

Okan U. Elci*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh

In clinical trials, we often compare two treatment groups using repeated binary measures over time. In such trials, we may encounter missing observations, adverse side effects, or non-responsiveness to therapy which for ethical reasons, may result in increased medical intervention beyond the protocol therapy. We developed a family of statistical tests based on the Wilcoxon statistic which orders the vectors of repeated binary observations and events where the ordering is determined by "clinical relevance". For some scenarios, clinically meaningful ordering of the vectors may be defined by a natural algorithm, while for other scenarios the ordering is obtained from a group of clinicians. We present the statistical development of the proposed method, effects of the variability of rankings among clinicians, examples of the application of the proposed method using data from a clinical trial on otitis media, and simulation studies comparing the statistical power of the proposed method to more traditional methods of analysis. Our simulation studies indicate that the proposed method is competitive with and, for some scenarios, is preferable to the traditional methods.

email: oue1@pitt.edu

8e. MODELING FOR LONGITUDINAL DATA WITH HIGH DROPOUT RATES

Sunkyung Yu*, Yale University
James Dziura, Yale University
Melissa M. Shaw, Yale University
Mary Savoye, Yale University

Many longitudinal studies such as randomized clinical trials (RCT) for weight loss have incomplete and unbalanced data due to dropout, resulting in loss of statistical power and biased inference on outcomes of interest. Currently, there are a number of alternative analytic methods available to handle missing data. Our goal is to identify analytic techniques that perform best when dropout is high. Using data from an RCT of a behavioral weight loss intervention in a pediatric population with >50% dropout at the 2-year endpoint, we will compare results and inferences applying several missing data techniques. Datasets will be generated from the raw data under null and alternative hypotheses. Type I and II error probabilities will be evaluated using complete case, Last Observation Carried Forward (LOCF), mixed model, multiple imputation and a pattern-mixture model.

email: sunkyung.yu@yale.edu

8f. A PERMUTATION TEST FOR RANDOM EFFECTS IN LINEAR MIXED MODELS

Oliver Lee*, University of Michigan
Thomas M. Braun, University of Michigan

Inference regarding the inclusion or exclusion of random effects in linear mixed models is challenging because the variance components are located on the boundary of their parameter space under the typical null hypothesis. As a result, the asymptotic null distribution of the Wald, score, and likelihood ratio tests will not have a chi-squared distribution under the null hypothesis. Although it has been proven that the correct null distribution is a mixture of chi-squared distributions, the appropriate mixture distribution is rather cumbersome and non-intuitive when the null and alternative hypotheses differ by more than one random effect. As an alternative, we present a permutation test whose statistic is a sum of weighted squared residuals, with the weights determined by the among- and within-subject variance components. The null permutation distribution of our statistic is computed by permuting the residuals both within-and among-subjects and is valid both asymptotically as well as in small samples. We examine the size of our test via simulation in a variety of settings and compare it to the size based upon the classical mixture-of-chi-squares approach.

email: oel@umich.edu

9. POSTERS: SPATIAL/TEMPORAL MODELING AND ENVIRONMENTAL/ECOLOGICAL APPLICATIONS

9a. ESTIMATION OF Br CONCENTRATION DISTRIBUTION IN GROUNDWATER

Yunjin Park*, Dongguk University, Korea
Yongsung Joo, Dongguk University, Korea

Br is a key element that indicates seawater intrusion in groundwater aquifer. However, distribution of Br concentration could not be properly estimated because of zero-inflation problem due to freshwater-dominant groundwater system and censoring problem due to low concentration level. In this paper, we solve these two problems using EM algorithm.

email: jin63945@dongguk.edu

9b. CUMULATIVE DIETARY EXPOSURE TO MALATHION AND CHLOYPYRIFOS IN THE NHEXAS-MARYLAND INVESTIGATION

Anne M. Riederer, Emory University
Ayona Chatterjee*, University of West Georgia
Scott M. Bartell, University of California
Barry P. Ryan, Emory University

Malathion and chlorpyrifos are pesticides widely used in homes and farms. A possible source of these pesticides entering the human body is by the consumption of contaminated food. This study looks at exposure assessment of these pesticides through the diet. Data from the NHEXAS-MD diet study provides information about the levels on malathion and chlorpyrifos for 405 individuals along with the amount of consumption for 157 foods for each individual. The aim of the study is to identify significant foods among these 157 foods that contribute to non-zero levels of both pesticides in the body. A Bayesian model is developed to model both the pesticide levels and the food consumption values. A Bayesian latent Gaussian model is established to account for the left censored pesticide levels. Since only a handful of food items were consumed by an individual, the consumption data set has large number of zero intakes. To model the presence of large amount of zero consumption values, we give each individual a particular propensity for consumption of a given food item. The analysis identifies significant food contributors for the individual and combined levels of malathion and chlorpyrifos. Predicted levels of pesticides are obtained using the significant foods and are compared to the observed data. The results obtained from the Bayesian model are also compared with those obtained from a Tobit regression.

email: achatter@westga.edu

9c. ADJUSTING FOR MEASUREMENT ERROR IN MATERNAL PM EXPOSURE AND BIRTH WEIGHT MODELS

Simone Gray*, Duke University
Alan Gelfand, Duke University
Marie Lynn Miranda, Duke University

In environmental health studies air pollution measurements from the closest monitor are commonly used as a proxy for personal exposure. This technique assumes that pollution concentrations are spatially homogeneous and consequently introduces measurement error into a model. To model the relationship between maternal exposure to air pollution and birth weight we build a hierarchical model that accounts for this associated measurement error. We allow four possible scenarios, with increasing flexibility, for capturing this uncertainty. In the two simplest cases we specify one model with a constant variance term and another with a variance component that allows the uncertainty in the exposure measurements to increase as the distance between maternal residence and the location of the closest monitor increases. In the second two models we introduce spatial dependence in these errors using spatial processes in the form of random effects models and kernel convolution process models. We detail the specification for the exposure measure to reflect the sparsity of monitoring sites and discuss the issue of quantifying exposure over the course of a pregnancy. The models are illustrated using data from the USEPA and the North Carolina Detailed Birth Records. Statistical analyses are implemented using hierarchical modeling within a Bayesian perspective.

email: simone@stat.duke.edu

9d. PRODUCT PARTITION MODELS WITH CORRELATED PARAMETERS

Joao VD Monteiro*, University of Minnesota
Rosangela H. Loschi, Universidade Federal de Minas Gerais
Renato M. Assuncao, Universidade Federal de Minas Gerais

In time series, Bayesian partition models aim to partition the entire observation period into disjoint temporal clusters. Each cluster is an aggregation of sequential observations and a simple model is adopted within each cluster. The main inferential problem is the estimation of the number and locations of the temporal clusters. We extend the well-known product partition model (PPM) by assuming that observations within the same cluster have their distributions indexed by correlated and different parameters. Such parameters are similar within a cluster by means of a Gibbs prior distribution. We carried out several simulations and real dataset analyzes showing that our model provides better estimates for all parameters, including the number and position of the temporal clusters, even for situations favoring the PPM. A free and open source code is available.

email: monte092@umn.edu

10. POSTERS: APPLICATIONS AND CASE STUDIES I

10a. PSYCHOLOGICAL CORRELATES OF PLACEBO RESPONSE

Hamdan Azhar*, University of Michigan
Christian S. Stohler, University of Maryland
Jon-Kar Zubieta, University of Michigan

The present study documents the relationship between general trait well-being and placebo responsivity in healthy controls. Forty-eight healthy young volunteers (20 males, 28 females, mean age 25.8±4.7) with no prior history of psychiatric illness or substance abuse were recruited. Subjects completed personality questionnaires and underwent a 20-minute standardized pain challenge both in the absence and presence of a placebo with expected analgesic properties. Pain intensity was rated every 15 seconds on a scale of 0 to 100. The percent difference in mean pain rating was used as the primary outcome measure of placebo response. Bivariate ordinary least squares regression models were fit for placebo response using the most significant psychological variables as predictors. Due to multicollinearity in the predictor space, a partial least squares regression (PLS) model was fit using leave-one-out cross validation to optimize predictive power for the response. Ego-resiliency and altruism were the most powerful individual positive predictors of response. The model weights for the PLS regression reveal a consistent effect of trait well-being. The overall variance explained by the model is nearly 30%, which provides significant evidence of trait and state effects on placebo response.

email: yousufh@umich.edu

10b. ASSESSMENT OF RELIABILITY, VALIDITY AND EFFECTS OF MISSING DATA OF THE FACT-M QUESTIONNAIRE

J. Lynn Palmer*, University of Texas M.D. Anderson Cancer Center

Our previous study showed the Functional Assessment of Cancer Therapy-Melanoma Module (FACT-M), which was developed at The University of Texas M.D. Anderson Cancer Center, is a reliable and valid instrument for patients with melanoma that can be used for the assessment of quality of life in clinical trials. In the prospective assessment of the reliability, validity and sensitivity to change of the instrument, we evaluated 273 patients with stages I-IV melanoma. This presentation summarizes the results of that study. In addition, because we observed a 40% attrition rate from baseline to 3 months later, we evaluate differences in attrition rates of patients with different stages of disease to determine the degree to which the data at later times may not be missing at random. We also examine one-week changes in the FACT-M scores to determine the availability and outcome of FACT-M scores at 3, 6 and 12 months.

email: lpalmer@mdanderson.org

10c. ESTIMATING DISEASE PREVALENCE USING INVERSE BINOMIAL POOLED SCREENING

Joshua M. Tebbs*, University of South Carolina
Nicholas A. Pritchard, Coastal Carolina University

Monitoring populations for vector-borne infections is an important part of agricultural and public health risk assessment. In such applications, it is common to test pools of subjects for the presence of infection, rather than to test subjects individually. This experimental design is known as pooled (group) testing. In this paper, we revisit the problem of estimating the population prevalence from pooled testing, but we consider applications where inverse binomial sampling is used. Our work is unlike previous research in pooled testing, which has largely assumed a standard binomial sampling model. Inverse sampling is natural to implement when there is a need to report estimates early on in the data collection process and has been used in individual testing applications when disease incidence is low. We consider point and interval estimation procedures in this new pooled testing setting, and we use example data sets from the literature to describe and to illustrate our methods.

email: tebbs@stat.sc.edu

10d. GENERATING A SIMULATED KIDNEY PAIRED DONATION PROGRAM

Yan Zhou*, University of Michigan
Yijiang (John) Li, University of Michigan
Jack D. Kalbfleisch, University of Michigan
Peter X. K. Song, University of Michigan

A kidney paired donation (KPD) program provides a unique and important platform for living incompatible kidney donor/recipient pairs to exchange organs in order to achieve mutual benefit. However, evaluating different strategies and policies of organ exchanges cannot be done via conventional clinical trials. Thus, simulation models become important in addressing many issues in KPD programs. We develop a set of basic elements required to construct a simulated kidney paired donation program. We begin with a Poisson process with empirically based arrival rates to generate a KPD pool of donor-recipient pairs. An incidence matrix characterizing virtual cross-matches among the pairs in the KPD pool is then computer generated based on rules currently utilized in paired donation programs. The set of all possible exchanges satisfying certain restrictions is then obtained by utilizing a Depth-First-Search (DFS) and a Breadth-First-Search (BFS) algorithm. These approaches are compared. Aspects of implementing a full simulation model are briefly discussed. Numerical illustrations are based on several simulation examples.

email: zhouyan@umich.edu

10e. FACTORS ASSOCIATED WITH DIFFICULTY IN USING COMMUNITY BASED SERVICES AMONG CHILDREN WITH SPECIAL HEALTH CARE NEEDS

Sreelakshmi Talasila*, University of North Texas Health Science Center
Kimberly Fulda, University of North Texas Health Science Center
Sejong Bae, University of North Texas Health Science Center
Karan Singh, University of North Texas Health Science Center

Despite various advances in the present health care system, children with special health care needs (CSHCN) face difficulty in accessing required services. The purpose of the study was to identify factors associated with difficulty in using community based services, individual barriers and institutional barriers for CSHCN. Data were obtained from the National Survey of CSHCN 2005-06. The Andersen Health Behavioral Model was used to identify predisposing, enabling and need factors. Odds ratios and 95% Confidence Intervals were calculated using logistic regression analysis. Among all risk factors, four predisposing factors (education, region, race/ethnicity, total number of children living in the household), two enabling factors (type of insurance, satisfaction with the services) and two need factors (functional limitation, severity of the child) were significantly associated with difficulty in using community based services. For Institutional barriers, two enabling factors and two need factors were significant. Results of this study suggest that functional limitations and severity of the child's illness increase the odds of difficulties in using community based services. Further investigation is recommended for making better policies to overcome barriers to access the health care system.

email: vtsreelakshmi@gmail.com

10f. USE OF MULTIPLE SINGULAR VALUE DECOMPOSITIONS TO ANALYZE COMPLEX CALCIUM ION SIGNALS

Josue G. Martinez*, Texas A&M University
Jianhua Z. Huang, Texas A&M University
Robert C. Burghardt, Texas A&M University
Rola Barhoumi, Texas A&M University
Raymond J. Carroll, Texas A&M University

Novel applications of singular value decompositions (SVD), and weighted versions of them (WSVD), to data available as time series of images, are proposed. The data, or image time series, capture the calcium ion (Ca2+) expression, or signal, of myometrial cells. The SVD and WSVD are used to harness these calcium ion signals so that comparisons between two treatments can be made. The approach is semi-automatic and tuned closely to the data and their many complexities. These complexities include the following. First, all interest focuses on the behavior of individual cells across time, thus the cells need to be identified and segmented. Second, once segmented, each cell is now represented by 100+ pixels which form 100+ curves measured over time. Data compression is required to extract the features of these curves. Third, some pixels in some of the cells are subject to image saturation due to inevitable bit depth limits, and this saturation needs to be accounted for if one is to interpret the images in a reasonably unbiased manner. The use of multiple weighted and standard singular value decompositions to detect, extract and clarify the Ca2+ signals is introduced. Our SVD based signal extraction method leads to simple although finely focused statistical tests used to compare calcium ion expressions across experimental conditions.

email: jgmartinez@stat.tamu.edu

10g. PSYCHOSOCIAL STRESS AND HEALTH DISPARITIES

Brisa N. Sanchez*, University of Michigan
Trivellore E. Raghunathan, University of Michigan
Meihua Wu, University of Michigan
Ana V. Diez-Roux, University of Michigan

Recent years have been marked by rapid growth of research into race/ethnic and socioeconomic disparities in health. One process which has been hypothesized to contribute to disparities in multiple health outcomes is psychosocial stress. Salivary cortisol has been proposed as a biological measure of the stress

response. However, the understanding of the best statistical methods to characterize features of the stress response is limited. We investigate and contrast statistical methods which may be useful in characterizing different features of the biological stress response, as assessed by repeat measures of salivary cortisol collected from population-based samples. In particular, we compare linear mixed models, parametric non linear mixed models, self modeling regression, and functional mixed effects models, in terms of the information they may extract from the data and their interpretation, and the ease of implementation in population data. We also expand upon available summaries of the cortisol response, and examine whether the use of different statistical methods has contributed to the presence of apparently discordant findings for some research questions addressing the link between stress and predictors of stress. This research enhances our understanding of the most appropriate statistical methods useful to analyze measures of salivary cortisol in population studies.

email: brisa@umich.edu

10h. SELECTION OF A WORKING CORRELATION STRUCTURE IN PEDIATRIC STUDIES OF RENAL AND CROHN'S DISEASE

Matthew White*, University of Pennsylvania School of Medicine
Justine Shults, University of Pennsylvania School of Medicine
Meena Thayu, University of Pennsylvania School of Medicine
Michelle Denburg, University of Pennsylvania School of Medicine
Mary Leonard, University of Pennsylvania School of Medicine

We consider two studies in children with renal disease and Crohn's disease. We implement a generalized estimating equation (GEE) and quasi-least squares (QLS) and demonstrates that the choice of working correlation structure to describe the pattern of association amongst the repeated measurements on the children can have an impact on the results. We compare several approaches for selection of a working correlation structure for GEE and QLS that have been proposed in the literature, via simulation and comparison of the analysis results for each structure.

email: mwhi@mail.med.upenn.edu

11. POSTERS: POWER/SAMPLE SIZE

11a. BAYESIAN SAMPLE SIZE DETERMINATION FOR STUDIES DESIGNED TO EVALUATE CONTINUOUS MEDICAL TESTS

Adam J. Branscum*, University of Kentucky
Dunlei Cheng, Baylor Health Care System
James D. Stamey, Baylor University

We develop a Bayesian approach to sample size and power calculations for cross-sectional studies that are designed to evaluate and compare continuous medical tests. For studies that involve one test or two conditionally independent or dependent tests, we present methods that are applicable when the true disease status of sampled individuals will be available and when it will not. Within a hypothesis testing framework we consider the goal of demonstrating that a medical test has area under the ROC curve that exceeds a minimum acceptable level or another relevant threshold, and the goal of establishing the superiority or equivalence of one test relative to another. A Bayesian average power criterion is used to determine a sample size that will yield high posterior probability, on average, of a future study correctly deciding in favor of these goals. A nonparametric method using Dirichlet process mixtures is also presented.

email: abran3@email.uky.edu

11b. BAYESIAN SAMPLE SIZE DETERMINATION FOR TWO INDEPENDENT POISSON RATES

Austin L. Hand*, Baylor University
James Stamey, Baylor University
Dean Young, Baylor University

Because of the high cost and time constraints for clinical trials, researchers often need to determine the smallest sample size n to provide accurate inferences and decisions for a chosen parameter of interest. Bayesian sample-size determination methods not only obtain starting values and incorporate for uncertainty, but also provide a variety of measurement criteria. Thus, they are becoming increasingly more popular in clinical trials because of their flexibility and their easily interpreted inferences. In this paper we are interested in two commonly implemented Bayesian methods, the average length criterion (ALC) method proposed by Joseph et al. (1995) and the average power (AP) method proposed by Wang and Gelfand (2002), for sample-size determination for clinical trials with count data featuring rare events. In particular we examine the comparison of two Poisson rate parameters. Our procedure uses prior information from previous clinical trials to define the parameters of the conjugate gamma prior through the implementation of marginal maximum likelihood estimation. Using these parameters, we derive algorithms for the ALC and AP methods for sample size determination of the ratio of two Poisson rates.

email: Austin_Hand@baylor.edu

11c. COMPARISON OF SAMPLE SIZE REQUIREMENTS IN 3-WAY COMPARISONS FOR FIXED-DOSE COMBINATION DRUG EFFICACY STUDIES

Linlin Luo*, University of Nebraska
Julia N. Soulakova, University of Nebraska

One of the most challenging problem related to combination drug efficacy trials is to illustrate that a drug combination is

effective (relatively to placebo) and is more effective than each drug component taken alone (superiority). This goal comes from the US FDA regulations. A number of testing methods suitable to assess this goal has been recently proposed. Nonetheless, there is a lack of statistical methodology that can be used at design stage. The main goal of out presentation is to discuss novel methods to perform the power/sample size calculations. Two approaches are considered as generalizations of Sidik's approach (Pharmaceutical Statistics, 2003; 273-278). One of the proposed methods in general results in smaller minimum required sample size but may be unsuitable in some special cases and the other one tends to overestimate the sample size but does guarantee that the test will achieve desired power. This research provides essential information for investigators and is anticipated to be used at the design of experiments stage not only in combination drug cases but also with respect to other disciplines, where assessing efficacy of a combination of two or more components is of interest. For example, in agronomy several fertilizers might be combined with a goal of evaluating the possible yield improvement.

email: linlin.luo1986@gmail.com

11d. A SIMPLE METHOD FOR APPROXIMATING EFFECT ON POWER OF LINEAR MODEL MISSPECIFICATION

T. Robert Harris*, University of Texas

It is well known that model misspecification (for example dichotomizing a continuous independent variable) may reduce or, in some cases, increase statistical power in linear models. Using large-sample methods, power depends on the noncentrality parameter of the F distribution, which in turn depends on variance explained by the predictor(s) being tested and other independent variables in the model. We show how to calculate the effect of model misspecification on variance explained and power. Simulations indicate conditions under which large-sample methods yield a satisfactory approximation.

email: TRobert.Harris@UTSouthwestern.edu

12. POSTERS: NONPARAMETRIC METHODS

12a. A BAYESIAN NONPARAMETRIC GOODNESS OF FIT TEST FOR LOGISTIC REGRESSION WITH CONTINUOUS RESPONSE DATA

Angela Schörgendorfer*, University of Kentucky
Adam J. Branscum, University of Kentucky
Timothy E. Hanson, University of Minnesota

Logistic regression models are a popular tool in medical and biological data analysis. With continuous response data, it is common to create a dichotomous outcome by specifying a threshold for positivity. Fitting a linear regression via least-squares to the original, non-dichotomized response assuming a logistic error distribution has previously been shown to yield more efficient estimators of odds ratios. We develop a novel test for logistic distribution based on a Bayesian nonparametric mixture of Polya trees model. Bayes factors are calculated using the Savage-Dickey ratio for testing the null hypothesis of logistic distribution versus a nonparametric generalization. An empirical Bayes approach is computationally efficient since it does not require MCMC sampling, and we show that results from it are equivalent to results from a fully Bayesian implementation. We also develop methods for nonparametric estimation of risks, risk ratios, and odds ratios that can be used if the hypothesis of a logistic error distribution is rejected.

email: angela.sch@uky.edu

12b. A STUDY OF BAYESIAN DENSITY ESTIMATION USING BERNSTEIN POLYNOMIALS

Charlotte C. Gard*, University of Washington
Elizabeth R. Brown, University of Washington

Bernstein densities are mixtures of beta densities in which the parameters of the component densities are completely determined by the number of mixture components. Petrone (Scandinavian Journal of Statistics, 1999; The Canadian Journal of Statistics, 1999) takes a Bayesian approach to density estimation using Bernstein polynomials, placing a prior on the number of mixture components and writing the mixture weights as increments of a distribution function G. Petrone assumes a Dirichlet process prior on G. She considers the parameters of the Dirichlet process, the baseline distribution and the variability, to be fixed. We extend Petrone's model, allowing the parameters of the baseline distribution (which we assume to be of a particular parametric form) and the variability to be random. We consider estimation of subject-specific probability density functions across a population, with subjects sharing a common baseline distribution. We discuss computation for our model and present the results of simulations exploring the relationship between the number of density components, the form of the baseline distribution, and the variability around the baseline distribution. We offer recommendations regarding prior choice based on our simulation study.

email: gardc@u.washington.edu

12c. NON PARAMETRIC ANALYSIS OF MULTIDIMENSIONAL PROFILES

Margo A. Sidell*, Tulane University
Leann Myers, Tulane University

Recent studies have compared different approaches to analysis of data with large number of measures (p) relative to n, the number

of subjects (e.g., Myers, et al 2009). This type of data is found in many fields including public health, biology and anthropology. Traditional multivariate analysis methods such as MANOVA are not optimal for data with high p/n ratios so appropriate analysis of these data presents a challenge. Alternative non parametric methods for testing equality of group centroids were explored using simulations. The robustness of these methods with respect to Type I error was assessed for various p to n ratios, overall sample sizes, and correlations between measures.

email: msidell@tulane.edu

12d. A FUNCTIONAL DATA ANALYSIS METHOD FOR EVALUATION OF INHERENCE OF MEDICAL GUIDELINE

Lingsong Zhang*, Purdue University

This study presents a functional data analysis method to evaluate whether inherence of medical guideline is associated with healthcare outcome. The study uses a 2-year longitudinal physician data for diabetic patients. We found that for those patients with fewer visits than the medical guidelines, their hemoglobin A1C level on average will have an increasing trend. And those patients following guidelines have a non-decreasing trend for A1C in these 2 years.

email: lingsong@purdue.edu

12e. ROBUST LOWER-DIMENSIONAL APPROXIMATION FOR SPARSE FUNCTIONAL DATA WITH ITS APPLICATION TO SCREENING YOUNG CHILDREN'S GROWTH PATHS

Wei Ying, Columbia University
Wenfei Zhang*, Columbia University

Growth charts are commonly used for screening children's growth. Current methods considered one measurement at a specific time. More informative screening can be achieved by studying the entire growth paths. We proposed the statistical methods to screening the growth paths based on finding the lower-dimensional approximation of the growth curves (sparse functional data). The methods are based on robust alternating regression, using B-splines to represent the growth curves. The growth curves can be ranked by the joint distribution of the projection scores obtained from our approximations. Additionally, we apply these methods to a real growth data and obtain some results on screening the growth paths.

email: wz2157@columbia.edu

12f. A COPULA APPROACH FOR ESTIMATING CORRELATION OF SHARED COUPLE BEHAVIORS

Seunghee Baek*, University of Pennsylvania
Scarlett L. Bellamy, University of Pennsylvania
Andrea B. Troxel, University of Pennsylvania
Thomas R. Ten Have, University of Pennsylvania
John B. Jemmott III, University of Pennsylvania

Copula-based approaches are becoming popular in multivariate modeling settings in various fields where multivariate dependency is of primary interest. This approach is flexible in measuring the effect of covariates on dependence and for estimating marginal probabilities for multiple outcomes simultaneously. We will apply the copula modeling approach to a study collecting self-reported data on shared sexual behaviors from couples (e.g., independently from male and female partners). We will estimate the reliability of couple reports using copulas, adjusting for key couple-level baseline covariates. We will do so by estimating measures of dependence using mixtures of max-infinitely divisible copulas, introduced by Joe and Hu. We focus on estimating the odds ratios and binary correlations, which can be easily derived from copula functions and marginal probabilities, and explore the influence of covariate information on the dependency. We apply these methods to data from the Multisite HIV/STD Intervention Trial for African American Couples (AAC) Study.

email: seunghee@mail.med.upenn.edu

12g. A COMPARATIVE STUDY OF NONPARAMETRIC ESTIMATION IN WEIBULL REGRESSION: A PENALIZED LIKELIHOOD APPROACH

Young-Ju Kim*, Kangwon National University

The Weibull distribution is popularly used to model lifetime distributions in many areas of applied statistics. This paper employs a penalized likelihood method to estimate the shape parameter and an unknown regression function simultaneously in a nonparametric Weibull regression. Four methods were considered: two cross-validation methods, a corrected Akaike information criterion, and a Bayesian information criterion. Each method was evaluated based on shape parameter estimation as well as selecting the smoothing parameter in a penalized likelihood model through a simulation study. Adapting a lower-dimensional approximation in the penalized likelihood, the comparative performances of methods using both censored and uncensored data were examined for various censoring rates. A real data example is presented to illustrate the application of the suggested methods.

email: ykim7stat@kangwon.ac.kr

12h. MULTIVARIATE SHAPE RESTRICTION REGRESSION WITH BERNSTEIN POLYNOMIAL

Jiangdian Wang*, North Carolina State University
Sujit Ghosh, North Carolina State University

There has been increasing interest in estimating a multivariate shape-restricted function, including monotonicity, convexity and concavity. The estimation is more challenging for multivariate predictors, especially for functions with compact support. Existing estimation methods for shape restricted regression functions are either computationally very intensive or have serious boundary biases. This article considers an application of multivariate Bernstein polynomials and proposes a sieved estimator drawn from a nested sequence of shape-restricted multivariate Bernstein polynomials. Three key features of the proposed method are: (1) the regression function estimate is shown to be the solution of a quadratic programming problem; making it computationally attractive (2) the estimate is shown to reduce the boundary bias; and (3) the estimation methodology is flexible in the sense that it can be easily adapted to accommodate many popular multivariate shape restrictions. Numerical results derived from simulated data sets are used to illustrate the superior performance of the proposed estimator compared to some of the existing estimators in terms of various goodness of fit metrics.

email: jwang8@ncsu.edu

13. POSTERS: MISSING DATA AND MEASUREMENT ERROR

13a. COMBINING DISPARATE MEASURES OF METABOLIC RATE DURING SIMULATED SPACEWALKS

Robert J. Ploutz-Snyder*, NASA Johnson Space Center
Alan H. Feiveson, NASA Johnson Space Center
Dan Nguyen, NASA Johnson Space Center
Lawrence Kuznetz, NASA Johnson Space Center

NASA is designing space suits for future missions in which astronauts will perform Extra Vehicular Activities (EVA) on Lunar or Martian surfaces. During EVAs, astronauts' integrated metabolic rates (MR) are used to predict how much longer activities can continue within acceptable safe margins. For EVAs in the Apollo era, physicians monitored heart rate, O2 consumption, and liquid-cooled garment (LCG) temperatures in real time, which were subjectively combined to estimate MR. But these data can be in conflict, making estimation of MR difficult. Current plans use a largely heuristic methodology for incorporating these measurements plus CO2 production, ignoring data that appears in conflict; however a more rigorous model-based approach is desirable. Here, we show how principal-axis factor analysis, in combination with OLS regression and LOWESS smoothing, can estimate metabolic rate as a weighted average of all four inputs with less sensitivity to data spikes and good within-subject reproducibility. These methods do not require physician monitoring and can be automated. Our models show promise for increasing safety and reducing errors from human integration of EVA data.

email: robert.ploutz-snyder-1@nasa.gov

13b. A BAYESIAN APPROACH TO MULTILEVEL POISSON REGRESSION WITH MISCLASSIFICATION

Monica M. Bennett*, Baylor University
John W. Seaman, Jr., Baylor University
James D. Stamey, Baylor University

In regression, coefficients that vary by group are often accounted for by using indicator variables. An alternate approach is to use multilevel models which give the varying coefficients a probability model. The model on the coefficients is considered the second-level model which has its own regression coefficients. The parameters for the second-level model can themselves be given a probability model, and so on. This model differs from the classical regression approach in that we are modeling the variation between groups. In this presentation, we consider the multilevel Poisson regression model. A common concern with Poisson regression is misclassification of the response variable. Thus we develop a Bayesian approach to the multilevel Poisson regression model that accounts for misclassified data.

email: monica_bennett@baylor.edu

13c. MISRECORDING IN THE NEGATIVE BINOMIAL REGRESSION MODEL

Mavis Pararai*, Indiana University of Pennsylvania

When count data is modeled, the assumption is that the response has been correctly reported. Sometimes there is over and underreporting in surveys due to the sensitive nature of the questions. A lot of underreporting goes with domestic violence data. The negative binomial regression model for underreported counts is used to illustrate such data. Some comparisons are made to the Poisson regression model for underreported counts. Underreporting should always be checked when dealing with count data.

email: pararaim@iup.edu

13d. A NON-PARAMETRIC METHOD FOR ESTIMATING A HEAPING MECHANISM FROM PRECISE AND HEAPED SELF-REPORT DATA

Sandra D. Griffith*, University of Pennsylvania
Saul Shiffman, University of Pittsburgh
Daniel F. Heitjan, University of Pennsylvania

One form of measurement error in self-report data is heaping, or an excess of values at round numbers. Daily cigarette counts, for example, commonly exhibit heaps at multiples of 20, and to a lesser extent, 2, 5, and 10, when measured by retrospective recall methods. Therefore, conclusions drawn from data subject to heaping are suspect. If we knew the mechanism behind heaping, we could account for the error. Methods for instantaneously recording self-report data could help us understand the heaping mechanism. A dataset with daily cigarette counts measured by both a retrospective recall method, timeline follow back (TLFB), and an instantaneous method, ecological momentary assessment (EMA), motivates our method. We have developed a non-parametric method to estimate the conditional distribution of the recall measurement (TLFB) given the instantaneous and presumably more precise measurement (EMA).

email: sgrif@upenn.edu

13e. MULTIPLE IMPUTATION IN GROUP RANDOMIZED TRIALS: THE IMPACT OF MISSPECIFYING CLUSTERING IN THE IMPUTATION MODEL

Rebecca R. Andridge*, The Ohio State University

In group randomized trials (GRTs), identifiable groups rather than individuals are randomized to study conditions. Resulting data consist of a small number of groups with correlated observations within a group. Usually the resulting intracluster correlation (ICC) is small, but can lead to large variance inflation factors and cannot be ignored. Missing data is a common problem with GRTs. I discuss strategies for accounting for clustering in multiply imputing a missing continuous outcome, focusing on a simple post-test only study design. Analysis with an adjusted two-sample t-test requires imputation with a mixed effects model in order for the analysis to be congenial; however this imputation procedure is not yet available in standard statistical software. An alternative approach readily available is to include fixed effects for cluster, but the impact of this misspecification has not been studied. I show that under this imputation model the MI variance estimator is biased and, somewhat counterintuitively, smaller ICCs lead to larger overestimation of the MI variance. Analytical expressions for the bias are derived for missingness completely at random (MCAR), and the case of missing at random (MAR) is illustrated through simulation. Methods are applied to data from a school-based GRT and differences in inference compared.

email: randridge@cph.osu.edu

13f. ESTIMATION IN HIERARCHICAL MODELS WITH INCOMPLETE BINARY RESPONSE AND BINARY COVARIATES

Yong Zhang*, University of Michigan
Trivellore Raghunathan, University of Michigan

Hierarchical models are often used when data are observed at different levels and the interaction effects on the outcomes between variables measured at different levels are of interest. Missing data can complicate analysis using hierarchical models and can occur at all levels, in both outcomes and covariates. Ignoring the subjects with missing data usually leads to biased estimates, yet less attention has been paid to the analysis based on hierarchical models with incomplete data. We use a combination of the EM algorithm and multiple imputations to develop approximate maximum likelihood estimates of the parameters in hierarchical models, assuming missing at random (MAR) and ignorable missing mechanism (Rubin, 1976; Little and Rubin, 2002). In this paper we consider a binary response with missing values as well as continuous and binary covariates with missing values at each level. Simulation study is used to demonstrate that our proposed method has desirable repeated sampling properties. The method is also applied to a survey data.

email: yonzhang@umich.edu

13g. BAYESIAN MULTILEVEL REGRESSION MODELS WITH SPATIAL VARIABILITY AND ERRORS IN COVARIATES

Theodore J. Thompson*, Centers for Disease Control and Prevention
James P. Boyle, Centers for Disease Control and Prevention

Estimates of diabetes and obesity prevalence are available for all 3141 counties in the United States. The (estimated) variances of these prevalence estimates are also available. We are interested in the relationship between diabetes and obesity and, possibly, other covariates. We describe a Bayesian multilevel regression model for these data. This model include convolution priors, i.e. spatially correlated error and unstructured error, in both the disease (diabetes) submodel and the covariate (obesity) submodel. Results show that a one percentage point increase in obesity prevalence is associated with a 0.28 percentage point increase in diabetes prevalence. We also show that, for these data, results are severely biased if one ignores the spatial correlation. Models were fit using the freely available WinBUGS software.

email: tat5@cdc.gov

14. POSTERS: STATISTICAL METHODS

14a. BAYESIAN INFERENCE FOR CENSORED BINOMIAL SAMPLING

Jessica Pruszynski*, Baylor University
John W. Seaman, Jr., Baylor University

Censored binomial data may lead to irregular likelihood functions and problems with statistical inference. We consider a Bayesian approach to inference for censored binomial problems and compare

it to non-Bayesian methods. We include examples and a simulation study in which we compare point estimation, interval coverage, and interval width for Bayesian and non-Bayesian methods.

email: jessica_pruszynski@baylor.edu

14b. A COMPARISON OF MODEL-BASED VERSUS MOMENT-BASED ESTIMATOR OF COMPLIER AVERAGE CAUSAL EFFECT (CACE)

Nanhua Zhang*, University of Michigan
Roderick J. A. Little, University of Michigan

We consider estimating the causal effect in randomized clinical trials with noncompliance. Under certain assumptions, the instrumental variable estimator is a valid estimator of the average causal effect for a subgroup of units, the compliers, thus we call this causal effect complier average causal effect (CACE). This estimator is a direct estimate of the causal effect and is protected from selection bias by randomization; however, it has large variance especially when the compliance rate is low. Model-based version of the IV estimator can be used to increase the efficiency of IV estimator. Simulation shows that when correctly specified, the model-based estimator can yield estimate with better coverage rate, less bias, and smaller root mean square error (RMSE). The effect of principal compliance rate and sample size on the relative efficiency of model-based versus moment-based IV estimator is also considered. These estimators are applied to a behavioral intervention trial called Making Effective Nutritional Choices for Cancer Prevention: the MENU study.

email: nhzhang@umich.edu

14c. ASSESSING THE EFFECT OF A TREATMENT WITH A CLUMP OF OBSERVATIONS AT ZERO

Jing Cheng*, University of Florida
Dylan Small, University of Pennsylvania

There are many studies with a clump of observations at zero. Several methods have been proposed for testing the treatment effect for these types of studies. In this work, in addition to testing the treatment effect, we are interested in understanding how the treatment works and measuring the magnitude of the treatment's effect. We develop an empirical likelihood based approach for this problem and demonstrate its advantages over existing approaches.

email: jcheng@biostat.ufl.edu

14d. NONCONVERGENCE IN LOGISTIC AND POISSON MODELS FOR NEURAL SPIKING

Mengyuan Zhao*, University of Pittsburgh
Satish Iyengar, University of Pittsburgh

Generalized linear models are an increasingly common approach for spike train data analysis. For the logistic and Poisson models, one possible difficulty is that iterative algorithms for computing parameter estimates may not converge because of certain data configurations. For the logistic model, these configurations are called complete and quasi-complete separation. We show that these features are likely to occur because of refractory periods of neurons. We use an example to study how standard software deals with this difficulty. For the Poisson model, we show that the same difficulties arise, this time possibly due to bursting or to specifics of the binning. We characterize the nonconvergent configurations for both models, show that they can be detected by linear programming methods, and discuss possible remedies.

email: mez25@pitt.edu

14e. APPROXIMATE INFERENCES FOR NONLINEAR MIXED-EFFECTS MODELS WITH SKEW-NORMAL INDEPENDENT DISTRIBUTIONS

Victor H. Lachos Davila*, Campinas State University, Brazil
Dipak K. Dey, University of Connecticut

Nonlinear mixed-effects models have received a great deal of attention in the statistical literature in recent years. A standard assumption in nonlinear mixed-effects models for continuous responses is the normal distribution for the random effects and the within-subject errors, making it sensitive to outliers. We present a novel class of asymmetric nonlinear mixed-effects models that provides for an efficient estimation of the parameters in the analysis of longitudinal data. We assume that, marginally, the random effects follow a multivariate skew--normal/independent distribution and that the random errors follow a symmetric normal/independent distribution providing an appealing robust alternative to the usual normal distribution in nonlinear mixed-effects models. We propose an approximate likelihood analysis for maximum likelihood estimation based on the EM algorithm that produce accurate maximum likelihood estimates and significantly reduces the numerical difficulty associated with the exact maximum likelihood estimation. Simulation studies indicate that our proposed methods work well for small, medium and large variability of the random effects. The methodology is illustrated through an application to theophylline kinetics data.

email: hlachos@ime.unicamp.br

15. POSTERS: SURVEY RESEARCH AND CATEGORICAL DATA ANALYSIS

15a. THE INTRACLUSTER CORRELATION AS A FUNCTION OF INHERENT AND DESIGN INDUCED COVARIANCES

Robert E. Johnson*, Virginia Commonwealth University
Tina D. Cunningham, Virginia Commonwealth University

The intracluster correlation (ICC) plays a significant role in the planning and analysis of cluster randomized studies. Clusters may represent primary care clinics in which patients are clustered. Clusters may also represent subjects on which repeated measures are obtained. Two measures sampled from the same randomly chosen cluster will be correlated. Typically we assume the within cluster sample, conditioned on the cluster, is independent. Thus the conditional covariance between two measures from the same cluster is zero. Here the ICC is a value induced by the two stage sample. Further, the ICC is the same for all pairs of measures taken from the same sample, leading to an exchangeable or central composite correlation structure. The magnitude of the ICC increases as the spread between the cluster means widens. However the assumption of a zero conditional correlation between pairs of measures within a cluster may not hold. For example, in repeated measures on a single patient one might assume an auto-regressive correlation structure. This correlation is inherent to the measures taken within a given cluster. In the presence of this inherent correlation, the ICC is more complex. In this presentation we will explore the general form of the ICC as a function of both the induced and inherent correlation and discuss implications on design and analysis.

email: rjohnson@vcu.edu

15b. IDENTIFIABILITY OF A RESTRICTED FINITE MIXTURE MODEL FOR FEW BINARY RESPONSES ALLOWING COVARIATE DEPENDENCE IN MIXING DISTRIBUTION

Yi Huang*, University of Maryland Baltimore County

For latent class models with very few binary outcomes, the identifiability of finite mixture models is often in question. Typical latent class models (J > 1) with only one binary response are known to be not identifiable (Goodman 1974). However, Dayton and Macready (1988) indicated: "in certain situations, concomitant-variable models can be fitted with only a single dichotomous variable." Although this was written 18 years ago, the supporting examples are rare to find. We proved the identifiability of our proposed restricted finite mixture models with one or two binary outcomes, and showed that the identifiability properties of those latent class models can be improved by incorporating covariate information, especially when the subclassification scheme is based on an underlying continuous latent variable.

email: yihuang@umbc.edu

15c. ESTIMATING DISEASE PREVALENCE WHEN TESTING CONSENT RATES ARE LOW: A POOLED TESTING APPROACH

Lauren Hund*, Harvard University
Marcello Pagano, Harvard University

When estimating the prevalence of a disease, blood samples are frequently combined into pools of size k and only the pool is tested. We propose a prevalence estimator which allows pool size to vary and choice of pool size for each subject is random. We plan to apply this method (with pool sizes 1 and k) to estimate HIV prevalence in a South African population-based survey with low testing consent rates. We hypothesize that those who will not test as individuals might be willing to provide a blood samples for pooled testing, since their individual test results are completely unknown. Asymptotic consistency and normality of the estimator are demonstrated when prevalence is bounded away from zero. A simulation study assessing bias, standard error, and confidence interval coverage is performed to determine relevant thresholds for maximum pool size choice in a relatively high prevalence setting; a jackknife correction is proposed to correct for upwards bias of the prevalence estimator for small sample sizes. Since some individuals will likely consent to neither individual nor pooled testing, we suggest corrections to the prevalence estimator if data is missing at random and if data is not missing at random.

email: lbhund@gmail.com

15d. TO WEIGHT OR NOT TO WEIGHT: A SURVEY SAMPLING SIMULATION STUDY

Marnie Bertolet*, University of Pittsburgh

Two fundamental approaches (design- and model-based) exist to incorporate sampling complexities into survey analyses. Design-based analysts use the sampling design as the sole source of variability. Model-based analysts incorporate the sampling design into the model that generated the data. A controversy exists regarding the use of inverse probability sampling weights in model-based analyses; design-based analysts believe the weights compensate for model misspecification and informative sampling while model-based analysts believe the weights cloud interpretation and inflate variances. Do sampling weights add value to a model that incorporates sampling complexities? A set of simulations was performed on linear mixed-effects models, whose methods for weighting have recently been published. The simulations varied the generating model, the estimating model, the informativeness of the sampling design and the sampling hierarchy. These simulations demonstrate that sampling weights reduce bias from model misspecification only when the misspecification induces informative sampling (e.g. sampling proportional to x1 and omitting x1 from the model). Bias related to a misspecified model that does not relate to the sampling design are unaffected by the weights (e.g. omitting x2 when sampling proportional to x1).

Finally, sampling weights can reduce, but not necessarily eliminate, bias induced by informative sampling.

email: bertoletm@edc.pitt.edu

16. POSTERS: BIOLOGICS, PHARMACEUTICALS AND MEDICAL DEVICES

16a. HIERARCHICAL MODELING TO ASSESS PRECISION AND TREATMENT EFFECTS IN AN INTERLABORATORY STUDY

Jason Schroeder*, Center for Devices and Radiological Health, U.S. Food and Drug Administration

In an effort to develop a standard test method for determining the compatibility of personal lubricants with latex condoms, an interlaboratory study (ILS) was conducted. During the study, a common lab test protocol was followed at nine laboratories and tensile and airburst properties of several condom-lubricant combinations were recorded. The objectives of the study included (i) determining the repeatability and reproducibility of the testing procedures, and (ii) assessing the effects of the lubricants on the tensile and airburst properties of the condoms. A hierarchical model is used to estimate the parameters of interest and to assess differences between labs. The results and inferences from this ILS will be discussed in the context of the proposed standard.

email: jason.schroeder@fda.hhs.gov

16b. A STATISTICAL TEST FOR EVALUATION OF BIOSIMILARITY IN VARIABILITY OF BIOLOGIC PRODUCTS

Tsung-Cheng Hsieh, National Taiwan University, Taiwan
Shein-Chung Chow, Duke University School of Medicine
Jen-Pei Liu, National Taiwan University, Taiwan
Chin-Fu Hsiao, National Health Research Institutes, Taiwan
Eric M. Chi*, Amgen Inc.

As more biologic products are going off patent protection, the development of follow-on biologic products has received much attention from both biotechnology industry and the regulatory agencies. Unlike small molecule drug products, the development of biologic products is very sensitive to the manufacturing process and environment. Thus, Chow and Liu (2009) suggested that the assessment of biosimilarity between biologic products be focused on variability rather than only average biosimilarity. In addition, it was also suggested that a probability-based criterion, which is more sensitive to variability, should be employed. In this article, we propose a probability-based asymptotic statistical testing procedure to evaluate biosimilarity in variability of two biologic products. A numerical study was conducted to investigate the relationship between the probability-based criterion in variability and various study parameters. Simulation studies were also conducted to empirically investigate the performance of the proposed probability-based asymptotic statistical testing procedure in term of sample size and power. A numerical example was provided to illustrate the proposed methods.

email: chi@amgen.com

16c. BAYESIAN HIERARCHICAL MONOTONE REGRESSION SPLINES FOR DOSE-RESPONSE ASSESSMENT AND DRUG-DRUG INTERACTION ANALYSIS

Violeta G. Hennessey*, University of Texas M.D. Anderson Cancer Center
Veerabhadran Baladandayuthapani, University of Texas M.D. Anderson Cancer Center
Gary L. Rosner, University of Texas M.D. Anderson Cancer Center

We provide a practical and flexible method for dose-response modeling and drug-drug interaction analysis. We developed a semi-parametric Bayesian hierarchical model that employs monotone regression I-splines for estimating the mean dose-response function. We use Markov chain Monte Carlo (MCMC) to fit the model to the data and carry out posterior inference on quantities of interest (e.g., inhibitory concentrations, Loewe Interaction Index). Our approach accounts for sources of variation inherent in the data, uncertainty in parameter values, and a monotone relationship between dose and response. We compare our approach to analysis using a parametric mean dose-response function.

email: v.g.hennessey@gmail.com

17. STATISTICAL METHODS FOR ESTIMATING THE PUBLIC HEALTH IMPACT OF POLICY INTERVENTIONS

ANALYSIS OF LONGITUDINAL DATA TO EVALUATE A POLICY CHANGE

Benjamin French*, University of Pennsylvania
Patrick J. Heagerty, University of Washington

Longitudinal data analysis methods are powerful tools for exploring scientific questions regarding change and are well suited to evaluate the impact of a policy intervention. However, there are challenging aspects of policy change data with respect to analysis and inference that require consideration: defining comparison groups, accounting for heterogeneity in the policy effect, and modeling longitudinal correlation. We compare currently available longitudinal data analysis methods to evaluate a policy change. We also illustrate issues specific to evaluating a policy change via a case study of laws eliminating gun-use restrictions and firearm-related homicide. We obtain homicide rate ratios estimating the effect of enacting a shall-issue law that vary between 0.903 and 1.101. However, using methods that are most appropriate implies that enacting such a law is associated with a non-significant increase in fire-arm-related

homicide. We conclude that in a policy change study it is essential to thoroughly model temporal trends and account for policy effect heterogeneity.

email: bcfrench@upenn.edu

A STATISTICAL APPROACH FOR ASSESSING THE PUBLIC HEALTH IMPACT OF SMOKING BANS

Christopher D. Barr*, Harvard University

Evaluating the effects of smoking bans on mortality and morbidity is an important public health question. To date, the few relevant studies have used small populations and provided inconclusive results. We have assembled monthly time series data on mortality and morbidity for all US counties implementing a smoking ban during the period 1987 - 2006. We plan to develop Bayesian hierarchical models for synthesizing the strength of evidence on the association between smoking bans and adverse health outcomes.

email: cdbarr@gmail.com

ESTIMATING LONGITUDINAL EFFECTS USING PROPENSITY SCORES AS REGRESSORS

Aristide C. Achy-Brou, JP Morgan
Constantine E. Frangakis*, Johns Hopkins University
Michael Griswold, University of Mississippi

We derive regression estimators that can compare longitudinal treatments using only the longitudinal propensity scores as regressors. These estimators, which assume knowledge of the variables used in the treatment assignment, are important for reducing the large dimension of covariates for two reasons. First, if the regression models on the longitudinal propensity scores are correct, then our estimators share advantages of correctly specified model-based estimators, a benefit not shared by estimators based on weights alone. Second, if the models are incorrect, the misspecification can be more easily limited through model checking than with models based on the full covariates. Thus, our estimators can also be better when used in place of the regression on the full covariates. We use our methods to compare longitudinal treatments for type I diabetes mellitus.

email: cfrangak@jhsph.edu

18. ADVANCES IN TIME SERIES ANALYSIS OF NEUROLOGICAL STUDIES

NONPARAMETRIC SPECTRAL ANALYSIS WITH APPLICATIONS TO SEIZURE CHARACTERIZATION USING EEG TIME SERIES

Li Qin*, Fred Hutchinson Cancer Research Center
Yuedong Wang, University of California-Santa Barbara

Understanding the seizure initiation process and its propagation pattern(s) is a critical task in epilepsy research. In this article, we analyze epileptic EEG time series using nonparametric spectral estimation methods to extract information on seizure-specific power and characteristic frequency (or frequency band(s)). Because the EEGs may become non-stationary before seizure events, we develop methods for both stationary and local stationary processes. Based on penalized Whittle likelihood, we propose a direct generalized maximum likelihood (GML) and generalized approximate cross-validation (GACV) methods to estimate smoothing parameters in both smoothing spline spectrum estimation of a stationary time series and smoothing spline ANOVA time-varying spectrum estimation of a locally stationary process.

email: lqin@scharp.org

CLASSIFICATION OF FAMILIES OF LOCALLY STATIONARY TIME SERIES

Robert T. Krafty*, University of Pittsburgh
Wensheng Guo, University of Pennsylvania

Existing methods in non-stationary time series classification assume time series from different units within a population are generated by the same underlying stochastic process characterized by a time-varying second order spectrum, and both the between-time-series variability and the within-time-series variability are results of the same underlying stochastic process. This is usually not true in real applications and can lead to misclassification. In this talk, we propose a model for a family of time series by imposing a hierarchical structure on their log-spectra. This model assumes that while a family of time series share some similarity characterized by the population-average spectrum, each time series has its own characteristics modeled by the unit-specific deviation in terms of its log-spectrum. We then propose nonparametric methods to estimate the population-average log-spectrum and the between-unit variance function. We develop a quadratic rule for discriminating between different populations based on the estimated mean log-spectra and the variance functions. A simulation study is presented to empirically demonstrate the benefits of accounting for the between-time-series variability and the proposed procedure is used to discriminate pre-seizure EEG time series from non-seizure baseline data.

email: krafty@pitt.edu

EVOLUTIONARY FACTOR ANALYSIS OF EEG DATA

Giovanni Motta, University of Maastricht
Hernando Ombao*, Brown University

Our goal is to characterize and estimate the dynamic structure of multi-channel electroencephalograms in a motor-visual task experiment. Preliminary analyses of our data indicate that both the variance of each channel and cross-covariance between a pair of channels evolve over time and that the cross-covariance profiles display a structure that is common across all pairs of channels. These observations suggest that the methods of evolutionary factor analysis which is a statistical tool recently developed to study multivariate non-stationary stochastic processes that are driven by common factors. EFA provides a new class of factor models with time-varying factor loadings. The factors will be modeled as stationary processes while the loadings are allowed to vary over time. The estimation of these non-stationary factor models makes use of the generalization of the properties of the principal components techniques to the time-varying framework. In our model, the factors share common features across several trials. We use result from EFA asymptotic theory to establish conditions for identification, estimation of the loadings, factors and common components using all trials. In our analysis, Common co-movements of EEG signals will be explained by latent factors that are primarily responsible for processing the visual-motor task.

email: ombao@stat.brown.edu

STIMULUS-LOCKED VAR MODELS FOR EVENT-RELATED fMRI

Wesley K. Thompson*, University of California-San Diego

We propose a model tailored for exploring effective connectivity of multiple brain regions in event-related fMRI designs – a semi-parametric adaptation of vector autoregressive (VAR) models, termed 'stimulus-locked VAR' (SloVAR). Connectivity coefficients vary as a function of time relative to stimulus onset, are regularized via basis expansions, and vary randomly across subjects. We demonstrate the SloVAR model on a sample of clinically depressed and normal controls, showing that early but not late cortico-amygdala connectivity appears crucial to emotional control and early but not late cortico-cortico connectivity predicts depression severity in the depressed group, relationships that would have been missed in a more traditional VAR analysis. SloVAR obtains flexible, data-driven estimates of effective connectivity and hence is useful for building connectivity models when prior information on dynamic regional relationships is sparse. Indices derived from the coefficient estimates can also be used to relate effective connectivity estimates to behavioral or clinical measures.

email: wktwktwkt@gmail.com

19. INFERENCE AND PREDICTION FOR SPARSE NETWORKS

VARIATIONAL EM ALGORITHMS FOR A CLASS OF NETWORK MIXTURE MODELS

Duy Vu, Penn State University
David R. Hunter*, Penn State University

We discuss a mixture of exponential-family models for networks whose edges take discrete values. The mixture structure assumes that each node is comes from one of several categories, but this category membership is unobserved. This model extends the network mixture models of Nowicki and Snijders (2001) and Daudin, Picard, and Robin (2008). Unlike the former, it is scalable to large networks due to the use of a variational EM algorithm; unlike the latter, it includes a dyadic independence, rather than an edge independence, assumption and therefore we are able to model a reciprocity effect. We discuss the application of these methods to network datasets with more than 100,000 nodes.

email: dhunter@stat.psu.edu

SPARSE REGRESSION MODELS FOR CONSTRUCTING GENETIC REGULATORY NETWORKS

Jie Peng*, University of California-Davis
Pei Wang, Fred Hutchinson Cancer Research Center

In this talk, we discuss several regression models utilizing sparse constraints. These models are motivated by reconstruction of genetic regulatory networks using high throughput genomics data. We focus our discussion on the use of multiple types of genomic data and the choice of sparse constraints which are suitable for the network structure that we envision. We also discuss the inference of the directions of interactions by utilizing multiple types of genomic data. Related issues such as computation and model tuning will also be discussed. We illustrate the performance of the methods through simulation studies and real applications.

email: jie@wald.ucdavis.edu

TIME VARYING NETWORKS: REVERSE ENGINEERING AND ANALYZING REWIRING SOCIAL AND GENETIC INTERACTIONS

Eric P. Xing, Carnegie Mellon University

A plausible representation of the relational information among entities in dynamic systems such as a social community or a living cell is a stochastic network that is topologically rewiring and semantically evolving over time. While there is a rich literature in modeling static or temporally invariant networks, until recently,

little has been done toward modeling the dynamic processes underlying rewiring networks, and on recovering such networks when they are not observable. In this talk, I will present a new formalism for modeling network evolution over time based on temporal exponential random graphs, and several new algorithms for estimating the structure of time evolving probabilistic graphical models underlying nonstationary time-series of nodal attributes. I will show some promising results on recovering the latent sequence of evolving social networks in the US Senate based it voting history, and the gene networks over more than 4000 genes during the life cycle of Drosophila melanogaster from microarray time course, at a time resolution only limited by sample frequency. I will also sketch some theoretical results on the asymptotic sparsistency of the proposed methods, which differ significantly from traditional sparsistency analysis of static structure estimation based on iid samples because of the temporal relatedness of samples.

email: epxing@cs.cmu.edu

PENALIZED REGRESSION WITH NETWORKED PREDICTORS AND ITS APPLICATION TO eQTL ANALYSIS

Wei Pan*, University of Minnesota
Benhua Xie, Takeda Global Research and Development
Xiaotong Shen, University of Minnesota

We consider the problem of conducting penalized regression analysis with predictors whose relationships are described a priori by a network. A class of motivating examples is to model a quantitative or categorical phenotype using gene expression profiles while accounting for coordinated functioning of genes in the form of biological pathways or networks. We introduce our new method and compare with some existing ones. We will discuss an application of the new method to expression quantitative trait loci (eQTL) analysis.

email: panxx014@umn.edu

20. RECENT DEVELOPMENTS IN HIGH DIMENSIONAL INFERENCE

OPTIMAL SCREENING FOR SPARSE SIGNALS

Tony Cai, University of Pennsylvania
Wenguang Sun*, North Carolina State University

In large scale statistical inference problems,it is common that signals are sparse and it is desirable to significantly reduce the original large data set to a much smaller subset for further study. In this talk, we consider two related data screening problems: One is to find the smallest subset such that it contains all signals with high probability and another is to find the largest subset so that it virtually contains only signals (i.e. the proportion of the nulls in the subset is negligible.) These screening problems are closely connected to but distinct from the more conventional detection or multiple testing problems. We shall discuss precise conditions under which these goals are achievable and construct screening procedures that have near optimality properties.

email: tcai@wharton.upenn.edu

REVISITING MARGINAL REGRESSION

Jiashun Jin*, Carnegie Mellon University
Christopher Genovese, Carnegie Mellon University
Larry Wasserman, Carnegie Mellon University

The lasso has become an important practical tool for high dimensional regression as well as the object of intense theoretical investigation. But despite the availability of efficient algorithms, the lasso remains computationally demanding in regression problems where the number of variables vastly exceeds the number of data points. A much older method, marginal regression, largely displaced by the lasso, offers a promising alternative in this case. Computation for marginal regression is practical even when the dimension is very high. In this paper, we compare the conditions for exact reconstruction of the two procedures, find examples where each procedure succeeds while the other fails, and characterize the advantages and disadvantages of each. Also, we derive and compare conditions under which the marginal regression will provide exact reconstruction with high probability. Last, we derive rates of convergence for the procedures and offer a new partitioning of the "phase diagram," that shows when exact or Haming reconstruction is effective.

email: jiashun@stat.cmu.edu

THEORETICAL SUPPORT FOR HIGH DIMENSIONAL DATA ANALYSIS BASED ON STUDENT'S t STATISTIC

Aurore Delaigle, University of Melbourne
Peter Hall*, University of Melbourne
Jiashun Jin, Carnegie Mellon University

Student's t statistic is finding applications today that were never envisaged when it was introduced more than a century ago. Many of these rely on properties, for example robustness against heavy tailed sampling distributions, that were not explicitly considered until relatively recently. In this talk we explore these features of the t statistic in the context of its application to high dimensional problems, including feature selection and ranking, highly multiple hypothesis testing, and sparse, high dimensional signal detection. Robustness properties of the t-ratio are highlighted, and it is established that those properties are preserved under applications of the bootstrap. In particular, bootstrap methods correct for skewness, and therefore lead to second-order accuracy, even in the extreme tails. Indeed, it is argued that the bootstrap, and also the more popular but less effective t-distribution and normal approximations, are more effective in the tails than towards the middle of the distribution. This leads to methods, for example

bootstrap-based techniques for signal detection, that confine attention to the significant tail of a statistic.

email: halpstat@ms.unimelb.edu.au

RISK PREDICTIONS FROM GENOME WIDE ASSOCIATION DATA

Hongyu Zhao*, Yale University
Jia Kang, Yale University
Ruiyan Luo, Yale University
Judy Cho, Yale University

Much progress has been made in recent years on using the genome-wide association study (GWAS) paradigm to identify genetic variants affecting individual's susceptibility to complex diseases. These studies generally involve hundreds to thousands of subjects and hundreds of thousands of genetic variants (potential risk predictors). To date, more than 1,000 replicated associations have been made on dozens of disorders and traits. To translate these exciting findings into clinical practice to benefit the general population, it is critical to mine the rich data that have been collected to develop risk models that relate each individual's genomic makeup with his/her disease risk. However, existing statistical approaches that have been developed in the context of a limited number of risk predictors are not well suited for the very high dimensional data from GWAS, where thousands or more genetic variants may provide information on disease risk. In this presentation, we will discuss statistical methods that have been developed in this context for disease risk predictions and demonstrate their usefulness through their applications to real GWAS data.

email: hongyu.zhao@yale.edu

21. IMAGE ANALYSIS

PREDICTING TREATMENT EFFICACY VIA QUANTITATIVE MRI: A BAYESIAN JOINT MODEL

Jincao Wu*, University of Michigan
Timothy D. Johnson, University of Michigan

The prognosis for patients with high-grade gliomas is poor, with a median survival of one year. Treatment efficacy assessment is typically unavailable until 5-6 months post diagnosis. Investigators hypothesize that quantitative MRI (qMRI) can assess treatment efficacy three weeks after therapy starts, thereby allowing salvage treatments to begin earlier. The purpose of this work is to build a predictive model of treatment efficacy using qMRI data and to assess its performance. The outcome is one-year survival status. We propose a joint, two-stage Bayesian model. In stage I, we smooth the image data with a spatio-temporal pairwise-difference prior. Four novel summary statistics are then calculated from the smoothed images. In stage II, these statistics enter a generalized non-linear model (GNLM) as predictors of survival status. We use the probit link and a Multivariate Adaptive Regression Spline basis. Gibbs sampling and reversible jump MCMC are applied iteratively between the two stages to estimate the posterior distribution. Through both simulation studies and model performance comparisons we find that we are able to attain lower overall misclassification rates by accounting for the spatio-temporal correlation in the images and by allowing for a more complex and flexible decision boundary provided by the GNLM.

email: jincaowu@umich.edu

A BAYESIAN HIERARCHICAL FRAMEWORK FOR MODELING OF RESTING-STATE fMRI DATA

Shuo Chen*, Emory University
DuBois F. Bowman, Emory University

Functional magnetic resonance imaging (fMRI) has emerged as a powerful technique to investigate the neuropathophysiology of major psychiatric disorders. Examining the so-called default mode of brain function, captured when subjects are left to rest and think for themselves in the scanner, has revealed a variety of brain networks that exhibit consistent properties across subjects. Moreover, altered resting-state fMRI characteristics are associated with mental illnesses such as major depressive disorder (MDD). fMRI data possess complex spatial and temporal dependence structures, and Bowman et al. (2008) proposed a Bayesian spatial model for detecting task-related changes in brain activity and functional connectivity between distinct brain locations. Despite the flexibility of this approach, it is primarily intended for task-induced neural processing rather than for resting-state fMRI profiles. A key limitation is that the resting-state fMRI profiles cannot be represented by a single activation statistic for subsequent spatial modeling. In this study, we first decompose the temporal profiles to coefficients based on orthogonal bases, then use a Bayesian hierarchical model to estimate the parameters of the variance-covariance matrix that reflect the connectivity between various brain regions. We also develop a framework for the prior distribution of the connectivity matrix that can incorporate information from previous resting-state fMRI studies and from multiple imaging modalities.

email: schen33@emory.edu

COVARIATE-ADJUSTED NONPARAMETRIC ANALYSIS OF MAGNETIC RESONANCE IMAGES USING MARKOV CHAIN MONTE CARLO

Haley Hedlin*, Johns Hopkins University
Brian Caffo, Johns Hopkins University
Ziyad Mahfoud, American University of Beirut
Susan S. Bassett, Johns Hopkins University School of Medicine

Permutation tests are useful for drawing inferences from imaging data because of their flexibility and ability to capture features of the brain under minimal assumptions. However, most implementations of permutation tests ignore important confounding covariates. To employ covariate control in a nonparametric setting we have developed a Markov chain Monte Carlo (MCMC) algorithm for conditional permutation testing using propensity scores. We present the first use of this methodology for imaging data. Our MCMC algorithm is an extension of algorithms developed to approximate exact conditional probabilities in contingency tables, logit, and log-linear models. An application of our nonparametric method to remove potential bias due to the observed covariates is presented.

email: hhedlin@jhsph.edu

META ANALYSIS OF FUNCTIONAL NEUROIMAGING DATA VIA BAYESIAN SPATIAL POINT PROCESSES

Jian Kang*, University of Michigan
Timothy D. Johnson, University of Michigan
Thomas E. Nichols, University of Warwick
Tor D. Wager, Columbia University

There is a growing interest in the meta analysis of functional neuroimaging studies. Typical neuroimaging meta analysis data consist of peak activation coordinates (PACs) from several studies. Most published methods only produce null-hypothesis inferences and do not provide interpretable, fitted model. To overcome these limitations, we propose a Bayesian hierarchical marked spatial Cox cluster process model. The posterior intensity function provides information on the most likely locations of population centers as well as the inter-study variability of PACs about the population centers. We model the PACs as offspring of latent realizations of a study center process for each study. Further, the study-level point processes are the offspring of latent realizations of a population center process. To reduce the bias in the results, our model incorporates weights for each study, based on the quality of the study, as marks of the process. We illustrate our model with a meta analysis consisting of 437 studies from 164 publications and assess our model via sensitivity analyses and simulation studies.

email: jiankang@umich.edu

MULTISCALE ADAPTIVE SUPERVISED FEATURE SELCTION FOR IMAGE DATA

Ruixin Guo*, University of North Carolina-Chapel Hill
Hongtu Zhu, University of North Carolina-Chapel Hill

Two challenges of classification for image data are the high dimensionality of the feature space (number of pixels) and the complex spatial structure on a two-dimensional (2D) surface or in a 3D volume. Thus, feature selection prior to classification becomes important and necessary for image data. However, commonly used feature selection methods do not take into account of the special underlying information possessed by image data: spatial information. The goal of this paper is to develop a Multiscale Adaptive Supervised Feature Selction (MASFS) method, which is able to incorporate the class label information as well as the spatial pattern of image data. MASFS adopts the idea of Multiscale Adaptive Regression Models (MARM) proposed by Li et al. (2009). MARM is a multiscale adaptive regression model designed for Magnetic Resonance Imaging (MRI) data. In this paper, we utilize and generalize this idea to the classification context as a feature selection tool, which is our proposed MASFS. We apply our method to different simulation studies. Our proposed MASFS demonstrates the substantial improvement over commonly used feature selection methods, which can effectively detect the informative region and further improve the classification performance.

email: rguo@bios.unc.edu

A MULTIRESOLUTION ANALYSIS OF ENVIRONMENTAL LEAD EXPOSURE'S IMPACT ON BRAIN STRUCTURE

Shu-Chih Su*, Merck & Co.
Brian Caffo, Johns Hopkins University

A variety of ideas have been proposed to isolate effects of interest on different resolutions scales, including the wavelet transform and scale-space filtering. Such multilevel classification has gained increasing importance in many real-world problems such as in image processing and compression. In this work, we aim to provide an approach to inform researchers on an appropriate image resolution to search for effects. The data come from an ongoing prospective study of lead's impact on the central nervous system structure and function. The approach allows for finding the optimal resolution within a scientific interpretation and investigates the effect at different resolution levels. It explored the use of hierarchical classification as a spatial analysis tool for modeling lead's effect on brain structure at different resolution levels. Our multi-resolution results confirm that VBM approaches to multi-subject analysis of structural magnetic resonance images allows for comparing gray and white matter volume or densities for a particular effect of interest. However, to fit voxel-wise model might lead to parameters superfluous parameters. In addition, generally the smoothing parameters used are chosen in an ad hoc manner. In contrast, our approach chooses a resolution level based on strict numerical criteria.

email: shu-chih_su@merck.com

SEMIPARAMETRIC APPROACHES TO SEPARATION OF SOURCES USING INDEPENDENT COMPONENT ANALYSIS

Ani Eloyan*, North Carolina State University
Sujit K. Ghosh, North Carolina State University

Data processing and representation using lower dimensional hidden structure plays an essential role in many fields of applications, including image processing, neural networks, genome studies, signal processing and other areas where large datasets are often encountered. One of the common methods for data representation using lower dimensional structure involves the use of parametric Independent Component Analysis (ICA), which is based on a linear representation of the observed data in terms of hidden sources. The problem then involves the estimation of the mixing matrix and the densities of the hidden sources. However the solution of problem depends on the identifiability of the sources. This work first presents a set of sufficient conditions to establish the identifiability of sources and the mixing matrix using restrictions on the moments of the hidden source variables. Under such sufficient conditions we then obtain semi-parametric maximum likelihood estimate of the mixing matrix using a class of mixture distributions. The method is illustrated and compared with existing methods using simulated and real datasets.

email: aeloyan@ncsu.edu

22. MISSING DATA IN LONGITUDINAL STUDIES

PSEUDO-LIKELIHOOD ESTIMATION FOR INCOMPLETE DATA

Geert Molenberghs*, Universiteit Hasselt & Katholieke Universiteit Leuven, Belgium
Michael G. Kenward, London School of Hygiene and Tropical Medicine
Geert Verbeke, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium
Teshome Birhanu, Universiteit Hasselt, Belgium

In applied statistical practice, incomplete measurement sequences are the rule rather than the exception. Fortunately, in a large variety of settings, the stochastic mechanism governing the incompleteness can be ignored without hampering inferences about the measurement process. While ignorability only requires the relatively general missing at random assumption for likelihood and Bayesian inferences, this result cannot be invoked when non-likelihood methods are used. A direct consequence of this is that a popular non-likelihood-based method, such as generalized estimating equations, needs to be adapted towards a weighted version or doubly-robust version, when a missing at random process operates. So far, no such modification has been devised for pseudo-likelihood based strategies. We propose a suite of corrections to the standard form of pseudo-likelihood, to ensure its validity under missingness at random. Our corrections follow both single and double robustness ideas, and is relatively simple to apply. When missingness is in the form of dropout in longitudinal data or incomplete clusters, such a structure can be exploited towards further corrections. The proposed method is applied to data from a clinical trial in onychomycosis and a developmental toxicity study.

email: geert.molenberghs@uhasselt.be

A BAYESIAN SHRINKAGE MODEL FOR LONGITUDINAL BINARY DATA WITH INTERMITTENT MISSINGNESS AND DROPOUT WITH APPLICATION TO THE BREAST CANCER PREVENTION TRIAL

Chenguang Wang *, University of Florida
Michael J. Daniels, University of Florida
Daniels O. Scharfstein, Johns Hopkins University
Stephanie Land, University of Pittsburgh

We consider inference in randomized longitudinal studies with informative intermittent missing and/or dropouts. In this setting, it is well known that full data estimands are not identified unless unverified assumptions are imposed. We assume a non-future dependence model for the drop-out mechanism and partial ignorability for the intermittent missingness. We posit an exponential tilt model that links non-identifiable distributions and distributions identified under partial ignorability. This exponential tilt model is indexed by non-identified parameters, which are assumed to have an informative prior distribution, elicited from subject-matter experts. Under this model, full data estimands are shown to be expressed as functionals of the distribution of the observed data. To avoid the curse of dimensionality, we model the distribution of the observed data using a Bayesian shrinkage Beta-binomial model. In a simulation study, we compare our approach to a fully parametric and a fully saturated model for the distribution of the observed data. Our methodology is motivated and applied to data from the Breast Cancer Prevention Trial.

email: cgwang@ufl.edu

A TEST OF MISSING COMPLETELY AT RANDOM FOR REGRESSION DATA WITH NONRESPONSE

Gong Tang*, University of Pittsburgh

We consider regression analysis of data with nonresponse. When the nonresponse is missing at random, the ignorable likelihood method yields valid inference. However, the assumption of missing at random is not testable in general. A stronger assumption, missing completely at random, is testable. Likelihood ratio tests have been discussed in the context of multivariate data with missing values but these tests require the specification of the joint distribution of all variables (Little, 1988). Subsequently Chen & Little (1999)

proposed a Wald-type test, and Qu & Song (2002) proposed a score test for generalized estimating equations with using the same fact that all sub-patterns follow the same distribution under MCAR. For regression analysis of data with nonresponse, here we propose a Wald-type test for missing completely at random by comparing two sets of consistent estimators of regression parameters. This method can be applied to longitudinal data with dropouts.

email: got1@pitt.edu

EVALUATING STATISTICAL HYPOTHESES FOR NON-IDENTIFIABLE MODELS USING GENERAL ESTIMATING FUNCTIONS

Guanqun Cao*, Michigan State University
David Todem, Michigan State University
Lijian Yang, Michigan State University
Jason P. Fine, University of North Carolina-Chapel Hill

Many statistical models in biomedical research contain non and weakly identified parameters under interesting parametric formulations. Due to identifiability concerns, tests concerning some model parameters cannot use conventional statistical theory to assess significance. This paper extends the literature by developing a test statistic that can be used to evaluate hypotheses for any nonidentifiable estimating function. We derive the limiting distribution of this test statistic, and propose resampling approaches to approximate its asymptotic distribution. The methodology's practical utility is illustrated in simulations and an analysis of quality-of-life outcomes from a longitudinal study on breast cancer.

email: cao@stt.msu.edu

GENERALIZED ANOVA FOR CONCURRENTLY MODELING MEAN AND VARIANCE WITHIN A LONGITUDINAL DATA SETTING

Hui Zhang*, University of Rochester
Xin M. Tu, University of Rochester

Although widely used for comparing multiple samples in biomedical and psychosocial research, the analysis of variance (ANOVA) model suffers from a series of flaws that not only raise questions about conclusions drawn from its use, but also undercut its potential applications to modern clinical and observational research studies. In this paper, we propose a generalized ANOVA model that speaks to the limitations of this popular approach to address some important age-old technical issues as well as cutting-edge methodological challenges arising from several timely applications. By integrating inverse probability weighted estimates within the context of U-statistics, we develop distribution-free inference for this new class of models to address missing data for longitudinal clinical trials and cohort studies. We illustrate the proposed model with both real and simulated study data, with the latter investigating behaviors of model estimates under small and moderate sample sizes.

email: hui_zhang@urmc.rochester.edu

SEMIPARAMETRIC REGRESSION MODELS FOR REPEATED MEASURES OF MORTAL COHORTS WITH NON-MONOTONE MISSING OUTCOMES AND TIME-DEPENDENT COVARIATES

Michelle Shardell*, University of Maryland School of Medicine
Gregory E. Hicks, University of Delaware
Ram R. Miller, University of Maryland School of Medicine
Jay Magaziner, University of Maryland School of Medicine

We propose a semiparametric marginal modeling approach for longitudinal studies of mortal cohorts to estimate regression parameters interpreted as conditioned on being alive. The method accommodates outcomes and time-dependent covariates that are missing not at random with non-monotone missingness patterns via inverse probability weighting. Missing covariates are replaced by consistent estimates derived from a simultaneously-solved inverse-probability weighted estimating equation. Thus, we utilize data points with observed outcomes and missing covariates beyond the estimated weights while avoiding numerical methods to integrate over missing covariates. The approach is applied to a cohort of elderly female hip fracture patients to estimate the prevalence of walking disability over time as a function of body composition, inflammation, and age.

email: mshardel@epi.umaryland.edu

MULTIPLE IMPUTATION METHODS FOR PREDICTION WITH MULTIPLE LEFT-CENSORED BIOMARKERS DUE TO DETECTION LIMITS

Minjae Lee*, University of Pittsburgh
Lan Kong, University of Pittsburgh

Increasingly used in biomedical studies for the diagnosis and prognosis of acute and chronic diseases, biomarkers provide insight into the effectiveness of treatments and potential pathways that can be used to guide future treatment targets. The measurement of these markers is often limited by the sensitivity of the given assay, resulting in data that are censored either at the lower limit or upper limit of detection. For the Genetic and Inflammatory Markers of Sepsis (GenIMS) study, many different biomarkers were measured to examine the effect of different pathways on the development of sepsis. In this study, the left censoring of several important inflammatory markers has lead to the need for statistical methods that can incorporate this censoring into any analysis of the biomarker data. This paper focuses on the development of multiple imputation methods for the inclusion of multiple left censored biomarkers in a logistic regression analysis. Multivariate normal

distribution is assumed to account for the correlations between biomarkers. Gibbs sampler is used for estimation of distributional parameters and imputation of censored markers. The proposed methods are evaluated and compared with some simple imputation methods through simulations. A data set of inflammatory markers and coagulation markers from GenIMS study is used for illustration.

email: leem2@upmc.edu

23. SURVIVAL ANALYSIS

RISK-ADJUSTED MONITORING OF TIME TO EVENT

Axel Gandy, Imperial College, London
Jan Terje*, Kvaløy University of Stavanger, Norway
Alex Bottle, Imperial College, London
Fanyin Zhou, Imperial College, London

Recently there has been increasing interest in risk-adjusted cumulative sum charts (CUSUM) to monitor e.g. the performance of hospitals, taking into account the heterogeneity of patients. Even though many outcomes involve time, only conventional regression models are being commonly used. In this presentation it will be discussed how survival analysis models can be used for monitoring purposes. We suggest to use CUSUM charts based on the partial likelihood ratio between an out-of-control state and an in-control state. Issues like how to choose thresholds, types of alternatives and whether to include head-starts will be briefly discussed. One example concerning length of stay in hospital and some simulations comparing the survival analysis based monitoring to more conventional approaches will be presented.

email: jan.t.kvaloy@uis.no

BAYESIAN INFERENCE FOR CUMULATIVE INCIDENCE FUNCTION UNDER ADDITIVE RISKS MODEL

Junhee Han*, University of Arkansas
Minjung Lee, National Cancer Institute

In the analysis of competing risks, regression models for cumulative incidence function under the proportional hazards assumption have been widely studied. However, such an assumption may not hold for some data. As an alternative, we consider the semiparametric additive risks model (Aalen, 1980; McKeague and Sasieni, 1994) on the cause-specific hazard function. Sinha et al. (2009) proposed an empirical Bayesian method for the semiparametric additive hazards regression model and we extend their method to modeling cumulative incidence function. We use Gamma process as a prior on the baseline cumulative hazard function for each cause. We present the implementation of the framework and illustrate our method with malignant melanoma data.

email: falllunar@gmail.com

SEMIPARAMETRIC HYBRID EMPIRICAL LIKELIHOOD INFERENCE FOR TWO-SAMPLE COMPARISON WITH CENSORED DATA

Mai Zhou, University of Kentucky
Haiyan Su*, Montclair State University
Hua Liang, University of Rochester

Two-sample comparison problems are often encountered in practical projects and have widely been studied in literature. Due to practical demands, the research for this topic under special settings such as a semiparametric framework have also attracted great attentions. In this study, we develop a new empirical likelihood-based inference under more general framework by using the hazard formulation of the censored data for two sample semi-parametric hybrid models. We demonstrate that our empirical likelihood statistic converges to a standard chi-squared distribution under the null hypothesis. We further illustrate the use of the proposed test by testing the ROC curve with censored data, among others. Numerical performance of the proposed method is also examined.

email: suh@mail.montclair.edu

SEMIPARAMETRIC TRANSFORMATION MODELS BASED ON DEGRADATION PROCESSES

Sangbum Choi*, University of Wisconsin-Madison
Kjell A. Doksum, University of Wisconsin-Madison

In this article we study a class of semiparametric transformation models based on degradation processes. In many failure mechanisms in medical, social, and economic settings, most items under study degrade physically over time, thus we may assume a physical deterioration precedes failure. Some stochastic processes can be adopted for modeling degradation, in which failure occurs when the degradation process first encounters a threshold. We consider Poisson, compound Poisson, Wiener, Gamma, and inverse Gamma processes as degradation models and present their failure time distributions. It turns out that transformation models naturally arise if an underlying degradation process is assumed, and this line of thinking provides a new rich class of transformation models. We show that the nonparametric maximum likelihood estimators (NPMLEs) for the parameters of these models are consistent and asymptotically normal. The limiting covariance matrices for the estimators of the finite dimensional parameters achieve the semiparametric efficiency bounds. We consider numerical methods to compute the NPMLEs and their covariance estimators. Simulation studies demonstrate that the proposed

methods perform well in moderate sample sizes. The methodology is applied to an analysis of malignant melanoma survival data.

email: sangbum@stat.wisc.edu

MULTIPLE IMPUTATION METHODS FOR INFERENCE ON CUMULATIVE INCIDENCE WITH MISSING CAUSE OF FAILURE

Minjung Lee*, National Cancer Institute
Kathleen A. Cronin, National Cancer Institute
Mitchell H. Gail, National Cancer Institute
Eric J. Feuer, National Cancer Institute

Analysis of cumulative incidence (sometimes called absolute risk or crude risk) can be difficult if the cause of failure is missing for some subjects. Assuming missingness is random conditional on the observed data, we develop asymptotic theory for multiple imputation methods to estimate cumulative incidence. Covariates affect cause-specific hazards in our model, and we assume that separate proportional hazards models hold for each cause-specific hazard. Simulation studies show that procedures based on asymptotic theory have near nominal operating characteristics in cohorts of 200 and 400 subjects. The methods are illustrated with colorectal cancer data obtained from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI).

email: leem5@mail.nih.gov

CONSTRAINED NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION OF SURVIVOR FUNCTIONS UNDER STOCHASTIC ORDERING IN ONE-SAMPLE AND TWO-SAMPLE CASES

Yong Seok Park*, University of Michigan
John D. Kalbfleisch, University of Michigan
Jeremy MG Taylor, University of Michigan

This article considers estimators of survivor functions subject to a stochastic ordering constraint based on right censored data. Dykstra (1982) developed the constrained nonparametric maximum likelihood estimator (C-NPMLE) for such problems. The estimator from Dykstra's method, however, does not always give an estimator with maximum likelihood. In this paper, we present methods to obtain the C-NPMLE of the survivor functions in one- and two-sample settings. As an extension, we propose a method to obtain maximum C-NPMLE (MC-NPMLE) of the survivor function S_1(t) under the constraint S_1(t) <= S_2(t), where S_2(t) is a known arbitrary survivor function. Simulation studies show improvement in efficiency compared to Dykstra's estimator. Consistency of the estimator is proved and data from a study on larynx cancer is analyzed to demonstrate the use of the method.

email: yongpark@umich.edu

24. MICROARRAYS: DIFFERENTIAL EXPRESSION AND REPRODUCIBILITY

A BAYESIAN MODEL AVERAGING APPROACH TO DIFFERENTIALLY EXPRESSED GENE DETECTION IN OBSERVATIONAL MICROARRAY STUDIES

Xi K. Zhou*, Weill Medical College of Cornell University
Fei Liu, University of Missouri-Columbia
Andrew J. Dannenberg, Weill Medical College of Cornell University

Identifying differentially expressed genes (DEGs) under two or more experimental conditions is the primary objective in many microarray studies. As more and more studies are carried out in observational rather than well controlled experimental samples, it becomes important to evaluate and control the impact of sample heterogeneity on DEG finding. Typical methods for identifying DEGs required ranking all the genes according to a pre-selected statistic for two or more group comparisons. Potential contributions from other covariates were either ignored all together or controlled for all the genes under investigation. We show that such "one model for all" approaches can result in increased false discovery rate or reduced sensitivity because of model misspecification. A Bayesian model averaging approach is proposed for ranking genes differentially expressed in observational microarray studies. This approach properly controls for sample heterogeneity and accounts for model uncertainty. We demonstrate through simulated microarray data that this novel approach resulted in improved performances compared to the "one model for all" approaches. We applied this approach to two observational microarray studies.

email: kaz2004@med.cornell.edu

GENE EXPRESSION BARCODES BASED ON DATA FROM 8,277 MICROARRAYS

Matthew N. McCall*, Johns Hopkins University
Michael J. Zilliox, Emory University School of Medicine
Rafael A. Irizarry, Johns Hopkins University

The ability to measure gene expression based on a single microarray hybridization is necessary for microarrays to be a useful clinical tool. In its simplest form, this amounts to estimating whether or not each gene is expressed in a given sample. Surprisingly, this problem is quite challenging and has been disregarded for the most part in favor of estimating relative expression. We purpose addressing this problem by: (1) using the distribution of observed log2 intensities across a wide variety of tissues to estimate an expressed and an unexpressed distribution for each gene, and (2) for each gene in a sample, denoting it as expressed if its observed log2 intensity is more likely under the expressed distribution than under the unexpressed distribution and as unexpressed otherwise. The first step is accomplished by fitting a hierarchical mixture model to the plethora of publicly available data. To guarantee that

each gene will be unexpressed in at least one tissue, we hybridized yeast samples to human microarrays and included these arrays when estimating the distributions. The output of our algorithm is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). We call this a gene expression barcode.

email: mmccall@jhsph.edu

A MULTIVARIATE EMPIRICAL BAYES MODELING APPROACH TO SIMULTANEOUS INFERENCE OF MULTI-CLASS COMPARISON PROBLEMS

Xiting Cao*, University of Minnesota
Baolin Wu, University of Minnesota

Empirical Bayes method has proven to be a very useful approach for studying the simultaneous significance testing problems encountered in large-scale biomedical data. The basic idea of empirical Bayes approach is to borrow information across different tests by utilizing and modeling their similarities to help individual significance testing. For example, for multiple gene differential expression detections in microarray data analysis, empirical Bayes approach models the summary statistics across genes, which are often some univariate test statistics (e.g., the widely used t/F-statistics), to share information and improve the individual gene inference and overall detection power. We propose a multivariate nonparametric statistical approach based on empirical Bayes modeling for simultaneous significance testing. Our intuitive idea is to increase the information sharing of empirical Bayes modeling across individual tests, which can be achieved by summarizing the sample observations with some multivariate instead of univariate statistics to be modeled across different tests. For multi-class differential gene expression detection, the proposed approach naturally increases the information sharing by using all the pairwise class differences instead of the univariate F-statistic. We conduct extensive simulation studies to show that the proposed multivariate approach could improve upon the traditional univariate empirical Bayes approaches. We also illustrate the competitive performance of the proposed method using some public microarray data.

email: caoxx060@umn.edu

TESTING MULTIPLE HYPOTHESES USING POPULATION INFORMATION OF SAMPLES

Mingqi Wu*, Texas A&M University
Faming Liang, Texas A&M University

Multiple hypothesis tests have been widely studied in the recent literature of statistics, however, most of the studies focus on how to control the false discovery rate for a given set of test scores or, equivalently, test p-values. Given the vast data involved in a multiple hypothesis test, it is natural to think about how to make use of population information of samples to improve the power of the test for each individual subject and thus to improve the power of the multiple hypothesis test. In this paper, we propose a nonparametric method for evaluation of test scores for each individual subject involved in a multiple hypothesis test. The method consists of two key steps, smoothing over neighboring subjects and density estimation over control samples, both of which allow for the use of population information of the subjects. The new method is tested on both the microarray data and the ChIP-chip data. The numerical results indicate that use of population information can significantly improve the power of multiple hypothesis tests.

email: mqwu@stat.tamu.edu

ESTIMATE OF TRANSCRIPT ABSOLUTE CONCENTRATION FROM DNA MICORARRAYS

Yunxia Sui*, Brown University
Zhijin Wu, Brown University

Microarrays quantify the abundance of nucleic acids via hybridization. As the binding efficiency of the probes can vary greatly, the apparent expression measurement is affected by the gene expression level as well as the probe effects. Thus most microarray studies provide only a relative measurement of gene expression for the same gene in different samples. The expression levels of different genes within a sample cannot be directly compared and measurements of gene expression taken on different microarray platforms cannot be directly compared. Methods using probe sequences to predict probe behavior in hybridization have been proposed to extract absolute concentration on gene expression. However, the small amount of calibration data limited the power of sequence-only models. We demonstrate that by taking advantage of a large database of samples in combination of sequence model, the probe efficiency can be estimated with smaller bias and variance. Gene expression measures adjusting for probe efficiency allow the comparison of expression between genes, as well as comparison of the same gene measured on different platforms.

email: ysui@stat.brown.edu

STATISTICAL PRACTICE IN HIGH-THROUGHPUT siRNA SCREENS IDENTIFYING GENES MEDIATING SENSITIVITY TO CHEMOTHERAPEUTIC DRUGS

Fei Ye*, Vanderbilt University
Joshua A. Bauer, Vanderbilt University
Huiyun Wu, Vanderbilt University
Jennifer A. Pietenpol, Vanderbilt University
Yu Shyr, Vanderbilt University

Chemotherapeutic drug resistance is a critical challenge in the treatment of cancer, accountable for most cases of cancer treatment

failure. High-throughput small interfering RNA (siRNA) screens have been used to find potential candidate genes that, when silenced, cause resistance to certain chemotherapy drugs; however, few statistical methods are currently available for analyzing siRNA data. In this work, we undertake an examination and evaluation of the currently applied and potential statistical approaches for identifying siRNAs that influence sensitivity to chemotherapeutic drugs using high-throughput siRNA screens. We focus on normalization techniques, combined RNA and drug effect on cell viability, and control of false-positive and false-negative rates.

email: fei.ye@vanderbilt.edu

BAYESIAN HIERARCHICAL MODELS FOR CORRELATING EXPRESSION DATA ACROSS CHIPS

Bernard Omolo*, University of North Carolina-Chapel Hill

We develop a class of Bayesian hierarchical models (BHM) to determine the reproducibility of microarray expression data across chips, accounting for within probe (or gene) correlation and across probe (or gene) heterogeneity. We apply the BHM approach to estimate the correlation between 16 melanoma cell-lines on Agilent 4x44K chips.

email: bomolo@bios.unc.edu

25. STEPHEN LAGAKOS: VIEWS FROM FORMER STUDENTS SPANNING 2.5 DECADES

STEVE LAGAKOS, A LEGACY OF INTERACTIONS

Roger S. Day*, University of Pittsburgh

Many years ago, Steve Lagakos connected me with Dr. Emil "Tom" Frei III, who had led the charge against childhood leukemia in the '60's. Early on, we studied how anti-cancer drugs might work together, or "interact", to help patients. Studies of drug interaction affecting heterogeneous cancer cell populations led to the "worst drug rule". Over the years, chasing after several forms and concepts of interaction has led my students and me in several fascinating directions. Studies of drug interaction from a biological mechanism viewpoint led to the "generalized additive effect model". Many disparate contexts suggested the "weakest link" paradigm, recently applied with intriguing success to hierarchical multiparameter flow cytometry data. These days, as researchers try to cope with massive sets of genome-wide association studies, the concept of "interactions" returns in the hunt for epistatic relationships. Genomic, proteomic, microRNA, methylation, SNP, and other biological data present huge challenges, where purely empirical approaches cannot hope to expose the biological realities, where our "features" must somehow interact in our analyses the way they do in life. This talk will touch on some curiosities along this journey of interactions originally instigated by Steve.

email: day01@pitt.edu

BIOSTATISTICS IN THE ERA OF INTERDISCIPLINARY SCIENCE

Melissa D. Begg*, Columbia University
Roger D. Vaughan, Columbia University

Recent years have seen an ever increasing emphasis on interdisciplinary approaches to solving problems in biomedical and public health research. This focus is reflected in a number of ways: the growing body of literature on interdisciplinarity, the increasing prominence of teams of researchers working collaboratively, and the intense interest on the part of public and private granting agencies to fund interdisciplinary research proposals. The reason most often cited for the mounting interest in interdisciplinarity is that we are addressing health problems that are highly complex; and more complex problems require more complex approaches to achieve solutions, often requiring the combined efforts of teams of scientists from multiple disciplines. To succeed as an investigator in this new, interdisciplinary environment, biostatisticians need to understand the rationale behind these new approaches, how they are distinct from more traditional approaches, and how the field of biostatistics can work to advance and support these approaches across a wide range of health research initiatives. This presentation will include: definitions for multidisciplinary, interdisciplinary and translational science; attributes of successful interdisciplinary collaboration; barriers to interdisciplinarity frequently encountered in academic settings; and implications for the training of future generations of biostatisticians.

email: mdb3@columbia.edu

CROSS-SECTIONAL PREVALENCE TESTING FOR ESTIMATING HIV INCIDENCE

Rui Wang*, Harvard University
Stephen W. Lagakos, Harvard University

HIV incidence estimation based on a cross-sectional sample of individuals evaluated with both a sensitive and less-sensitive test offers important advantages to a cohort study approach. However, two concerns have been raised regarding the reliability of the cross-sectional approach. One is the difficulty in obtaining a reliable external approximation for the mean window period between detectability of HIV infection with the sensitive and less-sensitive tests. The other is how to handle false negative results with the less-sensitive test; that is, subjects who may test negative--implying in the recent infection state--long after they are infected. We propose an augmented cross-sectional study design in which subjects who found in the recent infection state are followed for transition to the non-recent infection state. Inference is based on likelihood methods which account for the length-biased nature of the window periods

of subjects found in the recent infection state, and relate the distribution of their forward recurrence times, the time from the cross-sectional sample to become reactive to the less-sensitive test, to the population distribution of the window period. The approach performs well in simulation studies and eliminates the need for external approximations of the mean window period and the false negative rate.

email: rwang@hsph.harvard.edu

TRENDS AND CHALLENGES IN RESEARCH INVOLVING ELDERLY AND IMPAIRED DRIVERS

Jeffrey D. Dawson*, University of Iowa
Elizabeth Dastrup, University of Iowa
Amy M. Johnson, University of Iowa
Ergun Y. Uc, University of Iowa
Matthew Rizzo, University of Iowa

Motor vehicle driving is a complex process which may be studied at several levels. Our multi-disciplinary research team has been studying elderly and neurologically-impaired drivers (e.g., those with Alzheimer's disease and Parkinson's disease) for over 10 years. We are interested in how demographics, disease, and cognitive abilities predict driving behaviors in simulators, instrumented vehicles, and naturalistic settings. Our goals are a) to find tests of cognitive, visual, and motor skills that could be effective screening tools for driving safety, b) to find personal and vehicular interventions that could improve driver safety, and c) to use modern technology to modify vehicles to be diagnostics devices for neurological disease and progression. In this presentation, we highlight some of the methodological challenges that we face as our work advances from descriptive studies to predictive studies to interventional studies. This work is supported by NIH awards AG017177, AG015071, and NS044930.

email: jeffrey-dawson@uiowa.edu

26. ADAPTIVE CLINICAL TRIAL DESIGNS FOR DOSE FINDING

BAYESIAN ADAPTIVE DOSE-FINDING STUDIES WITH DELAYED RESPONSES

Haoda Fu*, Eli Lilly and Company
David Manner, Eli Lilly and Company

In recent years, Bayesian response adaptive designs have been used to improve the efficiency of learning in dose finding studies. Many current methods for analyzing the data at the time of the interim analysis only use the data from patients who have completed the study. However Slow enrollment rates can limit the number of patients who complete the study in a given period of time. Consequently, at the time of an interim analysis, there may be only a small proportion (e.g., 20%) of patients who have completed the study. In this paper, we propose a new Bayesian prediction model to incorporate all the data (from patients who have completed the study and those who have not completed) to make decisions about the study at the interim analysis. Examples of decisions made at the interim analysis include adaptive treatment allocation, dropping non-efficacious dose arms, stopping the study for positive efficacy, or stopping the study for futility. The model is able to handle incomplete longitudinal data including missing data considered Missing At Random (MAR). A utility function based decision rule is also discussed. The benefit of our new method is demonstrated through trial simulations. Three scenarios are examined and the simulation results demonstrate that this new method outperforms traditional design with the same sample size in each of these scenarios.

email: fuhaoda@gmail.com

BAYESIAN PHASE I DOSE-FINDING DESIGN MODELING DISCRETE-TIME TOXICITY GRADES

Lin Yang*, University of Texas M.D. Anderson Cancer Center
Nebiyou B. Bekele, University of Texas M.D. Anderson Cancer Center
Donald A. Berry, University of Texas M.D. Anderson Cancer Center

Toxicity in phase I oncology trials is commonly categorized by grade. Standard phase I dose-finding designs address early toxicity of sufficiently high grade. It is not standard to use early information about early lower grade toxicities though they may predict later higher grade toxicities. We propose a Bayesian phase I dose-finding design that models toxicity grade over time. The goal is to declare MTD based on average toxicity score (ATS), which is the weighted average of the probability of the various toxicity grades. The proposed method consists of a multinomial/Dirichlet multi-state model and a dose escalation/de-escalation algorithm. The algorithm allows early stopping for toxicity or success. This method does not require full follow-up of previous patients before assigning dose to the next patient. We conducted simulation studies for various toxicity scenarios, including those with probability of toxicity being constant, decreasing, and increasing with time. We compare operating characteristics with those of standard designs, including with other Bayesian designs. We describe settings in which the proposed design shortens trial duration and assigns fewer patients to risky doses.

email: echoyl@yahoo.com

BAYESIAN PHASE I/II DRUG-COMBINATION TRIAL DESIGN IN ONCOLOGY

Ying Yuan, University of Texas M.D. Anderson Cancer Center
Guosheng Yin*, University of Hong Kong

We propose a new integrated phase I/II trial design to identify the most efficacious dose combination that also satisfies certain safety requirements for drug-combination trials. We first take a Bayesian copula-type model for dose finding in phase I. After identifying a set of admissible doses, we immediately move the entire set forward in parallel to phase II. We propose a novel adaptive randomization scheme to favor assigning patients to more efficacious dose-combination arms. Based on the efficacy data also collected in the dose-finding stage, the adaptive randomization procedure is immediately invoked in phase II to randomize new patients to a more efficacious arm with a higher probability. Our adaptive randomization scheme takes into account both the rate and variability of efficacy. By using a moving reference to compare the relative efficacy among treatment arms, our method achieves a high resolution to distinguish different arms. We illustrate the proposed method using a phase I/II melanoma clinical trial, and conduct extensive simulation studies to examine the operating characteristics of the design.

email: gyin@hku.hk

BAYESIAN MODEL AVERAGING CONTINUAL REASSESSMENT METHOD IN PHASE I CLINICAL TRIALS

Guosheng Yin, University of Hong Kong
Ying Yuan*, University of Texas M.D. Anderson Cancer Center

The continual reassessment method (CRM) is a popular dose-finding design for phase I clinical trials. This method requires practitioners to prespecify the toxicity probability at each dose. Such prespecification can be arbitrary, and different specifications of toxicity probabilities may lead to very different design properties. To overcome the arbitrariness and further enhance the robustness of the design, we propose using multiple parallel CRM models, each with a different set of prespecified toxicity probabilities. In the Bayesian paradigm, we assign a discrete probability mass to each CRM model as the prior model probability. The posterior probabilities of toxicity can be estimated by the Bayesian model averaging (BMA) approach. Dose escalation or de-escalation is determined by comparing the target toxicity rate and the BMA estimates of the dose toxicity probabilities. We examine the properties of the BMA-CRM through extensive simulation studies, and also compare the new method and its variants with the original CRM. The results show that our BMA-CRM is competitive, robust, and eliminates the arbitrariness of the prespecification of toxicity probabilities.

email: yyuan@mdanderson.org

DOSE FINDING IN DRUG COMBINATIONS WITH DISCRETE AND CONTINUOUS DOSES

Lin Huo*, University of Texas M.D. Anderson Cancer Center
Ying Yuan, University of Texas M.D. Anderson Cancer Center
Guosheng Yin, University of Hong Kong

In cancer clinical trials, there has been a growing trend of treating patients with combined agents. The synergism of multiple drugs is often the primary motivation for such studies. We propose a Bayesian adaptive dose-finding design for a drug-combination trial with continuous and discrete doses. We jointly search for the maximum tolerated dose combination through a two-stage procedure. The first stage takes a continual reassessment method to locate the appropriate dose for the discrete agent, and the second stage estimates the best dose for the continuous agent by continuously updating the posterior estimates for the toxicity probabilities of the combined doses. To enhance the model fitting and predictability in stage 2, we take an adaptive model selection procedure when searching for the discrete dose in stage 1. Based on the accumulating data, we adaptively assign each new cohort of patients to the most appropriate dose. We conduct extensive simulation studies to examine the operating characteristics of the design and illustrate the proposed method under various practical scenarios based on a recent clinical trial at M.D. Anderson Cancer Center.

email: lin.huo@uth.tmc.edu

A TWO-STAGE DOSE-RESPONSE ADAPTIVE DESIGN METHOD FOR ESTABLISHING PROOF OF CONCEPT

Yoko Tanaka*, University of Pittsburgh
Stewart Anderson, University of Pittsburgh
Allan R. Sampson, University of Pittsburgh

In a two-stage dose-response adaptive design where both dropping and adding treatment arms are possible between the stages, we propose a method of extending the multiple comparison procedures-modeling approach (MCP-Mod) originally developed by Bretz, et al (2005). Among a set of potential candidates, several doses and a placebo are used in the first stage. In the second stage, the placebo and a set of doses are selected according to a pre-specified dose adaptation rule. In both stages, we use the same set of candidate dose-response models. Furthermore, in both stages, we test preliminary hypotheses to establish whether or not a dose-response relationship exists. In each stage, a maximum test statistic is selected from among the model-associated multiple contrast test statistics. Statistics in Stage 2 are weighted based on the dose adaptation result. The preliminary test results of both stages are combined to establish "global" dose-response evidence or a Proof of Concept (PoC) by use of a Conditional Error Function (CEF). A pre-specified decision rule utilizes the p-values associated with the maximum statistics obtained from the two stages and the CEF. Using simulations based on 10,000 trials, our method is evaluated by assessing the probability of detecting a dose-response curve.

email: yot3@pitt.edu

PROPORTIONAL ODDS MODEL FOR DOSE FINDING CLINICAL TRIAL DESIGNS WITH ORDINAL TOXICITY GRADING

Emily M. Van Meter*, Medical University of South Carolina
Elizabeth Garrett-Mayer, Medical University of South Carolina
Dipankar Bandyopadhyay, Medical University of South Carolina

Currently many dose finding clinical trial designs, including the continual reassessment method (CRM) and the standard 3 + 3 design, dichotomize toxicity outcomes based on pre-specified dose-limiting toxicity criteria. This loss of information is particularly inefficient due to the small sample sizes in phase I trials. Common Toxicity Criteria (CTCAEv3.0) classify adverse events into grades 1 through 5, which range from 1 as a mild adverse event to 5 as death related to an adverse event. In this paper, we extend the CRM to include ordinal toxicity outcomes as specified by CTCAEv3.0 using the proportional odds model and compare results with the dichotomous CRM. A sensitivity analysis of the new design compares various prior distributions, target dose-limiting toxicity rates, sample sizes, and cohort sizes. This design is also assessed under various dose-toxicity relationship models including proportional odds models as well as those that violate the proportional odds assumption. A simulation study shows that the proportional odds CRM performs as well as the dichotomous CRM on all criteria compared, and notably with more precision to estimate the maximum tolerated dose (MTD). These findings suggest that it is beneficial to incorporate ordinal toxicity endpoints into phase I trial designs.

email: EmilyVanMeter@gmail.com

27. BAYESIAN METHODS FOR COMBINING DATA FROM MULTIPLE SOURCES FOR CLINICAL TRIALS

SYNTHETIC PRIORS FROM ANALYSIS OF MULTIPLE EXPERTS' OPINIONS

Sourish Das*, Duke University
Hongxia Yang, Duke University
David Banks, Duke University

In this talk, we first present a brief review of Bayesian methods for prior elicitation from subject matter experts. But expert opinion is often internally inconsistent, and there is often substantial disagreement with the opinions of other experts. We therefore develop a statistical model for the elicited opinions, and use that to borrow strength across the responses through an exchangeable prior. Several versions of that prior are considered; the most advanced uses covariate information on the experts to characterize their areas of agreement and disagreement, which ultimately allows the estimation of the response from a synthetic expert whose covariates are selected by the analyst. To this end we present a novel technique to incorpoarte the background information of the expert through a hierarchical Dirichlet regression model and a hierarchical logistic-normal regression model.

email: sourish.das@gmail.com

A NOVEL APPROACH FOR ELICITING AN ODDS RATIO IN THE SETTING OF INCOMPLETE LONGITUDINAL DATA

Michael Daniels*, University of Florida
Chenguang Wang, University of Florida
Daniel Scharfstein, Johns Hopkins University

We consider inference in randomized studies, in which repeatedly measured outcomes may be informatively missing due to drop out. In this setting, it is well known that full data estimands are not identified unless unverified assumptions are imposed. We assume a non-future dependence model for the drop-out mechanism and posit an exponential tilt model that links non-identifiable and identifiable distributions. This model is indexed by non-identified parameters can be interpreted as odds ratios. We propose a novel approach to construct priors on these parameters by eliciting relative risks from subject matter experts. Our methodology is motivated and applied to data from the Breast Cancer Prevention Trial.

email: mdaniels@stat.ufl.edu

USING PRIOR INFORMATION AND ELICITED UTILITIES FOR ADAPTIVE DECISION MAKING IN PHASE I/II TRIALS

Peter F. Thall*, University of Texas M.D. Anderson Cancer Center

Adaptively choosing the best treatment for each patient in a phase I/II clinical trial is difficult because sample sizes are small and safety concerns dominate decision making. I will review a general Bayesian paradigm for constructing practical adaptive decision rules that combine prior information and elicited utilities with the data observed during the trial. Two applications will be presented. The first is a design that selects optimal dose pairs of two agents for bladder cancer by using elicited consensus utilities of bivariate ordinal toxicity and tumor response outcomes. The second design selects optimal combinations of dose and schedule of an agent given after an allogeneic stem cell transplant by using the joint utility of the time to toxicity and the time to disease progression. A graphical method for constructing a utility surface from elicited values will be illustrated.

email: rex@mdanderson.org

BORROWING STRENGTH WITH NON-EXCHANGEABLE PRIORS OVER SUBPOPULATIONS

Luis Gonzalo Leon-Novelo, University of Florida
B. Nebiyou Bekele, University of Texas M.D. Anderson Cancer Center
Peter Mueller*, University of Texas M.D. Anderson Cancer Center
Fernando Quintana, Pontificia Universidad Catolica de Chile
Kyle Wathen, University of Texas M.D. Anderson Cancer Center

We introduce a non-parametric Bayesian model for success rates in a phase II clinical trial with patients presenting different subtypes of the disease under study. The subtypes are not a priori exchangeable. The lack of prior exchangeability hinders straightforward use of traditional hierarchical models to implement borrowing of strength across disease subtypes. We introduce instead a random partition model for the set of disease subtypes. All subtypes within the same cluster share a common success probability. The random partition model is a variation of the product partition model that allows us to model a non-exchangeable prior structure.

The motivating study is a phase II clinical trial of patients with sarcoma. Each patient presents one subtype of the disease and subtypes are grouped by good, intermediate and poor prognosis. The prior model should respect the varying prognosis across disease subtypes. Two subtypes with equal prognosis should be more likely a priori to co-cluster than any two subtypes with different prognosis. The practical motivation for the proposed approach is that the number of accrued patients within each disease subtype is too small to asses the success rates with the desired precision if we were to analyze the data for each subtype separately.

Like a hierarchical model, the proposed clustering approach considers all observations, across all disease subtypes, to estimate individual success rates. But in contrast with standard hierarchical models, the model considers disease subtypes a priori non-exchangeable. This implies that when assessing the success rate for particular type, our model borrows more information from the outcome of the patients sharing the same prognosis than from the others.

email: pmueller@mdanderson.org

28. COMPETING RISKS IN ACTION

REGRESSION STRATEGY FOR THE CONDITIONAL PROBABILITY OF A COMPETING RISK

Aurelien Latouche*, University of Versailles

Competing risks are classically summarised by the cause-specific hazards and the cumulative incidence function. To get a full understanding of the competing risks, these quantities should be viewed simultaneously for all possible events. Another quantity is the conditional probability of a competing risk, (aka conditional cumulative incidence) which is defined as the probability of having failed from a particular cause given that no other (competing) events have occurred. This quantity provides useful insights and its interpretation may be preferable to communicate to clinicians. The use of the conditional probability has been limited by the lack of regression modelling strategy. In this work we apply recently developed regressions methodologies to the conditional probability function and illustrate the insights which can be gained using this methodology with a case study on conditioning regimens prior stem-cell transplantation (SCT) in acute leukemia.

email: aurelien.latouche@uvsq.fr

SUMMARIZING DIFFERENCES IN CUMULATIVE INCIDENCE FUNCTION

Mei-Jie Zhang*, Medical College of Wisconsin
Jason Fine, University of North Carolina-Chapel Hill

The cumulative incidence function is widely reported in competing risks studies, with group differences assessed by an extension of the logrank test. However, simple, interpretable summaries of group differences are not presented. An adaptation of the proportional hazards model to the cumulative incidence function is often employed, but the interpretation of the hazard ratio may be somewhat awkward, unlike the usual survival set-up. We propose nonparametric inferences for general summary measures, which may be time-varying, and for time-averaged versions of the measures. A real data example illustrates the practical utility of the methods.

email: meijie@mcw.edu

INFERENCE ON QUANTILE RESIDUAL LIFE UNDER COMPETING RISKS

Jong-Hyeon Jeong*, University of Pittsburgh

The ultimate goal in medical research is to extend a patient's remaining life years by an intervention. Especially, when a secondary therapy to be applied in the middle of follow-up after the initial therapy is considered, it would be reasonable to evaluate the treatment effect in terms of prolonging a patient's remaining life years conditional on survival beyond that time point. Popular summary measures to characterize a probability distribution of the remaining lifetimes are the mean or quantile residual life function. However, the quantile function is often preferred to summarize a residual life distribution, especially, under competing risks because the mean function does not exist theoretically in that case. A simple example of a competing risks analysis would be to infer the proportion of breast cancer related-deaths in the presence of non-breast cancer-related deaths due to heart failures, say. In this talk, we define the cause-specific residual subdistribution function under competing risks and propose a test statistic to compare the quantile

residual lifetimes between two groups. The proposed method is applied to a breast cancer dataset from a phase III clinical trial.

email: jeong@nsabp.pitt.edu

29. STUDYING GENETIC AND ENVIRONMENTAL RISK FACTORS OF COMPLEX HUMAN DISORDERS AND THEIR INTERACTIONS

DISCOVERING INFLUENTIAL VARIABLES: A METHOD OF PARTITIONS

Shaw-Hwa Lo*, Columbia University
Herman Chernoff, Harvard University
Tian Zheng, Columbia University

A trend in all scientific disciplines, based on advances in technology, is the increasing availability of high dimensional data in which are buried important information. A current urgent challenge to statisticians is to develop effective methods of finding the useful information from the vast amounts of messy and noisy data available, most of which are noninformative. This paper presents a general computer intensive approach, based on a method by Lo and Zheng for detecting which, of many potential explanatory variables, have an influence on a dependent variable Y. This approach is suited to detect influential variables, where causal effects depend on the confluence of values of several variables. It has the advantage of avoiding a difficult direct analysis, involving possibly thousands of variables, by dealing with many randomly selected small subsets from which smaller subsets are selected, guided by a measure of influence I. The main objective is to discover the influential variables, rather than to measure their effects. Once they are detected, the problem of dealing with a much smaller group of influential variables should be vulnerable to appropriate analysis. In a sense, we are confining our attention to locating a few needles in a haystack.

email: slo@stat.columbia.edu

COMBINING DISEASE MODELS TO TEST FOR GENE-ENVIRONMENT INTERACTION IN NUCLEAR FAMILIES

Thomas J. Hoffmann* University of California-San Francisco
Nan M. Laird, Harvard University

It is important to have robust gene-environment interaction tests that can utilize a variety of family structures in an efficient way. We focus on methods to test for a gene-environment interaction in the presence of main genetic and environmental effects. We first propose a relative risk method that can be applied to any family structure. We propose an approach that is more powerful when there are discordant sibs, but does not allow one to use cases when all offspring are affected, e.g. trios. Lastly, we propose using a hybrid of these approaches so that we can use the more powerful approach whenever applicable, and still obtain some information from families that do not have discordant offspring.

email: tjhoffm@gmail.com

THE VALUE OF SNPs FOR PROJECTING BREAST CANCER RISK

Mitchell H. Gail*, National Cancer Institute

I assessed the value of adding seven breast-cancer-associated SNPs to the Breast Cancer Risk Assessment Tool (BCRAT), which is based on ages at menarche and first live birth, family history, and breast biopsy examinations. The model with SNPs (BCRATplus7) had an area under the receiver operating characteristic curve (AUC) of 0.632, compared to 0.607 for BCRAT. This improvement is less than from adding mammographic density. I also assessed how much BCRATplus7 reduced expected losses in deciding whether a woman should take tamoxifen to prevent breast cancer, whether a woman should have a mammogram, and whether BCRATplus7 was more effective than BCRAT in allocating a scarce public health resource. In none of these applications did BCRATplus7 perform substantially better than BCRAT. A cross-classification of risk by the two models indicated that some women would change risk categories if BCRATplus7 were used, but it is not known if BCRATplus7 is well calibrated. These results were hardly changed if three recently found risk-associated SNPs were added. I conclude that available SNPs do not improve the performance of models of breast cancer risk enough to warrant their use outside the research setting.

email: gailm@mail.nih.gov

TESTING FOR THE EFFECT OF RARE VARIANTS IN COMPLEX TRAITS: A NOVEL APPROACH

Iuliana Ionita-Laza*, Columbia University
Christoph Lange, Harvard University
Nan M. Laird, Harvard University

Common diseases, such as bipolar disorder, asthma, cancer, etc. are caused by a complex interplay among multiple genetic and environmental risk factors. Both common and rare genetic variants are expected to influence risk to these traits. Thus far, most research in finding disease susceptibility variants has focused, out of necessity, on the discovery of common susceptibility variants (i.e. variants with a population frequency of at least 5%). However, taken together, the common variants identified so far to be associated with disease only explain a small fraction of the estimated trait heritability. Recent advances in sequencing technologies have brought along substantial reductions in cost and increases in genomic throughput by more than three orders of magnitude. These developments have lead to an increasing number of sequencing studies being performed, including the

1000 Genomes Project, with the main goal to identify rare genetic variants. Therefore, for the first time, it is now possible to systematically assess the role rare variants may play in various complex traits. In this talk I will discuss challenges in testing for the effect of rare variants in complex diseases, and propose a novel testing strategy, based on the Cochran-Armitage test for trend. I will show comparisons with existing methods on both simulated and real data.

email: ii2135@columbia.edu

30. SPATIAL/TEMPORAL: MODELING AND METHODOLOGY

VARIATIONAL BAYESIAN METHOD FOR SPATIAL DATA ANALYSIS

Qian Ren*, University of Minnesota
Sudipto Banerjee, University of Minnesota

With scientific data available at geocoded locations, investigators are increasingly turning to spatial process models for carrying out statistical inference. However, fitting spatial models often involves expensive matrix decompositions whose computational complexity increases in cubic order with the number of spatial locations. This situation is aggravated in Bayesian settings where such computations are required once every iteration of the Markov chain Monte Carlo (MCMC) algorithms. In this paper we describe the use of Variational Bayesian (VB) methods as an alternative to MCMC to approximate the posterior distributions of complex spatial models. Variational methods, which have been used extensively in Bayesian machine learning for several years, provide a lower bound on the marginal likelihood, which can be computed efficiently. We provide some results for the variational updates in several models especially emphasizing their use in multivariate spatial analysis. We demonstrate estimation and model comparisons from VB methods by using simulated data as well as some environmental datasets and compare them with inference from MCMC.

email: qian.ren@gmail.com

GAUSSIAN PREDICTIVE PROCESS MODEL FOR RANDOM KNOTS

Rajarshi Guhaniyogi*, University of Minnesota
Andrew O. Finley, Michigan State University
Sudipto Banerjee, University of Minnesota
Alan Gelfand, Duke University

With the increasing availability of geocoded scientific data, investigators are increasingly turning to spatial process models for carrying out statistical inference on environmental processes. Over the last few decades hierarchical spatial models implemented through Markov chain Monte Carlo (MCMC) have become especially popular as they enable richer modelling that would be infeasible otherwise. However, fitting hierarchical spatial models often involves matrix decompositions whose complexity increases in cubic order with the number of spatial locations, rendering such models infeasible for large datasets. One approach derives a "predictive process model" that alleviates computational bottlenecks by optimally projecting the spatial process using a smaller set of locations ("knots"). However, selecting these knots is a challenging problem and may become problematic for nonstationary data. To address this problem we devise two different methods for knot selection. The first employs a reversible jump MCMC which allows the Markov Chain to jump between models with different number of knots. The second approach induces a prior distribution on the knots using a point process that avoids reversible jump MCMC. Both these methods allow the knots to learn from the process, thereby considerably decreasing the number of knots needed for effective implementation and eliciting further computational benefits. Some theoretical aspects of the predictive process will be discussed along with practical illustrations.

email: guhan003@umn.edu

ADDITIVE MODELS WITH SPATIO-TEMPORAL DATA

Xiangming Fang*, East Carolina University
Kung-Sik Chan, University of Iowa

Additive models have been widely used. While the procedure for fitting an additive model to independent data has been well established, not as much work has been done when the data are correlated. The currently available methods for fitting additive models with correlated data are either computationally expensive or numerically unstable. We propose a new approach to fit additive models with spatio-temporal data via the penalized likelihood approach which estimates the smooth functions and covariance parameters by iteratively maximizing the penalized log likelihood. Both maximum likelihood (ML) and restricted maximum likelihood (REML) estimation schemes are developed. The asymptotic distribution of the estimates is studied in a Bayesian framework. Conditions for asymptotic posterior normality are investigated for the case of separable spatio-temporal data with fixed spatial covariance structure and no temporal dependence. We also propose a method to check the assumption of temporal independence and a new model selection criterion for comparing models with and without spatial correlation. The proposed methods are illustrated by both simulation study and real data analysis.

email: fangx@ecu.edu

COMPOSITE QUADRATIC INFERENCE FUNCTIONS AND APPLICATIONS IN SPATIO-TEMPORAL MODELS

Yun Bai*, University of Michigan
Peter X.K. Song, University of Michigan
Trivellore Raghunathan, University of Michigan

Spatio-temporal process modeling has received increasing attention in recent statistical research. However, due to the high dimensionality of the data, joint modeling of the spatial and temporal processes is computationally prohibitive for both likelihood-based and Bayesian approaches. In this paper, we propose a composite quadratic inference function approach to estimate spatio-temporal covariance structures, which significantly reduces the dimensionality and computational complexity and is more efficient than ordinary composite likelihood methods currently used in spatial/temporal literature. We construct three sets of estimating functions from spatial, temporal and cross pairs. This often results in a larger set of estimating equations than the number of parameters, so we form a quadratic inference function in a similar spirit to generalized method of moments (GMM) for estimation. We show that the proposed method yields consistent estimation and asymptotic distribution of the estimator is also derived. Simulations prove that our method performs very well and is more efficient than the weighted composite likelihood method. Finally, we apply our method to study the spatio-temporal dependence structure of PM10 particles in northeastern United States.

email: yunbai@umich.edu

NONPARAMETRIC HIERARCHICAL MODELING FOR DETECTING BOUNDARIES IN AREALLY REFERENCED SPATIAL DATASETS

Pei Li*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Timothy E. Hanson, University of Minnesota
Alexander M. McBean, University of Minnesota

With increasing accessibility to Geographical Information Systems (GIS) software, researchers and administrators in public health are increasingly encountering spatially referenced datasets. Inferential interest of spatial data analysis often resides not in the statistically estimated maps themselves, but on the formal identification of "edges ' or "boundaries" on the map. Boundaries can be thought of as a set of connected spatial locations that separate areas with different characteristics. A class of nonparametric bayesian models are proposed in this paper to account for uncertainty at various levels to elicit spatial zones of rapid change that suggest hidden risk factors driving these disparities. Simulation study are conducted to illustrate the new approaches and compare with existing methods. "Boundaries ' on Pneumonia and Influenza hospitalization map from the SEER-Medicare program in Minnesota are detected using the proposed approaches.

email: peili@biostat.umn.edu

BAYESIAN ANALYSIS OF HIGH-THROUGHPUT DATA VIA REGRESSION MODELS WITH SPATIALLY VARYING COEFFICIENTS

Xinlei Wang*, Southern Methodist University
Guanghua Xiao, University of Texas Southwestern Medical Center

High throughput technology, which allows for simultaneous acquisition of large amounts of data, has emerged over the last few years as an important tool in accelerating the pace of scientific discovery. The density and volume of data generated in a single experiment continue to grow exponentially due to rapid technology advances. Such high-density data are often spatially correlated with high noise levels. When there are only a few replicates available, as is typical in practice, modeling the spatial correlation carefully can greatly reduce noise, improve estimation efficiency, and lead to more reliable scientific findings. The objective of this study is to develop appropriate statistical methods to analyze high-density data, such as those generated from epigenetic studies, cell image-based high content screening and time course gene expression experiments, where real biological effects are often spatially dependent. We propose a set of new Bayesian regression models, which allow us to conduct spatial smoothing and statistical inference simultaneously to gain efficiency. These include normal linear models, generalized linear models and nonparametric regression models with basic functions, all having spatially varying coefficients whose covariance structures can be determined through autoregressive models or functions of a certain distance metric.

email: swang@smu.edu

MODELING TIME SERIES DATA WITH SEMI-REFLECTIVE BOUNDARIES WITH APPLICATION TO LATERAL CONTROL OF MOTOR VEHICLES

Amy M. Johnson*, University of Iowa
Jeffrey D. Dawson, University of Iowa

Some time series sequences have boundaries which tend to reflect the data towards mid-range values. For example, a motor vehicle is usually steered back towards the middle of a driving lane when the wheels approach or cross the lane boundaries. Dawson et al (2009) proposed a model to accommodate such semi-reflective boundaries, using weighted third-order polynomial projections and a signed error term that is a stochastic function of a re-centering parameter. This model allows the polynomial weights, the re-centering parameter, and the average level of the measured values to vary across subjects. In this report, we demonstrate how

to estimate the parameters of this model using standard statistical software, and we use simulations to illustrate the interpretation of the parameters as well as to investigate the statistical properties of the estimation procedure. We apply this model to vehicular lateral position data from 127 middle-aged and elderly drivers, and show that the re-centering parameter is associated with clinical predictors and on-road driving safety errors, suggesting that this model may be a useful tool in assessing the ability of drivers to safely operate a vehicle. This work was supported by NIH/NIA awards AG17177 and AG15071.

email: amy-m-johnson@uiowa.edu

31. PATHWAY AND NETWORK-BASED GENOMIC ANALYSIS

GENETIC NETWORK LEARNING IN GENETICAL GENOMICS EXPERIMENTS

Jianxin Yin*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

Data from genetical genomics experiments provide one possibility of constructing the transcriptional networks. To learn the structure of mixed Bayesian networks composed of phenotypes (continuous variables) and QTLs (discrete variables), a method based on constrained structure learning and penalized likelihood is proposed. Simulation study and a real data set analysis is showed to demonstrate our method.

email: yinj@mail.med.upenn.edu

ROBUST GENE PATHWAY TESTING

Hongyuan Cao*, University of North Carolina-Chapel Hill
Fred Wright, University of North Carolina-Chapel Hill
Michael Kosorok, University of North Carolina-Chapel Hill

In gene expression data, the inherent pathway structure can be used to test for its joint association with a phenotype of interest. In the literature, people either compare the association strength within the gene pathway or with its complement. But if there is a significant proportion of genes associated with the phenotype of interest, large gene sets corresponding to irrelevant pathways could contain many genes associated with the phenotype by chance. This motivates us to use the proportion of significantly expressed genes in a pathway as comparison criterion. The proportion estimates are derived for t-tests, F-tests and \chisqaure tests. This approach is shown to be robust to the size of the pathway. Subsampling and bootstrap are used to do inference.

email: hycao@email.unc.edu

STRUCTURED VARYING-COEFFICIENT MODEL FOR HIGH- DIMENSIONAL FEATURE DISCOVERY WITH APPLICATIONS IN GENOMIC ANALYSIS

Zhongyin J. Daye*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

Many biological processes are characterized with covariates that are structured. For example, genes in the same metabolic pathways or proteins that share similar functionalities often serve as covariates in genomic studies. Furthermore, the effects of biological processes often vary as functions depending upon some biological state, such as time. High-dimensional feature discovery where covariates are structured and the underlying model is nonparametric presents an important but largely unaddressed statistical challenge. In this talk, we propose the structured varying-coefficient estimator that can incorporate covariate structures for estimation and discovery of important features under high-dimensionality. We present an efficient algorithm for computing the structured varying-coefficient estimator. Finite-sample performances are studied via simulations, and the effects of high-dimensionality and structural information of covariates are especially highlighted. We further apply our method in a real-data application, in which we model transcription factor binding sites by incorporating structural information from DNA motif sequences. Our results demonstrate that our method is useful for high-dimensional feature discovery when the underlying model is nonparametric and covariates are structured.

email: zdaye@upenn.edu

AN INTEGRATIVE PATHWAY-BASED CLINICAL-GENOMIC MODEL FOR CANCER SURVIVAL PREDICTION

Xi Chen*, Vanderbilt University
Lily Wang, Vanderbilt University
Hemant Ishwaran, Cleveland Clinic

Although many prediction models that use gene expression levels have been proposed for personalized treatment of cancer, building accurate models that are easy to interpret remains a challenge. In this paper, we propose an integrative clinical-genomic approach that combines both genomic pathway and clinical information. First, we summarize information from genes in each pathway using Supervised Principal Components (SPCA) to obtain pathway-based genomic predictors. Next, we build a prediction model based on clinical variables and pathway-based genomic predictors using Random Survival Forests (RSF). Our rationale for this two-stage procedure is that the underlying disease process may be influenced by environmental exposure (measured by clinical variables) and perturbations in different pathways (measured by pathway-based genomic variables), as well as their interactions. Using two cancer microarray datasets, we show that the proposed pathway-based clinical-genomic model outperforms gene-based clinical-genomic models, with improved prediction accuracy and interpretability. Moreover, in addition to identifying important predictors for

predicting survival, the method also allows for identification of important interactions of predictors (pathway-pathway, pathway-clinical, or clinical-clinical).

email: steven.chen@vanderbilt.edu

ANALYSIS OF BIOLOGICAL PATHWAYS USING LAPLACIAN EIGENMAPS AND PENALIZED PRINCIPAL COMPONENT REGRESSION ON GRAPHS

Ali Shojaie*, University of Michigan
George Michailidis, University of Michigan

Gene, protein and metabolite networks provide valuable information on how components of biological systems interact with each other in order to carry out vital cell functions. The behavior of complex biological systems can only be understood by incorporating information on interactions among components of the system and by analyzing the effect of biological pathways, rather than individual components. In this paper, we propose a network-based approach for the analysis of significance of biological pathways using Laplacian eingenmaps. We establish a connection between Laplacian eigenmaps and principal components of the covariance matrix, and propose a dimension reduction method that directly incorporates the network information. Using this framework, the significance of biological pathways can then be analyzed by solving an eigenvalue problem with boundary conditions. We reformulate the problem of analysis of biological pathways as a principal regression problem on the graph, and use a group-lasso penalty to determine the significance of each subnetwork. The performance of the proposed method is evaluated using simulation studies as well as gene expression data in E-coli.

email: shojaie@umich.edu

MIXED EFFECTS COX MODELS FOR GENE SET ANALYSIS

Marianne Huebner*, Mayo Clinic
Terry Therneau, Mayo Clinic

Gene set analysis (GSA) methods take advantage of the fact that biological phenomena occur through interactions of multiple genes in functional relationships. Most GSA algorithms have been based on univariate test statistics that are aggregated for a gene set score (bottom-up), or combining gene expression levels within gene sets into a single covariate (top-down). For time-to-event data we propose to directly model the genes in a mixed effects Cox model. The gene set coefficients are treated as random effects with variance s and correlation r. As r-> 1 the coefficients are forced to be identical, mimicking the top-down model, ordinary shrinkage happens when r->0. In simulation results the approach is competitive with other methods, and has the advantage that other covariates fit naturally into the framework.

email: huebner.marianne@mayo.edu

NETWORK-BASED EMPIRICAL BAYES METHODS FOR LINEAR MODELS WITH APPLICATIONS TO GENOMIC DATA

Caiyan Li*, The U.S. Food and Drug Administration
Hongzhe Li, University of Pennsylvania
Zhi Wei, New Jersey Institute of Technology

Empirical Bayes methods are widely used in the analysis of microarray gene expression data in order to identify the differentially expressed genes or genes that are associated with other general phenotypes. Available methods often assume that genes are independent. However, genes are expected to function interactively and to form molecular modules to affect the phenotypes. In order to account for regulatory dependency among genes, we propose in this presentation a network-based empirical Bayes method for analyzing genomic data in the framework of linear models, where the dependency of genes is modeled by a discrete Markov random held defined on a pre-defined biological network. This method provides a statistical framework for integrating the known biological network information into the analysis of genomic data. Applications of the proposed methods in analysis of a human brain aging microarray gene expression data set and simulation studies will be presented.

email: caiyan.li@fda.hhs.gov

32. CAUSAL INFERENCE: METHODOLOGY

SENSITIVITY ANALYSIS FOR UNMEASURED CONFOUNDING IN PRINCIPAL STRATIFICATION

Scott Schwartz*, Duke University
Fan Li, Duke University
Jerry Reiter, Duke University

Intermediate variables are frequently required for correctly assessing treatment effects. Principal Stratification (PS) has become a standard framework to appropriately adjust for such intermediate variables. However, in the observational setting various types of confounding between treatment, intermediate variable and outcome can arise, threatening the conceptual and analytical validity of PS inference. Focusing on binary treatment and intermediate variable setting, we identify the various theoretical pathways of confounding present in the PS context as well as their implications for standard PS inference. We then represent these pathways as sensitivity parameters within a parametric model to allow for examination of result sensitivity to potential confounding scenarios. The methodology is validated using real data with introduced confounding and then applied to a medical example concerning the effects of influenza vaccination.

email: scott.schwartz@stat.duke.edu

INFERENCE FOR THE EFFECT OF TREATMENT ON SURVIVAL PROBABILITY IN RANDOMIZED TRIALS WITH NONCOMPLIANCE AND ADMINISTRATIVE CENSORING

Hui Nie*, University of Pennsylvania
Jing Cheng, University of Florida College of Medicine
Dylan S. Small, University of Pennsylvania

In many clinical studies with a survival outcome, follow-up ends at a pre-specified date when many subjects are still alive. This creates administrative censoring. An additional complication in some trials is that there is noncompliance with the assigned treatment. For this setting, we study the estimation of the causal effect on survival probability up to a given time point among those subjects who would comply with the assignment to both treatment and control. We first extend the standard instrumental variable method to survival outcomes. Then we propose the parametric maximum likelihood method under the Weibull distribution assumption. We further develop an efficient plug-in nonparametric empirical maximum likelihood estimation (PNEMLE) approach. Simulation studies show the efficiency gain of PNEMLE over the standard IV method. PNEMLE is applied to data from the HIP study to examine the effects of periodic screening on breast cancer mortality.

email: niehui@wharton.upenn.edu

BIAS ASSOCIATED WITH USING PROPENSITY SCORE AS A REGRESSION PREDICTOR

Bo Lu*, The Ohio State University
Erinn Hade, The Ohio State University

The use of propensity score methods to adjust for selection bias in observational studies has become increasingly popular in the public health and medical research. Our literature review shows that a substantial portion of studies using propensity score adjustment just treats propensity score as a regular regression predictor. We investigate the potential bias introduced by such adjustment with both a theoretical derivation under a simple parametric setup and an extensive simulation study comparing it to propensity score stratification, propensity score matching and regression adjustment after matching. Propensity score as a regression predictor may lead to serious bias under some common settings and regression adjustment after matching tend to produce the best overall result.

email: blu@cph.osu.edu

DOUBLY ROBUST INSTRUMENTAL VARIABLE ESTIMATION OF LATE(x)

Elizabeth L. Ogburn*, Harvard University
Andrea Rotnitzky, Harvard University
James Robins, Harvard University

Consider a study in which the effect of a binary treatment on a continuous outcome is confounded by variables which are unmeasured and therefore cannot be controlled for. An instrument is a variable Z that is related to the treatment but to neither unmeasured confounders or to the outcome (e.g. treatment assignment in a clinical trial with non-random non-compliance). Under certain assumptions instrumental variable methods give unbiased estimates of treatment effect where, due to unmeasured confounding, standard statistical methods cannot. In particular, under the monotonicity assumption that the instrument influences treatment in the same direction for all subjects (it may not affect treatment for some subjects), instrumental variable methods estimate the local average treatment effect (LATE), the effect of treatment on compliers (those subjects whose treatment is affected by the instrument). We present a new semiparametric method for estimating LATE for binary instruments in the presence of high dimensional covariates X, as a function of those covariates (LATE(x)). If Z is randomized or if f(Z|X) is known our method is guaranteed to be unbiased under the null hypothesis of no treatment effect among the compliers; if f(Z|X) is unknown our method is doubly robust.

email: ogburn@fas.harvard.edu

A POWERFUL AND ROBUST TEST STATISTIC FOR RANDOMIZATION INFERENCE IN GROUP-RANDOMIZED TRIALS

Kai Zhang*, University of Pennsylvania
Mikhail Traskin, University of Pennsylvania
Dylan Small, University of Pennsylvania

For group-randomized trials, randomization inference based on rank statistics provides robust, exact inference. However, in a matched pair design the currently available rank based statistics lose significant power compared to normal linear mixed model (LMM) test statistics when the LMM is true. In this research we investigate and develop the optimal test statistic over all statistics in the form of the weighted sum of signed Mann-Whitney-Wilcoxon statistics. This test is almost as powerful as the LMM even when the LMM is true, but much more powerful for heavy tailed distributions. Exact inference based on the statistics is considered and simulation is conducted to examine the power.

email: zhangk@wharton.upenn.edu

A MARKOV COMPLIANCE CLASSES AND OUTCOMES MODEL FOR CAUSAL ANALYSIS IN THE LONGITUDINAL STUDIES

Xin Gao*, University of Michigan
Michael R. Elliott, University of Michigan

We propose a Markov compliance classes and outcomes model for two-arm longitudinal randomized studies when noncompliance

is present. Under the potential outcome framework, our proposal model can provide causal effect of the treatment via principal stratification. Previous research (Lin, Ten Have, and Elliott, JASA 2007) considered the effect of subjects' joint compliance behavior on the joint distribution of the longitudinal outcomes, but not the impact of treatment effect and compliance behavior at time t-1 on the compliance behavior at time t. Our model can estimate the impact of the current treatment effect and compliance behavior on the future compliance behavior, as well as the treatment effect in each compliance class. We use data augmentation method for the unobservable variables and Markov Chain Monte Carlo algorithm for parameter estimation. Application of our model on the Suicide CBT study showed the significant effect of cognitive therapy on prevention of repeat suicide, and this effect increased as time increased. The results also showed the probability of compliance to the random assignment during the follow up period t increased when effect of cognitive therapy during the follow up period t-1 increased.

email: xingao@umich.edu

SEMIPARAMETRIC ESTIMATION OF CAUSAL MEDIATION EFFECTS IN RANDOMIZED TRIALS

Jing Zhang*, Brown University
Joseph Hogan, Brown University

In the context of randomized intervention trials of behavior science, such as those designed to increase physical activity or reduce substance abuse, the typical objective is to understand how the effect of an intervention operates on a primary outcome through potential mediating variables. Traditionally, mediation analysis is based on the Baron-Kenny method, a two-stage regression approach that requires randomization of the mediating variable within each intervention group. However, because mediating variables are observed after randomization, this assumption typically will not hold in practical settings. Moreover, when the mediator is reflects inherent characteristics of an individual (such as motivation to exercise), structural models formulated in terms of controlled direct and indirect effects may not be appropriate because they assume the mediator can be externally manipulated for each person. To address these shortcomings of existing methods for behavioral intervention trials, we propose three methods to estimate natural direct and indirect effects: inverse probability weighting (IPW), regression imputation (REG) and augmented inverse probability weighting (AIPW). We use baseline covariates to impute the unobserved potential mediator, along with a sensitivity parameter capturing the association between two potential mediators. Then, the unobserved potential primary outcome is treated as a missing value, whose expectation can be estimated by an observed outcome under some reasonable assumptions. We illustrate our methods in both simulation studies and an analysis of a recent intervention trial designed to increase physical activity.

email: jing_zhang@brown.edu

33. EPIDEMIOLOGICAL METHODS

ESTIMATION AND TESTING OF THE RELATIVE RISK OF DISEASE IN CASE CONTROL STUDIES WITH A SET OF k MATCHED CONTROLS PER CASE

Barry K. Moser*, Duke University Medical Center
Susan Halabi, Duke University Medical Center

Historically in case-control studies when a set of k-controls are matched to each case, an estimate of the odds ratio is produced. Under the rare disease assumption the odds ratio is equated with the more relevant parameter, the relative risk of disease. Therefore, the estimated odds ratio is typically used to estimate the relative risk of disease. The objective of this paper is to provide estimators for the relative risk of disease without making the rare disease assumption. To this end, algebraic processes are developed that produce parametric forms for the relative risk of disease, in case-control studies when a set of k-controls are matched to each case. One process is developed when the probability of exposure is constant for all cases (and constant for all controls). A more general process is then developed when the probability of exposure varies across cases and controls as a function of a set of covariates. Through these parametric forms estimators of the relative risk of disease are derived both when the probability of exposure is constant and when it varies. Closed form estimators of the variances of the estimators are then derived, along with confidence intervals, and hypothesis tests on the relative risk of disease. Through Monte Carlo simulation the new estimators of the relative risk of disease are shown to outperform estimators that assume the rare disease assumption.

email: moser004@mc.duke.edu

NESTED CASE-CONTROL ANALYSIS FOR OBSERVATIONAL DATA IN CARDIOVASCULAR DISEASE

Zugui Zhang*, Christiana Care Health System
Edward F. Ewen, Christiana Care Health System
Paul Kolm, Christiana Care Health System

Assessing cause and effect relationship in observational studies faces the issues of selection of controls and measurement of exposures before the occurrence of outcomes. Matching on potential confounding variables, the nested case-control study can be an efficient approach to solve these issues by combing the strengths of case-control study and cohort study designs. The purpose of this study was to apply nested case-control analysis to evaluate the association between cardiovascular (CV) risk factors and stroke in patients with atrial fibrillation or atrial flutter. Data of 841 patients were obtained from an electronic medical record encompassing office and hospital care. Fifty-five stroke cases were identified, and controls were selected for cases via incidence density sampling. Age and gender were the two primary matching factors, and length of

taking warfarin was the major exposure condition. Conditional logistic regression was applied to the matched data to estimate the association between occurrence of stroke and CV risk factors. Results are compared to those obtained from traditional case-control study and Cox proportional hazards regression.

email: ZZhang@ChristianaCare.org

SIMPLE ADJUSTMENTS TO REDUCE BIAS AND MEAN SQUARED ERROR ASSOCIATED WITH REGRESSION-BASED ODDS RATIO AND RELATIVE RISK ESTIMATORS

Robert H. Lyles*, Emory University
Ying Guo, Emory University

In most practical situations, maximum likelihood estimators of regression coefficients from logistic, Poisson, and Cox proportional hazards models are reasonably reliable in terms of their effective bias and approximate normal sampling distributions. However, a straightforward argument demonstrates that standard estimators of odds ratios and relative risks obtained by direct exponentiation are biased upward in an explicitly predictable way. This bias produces a propensity toward misleadingly large effect estimates in practice. We propose correction factors that apply in the same manner to each of these regression settings, such that the resulting estimators remain consistent and yield demonstrably reduced bias, variability, and mean squared error. Our initial estimator targets mean unbiasedness in the traditional sense, while the usual exponential transformation-based MLE is geared toward approximate median unbiasedness. We also propose a class of estimators that provide reduced mean bias and squared error, while allowing the investigator to control the risk of underestimating the true measure of effect. We discuss pros and cons of the usual estimator, and use simulation studies and real-data examples to compare its properties to those of the proposed alternatives.

email: rlyles@sph.emory.edu

MORTALITY MODEL FOR PROSTATE CANCER

Shih-Yuan Lee*, University of Michigan
Alexander Tsodikov, University of Michigan

Since the introduction of prostate cancer screening using the Prostate Specific Antigen (PSA), more than thirty percent prostate cancer mortality decline was observed. We propose a statistical model to assess and predict the effect of PSA screening on prostate cancer mortality in the United States. The model contains five major components. The marginal incidence model has age at prostate cancer diagnosis as an endpoint. Stage and grade (Z) specific incidence model predicts the probability of being diagnosed at a specific stage and grade at cancer incidence. The treatment model determines the probability of receiving a certain treatment combination at the time of cancer diagnosis. The disease progression model estimates the probability of disease progression after the screening diagnosis. Finally, the survival model calculates the survival time from the clinical diagnosis to death after adjusting for the lead time. The model was fit using Surveillance, Epidemiology and End Results (SEER) data and parameters were obtained using maximum likelihood methods. Age adjusted observed and predicted prostate cancer mortalities were compared.

email: shihylee@umich.edu

BINARY REGRESSION ANALYSIS WITH POOLED EXPOSURE MEASUREMENTS

Zhiwei Zhang*, Eunice Kennedy Shriver National Institute of Child Health of Human Development
Paul S. Albert, Eunice Kennedy Shriver National Institute of Child Health of Human Development

It has become increasingly common in epidemiologic studies to pool specimens across subjects in order to achieve accurate quantitation of biomarkers and certain environmental chemicals. In this paper, we consider the problem of fitting a binary regression model when an important exposure is subject to pooling. We take a regression calibration approach and derive several methods, including plug-in methods that use a pooled measurement and other covariate information to predict the exposure level of an individual subject, and normality-based methods that make further adjustments by assuming normality of calibration errors. Each type of methods can be implemented with different covariate configurations, two of which (i.e., covariate augmentation and imputation) are considered here. These methods are shown in simulation experiments to effectively reduce the bias associated with the naive method that simply substitutes a pooled measurement for all individual measurements in the pool. In particular, the normality-based imputation method performs reasonably well in a variety of settings, even under skewed distributions of calibration errors. The methods are illustrated using data from the Collaborative Perinatal Project.

email: zhiwei.zhang@nih.gov

ATTRIBUTABLE FRACTION FUNCTIONS FOR CENSORED EVENT TIMES

Li Chen*, University of North Carolina-Chapel Hill
Danyu Lin, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill

Attributable fractions are commonly used to measure the impact of risk on disease incidence in the population. These static measures can be extended to functions of time when the time to disease occurrence or event time is of interest. The present paper deals with nonparametric and semiparametric estimation of attributable fraction functions for cohort studies with potentially censored event time data. The semiparametric models include the familiar proportional hazards model and a broad class of transformation models. The proposed estimators are shown to be consistent,

asymptotically normal and asymptotically efficient. Extensive simulation studies demonstrate that the proposed methods perform well in practical situations. A cardiovascular health study is provided. Connection to casual inference is discussed.

email: lichen@email.unc.edu

A MULTIVARIATE NONLINEAR MEASUREMENT ERROR MODEL FOR EPISODICALLY CONSUMED FOODS

Saijuan Zhang*, Texas A&M University
Adriana Pérez, University of Texas Health Science Center at Houston
Victor Kipnis, National Cancer Institute
Laurence Freedman, Sheba Medical Center, Israel
Kevin Dodd, National Cancer Institute
Raymond J. Carroll, Texas A&M University
Douglas Midthune, National Cancer Institute

In the measurement error literature, there is a substantial work on corrections for attenuation and estimating the distribution of the unobserved latent variable when the data are necessarily continuous. However, our paper is based upon the observation that nutritional epidemiologists are also greatly interested in episodically consumed foods. Episodically consumed foods have zero-inflated skewed distributions. So-called two-part models have been developed for such data. However, in nutrition, along with amounts of a food, interest lies in the amount of an episodically consumed food adjusted for caloric intake. Hence, along with the episodically consumed food, models must account for energy (caloric) intake. We have recently developed such a model (Kipnis, et al., 2010), and have fit is using nonlinear mixed effects programs and methodology. There are technical challenges to this model because one of the covariance matrices is patterned having structural zeros. Such nonlinear mixed effects fitting is generally slow and there are times when the programs either fails to converge or converges to models with a singular covariance matrix. For these reasons we develop a Monte Carlo-based method of fitting this model, which allows for both frequentist and Bayesian inference. Our main application is to the NIH-AARP Diet and Health Study.

email: sjzhang@stat.tamu.edu

34. CATEGORICAL DATA ANALYSIS

KULLBACK LEIBLER RISK OF ESTIMATORS FOR UNIVARIATE DISCRETE EXPONENTIAL FAMILY DISTRIBUTIONS

Qiang Wu*, East Carolina University
Paul Vos, East Carolina University

For exponential families with discrete sample spaces, the Kullback Leibler (KL) risk of the maximum likelihood estimator is above 1/2 for much or all of the parameter space. We find estimators having corresponding risk that stays below 1/2 for much or all of the parameter space for the binomial, negative binomial, and poisson distributions. Since the KL risk can be defined without reference to parameterization, we require our estimators to be parameter invariant. For this reason, we define estimators that take values on the family of distributions without referring to the parameter. This construction allows us to decompose the KL risk in a fashion parallel to the decomposition of the mean squared error for real valued random variables. The decomposition consists of two nonnegative terms we call the KL-variance and the square of the KL-bias. Each of these is defined using the KL-mean which is a probability distribution. To choose among these estimators with a small KL risk we impose post-data restrictions: one restriction is on the median of the estimate and the other on its mode. Based on the KL risk and these restrictions we make recommendations for each of the exponential families considered.

email: wuq@ecu.edu

CORRELATED ORDINAL CATEGORICAL DATA ANALYSIS: COMPARING BRAUN-BLANQUET SEA GRASS COVERAGE ABUNDANCE SCORES

Nate Holt*, University of Florida
Mary Christman, University of Florida

Braun-Blanquet scoring is often used to assess sea grass cover and abundance. Common sampling strategies generate correlated ordinal categorical data. Procedures are considered that employ Dirichlet-multinomial models to compare Braun-Blanquet scores of sea grass cover abundance recorded by two different groups. This work proceeds that of Zhang and Boos (1997), who developed asymptotic tests for inference in similar problems. R functions written for this analysis will be discussed.

email: nateholt@ufl.edu

A BAYESIAN APPROACH FOR CORRECTING MISCLASSIFICATION IN BOTH OUTCOME VARIABLE AND COVARIATE

Sheng Luo*, University of Texas Health Science Center
Wenyaw Chan, University of Texas Health Science Center
Michelle Detry, University of Wisconsin-Madison

Misclassification occurring in either outcome variables or categorical covariates or both is a common issue in epidemiology. It leads to biased results and distorted disease-exposure relationship. A novel Bayesian approach is presented to address the misclassification in both outcome variables and covariates in logistic regression setting when neither gold standard nor prior knowledge

about the parameters exists. A simulated numerical example and a real clinical example are given to illustrate the proposed approach. The extension of the proposed approach to longitudinal setting is also discussed and proposed.

email: sheng.t.luo@uth.tmc.edu

ANALYSIS OF ZERO-INFLATED CLUSTERED COUNT DATA USING MARGINALIZED MODEL APPROACH

Keunbaik Lee*, Louisiana State University-New Orleans
Yongsuung Joo, Dongguk University, Korea
Joon Jin Song, University of Arkansas

Min and Agresti (2005) proposed random effect hurdle models for zero-inflated clustered count data with two-part random effects for a binary component and a truncated count component. In this talk, we propose new marginalized models for zero-inflated clustered count data using random effects. The marginalized models are similar to Dobbie and Welsh's (2001) model in which generalized estimating equations were exploited to find estimates. However, our proposed models are based on likelihood-based approach. Quasi-Newton algorithm is developed for estimation. We use these methods to carefully analyze two real datasets.

email: klee4@lsuhsc.edu

CONFIDENCE INTERVALS THAT MATCH FISHER'S EXACT OR BLAKER'S EXACT TESTS

Michael P. Fay*, National Institute of Allergy and Infectious Diseases

When analyzing a 2 by 2 table, the two-sided Fisher's exact test and the usual exact confidence interval (CI) for the odds ratio may give conflicting inferences; for example, the test rejects but the associated CI contains an odds ratio of 1. The problem is that the usual exact CI is the inversion of the test that rejects if either of the one-sided Fisher's exact tests rejects at half the nominal significance level. Further, the confidence set that is the inversion of the usual two-sided Fisher's exact test may not be an interval, so following Blaker (2000, Canadian Journal of Statistics, 783-798), we define the 'matching' interval as the smallest interval that contains the confidence set. We explore these two versions of Fisher's exact test as well as an exact test suggested by Blaker (2000) and discuss our R package (exact2x2) which automatically assigns the appropriate matching interval to the each of the three exact tests.

email: mfay@niaid.nih.gov

ON FINDING THE UPPER CONFIDENCE LIMIT FOR A BINOMIAL PROPORTION WHEN ZERO SUCCESSES ARE OBSERVED

Courtney Wimmer*, Medical College of Georgia

Confidence interval estimation for a binomial proportion is a long-debated topic, resulting in a wide range of exact and approximate methods. Many of these methods perform quite poorly when the number of observed successes in a sample of size n is zero. In this case, the main objective of the investigator is usually to obtain an upper bound, i.e., the upper limit of a one-sided confidence interval. Traditional notions of expected interval length and coverage probability are not applicable in this situation because it is assumed that the sample data have already been observed. In this paper we use observed interval length and p-confidence to evaluate nine methods for finding a confidence interval for a binomial proportion when it is known that the number of observed successes is zero. We also consider approximate sample sizes needed to achieve various upper bounds near the zero boundary. We show that many popular approximate methods perform poorly based on these criteria and conclude that the exact method has superior performance in terms of interval length and p-confidence.

email: slooney@mcg.edu

35. ADAPTIVE CLINICAL TRIAL DESIGN

HIERARCHICAL GAUSSIAN POWER PRIOR MODELS FOR ADAPTIVE INCORPORATION OF HISTORICAL INFORMATION IN CLINICAL TRIALS

Brian P. Hobbs*, University of Minnesota
Bradley P. Carlin, University of Minnesota
Daniel Sargent, Mayo Clinic
Sumithra Mandrekar, Mayo Clinic

Bayesian clinical trial designs offer the possibility of a substantially reduced sample size, increased statistical power, and reductions in cost and ethical hazard. However when prior and current information conflict, Bayesian methods can lead to higher than expected Type I error, as well as the possibility of a costlier and lengthier trial. This motivates an investigation of the feasibility of hierarchical Bayesian methods for incorporating historical data that are "adaptively robust" to prior knowledge that turns out to be inconsistent with the accumulating experimental data. In this paper, we present novel modifications to the traditional power prior approach for Gaussian data that allows the commensurability of the information in the historical and current data to determine how much historical information is used. We compare the frequentist performance of the various methods using simulation, and close with an example from the field of colon cancer that illustrates a linear models extension of our adaptive borrowing approach. This

design produces more precise estimates of the model parameters, in particular conferring statistical significance to the observed reduction in tumor size for the experimental regimen as compared to the control regimen.

email: hobbs040@umn.edu

EVALUATION OF VIABLE DYNAMIC TREATMENT REGIMES IN A SEQUENTIALLY RANDOMIZED TRIAL OF ADVANCED PROSTATE CANCER

Lu Wang*, University of Michigan
Peter Thall, University of Texas M.D. Anderson Cancer Center
Andrea Rotnitzky, Universidad Torcuato Di Tella
Xihong Lin, Harvard University
Randall Millikan, University of Texas M.D. Anderson Cancer Center

In this paper, we present new statistical analyses of data arising from a clinical trial designed to compare two-stage treatment strategies for advanced prostate cancer. The trial, conducted at M. D. Anderson Cancer Center from December 1998 to January 2006, was to mimic the way that oncologists actually behave when treating cancer patients. In this trial, the patients were randomized to four combination chemotherapies, denoted by the acronyms CVD, KA/VE, TEC and TEE, and later switched to a second different chemotherapies, based on their history of clinical outcomes. The goal of this paper is to compare 12 different sequential decision rules. We formally defined the dynamic treatment regimes that we compared, and defined the subjective, PI-specified, scoring function used to calculate the main endpoint of our analysis, as well as three additional endpoints. We discuss the inverse probability of treatment weighted methodology that we applied to estimate the mean overall score associated with each of the two-stage strategies. Our analyses also account for the possibility that drop-out may have been informative, in the sense of being explained by the history of recorded PSA values.

email: luwang@umich.edu

MEDIAN RESIDUAL LIFE TIME ESTIMATION IN SEQUENTIALLY RANDOMIZED TRIALS

Jin H. Ko*, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Adaptive treatment strategies are comprehensive methods for treating chronic diseases according to patients' needs and responses. Recently, sequentially randomized trials have drawn considerable attention as an effective way of comparing multiple treatment strategies. Analysis of data from such trials primarily focused on binary and survival outcomes. In survival analysis, it is often of interest to use median residual lifetime as the summary parameter to assess the treatment effectiveness. In this study, we propose methods for estimating strategy-specific median residual life function from a sequentially randomized trial. Three types of estimators are proposed by (i) inverting the inverse-probability-weighted estimated survival function, (ii) using the direct application of the mixture distribution, and (ii) using inverse-probability-weighted estimating equation function. We compare the three estimators through a simulation study. Our simulation study shows that (i) and (ii) produce approximately unbiased estimators in large samples. We demonstrate our methods by applying them to a sequentially randomized leukemia clinical trial data set.

email: jik16@pitt.edu

DESIGN OF DOSE-FINDING EXPERIMENTS WITH CORRELATED RESPONSES OF DIFFERENT TYPES

Valerii V. Fedorov, GlaxoSmithKline
Yuehui Wu, GlaxoSmithKline
Rongmei Zhang*, University of Pennsylvania

In dose-finding clinical studies, it is common that multiple endpoints are of interest. For instance, in phase I/II efficacy and toxicity are often the primary endpoints which need to be evaluated simultaneously. We discuss the dose-response model for categorical and continuous responses in which a latent multivariate normal distribution is used. We construct the benchmark locally optimal designs, and the more practical two-stage and adaptive designs. Various penalty functions are also considered to address ethical and economical concerns. While some theoretical results are derived, the main efforts are related to Monte Carlo simulations that allow us to analyze the properties of two-stage and adaptive designs.

email: yuehui.2.wu@gsk.com

ISSUES TO CONSIDER IN SELECTING A RESPONSE-ADAPTIVE DESIGN FOR DOSE-FINDING EXPERIMENTS

Nancy Flournoy*, University of Missouri-Columbia

Separately for design and analysis, one must choose between parametric and nonparametric, frequentist and Bayesian methods. Good choices depend on factors including the total sample size available and the experimental objectives. Prospects of significant toxicity bring multiple, competing objectives and dictate caution in the rules for transiting doses and/or suggest penalties be introduced. There is great heuristic appeal to using all the information obtained to date to select the next dose, but there are performance trade-offs between allocating subjects using long term memory versus short term memory procedures. Two stage designs at times are competitive with fully adaptive procedures. Procedures beckon for conditional power calculations, sample sizes

recalculation, and dropping (adding) treatments. We discuss these choices and reflect on how flexible an experiment should be.

email: flournoyn@missouri.edu

ESTIMATING THE DOSE-TOXICITY CURVE IN COMPLETED PHASE I STUDIES

Irina Ostrovnaya*, Memorial Sloan-Kettering Cancer Center
Alexia Iasonos, Memorial Sloan-Kettering Cancer Center

It has been shown previously that the MTD chosen by the 3+3 phase I design (standard method, SM) may be low, possibly leading to a non-efficacious dose. Additionally, when deviation from the original trial design occurs, the rules for determining MTD might be not applicable. We hypothesize that in these situations a retrospective analysis of toxicities from a completed trial should be used to determine or confirm the MTD. In this study we propose using constrained maximum likelihood estimation (CMLE) for these purposes. Such retrospective analysis might lead to at least as accurate or more accurate MTD than the one obtained by the SM. I will present a comparison of CMLE with the retrospective Continual Reassessment Method (O'Quigley 2005) in analyzing simulated SM trials as well as existing trials from Memorial Sloan Kettering Cancer Center. A framework for estimating confidence intervals around the toxicity probabilities at each dose level will also be presented.

email: ostrovni@mskcc.org

INFORMATION IN A SIMPLE ADAPTIVE OPTIMAL DESIGN

Ping Yao*, Northern Illinois University
Nancy Flournoy, University of Missouri-Columbia

This paper explores an important question concerning information as derived from sequentially implementing estimated optimal designs. Adaptive optimal designs are sequential experiments, each based on the optimal design estimated from the data obtained in all prior stages. The measure that is used in adaptive optimal designs to construct treatment allocation procedures is, by definition, neither the observed nor the expected (Fisher) information. We explore these three information measures in the context of a two-stage adaptive optimal design under a simple model. The simple model, taken to facilitate calculations, and hence understanding, is normal with mean the one parameter exponential function.

email: pyao@niu.edu

36. ROC ANALYSIS

BIOMARKER VALIDATION WITH AN IMPERFECT REFERENCE: BOUNDS AND ISSUES

Sarah C. Emerson*, Harvard University
Rebecca A. Betensky, Harvard University

Motivated by the goal of validating a newly developed marker for acute kidney injury, we consider the problem of assessing operating characteristics for a new biomarker when a true gold standard for disease status is unavailable. In this case, the new biomarker is typically compared to another imperfect reference test, and this comparison is used to estimate the performance of the new biomarker. However, errors made by the reference test can bias assessment of the new test. Analysis methods like latent class analysis have been proposed to address this issue, generally employing some strong and unverifiable assumptions regarding the relationship between the new biomarker and the reference test. We investigate the conditional independence assumption that is present in many such approaches, and demonstrate that this assumption may be violated even when the two tests are physiologically unrelated. We explore the information content of the comparison between the new biomarker and the reference test, and show that even if the operating characteristics of the reference test are known with certainty, it is often difficult to derive any useful information regarding the new test. In particular, we give bounds for the true sensitivity/specificity when operating characteristics for the reference test are known.

email: semerson@hsph.harvard.edu

LOGISTIC REGRESSION-BASED APPROACH TO BORROWING INFORMATION ACROSS COMMON-ROC POPULATIONS IN RISK PREDICTION

Ying Huang*, Columbia University
Ziding Feng, Fred Hutchinson Cancer Research Center

Characterizing the distribution of disease risk predicted by a biomarker is important for understanding the marker's capacity to stratify patients into different risk groups in a population of interest. In this research, we are interested in evaluating population-specific performance of a risk prediction marker with data from multiple populations when the marker's classification accuracy as characterized by the ROC curve is invariant across these populations. Instead of estimation using data from the target population only, we propose a logistic regression-based procedure to model disease risk based on standardized biomarker values using all available data. This procedure directly models the likelihood ratio and accommodates the common ROC assumption. The efficiency gain achieved by borrowing information across populations is demonstrated by simulation studies and in a real

dataset where PCA3 is evaluated as a risk prediction marker for prostate cancer among subjects with or without initial biopsy.

email: yh2441@columbia.edu

SUBJECT-SPECIFIC TYPE OF APPROACH IN FROC ANALYSIS

Andriy Bandos*, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
David Gur, University of Pittsburgh

Performance assessment analysis is an important part of evaluation of diagnostic systems in various fields. Free-response Receiver Operating Characteristic (FROC) analysis is a tool for the assessment of performance of a diagnostic system in a task of detection and localization of multiple targets per subject. A typical example of such a diagnostic task is the detection and localization of multiple abnormalities depicted on an image by a radiologist (with or without cues from an automated system). Traditionally FROC, as well as closely related clustered ROC analysis (e.g. ROI), addresses performance characteristics of a system from the population-averaged (or "marginal") perspective. However, in some applications, such as detection and localization of multiple nodules in diagnostic imaging, it may be more relevant to use a subject-specific approach which more closely reflects diagnostic accuracy for a subject randomly selected from the target population. Although population-averaged and subject-specific approaches frequently lead to similar conclusions, they may also lead to contradictory results (e.g. when detection rate varies together with the number of abnormalities). In this presentation we will describe a simple subject-specific type of approach to FROC analysis and discuss the differences from the conventional FROC methodology.

email: anb61@pitt.edu

CLASSIFICATION OF BINORMAL ROC CURVES WITH RESPECT TO IMPROPERNESS

Stephen L. Hillis*, Iowa City VA Medical Center
Kevin S. Berbaum, University of Iowa

The standard method for constructing an ROC curve is to use maximum likelihood estimation based on the assumption of a latent binormal model. However, a problem with the latent binormal model is that it produces an improper ROC curve in the sense that the ROC curve is not concave everywhere. This lack of concavity violates a basic assumption for a meaningful decision variable, that there is a monotone relationship between the decision variable and the likelihood of disease. In practice this is typically not a problem since the improperness is so small that it is not apparent when looking at the ROC curve. However, there are situations when the improperness is apparent, with the ROC curve visibly crossing below the chance line and having an obvious "hook." For these situations we deem the ROC curve to be "unacceptably improper." Presently, standard statistical software does not provide any diagnostics for assessing the magnitude of the improperness. We show how the mean-to-sigma ratio can be a useful, easy-to-understand, and easy-to-use diagnostic for detecting unacceptably improper binormal ROC curves by showing how it is related to the chance-line crossing. We suggest an improperness criteria based on the absolute value of the mean-to-sigma ratio.

email: steve-hillis@uiowa.edu

NONPARAMETRIC ESTIMATION OF TIME-DEPENDENT PREDICTIVE ACCURACY CURVE

Paramita Saha*, National Institute of Environmental Health Sciences
Patrick J. Heagerty, University of Washington

A major biomedical goal for developing predictive survival model is to accurately distinguish between incident cases at t from the subjects surviving beyond t. Extensions of standard binary classification measures like time-dependent True & False Positives & ROC curve have become popular in this context. AUC curve has been introduced as a measure of discrimination of the marker throughout the entire study period. However, the existing AUC curve estimators are not estimated directly; they are derived from a semi-parametric estimate of ROC curve via numerical integration. We propose a direct, nonparametric estimator of the time-dependent AUC curve forgoing the step of estimating the ROC curve first. The proposed method extends nonparametric AUC estimator arising from a binary data context & possesses desirable asymptotic properties. An overall measure of concordance (similar to Harrell's C-index) is proposed. Time-dependent marker can also be accommodated. We also show that marker comparison can also be done via this approach.

email: sahap@niehs.nih.gov

OPTIMAL COMBINATIONS OF DIAGNOSTIC TESTS BASED ON AUC

Xin Huang*, Georgia State University
Yixin Fang, Georgia State University
Gengsheng Qin, Georgia State University

When several diagnostic tests are available, one can combine them to achieve better diagnostic accuracy. This paper considers the optimal linear combination that maximizes the area under the receiver operating characteristic curve (AUC); the estimates of the combination's coefficients can be obtained via a non-parametric procedure. However, for estimating the AUC associated with the estimated coefficients, the apparent estimation by re-substitution is too optimistic. To adjust for the upward bias, several methods are proposed. Among them the cross-valiation approach is especially advocated, and an approximated cross-validation is developed to reduce the computational cost. Furthermore, these proposed methods can be applied for variable selection to select important

diagnostic tests. The proposed methods are examined through simulation studies and applications to three real examples.

email: xhuang4@student.gsu.edu

37. COMPARATIVE EFFECTIVENESS RESEARCH: THE METHODOLOGIC CHALLENGES

COMPARATIVE EFFECTIVENESS RESEARCH: PROMISES AND CHALLENGES

Kathleen N. Lohr*, RTI International

Comparative Effectiveness Research: Promises and Challenges Comparative effectiveness research compares clinical and patient outcomes, effectiveness, and appropriateness of services and procedures that clinicians can use to prevent, diagnose, or treat all types of diseases and health conditions. Supporters cite numerous benefits: e.g., showing what is the right care at the right time for patients; supporting informed decision making; contributing to personalized medicine; assisting efforts to control health care costs; and providing information to help decision makers know whether health care services bring value and are worth the costs. Nonetheless, various criticisms and challenges remain. Some relate to policy, such as concerns about rationing or impact on innovation. Others involve methods questions, such as choosing appropriate study designs, applying the right statistical methods, and addressing issues of clinical heterogeneity (e.g., whether populations studied cover an adequate range of the patients whom clinicians typically see). Syntheses of CE research into comparative effectiveness reviews poses yet other methods challenges, such as rating quality of individual studies and grading strength of bodies of evidence. This presentation will introduce key perspectives on this emerging field of investigation.

email: klohr@rti.org

COMPARATIVE EFFECTIVENESS RESEARCH: THE ROLE OF RESEARCH SYNTHESIS

Joel B. Greenhouse*, Carnegie Mellon University

The synthesis of evidence comparing different health interventions has been identified by the IOM and others as a core methodology for comparative effectiveness research. More than meta-analysis which is the synthesis of information from similar studies, research synthesis is defined as the use of multiple data sources, including randomized clinical trials, observational studies, and administrative databases, to learn about what works for whom and under what conditions. In this talk we discuss different approaches to research synthesis and consider the strengths and weaknesses of evidence generated from these types of studies.

email: joel@stat.cmu.edu

COMPARATIVE EFFECTIVENESS OF HIP REPLACEMENT SYSTEMS

Sharon-Lise T. Normand*, Harvard Medical School and Harvard School of Public Health
Danica Marinac-Dabic, Center for Devices and Radiological Health, U.S. Food and Drug Administration
Art Sedrakyan, Center for Devices and Radiological Health, U.S. Food and Drug Administration

Randomized controlled trials that may serve as the basis for device approval can be small, short-term, and generalizable to an increasingly smaller percentage of patients. Some of the most common and challenging devices are those used in hip replacement. With an aging US population and increasing obesity, the incidence of hip replacement will increase despite little information on long term outcomes. In this talk, we propose a statistical framework for integrating post-market information with pre-market information via hierarchical generalized linear models in order to provide an enhanced understanding of the comparative effectiveness of different hip systems. Our approach capitalizes on methods for cross-design synthesis, meta-analysis, and network meta-analysis. Our key assumption is that device performance characteristics and outcomes obtained from one cohort are related to device performance characteristics and outcomes of the same device or similar devices observed in other cohorts.

e-mail: sharon@hcp.med.harvard.edu

HOW TO COMPARE THE EFFECTIVENESS OF HYPOTHETICAL INTERVENTIONS (HINT: FIRST SPECIFY THE INTERVENTIONS)

Miguel A. Hernan*, Harvard University

Observational studies are often used to make inferences regarding the comparative effectiveness of clinical interventions. Most discussions about the relative advantages and disadvantages of observational studies, compared with randomized clinical trials, have focused on potential confounding bias arising from lack of randomization. Interestingly, these discussions usually ignore the fact that conventional analyses of observational studies and randomized clinical trials may answer different clinical questions. In some cases, the question implicitly asked in observational studies is not clinically relevant or even well-defined, which makes it difficult to assess the validity and implications of a particular analysis of observational data. This problem is more severe when estimating effects from complex longitudinal data with time-varying treatments. I will discuss how the absence of a well-defined clinical question affects causal analyses of commonly used pharmacological treatments and coronary heart disease.

e-mail: mhernan@hsph.harvard.edu

38. ADVANCES/CHALLENGES IN JOINTLY MODELING MULTIVARIATE LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA

SAMPLE SIZE AND POWER DETERMINATION IN JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA

Joseph G. Ibrahim*, University of North Carolina-Chapel Hill
Liddy Chen, University of North Carolina-Chapel Hill
Haitao Chu, University of North Carolina-Chapel Hill

Due to the rapid development of biomarkers in clinical trials, joint modeling of survival and longitudinal data has gained its popularity in recent years because it reduces bias and provides improvements of efficiency in the assessment of treatment effects and other prognostic factors. Statistical design, such as sample size and power calculations, is a crucial first step in clinical trials. Although much effort has been put into inferential methods in joint modeling, such as estimation and hypothesis testing, design aspects have not been formally considered. We derive a closed form sample size formula for estimating the effect of the longitudinal process in joint modeling, and extend Schoenfeld's (1983) sample size formula to the joint modeling setting for estimating the overall treatment effect. We discuss the impact of the within subject variability on power, and data collection strategies, such as spacing and frequency of repeated measurements, in order to maximize power. We also show that a small number of measurements can lead to a biased estimate of the longitudinal effect and result in a significant loss of power.

email: ibrahim@bios.unc.edu

PREDICTING RENAL GRAFT FAILURE USING MULTIVARIATE LONGITUDINAL PROFILES

Geert Verbeke*, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium
Steffen Fieuws, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium

In many medical studies repeatedly measured biomarker information is gathered together with a time to an event, e.g. occurrence of a disease. In such situations, the biomarker information serves as a health indicator representing the progression of the disease, and can therefore be used to predict the event of interest. The application motivating this presentation considers patients who received a kidney transplant and who are intensively monitored during the years after the transplant. It is of interest to predict graft failure from chronic rejection or recurrent disease within 10 years after the transplantation based on longitudinal information about serum creatinine, urine proteinuria, mean of systolic and diastolic blood pressure, and blood haematocrit level. Our aim is to construct a model that allows prediction of graft failure based on all repeated measurements of all 4 markers. Furthermore, it is of interest to investigate how the multivariate information outperforms the information in each marker separately, when it comes to predicting the event of interest. The proposed combines linear, generalised linear and nonlinear mixed models into one multivariate longitudinal model by specifying a joint distribution for all random effects. Due to the high number of markers, a pairwise model fitting approach, where all possible pairs of bivariate mixed models are fitted, is used.

email: geert.verbeke@med.kuleuven.be

AN ANALYSIS FOR JOINTLY MODELING MULITVARIATE LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA

Paul S. Albert*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Joanna H. Shih, National Cancer Institute

In many medical studies, patients are followed longitudinally and interest is on assessing the relationship between longitudinal measurements and time to an event. Recently, various authors have proposed joint modeling approaches for longitudinal and time-to-event data for a single longitudinal variable. These approaches become intractable for even a few longitudinal variables. We propose a two-stage regression calibration approach which appropriately accounts for informative dropout in the longitudinal measurements. Specifically, we approximate the conditional distribution of the multiple longitudinal variables given the event time by modeling all pairwise combinations of the longitudinal measurements using a bivariate linear mixed model which conditions on the event time. Complete data are then simulated based on estimates from these pairwise conditional models, and regression calibration is used to estimate the relationship between longitudinal data and time-to-event data using the complete data. We illustrate this methodology with simulations and with an analysis of primary biliary cirrhosis (PBC) data.

email: albertp@mail.nih.gov

39. STATISTICAL MODELS AND PRACTICE THAT IMPROVE REPRODUCIBILITY IN GENOMICS RESEARCH

THE IMPORTANCE OF REPRODUCIBILITY IN HIGH-THROUGHPUT BIOLOGY: SOME CASE STUDIES

Keith A. Baggerly*, University of Texas M.D. Anderson Cancer Center
Kevin R. Coombes, University of Texas M.D. Anderson Cancer Center

Over the past few years, microarray experiments have supplied much information about the disregulation of biological pathways associated with various types of cancer. Many studies focus on

identifying subgroups of patients with particularly agressive forms of disease, so that we know who to treat. A corresponding question is how to treat them. Given the treatment options available today, this means trying to predict which chemotherapeutic regimens will be most effective. Several microarray studies have provided such predictions. Unfortunately, ambiguities associated with analyzing the data have made many of these results difficult to reproduce. In this talk, we will describe how we have analyzed the data, and reconstructed aspects of the analysis from the reported results. In some cases, these reconstructions reveal inadvertent flaws that affect the results. Most of these flaws are simple in nature, but their simplicity is obscured by a lack of documentation. We briefly discuss the implications of such ambiguities for clinical findings. We will also describe approaches we now follow for making such analyses more reproducible, so that progress can be made more steadily.

email: kabagg@mdanderson.org

NORMALIZATION USING NEGATIVE CONTROL FEATURES

Zhijin Wu*, Brown University
Yunxia Sui, Brown University

Normalization has been recognized as a necessary preprocessing step in a variety of high throughput biotechnologies. A number of normalization methods have been developed specifically for microarrays, some general and others tailored for certain experimental designs. All methods rely on assumptions on what values are expected to stay constant across samples. Most assume some quantities related to the specific signal stay the same, which is usually not verifiable. Some recent platforms of short oligonucleotide arrays include a large number of negative control probes that cover a great range of the measured intensity. We present normalization methods that utilize the negative controls and show decreased variation among technical replicates while maintaining biological variation. We also demonstrate the normalization across platform using common negative control probes.

email: zhijin_wu@brown.edu

STATISTICAL REPRODUCIBILITY IN CLINICAL GENOMICS

Jeffrey T. Leek*, Johns Hopkins University
John D. Storey, Princeton University

There is a growing interest in reproducibility in the analysis of genomic data, particularly because of increased clinical applications of genomics. There are many types of reproducibility, ranging from replicating the numbers in a published paper to observing a significant result in multiple different experiments. I will introduce the concept of statistical reproducibility in genomics experiments and the sources of variation that may lead to statistical irreproducibility. Statistical reproducibility has consequences for both the significance and biological conclusions of a high-throughput study. I will illustrate the benefits of improving statistical reproducibility with data from a large collaborative study of the genomic response to trauma.

email: jleek@jhsph.edu

40. DYNAMIC TREATMENT REGIMES AND REINFORCEMENT LEARNING IN CLINICAL TRIALS

MODEL-CHECKING FOR SEMIPARAMETRIC ESTIMATION OF OPTIMAL DYNAMIC TREATMENT REGIMES

Erica Moodie*, McGill University
Benjamin Rich, McGill University
David A. Stephens, McGill University

To estimate the sequence of actions that optimizes response in a longitudinal setting, it is important to study the actions as a set of decision rules rather than as single-action comparisons. There are many statistical challenges that arise in estimation of dynamic regimes and semiparametric methods have found favour in recent years although little attention has been paid to model adequacy. We propose residual diagnostic plots for semiparametric estimation of optimal dynamic treatment regimes, and consider the utility of such approaches in the face of non-regularity. We are motivated by the estimation of decision rules for the duration of breastfeeding with a view to optimizing infant growth. Breastfeeding has many well-recognized health benefits, although the long-term consequences for stature and adiposity remain controversial. The Promotion of Breastfeeding Intervention Trial (PROBIT) recruited 17,046 women in which Belarus who were randomized to a breastfeeding promotion intervention or to standard care. There are many covariates to consider in this trial and previous work has shown no effect of breastfeeding from 9-12 months of age, indicating non-regularity.

email: erica.moodie@mcgill.ca

INFERENCE FOR NON-REGULAR PARAMETERS IN OPTIMAL DYNAMIC TREATMENT REGIMES

Bibhas Chakraborty*, Columbia University
Susan A. Murphy, University of Michigan
Victor J. Strecher, University of Michigan

Dynamic treatment regimes are individually tailored treatments. They offer a way to operationalize the adaptive multistage decision making in clinical practice, thus providing an opportunity to improve such decision making. However, when using longitudinal data on patients to construct these treatment regimes, hypotheses concerning the choice of the optimal treatment at each stage may

involve non-regular parameters. The non-regularity stems from the fact that parameters of interest are functions of maxima. As a result, the parameter estimates can be biased, and traditional methods of constructing confidence intervals can have poor frequentist properties. In this paper, we present and evaluate a method that adapts to this non-regularity by the use of a thresholding (empirical Bayes) approach, and compare with other available approaches. Analysis of data from a two-stage smoking cessation trial is presented as an illustration.

email: bc2425@columbia.edu

REINFORCEMENT LEARNING STRATEGIES FOR CLINICAL TRIALS IN NON-SMALL CELL LUNG CANCER

Yufan Zhao*, Amgen Inc.
Michael R. Kosorok, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill
Mark A. Socinski, University of North Carolina-Chapel Hill

We present a reinforcement learning design to discover optimal individualized treatment regimens for a non-small cell lung cancer trial. In addition to the complexity of the problem of selecting optimal compounds for first and second-line treatments based on prognostic factors, another primary scientific goal is to determine the optimal time to initiate second-line therapy, either immediately or delayed after induction therapy, yielding the longest overall survival time. Q-learning is utilized and approximating the Q-function with time-indexed parameters can be achieved by using support vector regressions. A simulation study shows that the procedure not only successfully identifies optimal strategies of two lines treatment from clinical data, but also reliably selects the best time to initial second-line therapy while taking into account heterogeneities of NSCLC across patients.

email: yufanzz@gmail.com

41. MULTIVARIATE ANALYSIS

CONTINUITY AND ANALYSIS OF PRINCIPAL COMPONENTS

Ahmad Reza Soltani*, Kuwait University
Fatemah Alqallaf, Kuwait University
Noriah Alkandari, Kuwait University

Double arrays of n rows and p columns can be regarded as n drawings from some p-dimensional population. A sequence of such arrays is considered. Principal component analysis for each array forms sequences of sample principal components and eigenvalues. The continuity of these sequences, in the sense of convergence with probability one and convergence in probability, is investigated, that appears to be informative for pattern study and prediction of principal components. This allows to study the turbulence in sequences in repeated measurements, and to measure the contributions of the factors to the PC components along certain duration. Applications to real bio-data as repeated measurements and longitudinal data are given and discussed.

email: soltani@kuc01.kuniv.edu.kw

CONVERGENCE AND PREDICTION OF PRINCIPAL COMPONENT SCORES IN HIGH-DIMENSIONAL SETTINGS

Seunggeun Lee*, University of North Carolina-Chapel Hill
Fei Zou, University of North Carolina-Chapel Hill
Fred A. Wright, University of North Carolina-Chapel Hill

A number of settings arise in which it is of interest to predict Principal Component (PC) scores for new observations using data from an initial sample. In this paper, we demonstrate that naive approaches to PC score prediction can be substantially biased towards 0 in the analysis of large matrices. This phenomenon is largely related to known inconsistency results for sample eigenvalues and eigenvectors as both dimensions of the matrix increase. For the spike eigenvalue model for random matrices, we expand the generality of these results, and propose bias-adjusted PC score prediction. In addition, we compute the asymptotic correlation coefficient between the population and sample PC scores. Simulation and real data examples from the genetics literature show the improved bias and numerical properties of our estimators.

email: slee@bios.unc.edu

SUFFICIENT DIMENSION REDUCTION IN REGRESSION AND APPLICATIONS TO SNP DATASETS

Kofi P. Adragni*, University of Alabama-Birmingham

With technological advances, high throughput datasets are becoming frequent and will certainly be more prevalent in the future. Researchers are now formulating regressions with thousands of predictors. Several methods have been proposed to deal with such regressions. To help reduce the dimensionality of such massive datasets without altering the regression information, dimension reduction methods can be applied. Several dimension reduction methods are found in the literature, but few can handle large datasets where the number of predictors p (genes) is larger than the number of observations n. Principal fitted components (PFC) models, developed by Cook in 2007, are to yield the sufficient reduction. PFC models assume that the predictors are random and are not hindered by large dimensionality when the number of predictors exceeds greatly the number of observations. In this presentation, we give an overview of PFC models for sufficient dimension reduction and also introduce a novel prediction method

based on PFC models in large p regressions. We present an application to a SNP dataset.

email: kadragni@ms.soph.uab.edu

BETWEEN ESTIMATOR IN THE INTRACLASS CORRELATION MODEL WITH MISSING DATA

Mixia Wu, Beijing University of Technology
Kai F. Yu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development

The between estimator for the intraclass correlation model with missing data in longitudinal studies is investigated. A necessary and sufficient condition is given for the existing exact simultaneous confidence intervals for all contrasts in the means under the between transformed model. This clarifies the validity of the simultaneous confidence intervals for all contrasts in the means of the intraclass correlation model with missing data in the literature. The true distribution of the between estimator can be derived as a result. In addition, the exact test statistic and confidence intervals for partial contrasts can be constructed.

email: yukf@mail.nih.gov

DISTRIBUTION-FREE TESTS OF MEAN VECTORS AND COVARIANCE MATRICES FOR MULTIVARIATE PAIRED DATA

Erning Li*, Texas A&M University
Johan Lim, Seoul National University, Korea
Kyunga Kim, Seoul National University, Korea
Shin-Jae Lee, Seoul National University, Korea

We study a permutation procedure to test the equality of mean vectors, homogeneity of covariance matrices, or simultaneous equality of both mean vectors and covariance matrices in multivariate paired data. We propose to use two separate test statistics for the equality of mean vectors and the homogeneity of covariance matrices, respectively, and combine them to test the simultaneous equality of both mean vectors and covariance matrices. Since the combined test has composite null hypothesis, we control its type I error probability and theoretically prove the unbiasedness and consistency of the combined test. The new procedure requires neither distributional assumption on the data nor structural assumption on the covariances. We illustrate the good performance of the proposed approach with comparison to competing methods via simulations. We apply the proposed method to testing the symmetry of tooth size in a dental study and to finding differentially expressed gene sets with dependent structures in a microarray study of prostate cancer.

email: eli@stat.tamu.edu

INFERENCE FOR FACTOR ANALYSIS WITH LARGE p AND SMALL n: UNDERSTANDING THE CHANGE PATTERNS OF THE US CANCER MORTALITY RATES

Miguel Marino*, Harvard University
Yi Li, Harvard University

Cancer researchers track cancer mortality rate trends and study the cross relationship of these trends not only for scientific reasons of understanding cancer as a complex dynamical system, but also for practical reasons such as planning and resource allocation. Factor analysis which studies these matrices is an effective means of data reduction, which typically requires the number of random variables, p, to be relatively small and fixed, and the sample size, n, to be approaching infinity. However, contemporary surveillance techniques have yielded large matrices in both dimensions, limiting existing factor analysis techniques due to the poor estimate of the correlation matrix. We develop methods, in the framework of random matrix theory, to study the cross-correlation of cancer mortality annual rate changes in the setting where $p > n$. We propose to use the Tracy-Widom test to test complete independence across cancer sites. We develop methodology based on group sequential theory to determine the number of significant factors in a factor model. Sparse principal components analysis is studied on the principal components deemed to be significantly different than random matrix theory prediction to aid in the interpretation of the underlying factors. Methods are implemented on SEER cancer mortality rates from 1969-2005.

email: mmarino@hsph.harvard.edu

USE OF FACTOR ANALYSIS IN MEDICAL EDUCATION

Jay Mandrekar*, Mayo Clinic

An area of research that has received much attention in the literature is the evaluation of faculty in clinical teaching. These evaluations are used to make critical decisions regarding faculty promotion, and improvement in the quality of instruction, and patient care. This data is often collected in the form of questionnaires, with individual item quantification using Likert scales. The focus of this presentation is to provide brief illustrative examples of projects with applications from medical education utilizing factor analysis as one of the primary analytical tool. Specifically, 1) to determine whether the factorial structure of clinical teaching assessments remained stable among medical specialties and 2) to determine the validity of an instrument for assessing residents' reflection on quality improvement opportunities.

email: mandrekar.jay@mayo.edu

42. BIOPHARMACEUTICAL METHODS

AN EXTENDED F-TEST FOR BIOSIMILARITY OF VARIABILITY TO ASSESS FOLLOW-ON BIOLOGICS

Jun Yang*, Amgen, Inc.
Nan Zhang, Amgen, Inc.
Shein-Chung Chow, Duke University
Eric Chi, Amgen, Inc.

As more biologic products are going off patent protection, the development of follow-on biologic products (biosimilars) has received much more attention from both biotechnology industry and the regulatory agencies. Unlike small molecule drug products, the development of biologic products is very different and variable to the manufacture process and environment. Thus, Chow et al. (2009) suggested that the assessment of biosimilarity between biologic products should be conducted based on variability in addition to biosimilarity in average of endpoints of interest. They also recommended that a more stringent probability-based criterion, which is shown to be sensitive to a small change in variability, should be employed. In this article, alternatively, we extend the traditional F-test for homogeneity of variances to evaluation of biosimilarity between biologic products. Extensive simulation studies are conducted to compare relative performance of the proposed method with the probability-based method proposed by Hsieh et al. (2009) for assessment of biosimilarity in variability of follow-on biologics in terms of consistency/ inconsistency for correctly concluding biosimilarity in variability.

email: juny@amgen.com

BIOEQUIVALENCE ANALYSES FOR REPLICATED CROSSOVERS: STRUCTURED COVARIANCE?

Donna L. Kowalski*, Astellas Pharma
Devan V. Mehrotra, Merck & Co., Inc.

Replicated crossover designs (e.g., TRR, RTT) are sometimes used to demonstrate average bioequivalence of a test (T) and reference (R) treatment. Standard pharmacokinetic summary measures from such designs (e.g., log(AUC)) are commonly analyzed using a linear mixed effects model described in a 2001 FDA guidance document on bioequivalence analyses. The document recommends a factor analytic variance-covariance structure for the vector of within-subject responses (TYPE=FA0(2) in SAS PROC MIXED terminology), and explicitly notes that use of an unstructured covariance (TYPE=UN) should be avoided. In this talk, we take a closer look at the FDA guidance, and use theoretical arguments along with simulation results to support our preference for using an unstructured covariance.

email: donna.kowalski@us.astellas.com

STOCHASTIC DIFFERENTIAL EQUATIONS WITH POSITIVE SOLUTIONS IN MODELING AND DESIGN OF PHARMACOKINETIC STUDIES

Valerii Fedorov, GlaxoSmithKline
Sergei Leonov*, GlaxoSmithKline
Vyacheslav Vasiliev, Tomsk State University

In compartmental pharmacokinetic (PK) modeling, ordinary differential equations are traditionally used with two sources of randomness, measurement error and population variability. Over the last few years a number of papers were published which addressed the intrinsic variability of stochastic systems modelled via stochastic differential equations (SDE). In most of these publications insufficient attention was given to the fact that the trajectories of such stochastic systems may become negative with positive probability. The latter is counterintuitive from physiological perspective. We explore the SDE models with positive solutions and consider the optimal design problem, i.e. finding sequences of sampling times in PK studies that provide the most precise estimation of unknown model parameters. Examples of optimal designs are provided including those which maximize information under cost constraints.

email: Sergei.2.Leonov@gsk.com

A LIKELIHOOD BASED METHOD FOR SIGNAL DETECTION IN SAFETY SURVEILLANCE WITH APPLICATION TO FDA'S DRUG SAFETY DATA

Lan Huang*, U.S. Food and Drug Administration
Jyoti Zalkikar, U.S. Food and Drug Administration
Ram Tiwari, U.S. Food and Drug Administration

Post-market drug safety surveillance has become very important in recent decade. In order to monitor the safety issues of the drugs on the market after approval, FDA collects Adverse Event Reporting System (AERS) data including adverse events (AEs) reported by patients, health care providers, and other sources through a spontaneous reporting system. Computational and statistical methods that are available in literature to systematically identify drug-event combinations with disproportionately high frequencies in large safety database including AERS database, are subject to high false discovery rates and some methods use ad-hoc thresholds for signal detection. Here, we propose a method, based on the likelihood ratio test (LRT) theory, to analyze the AERS data for identifying drug-event combinations with unusually high reporting rates. We conduct an extensive simulation study to evaluate and compare the performance of the proposed method with some existing methods using operating characteristics such as power, type I error, sensitivity, and false discovery rate. We illustrate the application of the proposed method using a dataset for Singulair consisting of suicidal behavior and mood change-related AE cases reported to FDA during 2004-2007 and a dataset for Heparin

consisting of all possible AE cases reported to FDA during 2004-2008.

email: lannecessary@gmail.com

GROUP SEQUENTIAL METHODS FOR OBSERVATIONAL DATA INCORPORATING CONFOUNDING THROUGH ESTIMATING EQUATIONS WITH APPLICATION IN POST-MARKETING VACCINE/DRUG SURVEILLANCE

Andrea J. Cook*, University of Washington
Jennifer C. Nelson, University of Washington
Ram C. Tiwari, U.S.Food and Drug Administration

Conducting observational post-marketing vaccine and drug safety surveillance is important for detecting rare adverse events not identified pre-licensure. The data that is currently collected for this type of surveillance is prospective observational data that is updated as often as weekly when new subjects are exposed (i.e. vaccinated). We propose a new statistical method for surveillance utilizing estimating equations in a group sequential monitoring framework to account for confounding with limited model assumptions. Current methods account for confounding through matching or stratification and therefore are constrained in their availability to truly to control for confounding compared to using a regression approach for adjustment. Further, the method proposed uses estimating equations and therefore can easily handle outcomes of any time type including binary, count, and continuous. The method is also able to handle aggregate data through incorporation of weights, which is important since individual level data may be unavailable in surveillance studies due to data confidentiality issues. A simulation study will be presented to evaluate the performance of the method. The proposed method will then be applied to a dataset from the vaccine safety datalink (VSD) evaluating the relationship between the vaccine MMRV and several adverse events.

email: cook.aj@ghc.org

IMPROVED ANALYSIS OF 2x2 CROSSOVER TRIALS WITH POTENTIALLY MISSING DATA

Devan V. Mehrotra*, Merck Research Laboratories
Yu Ding,
Yang Liu,
John Palcza,

The two period, two treatment crossover design is often used to compare within-subject responses to a test (T) and a reference (R) treatment. The resulting data are commonly analyzed using a standard linear mixed effects model with fixed effect terms for treatment, period and sequence, and random effect terms for subject and residual error. With no missing values, the standard analysis generally delivers acceptable results. However, when some values are missing, the performance of the standard analysis can sometimes be adversely affected in a non-trivial manner. In this talk, we (i) explain when and why the standard analysis breaks down, and quantify the resulting detrimental impact on type 1 error rate, power, and parameter estimates using simulations, and (ii) propose a simple alternate method of analysis with desirable properties. Our results are applicable regardless of whether the goal is to establish that the population mean responses for T and R are different (superiority trials) or similar (equivalence trials).

email: devan_mehrotra@merck.com

ON THE RELATIONSHIP BETWEEN THE DISTRIBUTION OF BATCH MEANS AND THE DISTRIBUTION OF BATCH SHELF LIVES IN ESTIMATING PRODUCT SHELF LIFE

Michelle Quinlan*, University of Nebraska-Lincoln
Walt Stroup, University of Nebraska-Lincoln
Dave Christopher, Schering-Plough Corporation
James Schwenke, Boehringer Ingelheim Pharmaceuticals, Inc.

ICH Q1E prescribes estimating shelf life using a 95% confidence interval for the batch mean of a stability limiting characteristic, treating batches as fixed effects. Implicit in the ICH prescription is definitions of batch shelf life and product shelf life. Using these definitions, one can focus on batch means, or alternatively on the distribution of shelf lives. Following ICH, change in batch mean over time can be modeled as b0 + b1t, where t denotes time. The resulting batch shelf life is (A-b0)/b1, where A denotes acceptance criterion. The distribution of batch means on the y-axis projects to a distribution of batch shelf lives on the x-axis. Assuming b0, b1 have a multivariate normal distribution, shelf life is the ratio of two correlated Gaussian variables. Using Hinkley (1969), we describe the relationship between quantiles of the distributions of batch shelf lives (x-axis) and means (y-axis). Exploiting this relationship, a mixed model (batches random) can be used to estimate a target quantile of batch shelf lives, the target being a suitably small quantile consistent with ICH. The distinction between the distribution of batch mean and shelf life are discussed. Simulation results are presented.

email: mquinlan22@yahoo.com

43. CLINICAL TRIALS WITH ADAPTIVE SAMPLE SIZES

APPLICATION OF GROUP SEQUENTIAL METHODS WITH RESPONSE ADAPTIVE RANDOMIZATION DESIGN FOR COMPARING THE TREATMENT EFFECT WITH BINARY OUTCOMES: AN EVALUATION OF BAYESIAN DECISION THEORY APPROACH

Fei Jiang*, University of Texas M.D. Anderson Cancer Center
J. Jack Lee, University of Texas M.D. Anderson Cancer Center
Peter Mueller, University of Texas M.D. Anderson Cancer Center

Group sequential methods and response adaptive randomization (RAR) procedure have been applied in clinical trials due to economical and ethical considerations. Group sequential methods are able to reduce the average sample size by inducing early stopping. RAR procedure allocates more patients to better arm, however it requires more sample size to obtain a certain power. This study intends to combine these two procedures. We apply the Bayesian decision theory approach to define our group sequential stopping rules and evaluate the operating characteristics under RAR setting. The results show that Bayesian decision theory method is able to preserve the type I error rate as well as achieve a favorable power. In addition, compared with the error spending function method, Bayesian decision theory approach is more effective on reducing average sample size.

email: homebovine@hotmail.com

ADJUSTMENT OF PATIENT RECRUITMENT IN THE BAYESIAN SETTING

Frank V. Mannino*, GlaxoSmithKline
Valerii Fedorov, GlaxoSmithKline
Darryl Downing, GlaxoSmithKline

Clinical trials are a vital, but expensive part of the process of bringing new drugs to the market. Poor designs can lead to substantial losses when, for example, recruitment is slower than expected leading to the reduced revenues. Stochastic models of the recruitment process allow us to build the predictive distribution of the time needed to recruit a desired number of patients. This allows us to construct risk functions which can be minimized with respect to various enrollment scenarios. Sponsors frequently push to open fewer centers in order to reduce their costs without consideration of the effects on the length of the trial, which can increase total losses. Using the risk function that includes costs of enrollment and potential losses due to delays we optimize the number of centers from the start of a trial and select the best decision rules to update a trial as interim information becomes available.

email: frank.v.mannino@gsk.com

BAYESIAN ADAPTIVE DESIGNS FOR PHASE III CARDIOVASCULAR SAFETY

Jason T. Connor*, Berry Consultants
Scott M. Berry, Berry Consultants

In the wake of rofecoxib and rosiglitazone, FDA now requires pharmaceutical companies to demonstrate the cardiovascular safety of many new drugs (e.g. diabetes drugs) before approval. The power of such study designs is highly dependent upon the event rate a CV event rate that is likely rare and unknown for each different non-CV application. We have designed multiple large pre-market, randomized, double-blind, controlled trials testing the hypotheses that the new drugs do not increase the likelihood of CV events. We use an adaptive Bayesian design with a parametric survival model and pre-scheduled unblinded interim looks to select the sample size. Accrual stops when the current sample produces a high predictive probability that once all enrolled patients have two years follow-up the Pr(risk ratio < R or risk difference < D) > 0.95 or stops the trial for futility if the probability of success if enrolling to the maximum trial size is < 0.05. The design maintains high power over a broad range of equivalent event rates and minimizes sample size, study cost and duration compared to defined fixed number of unblinded events.

email: jason@berryconsultants.com

ADAPTIVE INCREASE IN SAMPLE SIZE WHEN INTERIM RESULTS ARE PROMISING

Cyrus R. Mehta*, President, Cytel Inc.
Stuart J. Pocock, London School of Hygiene and Tropical Medicine

In major clinical trials there is considerable interest in adaptive sample size re-estimation using an unblinded interim estimate of the primary effect size. However, most existing statistical methods entail the inconvenience and complexity of not permitting conventional p-values and estimation at the final analysis. In the spirit of making adaptive increases in trial size more widely appealing and readily implementable, we here define those promising circumstances in which conventional final inference can be performed while preserving the overal type-1 error. We illustrate with examples of real clinical trials in which the methods have been applied, backed up by simulation results.

email: mehta@cytel.com

DESIGN OF SEQUENTIAL PROBABILITY LIKELIHOOD RATIO TEST METHODOLOGY FOR POISSON GLMM WITH APPLICATIONS TO MULTICENTER RANDOMIZED CLINICAL TRIALS

Judy X. Li*, University of California-Riverside
Daniel R. Jeske, University of California-Riverside
Jeffrey A. Klein, University of California-Irvine

Sequential analyses in clinical trials have ethical and economic advantages over fixed sample size methods. We consider the problem of sequential hypothesis testing with data based on a generalized linear model and a generalized linear mixed model. Both Wald SPRT and Bartlett SPRT are discussed. A new extended Bartlett SPRT method is proposed and compared to a fixed sample size test and the Wald SPRT. The methodology is illustrated in the context of a multi-center randomized clinical trial that compares two preventive treatments for surgical site infections.

email: xiangli0422@hotmail.com

PREDICTING NUMBER OF EVENTS IN CLINICAL TRIALS WHEN TREATMENT ARMS ARE MASKED

Yu Gu*, Florida State University
Liqiang Yang, Pfizer Inc.
Ke Zhang, Pfizer Inc.
Debajyoti Sinha, Florida State University

In randomized clinical trials, interim analyses are often scheduled for both efficacy and safety reasons. For clinical trials with time-to-event outcomes, the statistical information collected is strongly related to the number of events occurred. Early and accurate prediction of number of events will help timeline planning and reduce the waste of sources that are involved with interim analyses. We propose frequentist and Bayesian approaches using EM algorithm to estimate the distribution parameters based on the actual enrollment and event data, which does not require the unblinding of the treatment assignment. Prediction of the number of events at any future time point during the study is obtained through the estimation of the parameters. Simulation results are given to show that the prediction is more accurate with more data being collected and asymptotically unbiased.

email: yugu@stat.fsu.edu

OPEN-SOURCE SIMULATION EXPERIMENT PLATFORM FOR EVALUATING CLINICAL TRIAL DESIGNS

Yuanyuan Wang*, University of Pittsburgh
Roger S. Day, University of Pittsburgh

Trial simulation is used by pharmaceutical companies to improve the efficiency and accuracy of drug development. Sophisticated commercial software for trial simulations is available for those with resources to cover fees and with design challenges that happen to match the software's capabilities. Academic research centers usually use locally developed or shared software for study design, mainly due to cost and flexibility considerations. Inspired by the success of open-source software development projects, we are building an open-source simulation experiment platform with the intention of utilizing the power of distributed study design expertise, development talent, and code peer review. The code base relies on S4 classes and methods within R. Four key classes define the specifications of the population models, clinical trial designs, outcome models and evaluation criteria. Five key methods define the interfaces for generating patient baseline characteristics, applying a stopping rule, assigning treatments, generating patient outcomes and calculating criteria. Documentation of their connections with the user input screens, with the central simulation loop, and with each other will facilitate extensibility. To illustrate the application, we evaluate the effect of patient pharmacokinetic heterogeneity on the robustness of common Phase I designs.

email: ywangpitt@gmail.com

44. CAUSAL INFERENCE: METHODOLOGY AND APPLICATIONS

VALIDATION OF SURROGATE OUTCOMES USING A CAUSAL INFERENCE FRAMEWORK

Andreas G. Klein*, University of Western Ontario

This paper proposes a surrogacy measure that is based on potential outcome notation. Individual-level surrogacy is defined as an association between individual causal effects of a treatment on an intermediate variable and a clinical outcome. The concept that a surrogate marker is an indicator of an unobserved intermediate causal process is given a formal representation. The approach is compared to Prentice's concept of surrogacy and other known methods that assess subgroup-level surrogacy. A procedure is presented which - under certain assumptions about the structure of the causal process - produces bound estimates for the proposed surrogacy measure. The application of the new procedure is illustrated by an empirical example.

email: aklein25@uwo.ca

ENTIRE MATCHING WITH FINE BALANCE AND ITS APPLICATION IN PSYCHIATRY

Frank B. Yoon*, Harvard Medical School

Matching with fine balance creates treated and control samples with the same marginal distributions of discrete covariates, without needing to match exactly on those covariates. In the standard approach each treated subject is matched to a constant number of controls, which is constrained by the size of the available control pool and its covariate distribution. As a better alternative entire matching with fine balance uses a flexible match ratio, called the entire number, that is determined by using propensity scores in a new way. The entire number: (1) permits fine balance; (2) stochastically balances other continuous covariates; and (3) minimizes the standard error of the estimated treatment effect. While pair matching with fine balance is in many situations possible, it does not utilize the maximum sample size; on the other hand, a larger matching ratio might not permit fine balance. Entire matching with fine balance uses the maximum sample size while permitting fine balance, so that it outperforms standard methods. This claim is demonstrated theoretically and by simulation. In an application to a study of apsychotic treatments for schizophrenia, entire matching with fine balance is illustrated along with analysis of multiple endpoints through the matched design.

email: yoon@hcp.med.harvard.edu

CAUSAL SURROGACY ASSESSMENT IN A META ANALYSIS OF COLORECTAL CLINICAL TRIALS

Yun Li*, University of Michigan
Jeremy MG Taylor, University of Michigan
Michael R. Elliott, University of Michigan
Bhramar Mukherjee, University of Michigan

When the true endpoints (T) are difficult or costly to measure, surrogate markers (S) are often collected in clinical trials to help predict the treatment effect (Z). There is a vast interest in understanding the relationship among S, T and Z. Traditional models have been used to assess surrogacy; however, these models often condition on a post-randomization variable S which may cause bias. A counterfactual-based principal stratification concept has been proposed by Frangakis and Rubin (2000) to study their causal associations. In this paper, we propose a Bayesian estimation method to assess surrogacy in a multiple trial setting and obtain the trial-specific and overall causal associations among S, T and Z in this counterfactual framework. The method allows information sharing across trials and improves the estimation precision. All S, T and Z are binary. We extend our method to the settings with and without the monotonicity assumption. We assess the method using simulations. We then apply it to two colon cancer examples and evaluate the goodness-of-fit of the models.

email: yunlisph@umich.edu

CHALLENGES IN EVALUATING THE EFFICACY OF A MALARIA VACCINE

Dylan S. Small*, University of Pennsylvania
Jing Cheng, University of Florida
Thomas R. Ten Have, University of Pennsylvania

An effective vaccine against malaria is actively being sought. We formulate a potential outcomes definition of the efficacy of a malaria vaccine for preventing fever. A challenge in estimating this efficacy is that there is no sure way to determine whether a fever was caused by malaria. We study the properties of two approaches for estimating efficacy: (1) use a deterministic case definition of a malaria caused fever as a fever plus a parasite density above a certain cutoff; (2) use a probabilistic case definition in which we estimate the probability that each fever was caused by malaria. We show in a simulation study that both approaches potentially have large bias.

email: dsmall@wharton.upenn.edu

CROSS-TIME MATCHING IN AN OBSERVATIONAL STUDY OF SMOKING CESSATION

Bo Lu, The Ohio State University
Chih-Lin Li*, The Ohio State University

In observational studies, because of the lack of randomization, the estimated effect may be potentially confounded by the pretreatment differences. Matching is a popular method for removing biases in observed covariates. However, most matching designs focus on observational studies at one time point. In this talk, we introduce cross-time matching design for observational studies repeated at multiple time points. This research is motivated by an observational study of comparing efficacy of drug plus counseling and counseling only strategies on smoking cessation in Italy. The smoking cessation program with the same study protocol was conducted every year from 2001 to 2006. Early 2005, a public smoking ban was enacted in Italy. The main research question is the impact of the smoking ban on the effectiveness of the smoking cessation program. We re-structure the data to a four-group set up, pre-/post-ban and treatment/control. We propose a four-group matching design to balance the covariate distribution for all treatment and time combination. A more robust difference-in-difference type of estimator is used in post-matching analysis. We also propose two matching algorithms for unordered four-group matching and the simulation studies show that the suboptimal algorithm produces matched sets with smaller total distance than the nearest neighbor algorithm in most of the scenarios.

email: li.698@buckeyemail.osu.edu

G-COMPUTATION ALGORITHM FOR SMOKING CESSATION DATA

Shira I. Dunsiger*, Centers for Behavioral and Preventative Medicine, The Miriam Hospital
Joseph W. Hogan, Brown University
Bess H. Marcus, Brown University

Smoking cessation trials often collect longitudinal records on both cessation status and compliance with assigned treatment. Since many of these studies are subject to non-compliance, it may be of interest to estimate both the effectiveness and the efficacy of the intervention. The latter poses challenges to the researcher since compliance is a post-randomization variable.In this paper, we describe the G-Computation Algorithm (GCA) for estimating the effect of receiving treatment and illustrate it with a detailed hypothetical example. Next, we apply the GCA to weekly data from a smoking cessation trial (Commit to Quit), in which participants were randomized to a smoking cessation program plus either exercise or contact control (wellness). Results suggest that if participants had been fully compliant with assigned treatment, the difference in weekly quit rates between the exercise and wellness arms would have ranged from 0.09-0.19. We compare the GCA estimates and those from the intent to treat analyses and highlight specific advantages of the GCA for use in smoking cessation trials.

email: shira@stat.brown.edu

BUILDING A STRONGER INSTRUMENT IN AN OBSERVATIONAL STUDY OF PERINATAL CARE FOR PREMATURE INFANTS

Mike Baiocchi*, University of Pennsylvania
Dylan Small, University of Pennsylvania
Scott Lorch, University of Pennsylvania
Paul Rosenbaum, University of Pennsylvania

An instrument is a haphazard push towards acceptance of a treatment which affects outcomes only to the extent that it affects acceptance of the treatment. In settings in which treatment assignment is mostly deliberate and not random, there may nonetheless exist some essentially random pushes to accept treatment, so that use of an instrument may extract bits of random treatment assignment from a setting that is otherwise quite biased in its treatment assignments. An instrument is weak if the haphazard pushes barely influence treatment assignment or strong if they are often decisive in influencing treatment assignment. Although one hopes that an ostensibly haphazard instrument is perfectly random and not biased, it is not possible to be certain of this, so a typical concern is that even the instrument is biased to some degree. It is known from theoretical arguments that weak instruments are invariably sensitive to extremely small biases, so for this reason, strong instruments are preferred. The strength of an instrument is often taken as a given. It is not. In an evaluation of effects of perinatal care on the mortality of premature infants, we show that it is possible to build a stronger instrument, we show how to do it, and we show that success in this task is critically important. Also, we develop methods of permutation inference for effect ratios, a key component in an instrumental variable analysis.

email: mike.baiocchi@gmail.com

45. SURVIVAL ANALYSIS METHODOLOGY

LANDMARK PREDICTION OF SURVIVAL

Layla Parast*, Harvard University
Tianxi Cai, Harvard University

Recent advancement in technology has lead to a wide range of genetic and biological markers that hold great potential in improving the prediction of survival outcomes. It is of clinical interest to construct an optimal prognostic score when there are multiple such markers available for prediction. Although such new classifiers promise better disease prognosis, the accuracy in identifying short term and long term survivors remains unsatisfactory for most complex diseases. It has often been argued that short term clinical outcomes may have potential in predicting long term outcomes. In this paper, we develop conditional prognostic rules for the prediction of long term outcomes based on baseline marker information along with the event status at an earlier landmark time point. When there are multiple markers available, we construct an optimal composite score by fitting a proportional hazards working model for the conditional survival distribution. The accuracy of the score was evaluated non-parametrically based on inverse probability weighting. Resampling procedures were proposed to derive estimation procedures for the accuracy measures. Numerical studies suggest that the proposed procedures perform well in finite samples.

email: lparast@hsph.harvard.edu

PARAMETER ESTIMATIONS FOR GENERALIZED EXPONENTIAL DISTRIBUTION UNDER PROGRESSIVE TYPE-I INTERVAL CENSORING

Din Chen*, Georgia Southern University
Yuhlong Lio, University of South Dakota

In this talk, the estimations of parameters in generalized exponential distribution based on a progressively type-I interval censored sample are studied. The maximum likelihood estimation, the moment of method and the estimation based on the probability plot are developed. Through a computer simulation, these methods are compared in terms of mean squared errors and biases, respectively. Finally, these methods are applied to a real data which contains patients with plasma cell myeloma for illustration.

email: dchen@georgiasouthern.edu

PARTLY PROPORTIONAL SINGLE-INDEX MODEL FOR CENSORED SURVIVAL DATA

Kai Ding*, University of North Carolina-Chapel Hill
Michael R. Kosorok, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill
David B. Richardson, University of North Carolina-Chapel Hill

In this paper, we propose a partly proportional single-index model for censored survival data. This model is an extension of the Cox model and allows flexible semiparametric modeling of covariate effects in a parsimonious way via a single-index. We consider two commonly used profile likelihood methods for parameter estimation. One is based on the local likelihood and the other is based on stratification. We show that both commonly used approaches lead to biased estimation. A bias correction is then proposed for each method and the corrected profile likelihood estimators are shown to be consistent. We evaluate the finite-sample properties of our estimators through simulation studies and illustrate the proposed methods with an application to the Mayo PBC data.

email: kding@bios.unc.edu

STOCHASTIC FRAILTY MODEL INDUCED BY TIME DEPENDENT COVARIATES

Lyrica Xiaohong Liu*, University of Michigan
Alex Tsodikov, University of Michigan
Susan Murray, University of Michigan

Frailty model is an extension of Cox model when hazards of population demonstrate heterogeneity. Most research treat frailty as some random variable, the underlying assumption is that frailty, though unobserved, is a fixed quantity over time. However, sometimes latent frailty might function as a stochastic process due to the nature of disease incidence, for example, tumor growth, or, due to a dynamic treatment assignment. In both cases, the frailty process changes its distribution characteristics as time goes by. Traditional frailty modeling approach will not be adequate for those types of scenario. In this research, we propose a frailty process model induced by time changing covariates within one subject. We establish the properties of estimates using counting process and Martingale related theories. Finally we apply our method to prostate cancer data from SEER (Surveillance, Epidemiology, and End Results, http://seer.cancer.gov) database.

email: lyrica@umich.edu

46. STATISTICAL GENETICS: COMPLEX DISEASES AND LINKAGE ANALYSIS

RANDOM FORESTS AND MARS FOR DETECTING SNP-SNP INTERACTIONS IN COMPLEX DISEASES

Lin Hui-Yi*, Moffitt Cancer Center & Research Institute

A growing number of evidences show that single nucleotide polymorphisms (SNP) interactions are more important than single genetic factor in complex diseases. Several data mining methods have been proposed to analyze high-dimensional SNP data. In this study, we evaluated two data mining methods for detecting SNPs interactions in complex diseases, such as asthma and cancer. Multivariate Adaptive Regression Splines (MARS) combines the advantages of recursive partitioning and spline fitting. MARS can effectively reduce the number of terms in a model by automatically categorizing a three-level categorical SNP into different inherited modes and detecting specific interaction patterns. Random Forests (RF) method is a collection of classification trees grown on bootstrap samples. RF generates variable importance measures, which take into account interactions among variables. A simulation study was conducted to compare the performance of MARS and RF in detecting SNP-SNP interactions. Four hundred subjects (200 cases and 200 controls) were generated with non-missing 10 or 100 candidate SNPs. Two 2-way interaction models and one 3-way interaction model were evaluated. Results showed that RF is a powerful tool for screening candidate SNPs. Combing two data methods (RF and MARS) may be a good approach for detect SNP-SNP interactions in complex diseases.

email: hui-yi.lin@moffitt.org

THEORETICAL BASIS FOR HAPLOTYPING COMPLEX DISEASES

Li Zhang*, Cleveland Clinic
Jiangtao Luo, University of Florida
Rongling Wu, Penn State University

To determine specific DNA sequences directly associated with the phenotypic variation of disease risk, Liu et al. (2004) developed a model based on the haplotype structure of the genome in the maximum-likelihood context, implemented with the EM algorithm. We theoretically verify the EM steps, and prove and obtain the asymptotic distributions of the log-likelihood ratio test statistics of the corresponding genetic tests. The related work provides the theoretical basis for Liu et al. model, and investigates its statistical properties and validate its usefulness. We performed simulation studies to numerically evaluate conclusion with validation by a worked example, in which a DNA sequence variant is detected to significantly reduce human obesity.

email: mlizhang@gmail.com

BAYESIAN VARIABLE SELECTION WITH BIOLOGICAL PRIOR INFORMATION

Deukwoo Kwon*, National Cancer Institute

Complex diseases are functionally caused by a combination of environmental and genetic factors. Epidemiologists are posed with the problem of determining the specific causes and combinations of risk factors. With inexpensive genotyping technology available, in genetic association studies we analyze several thousands single nucleotide polymorphisms for each individual. Due to the large numbers of predictors available in genetic studies, we need to use variable selection techniques to decide which effects to include in a model that relates risk factors to phenotypic outcomes. Therefore, we present a hierarchical Bayesian variable selection method, which is an extension of the stochastic search variable selection. We introduce two latent binary vectors in a hierarchical manner to model the relationship between genes and SNPs and use generalized linear models to relate them to a phenotype. When biological pathways information is available we need to incorporate this prior information into variable selection method.

email: kwonde@mail.nih.gov

ASSESSING STATISTICAL SIGNIFICANCE IN GENETIC LINKAGE ANALYSIS WITH THE VARIANCE COMPONENTS MODEL

Gengxin Li*, Michigan State University
Yuehua Cui, Michigan State University

Variance components (VC) analysis has been a standard means in genetic linkage analysis. When a QTL has pleiotropic effect on several phenotypic traits, multivariate approaches in genetic linkage analysis can increase the power and precision to identify genetic effects. The VC technique treats genetic effects as random, and tests whether variance terms are zero using the likelihood ratio test (LRT). In the literature, the asymptotic distribution of the LRT is claimed to follow a mixture chi-square distribution, where the mixture proportions are calculated with standard binomial coefficients, a special case in Self and Liang (1987). This threshold calculation, however, often yields conservative hypothesis tests as discussed in a number of studies, especially in multivariate traits cases. In this work we show that the chi-square mixture proportions depend on the estimated Fisher information matrix in both univariate and multivariate trait analysis, and provide a general approximation form of the LRT under the null hypothesis of no linkage. We illustrate our idea with three commonly used VC models in genetic linkage analysis. The superiority of the new threshold calculation method is demonstrated by simulation studies.

email: ligengxi@stt.msu.edu

ASSESSING GENETIC ASSOCIATION IN CASE-CONTROL STUDIES WITH UNMEASURED POPULATION STRUCTURE

Yong Chen*, Johns Hopkins University
Kung-Yee Liang, Johns Hopkins University
Terri H. Beaty, Johns Hopkins University
Kathleen C. Barnes, Johns Hopkins University

The case-control study design is one of the main tools for detecting associations between genetic markers and disease. It is well known population structure (PS) can lead to spurious association between disease status and a genetic marker if the prevalence of disease and the marker allele frequency vary across subpopulations. In this paper, we proposed a novel statistical method to estimate the association in case-control studies with potential population structure. The proposed method takes two steps. First, the information on genomic markers and disease status is used to infer population structure; second, the association between disease and any one marker adjusting for the population structure is modeled and estimated parametrically through polytomous logistic regression. The performance of the proposed method, relative to others, on bias, coverage probability and computational time, is assessed through simulations. Finally, the method is applied to an asthma study in an African Americans population.

email: yonchen@jhsph.edu

A DATA-ADAPTIVE SUM TEST FOR DISEASE ASSOCIATION WITH MULTIPLE COMMON OR RARE VARIANTS

Fang Han*, University of Minnesota
Wei Pan, University of Minnesota

Given typically weak associations between complex diseases and common variants, and emerging approaches to genotyping rare variants (e.g. by next-generation se-quencing), there is an urgent demand to develop powerful association tests that are applicable to detecting disease association with both common and rare variants. In this article we develop such a test. It is based on a data-adaptive modification to a so-called Sum test originally proposed for common variants, which aims to strike a balance in utilizing information in multiple correlated variants while reducing the cost of large degrees of freedom (DF) or of multiple testing adjustment. The proposed test is easy to use with DF=1 and no need to adjust for multiple testing. We show that the proposed test has high power across a wide range of scenarios with either or both of common and rare variants. In particular, under some situations the proposed test performs better than several commonly used methods.

email: fanghan@biostat.umn.edu

EPISTATIC INTERACTIONS

Tyler J. VanderWeele*, Harvard University

The term epistasis is sometimes used to describe some form of statistical interaction between genetic factors and is alternatively sometimes used to describe instances in which the effect of a particular genetic variant is masked by a variant at another locus. Heather Cordell has argued that statistical tests for interaction are of limited use in detecting epistasis in the sense of masking. The paper shows there are relations between empirical data patterns and epistasis that have not been previously noted. These relations give rise non-standard interaction tests and can sometimes be exploited to empirically test for epistasis in the sense of the masking of the effect of a particular genetic variant by a variant at another locus. The paper discusses how these tests can be employed in case-control, case-only and cohort study designs. The results clarify when interaction tests can be interpreted as evidence for epistasis in the biologic sense of masking and show how tests for epistasis in sense of masking can be conducted even when standard interaction tests do not have this interpretation.

email: tvanderw@hsph.harvard.edu

47. FUNCTIONAL DATA ANALYSIS

ROBUST FUNCTIONAL MIXED MODELS

Hongxiao Zhu*, University of Texas M.D. Anderson Cancer Center
Philip J. Brown, University of Kent, Canterbury, UK.
Jeffrey S. Morris, University of Texas M.D. Anderson Cancer Center

The Bayesian wavelet-based functional mixed models (WFMM) proposed by Morris (2006) provide a flexible way to analyze complex functional data. The Gaussian assumptions for the priors and likelihood of WFMM, however, can be inadequate in many real applications. In this paper we present an improved robust functional mixed model (R-FMM) to accommodating heavier-than-normal tailed distributions. The robustness is achieved by adopting scale mixtures of normal, which embraces a large scope of heavier tailed distributions, and is computationally efficient for posterior sampling. The model is presented under the general framework of isomorphic basis-space (IBS) approach, with special focus on discrete wavelet transform. We design a simulation study based on a real mass spectrometry dataset and compared the performance of the proposed R-FMM with WFMM. Simulation results show that for data with heavier tailed random effect and error distributions, the R-FMM can dramatically improves the estimation precision and provides more adaptive regularization than WFMM. By controlling expected Bayesian false discovery rate (FDR), posterior inferences can also be conducted to flag out significant regions for factors of interested. The R-WFMM is finally applied to the real mass spectrometry dataset and the results are compared with that of WFMM.

email: hzhu1@mdanderson.org

MULTILEVEL FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS FOR SPARSELY SAMPLED HIERARCHICAL CURVES

Chongzhi Di*, Fred Hutchinson Cancer Research Center
Ciprian M. Crainiceanu, Johns Hopkins University

Sparsely sampled curves were traditionally viewed as longitudinal data and analyzed using generalized estimating equations or mixed effects models. Recently, it is becoming popular to view this type of data as sparse observations of underlying smooth functions, and utilize functional approaches. In this paper, we consider statistical analysis of sparsely sampled hierarchical functions or curves. Existing approaches based on FPCA do not work in this case because of the multilevel structure among the curves. Instead, we adapt the recently proposed multilevel functional principal component analysis (MFPCA; Di et al. 2009) to this setting, and discuss statistical issues on estimation, inference and prediction while accounting for sparsity. The MFPCA method extracts dominating modes of variations at both between and within subject levels, and summarizes each curve by two sets of principal component scores. It is a nonparametric functional approach, and does not rely on parametric assumptions on the shapes of curves. This approach is illustrated by applications to sleep studies.

email: cdi@fhcrc.org

FUNCTIONAL DATA ANALYSIS VIA MULTIPLE PRINCIPLE COMPONENTS VARIABLES

Andrew Redd*, Texas A&M University

Functional principle component analysis is a valuable assent for statisticians when dealing with irregular and sparse longitudinal or functional data. I propose an additive model for functional data analysis, which extends the previous work of Zhou et. al. (2008) and James et, al. (2001) into a useful analysis tool as well as an efficient data reduction technique. The motivating example for the method comes from an experiment measuring the effect of toxins on calcium ion signals.

email: aredd@stat.tamu.edu

SEMIPARAMETRIC BAYES MULTIVARIATE FUNCTIONAL DATA CLUSTERING WITH VARIABLE SELECTION

Yeonseung Chung*, Harvard University
Brent Coull, Harvard University

This research proposes a semiparametric Bayes multivariate functional data analysis methodology. The proposed method models the multivariate functional trajectories using a basis function expansion model with the basis coefficients taken as random effects. To resolve the high-dimensionality of the random effects, dimension reduction is applied using a factor analysis taking into account the correlations both among different basis coefficients and across multivariate outcomes. The trajectories are then associated with potential predictors by modeling the factor distribution as nonparametric distributions flexibly changing across the predictors. The model essentially relies on the probit stick-breaking process (PSBP) mixture for the random effects distribution which relaxes a single Gaussian assumption and linearity. The PSBP mixture also allows for selecting the predictors for random effects and correspondingly for multivariate trajectories. Posterior computation relies on Gibbs sampling with a stochastic search variable selection (SSVS) algorithm. The methods are illustrated through simulations and applied to a particulate matter (PM) animal study conducted by Environmental Health researchers in Harvard School of Public Health.

email: ychung@hsph.harvard.edu

CLUSTERING ANALYSIS OF fMRI TIME SERIES USING WAVELETS

Cheolwoo Park*, University of Georgia
Jinae Lee, University of Georgia
Benjamin Austin, University of Georgia
Kara Dyckman, University of Georgia
Qingyang Li, University of Georgia
Jennifer McDowell, University of Georgia
Nicole A. Lazar, University of Georgia

In functional Magnetic Resonance Imaging (fMRI) studies clustering methods are used to detect similarities in the activation time series among voxels. It is assumed that the temporal pattern of activation is organized in a spatially coherent fashion such that clustering will extract the main temporal patterns and partition the dataset by grouping similarly behaved functions together. In this work fMRI data are acquired on two occasions while participants are engaged in saccade tasks (anti-saccade, pro-saccade, and fixation) from 37 undergraduate women. We attempt to aggregate voxel time series into a small number of clusters and compare the clustered maps for the three practice groups and between the two scan time points. Prior to clustering analysis we perform wavelet transformation to decorrelate the temporal dependence. In addition, we apply the adaptive pivot test based on wavelets to exclude voxels that are considered as pure noises. We cluster the wavelet coefficients of the remaining voxels using principal component analysis K-means clustering. The resulting clustered maps are compared using ANOVA analysis.

email: cpark@uga.edu

A BAYESIAN APPROACH ON SMOOTHING AND MAPPING FUNCTIONAL CONNECTIVITY FOR EVENT-RELATED fMRI TIME SERIES

Dongli Zhou*, University of Pittsburgh
Wesley Thompson, University of California-San Diego

Neuroscientists have become increasingly interested in exploring dynamic relationships among brain regions. The presence of such a relationship is denoted by the term "functional connectivity." We propose a methodology for exploring functional connectivity in event-related designs, where stimuli are presented at a sufficient separation to examine dynamic responses in multiple brain regions. Our methodology simultaneously determines the level of smoothing to obtain the underlying noise-free BOLD response and functional connectivity among several regions. Smoothing is accomplished through an empirical basis via functional principal components analysis. The coefficients of the basis are assumed to be correlated across regions, and the nature and strength of functional connectivity is derived from this correlation matrix. The method is implemented via Bayesian MCMC on an fMRI data set.

email: doz5@pitt.edu

CROSS-CORRELATION ANALYSIS OF SPATIO-TEMPORAL PROCESSES

Huijing Jiang*, Georgia Institute of Technology
Nicoleta Serban, Georgia Institute of Technology

Cross-correlation analysis of processes varying over two different continuum domains, for example, geographic space and time, is becoming increasingly important in a wide range of application fields. In this paper, we introduce a computational efficient and theoretically-founded cross-correlation analysis for bivariate spatio-temporal processes with general applicability. We introduce space-varying and time-varying correlation measures to model different aspects of the local association between spatio-temporal processes. We use a semiparametric model for partitioning the spatio-temporal processes into global and local trends. Under this model, we show that the cross-correlation estimators are asymptotically unbiased under the conditions that the sample size is large and the intrinsic dimensionality of the spatio-temporal processes is much smaller than the sample size. In a simulation study, we evaluate the accuracy of the correlation estimates with respect to the sample size as well as the robustness of the correlation estimates to varying model dimensionality. We illustrate the correlation analysis within a demographic study, in which we analyze the association between per capita income and racial-ethnic diversity for five southeast states at a low spatial aggregation level over the past 11 years.

email: hjiang@isye.gatech.edu

48. CELEBRATING 70: THE CONTRIBUTIONS OF DONALD A. BERRY TO STATISTICAL SCIENCE AND STATISTICAL PRACTICE IN ACADEMICS, INDUSTRY AND GOVERNMENT

DON BERRY, STATISTICS, AND OTHER DANGEROUS THINGS

Michael Krams*, Pfizer Inc.

Don Berry has inspired and energized generations of clinical trialists. This presentations will offer a personal account on how interacting with Don over the past 12 years has led to implementing transformational change in the arena of biopharmaceutical research and development.

email: kramsm@wyeth.com

DON BERRY'S IMPACT IN THE DESIGN AND ANALYSIS OF MEDICAL DEVICE CLINICAL TRIALS IN THE REGULATORY SETTING

Telba Irony*, Center for Devices and Radiological Health, U.S. Food and Drug Administration

Don Berry's contributions have been a driving force behind the dramatic increase in the use of Bayesian methods in the design and analysis of medical device clinical trials for submission to the Food and Drug Administration. Bayesian methods have been particularly helpful, not only due to the availability of prior information, but mainly because they provide flexibility with respect to interim analyses, prediction, meta- analysis, and missing data. Currently, the Center for Devices and Radiological Health at the FDA is also exploring the use of formal Decision Analysis methodology which is inherent to the Bayesian approach. In this presentation we will talk about Professor Don Berry's contributions as an expert, advisor, educator, and promoter of Bayesian method to improve the design and analysis of medical device clinical trials for submission to the Food and Drug Administration.

email: telba.irony@fda.hhs.gov

USING STATISTICS TO FIGHT CANCER: EXAMPLES FROM DON BERRY'S CAREER

Giovanni Parmigiani*, Dana-Farber Cancer Institute

The presentation will be a personal tribute to Don Berry as well as an opportunity to reinforce the fundamental role that our profession can play in discovering new knowledge, in using it to establish policies that affect a large number of individuals, and in communicating both of these steps to large nontechnical audiences. I will present examples that highlight Don Berry's unique approach to using statistical thinking in the ongoing fight against cancer. Without trying to compile a comprehensive catalogue of Don's contributions to cancer research (an impossible task!) I will focus on his role in the discovery of Her2-Neu as one of the earliest molecular markers of response to chemotherapy; his role in the controversies about screening for early detection for both breast and prostate cancer; and his role in initiating modern familial risk prediction.

email: gp@jimmy.harvard.edu

BANDITS, STOPPING RULES AND MULTIPLICITY

James Berger*, Duke University

Don Berry has made fundamental advances in our understanding of statistics. This talk will review some of the highlights of his foundational work on bandit problems, stopping rules and multiplicity. Part of Don's genius is that, while ostensibly theoretical, this work has had a profound effect on the practice of statistics in clinical trials and other areas.

email: berger@stat.duke.edu

49. CURRENT ISSUES IN STATISTICAL PROTEOMICS

INTERNATIONAL COMPETITION ON PROTEOMIC DIAGNOSIS

Bart Mertens*, Leiden University Medical Center

We present a summary overview on the International Competition on Proteomic Diagnosis which was recently organized by the Department of Medical Statistics and Bioinformatics of the Leiden University Medical Centre. The design of this comparative study is discussed and we provide some details on the mass spectrometric data on which this competition was based. A summary of results is presented as well as a reflection on lessons learned.

email: b.mertens@lumc.nl

MONOISOTOPIC PEAK DETECTION AND DISEASE CLASSIFICATION FOR MASS SPECTROMETRY DATA

Susmita Datta*, University of Louisville
Mourad Atlas, U.S. Food and Drug Administration

Mass spectrometry has emerged as a core technology for high throughput proteomics profiling. It has enormous potential in biomedical research specifically for biomarker detection for complex diseases like cancer. However, the complexity of the data poses new statistical challenges for the analysis. Statistical methods and software developments for analyzing proteomics data are likely to continue to be a major area of research in the coming years. In this work we develop a novel statistical method for analyzing matrix assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry data. We propose to use the chemical knowledge regarding isotopic distribution of the peptide molecules along with statistical modeling to detect chemically valuable peaks from each spectrum. We discuss the varying nature of the model fitting procedure in different mass regions of the spectrum. We provide comparative performance of our peak detection method with relatively new other peak detection methods. We demonstrate the superiority of our peaks in the context of classification study for case control data.

email: susmita.datta@louisville.edu

MULTIPLE TESTING ISSUES AND DIMENSION REDUCTION IN PROTEOMICS

Francoise Seillier-Moiseiwitsch*, Georgetown University Medical Center

We describe novel protocols used for analyzing two-dimensional gel images and LC-MS maps. In the resulting protein maps for groups

of patients, we seek to identify proteins that are differentially expressed. We have developed comprehensive analytical approaches that deal with preprocessing, alignment and differential analysis. Preprocessing removes the bulk of the background noise. It involves smoothing, selecting regions containing spots and gradient thresholding. Images are aligned using cubic-spline transformations. The alignment is formulated as a quadratic programming problem that is optimized using an interior-point method. In the global approach, wavelets are utilized to summarize the aligned images, and statistical tests performed on the wavelet coefficients. In the region-based approach, the images are segmented using the watershed algorithm and summary statistics are computed on each region. Statistical tests are applied to these summary statistics. The two-component empirical Bayes model is utilized to estimate the local false-discovery rate. A novel estimation procedure is proposed.

email: seillier@georgetown.edu

50. SURVEY OF METHODOLOGIES FOR POPULATION ANALYSIS IN PUBLIC HEALTH

STATISTICS CAN LIE BUT CAN ALSO CORRECT FOR LIES: REDUCING RESPONSE BIAS IN NLAAS VIA BAYESIAN IMPUTATION

Jingchen Liu*, Columbia University
Xiao-Li Meng, Harvard University
Margarita Alegria, Cambridge Health Alliance and Harvard University
Chih-Nan Chen, Cambridge Health Alliance

National Latino and Asian American Study (NLAAS) is a large scale survey of psychiatric epidemiology, the most comprehensive survey of this kind. Its data were made public in July 2007. A unique feature of NLAAS is its embedded experiment for estimating the effect of alternative orderings of interview questions. The findings from the experiment are not completely unexpected, but nevertheless astonishing. Compared to the survey results from the widely used traditional ordering, the self-reported psychiatric service-use rates are often doubled or even tripled under a more sensible ordering introduced by NLAAS. These findings explain certain perplexing empirical findings in literature, but at the same time impose some grand challenges. In this talk, we present models for imputing the original responses had the respondents under the traditional survey not taken advantage of the skip patterns to reduce interview time, which resulted in increased rates of incorrect negative responses over the course of the interview. The imputation modeling task is particularly challenging because of the complexity of the questionnaire, the small sample sizes for subgroups of interests, and the need of providing sensible imputation for whatever subpopulation a future user might be interested in studying.

email: jcliu@stat.columbia.edu

LATENT SPACE MODELS FOR AGGREGATED RELATION DATA FOR THE STUDY OF HIGH RISK POPULATIONS OF HIV+/AIDS

Tian Zheng*, Columbia University
Tyler H. McCormick, Columbia University

Aggregated Relational Data (ARD), originally introduced by Killworth et al. (1998) as "How many X's do you know" survey questions, are a common tool for observing social networks indirectly, especially for subpopulations that are hard to study directly. Previous methods for ARD estimate specific network features, such as overdispersion. We suggest a more general approach to understanding social structure using ARD based on a latent space framework. In this talk, we will discuss latent space models as a unified framework for inference with ARD by demonstrating that the network features estimated using previous methods can be represented as latent structure. Using data from McCarty et al. (2001), we further demonstrate the utility of a latent space model in extracting social structure information of the networks of individuals who are difficult to reach with traditional surveys, such as those with HIV/AIDS, the homeless, or injection drug users.

email: tzheng@stat.columbia.edu

SAMPLING UNSETTLED POPULATIONS

David Banks, Duke University

After a disaster, or in the wake of armed conflict, it is difficult to obtain urgent public health information needed to target relief. And even when the instability has abated, there is a duty to count the casualties and to support long-term recovery that may require decades. From a statistical standpoint, this poses novel methodological problems: we must survey populations whose members may be dead, displaced, or missing. Traditional sampling starts with a frame, but in these contexts, no usable frame exists. Drawing upon experience in surveying Katrina refugees, and familiarity with several post-conflict surveys for Truth and Reconciliation Committees, this talk compares the pros and cons of respondent-driven sampling, multiple systems estimation, demographic backcasting, and new technological capabilities for making the statistical inferences needed for public health, recovery, and historical understanding.

email: banks@stat.duke.edu

51. PREDICTION AND MODEL SELECTION WITH APPLICATIONS IN BIOMEDICINE

FEATURE SELECTION WITH IN ULTRADIMENSIONAL STATISTICAL PROBLEMS

Jianqing Fan*, Princeton University

Ultrahigh-dimensionality characterizes many contemporary statistical problems from genomics and genetics to finance and economics. We first outline a unified approach to ultrahigh dimensional variable selection problems and then focus on penalized likelihood methods which are fundamentally important building blocks to ultra-high dimensional variable selection. How high dimensionality can such methods handle? What is the role of penalty functions? How to analyze ultrahigh dimensional data and what are possible suprious relations due to ultrahigh dimensionality? This talk will provide some insights into these problems. The focus will on the model selection consistency and oracle properties for a class of penalized likelihood approaches using folded-concave penalty functions. The advantages over convex penalty will be clearly demonstrated. The coordinate optimization is implemented for finding the solution paths, whose performance is evaluated by a few simulation examples and the real data analysis. The recent results on independence screening will also be summarized.

email: jqfan@princeton.edu

ADAPTIVE INDEX MODELS

Lu Tian*, Stanford University
Robert Tibshirani, Stanford University

We use the term index predictor to denote a score that consists of K binary rules such as “age >60” or “blood pressure >120 mm Hg”. The index predictor is the sum of the scores, yielding a value from 0 to K. Such scores as often used in clinical studies to stratify population risk are usually derived from subject area considerations. In this paper we propose a fast procedure for automatically constructing such indices based on a training dataset, for linear, logistic and Cox regression models. We also extend the procedure to create indices for detecting treatment-marker interactions. The methods are illustrated on a study with protein biomarkers as well as two microarray gene expression studies.

email: lutian@stanford.edu

VARIABLE SELECTION IN CENSORED QUANTILE REGRESSION

Huixia Judy Wang*, North Carolina State University

Quantile regression provides a valuable supplement to Cox proportional hazards model for analyzing survival data where censoring is common. In contrast to conventional statistical methods, quantile regression models can help discover heterogeneous effects of drug treatments on survival times of both high and low risk patients. Existing methods for censored quantile regression often require stringent assumptions such as linearity of all quantile functions, which restrict model flexibility and complicate computation. In this talk, I will first present an index-based estimation method for censored quantile regression to accommodate high dimensional covariates. Then I will discuss penalization methods for variable selection, including selection of groups of correlated covariates, in censored quantile regression.

email: wang@stat.ncsu.edu

52. MISSING DATA

BINARY REGRESSION ANALYSIS WITH COVARIATE SUBJECT TO DETECTION LIMIT

Chunling Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Paul Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development

In epidemiologic studies the association between the disease and a continuous exposure is frequently evaluated using a binary regression model with a specified link function. When measuring the exposure level is subject to a limit of detection below which the levels of the exposure can not be quantified, the conventional approach to estimating the regression parameters is not applicable. In this talk we propose a two-stage maximum likelihood estimation approach to estimating the regression parameters, assuming that the exposure levels follow a distribution in the Box-Cox transformation family. The proposed method is appealing in that it provides some flexibility in modeling the disease-exposure data and is more robust than simply assuming that data are normally/log-normally distributed. The methods are exemplified using data from a study on children with autism and autism spectrum disorder to investigate the association of the disease with growth-related hormones.

email: liuc3@mail.nih.gov

PARAMETRIC FRACTIONAL IMPUTATION FOR MISSING DATA ANALYSIS

Jae-kwang Kim*, Iowa State University

Imputation is a popular method of handing missing data in practice. Imputation, when carefully done, can be used to facilitate the parameter estimation by applying the complete-sample

estimators to the imputed dataset. The basic idea is to generate the imputed values from the conditional distribution of the missing data given the observed data. In this article, parametric fractional imputation is proposed for generating imputed values. Using fractional weights, the observed likelihood can be approximated by the weighted mean of the imputed data likelihood. Some computational efficiency can be achieved using the idea of importance sampling and calibration weighting. The proposed imputation method provides efficient parameter estimates for the model parameters specified in the imputation model and also provides reasonable estimates for parameters that are not part of the imputation model. Variance estimation is covered and results from a limited simulation study are presented.

email: jkim@iastate.edu

THE CONVERGENCE OF MULTIPLE IMPUTATION ALGORITHMS USING A SEQUENCE OF REGRESSION MODELS

Jian Zhu*, University of Michigan
Trivellore E. Raghunathan, University of Michigan

The convergence of multiple imputation algorithms using a sequence of regression models is often arguable because the fully conditionally specified models may be incompatible and then the underlying joint model of all variables does not exist. In this paper, we focus on sequential regression imputation algorithms applied on bivariate data with ignorable missing values and assess their convergence properties based on theoretical work and simulation studies. For imputation models that can be derived from a joint distribution, we demonstrate that the model incompatibility due to overparameterization can be ignored and imputation results converge to the targeting joint distribution. For incompatible imputation models that can not form a joint distribution, we show that the bias introduced by model incompatibility can be reduced to a minimum if the conditional models fit the data well. We conclude that sequential regression imputation algorithms perform well when the regression models are correctly specified, and researchers applying such algorithms should focus on improving model specification in practice.

email: jianzhu@umich.edu

LOGISTIC REGRESSION MODELS WITH MONOTONE MISSING COVARIATES

Qixuan Chen*, Columbia University
Myunghee C. Paik, Columbia University

We propose a new method to handle monotone missing covariates in the logistic regression model for a binary outcome when the probability of missingness depends on the observed outcome or covariates. The proposed estimating equation presents alternative to inverse probability weighting, imputation, or likelihood-based approaches when missing covariates arise from exponential family distributions. Under certain regularity conditions, the estimates of the regression coefficients obtained by the proposed method are consistent and asymptotically normally distributed. This method can be extended to GEE models for binary outcomes as well as logistic regression models for complex survey data. We illustrate this method using a study of environmental exposure to dioxin in Michigan.

email: qc2138@columbia.edu

SUBSAMPLE IGNORABLE MAXIMUM LIKELIHOOD FOR REGRESSION WITH MISSING DATA

Nanhua Zhang, University of Michigan
Roderick J. Little, University of Michigan

Let X be an n by q data matrix, and X = (X1, X2, X3) be a column partition of X.. Interest concerns the parameters of the regression of X3 on X1 and X2, or regression parameters obtained by further conditioning on components of X3. The submatrix X1 is completely observed, but X2 and X3 contain missing values. Two standard analyses are to estimate the target distribution using the set of complete cases (CC), or to estimate it using all the data by maximum likelihood, assuming the missing data mechanism is ignorable (IML). We propose subsample IML (SSIML), a hybrid method that computes the target distribution by IML, restricted to the subsample of cases where X2 is fully observed. Conditions on the missing data mechanism are presented under which SSIML gives consistent estimates, but both IML and CC analysis are inconsistent. In other circumstances, IML is inconsistent and SSIML and CC are consistent, but SSIML is more efficient than CC since it uses more of the data. We apply the proposed method to regression analysis with missing covariates, and to data from the National Health and Nutrition Examination Survey and a liver cancer study.

email: rlittle@umich.edu

MULTIPLE IMPUTATION FOR REGRESSION ANALYSIS WITH MEASUREMENT ERROR IN A COVARIATE

Ying Guo*, University of Michigan
Roderick Little, University of Michigan

Covariate measurement error is very common in empirical studies, and currently information about measurement error provided from calibration samples is insufficient to provide valid adjusted inferences. We consider the problem of estimating the regression of an outcome Y on covariates X and Z, where Y and Z are observed, X is unobserved, but a proxy variable W that measures X with error is observed. Data on the joint distribution of X and W (but not Y and Z) are recorded in a calibration experiment. The data from this experiment are not available to the analyst, but summary statistics for the joint distribution of X and W are provided. We

describe a multiple imputation (MI) method that provides multiple imputations of the missing values of X in the regression sample, so that the regression of Y on X and Z and associated standard errors are estimated correctly using standard multiple imputation (MI) combining rules, under normal assumptions. Parameters are identified by assuming non-differential measurement error, that is, Y and Z are independent of W given X. The proposed method is shown by simulation to provide better inferences than existing methods, namely the naïve method and regression calibration.

email: guoy@umich.edu

INFORMATIVE MODEL SPECIFICATION TEST USING COARSENED DATA

Xianzheng Huang*, University of South Carolina

We propose novel methods based on coarsened data to assess the validity of model specifications. The methodological development is initially set on the platform of mixed effects models, and can potentially be extended to a more general context of model-assumptions checking. The idea underlying the method is that, in the presence of model misspecification, likelihood inference based on the raw observed data can differ from the counterpart inference resulting from a coarsened data induced from the raw data. By monitoring the change in inference as the data is coarsened, one can detect violation to model assumptions for a particular data set. Moreover, when multiple assumptions may be violated, one can reveal the specific source(s) of misspecification by examining the pattern of discrepancy between these two sets of inference. A strategically designed coarsening mechanism can improve the power to detect model missepcification.

email: huang@stat.sc.edu

53. INFERENCE FOR CLINICAL TRIALS

METHODS TO TEST MEDIATED MODERATION IN LOGISTIC REGRESSION: AN APPLICATION TO THE TORDIA CLINICAL TRIAL

Kaleab Z. Abebe*, University of Pittsburgh
Satish Iyengar, University of Pittsburgh
David A. Brent, University of Pittsburgh

Currently there is little discussion about methods to explain treatment-by-site interaction in multisite clinical trials, so investigators must explain these differences post-hoc with no formal statistical tests in the literature. An example is the Treatment of SSRI-Resistant Depression in Adolescents (TORDIA) study, which concluded that the combination of cognitive behavioral therapy and antidepressant medication (versus only medication) had an effect on clinical response that was highly variable across sites. A secondary paper sought to explain these differences using a variety of univariate analyses, and came to the conclusion that differences in baseline clinical characteristics across sites were to blame. Here, we extend mediated moderation techniques to the logistic regression model with two treatments, multiple sites, and multiple mediator variables. Test statistics and critical values are derived for difference-in-coefficients and product-of-coefficients tests, and power is estimated using simulation. In the single mediator case, the former test does well in terms of approximating type I error and power, while the latter suffers from a slight inflation of type I error. Finally, the proposed methodology is applied to the TORDIA study. The contribution of this is formal significance tests for explaining treatment-by-site interaction in multisite clinical trials.

email: kza3@pitt.edu

EXACT TESTS USING TWO CORRELATED BINOMIAL VARIABLES IN CONTEMPORARY CANCER CLINICAL TRIALS

Jihnhee Yu*, University at Buffalo
James Kepner, American Cancer Society
Renuka Iyer, Roswell Park Cancer Institute

New therapy strategies for the treatment of cancer are rapidly emerging because of recent technology advances in genetics and molecular biology. Although newer targeted therapies can improve survival without measurable changes in tumor size, clinical trial conduct has remained nearly unchanged. When potentially efficacious therapies are tested, current clinical trial design and analysis methods may not be suitable for detecting therapeutic effects. We propose an exact method with respect to testing cytostatic cancer treatment using correlated bivariate binomial random variables to simultaneously assess two primary outcomes. The method is easy to implement. It does not increase the sample size over that of the univariate exact test and in most cases reduces the sample size required. Sample size calculations are provided for selected designs.

email: jinheeyu@buffalo.edu

HYPOTHESIS TESTING PROBLEM FOR THREE-ARM NONINFERIORITY TRIALS

Xiaochun Li*, New York University
Yongchao Ge, Mount Sinai School of Medicine
Judith D. Goldberg, New York University

In three-arm randomized clinical trials that consists of placebo, reference control, and experimental treatments, we need to demonstrate that: (A) the reference control is better than the placebo; and (B) the experimental treatment is not worse than the reference control. Pigeot's paper (Statistics in Medicine 2003, 22:883-899) and Hasler's paper (Statistics in Medicine 2008, 27:490-503) consider a t-statistic for the non-inferiority testing problem (B) to compute the type I error rates and power. These methods require the rejection of the null hypothesis of (A) in order to proceed to test hypothesis (B); however, they have not formally

incorporated these two hypotheses into the computation of type I error rates and power. In this talk, we consider three methods to test the hypotheses (A) and (B) jointly. These methods are: (1) p-value rejection method that is implicitly done by Pigeot and Hasler; (2) our proposed Maxp method; and (3) our proposed likelihood ratio test (LRT) approach. The type I error rates and power of the three methods are compared by simulation. We find that the LRT and Maxp methods have increased power compared with the p-value rejection method for the cases considered.

email: xl303@nyu.edu

AN EMPIRICAL LIKELIHOOD APPROACH TO NONPARAMETRIC COVARIATE ADJUSTMENT IN RANDOMIZED CLINICAL TRIALS

Xiaoru Wu*, Columbia University
Zhiliang Ying, Columbia University

Covariate adjustment is an important tool in the analysis of randomized clinical trials as well as observational studies. It can be used to increase efficiency and thus power, and to reduce possible bias. While most statistical tests for randomized clinical trials are nonparametric in nature, approaches for covariate adjustment typically rely on specific regression models, such as the linear model for continuous outcome variable, the logistic regression for dichotomous outcome and the Cox model for survival time. This paper makes use of the empirical likelihood method and proposes a nonparametric approach to covariate adjustment. A major advantage of the new approach is that it automatically utilizes covariate information in an optimal way without fitting nonparametric regression. The usual asymptotic properties, including the Wilks-type result of convergence to chi-square distribution for the empirical likelihood ratio based test and asymptotic normality for the corresponding maximum empirical likelihood estimator, are established. It is also shown that the resulting test is asymptotically most powerful and that the estimator for treatment effect achieves semiparametric efficiency bound.

email: xw2144@columbia.edu

INCORPORATING THE RISK DIFFERENCE INTO NON-INFERIORITY TRIALS OF SAFETY

Kristine R. Broglio*, Berry Consultants and Texas A&M University
Jason T. Connor, Berry Consultants
Scott M. Berry, Berry Consultants

In response to growing safety concerns, many divisions of the FDA are requiring premarketing safety studies. Guidance is based on ensuring the risk ratio of treatment to an appropriate control is less than a clinically unacceptable value. The restriction to a risk ratio has several important consequences in drug and device development. A trial's power is purely a function of the number of events. This causes the therapy to be investigated in a high-risk population, which may not be the intended treatment population. A treatment effect may not be seen because the high-risk population will experience events early, before treatments have had an effect. While a high-risk population is used to avoid enormous trial sample sizes, paradoxically, extremely large trials in a lower risk or the intended treatment population may provide very few events. Large trials should appropriately categorize the possible safety risks of the new therapy, and yet are inconclusive. The risk difference may be a more appropriate measure of risk when event rates are very low. Therefore, we propose using both the risk ratio and the risk difference in the definition of non-inferiority. We explore this altered definition of non-inferiority, which has been implemented in multiple Bayesian adaptive clinical trials.

email: kristine@berryconsultants.com

POST-RANDOMIZATION INTERACTION ANALYSES IN CLINICAL TRIALS WITH STANDARD REGRESSION

Rongmei Zhang*, University of Pennsylvania
Jennifer Faerber, University of Pennsylvania
Marshall Joffe, University of Pennsylvania
Tom Ten Have, University of Pennsylvania

We address several questions on analyzing how post-randomization factors may modify the intent-to-treat effects of randomized interventions. We investigate the assumptions underlying the standard regression model with main effects and interactions for the baseline randomized intervention and post-randomization effect modifier. The crucial assumptions are sequential ignorability and no effect of the baseline intervention on the post-randomization. We present analytic results for all terms in the model under different combinations of the assumptions. In addition, we confirm our results with simulations and further assess our results through a randomized cognitive therapy trial example. We show that there are different biases for the interaction term and the stratified intervention effect when the above assumptions are violated.

email: rongmei@mail.med.upenn.edu

ESTIMATING TREATMENT EFFECTS IN RANDOMIZED CLINICAL TRIALS WITH NON-COMPLIANCE AND MISSING OUTCOMES

Yan Zhou*, U.S. Food and Drug Administration
Jack Kalbfleisch, University of Michigan
Rod Little, University of Michigan

We analyze randomized trials with active treatment verses control treatment, where treatments are subject to all-or-none compliance and outcomes have missing values. In addition to latent ignorability (Frangakis and Rubin, 1999), we further specify two assumptions for principal compliance and two assumptions for missing outcome to identify the model. In each of four scenarios defined by combinations of these assumptions, we derive maximum

likelihood (ML) estimates by using the EM algorithm, as well as non-iterative ML estimates by implementing pattern-mixture models with covariates (Little and Wang, 1996). This shows that, under certain conditions, the method-of-moments (MOM) estimates are ML estimates. We show that the models of principal compliance determine which type of analysis is used to estimate treatment efficacy, per-protocol analysis or IV estimation with the treatment assignment indicator as the instrumental variable. On the other hand, we show that the assumptions for missing outcome determine whether MOM estimates are ML estimates or not. We apply our methods to data from a double-blinded randomized clinical trials with clozapine vs. haloperidol for patients with refractory schizophrenia.

email: yan.zhou@fda.hhs.gov

54. METHODS FOR IMAGE DATA AND TIME SERIES

fMRI ANALYSIS VIA BAYESIAN VARIABLE SELECTION WITH A SPATIAL PRIOR

Jing Xia*, University of Illinois at Urbana-Champaign
Feng Liang, University of Illinois at Urbana-Champaign
Yongmei M. Wang, University of Illinois at Urbana-Champaign

Applications of functional magnetic resonance imaging (fMRI) provide insights into the neuroscience. Especially, more and more interests have been put into the research on hemodynamic response function (HRF). This paper presents a novel spatial Bayesian method for simultaneous HRF estimation and activation detection for fMRI data. A Bayesian variable selection approach is used to induce shrinkage and sparsity; moreover, a spatial prior on latent variables is used to represent activated hemodynamic response components. Then, the activation map is generated from the full spectrum of posterior inference constructed through a Markov chain Monte Carlo scheme, and HRFs at different voxels are estimated nonparametrically with information pooling from neighboring voxels. By integrating functional activation detection and HRFs estimation in a unified framework, our method is more robust to noise and less sensitive to model mis-specification.

email: jingxia2@illinois.edu

WAVELET THRESHOLDING USING ORACLE FALSE DISCOVERY RATE WITH APPLICATION TO FUNCTIONAL MAGNETIC RESONANCE IMAGING

Nan Chen*, George Mason University
Edward J. Wegman, George Mason University

The detection of differentiated voxels for a specified task is difficult in the functional magnetic resonance imaging (fMRI) study due to many reasons. One of the major reasons is the poor signal-to-noise ratios (SNRs) in the typical fMRI data. As an effort to improve SNRs, we study the wavelet thresholding problem in the fMRI study. We propose to conduct wavelet thresholding procedure using an oracle false discovery rate approach (Sun and Cai, 2007). This involves extracting wavelet coefficients resulting from images and can be formulated as a multiple hypotheses testing problem. We conduct a number of fMRI-type simulations to compare the numerical performance between our approach and two other well adapted approaches in the current literature including the FDR procedure (Benjamini and Hochberg, 1995) and adaptive FDR procedure (Benjamini and Hochberg, 2000). We also illustrate the proposed method on a real fMRI data.

email: chennan93@yahoo.com

A SEMIPARAMETRIC HETEROGENEOUS-MIXTURE MODEL FOR CERTAIN QUANTUM-DOT IMAGES, AND LIKELIHOOD INFERENCE FOR DOT COUNT AND LOCATION

John Hughes*, Penn State University
John Fricks, Penn State University

We introduce a procedure to automatically locate and count the quantum dots in certain microscopy images. Our procedure employs an approximate likelihood estimator based on a two-component heterogeneous-mixture model for the image data; the first component is normal, and the other component is a normal plus an exponential. The normal component has an unknown variance function, which we model as a function of the mean. We use B-splines to estimate the variance function during a training run on a suitable image, and the estimate is used to process subsequent images. Estimates of standard errors are generated for each image along with the parameter estimates, and the number of dots in the image is determined using an information criterion and likelihood-ratio tests. Realistic simulations show that our procedure is robust and that it leads to accurate estimates, both of parameters and of standard errors.

email: jph264@psu.edu

THE GENERALIZED SHRINKAGE ESTIMATOR FOR PARTIAL COHERENCE ESTIMATION IN MULTIVARIATE TIME SERIES

Mark Fiecas*, Brown University
Hernando Ombao, Brown University
We develop a new statistical method for the estimation of functional connectivity between neurophysiological signals represented by a multivariate time series. We use partial coherence as the measure of functional connectivity. Partial coherence identifies the frequency band that drives the direct linear association between any pair of channels. To estimate partial coherence, one would first need an estimate the spectral density matrix of the multivariate time series. In this work, we develop the

generalized shrinkage estimator, which is a weighted average of a parametric estimator and a nonparametric estimator. We derive the optimal weights under the expected L_2 loss criterion. We validate the generalized shrinkage estimators through simulated data sets and apply it on an EEG data set.

email: mfiecas@stat.brown.edu

ELASTIC-nt BASED MODEL FOR IMAGING MS PROTEOMIC DATA PROCESSING

Fengqing Zhang*, Middle Tennessee State University
Don Hong, Middle Tennessee State University

Imaging Mass Spectrometry (IMS) has shown a great potential of direct examination of biomolecular patterns from cells and tissue. However, challenges remain in data processing due to the difficulty of high dimensionality, the fact that the number of predictors is significantly larger than the sample size, and the needs to consider both spectral and spatial information in order to represent the advantage of the equipment technology well. A very recently developed elastic-net (EN) method, produces a sparse model with admirable prediction accuracy, can be an effective tool for IMS data processing. In this article, we incorporate a spatial penalty term into the EN model and develop a new tool for IMS biomarker selection and classification. The EN-based model outperforms many other popular statistical methods for IMS data analysis. A software package, called EN4IMS, is also presented. IMS data analysis results show that EN4IMS helps in confirming new biomarkers, producing a more precise peak list, and providing more accurate classification results. The EN-based model takes data without peak binning beforehand and thus saves a signficant amount of time for data processing. A more advanced model by incorporating weights into EN method, called WEN, will be also discussed.

email: fengqingbuaa@gmail.com

SUBDIFFUSION DETECTION IN MICRORHEOLOGICAL EXPERIMENTS

Gustavo Didier*, Tulane University
John Fricks, Penn State University

The widespread availability of high quality light microscopy combined with high speed digital camera recording and automated tracking tools has allowed for experiments which track single particles passively diffusing in complex fluids (Microrheology). In a Newtonian fluid, the mean squared displacement (MSD), i.e., the second moment, of a particle's position grows linearly in time, a situation called diffusion. However, for other viscoelastic materials such as biological fluids, one may observe a MSD that grows slower than linearly, also called subdiffusion. Detecting subdiffusion in a statistically sound manner can be biologically relevant. For example, knowing that a virus or other parasite in a complex biological fluid such as lung mucus diffuses out relatively slowly can have important clinical ramifications. In this talk, we propose the use of the Local Whittle Estimator to estimate and test for subdiffusivity in data from human lung mucous. Moreover, we propose a fast wavelet-based simulation method for the velocity process of the particle. This allows us to study the finite-sample properties of the Local Whittle Estimator as well as the power of the associated hypothesis testing procedure.

email: gdidier@tulane.edu

55. SURVIVAL ANALYSIS: CURE MODELS AND COMPETING RISKS

SEMIPARAMETRIC REGRESSION CURE MODELS FOR INTERVAL-CENSORED DATA

Hao Liu*, Baylor College of Medicine
Yu Shen, University of Texas M. D. Anderson Cancer Center

We present semiparametric non-mixture cure models for the regression analysis of interval-censored time-to-event data. Motivated by medical studies in which patients could be cured of disease but the disease event time may be subject to interval censoring, we develop semiparametric maximum likelihood estimation for the model using the expectation-maximization method. The maximization step for the baseline function is nonparametric and numerically challenging. We develop an efficient and numerically stable algorithm via modern convex optimization techniques, which yields a self-consistency algorithm. We prove the strong consistency of the maximum likelihood estimators under the Hellinger distance. We assess the performance of the estimators in a simulation study with small to moderate sample sizes. To illustrate the method, we analyze a real data set from a medical study for the biochemical recurrence of prostate cancer among patients who have undergone radical prostatectomy.

email: haol@bcm.edu

INCORPORATING SHORT-TERM EFFECTS IN CURE MODELS WITH BOUNDED CUMULATIVE HAZARDS

Xiang Liu*, University of Rochester Medical Center
Li-Shan Huang, University of Rochester Medical Center

Survival models incorporating a surviving (cure) fraction are becoming increasingly popular in analyzing data from cancer clinical trials. The bounded cumulative hazard model is an appealing alternative to the widely used two-component mixture cure model. Making the assumption of bounded cumulative hazards in a proportional hazards (PH) model leads to the so-called improper PH model. Such an improper PH model considers the covariate effects on the long-term surviving fraction and has a biologically meaningful interpretation from the view of a simple mechanistic model. Motivated by a more general mechanistic model of tumor recurrence, we extend the improper PH model

to further take into account the short-term covariate effects in survival. Simulations and a real data example are presented for illustration.

email: xliu@urmc.rochester.edu

CURE RATE MODEL WITH NONPARAMETRIC SPLINE ESTIMATED COMPONENTS

Lu Wang, Virginia Tech University
Pang Du*, Virginia Tech University

In some medical studies, there are often long term survivors who can be considered as permanently cured. The goals in these studies include the understanding of the covariate effect on both the cure probability of the whole population and the hazard rate of the non-cured subpopulation. We propose a two-component mixture cure model with nonparametric forms for both the cure probability and the hazard rate function. Identifiablity of the model is guaranteed by an additive assumption on hazard rate. Estimation is carried out by an EM algorithm on maximizing a penalized likelihood. Consistency and convergence rate are established. We then evaluate the proposed method by simulations and application to a melanoma study.

email: pangdu@vt.edu

CAUSE-SPECIFIC ASSOCIATION MEASURES FOR MULTIVARIATE COMPETING RISKS DATA AND THEIR NONPARAMETRIC ESTIMATORS

Yu Cheng, University of Pittsburgh
Hao Wang*, University of Pittsburgh

There are some recent developments in association analysis of bivariate competing risks data. For examples, Bandeen-Roche and Liang (2002) and Bandeen-Roche and Ning (2008) extended Oakes (1989) cross hazard ratio to cause-specific competing risks settings, and Cheng and Fine (2008) proposed an equivalent association measure based on bivariate cause-specific hazards. These approaches take into account the dependent structure between the risk of interest and competing risks which is the obstacle to utilizing standard methods for bivariate survival analysis. To broaden their applications, Cheng et al. (2009) further extended the cause-specific cross hazard ratio to more complicated family structures, e.g., exchangeable sibship data and multiple mother-child pairs. In line with this research, we propose a pseudo-likelihood estimator (Clayton 1978, Oakes 1982) for the extended cause-specific cross hazard ratio. We also adapt the association measure proposed by Cheng and Fine (2008) to the clustered data and develop a plug-in estimator. Asymptotic properties of the two estimators are established by using empirical processes techniques and their practical performances are compared with that of the U-statistic estimator (Cheng and Fine 2009) by simulation studies. The practical utility of the three approaches is illustrated in an analysis of the Cache County Study of Dementia.

email: haw21@pitt.edu

SEMIPARAMETRIC ANALYSIS OF COMPETING RISKS MODEL WITH A MISATTRIBUTION OF CAUSE OF DEATH

Jinkyung Ha*, University of Michigan
Alex Tsodikov, University of Michigan

Missing failure type is a very common phenomenon due to various reasons in competing risks data. Under the missing-at-random (MAR) assumption, this problem has received considerable attention. However, the MAR assumption is often unrealistic. More specifically, many authors have pointed out that with the introduction of prostate-specific antigen (PSA) screening in the late 1980s, a proportion of deaths may be mistakenly classified to prostate cancer just because the men were diagnosed with prostate cancer. We first show that the more informative partial likelihood is equivalent with the profile likelihood and its score function is semiparametric efficient if the ratio of baseline hazard rates is parametrically modeled. We then introduce a Kullback-Leibler's type function whose empirical counterpart is a full likelihood. By making some adjustments to Kullback-Leibler's type function, we derive two estimating equations which do not require any parametric assumption for the ratio of baseline hazard rates. We then propose the weighted estimating equation which gains more efficiency by allowing a parametric model for the ratio. The corresponding estimator is consistent even if the model is not correctly specified.

email: jinha@umich.edu

REGRESSION STRATEGY FOR THE CONDITIONAL PROBABILITY OF A COMPETING RISK

Aurelien Latouche*, University of Versailles

Competing risks are classically summarized by the cause-specific hazards and the cumulative incidence function. To get a full understanding of the competing risks, these quantities should be viewed simultaneously for all possible events. Another quantity is the conditional probability of a competing risk, (aka conditional cumulative incidence) which is defined as the probability of having failed from a particular cause given that no other (competing) events have occurred. This quantity provides useful insights and its interpretation may be preferable to communicate to clinicians. The use of the conditional probability has been limited by the lack of regression modeling strategy. In this work we apply recently developed regressions methodologies to the conditional probability function and illustrate the insights which can be gained using this

methodology with a case study on conditioning regimens prior stem-cell transplantation (SCT) in acute leukemia.

email: aurelien.latouche@uvsq.fr

NUMBER NEEDED TO TREAT FOR TIME TO EVENT DATA WITH COMPETING RISKS

Suprateek Kundu*, University of North Carolina-Chapel Hill
Jason P. Fine, University of North Carolina-Chapel Hill

The Number Needed to Treat(NNT) has been a popular choice amongst clinicians for summarizing the efficacy of a test treatment versus a control or an active treatment,where the outcome is binary,say success and failure. Although such methods have been extended to time to event data with independent right censoring, their application to survival data with multiple causes of failure has not be examined. Competing risks data are common in clinical trial applications, like those in cancer, in which failure may occur due to either disease or non-disease related causes. In such time to event settings, the NNTs are functions of time, with the resulting estimates varying across time. Extending methods for independently right censored data, we develop methods for estimating NNT separately for each cause of failure, thus giving us better insight into the efficacy of the treatments on the competing risk end-points. We provide both a non-parametric method involving cumulative incidence functions and a semi-parametric method involving proportional hazards model for the subdistribution to come up with ways for computing NNTs. Furthermore we also provide ways to compute condence intervals for NNTs. The methods are illustrated with data from a breast cancer trial.

email: sups1984@yahoo.co.in

56. STATISTICAL GENETICS: QUANTITATIVE TRAIT LOCI

BAYESIAN NONPARAMETRIC MULTIVARIATE STATISTICAL MODELS FOR QUANTITATIVE TRAITS AND CANDIDATE GENES ASSOCIATION TESTS IN STRUCTURED POPULATIONS

Meijuan Li*, U.S. Food and Drug Administration
Tim Hanson, University of Minnesota

Population-based Linkage-Disequilibrium (LD) mapping permits finer-scale mapping than linkage analysis. However, the population-based association mapping is subject to false positives that due to the population structure and the kinship between the samples. While there is interest in simultaneously testing the association between a candidate gene and the multiple phenotypes of interest, the current available association mapping methods are limited to univariate traits only. Here we present a new method for population-based multi-trait candidate gene association mapping via a Bayesian semiparametric approach, where the error distribution is flexibly modeled via a multivariate mixture of Polya trees centered around a family of multivariate normal distributions. The method we developed accounts for the population structure and the complex relatedness between the samples. We will compare the new proposal in the type I error rate and power to the existing multivariate version of Yu's parametric model using the previously published association mapping for two types of \emph{Arabidopsis} \emph {thaliana} flowering data and simulated data as well.

email: meijuan.li@fda.hhs.gov

ROBUST SCORE STATISTICS FOR QTL LINKAGE ANALYSIS USING EXTENDED PEDIGREES

Chia-Ling Kuo*, University of Pittsburgh
Eleanor Feingold, University of Pittsburgh

Score statistics for quantitative trait locus (QTL) linkage analysis have been proposed by many authors as an alternative to variance components (VC) and/or Haseman-Elston (HE) type methods because they have high power and can be made robust to selected samples and/or non-normal traits. But most literature exploring the properties of these statistics has focused on nuclear families. There are a number of computational complexities involved in implementing the score statistics for extended pedigrees, primarily having to do with computation of the statistic variance. In this paper, we propose several different practical methods for computing this variance in general pedigrees, some of which are based only on relative pairs and some of which require working with the overall pedigree structure, which is computationally more difficult. We evaluate the performance of these different score tests using various trait distributions, ascertainment schemes, and pedigree types.

email: chialing.kuo@gmail.com

IDENTIFYING QTLs IN CROP BREEDING POPULATIONS USING ADAPTIVE MIXED LASSO

Dong Wang*, University of Nebraska
Kent M. Eskridge, University of Nebraska
Jose Crossa, International Maize and Wheat Improvement Center

Recently, there has been heightened interest in performing association analysis in important crop species for its significant potential in dissecting complex traits by utilizing diverse mapping populations. Notable examples include studies in maize, wheat, barley, sorghum, and potato. However, the mixed linear model approach is currently limited to single marker analysis, which is not suitable for studying multiple QTL effects, epistasis and gene by environment interactions. In this talk, we report the development of the adaptive mixed LASSO method that can incorporate a

large number of predictors (genetic markers, epistatic effects, environmental covariates, and gene by environment interactions) while simultaneously accounting for the population structure. We have proved that the adaptive mixed LASSO estimator is consistent under mild conditions. Algorithms have been developed to iteratively estimate the regression coefficients and variance components. Our results show that the adaptive mixed LASSO method is very promising in modeling multiple genetic effects as well as modeling gene by environment interactions when a large number of markers are available and the population structure cannot be ignored. It is expected to advance the study of complex traits in important crop species.

email: dwang3@unl.edu

AN APPROACH TO TESTING PLEIOTROPY WITH QUANTITATIVE TRAITS IN GENOME-WIDE ASSOCIATION STUDIES

Emily Kistner-Griffin*, Medical University of South Carolina
Nancy J. Cox, University of Chicago
Dan L. Nicolae, University of Chicago

Recently several genome-wide association studies have been completed for studies of disease states that are best described by multiple quantitative measurements. One such disease of interest is Type I Diabetes, in which severity of disease may be represented by summarizing numerous biologic variables, such as HDL, cholesterol, hemoglobin A1C, and serum creatinine. Investigators may explore each quantitative trait for genetic associations that predispose to an elevated biomarker, although it would be advantageous to consider the traits simultaneously. Here, we propose a method for testing pleiotropy, in which in the phenotypic traits are correlated because a mutation in a single gene signals alterations in various biologic pathways. The proposed method allows testing for differences in correlations across genotype groups, assuming a bivariate normal distribution and using a weighted linear combination of Fisher's z transformations of the estimated genotype-specific correlations. Simulations are described for phenotypes violating the bivariate normality assumption and an adjusted test statistic is proposed. We demonstrate the method using Affymetrix SNP Array 5.0 data from the Genetics of Kidney in Diabetes (GoKinD) study. An R-plugin for PLINK, software used for genome-wide association studies (GWAS), is developed to allow efficient testing of pleiotropy for GWAS data.

email: kistner@musc.edu

ENRICHING OUR KNOWLEDGE IN GENE REGULATION VIA eQTL MAPPING: A COMBINED P-VALUE APPROACH

Shaoyu Li*, Michigan State University
Yuehua Cui, Michigan State University

The genetic bases of complex traits often involve multiple inherited genetic factors that function in a network basis. By changing the expression of functional genes related to a trait, gene regulations have been thought to be a major player in determining the trait variations. The combined analysis of genetic and gene expression, termed eQTL mapping, holds great promise in this regard. Known that genes function in a network basis, the detection of overall signal of the system could shed new light on the role of genetic regulation. We propose to identify novel regulators that mediate the expression changes by combining evidences to study gene regulations in an eQTL mapping framework. We hypothesize that gene expression changes are due to the regulation of a set of variants that belongs to a common system (e.g., network/pathway), and combine individual p-values in the system to form an overall signal while considering correlations between variants. Both simulation and real data analysis show the relative merits of the combined method. The proposed method provides an alternative strategy in addressing questions related to gene regulations from a systems biology perspective.

email: lishaoyu@stt.msu.edu

COMPARISON OF METHODS FOR gXg INTERACTION FOR QUANTITATIVE TRAITS IN CASE-CONTROL ASSOCIATION STUDIES

Raymond G. Hoffmann*, Medical College of Wisconsin
Soumitra Ghosh, Medical College of Wisconsin
Thomas J. Hoffmann, University of California-San Francisco
Pippa M. Simpson, Medical College of Wisconsin

Identifying g x g interaction for quantitative traits is a difficult problem because of the many potential forms for interaction. Multiple methods have been used to model these relationships: linear regression, survival analysis, Classification and Regression Trees (CART), Random Forests, etc. The goal of this study is to compare these methods for different types of interaction -- linear by linear, a PK/PD pharmocologic model, as well as some of the more complex forms found in genetics such as a positive interaction at one level and a negative interaction at another level.

email: hoffmann@mcw.edu

MAPPING QUANTITATIVE TRAIT LOCI FOR TIME-TO-EVENT PHENOTYPE WITH CURED INDIVIDUALS

Chi Wang*, University of California-Riverside
Zhiqiang Tan, Rutgers University
Thomas A. Louis, Johns Hopkins University

Time-to-event is an important trait in many genetic studies. It is frequently observed that a substantial proportion of individuals do not experience the event by the end of study. Among these individuals, some may be considered as cured in the sense that

they are free of the event even with extended follow-up time. The proportional hazards cure model has been used to model the phenotype distribution in presence of cured individuals. But the model formulation is different from the regular proportional hazards model for data without cured individuals. The application of this method is limited since researchers are usually uncertain about the presence of cured individuals. In this presentation, we propose a unified semiparametric model suitable for both the presence and absence of cured individuals. We develop a genome-wide screening procedure based on the proposed model. Our method is illustrated using a Listeria infection data set to identify QTLs associated with survival times of mice after the infection.

email: chiwang@ucr.edu

57. MICROARRAYS: TIME COURSE AND GENE SETS

A NOVEL APPROACH IN TESTING FOR PERIODICITY IN CELL-CYCLE GENE EXPRESSION PROFILES

Mehmet Kocak*, St. Jude Children's Research Hospital
E. Olusegun George, University of Memphis
Saumyadipta Pyne, MIT and Harvard University

Investigating the cyclic behavior of genes during cell cycle has been of interest for a long time. The permutation test described by Lichtenberg et.al. and Fisher's G-test are among the most commonly used methods to test whether or not a given gene has significantly cyclic pattern. Fisher's G-test does not utilize the exact timing of the time-course gene expression profile of a given gene as it only utilizes the rank of the exact time points. On the other hand, the permutation test may be inefficient when one wants to perform a large number of permutations. Therefore, in this study, we propose a novel approach, based on a non-linear regression model, to test whether or not a gene has periodic behavior. We compare the sensitivity and specificity of our novel approach with the permutation test and Fisher's G-test using extensive simulations. We, then, apply the nonlinear approach to real gene expression time-course data on Schizosaccharomyces pombe (Rustici et. al. 2004; Oliva et. al. 2005; Peng et. al. 2005).

email: mehmet.kocak@stjude.org

A UNIFIED MIXED EFFECTS MODEL FOR GENE SET ANALYSIS OF TIME COURSE MICROARRAY EXPERIMENTS

Lily Wang*, Vanderbilt University
Xi Chen, Vanderbilt University
Russell D. Wolfinger, SAS Institute Inc.

Methods for gene set analysis test for coordinated changes of a group of genes involved in the same biological process or molecular pathway. Higher statistical power is gained for gene set analysis by combining weak signals from a number of individual genes in each group. Although many gene set analysis methods have been proposed for microarray experiments with two groups, few can be applied to time course experiments. We propose a unified statistical model for analyzing time course experiments at the gene set level using random coefficient models, which fall into the more general class of mixed effects models. These models include a systematic component that models the mean trajectory for the group of genes, and a random component (the random coefficients) that models how each gene's trajectory varies about the mean trajectory. The proposed methodology provides a unified statistical model for systems analysis of microarray experiments with complex experimental designs when re-sampling based methods are difficult to apply.

email: lily.wang@vanderbilt.edu

RANK BASED GENE SELECTION FOR CLASSIFICATION

Shuxin Yin*, Auburn University
Asheber Abebe, Auburn University

One important application of gene expression microarray data is classification of samples into categories, such as types of tumor. Gene selection procedures become crucial since gene expression data from DNA miscroarrays are characterized by thousands measured genes on only a few subjects. Of these, only a few genes are thought to determine a specific genetic trait. In this presentation, we develop a novel nonparametric procedure for selecting such genes. This rank-based forward selection procedure rewards genes for their contribution towards determining the trait but penalizes them for their similarity to genes that are already selected. We will show that our method gives lower misclassification error rates in comparison to dimension reduction using principal component analysis.

email: yinshux@auburn.edu

ADAPTIVE PREDICTION IN GENOMIC SIGNATURES-BASED CLINICAL TRIALS

Yang Xie*, University of Texas Southwestern Medical Center
Guanghua Xiao, University of Texas Southwestern Medical Center
Chul Ahn, University of Texas Southwestern Medical Center
Luc Girard, University of Texas Southwestern Medical Center
John Minna, University of Texas Southwestern Medical Center

Personalized medicine is defined by predicting the patients' clinical outcomes using their molecular profiling data and other clinical information before treatment and thereby selecting the best possible therapies. To prove the worth of personalized medicine and bring it into clinical practice, well-designed clinical trials are essential steps. I will present a procedure that builds prediction model based on training data, uses this model to predict the best

treatment for individual patients enrolled in the trial, and then updates the model once the outcome of patient is available. The updating is conducted through a re-weighted random forest model accounting for the heterogeneity between training and testing data. Both simulation data sets and oncology data sets are used to show the performance of the method.

email: yang.xie@utsouthwestern.edu

STEREOTYPE LOGIT MODELS FOR HIGH-DIMENSIONAL DATA

Andre A.A. Williams*, Virginia Commonwealth University
Kellie J. Archer, Virginia Commonwealth University

Gene expression studies are of growing importance in the field of medicine. In fact, subtypes within the same disease have been shown to have differing gene expression profiles (Golub et al., 1999). Often rather than differentiating disease subclasses researchers are interested in differentiating the same disease by a categorical classification of disease progression. Specifically, it is likely of interest to identify genes that are associated with progression and to accurately predict the state of progression within a disease using gene expression data. One problem when modeling microarray gene expression data is that there are more genes (variables) than there are observations. In addition, the genes usually demonstrate a complex variance-covariance structure. Therefore, modeling a categorical classification of disease progression using gene expression data presents the need for methods capable of modeling high dimensional data with an ordinal outcome. In an attempt to overcome the aforementioned problems, we propose a method that combines the Stereotype logistic model (Anderson, 1984) with an elastic net penalty (Friedman et al. 2009), which is a combination of a ridge and lasso penalty. The proposed method will be applied to simulated data designed to mimic gene expression data and the results will be reported.

email: williamsaa4@mymail.vcu.edu

LIKELIHOOD BASED APPROACH TO GENE SET ENRICHMENT ANALYSIS WITH A FINITE MIXTURE MODEL

Sang Mee Lee*, University of Minnesota
Baolin Wu, University of Minnesota

In this paper, we study a parametric modeling approach to gene set enrichment analysis. Existing methods have largely relied on nonparametric approaches employing, e.g., categorization, permutation or resampling based significance analysis methods. These methods have proven useful yet might not be powerful. By formulating the enrichment analysis into a model comparison problem, we adopt the likelihood ratio based testing approach to assess significance of enrichment. Through simulation studies and application to gene expression data, we will illustrate the competitive performance of the proposed method.

email: leex2919@umn.edu

58. EXPERIMENTAL DESIGN, POWER/SAMPLE SIZE AND SURVEY RESEARCH

SEQUENTIAL DESIGN FOR MICROARRAY STUDIES

Laurent Briollais*, Mount Sinai Hospital, Toronto
Gilles Durrieu, University of Bordeaux, France

A critical aspect in the design of microarray studies is the determination of the sample size necessary to declare genes differentially expressed across different experimental conditions. Here, we propose a sequential approach where the decision to stop the experiment depends on the accumulated microarray data. The study could stop whenever sufficient data have been accumulated to identify gene expression changes across several experimental conditions. The gene expression response is modeled by a robust linear regression model. We then construct a sequential confidence interval for the intercept of this model, which represents the median gene expression at a given experimental condition. We derive the stopping rule of the experiment for both continuous and discrete sequential approaches and give the asymptotic properties of the stopping variable. In our application to a study of hormone responsive breast cancer cell lines, we estimated the stopping variable for the sample size determination to be smaller than the actual sample size available to conduct the experiment. This means that we can obtain an accurate assessment of differential gene expression without compromising the cost and size of the study. Altogether, we anticipate that this approach could have an important contribution to microarray studies by improving the usual experimental designs and methods of analysis.

email: laurent@lunenfeld.ca

TESTS FOR UNEQUAL TREATMENT VARIANCES IN CROSSOVER DESIGNS

Yoon-Sung Jung*, Alcorn State University
Dallas E. Johnson, Kansas State University

Treatments and periods at crossover design are compared within subjects, i.e. each subject serves as his/her own control. Therefore, any effect that is related to subject differences is removed from treatment and period comparisons. Crossover designs both with and without carryover are traditionally analyzed assuming that the response due to different treatments have equal variances. The effects of unequal variances on traditional tests for treatment and carryover difference were recently considered in crossover designs assuming that the response due to treatments have unequal variances with a compound symmetry correlation structure.

An iterative procedure is introduced to estimate the parameters for the two and three treatment crossover designs. To check the performance of the likelihood ratio tests, Type I error rates and power comparisons are explored using simulations.

email: yjung@alcorn.edu

OPTIMAL DESIGNS FOR RESPONSE FUNCTIONS WITH A DOWNTURN

Seung Won Hyun*, University of Missouri
Min Yang, University of Missouri
Nancy Flournoy, University of Missouri

In many toxicological assays, interactions between primary and secondary effects may cause a downturn in mean responses at high doses. In this situation, the typical monotonicity assumption is invalid and may be quite misleading. Prior literature addresses the analysis of response functions with a downturn, but so far as we know, this paper initiates the study of experimental design for this situation. A growth model is combined with a death model to allow for the downturn in mean doses. Several different objective functions are studied. When the number of treatments equals the number of parameters, Fisher information is found to be independent of the model of the treatment means and on the magnitudes of the treatments. In general, A- and DA-optimal weights for estimating adjacent mean differences are found analytically for a simple model and numerically for a biologically motivated model. Results on c-optimality are also obtained for estimating the peak dose and the EC50 (the treatment with response half way between the control and the peak response on the increasing portion of the response function).

email: swhrg7@mail.missouri.edu

POWER ANALYSIS FOR LONGITUDINAL STUDIES WITH TIME DEPENDENT COVARIATE

Cuiling Wang*, Albert Einstein College of Medicine

The research on power analysis for longitudinal studies with time dependent covariate is limited. In this paper we consider power analysis for longitudinal studies in which the associationbetween a time-dependent covariate with a continuous outcome is of primary interest, in presence of drop out. Sample size calculation formulae are provided. Simulation studies show that the formulae perform well. The method is applied to an example from a real epidemiological study design.

email: cuwang@aecom.yu.edu

THE USE OF PERCENTILES FOR ESTIMATING VARIABILITY OF NORMAL AND UNIFORM DISTRIBUTIONS

Chand K. Chauhan*, Indiana University-Purdue University, Fort Wayne
Yvonne M. Zubovic, Indiana University-Purdue University, Fort Wayne

Consider a situation in which the standard deviation of a population is unknown. Further suppose that the only information available from a population is summary statistics of a sample instead of a complete data set. In this paper the authors propose estimators of the population standard deviation when only two percentile values are known. This approach of estimation has practical significance in situations where only certain percentiles (such as 25th and 75th) of a data set are known. The properties of the proposed estimators are investigated. The standard deviations of the proposed estimators are compared with that of the well known estimate, s, calculated from complete data set. The results will be discussed for normal and uniform distributions.

email: chauhan@ipfw.edu

THE IMPACT OF SURVEY ORDER ON IDENTIFIABILITY OF RESPONSE FATIGUE AND ESTIMATION OF TREATMENT EFFECTS

Brian L. Egleston*, Fox Chase Cancer Center

The order of survey questions can affect response patterns. Questions asked later in a long survey are often prone to more measurement error or misclassification. The response given is a function of both the true response and participant response fatigue. We investigate the identifiability of survey order effects and their impact on estimators of treatment effects. We consider linear, Gamma, and logistic models of response that incorporate both the true underlying response and the effect of question order. For continuous data, survey order effects have little impact on study power when all participants are asked questions in the same order. For binary data and for decreasing chance of a positive response (<1/2), order effects cause power to increase under a linear probability (risk difference) model, but decrease under a logistic model. The results suggest that measures designed to reduce survey order effects might have unintended consequences. A data example is discussed.

email: brian.egleston@fccc.edu

59. COMBINING MULTIPLE SOURCES OF EXPOSURE DATA IN HEALTH ENVIRONMENT STUDIES

BAYESIAN GRAPHICAL MODELS FOR COMBINING MULTIPLE DATA SOURCES, WITH APPLICATIONS IN ENVIRONMENTAL EPIDEMIOLOGY

Sylvia Richardson*, Imperial College London, UK
Alexina Mason, Imperial College London, UK
Lawrence McCandless, Simon Fraser University, Canada
Nicky Best, Imperial College London, UK

The study of the influence of environmental and socio-economic risk factors on health is typically based on observational data, and typically a single data set may not provide sufficient information for valid inference. So multiple data sources are often required, some will contain detailed information on a small sample of individuals, while others will have only a limited number of variables for a large population, and miss important confounders. Building models that can link various sources of data and adequately account for uncertainty arising from missing or partially observed confounders in large data bases present a number of statistical challenges. Bayesian graphical models can be used to fit a common regression model to a combination of data sets with different sets of covariates, with propagation of information between the model components. We will discuss the benefits and difficulties of using these models for carrying out the synthesis of datasets of different designs, with particular emphasis on issues of propagation of information using different factorisations of the marginal likelihood. The discussion will be illustrated by a case study on the effect of water disinfection by-products on low birth weight, where different strategies for combining the data sets and adjusting for partially observed confounders will be explored and compared.

email: sylvia.richardson@imperial.ac.uk

NONLINEAR LATENT PROCESS MODELS FOR ADDRESSING TEMPORAL CHANGE OF SUPPORT IN SPATIO-TEMPORAL STUDIES OF ENVIRONMENTAL EXPOSURES

Nikolay Bliznyuk*, Texas A&M University
Christopher Paciorek, University of California-Berkeley
Brent Coull, Harvard University

Spatio-temporal prediction of levels of an environmental exposure is an important problem in environmental epidemiology. When multiple sources of exposure information are available, a joint model that pools information across sources maximizes data coverage over both space and time, thereby reducing the prediction error. We consider a Bayesian hierarchical framework where a joint model consists of a set of submodels, one for each data source, and a model for the latent process that serves to relate the submodels to one another. However, if a submodel depends on the latent process nonlinearly, inference using standard MCMC techniques can be computationally prohibitive. To make such problems tractable, we 'linearize' the nonlinear components with respect to the latent process and induce sparsity in the covariance matrix of the latent process using compactly supported covariance functions. We propose an efficient MCMC scheme that takes advantage of these approximations. We then apply our methods to motivating data on the spatio-temporal distribution of mobile source particles in the greater Boston area. We use our model to address a temporal change of support problem whereby interest focuses on pooling daily and weekly black carbon readings in order to maximize the spatial coverage of the study region.

email: nab36.cornell@gmail.com

MORTALITY RISKS OF SHORT AND LONG-TERM EXPOSURE TO CHEMICAL COMPOSITION OF FINE PARTICULATE AIR POLLUTION (2000-2007): STATISTICAL CHALLENGES

Francesca Dominici*, Harvard University

Population-based studies have estimated health risks of short-term and long-term exposure to fine particles using mass of PM2.5 (particulate matter < 2.5 micrometers in aerodynamic diameter) as the indicator. Evidence regarding the toxicity of the chemical components of the PM2.5 mixture is limited. We used a national dataset comprising daily data for 2000-2007 on mortality and hospital admissions for cardiovascular and respiratory outcomes, ambient levels of major PM2.5 chemical components (sulfate, nitrate, silicon, elemental carbon, organic carbon matter, sodium and ammonium ions), and weather. By linking chemical components of PM2.5 data to the Medicare billing claims by zip code of residence of the enrollees, we have developed a new retrospective cohort study, the Medicare Cohort Air Pollution Cohort Study. We develop regression models for estimating relative risks associated with short and long-term exposure to chemical components adjusted by temporal and spatially varying confounders. We found that ambient levels of elemental carbon and organic carbon matter, which are generated primarily from vehicle emissions, diesel, and wood burning, were associated with the largest risks of emergency hospitalization across the major chemical constituents of fine particles.

email: fdominic@hsph.harvard.edu

60. IMAGINING 2025: THE HOPES FOR CLINICAL TRIALS

PREDICTING THE PREDICTABLE? CLINICAL TRIALS IN 2025

Steven N. Goodman*, Johns Hopkins University

As with the stock market and global warming, it can be easier to predict long term than short term trends. It is therefore not hard to predict what the world of clinical trials will look like decades hence. The global information economy will affect clinical trials as they have every other domain, and they will also be affected by changes in the US health care system. Clinical trials will be more carefully designed to provide answers to questions asked by decision makers, with results both more widely accessible and reproducible. There will be far more emphasis on efficient designs, and social-scientific structures are developing that will minimize unnecessary experimentation and maximize the value of information that is produced. This talk will sketch out what this clinical trials future might look like, based both on initiatives happening today, and new Apple technologies scheduled for release in 2024.

email: sgoodman@jhmi.edu

PEERING INTO THE HOPEFUL CRYSTAL BALL: CLINICAL TRIALS IN 2025

Janet Wittes*, Statistics Collaborative

World population, 6.5 billion now; is projected to increase to 8 billion (plus or minus) in 2025. In 1950, an estimated 20 percent of the world's population was malnourished; today about 50% are; in 2025, who knows? Today, about half of the world's population live in cities of more than one million people; by 2025, that figure is expected to grow to two-thirds. Thus the diseases attendant on urbanization are likely to become even more important than they are now. In the developed world, under 2 percent of deaths are caused by infectious disease; in the third world, the figure is closer to 40%. What will be the figure in 2025? When we statisticians think of clinical trials, we often focus on methodology. We wonder about such questions as: what techniques will be available for analyzing missing data? for dealing with non-proportional hazards? and will the Bayes-classical chasm be closed? But perhaps more important are questions related to the type oft trials that will be carried out, for what populations, and under whose sponsorship. In this talk, I make some conjectures about the types of trials that will be important in the next decades and pose some questions addressing how we as statisticians can contribute most effectively.

email: janet@statcollab.com

A BAYESIAN 21ST CENTURY?

Donald A. Berry*, University of Texas M D Anderson Cancer Center

The Bayesian approach is being used increasingly as a tool for developing more efficient clinical trial designs. The designs tend to be more complicated than those for standard trials with fixed allocations to treatment. However, they are completely specified prospectively and so their operating characteristics (type I error, power, etc.) can be found, if only by simulation. So the Bayesian approach provides a tool for building an efficient frequentist design. This is a positive step and it is not very controversial. The question is whether it is the limit of the extent to which the Bayesian approach will be used in medical research. I will address this question, especially in the context of decision analysis. For example, by 2025 will we be taking a decision-analytic approach in rare diseases (such as most pediatric cancers) where the explicit goal will be to treat as many patients as effectively as possible and only incidentally to achieve some level of power for distinguishing treatments? And how far will we have gotten in our quest of addressing personalized medicine within the context of clinical trials?

email: dberry@mdanderson.org

61. INNOVATIVE METHODS IN BIOSURVEILLANCE: ACCURATE METHODS FOR RAPID DETECTION OF EVENTS AND PATTERNS

USING SPATIOTEMPORAL REGRESSION METHODS TO IDENTIFY CAUSES OF DISEASE OUTBREAKS

Michael R. Kosorok*, University of North Carolina-Chapel Hill
Yingqi Zhao, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill
Amy H. Herring, University of North Carolina-Chapel Hill
David Richardson, University of North Carolina-Chapel Hill

The goal of a disease surveillance system is to detect outbreaks, or excesses of disease. Once detected, often the next step is to identify the causes of an outbreak. One approach is to conduct a review of available records to see if one or more possible causes emerge. A second, more principled approach is to use spatiotemporal regression to identify unusual conditions among potential explanatory variables such as temperature, precipitation, ozone level, etc. Those variables with unusual values occurring before the disease outbreak under investigation could be considered possible causes. We develop a new methodology to identify the causes of an outbreak and present some initial results on performance. We also apply our methodology to surveillance data from North Carolina.

email: kosorok@unc.edu

FAST SUBSET SCANNING FOR MULTIVARIATE EVENT DETECTION

Daniel B. Neill*, Carnegie Mellon University

This talk will present several new approaches for rapid detection of emerging events (e.g. outbreaks of disease) in multivariate space-time data. I will briefly review our recently proposed method of 'linear-time subset scanning' (LTSS), which enables very fast identification of the most significant (unconstrained) subset of the monitored locations. I will present three main extensions of

LTSS which incorporate constraints on spatial proximity, graph connectivity, and data similarity respectively. The resulting methods include 'fast localized scan' for detection of irregularly shaped but proximity-constrained spatial clusters, 'fast graph scan' for detection of significant connected subgraphs, and 'fast generalized scan' for identification of anomalous groups of data records in arbitrary multivariate datasets.

email: neill@cs.cmu.edu

ADJUSTMENTS FOR SECONDARY MULTIPLICITY PROBLEMS WITH SCAN STATISTICS

Ronald Gangnon*, University of Wisconsin-Madison

Cluster detection approaches based on scan statistics often raise secondary multiple comparison problems in addition to the primary multiple comparison problem. Examples of these secondary multiple comparison problems include adjustments for local multiplicities and significance testing based on the Pareto set of cluster solutions. We discuss a general approach for addressing both the primary and secondary multiplicity problems which combines a one- or two-parameter Gumbel distribution approximations based on a small number of Monte Carlo simulations to address the secondary multiple comparison problems with a larger, second-stage Monte Carlo simulation to address the primary multiple comparison problem.

email: ronald@biostat.wisc.edu

62. INNOVATIVE STATISTICAL METHODS FOR FUNCTIONAL AND IMAGE DATA

FDA FOR TREE-STRUCTURED DATA OBJECTS

J. S. Marron*, University of North Carolina-Chapel Hill
(Not Presenting)

The field of FDA has made a lot of progress on the statistical analysis of the variation in a population of curves. A particularly challenging extension of this set of ideas, is to populations of tree-structured objects. Deep challenges arise, which involve a marriage of ideas from statistics, geometry, and numerical analysis, because the space of trees is strongly non-Euclidean in nature. These challenges, together with some first approaches to addressing them, are illustrated using a real data example, where each data point is the tree of blood vessels in one person's brain.

email: marron@email.unc.edu

BAYESIAN SPARSE FACTOR MODELS FOR MULTIVARIATE FUNCTIONAL DATA

David B. Dunson*, Duke University

In analyzing multivariate functional data, there is a need for models that sparsely characterize heterogeneity in the individual functions across individuals, while characterizing the within- and between-function dependence. In addition, covariate effects are of interest. To address these challenges, we propose a flexible class of sparse Bayesian latent factor models, which characterize the high-dimensional basis coefficients specific to the different functions as dependent on a low-dimensional set of latent factors. These latent factors can be infinite-dimensional, but with the loadings are higher indexed factors shrunk to near zero. Efficient adaptive blocked Gibbs algorithms are developed to model average across low rank approximations. By mixing low rank Gaussian factor models, we obtain a flexible specification in which the multivariate functions can be characterized approximately as points on a low-dimensional manifold.

email: dunson@stat.duke.edu

AUTOMATED, ROBUST ANALYSIS OF FUNCTIONAL AND QUANTITATIVE IMAGE DATA USING FUNCTIONAL MIXED MODELS AND ISOMORPHIC BASIS-SPACE MODELING

Jeffrey S. Morris*, University of Texas M.D. Anderson Cancer Center
Veerabhadran Baladandayuthapani, University of Texas M.D. Anderson Cancer Center
Hongxiao Zhu, University of Texas M.D. Anderson Cancer Center

In this talk, I will describe flexible new automated methods to analyze functional and quantitative image data. The methods are based on functional mixed models, a framework that can simultaneously model multiple factors and account for between-image correlation. We use an isomorphic basis-space approach to fitting the model, which leads to efficient calculations and adaptive smoothing yet flexibly accommodates the complex features characterizing these data. The method is automated and produces inferential plots indicating regions of the function or image associated with each factor, simultaneously considering practical and statistical significance, and controlling the false discovery rate. We discuss a robust modeling approach that allows the method to accommodate outlying curves or images, and is flexible enough to even handle data with very heavy tails. Our simulation studies demonstrate how the method is able to down-weight the outliers and obtain effective inference in heavy-tailed settings.

email: jefmorris@mdanderson.org

REDUCED-RANK t MODELS FOR ROBUST FUNCTIONAL DATA ANALYSIS

Daniel Gervini*, University of Wisconsin-Milwaukee

Outlying curves often occur in functional or longitudinal datasets, and can be very influential on parameter estimators and very hard to detect visually. In this talk we present estimators of the mean and the principal components that are resistant to, and can be used for detection of, outlying trajectories. The estimators are based on reduced-rank t-models and are specifically aimed at sparse and irregularly sampled functional data. We will show applications to the analysis of Internet traffic data and of glycated hemoglobin levels in diabetic children.

email: gervini@uwm.edu

63. JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

AN ESTIMATION METHOD OF MARGINAL TREATMENT EFFECTS ON CORRELATED LONGITUDINAL AND SURVIVAL OUTCOMES

Qing Pan*, George Washington University
Grace Y. Yi, University of Waterloo

We study correlated longitudinal and survival processes. The marginal mean of the longitudinal outcome is of primary interest. The difference between treatment-specific longitudinal measures are often not constant over time in the presence of treatment-specific survival distributions. Hence we propose to quantify treatment effect based on the cumulative differences in the longitudinal outcomes averaging over subjects with and without the event. We separately estimate the event probabilities, rate of change in longitudinal measures given survival, and rate of change in longitudinal measures given events, then take the average rate of change weighted by event and survival probabilities and integrate over time. Generalized linear mixed model for the longitudinal outcome and piecewise exponential proportional hazards model for the survival outcomes with correlated frailty terms are estimated jointly. The subject-level treatment effects on the survival process and two longitudinal processes are estimated together with the proposed marginal difference. The method is motivated by the study of weight loss resulted from Diabetes Prevention Program where weight after diabetes occurrence is affected by medications and systematically different from diabetes-free weights.

email: qpan@gwu.edu

JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA SUBJECT TO NON-IGNORABLE MISSING AND LEFT-CENSORING IN REPEATED MEASUREMENTS

Abdus Sattar*, University of Pittsburgh

Objective of this study is to analyze differential reasons for missing and left-censored longitudinal continuous biomarker and time-to-mortality data using joint modeling. In this joint modeling approach, we have utilized weighted random effects Tobit model (Sattar, 2009) to account for the missing due to dropouts, deaths, administrative reasons, etc, and left-censored observations. Weights were computed using inverse probability weighting methodology and considered as nuisance parameters in the estimation and inference process. So, pseudo likelihood theory (Gong and Samaniego, 1981) has been used in making inferences for the parameter of interests in the presence of infinite number of these nuisance parameters. This work has been applied to the genetic and inflammatory marker of sepsis study which was a large cohort study of patients with community acquired pneumonia (Kellum et al, 2007). A simulation study is in underway to measure the performance of the intended joint model.

email: sattarphd@gmail.com

JOINT MODELING OF LONGITUDINAL AND TIME TO EVENT DATA WITH RANDOM CHANGEPOINTS

Chengjie Xiong*, Washington University
Yuan Xu, Washington University

Longitudinally observed disease markers often provide crucial information on the antecedent progression of many diseases such as Alzheimer's disease (AD) and HIV prior to the disease onset. We propose a joint model of longitudinal disease marker data and time to disease onset data which allows a possible antecedent acceleration on the rate of changes on disease markers prior to the disease onset. We rely on the standard general linear mixed models for longitudinal data and the standard Cox proportional hazards model for time to event data and link them with a random changepoint model on the rate of change for disease markers. We provide estimates to the regression parameters in these models as well as to the parameters associated with the changepoints. The proposed model is then demonstrated by using a real world study that seeks to understand the antecedent cognitive changes before the onset of Alzheimer's disease.

email: chengjie@wubios.wustl.edu

JOINT MODELING OF THE RELATIONSHIP BETWEEN LONGITUDINAL AND SURVIVAL DATA SUBJECT TO BOTH LEFT TRUNCATION AND RIGHT CENSORING WITH APPLICATIONS TO CYSTIC FIBROSIS

Mark D. Schluchter, Case Western Reserve University
Annalisa VanderWyden Piccorelli*, Case Western Reserve University

Methods for joint analysis of longitudinal measures of a continuous outcome Y and a time-to-event outcome T have recently been developed either to focus on the longitudinal data Y while correcting for non-ignorable dropout, to predict the survival outcome T using the longitudinal data Y, or to examine the relationship between Y and T. The motivating problem for our work is in joint modeling the serial measurements of pulmonary function (FEV1 % predicted) and survival in cystic fibrosis (CF) patients using registry data, where an additional complexity is that some patients have not been followed from birth, and thus their survival times are left truncated. We assume a linear random effects model for FEV1 % predicted, where the random intercept and slope of FEV1 % predicted, along with a specified transformation of the age at death follow a trivariate normal distribution. We develop an EM algorithm for maximum likelihood estimation of parameters, which takes left truncation as well as right censoring of survival times into account. The methods are illustrated using simulation studies and using data from CF patients followed at Rainbow Babies and Children's Hospital, Cleveland, OH.

email: mds11@case.edu

SEMIPARAMETRIC ESTIMATION OF TREATMENT-FREE RESTRICTED MEAN LIFETIME USING LANDMARK ANALYSIS WITH A PARTLY CONDITIONAL MODEL

Qi Gong*, University of Michigan
Douglas E. Schaubel, University of Michigan

We propose semiparametric methods for predicting restricted mean lifetime in the absence of treatment. The data structure of interest consists of potentially censored survival times and a longitudinal sequence of measurements. In addition, patients may be removed from consideration for treatment. Treatment-free survival time is of interest, and may be dependently censored by the receipt of treatment. The proposed methods involve landmark analysis and partly conditional hazard regression. Dependent censoring is overcome by Inverse Probability of Censoring Weighting. The proposed methods circumvent the need for explicit modeling of the longitudinal covariate process. The predicted quantities are marginal in the sense that time-varying covariates are taken as fixed at each landmark. The proposed estimators are shown to be consistent and asymptotically normal, with consistent covariance estimators provided. Simulation studies reveal that the proposed estimation procedures are appropriate for practical use. We present an application of the proposed method to Scientific Registry of Transplant Recipients data.

email: gongqi@umich.edu

JOINT MODELS OF LONGITUDINAL DATA AND RECURRENT EVENTS WITH INFORMATIVE TERMINAL EVENT

Se Hee Kim*, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill
Lloyd Chambless, University of North Carolina-Chapel Hill

In many biomedical studies, data are collected from patients who experience the same event at multiple times along with longitudinal biomarkers. Additionally, some subjects may experience some terminal event such as death. In this paper, we propose semiparametric joint models to analyze such data. A broad class of transformation models for the cumulative intensity of the recurrent events and the cumulative hazard of the terminal event is considered. We propose to estimate all the parameters using the nonparametric maximum likelihood estimators (NPMLE). We provide the simple and efficient EM algorithms to implement the proposed inference procedure. The asymptotic properties of the estimators are shown to be asymptotically normal and semiparametrically efficient. Finally, we evaluate the performance of the method through extensive simulation studies and a real-data application.

email: skim@bios.unc.edu

64. HEALTH SERVICES RESEARCH

A LOGISTIC REGRESSION MODEL OF OBESITY IN PRE-SCHOOL CHILDREN

MaryAnn Morgan-Cox*, Baylor University
Veronica Piziak, M.D., Scott & White Medical Center
Jack D. Tubbs, Baylor University
James D. Stamey, Baylor University
John W. Seaman, II, Baylor University

Body mass index (BMI) for age is the standard method to identify and follow overweight children. The population as a whole has grown more obese and obesity in children has increased alarmingly since that time. We present a study in which age, gender, height, and weight measurements were obtained from 18,462 children who participated in the Head Start program from Fall 2003 through Spring 2008. Specifically, data was collected from the Head Start centers in several South Texas border counties and one Central Texas county. We compare our results to the Mexican American cohort of the National Health and Nutrition Examination Survey (NHANES) sample consisting of 2-5 year old children. We implement logistic regression, including year, gender, and county as covariates in the initial model. No statistically significant increase

is found to exist between estimates for the years, and the data for the six years are combined. Differences were found to exist between some of the counties, and a gender effect was found to be significant at the 85 percent cut point. We find the prevalence of high BMI-for-age to be much higher in each of the Texas counties than that reported in the JAMA study. A Bayesian logistic regression model is also considered.

email: maryann_morgan-cox@baylor.edu

ASSESSING CONTENT VALIDITY THROUGH CORRELATION AND RELEVANCE TOOLS: A BAYESIAN RANDOMIZED EQUIVALENCY EXPERIMENT

Byron J. Gajewski*, University of Kansas
Valorie Coffland, University of Kansas
Diane K. Boyle, University of Kansas
Marjorie Bott, University of Kansas
Jamie Leopold, University of Kansas
Nancy Dunton, University of Kansas

A criticism of content validity indices of psychometric instruments is that they provide weak justification for acceptable scores. Traditionally, content validity elicits expert opinion regarding items' relevancy to a domain. This study developed an alternative tool that elicits expert opinion regarding correlations between each item and its respective domain. With 109 Registered Nurse (RN) site coordinators from National Database of Nursing Quality Indicators® (NDNQI®), we implemented a randomized Bayesian equivalency trial with coordinators completing 'relevance' or 'correlation' content tools regarding the RN Job Enjoyment (JE) scale. We confirmed our hypothesis that the two tools would result in equivalent content information. A Bayesian ordered analysis model supported the results, suggesting traditional content validity can be justified using correlation arguments.

email: bgajewski@kumc.edu

HIERARCHICAL BAYESIAN MODELS TO QUANTIFY HOSPITAL PERFORMANCE

Yulei He*, Harvard Medical School
Sharon-Lise T. Normand, Harvard Medical School
Robert Wolf, Harvard Medical School

The study of health care quality is a central activity of health services and outcomes research. Process-based measures are often used as quality indicators to quantify the extent to which a provider complies with evidence-based care guidelines (e.g., the administration of aspirin to reduce mortality in acute myocardial infarction). Hospital performance cards based on such process measures typically include the number of eligible patients for evidence-based therapies ("denominator") and the number of patients receiving therapies among the number eligible for each therapy ("numerator"). A natural hospital quality indicator is the treatment rate among those eligible for a particular therapy. A popular estimation approach is to use the fraction of the treated among those eligible. An alternative approach involves estimation of a hierarchical Bayesian model for the number receiving therapy conditional on number eligible and then permit between-hospital variation in the true rates. However, both approaches ignore the information on patient eligibility. To incorporate the latter, we propose a two-stage hierarchical Bayesian model in which the first stage (the eligibility stage) characterizes the inclusion probability for measure eligibility conditional on patient admission, and the second stage (the treatment stage given eligibility) characterizes the probability for receipt of therapy conditional on patient eligibility. The two probabilities are correlated across hospitals and the underlying true rates vary across hospitals. We compare the performance among different approaches using theoretical derivations and Monte Carlo simulations. Our study suggests that the least preferred approach is using the raw fraction. In addition, estimates of the treatment rate can be adequately obtained using the treatment-stage model. Using the two-stage model can lead to further improved estimates. We illustrate findings using the National Hospital Compare data.

email: he@hcp.med.harvard.edu

ANALYSIS OF INTERVAL-GROUPED RECURRENT EVENT DATA WITH APPLICATION TO NATIONAL HOSPITALIZATION DATA

Dandan Liu*, University of Michigan
Jack D. Kalbfleisch, University of Michigan
Douglas E. Schaubel, University of Michigan

Centers are often evaluated based on the post-treatment outcome rates experienced by their patients (e.g., comparison with an overall average). The use of large observational databases for such purposes may introduce computational difficulties, particularly when the event of interest is recurrent. In such settings, grouping the recurrent event data according to pre-specified intervals leads to a flexible event rate model and a data reduction which remedies the computational issues. We propose a possibly stratified marginal Poisson model with a piecewise constant baseline event rate. Large-sample distributions are derived for the proposed estimators. Simulation studies are conducted under various data configurations, including settings in which the model is misspecified. We then show that the proposed procedures can be carried out using standard statistical software (e.g., SAS, R). An application based on national hospitalization data is provided.

email: dandanl@umich.edu

IMPLEMENTATION OF A KRONECKER PRODUCT CORRELATION STRUCTURE FOR THE ANALYSIS OF UNBALANCED DATA

Arwin M. Thomasson*, University of Pennsylvania
Hanjoo Kim, Forest Labs
Justine Shults, University of Pennsylvania
Russell Localio, University of Pennsylvania
Harold I. Feldman, University of Pennsylvania
Peter P. Reese, University of Pennsylvania

Medical studies often yield data with multiple sources of correlation. For example, a quality-of-care measure applied to multiple transplant centers over multiple years could be correlated in two ways--the scores for each center could be correlated over time, and the center measurements could be correlated within regions. We present an approach for analysis of unbalanced multi-level correlated data that implements a Kronecker product (KP) structure with quasi-least squares (QLS). While previous researchers primarily implemented the KP structure for analysis of balanced data with a constant number of measurements per subject, our approach allows for consideration of unbalanced data (Kim, et al., 2009). We implement our approach in an analysis of a standardized live-donor transplant rate (SLDTR) that identifies characteristics of transplant centers that do not fully utilize live-donor kidney transplants. We also explore the issue of grouping transplant centers into donor service areas, and we describe SAS software that we have developed for implementation of our approach.

email: arwin@mail.med.upenn.edu

POWER IN THE DESIGN OF TWO-LEVEL RANDOMIZED TRIALS

Yongyun Shin*, Virginia Commonwealth University

Experimental research studies involve randomization of individuals to treatments who are nested within a cluster or a site. In cluster-randomized trials, dominant designs and interventions involve random assignment of whole hospitals or schools, rather than patients or children, to treatments. In designing a multilevel randomized trial, it is essential to select optimal sample sizes that achieve the desired power of the test for a factor effect at any of the involved levels and minimize cost. Computation of such power may involve a cluster-level factor, an individual-level factor, or their interaction between the same-level factors or cross-level factors. A factor may take two or more values. This multilevel setting also arises for repeated measures within a patient, workers within a firm, doctors within a hospital, and adults living in a neighborhood. This paper is concerned with a general method for power analysis of balanced two-level hierarchical linear models.

email: yshin@vcu.edu

65. MODELS INVOLVING LATENT VARIABLES

A PENALIZED MAXIMUM LIKELIHOOD APPROACH TO SPARSE FACTOR ANALYSIS

Jang Choi*, University of Minnesota
Hui Zou, University of Minnesota
Gary Oehlert, University of Minnesota

Factor analysis is used to analyze the data in terms of a smaller number of random quantities than variables. One popular method of estimation is obtained by factor rotation. However, factor rotation does not estimate factor loadings properly if data are sparse. In this paper, penalized maximum likelihood methods of factor analysis are used to obtain the loadings, and show proper interpretation. Methods for selecting the number of common factors and penalization parameter are suggested. An efficient algorithm is developed.

email: choix122@umn.edu

MODERN CLUSTER METHODS FOR INCOMPLETE LONGITUDINAL DATA

Liesbeth Bruckers*, Hasselt University
Geert Molenberghs, Hasselt University

Recently, a lot of work has been done from a methodological perspective so as to allow cluster analysis to cope with complex data structures, such as repeated measures Muthén and Muthén (1998-2001) proposed a general growth mixture modeling (GGMM) framework. Conventional growth models are used to describe individual and average trajectory of a measurement over time (Verbeke and Molenberghs,2000). Individual differences in evolution are captured by introducing patient-specific parameters. These random effects are assumed to follow a normal distribution. GMM relaxes the assumption of a single population by allowing for different, unknown classes of individuals to vary around different mean growth curves. Missing data are inevitable when collecting information about patients. Proper handling of missing values in statistical analyses is important. Estimation of the GGMM parameters can be done conveniently via maximum likelihood, and thus yield correct inferences for data which are MAR. However, different techniques can be used to tackle the problem of missing data. The challenge is to incorporate them in GGMM and to study the effect of missing data on the number of clusters, the estimated trajectories, the posterior membership probabilities?

email: liesbeth.bruckers@uhasselt.be

LATENT VARIABLE MODELS FOR DEVELOPMENT OF COMPOSITE INDICES

Xuefeng Liu*, East Tennessee State University
Meng Liu, East Tennessee State University
Kesheng Wang, East Tennessee State University
Jeffray Roth, University of Florida

Composite scores are usually developed to capture common characteristics which are not easy to measure in areas of human behavior, psychology, health and clinical sciences. Latent variable modes with normally-distributed trait can be used to construct the comprhensive index. Howerver, normal assumption does not hold in many cases. We propose an extended latent trait model in which the latent trait is non-normal and the conditional probability of each outcome is modeled as a nonlinear function of the latent trait which has properties similar to the logistic function. A modified Gauss-newton algorithm for multiple multinomial outcomes is developed for parameter estimation. The model is applied to an infant morbidity study in which there are four manifest morbidity outcomes. A composite variable, called infant mobidity index (IMI) which is a summary of these four infant morbidity outcomes and represents propensity for infant morbidity, is developed. It has been shown that IMI is correlated with each of individual outcomes, with infant mortality and with a face-valid index of morbidity outcomes, and could be used in future research as a measure of infants propensity for morbidity.

email: lix01@etsu.edu

EFFICIENCY OF LIKELIHOOD BASED ESTIMATORS IN THE ANALYSIS OF FAMILIAL BINARY VARIABLES

Yihao Deng*, Indiana University Purdue University-Fort Wayne
Roy T. Sabo, Virginia Commonwealth University
N. Rao Chaganty, Old Dominion University

Data consisting of clusters of familial responses typically exhibit some sort of dependence. Correlated binary outcome measures on those family members can lead to problematic analysis using standard non-likelihood-based repeated measure methodologies. In this paper, we derive maximum likelihood estimators based on the multivariate probit model, and compare its efficiency with standard estimation methodologies. We motivate this analysis with a real life data example on the effect of erythrocyte adenosine triphosphate (ATP) levels among family members.

email: dengy@ipfw.edu

A STRUCTURED LATENT CLASS MODEL VERSUS A FACTOR MIXTURE MODEL FOR EXPLORING CLUSTERS OF OBESOGENIC BEHAVIORS AND ENVIRONMENTS IN ADOLESCENTS

Melanie M. Wall*, University of Minnesota

The latent class model (LCM) is a model based clustering method typically used for clustering individuals into k distinct classes (clusters) based on p ordered categorical variables. The basic assumption of LCM is that all p variables are conditionally independent given class (cluster) membership. Factor mixture models (FMM) extend the traditional factor analysis model for p continuous or ordered categorical variables such that the q (q less than p) underlying continuous latent factors are assumed to come from a k component mixture. Unlike LCM which directly clusters individuals based on all p variables, the FMM reduces the dimension from p observed variables to q latent factors and then clusters individual on the q continuous latent factors. Motivated by the FMM, a structured LCM is considered in this talk that partitions the p variables into smaller groups of variables that are indicative of particular categorical aspects of the underlying clusters. The performance of these models will be compared via simulated data and used in an epidemiological example where it is desired to explore potential clusters of adolescents based on over 50 self-reported survey variables identified as obesity risk factors measuring adolescents' behaviors and environments.

email: melanie@biostat.umn.edu

CONFIDENCE INTERVALS FOR THE DIFFERENCE IN TPRs AND FPRs OF TWO DIAGNOSTIC TESTS WITH UNVERIFIED NEGATIVES

Eileen M. Stock*, Baylor University

We derive two likelihood-related confidence intervals and a pseudo likelihood-based confidence interval for estimating the difference in the true positive rates and false positive rates of two dichotomous screening or diagnostic tests applied to two populations. The populations have varying disease prevalences with unverified negatives. In particular, we examine the efficacy of a Wald-type confidence interval, a score-based confidence interval, and a profile-likelihood confidence interval by comparing interval widths and coverage properties for a spectrum of different specificities and sensitivities for varying sample sizes.

email: Eileen_Stock@baylor.edu

SPARSE BAYESIAN INFINITE FACTOR MODELS

Anirban Bhattacharya*, Duke University
David B. Dunson, Duke University

We focus on sparse modeling of high-dimensional covariance matrices using Bayesian latent factor models. We propose a multiplicative gamma process shrinkage prior on the factor loadings which allows introduction of infinitely many factors, with the loadings increasingly shrunk toward zero as the column index increases. We use our prior on a parameter expanded loadings matrix to avoid the order dependence typical in factor analysis models and develop a highly efficient Gibbs sampler that scales well as data dimensionality increases. The gain in efficiency is achieved by the joint conjugacy property of the proposed prior, which allows block updating of the loadings matrix. We propose an adaptive Gibbs sampler for automatically truncating the infinite loadings matrix through selection of the number of important factors.

Theoretical results are provided on the support of the prior and truncation approximation bounds. A fast algorithm is proposed to produce approximate Bayes estimates. Latent factor regression methods are developed for prediction and variable selection in applications with high-dimensional correlated predictors. Operating characteristics are assessed through a number of simulation studies.

email: ab179@stat.duke.edu

66. NEXT GENERATION SEQUENCING AND TRANSCRIPTION FACTOR BINDING SITES

IDENTIFICATION OF miRNAs IN NEXT-GENERATION SEQUENCING DATA

W. Evan Johnson*, Brigham Young University

We present a method for identifying small RNA molecules, called miRNAs, that regulate genes in the cell by interfering with the gene's transcribed mRNAs and targeting them for degradation. In the rst step of our modeling procedure to identify miRNAs, we apply an innovative dynamic linear model that identies candidate miRNA genes in high-throughput sequencing data. The model is very flexible and can accurately identify interesting biological features while naturally accounting for both the read count, read spacing, and sequencing depth. Additionally, miRNA candidates are also processed using a modied Smith-Waterman sequence alignment that scores the regions for potential RNA hairpins, which one of the major characteristics of miRNAs. We illustrate our method simulated data sets as well as on a small RNA Caenorhabditis elegans data set from the Solexa/Illumina sequencing platform. These examples show that our method is highly sensitive for identifying known and novel miRNA genes.

email: evan@stat.byu.edu

BAYESIAN HIERARCHICAL MODELS FOR QUANTIFYING METHYLATION LEVELS BY NEXT-GENERATION SEQUENCING

Guodong Wu, University of Alabama-Birmingham
Nengjun Yi, University of Alabama-Birmingham
Devin Absher, HudsonAlpha Institute for Biotechnology
Degui Zhi*, University of Alabama-Birmingham

DNA methylation is an important epigenetic phenomenon that implicated in various aspects of gene regulation and diseases. Recently, next-generation sequencing-based technologies enable DNA methylation profiling at high resolutions and low costs. Methyl-Seq (Brunner, et al., 2009) and Reduced Representation Bisulfite Sequencing (RRBS) (Meissner, et al., 2005) are two such technologies allowing for interrogating the methylation levels of tens of thousands of CpG sites throughout the entire human genome. The rapid development of these technologies promises the prospective of genome-wide association studies for epigenetic changes in a near future. For a biological sample, the methylation level at each CpG sites is quantified by the Beta-value, the percent of DNA molecules being methylated. Current methylation quantification protocols only estimate the mean of Beta-value. However, this estimate can have large and non-uniform variances due to the non-uniform sequencing coverage. Estimating the variance of Beta-values is a prerequisite for epigenetic association studies. We developed new Bayesian hierarchical models for quantifying methylation levels in Methyl-Seq and RRBS. With the Poisson assumption of tag counts at each CpG sites (Lander and Waterman, 1988), we apply MCMC to update un-informative priors on Beta-values and obtain their posterior distribution. We compare our methylation quantifications with existing experimental data.

email: dzhi@uab.edu

GENE CLASS ENRICHMENT ANALYSIS FOR RNA-SEQUENCING

Liyan Gao, University of Alabama-Birmingham
Degui Zhi, University of Alabama-Birmingham
Kui Zhang, University of Alabama-Birmingham
Xiangqin Cui*, University of Alabama-Birmingham

The rapid growing next-generation sequencing technology has the potential to replace the microarray technology in measuring genome-wide gene expression. By obtaining tens of millions of short sequence reads from atranscript population and by mapping these reads to the genome, RNA-seq produces digital (counts) rather than analog signals, which lead to highly replicable results with relatively little technical variation. The RNA-seq data also has the advantage of discovering alternative splice variants and novel transcripts simultaneously with the gene expression measurements. Gene class enrichment analysis has played an important role in identifying associated pathways and biological processes in microarray data analyses. Many gene class analysis methods have been developed based on microarray data. However, the unique properties of RNA-seq data make these methods unfit for the RNA-seq data analysis. Unlike microarray data, the amount of information obtained in RNA-seq data depends on transcript length. Longer transcripts have more reads, and therefore, more information and more power to detect differential expression cross conditions/treatment. To solve this problem, we modified one of the gene class enrichment analysis methods, fisher's exact test, to incorporate the transcript length for a more appropriate test for gene class enrichment in the RNA-seq analysis.

email: xcui@uab.edu

A STATISTICAL FRAMEWORK FOR THE ANALYSIS OF ChIP-SEQ DATA

Pei Fen Kuan*, University of North Carolina-Chapel Hill
Guangjin Pan, Genome Center of Wisconsin
James A. Thomson, University of Wisconsin-Madison
Ron Stewart, University of Wisconsin-Madison
Sunduz Keles, University of Wisconsin-Madison

Chromatin immunoprecipitation followed by direct sequencing (ChIP-Seq) has revolutionized the experiments in profiling DNA-protein interactions and chromatin remodeling patterns. Although this technology offers promising results for surveying large genomes at higher resolution, it is not free of sequencing and other sources of biases. Despite this, most of the existing tools do not consider such biases. We carefully study sources of bias in the underlying data generating process of ChIP-Seq technology by utilizing sequenced naked DNA (non-cross-linked, deproteinized DNA) and develop a model that captures the background signal in the ChIP-Seq data. We then proposed mixture models for analyzing ChIP-Seq data. Our modeling framework incorporates the variability in both the mappability and GC-content of regions on the genome and sequencing depths of the samples. We show that our model fits very well on real data and provides a fast model-based approach for ChIP-Seq data analysis.

email: kuanp@stat.wisc.edu

A COMPARISON OF MONTE-CARLO LOGIC AND LogicFS REGRESSION METHODS FOR IDENTIFYING IMPORTANT CO- REGULATORS OF GENE EXPRESSION WITH APPLICATION TO A STUDY OF HUMAN HEART FAILURE

Yun Lu*, University of Pennsylvania School of Medicine
Sridhar Hannenhalli, University of Pennsylvania School of Medicine
Thomas Cappola, University of Pennsylvania School of Medicine
Mary Putt, University of Pennsylvania School of Medicine

Multiple transcription factors (TFs) are thought to co-regulate gene expression in human heart failure. Logic regression is an adaptive regression method for identifying Boolean combinations of important binary predictors of outcome. Logic regression, in its original form, was previously suggested as a method for identifying combinations of TFs that co-regulate gene expression. Here we compare the ability of two extensions of logic regression, Monte-Carlo Logic (MC Logic) regression and LogicFS regression, to both identify important predictors and to do so in the predictors' correct form. We use a novel simulation where we seed simulated main effects and interactions, of a realistic effect size, into an existing human heart failure whole-genome expression data set. We propose two new metrics to facilitate direct comparison of the methods. Simulation results confirm that MC Logic is able to detect important predictors in their correct form. In contrast, LogicFS frequently overfits the model. The resulting output is a list of spurious combinations of predictors of interest. The "importance score" from LogicFS, proposed as a metric for ranking and evaluating combinations of predictors, is inconsistent, and sometimes misleading. We conclude that MC Logic is a superior tool for identifying promising regulators and co-regulators of gene expression.

email: luyunpku@yahoo.com

DETECTION AND REFINEMENT OF TRANSCRIPTION FACTOR BINDING SITES USING HYBRID MONTE CARLO METHOD

Ming Hu*, University of Michigan
Jindan Yu, University of Michigan and Northwestern University
Jeremy Taylor, University of Michigan
Arul Chinnaiyan, University of Michigan
Zhaohui Qin, University of Michigan

Coupling chromatin immunoprecipitation (ChIP) with recently developed massively parallel sequencing technologies has enabled genome-wide detection of protein-DNA interactions with unprecedented sensitivity and specificity. In this study, we explore the value of using ChIP-Seq data to better detect and refine transcription factor binding sites (TFBS). We introduce a novel computational algorithm named Hybrid Motif Sampler (HMS), specifically designed for TFBS motif discovery in ChIP-Seq data. Simulation studies demonstrate favorable performance of HMS compared to other existing methods. When applying HMS to real ChIP-Seq datasets, we find that (i) the accuracy of existing TFBS motif patterns can be significantly improved; and (ii) there is significant intra-motif dependency inside all the TFBS motifs we tested. These findings may offer new biological insights into the mechanisms of transcription factor regulation.

email: hming@umich.edu

67. VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA

RANDOM FORESTS: A SUMMARY OF THE ALGORITHM AND A DESCRIPTION OF THE FEATURES THAT ARE AVAILABLE, INCLUDING A PRESENTATION OF RECENT WORK ON VARIABLE IMPORTANCE AND PROXIMITIES

Adele Cutler*, Utah State University

Leo Breiman and I were working together on random forests from late in 2000 to his death in 2005. A Random Forest is a collection of classification trees, each generated by bootstrap sampling from a training set with random sampling of predictor variables at each node. It has been widely adopted in many fields, and especially in the Biomedical and Pharmaceutical arena, because it is not only an

extremely accurate prediction machine, but it is also able to give the user valuable insights into the data. Random Forests can handle thousands of variables with small or large sample sizes. This talk gives a summary of the random forests algorithm and a description of the features that are available, including a presentation of recent work on variable importance and proximities. I will describe the methods we use to compute proximities and variable importance measures and illustrate their use with case studies showing important steps in Random Forests data analysis with examples related to Autism, Multiple Sclerosis and Microarray Data. Background: Comparative tests show Random Forests to be neck and neck with the best current prediction algorithms such as Support Vector Machines, but RF is more suited to statistical applications because: it is interpretable; it has an uncanny ability to detect and rank the important variables; and, it derives an intrinsic similarity between cases and uses the similarity to project down to two dimensions showing fascinating and unsuspected aspects of the data. These similarities are also used to detect outliers and provide a very effective method for filling in missing data. Another use lets the analyst focus on high class-density areas and see the distribution of variables in these areas, thus giving the user understanding of which variables are driving the classification. Unbalanced data sets, where the class of interest is much smaller than the other classes are becoming more frequent. An innocent classifier will work on getting the large classes right while getting a high error rate on the small class. RF has an effective method for giving balanced results in highly unbalanced data. Variable importance can be measured both locally and globally. Proximities allow us to view the data in illuminating ways and are also useful for detecting outliers, imputing missing values, and extracting clustering information.

email: mfajardo@salford-systems.com

NONPARAMETRIC INDEPENDENCE SCREENING IN ULTRA-HIGH DIMENSIONAL ADDITIVE MODELS

Jianqing Fan, Princeton University
Yang Feng*, Princeton University
Rui Song, Colorado State University

A variable screening procedure via correlation learning was proposed in Fan and Lv (2008) to reduce dimensionality in sparse ultra-high dimensional models. Even when the true model is linear, the marginal regression can be highly nonlinear. To address this issue, we further extend the correlation learning to marginal nonparametric learning. Our nonparametric independence screening is called NIS, a specific member of the sure independence screening. Several closely related variable screening procedures are proposed. Under the nonparametric additive models, it is shown that under some mild technical conditions, the proposed independence screening methods enjoy a sure screening property. The extent to which the dimensionality can be reduced by independence screening is also explicitly quantified. As a methodological extension, an iterative nonparametric independence screening (INIS) is also proposed to enhance the finite sample performance for fitting sparse additive models. The simulation results and a real data analysis demonstrate that the proposed procedure works well with moderate sample size and large dimension and performs better than competing methods.

email: yangfeng@princeton.edu

GROUPED VARIABLE SELECTION IN HIGH-DIMENSIONAL PARTIALLY LINEAR ADDITIVE COX MODEL

Li Liu*, University of Iowa
Jian Huang, University of Iowa

We study the problem of variable selection and estimation in the partially linear Cox model with high-dimensional data. We approximate the nonparametric components by truncated series expansions with B-spline bases. With this approximation, the problem of variable selection becomes that of selecting the groups of coefficients in the expansion. We apply the group Lasso to obtain an initial solution path and reduce the dimension of the problem and then update the whole solution path with the adaptive group Lasso. The group coordinate descent algorithm is implemented for stable and rapid computation. Simulation studies are carried out to evaluate the finite sample performance of the proposed procedure using several tuning parameter selection methods for choosing the point on the solution path as the final estimator. We demonstrate the proposed approach on two real data examples. We also investigate the theoretical properties regarding selection and estimation consistency of the proposed procedure.

email: li-liu@uiowa.edu

MODEL-FREE FEATURE SELECTION

Liping Zhu*, Penn State University and East China Normal University
Runze Li, Penn State University

Variable selection in ultrahigh dimensional feature space characterizes various contemporary problems in scientific discoveries. In this paper, we propose a model-free independence screening procedure to select the subset of active predictors by using the diagonal elements of an average partial mean estimation matrix. The new proposal possesses the sure independence screening property for a wide range of semi-parametric regressions, i.e. it guarantees to select the subset of active predictors with probability approaching to one as the sample size diverges. In addition, it is computationally efficient in the sense that it is free of tuning and avoids completely iterative algorithm. By adding a series of auxiliary variables to set up a benchmark for screening, a new technique is introduced to reduce the false discovery rate in the feature screening stage. Numerical studies through several synthetic examples and a real data example are presented to illustrate the methodology. The empirical investigations found that the new

proposal allows strong correlations within the group of inactive features, and works properly even when the number of active predictors is fairly large.

email: zhulp1@gmail.com

GENERALIZED FORWARD SELECTION: SUBSET SELECTION IN HIGH DIMENSIONS

Alexander T. Pearson, University of Rochester
Derick R. Peterson*, University of Rochester

We consider the problem of subset selection in high dimensions. Faced with a huge number of candidate predictors, as when constructing a prognostic gene expression signature, the goal is to select a parsimonious subset of variables to include in a Cox, logistic, linear, or other generalized linear regression model. We propose a generalized forward selection procedure that conceptually lies between the greedy traditional forward selection method and the computationally infeasible all-subsets search. In contrast to LASSO and other shrinkage-based approaches to this problem, we allow standard unconstrained parameter estimation in the selected models, and we further allow prespecified predictors to be forced into all candidate models. Since it offers more complexities than most other model frameworks, we focus our simulations and data analyses on the Cox model for censored survival outcomes. Our simulation results demonstrate that our generalized forward selection method has a number of advantageous properties for selecting sets of predictive variables, compared with forward selection, univariate screening, and the LASSO. We apply our method to lymphoma gene array data with 7399 genes and 240 patients. Training our method on a subsample of the data leads to predictive results in the independent validation data.

email: peterson@bst.rochester.edu

FEATURE SELECTION IN MICROARRAY DATA USING TIGHT CLUSTERING

Ami Yu*, Korea University
Jae Won Lee, Korea University

In clustering microarray data many clustering algorithms include all genes into clusters. Tight clustering(Tseng and Wong, 2005, Biometrics, 10-16) on the other hand, can find more biologically meaningful gene clusters because it clusters the most informative genes only and excludes genes being included in the clusters unnecessarily. Tseng and Wong applied tight clustering to microarray data and they found some tight and stable clusters that all samples commonly have. In this study we extended their idea and propose a new method that finds tight clusters of genes for each sample and then also clusters samples based on the tight clustering results for feature selection. Through sample clustering features which show similar expression pattern within sample clusters can be obtained. We performed tight clustering for each sample and calculated adjusted Rand index comparing their tight clusters in order to measure degrees of similarity for samples using a simulated dataset. We also applied hierarchical clustering to the samples according to their degrees of similarity.

email: lilyu@korea.ac.kr

68. QUANTILE REGRESSION AND SYMBOLIC REGRESSION

SPATIAL QUANTILE REGRESSION

Kristian Lum*, Duke University
Alan Gelfand, Duke University

Quantile regression aims to explain the quantiles of an outcome variable conditional on covariates. This has typically been done assuming that the outcomes are iid. We will discuss Bayesian quantile regression in the setting of spatially correlated data. We first demonstrate the possible advantage gained by adding the spatial component to this model through simulation. We will discuss this method in the context of explaining the lower quantiles of birth weight relative to the upper quantiles for a sample of infants born in North Carolina from 1997 to 2000. We will also provide a comparison of the effects of smoking on the quantiles of birth weight as gauged through non-spatial and spatial methods. We accommodate this large data set, for which standard spatial models are computationally infeasible, by extending the predictive process model to this quantile model.

email: kcl12@stat.duke.edu

ON RANK SCORE TEST FOR LONGITUDINAL BEST LINE QUANTILE MODEL

Nanshi Sha*, Columbia University
Ying Wei, Columbia University

Bent line quantile model has been shown useful in many applications such as physiology and epidemiology. In this article we consider developing a robust rank score test for longitudinal bent line change-point quantile regression model. We propose two reliable tests for detecting location of change-point and slope coefficient. Through a series of simulation studies we demonstrate that the confidence intervals generated by inverting the proposed rank score tests have certain advantages over those obtained from bootstrap method in terms of shorter lengths, less computational burden and highly robustness against heteroscedasticity. We illustrate the use of the proposed tests by applying them to a real longitudinal HIV study. Estimates and confidence intervals of the treatment-plateau-point are obtained and the long-term treatment effect are evaluated.

email: ns2397@columbia.edu

AN EXACT BOOTSTRAP APPROACH TOWARDS MODIFICATION OF THE HARRELL-DAVIS QUANTILE FUNCTION ESTIMATOR FOR CENSORED DATA

Dongliang Wang*, University of Buffalo
Alan D. Hutson, University of Buffalo
Daniel P. Gaile, University of Buffalo

A new kernel quantile estimator is proposed for right-censored data. The advantage of this estimator is that exact bootstrap methods may be employed to estimate the mean and variance. It follows that a novel solution for finding the optimal bandwidth may be obtained through minimization of the exact bootstrap MSE. We prove the large sample consistency of this estimator for fixed values of the bandwidth parameter. A Monte Carlo simulation study shows that our estimator is significantly better than the product-limit quantile estimator, with respect to various MSE criteria. For general simplicity, setting the bandwidth parameter equal to 1 leads to an extension of classical Harrell-Davis estimator for censored data and performs well in simulations. The procedure is illustrated by an application to lung cancer survival data.

email: dw39@buffalo.edu

BENT LINE QUANTILE REGRESSION WITH APPLICATION TO AN ALLOMETRIC STUDY OF LAND MAMMALS' SPEED AND MASS

Chenxi Li*, University of Wisconsin-Madison
Ying Wei, Columbia University
Rick Chappell, University of Wisconsin-Madison
Xuming He, University of Illinois at Urbana-Champaign

Quantile regression, which models the conditional quantiles of the response variable given covariates, usually assumes a linear model. However, this kind of linearity is often unrealistic in real life. One situation where linear quantile regression is not appropriate is when the response variable is piecewise linear but still continuous in covariates. To analyze such data, we propose a bent line quantile regression model. We derive its parameter estimates, prove that they are asymptotically valid given the existence of a change-point, and discuss several methods for testing the existence of a change-point in bent line quantile regression together with a power comparison by simulation. An example of land mammal maximal running speeds is given to illustrate an application of bent line quantile regression in which this model is theoretically justified and its parameters are of direct biological interests.

email: chenxili@stat.wisc.edu

USING LOGISTIC REGRESSION TO CONSTRUCT CONFIDENCE INTERVALS FOR QUANTILE REGRESSION COEFFICIENTS

Junlong Wu*, University of South Carolina
Matteo Bottai, University of South Carolina

Quantile regression has emerged as a powerful complement to linear mean regression. In recent years various methods for constructing confidence interval for its coefficients have been developed. These methods can be classified into three main categories: methods that estimate the sparsity directly, rank-score methods, and resampling methods. The latter have been generally recommended, though they can be computationally intensive with large sample sizes or numbers of covariates. Inspired by the testing procedure proposed by Redden et al (Stat Med 2004), we develop a new and simple method to construct the confidence interval for quantile regression coefficients based on inverting a score test from a logistic regression likelihood. The testing procedure comprises three steps: (1) estimate the regression model under the null value of the coefficient being tested, (2) create an indicator variable for the outcome above the predicted value, (3) apply a score test from the logistic regression that has the newly created indicator variable as dependent variable and the covariate being tested as the only independent variable. Based of the results of a simulation study the proposed confidence intervals seem to have correct coverage probability, and shorter computation time and similar or sometimes narrower length than some of those currently available.

email: junlongwu5@gmail.com

DETECTION OF TREATMENT EFFECTS BY COVARIATE-ADJUSTED EXPECTED SHORTFALL AND TRIMMED RANKSCORE

Ya-Hui Hsu*, University of Illinois at Urbana-Champaign

The statistical tests that are commonly used in detecting treatment effects suffer from low power when the two distribution functions differ only in the upper (or lower) tail, as in the assessment of the Total Sharp Score (TSS) under different treatments for rheumatoid arthritis. In this talk, we propose two more powerful tests that detect treatment effects through the expected shortfalls and the trimmed rankscores. We examine the validity and the efficiency of the tests under iid as well as more general error structures.

email: yhsu6@illinois.edu

REGRESSION FOR INTERVAL-VALUED SYMBOLIC DATA

Wei Xu*, University of Georgia
Lynne Billard, University of Georgia

The concept of symbolic data was introduced by Diday (1987). Symbolic data include list, intervals, histograms or even distributions. Unlike classical data, symbolic data have internal variations. Therefore, classical analytical methods can not be applied readily. This study focuses on symbolic data linear regression. It begins with the concept of symbolic data, its descriptive statistics and existing linear regression approaches. It then introduces two new approaches, symbolic covariance method and symbolic likelihood method. The symbolic covariance method combines a symbolic covariance structure with least-square techniques to find the regression model. The symbolic likelihood method builds a symbolic likelihood function of the regression model, then transforms it into a classical likelihood function. An algorithm to obtain the parameter estimators is also provided by borrowing ideas from linear mixed models. Examples compare the different methods.

email: xuwei@uga.edu

69. PERSONALIZED THERAPY AND VARIABLE SELECTION IN CLINICAL APPLICATIONS

DEVELOPING ADAPTIVE PERSONALIZED THERAPY FOR CYSTIC FIBROSIS BY REINFORCEMENT LEARNING

Yiyun Tang*, University of North Carolina-Chapel Hill
Michael R. Kosorok, University of North Carolina-Chapel Hill

Personalized medicines hold the promise of being proactive in predicting individual susceptibility and targeting medicines and dosages more precisely and safely for each patient. Besides the need to incorporate genetic and other biomarkers, optimal clinical management of the inherited chronic diseases, such as cystic fibrosis (CF), requires a dynamic approach to cope with the evolving course of illness. In this paper, we propose to use reinforcement learning for constructing optimal adaptive personalized therapy, which uses genetic biomarkers and time varying covariates, alters treatment decisions to achieve a favorable ultimate outcome. We conduct a simulation study of virtual CF patients with Pseudomonas aeruginosa infection and antibiotic therapy with a discrete time non-homogeneous Markov model and parameters tuned to approximately match real studies of Wisconsin CF neonatal screening project. A temporal difference reinforcement learning method with a Markovian assumption called fitted Q-iteration is utilized to discover the optimal treatment regimen from this disease progression. Our simulation results show the great capacity of reinforcement learning for discovering personalized therapy which optimise benefit-risk trade off and in a multi-stage decision making context to improve long term outcomes in chronic disease.

email: ytang@bios.unc.edu

PENALIZED MODELS FOR ORDINAL RESPONSE PREDICTION: APPLICATION DISCRIMINATING PATIENTS WITH EARLY-STAGE PARKINSON'S DISEASE

Kellie J. Archer*, Virginia Commonwealth University
Andre A.A. Williams, Virginia Commonwealth University

Recently, penalized methods have been successfully applied to high-throughput genomic datasets in fitting linear, logistic, and Cox proportional hazards models. However, extensions for fitting penalized models for predicting ordinal responses have not yet been fully characterized, even though clinical and histological outcomes are frequently recorded using an ordinal scale. Herein we describe a penalized continuation ratio model capable of predicting an ordinal response when high-dimensional genomic data comprise the predictor space. The model is applied for the purpose of predicting early stage Parkinson's disease using non-invasively acquired blood samples that were profiled using gene expression microarrays.

email: kjarcher@vcu.edu

SCREENING SUBOPTIMAL TREATMENTS USING THE FUSED LASSO

Eric B. Laber*, University of Michigan
Mahdi Fard, McGill University
Joelle Pineau, McGill University
Susan A. Murphy, University of Michigan

In personalized medicine the goal is to recommend treatment based on patient characteristics thus hopefully leading to more favorable clinical outcomes. Different patterns of patient characteristics can lead to different treatment recommendations. Because there is likely to be insufficient evidence that one and only one treatment is best for each pattern of patient characteristics, we focus instead on screening out suboptimal treatments for each given pattern of patient characteristics. This approach recommends a class of treatments among which there is insufficient evidence to prefer one treatment over the others. The class of treatments can vary by the pattern of patient characteristics. In this setting given a particular patient, clinical decision makers would select from among the recommended class of treatments using considerations such as individual preference, clinical expertise, cost, and local availability. Our approach works by effectively "fusing together" likely optimal treatments using a novel lasso-type penalty. In simple settings the penalty can be directly related to controlling the false discovery rate of a series of comparisons of each treatment with the best. Experiments on real and simulated data show promising results.

email: laber@umich.edu

IDENTIFY INTERESTING INTERACTIONS IN DECISION MAKING

Peng Zhang*, University of Michigan
Susan A. Murphy, University of Michigan

In this article we discuss identifying variables that are important for adapting or personalizing treatment. Most variable selection techniques focus on variable selection for the prediction of the response in a supervised learning setting. However, as noted by many, very few of these variables are likely to be useful for deciding which treatment to provide to which patient. In constructing personalized treatments or dynamic treatment regimes, we are most interested in variables that have qualitative interaction with the treatment. Variables that qualitatively interact with treatment not only inform us about the magnitude of treatment effect, but also differentiate between patients who should be offered different treatments. We will discuss methods identifying such interesting interactions and how to make the statistical inference.

email: pczhang@umich.edu

EXAMPLES IN EPIDEMIOLOGY USING ADVANCED DATA MINING TECHNIQUES: CART, MARS AND TreeNet/MART

Shenghan Lai*, Johns Hopkins University
Mikhail Golovnya, Salford Systems

Advanced data mining tools can be exceptionally powerful techniques in analyzing massive epidemiologic data. In this presentation, several examples from epidemiologic research at Johns Hopkins University's Bloomsburg School of Public Health are used to illustrate the usefulness of MARS, CART and TreeNet/MART data mining techniques. The first example uses MARS to explore the association between regional heart function and coronary calcification. This example demonstrates that without MARS analysis, the conventional approach fails to identify the association. The second example uses CART to classify the study participants. This example shows that CART is more powerful than the conventional logistic regression analysis. The third example uses MART to explore the association between vitamin E and the development of myocardial infarction. Again, this example suggests that without MART, the relationship may never be able to be identified. These approaches were developed at Stanford University(CART, MARS, TreeNet) and Berkeley (CART) by world-renowned statisticians Leo Breiman and Jerome Friedman. TreeNet/MART attempts to leverage predictive power of traditional CART (Classification and Regression Trees) models by combining a large number of trees together using either bootstrap aggregation or boosting approaches.

email: mwoodward@salford-systems.com

CHANGE-LINE CLASSIFICATION AND REGRESSION FOR CHEMICAL TOXICITY ANALYSIS

Chaeryon Kang*, University of North Carolina-Chapel Hill
Fei Zou, University of North Carolina-Chapel Hill
Hao Zhu, University of North Carolina-Chapel Hill
Michael R. Kosorok, University of North Carolina-Chapel Hill

We introduce the 'change-line' classification and regression method to study latent subgroups. The proposed method finds a line which optimally divides a feature space into two heterogeneous subgroups, each of which yields a response having a different probability distribution or having a different regression model. The procedure is useful for classifying biochemicals on the basis of toxicity, where the feature space consists of chemical descriptors and the response is toxicity activity. In this setting, the goal is to identify subgroups of chemicals with different toxicity profiles. The split-line algorithm is utilized to reduce computational complexity. A two step estimation procedure, using either least squares or maximum likelihood for implementation, is described. Two sets of simulation studies and a data analysis applying our method to rat acute toxicity data are presented to demonstrate utility of the proposed method. A graphical examination is also performed to verify the existence of an underlying change in the distribution of toxicity activity.

email: ckang@bios.unc.edu

70. PRESIDENTIAL INVITED ADDRESS

BAYRS, BARS, AND BRAINS: STATISTICS AND MACHINE LEARNING IN THE ANALYSIS OF NEURAL SPIKE TRAIN DATA

Robert E. Kass, Department of Statistics, Center for the Neural Basis of Cognition, Machine Learning Department, Carnegie Mellon University

One of the most important techniques in learning about the functioning of the brain has involved examining neuronal activity in laboratory animals under varying experimental conditions. Neural information is represented and communicated through series of action potentials, or spike trains, and the central scientific issue in many studies concerns the physiological significance that should be attached to a particular neuron firing pattern in a particular part of the brain. In addition, a major comparatively new effort in neurophysiology involves the use of multielectrode recording, in which responses from dozens of neurons are recorded simultaneously. Among other things, this has made possible the construction of brain-controlled robotic devices, which could benefit people whose movement has been severely impaired.

In my talk I will briefly outline the progress made, by many people, over the past 10 years, highlighting some of the work my colleagues and I have contributed. The new methods of neural data analysis have resulted largely from adapting standard statistical approaches

(additive models, state-space models, Bayesian inference, the bootstrap) to the context of neural spike trains, which may be considered point processes. Methods commonly associated with machine learning have also been applied. I will make some comments about the perspectives of statistics and machine learning, and will indicate current status and future challenges.

kass@stat.cmu.edu

71. BAYESIAN METHODS IN GENOMIC RESEARCH

BAYESIAN APPROACHES FOR INCORPORATING INTERMEDIATE BIOMARKERS IN GENETIC ASSOCIATION STUDIES

David V. Conti*, University of Southern California
Wonho Lee, University of Southern California
Rachel Tyndale, University of Toronto
Andrew Bergen, SRI International
Gary Swan, SRI International

Since success with available pharmacotherapies for smoking cessation varies across individuals, the potential to guide treatment is of great interest. Studies have implicated genetic factors with different underlying processes. A pharmacokinetic (PK) mechanism is implied with the association of CYP2A6, a gene that is critical in the nicotine metabolism. The association of a SNP in CHRNB2 with quit rates and severity of withdrawal symptoms (WS) suggests a pharmacodynamic (PD) response. In addition to this genetic variation, biomarkers may be used for prediction, such as the nicotine metabolite ratio (NMR), a biomarker reflecting both CYP2A6 genetic and environmental influences. Using data from two clinical trials of smoking cessation with measured NMR, WS, and genotypes, we present analyses of gene-biomarker and biomarker-cessation associations. To better understand effects, we use a Bayesian approach treating biomarkers as flawed measures of underlying mechanisms. The model uses genetic and environmental factors to model latent factors for PK and PD processes with additional refinement using measured NMR and WS, respectively. We estimate genetic effects on cessation and on the latent processes, as well as pathway effects on cessation. We discuss extensions to multiple biomarkers and numerous genetic variants.

email: dconti@usc.edu

PICS: PROBABILISTIC INFERENCE FOR ChIP-SEQ

Raphael Gottardo*, Clinical Research Institute of Montreal and University of British Columbia

ChIP-seq, which combines chromatin immunoprecipitation with massively parallel short-read sequencing, can profile in vivo genome-wide transcription factor-DNA association with higher sensitivity, specificity and spatial resolution than ChIP-chip. While it presents new opportunities for research, ChIP-seq poses new challenges for statistical analysis that derive from the complexity of the biological systems characterized and the variability and biases in its digital sequence data. We propose a method called PICS (Probabilistic Inference for ChIP-seq) for extracting information from ChIP-seq aligned-read data in order to identify regions bound by transcription factors. PICS identifies enriched regions by modeling local concentrations of directional reads, and uses DNA fragment length prior information to discriminate closely adjacent binding events via a Bayesian hierarchical t-mixture model. PICS uses pre-calculated, whole-genome read mappability profiles and a truncated t-distribution to adjust binding event models for reads that are missing due to local genome repetitiveness. It estimates uncertainties in model parameters that can be used to define confidence regions on binding event locations and to filter estimates. Finally, PICS calculates a per-event enrichment score relative to a control sample, and can use a control sample to estimate a false discovery rate. PICS performs favorably compared to other popular analysis methods.

email: raphael.gottardo@ircm.qc.ca

BAYESIAN MODEL-BASED METHODS FOR ANALYZING ChIP SEQUENCING DATA

Ming Hu, University of Michigan
Jindan Yu, Northwestern University Feinberg Medical School
Jeremy MG Taylor, University of Michigan
Arul M. Chinnaiyan, University of Michigan
Zhaohui S. Qin*, University of Michigan

Protein-DNA interaction constitutes a basic mechanism for genetic regulation of target gene expression. Deciphering this mechanism is challenging due to the difficulty in characterizing protein-bound DNA on a genomic scale. The recent arrival of ultra-high throughput sequencing technologies has revolutionized this field by allowing quantitative sequencing analysis of target DNAs in a rapid and cost-effective way. ChIP-Seq, which couples chromatin immunoprecipitation (ChIP) with next-generation sequencing, provides millions of short-read sequences, representing tags of DNAs bound by specific transcription factors and other chromatin-associated proteins. The rapid accumulation of ChIP-Seq data has created a daunting analysis challenge. Here we propose a hidden Markov model (HMM)-based algorithms to detect genomic regions that are significantly enriched by ChIP-Seq. We also propose a multi-level hierarchical HMM that will allow integration of data from both ChIP-Seq and ChIP-chip experiments. Finally, we will discuss some issues related to post-processing ChIP-Seq data to obtain new biological insights.

email: qin@umich.edu

MODELING POPULATION HAPLOTYPE VARIATION

Paul Scheet*, University of Texas M. D. Anderson Cancer Center

Large-scale genotyping and sequencing technologies are facilitating the collection of an unprecedented amount of genetic data. Statistical methods may account for the poor resolution of genotypes that result from certain technologies or study designs. These methods rely on models for the dependence of alleles at nearby loci (linkage disequilibrium; LD). Our model for LD is statistical and allows a convenient framework for imposing parametric constraints, including in the form of distributions on parameters, to accomplish the incorporation of knowledge from other individuals within the population, or from other populations, as well as information about the specific phenomena of interest. We apply our method to real and simulated data for the purposes of estimating and correcting individual genotypes.

email: pascheet@mdanderson.org

72. INTERFERENCE AND SPILLOVER EFFECTS IN CAUSAL INFERENCE

ON INTERFERENCE IN INFERENCE FOR CAUSAL EFFECTS AND EXTENSIONS WITH APPLICATION TO INFECTIOUS DISEASES

Eric J. Tchetgen*, Harvard University
Tyler VanderWeele, Harvard University

The first part of the presentation will provide an overview of recent work on interference in causal inference. The second part of the presentation will focus on possible extensions of existing work to the infectious disease context and discuss the relevance of distance and of modeling to causal inference under interference.

email: etchetgen@gmail.com

STRATEGIES FOR MODELING INFERENCE BETWEEN UNITS IN MULTI-SITE TRIALS

Stephen Raudenbush*, University of Chicago

This paper considers the phenomenon of interference between units that arises in assessing the impact of an intervention. Interference arises when the potential outcomes of a unit are influenced by the treatment assignment of other units. In many social settings, the interveners deliberate exploit this possibility to strengthen the impact of the intervention. For example, a teacher's attempt to reduce aggressive behavior of students in her classroom may be enhanced if teachers in other classrooms of the same school are effectively intervening to reduce aggressive behavior of their own students. If so, a school-wide intervention to reduce aggression may be more effective than an intervention that focuses on a single classroom. I will draw on recent work that shows how to represent such a theory mathematically. When the intervention is replicated in multiple sites, new opportunities arise to identify the impact of such "benign" interference. I will compare two strategies for identification: the use multiple instrumental variables and the use of adjustment by observed pre-treatment covariates. The assumptions and data requirements for the two approaches are quite different, as illustrated in several recent examples.

email: sraudenb@uchicago.edu

MEASURING THE AVERAGE OUTCOME AND INEQUALITY EFFECTS OF SEGREGATION IN THE PRESENCE OF SOCIAL SPILLOVERS

Bryan S. Graham*, New York University
Guido W. Imbens, Harvard University
Geert Ridder, University of Southern California

In this paper we analyze the causal effects of reallocating individuals across social groups in the presence of social interactions or social spillovers. We consider the case where individuals are either "high" or "low" types. Own outcomes may depend on the fraction of high types in one's social group. We characterize the average outcome effect and inter-type inequality effects of "local" increases in segregation. We also characterize the average outcome-maximizing allocation of individuals to groups. We relate our estimands to the theory of sorting in the presence of social spillovers. For each estimand we provide conditions for identification. We also propose nonparametric estimators and characterize their large sample properties.

email: bryan.graham@nyu.edu

73. STATISTICAL METHODS IN NEUROSCIENCE

A NEW LOOK AT STATE-SPACE MODELS IN NEUROSCIENCE

Liam Paninski*, Columbia University

State space methods have proven indispensable in neural data analysis. However, common methods for performing inference in state-space models with non-Gaussian observations rely on certain approximations which are not always accurate. Here we review direct optimization methods that avoid these approximations, but that nonetheless retain the computational efficiency of the approximate methods. We discuss a variety of examples, applying these direct optimization techniques to problems in spike train smoothing, stimulus decoding, parameter estimation, and inference of synaptic properties. Along the way, we point out connections to some related standard statistical methods, including spline smoothing and isotonic regression. Finally, we note that the computational methods reviewed here do not in fact depend on the

state-space setting at all; instead, the key property we are exploiting involves the bandedness of certain matrices. We close by discussing some applications of this more general point of view, including Markov chain Monte Carlo methods for neural decoding and efficient estimation of spatially-varying firing rates.

email: liam@stat.columbia.edu

MULTI-SCALE MULTIPLE HYPOTHESIS TESTING FOR SPIKE TRAINS

Matthew T. Harrison*, Brown University
Asohan Amarasingham, Rutgers University

A recurring statistical problem in neuroscience and other fields is the identification of differences across experimental conditions. For neural spike trains, this often means identifying the location(s) in time (relative to some event) and the corresponding time scale(s) for which the firing rates are different across conditions. The multitude of locations, scales, and neurons creates a large multiple-testing problem. We observe that permutation tests using the well-known max-T or min-p methods are well suited for this situation. Unlike traditional permutations tests, however, the multiple testing corrections are not distribution free. We discuss robustness of the procedures to these assumptions.
email: matthew_harrison@brown.edu

TESTS FOR DIFFERENTIAL SPIKING ACTIVITY BASED ON POINT PROCESS MODELS

Uri Eden*, Boston University

A central problem in neuroscience is to determine whether two sets of spike trains represent information about the outside world in the same way. Such questions arise in assessing whether different neurons maintain similar representations of biological and behavioral signals, or in establishing whether a single neuron responds differently to changing stimuli or contexts. Previously, point process modeling has been used successfully to characterize the statistical properties of neural firing activity. We expand on the point process modeling framework, and develop a general testing paradigm to determine whether two collections of spike trains are likely to have been generated from the same process. The testing procedure involves fitting conditional intensity models to the observed spiking data and constructing test statistics from the resulting model fits. We identify some useful test statistics: the Integrated Squared Error (ISE), the maximum difference (MD), and the likelihood ratio (LR) statistic. The sampling distributions associated with each of these test statistics can be estimated using bootstrap methods, or in some cases, the asymptotic analytical distribution can be computed exactly. A simulation study and analysis of real data from rat hippocampus suggest that this testing procedure is able to detect differential firing robustly.

email: tzvi@bu.edu

MOTOR CORTICAL DECODING USING HIDDEN STATE MODELS

Vernon Lawhern, Florida State University
Wei Wu*, Florida State University
Nicholas Hatsopoulos, University of Chicago
Liam Paninski, Columbia University

Classical generalized linear models (GLMs) have been developed for modeling and decoding neuronal spiking activity in the motor cortex. These models are based on various forms of Poisson or non-Poisson processes, and provide reasonable characterizations between neural activity and motor behavior. However, they lack a description of other movement-related terms, such as joint angles at the shoulder and elbow, muscular activation, and other internal or external states. Here we propose to include a multi-dimensional hidden state to address these states in a GLM framework. The model can be identified by an Expectation-Maximization algorithm or a direct Laplace approximation method. We tested this new method in two datasets where spikes were simultaneously recorded using a multi-electrode array in the primary motor cortex of two monkeys. It was found that this method significantly improves the model-fitting over the classical GLM model for each hidden dimension varying from 1 to 4. This method also provides more accurate decoding of hand state (lower the Mean Square Error by up to 29%), while keeping real-time efficiency. These improvements on representation and decoding over the classical GLM model suggest that this new approach could contribute as a useful tool to neural coding and prosthetic applications.

email: wwu@stat.fsu.edu

74. RECENT ADVANCES IN VARIABLE SELECTION METHODOLOGY

PENALIZED REGRESSION METHODS FOR RANKING VARIABLES BY EFFECT SIZE, WITH APPLICATIONS TO GENETIC MAPPING STUDIES

Nam-Hee Choi, University of Michigan
Kerby Shedden, University of Michigan
Ji Zhu*, University of Michigan

Multiple regression can be used to rank predictor variables according to their 'unique' association with a response variable - that is, the association that is not explained by other measured predictors. Such a ranking is useful in applications such as genetic mapping studies, where one goal is to clarify the relative importance of several correlated genetic variants with weak effects. The use of classical multiple regression to rank the predictors according to their unique associations with the response is limited by difficulties due to collinearities among the predictors. Here we show that regularized regression can improve the accuracy of this ranking, with the greatest improvement occurring when the pairwise correlations among the predictor variables are strong and

heterogeneous. Considering a large number of examples, we found that ridge regression generally outperforms regularization using the L1 norm for variable ranking, regardless of whether the true effects are sparse. In contrast, for predictive performance, L1 regularization performs better for sparse models and ridge regression performs better for non-sparse models. Our findings suggest that the prediction and variable ranking problems both benefit from regularization, but that different regularization approaches tend to perform best in the two settings.

email: jizhu@umich.edu

VARIABLE SELECTION AND TUNING VIA CONFIDENCE REGIONS

Howard D. Bondell*, North Carolina State University
Funda Gunes, North Carolina State University

A new approach to variable shrinkage and selection is proposed. The idea is based on choosing a sparse solution within a given joint confidence region. It is shown that the approach yields an interpretable and automatic tuning procedure for some existing selection methods. The proposed method can obtain selection consistency in both linear and generalized linear models. In simulation studies, the confidence region based approach has been shown to compare favorably with other methods of tuning.

email: bondell@stat.ncsu.edu

COMPUTING THE SOLUTION PATH OF PENALIZED COX REGRESSION

Hui Zou*, University of Minnesota

Penalized Cox's proportional hazard model is often used in survival analysis when the number of covariates is large. Penalty functions that encourage sparse solutions are particularly preferred in the high-dimension scenario. Several papers have been devoted to the Lasso-type penalized Cox regression. In this talk we introduce a new efficient algorithm that can compute the solution path of penalized Cox regression. Our method combines two optimization techniques: majorization-minimization and coordinate descent. Some numerical examples are used to demonstrate its utility.

email: hzou@stat.umn.edu

COORDINATE DESCENT ALGORITHMS FOR VARIABLE SELECTION

Trevor J. Hastie*, Stanford University
Rahul Mazumder, Stanford University
Jerome Friedman, Stanford University

The Lasso is a popular regularizer, since it combines both variance reduction with variable selection. Very efficient algorithms exist for computing the lasso regularization path. Coordinate descent appears to be the most efficient, and works well for a variety of loss functions. For very sparse models, the lasso is not aggressive enough in its selection, and attention has focused on non-convex penalties. We show that again coordinate descent provides a very attractive framework for computing entire regularization paths for families of non-convex penalties.

email: hastie@stanford.edu

75. BIOMARKERS AND DIAGNOSTIC TESTS

USING TUMOR RESPONSE IN DESIGNING EFFICIENT CANCER CLINICAL TRIALS WITH OVERALL SURVIVAL AS PRIMARY ENDPOINT

Donald A. Berry, University of Texas M. D. Anderson Cancer Center
Haiying Pang*, University of Texas M. D. Anderson Cancer Center

Tumor response is a standard primary endpoint to demonstrate anti-tumor activity in phase II oncology clinical trials. In conventional phase III oncology trials, the standard primary endpoint is either progression-free survival or overall survival. Possible roles for tumor response in phase III studies and in addressing the question of survival benefit have not been adequately explored. Ignoring tumor response information in phase III can waste information, prolong trial duration, and require a larger sample size. We propose a method to model the relationship between survival and tumor response. The goal is to make more informed conclusions about the effect of a treatment on overall survival. We conduct extensive simulation studies to assess the operating characteristics of the proposed method. The results show that the proposed method is able to detect treatment efficacy better, and can shorten trial duration.

email: haiying.pang@uth.tmc.edu

A STRATEGY TO IDENTIFY PATIENTS SENSITIVE TO DRUGINDUCED ADVERSE EVENTS

Wei-Jiun Lin*, National Center for Toxicological Research, U.S. Food and Drug Administration
James J. Chen, National Center for Toxicological Research, U.S. Food and Drug Administration

The occurrence of a drug-induced adverse event that is recognized only after marketing a drug generally has an incidence too low to be detected in common pre-clinical and clinical trials. It may result in withdrawal of the drug from the market, even though the drug benefits the vast majority of those taking it, without increased risk of adverse effect. Sensitive subpopulations of patients are not presently identifiable in advance using conventional diagnostic

criteria. If sensitive subpopulations could be identified in advance, drugs could be approved selectively for the majority of patients who do not belong to sensitive subpopulations. This study presents a strategy to identify sensitive subpopulations for high-dimensional data such as microarray gene expression data. Classification procedures are developed to distinguish sensitive and non-sensitive populations. In silico approach is used to estimate the needed sample size for identifying a small portion of sensitive samples and to evaluate the performance of the classification algorithms. A simulation dataset based on the data from a pre-clinical liver toxicity biomarker experiment will be used for illustration.

email: WeiJiun.Lin@fda.hhs.gov

EARLY DETECTION OF ALZHEIMER'S DISEASE USING PARTIALLY ORDERED CLASSIFICATION MODELS

Curtis Tatsuoka*, Case Western Reserve University
Huiyun Tseng, Columbia University
Judith Jaeger, AstraZeneca Pharmaceuticals
Alan Lerner, Case Western Reserve University

Despite its widespread clinical importance in neurological applications, there do not appear to be established statistical methods that are specifically tailored to untangle the complex links between neuropsychological (NP) assessment performance and discrete cognitive functions. Partially ordered set (poset) classification models directly attempt to address the complexities that arise in NP assessment data analysis. An overview of these methods will be given. In addition, an application in the early detection of Alzheimer's disease through cognitive markers will be described. Implications for reducing sample size requirements in clinical trials for Alzheimer's disease also will be discussed.

email: curtis.tatsuoka@case.edu

CHALLENGES AND TECHNICAL ISSUES OF ASSESSING TRIALLEVEL SURROGACY OF PUTATIVE SURROGATE ENDPOINTS IN THE META-ANALYTIC FRAMEWORK FOR CLINICAL TRIALS

Qian Shi*, Mayo Clinic
Lindsay A. Renfro, Baylor University
Brian M. Bot, Mayo Clinic
Daniel J. Sargent, Mayo Clinic

Various meta-analytic approaches have been developed and applied to evaluate putative surrogate endpoints (S) of primary endpoints (T) in oncology clinical trials. The estimation performance (EP) of conventional and systematic trial-level surrogacy (TLS) measures were assessed and compared through a large scale simulation study. Previously, we demonstrated that the number of trails included in the meta-analysis, the degree of variation in treatment effect across trials, and effective sample sizes greatly impact on the EPs. In the current presentation, we expand our study to include a broader content of challenges and technical issues in the assessment of the TLS. These challenges include 1) the variability of surrogacy estimation by splitting trials into subunits when the number of trials is limited; 2) robustness of the underling hierarchical structure and distributional assumptions; 3) complications due to the ignorance of the natural constraints between S and T; 4) challenges of jointly modeling S and T when T is one component of S.

email: shi.qian2@mayo.edu

PROBIT LATENT CLASS MODELS FOR EVALUATING ACCURACY OF DIAGNOSTIC TESTS WITH INDETERMINATE RESULTS

Huiping Xu*, Mississippi State University
Bruce A. Craig, Purdue University

Indeterminate or inconclusive test results often occur with diagnostic tests. When assessing diagnostic accuracy, it is important to properly report and account for these results. In the literature, however, these results have most commonly been discarded prior to analysis or treated as either a positive or negative result. While these adjustments allow accuracy to be computed in the standard way, these forced decisions limit the interpretability and usefulness of the results. Estimation of diagnostic accuracy is further complicated when a gold standard is not available. In this situation, multiple diagnostic tests are usually used to better understand the test characteristics. Traditional latent class modeling can be readily applied to analyze these data and account for the indeterminate results. These models, however, assume that tests are independent conditional on the true disease status, which is rarely valid in practice. We propose a polytomous probit latent class model, which allows arbitrarily complicated correlation structures among multiple tests, while taking into consideration the indeterminate results. To obtain the maximum likelihood estimates, we implement a Monte Carlo EM algorithm and accelerate the convergence rate using the idea of dynamic EM algorithm. We demonstrate our model using a simulation study and two published medical studies.

email: hxu@math.msstate.edu

BAYESIAN ANALYSIS AND CLASSIFICATION OF TWO QUANTITATIVE DIAGNOSTIC TESTS FOR OCCASIONALLY ABSENT MARKERS AND NO GOLD STANDARD

Jingyang Zhang*, University of Iowa
Kathryn Chaloner, University of Iowa
Jack T. Stapleton, University of Iowa

Two diagnostic tests are used to detect antibodies to the GB Virus type C (GBV-C) envelope glycoprotein E2. There is no reference test (gold standard). Each test looks for a characteristic that is

typically present if the antibodies are present, but is occasionally absent. A model is developed reflecting the absence of either characteristic independently. A mixture of four bivariate normal distributions is used along with a prior distribution and a Bayesian analysis. Test results from 100 subjects are analyzed using the Metropolis-within-Gibbs sampler. Subjects are classified by a statistical decision rule, and the classification appears to separate the subjects well into those that are positive and negative for GBV-C antibodies. Simulation studies are conducted to assess the accuracy of the classification. Issues related to the choice of prior distributions are also discussed.

email: jingyang-zhang@uiowa.edu

76. SURVIVAL ANALYSIS IN CLINICAL TRIALS

A TEST FOR EQUIVALENCE OF TWO SURVIVAL FUNCTIONS IN PROPORTIONAL ODDS MODEL

Wenting Wang*, Florida State University
Debajyoti Sinha, Florida State University
Stuart R. Lipsitz, Harvard Medical School
Richard J. Chappell, University of Wisconsin-Madison

To show a new treatment's equivalent therapeutic effect to an existing standard treatment for survival outcomes, the usual statistical equivalent tests are based on Cox's (1972) proportional hazards assumption. We present the alternative method based on the proportional odds survival models (POSM) for the new and the standard treatment arms, and show the advantages of using this equivalence test instead of tests based on Cox's model. We first show that effective tests procedures for equivalence of treatment arms under POSM can be implemented when we are either interested in maximum difference in survival functions or in difference in hazard functions from two treatment arms. We propose different test procedures to deal with survival function for the standard treatment being unknown and known. The simulation study of the relationship between the sample size and the power show that, the equivalence tests under Cox model has an inflated type I error rate when the true model is POSM; However, our procedures have correct Type I error rates under both proportional odds as well as Cox's model. Further, when the true model is Cox model, our procedure has a high power comparable to the equivalence tests based on Cox model. In practice where only a fraction of new treatments are clinically equivalent to the standards, our another simulation study shows that repeated use of our test (compared to log-rank based tests) will be a safer statistical practice, because fewer numbers and percentages of statistically accepted (via our tests) equivalent treatments are going to be actually clinically non-equivalent.

email: wwang@stat.fsu.edu

A NONPARAMETRIC TEST FOR EQUALITY OF SURVIVAL MEDIANS

Mohammad H. Rahbar*, University of Texas Health Science Center at Houston and University of Texas School of Public Health
Zhongxue Chen, Florida International University
Sangchoon Jeon, Yale University
Joseph C. Gardiner, Michigan State University

In clinical studies testing for equality of survival medians across treatment arms is often useful. Procedures which are designed for testing the homogeneity of survival curves, such as the log-rank, Wilcoxon, and the Cox model, have been applied. However, in practice they lead to inflation of type I error, particularly when the underlying assumptions are not met. We propose a new nonparametric method for testing the equality of several survival medians based on randomly right censored data. We develop a new test statistic based on the Kaplan-Meier method and derive asymptotic properties of this test statistic. Through simulations we compare the power of our new procedure with that of the log-rank, Wilcoxon, the Cox model, and the Brookmeyer-Crowley method. Our results suggest that performance of some of these estimation procedures depend on the level of censoring and appropriateness of the underlying assumptions for each procedure. When the objective is testing for homogeneity of survival curves, the log-rank test and the Cox model are more powerful than our proposed test. However, when the objective is testing for equality of survival medians, the Brookmeyer-Crowley and our new test statistic seem to have some advantages.

email: Mohammad.H.Rahbar@uth.tmc.edu

EVALUATION OF TREATMENT-EFFECT HETEROGENEITY IN THE AGE OF BIOMARKERS

Ann A. Lazar*, Harvard University and Dana-Farber Cancer Institute
Bernard F. Cole, University of Vermont and Dana-Farber Cancer Institute
Marco Bonetti, Bocconi University
Richard D. Gelber, Harvard School of Public Health, Harvard Medical School and Dana-Farber Cancer Institute

Randomized clinical trials, particularly in oncology, collect information on relevant covariates to identify factors that predict treatment response and prognostic factors for risk of disease progression or relapse. A traditional analytical approach evaluates the treatment-covariate interaction to explore potential heterogeneity of therapeutic effect for different patient subgroups. However, in the age of continuously measured biomarkers, traditional modeling approaches sometimes fail to reveal the patterns of treatment effects associated with biomarker values. The purpose of this paper is to provide an overview of a statistical approach, Subpopulation Treatment Effect Pattern Plots (STEPP), for evaluating treatment-effect heterogeneity by estimating treatment effects within overlapping subgroups defined along

the biomarker-covariate continuum. The treatment effects can be measured using a variety of clinically relevant endpoints including Kaplan-Meier survival estimates at a particular time from randomization. We propose extending the STEPP methodology to measures of treatment effect from hazard ratio values based on observed minus expected estimation and measures obtained in the competing risk setting, which is particularly relevant when considering biomarkers because these predictors are more likely to affect disease-specific events rather than other competing events. We illustrate how the STEPP methodology can explore patterns of treatment effect for varying levels of biomarkers by using the Breast International Group (BIG) 1-98 randomized clinical trial evaluating adjuvant therapy with letrozole versus tamoxifen for postmenopausal women with hormone-receptor-positive breast cancer.

email: alazar@hsph.harvard.edu

STRATIFIED AND UNSTRATIFIED LOG-RANK TESTS IN MULTICENTER CLINICAL TRIALS

Changyong Feng*, University of Rochester

The log-rank test is widely used in multicenter clinical trials with time to event as the pirmary outcome variable. Due to the heteogeneity among differnt centers, the stratified log-rank test is usually more powerful. In this talk we discuss the power loss of stratfied and unstratified log-rank tests and develop a linear combinations fo these two test which has more power than either of thm in genral cases. Our result is used to a multicenter clinical trial of heart disease study.

email: feng@bst.rochester.edu

HYPOTHESIS TESTING IN RANDOMIZED TRIALS FOR SURVIVAL DATA WITH MISSPECIFIED REGRESSION MODELS

Jane Paik*, Stanford University

For a large class of commonly used regression models, standard hypothesis tests in randomized clinical trials that are based on incorrectly specified models are guaranteed to have asymptotically correct Type I error under the null hypothesis, whether or not the actual data generating distribution is different from the model (Rosenblum and van der Laan, 2009). In the setting of analyzing survival data in clinical randomized trials using regression models, we consider the null hypothesis in which the treatment has no effect on the survival distribution for a subpopulation defined by baseline variable. We show that the results of Rosenblum and van der Laan (2009) can be directly applied to show that standard hypothesis tests based on a misspecified Cox model have asymptotically correct Type I error when the censoring distribution is independent of the treatment assignment. In addition, the asymptotic robustness property holds for misspecified parametric regression models such as the exponential and Weibull model, under the case of independence between censoring and treatment assignment. The same arguments can be modified and extended to show asymptotic robustness when censoring is dependent on treatment, when a multiplicative form for the censoring distribution is assumed.

email: janepaik@stanford.edu

SEMIPARAMETRIC ESTIMATION OF TREATMENT EFFECT WITH TIME-LAGGED RESPONSE IN THE PRESENCE OF INFORMATIVE CENSORING

Xiaomin Lu*, University of Florida
Anastasios Tsiatis, North Carolina State University

In many randomized clinical trials, the primary response variable, for example, the survival time, is not observed directly after the patients enroll in the study but rather observed after some period of time (lag time). It is often the case that such a response variable is missing for some patients due to censoring that occurs when the study ends before the patient's response is observed or when the patients drop out of the study. It is often assumed that censoring occurs at random which is referred to as on informative censoring; however, in many cases such an assumption may not be reasonable. If the missing data are not analyzed properly, the estimator or test for the treatment effect may be biased. In this paper, we use semiparametric theory to derive a class of consistent and asymptotically normal estimators for the treatment effect parameter which are applicable when the response variable is right censored. The baseline auxiliary covariates and post-treatment auxiliary covariates, which may be time-dependent, are also considered in our semiparametric model. These auxiliary covariates are used to derive estimators that both account for informative censoring and are more efficient then the estimators which do not consider the auxiliary covariates.

email: xlu2@ufl.edu

A GENERALIZED COX PROPORTIONAL HAZARD MODEL FOR COMPARING DYNAMIC TREATMENT REGIMES

Xinyu Tang*, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Dynamic treatment regimes are algorithms for assigning treatments to patients with complex diseases, where treatment consists of more than one episode of therapy, potentially with different dosages of the same agent or different agents. Sequentially randomized clinical trials are usually designed to evaluate and compare the effect of different treatment regimes. In such designs, eligible patients are first randomly assigned to receive one of the initial treatments. Patients meeting some criteria (e.g. no progressive disease) are then randomized to receive one of the maintenance treatments. Usually,

the procedure continues until all treatment options are exhausted. Such multistage treatment assignment results in dynamic treatment regimes consisting of initial treatments, intermediate responses and second stage treatments. However, methods for efficient analysis of sequentially randomized trials have only been developed very recently. As a result, earlier clinical trials reported results based only on the comparison of stage-specific treatments. In this article, we propose a generalized Cox proportional hazard model that applies to comparisons of any combination of any number of treatment regimes regardless of the number of stages of treatment. Contrasts of dynamic treatment regimes are tested using the Wald chi-square method. Both the model and Wald chi-square tests of contrasts are illustrated through a simulation study and an application to a high risk neuroblastoma study to complement the earlier results reported on this study.

email: xit11@pitt.edu

77. HIGH DIMENSIONAL MODELING: SEGMENT DETECTION AND CLUSTERING

OPTIMAL SPARSE SEGMENT IDENTIFICATION

Jessie Jeng*, University of Pennsylvania
Tony Cai, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

We consider the problem of detecting and identifying sparse segments in a long sequence of data with additive Gaussian white noise, where the number,lengths and the positions of the segments are unknown. The problem can be formulated as a multiple hypothesis testing problem. We present the conditions for the existence of a consistent testing procedure where the detection of segments in Gaussian noise data is possible. We also present a restricted likelihood selection procedure for identifying the segments and show the optimality of this method. We demonstrate the proposed methods with simulations and analysis of a real data set related to identification of copy number variants.

email: xjeng@upenn.edu

USING SCAN STATISTICS ON DEPENDENT SIGNALS AND ASSESSING ITS DISTRIBUTION, WITH APPLICATION TO SEARCHING SEQUENCES OF INTEREST ALONG THE GENOME

Anat Reiner-Benaim*, Haifa University

The attempt to locate sequences of interest along the genome is a familiar problem that is frequently confronted by genome researchers. The challenge here is to identify short intervals of nucleotides on the genome, within noisy and much longer sequences, such as genes. One example in which the problem occurs is the search for transcription factor binding sites within a group of functionally related genes. Another challenging example, which will be discussed here, is the search for intronic regions. Inference on the presence of intronic regions can be made based on continuous monitoring of expression level across the genomic sequence, using a tiling array experiment, which can facilitate detection of sudden changes or occurrences of expression. Here, we suggest using a scan statistics to test whether an interval, within a specified gene, is showing the biological effect expected to occur in an intronic region. We offer a statistic that integrates several important considerations: the multiple testing of many genes; the bias caused by the difference between gene lengths; the dependence between adjacent measures of expression along the genomic sequence. We also offer an analytical assessment of the scan statistics distribution considering this dependence under a normal stochastic process.

email: areiner@stat.haifa.ac.il

A FRAMEWORK FOR DENSITY ESTIMATION FOR BINARY SEQUENCES

Xianyun Mao*, Penn State University
Bruce Lindsay, Penn State University

We present some new methods based on kernel density estimation and a modal expectation-maximization (MEM) method for clustering DNA haplotype sequences. For the simple mission of clustering binary sequences, we define a kernel density estimator for the sequences and use it to define a weight function for each sequence. Then we start from each data sequence and examine all other sequences along with the weight function to find the nearest mode of the density. We then cluster the sequences that share the same mode. For the haplotype problem, we construct a likelihood function for the genotype-haplotype problem that depends on the unknown haplotype type density and then use likelihood EM to create an updated density that partially maximizes the likelihood. The performance of the method regarding haplotype inference is tested on large datasets with the comparison to the available methods such as Phase. It shows that the new method yield comparable results while requiring less computational time. In a similar fashion, we can define a density estimator for the binary sequences (haplotypes) in the presence of recombination and mutation. One direct outcome is the resulting tree structure converges to a single ancestor faster than the ones that are based on a model of mutation alone.

email: xxm106@psu.edu

BAYESIAN SPECIES CLUSTERING VIA DP AND SAMPLING DESIGNS VIA MONTE CARLO

Hongmei Zhang*, University of South Carolina
Kaushik Ghosh, University of Nevada-Las Vegas
Pulak Ghosh, Novartis Pharmaceuticals

In a sample of DNA segments, sequences without duplicates or with small number of copies are likely to carry information related

to mutations or diseases and can be of special interest. Each distinct DNA sequence is referred to as a species in the article. Assuming the species abundance unknown, we propose a two-step Bayesian sampling design to collect species of interest. The first phase of the design is used to infer the species abundance through clustering analysis based on a multivariate hypergeometric model with Dirichlet Process prior for species frequencies; and the second phase is to infer the sample size using Monte Carlo simulations to collect a certain number of species of interest. The proposed method is demonstrated and evaluated via various simulations from different directions. A real DNA segment data set is used to illustrate and motivate the proposed sampling method.

email: hzhang@sc.edu

FEATURE-SUBSET-SPECIFIC CLUSTERING USING STOCHASTIC SEARCH

Younghwan Namkoong, University of Florida
Yongsung Joo*, Dongguk University, Korea
Douglas D. Dankel, University of Florida

The majority of the prior research related to the feature selection were interested in dividing features into a contributing and a non-contributing subset for clustering. However, data can often comprise several feature subsets where each feature subset constructs clusters differently. In this paper, we present a novel model-based clustering approach using stochastic search to discover the multiple feature subsets that have different model-based clusters. Using the United Nations (UN) World Statistics dataset, we demonstrate that the proposed method can be successfully applied to social and economical research.

email: yongsungjoo@dongguk.edu

78. LONGITUDINAL DATA ANALYSIS

EFFICIENT SEMIPARAMETRIC REGRESSION FOR LONGITUDINAL DATA WITH NONPARAMETRIC COVARIANCE ESTIMATION

Yehua Li*, University of Georgia

In order to have efficient estimators in semiparametric regression for longitudinal data, it is important to take into account of the within-subject covariance. In existing literature, the covariance is usually modeled either parametrically or semiparametrically. We show in this paper that, when the covariance or correlation model is mis-specified, the semiparametric regression estimator could lose efficiency. Instead, we propose a method that combines efficient semiparametric estimator and nonparametric covariance estimation. We show that the kernel covariance estimator provides uniformly consistent estimators for the within-subject covariance matrices, and the semiparametric profile estimator with plugged in nonparametric covariance estimator is still semiparametrically efficient. The proposed method is robust against mis-specification of covariance models. Finite sample performance of the proposed estimator is illustrated by simulation studies. In an application to the CD4 count data from an AIDS clinical trial, the proposed method is further extended to a functional analysis of covariance (fANCOVA) model.

email: yehuali@gmail.com

A DISTRIBUTION-FREE ASSOCIATION MEASURE FOR LONGITUDINAL DATA WITH APPLICATIONS TO HIV/AIDS RESEARCH

Sujay Datta*, Fred Hutchinson Cancer Research Center and the University of Washington
Li Qin, Fred Hutchinson Cancer Research Center and the University of Washington
Stephen G. Self, Fred Hutchinson Cancer Research Center and the University of Washington

Measuring association between two variables is a fundamental task in statistical inference and there is extensive literature on how to do it in the univariate case, including Pearson's correlation, Spearman's rank correlation and Kendall's tau coefficients. In the case of multivariate longitudinal or timecourse data, however, relatively little is available for this purpose. The few available methods (e.g. dynamical correlation by Dubin and Muller (2005) or odds ratio for correlated binary data by Lipsitz et al. (1991)) are either computationally challenging or dependent on specific assumptions about the underlying distribution and data-type. Here we introduce an association measure for longitudinal/timecourse data (continuous, discrete numerical or ordinal) which is conceptually simple and distribution-free. It is quite flexible regarding the sequences of time-points at which, the two variables are observed. After discussing its asymptotic property, we demonstrate its small-sample performance via simulation. We then apply it to measure the association between temporal expression profiles of genes and temporal measurements of viral load (viral RNA counts per unit of blood) in a group of acutely HIV-infected individuals. Finally we apply it to measure the association between temporal levels of pairs of cytokines in another group of HIV-infected individuals.

email: sdatta@fhcrc.org

FLEXIBLE BENT-CABLE MODELS FOR MIXTURE LONGITUDINAL DATA

Shahedul A. Khan*, University of Waterloo
Grace Chiu, University of Waterloo
Joel A. Dubin, University of Waterloo

Data showing a trend that characterizes a change due to a shock to the system are a type of changepoint data, and may be referred to as shock-through data. As a result of the shock, this type of data may exhibit one of two types of transitions: gradual or abrupt.

Although shock-through data are of particular interest in many areas of study such as biological, medical, health and environmental applications, modeling the trend is challenging in the presence of discontinuous derivatives. Further complications arise when we have (1) longitudinal data, and/or (2) samples which come from two potential populations: one with a gradual transition, and the other abrupt. Bent-cable regression is an appealing statistical tool to model shock-through data. We develop extended bent-cable methodology for longitudinal data in a Bayesian framework to account for both types of transitions; inference for the transition type is driven by the data rather than a presumption about the nature of the transition. We demonstrate our methodology by a simulation study, and with two applications: (1) assessing the transition to early hypothermia in a rat model, and (2) understanding CFC-11 trends monitored globally.

email: sa4khan@math.uwaterloo.ca

A CIRCULAR LEAR CORRELATION STRUCTURE FOR CYCLICAL LONGITUDINAL DATA

Sean L. Simpson*, Wake Forest University School of Medicine
Lloyd J. Edwards, University of North Carolina-Chapel Hill

Circular covariance patterns arise naturally from many important biological and physical processes when there is either an outcome measure with a temporal cycle or spatial measurements taken in a circular fashion. Modeling these patterns can be immensely important for proper estimation, inference, and model selection. We propose a circular linear exponent autoregressive (LEAR) correlation structure for cyclical longitudinal data which extends the standard LEAR model to the circular context and allows for the modeling of data that have within-subject correlation decreasing exponentially as a function of cyclical distance (distance between two measurements in a cycle). Special cases of this parsimonious correlation model include the equal correlation and first-order moving average (MA(1)) correlation structures and a circular analog of the continuous-time AR(1) model. We discuss properties and estimation of the circular LEAR model in the context of cyclical longitudinal data concerning diet and hypertension (the DASH study). Analysis of these data exemplifies the benefits of the circular LEAR correlation structure.

email: slsimpso@wfubmc.edu

DISCRETE TIME-TRANSFORMATION MODEL WITH RANDOM EFFECTS AND STERILE FRACTION: AN APPLICATION TO TIME TO PREGNANCY

Alexander McLain*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Many interesting statistical issues arise when analyzing time to pregnancy (TTP) or fecundability data. Scheike and Jensen (1997) proposed using the discrete time equivalent to the proportional hazards model to analyzing TTP data. While the discrete survival approach has had some success in the TTP literature, some have critiqued its lack of biological validity. Current discrete survival models do not allow for the incorporation of day-level lifestyle variable effects that occur within the "fertile window." In this paper, we propose a flexible class of discrete-time transformation models that addresses biological validity by allowing day-level covariates and coefficients to be included. Common issues associated with TTP data, such as allowing for a sterile fraction and unobserved covariates, are included. We illustrate our method on simulated and real data.

email: mclaina@mail.nih.gov

A CORRELATED BAYESIAN HUMAN FECUNDABILITY MODEL WITH MISSING COVARIATES

Sungduk Kim*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Germaine B. Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Human fecundability is defined as the probability of pregnancy in a menstrual cycle given unprotected sexual intercourse, and is used to identify toxicants that adversely impact human reproduction. Models for estimating fecundability have slowly evolved, beginning with those developed by Barrett and Marshall (1969). However, such models have assumed that the baseline day-specific probabilities of pregnancy are independent. The extent to which such assumptions reflect biologically relevant models inclusive of behaviors such as intercourse remain to be established and served as the motivation for this work. In this paper, we consider the gamma process prior to account for the dependence between baseline day-specific probabilities of pregnancy. Also, we offer a model applicable for missing covariates when estimating human fecundability. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. We use data from the New York State Angler Prospective Pregnancy Study with preconception enrollment of women.

email: kims2@mail.nih.gov

BAYESIAN CHANGEPOINT MODELS IN DETECTING WOMEN'S MENOPAUSAL TRANSITION

Xiaobi Huang*, University of Michigan
Siob'an D. Harlow, University of Michigan
Michael R. Elliott, University of Michigan

As women approach menopause, the patterns of their menstruation segment lengths change. In order to study when changes in menstrual length happen, we usebuild Bayesian linear change point models to jointly model both the mean as well as the variability of the segment length. The model incorporates separate mean and variance change points for each woman and a hierarchical model to link them together, along with regression components to include predictors of menopausal onset such as age at menarche and parity. Data are from TREMIN, an ongoing 70-year old longitudinal study that has obtained menstrual calendar data of women throughout their life course. Our study cohort includes nearly 1000 women, many of whom have missingness due to hormone use, surgery, random missingness and loss of contact. We integrate multiple imputation in our Bayesian estimation procedure to deal with different forms of the missingness. Posterior predictive model checks are applied to evaluate the model fit.

email: xiaobih@umich.edu

79. STATISTICAL GENETICS: EPISTASIS, GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS

EPISTASIS IN GENOME-WIDE ASSOCIATION STUDIES

Huei-Wen Teng*, Penn State University
Yu Zhang, Penn State University

The importance of understanding epistasic interaction in molecular and quantitative genetics has led to many competing methods to identify epistasis. Existing approaches are however restrictive. For example, when a group of single-nucleotide polymorphisms (SNPs) are marginally insignificant but jointly associated with the disease, traditional step-wise method may have difficulty in detecting them. In addition, searching the space of possible epistasis structure is computationally infeasible, particularly when the size of the dataset set is huge. To solve these problems, we propose a two-stage procedure in a Bayesian framework to capture disease associated SNPs and identify epistasis. In the first stage, we propose a Bayesian partition model with a modified Markov random field (MRF) with two features: (i) We define a supernode as a group of SNPs, in which SNPs are jointly associated with the disease. (ii) A MRF structure is placed to incorporate pairwise interaction between supernodes. In the second stage, we focus on a few number of SNPs identified in the first stage, and propose a fast algorithm to identify the interaction structure using the probabilistic framework of Bayesian networks. Our approach is implemented to synthetic datasets consisting of a verity of high-order interaction structures, and to a real dataset consisting many thousands of SNPs.

email: hut118@psu.edu

ENTROPY-BASED TESTS FOR GENETIC EPISTASIS IN GENOME WIDE ASSOCIATION STUDIES

Xin Wang*, Mayo Clinic College of Medicine
Mariza de Andrade, Mayo Clinic College of Medicine

In the past few years, several entropy-based tests were proposed for testing either gene-gene interaction or gene association. These tests are mainly based on Shannon entropy and compared to standard chi-square tests, they have higher statistical power, especially when the number of marker loci is large. In our study, we extend some of these tests using a more generalized entropy definition, Rényi entropy, where Shannon entropy is as a special case of order 1. The order alpha (>0) of Rényi entropy, weights the events (genotype/haplotype) according to their probabilities (frequencies). Higher alpha emphasis more on high probability events while smaller alpha approaching 0 tends to assign weights more equally. Thus, by properly choosing the order, one can potentially increase the power of the tests. It is also informative to see how p-value changes along with order alpha. We conducted simulation studies as well as real data studies to assess the impacts of order alpha and the performances of these generalized tests.

email: mandrade@mayo.edu

A DIMENSION REDUCTION APPROACH TO DETECT MULTILOCUS INTERACTION IN A CASE CONTROL STUDY

Saonli Basu*, University of Minnesota

Studying one single nucleotide polymorphism (SNP) at a time may not be sufficient to understand complex diseases. A SNP or a gene alone may have little or no effect on risk of disease, but together may increase the risk substantially. The joint behavior of genetic variants is often referred to as epistasis or multilocus interaction. We have proposed a dimension reduction approach to model such multilocus interaction. The model offers a data reduction strategy that substantially reduces the estimation of a large number of parameters corresponding to a large number of SNPs. Our method is based on a likelihood approach, and estimation and inference can be conducted in a systematic manner within the likelihood framework. We also propose a formal statistical test for the significance of the effect of a group of SNPs on the disease. This proposed approach also provides a way to capture the uncertainties regarding the choice of the model, which most of the current approaches thrives to capture. We illustrate and compare our model with existing approaches through extensive simulations and demonstrate the superiority of our model in detecting multilocus interaction.

email: saonli@umn.edu

ESTIMATING GENE-ENVIRONMENT INTERACTION BY POOLING BIOMARKERS

Michelle R. Danaher*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Anindya Roy, University of Maryland, Baltimore County
Paul Albert, Eunice Kennedy Shriver National Institute of Child Health and Human
Development
Enrique Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human
Development

The cost of assays for genotyping often limits researchers' ability to examine interesting hypotheses about a gene-environment interaction for a disease due to the large number of assays which are necessary to obtain cases and controls in all combinations of genotypes and exposures for a reasonable statistical power of the gene-environment interaction estimate. To address this problem, we propose a new study design where we strategically pool biospecimens to obtain an estimate of the gene-environment interaction for a rare disease. By pooling, we increase the information obtained, while holding the number of assays fixed. We explore five methods that have been proposed to estimate the gene-environment interaction including: case-control, case only, bayes model averaging, empirical bayes-type shrinkage estimator, and a two stage case-control case-only. All five methods benefit from an increased efficiency due to pooling for a fixed number of assays. We focus on a special, though realistic case, where the data for disease, environment (e.g. smoking yes/no), and genotype status are binary. With a fixed number of assays available, and accounting for varying levels of measurement error due to pooling, we explore the power and robustness the five methods using simulations. We also compare the methods using an interesting epidemiologic study.

email: dan87her@yahoo.com

GENE-ENVIRONMENT INTERACTION TESTING IN FAMILY-BASED ASSOCIATION STUDIES WITH PHENOTYPICALLY ASCERTAINED SAMPLES: A CAUSAL INFERENCE APPROACH

David Fardo*, University of Kentucky
Yan Huang, University of Kentucky
Stijn Vansteelandt, Ghent University

The class of family-based association tests (Laird et al, 2000) provides strategies for testing main genetic effects that are robust to undetected/unaccounted for population substructure. Conditioning on parental/founder genotypes (or the corresponding sufficient statistics when parental genotypes are missing; Rabinowitz and Laird, 2000) insulates these testing strategies from the bias due to ancestry-driven confounding. However, once a main genetic effect must be estimated, as in the case of testing for GxE and GxG interactions, ascertainment conditions for sample recruitment must appropriately be taken into account. The calculus of directed acyclic graphs, specifically rules of d-separation (Pearl, 2000; Robins, 2001), helps identify estimating equations that can properly incorporate ascertainment criteria. We employ the concept of principal stratification (Frangakis and Rubin, 2002) and G-estimation techniques (Robins et al, 1992) to estimate main genetic effects consistently and are able, then, to test for interactions. The resulting test maintains robustness to population stratification, avoids assumptions on the phenotypic and allele frequency distributions and accounts for sample ascertainment. We assess the performance of this test empirically through extensive simulation studies. We apply these new techniques to a study of COPD.

email: david.fardo@uky.edu

A GENERAL FRAMEWORK FOR STUDYING GENETIC EFFECTS AND GENE-ENVIRONMENT INTERACTIONS WITH MISSING DATA

Yijuan Hu*, University of North Carolina-Chapel Hill
Danyu Lin, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill

Missing data arise in genetic association studies when genotypes are unknown or when haplotypes are of direct interest. We provide a general likelihood-based framework for making inference on genetic effects and gene-environment interactions with such missing data. We allow genetic and environmental variables to be correlated while leaving the distribution of environmental variables completely unspecified. We consider three major study designs ---cross-sectional, case-control, and cohort designs ---and construct appropriate likelihood functions for all common phenotypes (e.g., case-control status, quantitative traits, and potentially censored ages at onset of disease). The likelihood functions involve both finite- and infinite-dimensional parameters. The maximum likelihood estimators are shown to be consistent, asymptotically normal, and asymptotically efficient. EM algorithms are developed to implement the corresponding inference procedures. Extensive simulation studies demonstrate that the proposed inferential and numerical methods perform well in practical settings. Illustration with a genome-wide association study of lung cancer is provided.

email: yhu@bios.unc.edu

80. BAYESIAN METHODS AND APPLICATIONS

ROBUSTNESS OF NONPARAMETRIC BAYESIAN METHODS

Steven N. MacEachern, The Ohio State University

Nonparametric Bayesian methods provide a means of flexibly modelling data. They have been used successfully for problems ranging from exploratory data analysis to sharp, focused inference in sophisticated hierarchical models. They now come in a wide

variety of forms: One is able to capture both a single nonparametric distribution and a collection of nonparametric distributions, whether the collection be finite, countable or uncountable. An oft-touted benefit of the methods is their robustness to a violation of parametric assumptions. While the models can fit data quite well, their very flexibility can render these fits non-robust. This talk provides an overview of situations where the methods have shown a lack of robustness. The mechanisms which lead to the lack of robustness are described, concepts are formalized, and results given. Strategies for improving the robustness of the models are provided. Particular attention is given to the structures that distinguish robust from non-robust models.

email: snm@stat.osu.edu

MODELING RELATIONAL DATA USING NESTED PARTITION MODELS

Abel Rodriguez, University of California-Santa Cruz
Kaushik Ghosh, University of Nevada Las Vegas

This paper introduces a flexible class of models for relational data based on a hierarchical extension of the two-parameter Poisson-Dirichlet process. The model is motivated by two different applications: 1) A study of cancer mortality rates in the U.S., where rates for different types of cancer are available for each state, and 2) the analysis of microarray data, where expression levels are available for a large number of genes in a sample of subjects. In both these settings, we are interested in improving estimation by flexibly borrowing information across rows and columns while partitioning the data into homogeneous subpopulations. Our model allows for a novel nested partitioning structure in the data not provided by existing nonparametric methods, in which rows are clustered while simultaneously grouping together columns within each cluster of rows.

email: kaushik.ghosh@unlv.edu

BAYESIAN INFERENCE ON PARAMETERS OF A ZERO INFLATED NEGATIVE MULTINOMIAL DISTRIBUTION

Santanu Chakraborty*, University of Texas Pan American

Generally for modeling count data, integer valued random variables like a Poisson or a Negative Binomial are the ideal distributions. But on certain occasions, there are too many zeros in a count data set compared to a usual Poisson or a Negative Binomial data set. In those case, one uses zero inflated versions of these distributions which already exist in the literature. For example, there had been experiments to estimate the proportion of sterile couples using zero inflated negative binomial distribution. In such an experiment, if a couple is not sterile, it may be considered to be a failure. Now one can think of classifying these failures - like couples having one child, two children, three children etc. So, the failure can be thought of as a vector. What could be its distribution? Negative multinomial distribution already exists in the literature. But in this particular experiment, zeros would be occurring more frequently for each coordinate of the failure vector. In such a case, a zero inflated version of the negative multinomial would be ideal. In this talk, we formalize the concept of zero inflated negative multinomial distribution and do some Frequentist and Bayesian inferential studies.

email: schakraborty@utpa.edu

PERFORMANCE OF BAYESIAN RANKING METHODS FOR IDENTIFYING THE EXTREME PARAMETER

Yi-Ting Chang*, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University

The Bayesian approach provides a unified framework for estimating unit specific parameters in multilevel models. Several optimal estimators of ranks under different loss functions have been proposed, and performance evaluations have been conducted as well. However, in some cases, our focus is only on identifying the unit with the largest underlying value rather than the whole population. Thus, we compare the performances of the different rank estimators in the literature by simulations in terms of identifying the top-ranked (spiked) unit. We also apply these ranking methods to simulated genetic outcomes from a real data drawn from the International HapMap Project.

email: ychang@jhsph.edu

AN EFFICIENT MARKOV CHAIN MONTE CARLO METHOD FOR MIXTURE MODELS BY NEIGHBORHOOD PRUNING

Youyi Fong*, University of Washington
Jon Wakefield, University of Washington
Ken Rice, University of Washington

Inference on both finite mixture models and infinite mixture models involves a partition parameter, which is the clustering of the observations into groups. Because the partition is a discrete parameter and can take an massive number of values, it presents a challenging task for making inference. In this paper, we focus on devising efficient Metropolis-Hastings methods for models in which parameters other than the partition parameter can be integrated out. By drawing an analogy between the Metropolis-Hastings (MH) algorithm and the Stochastic Local Search (SLS) algorithm, we introduce several concepts from the SLS to guide the design of MH algorithms, including neighborhood pruning. We propose a new neighborhood pruning method for clustering problem based on bottom-up hierarchical clustering, and use it to design a Markov chain Monte Carlo method for mixture models. Through experiments on four datasets of varying complexity, we

demonstrate that our method improves the mixing for all datasets. The improvement is significant in some cases.

email: youyifong@gmail.com

NONPARAMETRIC BAYES STOCHASTICALLY ORDERED LATENT CLASS MODELS

Hongxia Yang*, Duke University
David Dunson, Duke University
Sean OBrien, Duke University

Latent class models (LCMs) are used increasingly for addressing a broad variety of problems, including sparse modeling of multivariate and longitudinal data, model-based clustering, and flexible inferences on predictor effects. Typical frequentist LCMs require estimation of a single finite number of classes, which does not increase with the sample size, and have a well-known sensitivity to parametric assumptions on the distributions within a class. Bayesian nonparametric methods have been developed to allow an infinite number of classes in the general population, with the number represented in a sample increasing with sample size. However, Bayes methods relying on Markov chain Monte Carlo sampling encounter a challenging label ambiguity problem, which makes it difficult to perform inferences on class-specific quantities. In this article, we propose a new nonparametric Bayes model that allows predictors to flexibly impact the allocation to latent classes, while limiting sensitivity to parametric assumptions and label switching problems by allowing class-specific distributions to be unknown subject to a stochastic ordering constraint. An efficient MCMC algorithm is developed for posterior computation. The methods are validated using simulation studies and applied to the problem of ranking medical procedures in terms of the distribution of patient morbidity.

email: hy35@duke.edu

81. SPATIAL/TEMPORAL APPLICATIONS AND INFECTIOUS DISEASE MODELING

BAYESIAN GEOSTATISTICAL MODELING WITH INFORMATIVE SAMPLING LOCATIONS

Debdeep Pati*, Duke University
Brian J. Reich, North Carolina State University
David B. Dunson, Duke University

We consider geostatistical models that allow the locations at which data are collected to be informative about the outcomes. Diggle et al. (2009) refer to this problem as preferential sampling, though we use the term informative sampling to highlight the relationship with the longitudinal data literature on informative observation times. In the longitudinal setting, joint models of the observation times and outcome process are widely used to adjust for informative sampling bias. We propose a Bayesian geostatistical joint model, which models the locations using a log Gaussian Cox process, while modeling the outcomes conditionally on the locations as Gaussian with a Gaussian process spatial random effect and adjustment for the location intensity process. We prove posterior propriety under an improper prior on the parameter controlling the degree of informative sampling, demonstrating that the data are informative. In addition, we show that the density of the locations and mean function of the outcome process can be estimated consistently under mild assumptions. The methods are applied to ozone data.

email: dp55@stat.duke.edu

INDEPENDENT COMPONENT ANALYSIS FOR COLORED SOURCES WITH APPLICATION TO FUNCTIONAL MAGNETIC RESONANCE IMAGING

Seonjoo Lee*, University of North Carolina-Chapel Hill
Haipeng Shen, University of North Carolina-Chapel Hill
Young Truong, University of North Carolina-Chapel Hill

Independent component analysis (ICA) is an effective data-driven method for blind source separation. Most existing source separation procedures are carried out by relying solely on the estimation, parametrically or non-parametrically, of the marginal density functions. In many of these applications, the correlated structures within each source signal can play a more important role than the marginal distributions. This is often the case in functional magnetic resonance imaging (fMRI) analysis where the brain-function-related signals are temporally correlated. In this paper, we consider a novel approach to ICA that will fully exploit the temporal correlation of the source signals. Specifically, we propose to estimate the spectral density functions of the source signals instead of the marginal density functions. This is possible by virtue of the intrinsic relationship between the (unobserved) source and the (observed) mixed signals. A methodology will be described and implemented using spectral density functions from frequently used time series models such as autoregressive and moving-average (ARMA) processes. The time series parameters and the mixing coefficients will be estimated via the Whittle likelihood function. Extensive numerical results indicated that our approach has outperformed several popular methods including the most widely used fast ICA.

email: seonjool@email.unc.edu

A HIERARCHICAL MODEL FOR PREDICTING FOREST VARIABLES OVER LARGE HETEROGENEOUS DOMAINS

Andrew O. Finley*, Michigan State University
Sudipto Banerjee, University of Minnesota

We are interested in predicting one or more continuous forest variables (e.g., biomass, volume, age) at a fine resolution (e.g., pixel-level) across a specified domain. Given a definition of forest/

non-forest, this prediction is typically a two step process. The first step predicts which locations are forested. The second step predicts the value of the variable for only those forested locations. Rarely is the forest/non-forest predicted without error. However, the uncertainty in this prediction is typically not propagated through to the subsequent prediction of the forest variable of interest. Failure to acknowledge this error can result in biased and perhaps falsely precise estimates. In response to this problem, we offer a modeling framework that will allow propagation of this uncertainty. Here we envision two latent processes generating the data. The first is a continuous spatial process while the second is a binary spatial process. We assume that the processes are independent of each other. The continuous spatial process controls the spatial association structure of the forest variable of interest, while a binary process indicates presence of a 'measurable' quantity at a given location. Finally, we explore the use of a predictive process for both the continuous and binary processes to reduce the dimensionality of the data and ease the computational burden. The proposed models are motivated using georeferenced National Forest Inventory (NFI) data and coinciding remotely sensed predictor variables.

email: finleya@msu.edu

ESTIMATING CASE FATALITY RATIOS FROM INFECTIOUS DISEASE SURVEILLANCE DATA

Nicholas G. Reich*, Johns Hopkins University
Justin Lessler, Johns Hopkins University
Ron Brookmeyer, Johns Hopkins University

Understanding the virulence of an emerging infectious disease such as H1N1 can help direct resources during an outbreak response. The case fatality ratio, a measure of virulence, is the fraction of cases who die after contracting a disease. Incomplete reporting of the number of infected individuals, both recovered and dead, can lead to biased estimates of the case fatality ratio when using a naive estimator. We propose a simple estimator for the relative case fatality ratio which controls for reporting rates that may vary across time and is asymptotically unbiased in these situations. Further, we generalize our methods to account for elapsed time between infection and death. We conduct a simulation study to evaluate the performance of our methods across a range of realistic outbreak scenarios. We also evaluate the sensitivity of our estimator to the model assumptions. Our new methods could provide valuable information early in an epidemic about which subgroups of the population are most vulnerable to dying from infection with an emerging pathogen such as the current H1N1 pandemic virus.

email: nreich@jhsph.edu

MODELING THE SPATIO-TEMPORAL TRANSMISSION OF INFECTIOUS DISEASES IN ANIMALS

Christel Faes*, Hasselt University
Marc Aerts, Hasselt University
Niel Hens, Hasselt University

Bluetongue serotype 8 was introduced into North-West Europe in 2006 and spread to several European countries, affecting mainly sheep and cattle. Individual-based spatio-temporal models can be used to assess the degree of spread between farms, by modeling the spatial interactions between farms. This can become computationally very complex when the spatial window and the number of farms are large. Meta-population models treating each municipality as a subpopulation, is an attractive alternative method. The force of infection can be modeled as a weighted sum of the prevalence's in all neighboring areas. Several factors have to be taken into account in such a model. First, environmental risk factors for the spread of the disease, such as the animal density, the temperature, precipitation, and land use have to be considered. These risk factors describe the activity of the vector. Second, the wind-direction enhances the spread of the disease since the vector can move over long distances via the wind. Third, the transport rate of animals between areas needs to be considered as well. It is investigated how each of these factors can be accounted for in a spatio-temporal meta-population model. The methods are illustrated using the 2006 outbreak of Bluetongue in Belgium, the Netherlands, Luxembourg and Germany.

email: christel.faes@uhasselt.be

ASSESSING THE SPATIAL VARIABILITY OF SYPHILIS IN BALTIMORE IN THE 1990S USING GEOGRAPHICALLY WEIGHTED REGRESSION

Jeffrey M. Switchenko*, Emory University
Lance A. Waller, Emory University

Syphilis remains a significant cause of morbidity in many developing countries and in some areas within North America and Europe. Past studies have noted substantial associations between demographic variables and high disease incidence. The analyses in this presentation will build on earlier studies which have focused on the presence of core areas of STD transmission in Baltimore, Maryland in the mid-1990s. Core areas are primarily defined geographically and can be characterized by socioeconomic factors such poverty and poor health care access. We will illustrate spatial variations on the significant demographic characteristics over the study area, and determine not only the strongest risk factors for disease transmission, but also how those effects vary over Baltimore City County. Geographically weighted regression (GWR) is a technique for exploratory data analysis, which allows the relationships of interest to vary over space. With GWR, instead of assuming fixed global parameter estimates, estimates can vary according to a position in space, characterized by latitudinal and

longitudinal coordinates. We will explore how certain demographic variables vary spatially through the use of GWR in both linear and Poisson regression forms.

email: jswitch@emory.edu

DETECTING DISEASE OUTBREAKS USING LOCAL SPATIOTEMPORAL METHODS

Yingqi Zhao*, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill
Amy H. Herring, University of North Carolina-Chapel Hill
David Richardson, University of North Carolina-Chapel Hill
Michael R. Kosorok, University of North Carolina-Chapel Hill

A real-time surveillance method is developed with emphasis on rapid and accurate detection of emerging outbreaks. We develop a model with relatively weak assumptions regarding the latent processes generating the observed data, ensuring a robust prediction of the spatiotemporal incidence surface. Estimation occurs via a local linear fitting combined with day-of-week effects, where spatial smoothing is handled by a novel distance metric that adjusts for population density. Detection of emerging outbreaks is carried out via residual analysis. Both daily residuals and AR model-based de-trended residuals are used for detecting abnormalities in the data given that either a large daily residual or a large temporal change in the residuals signals a potential outbreak, with the threshold for statistical significance determined using a resampling approach.

email: yqzhao@email.unc.edu

82. IMS MEDALLION LECTURE

A STATISTICIAN'S ADVENTURES IN COLLABORATION: DESIGNING BETTER TREATMENT STRATEGIES

Marie Davidian, North Carolina State University

Over the past decade, there has been an increasing focus in a host of disease and disorder areas on not only traditional development and evaluation of new interventions but on elucidating more broadly the best ways to use both new and existing treatments. The recent emphasis by policymakers on comparative effectiveness research has only heightened this interest. Statisticians, mathematicians, and other quantitative scientists have developed various methodological approaches that can inform and guide this endeavor, particularly in the context of designing strategies for best using available treatment options over the entire course of a disease or disorder. Adopting the perspective that "treatment" of chronic diseases and disorders really involves a series of therapeutic decisions over time, each of which should ideally be made adaptively based on the evolving health status of the patient, holds the promise of developing treatment regimens that in this sense move toward "individualizing" treatment to the patient, thereby improving health outcomes. Multidisciplinary collaborations involving disease/disorder area specialists, statisticians, and other quantitative scientists to exploit these methodological advances are essential if this goal is to be achieved. I am fortunate to be involved in several collaborative research projects focused on development of such adaptive treatment strategies in areas ranging from HIV infection to alcohol abuse to organ transplantation. I will describe these exciting research projects and the statistical and mathematical methods that are being brought to bear to address this challenge.

davidian@stat.ncsu.edu

83. SHRINKAGE ESTIMATION IN MICROARRAY DATA ANALYSIS

LEMMA: LAPLACE APPROXIMATED EM MICROARRAY ANALYSIS

Bar Haim, Cornell University
James G. Booth*, Cornell University
Elizabeth Schifano, Cornell University
Martin T. Wells, Cornell University

A mixture of mixed-effects model for the analysis of microarray data is proposed. Approximate maximum likelihood fitting is accomplished via a stable and fast EM-type algorithm. Posterior odds of treatment/gene interactions, derived the model, involve shrinkage estimates of both the interactions and of the gene specific error variances. Genes are classified as being associated with treatment based on the posterior odds or equivalently using local FDR with a fixed q-value cutoff. Simulation studies show that the approach outperforms some well-known competitors.

email: jim.booth@cornell.edu

BORROWING INFORMATION ACROSS GENES AND ACROSS EXPERIMENTS FOR IMPROVED RESIDUAL VARIANCE ESTIMATION IN MICROARRAY DATA ANALYSIS

Tieming Ji, Iowa State University
Peng Liu, Iowa State University
Dan Nettleton*, Iowa State University

Statistical inference for microarray experiments usually involves the estimation of residual variance for each gene. Because the sample size available for each gene is often low, the usual unbiased estimator of the residual variance based on a linear model fit can be unreliable. Shrinkage methods, including empirical Bayes approaches that borrow information across genes to produce more stable estimates, have been developed by several research groups in recent years. Because a single laboratory will often conduct a sequence of similar experiments using the same microarray technology, there is an opportunity to improve variance estimation further by borrowing information not only across genes but also

across experiments. We propose a log-normal model for residual variances that involves a sum of random gene effects and random experiment effects. Based on the model, we develop an empirical Bayes estimator of the residual variance for each combination of gene and experiment. Statistical inference is carried our via a permutation testing strategy. We illustrate the advantages of our method over existing approaches via simulation.

email: dnett@iastate.edu

MIXTURE PRIORS FOR VARIABLE SELECTION WITH APPLICATION IN GENOMICS

Marina Vannucci*, Rice University

This talk will start with a brief review of Bayesian methods for variable selection in linear models that use mixture priors. Models and inferential algorithms are quite flexible and allow to incorporate additional information, such as data substructure and/or knowledge on relationships among the variables. Specific interest will be towards high-dimensional genomic data, and in particular DNA microarrays. The vast amount of biological knowledge accumulated over the years has allowed researchers to identify various biochemical interactions among genes and to define pathways as groups of genes that share same functions. There is an increased interest in identifying pathways and pathway elements involved in particular biological processes. Drug discovery efforts, for example, are focused on identifying biomarkers as well as pathways related to a disease. In the talk we show how information on pathways and gene networks can be incorporated into a Bayesian model. We illustrate the method with an application to gene expression data with censored survival outcomes. In addition to identifying markers that would have been missed otherwise and improving prediction accuracy, the integration of existing biological knowledge into the analysis provides a better understanding of underlying molecular processes.

email: marina@rice.edu

ASSESSING DIFFERENTIAL GENE EXPRESSION USING A NONPARAMETRIC MEAN-VARIANCE SMOOTHING: APPLICATION TO ARABIDOPSIS THALIANA ABIOTIC STRESS MICROARRAY EXPERIMENTS

Taps Maiti*, Michigan State University
Pingsha Hu, Michigan State University

Differential gene identification has generated numerous statistical inferential issues in recent years. “Borrowing strength” is a common technique to avoid small sample size problem and to exploit gene dependency structure. Shrinking variance using a simple parametric model proved to be useful in gene expression data analysis. In this talk we propose a spline based shrinkage estimation of gene specific variances and develop methods to assess differential gene expressions using empirical Bayes techniques. The method is applied to transcriptional profiling data sets of Arabidopsis thaliana from various stress conditions.

email: maiti@stt.msu.edu

84. OPPORTUNITIES FOR BIOSTATISTICIANS INSIDE (RESEARCH) AND OUTSIDE (FUNDING) OF NIH

BIOSTATISTICS AT THE NATIONAL HEART, LUNG, AND BLOOD INSTITUTE OF THE NATIONAL INSTITUTES OF HEALTH (NIH)

Nancy L. Geller*, National Heart, Lung, and Blood Institute

The Office of Biostatistics Research at the National Heart, Lung, and Blood Institute (NHLBI) has a three part mission: collaboration in the design and analyses of studies funded by NHLBI, collaboration in data management and analysis of studies sponsored by the Division of Intramural Research and undertaking methodological research. The NHLBI environment is conducive to collaboration and team science and statistical collaborations result from this interactive environment. Good communication skills are essential. The career path of NHLBI statisticians is described and suggestions for achieving success in this environment are given.

email: ng@helix.nih.gov

MY RESEARCH EXPERIENCE AT THE NIH AND UTSPH

Sheng Luo*, University of Texas School of Public Health

In this talk, I will share my experience as a Pre-doctoral fellow in Biostatistics Branch at NCI while I was a PhD candidate of Biostatistics department of Johns Hopkins University. I will discuss the opportunities of conducting statistical methodological and collaborative epidemiological research in the intramural research environment within NCI. I will also discuss how the NCI experience helped me in my transition into the position of assistant professor at UTSPH, e.g., develop local collaboration, identify workable research problems, recruit graduate students, etc.

Sheng.T.Luo@uth.tmc.edu

NIH FUNDING OPPORTUNITIES FOR BIOSTATISTICIANS

Michelle C. Dunn*, National Cancer Institute

An overview of NIH funding opportunities for biostatisticians will be given. The first topic will be funding mechanisms (including those for new investigators), from the well-known R01 research grant to the more obscure K25 mentored quantitative research award. Next, the NIH application and review process will be described, particularly for research grants. Finally, resources for seeking help and further information will be shared.

email: dunnm3@mail.nih.gov

APPLYING FOR BIOSTATISTICAL RESEARCH GRANTS FROM NIH: WHY? WHAT? AND HOW?

Hulin Wu*, University of Rochester

Many statisticians are now working in the environment of the Department of Biostatistics at a Medical School or a School of Public Health with "soft money" support, which means that the major portion of our salary is supported by research grants. The major source of biomedical and biostatistical research grants, which are quite competitive, is the National Institutes of Health (NIH). Usually as a biostatistician, if we can provide statistical support to biomedical investigators' research grants from which we can share a piece of the pie to support our effort and salary, it is good enough for us to survive. However, nowadays more and more academic departments of Biostatistics require their faculty to secure their own research grants as a Principle Investigator (PI), instead of Statistician or Co-Investigator in biomedical investigators' grants, in order to get promotion or some merit reward. However, with low NIH paylines, it is quite challenging for biostatisticians to get our own grants. What are good strategies for us to face this challenge? There are several channels to achieve our goals, which include the general statistical methodology grant, biomedical sciences-oriented statistical grants, and statistical cores in large center or program grant applications. We will exchange ideas and share our experience on how to optimize our chance to get NIH grants.

email: hwu@bst.rochester.edu

85. ROC METHODS: EXPERIMENTS WITH TIMEDEPENDENT AND CLUSTERED DATA

STRATIFIED AND CLUSTERED RECEIVER OPERATING CHARACTERISTIC ANALYSIS

Kelly H. Zou*, Pfizer Inc.
Simon K. Warfield, Children's Hospital Boston and Harvard Medical School

Post-hoc analyses often deals with subgroups, particularly the effect of strata or clusters. In this research, both nonparametric and parametric methods for estimating the receiver-operating characteristics measures, along with the area under the curve, will be presented and contrasted. The overall unadjusted and adjusted measures will be compared both theoretically and via Monte-Carlo simulations. An example on medical imaging will be provided for illustrations.

email: Kelly.Zou@pfizer.com

EVALUATION OF BIOMARKER ACCURACY UNDER NESTED CASE-CONTROL STUDIES

Tianxi Cai*, Harvard University
Yingye Zheng, Fred Hutchinson Cancer Research Institute

Accurate prediction has significant bearing on the optimization of therapeutic and monitoring plans for each individual. The rapid emergence of new biological and genetic markers holds great promise for improving risk prediction. To determine the clinical utility of these markers, a crucial step is to evaluate their predictive accuracy with prospective studies. However, due to the financial and medical cost associated with marker measurement, it is often undesirable and/or infeasible to obtain marker values for the entire study population. To overcome such difficulties, the nested case-control design is often employed as a cost-effective strategy for conducting studies of biomarker evaluation. Under such a study design, the biomarkers are only ascertained for the cases who developed events as well as a fraction of controls. In this research, we proposed estimation procedures for commonly used accuracy measures with data from NCC studies. The accuracy estimators were shown to be consistent and asymptotically normal. Simulation results suggest that the proposed procedures perform well in finite samples. The proposed procedures were illustrated with data from the Framingham Offspring study to evaluate the accuracy of a recently develop risk score with C-reactive protein information for predicting cardiovascular disease in women.

email: tcai@hsph.harvard.edu

SUMMARIZING PERFORMANCE IN FROC EXPERIMENTS

Howard E. Rockette*, University of Pittsburgh
Andriy Bandos, University of Pittsburgh

The common method of evaluating detection performance of diagnostic imaging systems is based on summary indices such as the area under the Receiver Operating Characteristics (ROC) curve. A free-response ROC (FROC) experiment, which addresses the detection and localization performance, allows identifying and scoring sections of the image. Thus, each image has an apriori unknown number of marks that were found suspicious and were assigned a rating. Unlike ROC, FROC analysis emphasizes separability of the ratings as well as the average number of

false positive and true positive marks. Summary indices that inadequately reflect any of these three aspects can lead to results inconsistent with the empirical FROC curve. Sensitivity to all three aspects as well as a useful interpretation are important considerations in constructing a summary index. Previously we have proposed a conveniently interpretable index. Here we will discuss a flexible index that permits conducting efficient assessments for each of the three characteristics separately or "trading off" sensitivity to one characteristic in order to increase sensitivity to another. We will also consider the relationship to other currently used FROC summary statistics.

email: herbst@pitt.edu

SAMPLE SIZE CONSIDERATIONS FOR TIME-DEPENDENT ROC ESTIMATION

Hong Li, Harvard University
Constantine Gatsonis*, Brown University

In contrast to the usual ROC analysis with a contemporaneous reference standard, time-dependent ROC methods are applicable to settings in which the reference standard depends on a future event. In such settings, the reference standard may not be known for every patient because of censoring. The goal of this research is to determine the required sample size for estimating the area under the curve (AUC) in time-dependent ROC analysis. We adopt a previously published estimator of the time-dependent AUC, which is a function of the expected conditional survival functions. The calculation of the required sample size is based on approximations of the expected conditional survival functions and their variances, derived under parametric assumptions of an exponential failure time and an exponential censoring time. We consider alternative patterns for patient entry into the study and present results of a simulation designed to assess the accuracy of the method and its robustness to departures from the parametric assumptions. We apply the proposed method to the design of a study of PET as predictor of disease free survival in women undergoing therapy for cervical cancer.

email: gatsonis@stat.brown.edu

86. ANALYSIS OF CLINICAL TRIALS AND BIOPHARMACEUTICAL STUDIES

COMPARISON OF ESTIMATES FOR THE COMMON CORRELATION COEFFICIENT IN A STRATIFIED EXPERIMENT

Yougui Wu, University of South Florida
Jason Liao*, Merck Research Laboratories

Many experiments are conducted in a stratified manner in practice and it is often of the interest to estimate the common correlation coefficient in all strata. In this paper, the maximum likelihood estimate (MLE), the commonly used sample size weighted linearly combined estimate, and the newly proposed model based estimate are compared. The MLE does not have a closed form and needs an iteration to obtain the estimate, while the other two have a closed form. Instead of using the complicated second derivatives from the Fisher information matrix as the variance, an approximation for the variance of MLE is derived. The simulation study indicates that the newly proposed model based estimate and the MLE are very comparable and both perform better than the commonly used sample size weighted linearly combined estimate in terms of the bias and mean squared error (MSE). Due to the closed form and computation simplicity, the newly proposed model based estimate is recommended for estimating the common correlation coefficient in a stratified experiment. A real data set is used to illustrate these estimates.

email: jason_liao@merck.com

A CONSISTENCY-ADJUSTED STRATEGY FOR TESTING ALTERNATIVE ENDPOINTS IN A CLINICAL TRIAL

Mohamed Alosh,* U.S. Food and Drug Administration

A clinical trial might involve more than one clinically important endpoint (subgroup) each of which can characterize the treatment effect of the experimental drug under investigation. For prespecifeid alternative endpoints (subgroups) there are several approaches which can be used for testing for efficacy for the alternative endpoints or the subgroup and total study population. Traditional multiplicity approaches use constant significance levels for these alternative endpoints. However, some recent multiplicity strategies allow the alpha-level allocated to testing a subsequent endpoint to be dependent on the results of the previous endpoint. In this presentation we discuss the need for establishing a minimum level of efficacy for the previous endpoint before proceeding to test for the subsequent alternative endpoint (subgroup) so that potential problems in interpreting study findings can be avoided. We consider implementing such requirements, which we call consistency criteria, along with adaptation of the significance level for subsequent endpoints at the study design stage and investigate its impact on study power. In addition, we consider its application to actual clinical trial data.

email: Mohamed.Alosh@fda.hhs.gov

ADAPTIVE CONFIDENCE INTERVALS FOR NON-REGULAR PARAMETERS IN DYNAMIC TREATMENT REGIMES

Min Qian*, University of Michigan
Eric B. Laber, University of Michigan
Susan A. Murphy, University of Michigan

Dynamic treatment regimes are often used to operationalize multi-stage decision making in the medical field. A dynamic

treatment regime is a sequence of decision rules that specify how the intensity or type of treatment should change depending on patient characteristics. Common approaches to constructing dynamic treatment regimes, such as Q-Learning, employ non-smooth functionals of the data. Due to the non-smooth operation, the fitted coefficients in a decision rule prior to the last stage have non-regular asymptotic distributions. In particular, if the optimal treatment at the last stage is unique then the limiting distributions are normal, however, if two or more treatments in the last stage have comparable effects then the limiting distributions of the fitted coefficients are non-normal. As a consequence, standard approaches to forming confidence intervals (e.g. bootstrap, Taylor series arguments) may fail to provide nominal coverage. Existing methods in the literature are either too conservative or lacking in theoretical justification. In this talk, we present a bootstrap based method for constructing asymptotically valid confidence intervals. This method is adaptive in the sense that it provides exact coverage when the last stage optimal treatment is unique and is conservative otherwise. Empirical studies show that the amount of conservatism is small.

email: minqian@umich.edu

PREDICTION OF THE BIGGEST LOSER: A BRIDGE CONNECTING OBESITY CLINICAL OUTCOMES AND ANIMAL MODELS

Yuefeng Lu*, Eli Lilly and Company

The pharmaceutical indsutry will be experiencing an overwheleming patent loss over the next decade. A billion dollar question is to unveil the connection (or lack of connection) between animal models and clinical outcomes, and truly make innovations to predict the probability technical success of clinical trials with preclinical data package beyond the allometric scaling. In this talk, I will focus on the translational research in the obesity area. The connection between the animal models and clincal outcomes is modeled by a functional statistical model and a mechanism-based physioloigcal model. I will discuss the values and limits of these models for practical uses.

email: lu_yuefeng@lilly.com

SEMIPARAMETRIC CAUSAL INFERENCE FOR RANDOMIZED CLINICAL TRIALS WITH A TIME-TO-EVENT OUTCOME AND ALL-OR-NONE TREATMENT-NONCOMPLIANCE

Ying Zhou*, University of California-Los Angeles
Gang Li, University of California-Los Angeles
Huazhen Lin, Sichuan University, China

To evaluate the biologic efficacy in two-arm randomized clinical trials with all-or-none treatment noncompliance, one needs to study the treatment effect relative to the control in the treatment compliance subgroup of the study population. We develop a new semiparametric method to estimate and compare treatment and control survival functions among treatment compliers who are typically not identifiable in the control arm. Unlike a naive estimator of the survival function commonly used in the literature that is not necessarily monotonically non-increasing, our new estimator is a proper survival function. Large sample properties and inference are developed. We illustrate our method using both simulated and real data.

email: yzhou7@ucla.edu

POOLING STRATEGIES FOR OUTCOME UNDER A GAUSSIAN RANDOM EFFECTS MODEL

Yaakov Malinovsky*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Paul S. Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development

Longitudinal studies of biomarkers outcome are often challenging to conduct due to the cost of the assays. This work investigates the efficiency of different pooling strategies for estimating the mean structure, the random effect variance, the residual error variance, as well as individual estimation of the random effects. We investigate optimal design strategies for the variances component estimation using analytic results and simulations. We illustrate our proposed design using a longitudinal cohort study.

email: Yaakov.Malinovsky@nih.gov

M-ESTIMATION PROCEDURES IN HETEROSCEDASTIC NONLINEAR REGRESSION MODELS WITH PARAMETRIC VARIANCE MODEL

Changwon Lim*, National Institutes of Health
Pranab K. Sen, University of North Carolina-Chapel Hill
Shyamal D. Peddada, National Institutes of Health

Nonlinear regression models are commonly used in dose-response studies, especially when researchers are interested in determining various toxicity characteristics of a chemical or a drug. When fitting nonlinear models for toxicology data, one needs to pay attention to error variance structure in the model and the presence of possible outliers or influential observations. In this talk, M-estimation procedures are considered in heteroscedastic nonlinear regression models for the case where the error variance is modeled by a nonlinear function that may be appropriate in toxicological data. We propose M-estimators for various parameters and derive their asymptotic properties under suitable regularity conditions. The proposed methodology is illustrated using a toxicological data.

email: limc2@niehs.nih.gov

87. CLUSTERED DATA METHODS

ANALYSIS OF UNBALANCED AND UNEQUALLY-SPACED FAMILIAL DATA USING GENERALIZED MARKOV DEPENDENCE

Roy T. Sabo*, Virginia Commonwealth University

Familial correlations between age-dependent health outcomes often exhibit dependence that is inversely related to age differences between family members. These relationships are further complicated by varying age distributions between families, as well as unequal family sizes. Rather than impose a somewhat arbitrary and possibly ill-fit auto-regressive structure, familial dependence is here modeled using an extension of the generalized Markov structure, which incorporates age differences between family members and allows for a dampening parameter to modify the effect of those differences. General properties of this structure for one-parent families are discussed, and both maximum likelihood and quasi-least-squares estimators are derived. Metabolic data from the Fels Longitudinal Study are included as an example, where we examine familial relationships between childhood metabolic biomarkers and adult obesity.

email: rsabo@vcu.edu

IDENTIFYING DISTINCT SUBTYPES IN ACUTE MYELOID LEUKEMIA: A MODEL BASED CLUSTERING APPROACH

Matthias Kormaksson*, Cornell University

A recent study on a cohort of 344 well-characterized patients with acute myeloid leukemia suggests that subjects can be segregated into distinct groups using unsupervised clustering based on their DNA methylation profiles. Simple hierarchical clustering methods based on a Euclidean distance metric or correlation similarity have given some promising results. We suggest a model based approach, where we introduce latent cluster specific methylation indicators on each gene. These indicators along with some standard assumptions impose a specific mixture distribution on each cluster and the parameters of the induced model are estimated using the EM algorithm. The results of our method compare well with biological traits of the patients and also provide output that give insight into which genes are driving the differences between clusters.

email: mk375@cornell.edu

WEIGHTED SCORES METHOD TO ANALYZE MULTIVARIATE OVERDISPERSED COUNT DATA

Aristidis K. Nikoloulopoulos, Athens University of Economics and Business
Harry Joe, University of British Columbia
N. Rao Chaganty*, Old Dominion University

There exists multivariate probability models based on copulas for clustered and longitudinal overdispersed counts. Continuously and rapidly changing technological advances in computer hardware and software are making it possible to fit these models for data using maximum likelihood method, which is the optimal procedure for parameter estimation and inference. However, if the main interest is in the regression and other univariate parameters and not the dependence, then we propose a 'weighted scores method', which is based on weighting score functions of the univariate margins. The weights for the univariate scores are obtained fitting the multivariate normal copula model. Our method can be viewed as a generalization of the generalized estimating equations, and it is also applicable to families that are not in the GLM class. We present the application of our general methodology to negative binomial regression. Asymptotic and small sample efficiency calculations show that the weighted scores method is robust and nearly as efficient as the maximum likelihood for fully specified copula models. An illustrative example is given to show the use of our weighted scores method to analyze utilization of health care based on characteristics of the families.

email: rchagant@odu.edu

RE-SAMPLING BASED METHOD TO ESTIMATE INTRA-CLUSTER CORRELATION FOR CLUSTERED BINARY DATA

Hrishikesh Chakraborty*, RTI International
Pranab K. Sen, University of North Carolina-Chapel Hill

Various methods have been proposed to estimate intra-cluster correlation (ICC) for correlated binary data and they are very specific to the type of design and underlying distributional assumptions. The analysis of variance (ANOVA) based estimation technique to estimate ICC is the most widely used method in cluster randomized trials. We proposed a new method to estimate intra-cluster correlation (ICC) and its variance using re-sampling without replacement principle and U-statistics. The main advantage of this proposed method is that it can be used for any type of categorical variable without making any additional distributional assumptions. We created a Monti Carlo simulation exercise and compared our ICC estimates to those estimates by the most widely used ANOVA method. We found that if the binary proportion is large then both methods provide similar ICC estimates for varying number of clusters and sizes. However, for small binary proportions our method provides more accurate ICC estimates than to the

ANOVA method does for different numbers of clusters and cluster sizes.

email: hchakraborty@rti.org

A BI-DIRECTIONAL RANDOM EFFECTS SPECIFICATION FOR HETEROGENEITY IN MIXED EFFECTS MODELS

Zhen Chen*, National Institutes of Health

In analyzing longitudinal and repeated measure data, mixed effects models are often used to account for dependence due to clustering. However, the heterogeneity structure implied by a mixed effects specification may not conform with what the data present, which can lead to incorrect inferences. For example, in a regression model with a binary (0/1) group predictor, a random intercept model assumes equal heterogeneity in the two groups. When observed heterogeneity differ between the two groups, a random intercept model may produce erroneous inference results. A model with both random intercept and random slope is not foolproof either, as it assumes a higher level of heterogeneity in group 1 than in group 0. When the heterogeneity structure in the data goes the other direction, biased estimates and/or incorrect type I error may result. In this paper, I propose a bi-directional random effects specification that accommodates flexible heterogeneity structure in correlated data. Through simulations, I show that the proposed approach has good performance in terms of coverage probabilities and biases. I also demonstrate the use of the bi-directional random effects model in examples of teen driving behavior and lung cancer interventions.

email: chenzhe@mail.nih.gov

REGRESSION ANALYSIS OF CLUSTERED INTERVAL-CENSORED FAILURE TIME DATA WITH INFORMATIVE CLUSTER SIZE

Xinyan Zhang*, University of Missouri
Jianguo Sun, University of Missouri

Correlated or clustered failure time data often occur in medical studies among other fields and sometimes such data arise together with interval censoring. Furthermore, the failure time of interest may be related to the cluster size. A simple and common approach to the analysis of these data is to simplify or reduce interval-censored data to right-censored data as there does not seem to exist proper inference procedures that can be used for the direct analysis of the data. In this paper, two procedures are presented for regression analysis of clustered failure time data that allow both interval censoring and informative cluster size. Simulation studies are conducted to evaluate the presented approaches and they are applied to a lymphatic filariasis example.

email: xzmpc@mail.missouri.edu

LIKELIHOOD METHODS FOR BINARY RESPONSES OF PRESENT COMPONENTS IN A CLUSTER

Xiaoyun Li*, Florida State University
Dipankar Bandyopadhyay, Medical University of South Carolina
Stuart Lipsitz, Harvard Medical School
Debajyoti Sinha, Florida State University

In some biomedical studies involving clustered binary responses (say, disease status) the cluster sizes can vary because some components of the cluster can be absent. In this paper, we propose a novel random effects logistic regression framework where both the presence of a cluster component and the binary response of disease status for a present component are treated as responses of interest. For the ease of interpretation of regression effects, both the marginal probability of presence/absence of a component as well as the conditional probability of disease status of a present component, integrated over the cluster random effect, preserve the logistic regression forms. We present a maximum likelihood method of estimation implementable using standard statistical software. We compare our models and the physical interpretation of regession effects with corresponding marginal GEE-based methods. The methodology is illustrated via analyzing a study of the periodontal health status in a diabetic Gullah population in South Carolina. We also present a simulation study to assess the robustness of our procedure to missspecification of the random effect distribution and to compare finite sample performances of estimates with existing methods.

email: xli@stat.fsu.edu

88. MULTIVARIATE SURVIVAL

INCORPORATING SAMPLING BIAS IN ANALYZING BIVARIATE SURVIVAL DATA WITH INTERVAL SAMPLING AND APPLICATION TO HIV RESEARCH

Hong Zhu*, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University

In biomedical cohort studies, it is common to collect data with incidence of disease occurring within a calendar time interval. Bivariate or multivariate survival data arise frequently in these studies and are of interest when the ordered multiple failure events are considered as the major outcomes to identify the progression of a disease. This paper considers a sampling scheme, where the first failure event (i.e., HIV infection) is identified within a calender time interval, the time of the initiating event (i.e., birth) can be retrospectively confirmed, and the second failure event (i.e., death) is observed subject to right censoring. The focus is on incorporating sampling bias in analyzing this type of bivariate survival data with interval sampling. We develop statistical estimation methods for failure time distributions under stationary and semi-stationary conditions correcting sampling bias. And the dependency structure

of the bivariate survival data is studied by semiparametric copula models and through "two-stage" estimation procedures. The proposed methods are evaluated by simulations and illustrated by Rakai Human Immunodeficiency Virus (HIV) seroconversion data to study the disease progression of HIV for treatment-naive individuals.

email: hongzhu@jhsph.edu

ESTIMATION OF THE GAP TIME DISTRIBUTIONS FOR ORDERED MULTIVARIATE FAILURE TIME DATA: A SIEVE LIKELIHOOD APPROACH

Chia-Ning Wang*, University of Michigan
Bin Nan, University of Michigan
Roderick Little, University of Michigan
Harlow Sioban, University of Michigan

In many scientific investigations, a subject can experience multiple events or transit form one state to another state, and researchers are often interested in the gap time between two successive events or states. In our specific application, the time to a menopausal transition marker and the time form the marker to the final menstrual period (FMP) are two primary events of interest. One major statistical challenge in the analysis of gap times is the induced dependent censoring that is, when two gap times are correlated, the probability of the second gap being censored depends on the length of the first gap. Hence standard approaches, such as Kaplan-Meier estimator, cannot be applied to the gap time. Several nonparametric methods based on inverse weighting have been proposed for estimating the gap time distribution. In this study, we propose an estimation method by maximizing the sieve likelihood function. The likelihood function is represented by the hazard function of the first gap time and the conditional hazard function of the second gap time given the first gap time, both of which are estimated by the tensor product cubic spline. Simulations are performed to evaluate the proposed method in different scenarios.

email: cnwang@umich.edu

SEMIPARAMETRIC INFERENCE FOR SUCCESSIVE DURATIONS

Jing Qian*, Harvard University
Yijian Huang, Emory University

For many chronic diseases, a bi-state progressive process provides a useful model for the disease progression before reaching death. For example, after surgery, colon cancer patients progress through cancer-free and caner-recurrence states. Often, scientific interests lie in the successive durations. For the one-sample problem with incomplete follow-up data, recent investigations have focused on nonparametric inference. However, in many practical situations, the distribution of the second duration is nonparametrically nowhere identifiable. Furthermore, most existing approaches require a rather restrictive censoring mechanism and have difficulty in predicting the process with given history. To address these issues, we suggest a semiparametric model that postulates normal copula for the association between the two durations, while leaving the marginals unspecified. Motivated by the colon cancer data example, we allow our model to accommodate the situation where the second duration has a probability mass at zero. We propose an inference procedure for estimation. The finite sample performance of the proposed method is evaluated via the simulation studies and illustrated with the data from a colon cancer study.

email: jqian@hsph.harvard.edu

A JOINT MODELING OF CORRELATED RECURRENT AND TERMINAL EVENTS WITH MULTIVARIATE FRAILTY IN THE ANALYSIS OF DRIVING SAFETY DATA IN OLD DRIVERS WITH NEUROLOGICAL DISEASES

Dawei Liu*, University of Iowa
Ergun Uc, University of Iowa
Elizabeth Dastrup, University of Iowa
Aaron Porter, University of Iowa
Jeff Dawson, University of Iowa
Matt Rizzo, University of Iowa

Automobile driving has become an integral activity of daily life, yet the risk of unsafe driving increases with age and neurological diseases, such as Parkinson's disease, Alzheimer's disease, and stroke. With the progression of the disease, some old drivers may have to give up the privilege of driving, which may result in depression and social isolation and hence reduced quality of life. In a driving safety study on old drivers with neurological diseases, data are available on a subject's past unsafe driving records, such as crashes and citations, and on that subject's current driving status, that is whether the subject has ceased driving or not. One feature of the data is that citations and crashes may occur repeatedly over time but driving cessation can occur only once, and once a subject ceased driving, no unsafe outcomes could be observed. In this talk, we propose a semiparametric model with multivariate frailty to jointly model these outcomes by treating citations and crashes as recurrent events and driving cessation as terminal event. Maximum likelihood estimates of model parameters are obtained through Monte Carlo EM. The proposed method is illustrated in the analysis of driving safety data in old drivers with neurological diseases.

email: dawei-liu@uiowa.edu

EXPONENTIAL GAP-TIME ESTIMATION FOR CORRELATED RECURRENT EVENT DATA UNDER INFORMATIVE MONITORING

Akim Adekpedjou*, Missouri University of Science and Technology
Gideon Zamba, University of Iowa

Time to occurrence of an event in a recurrent event data could be affected by many factors--chiefs among which are the within-subject unobservable frailty and informative censoring. On one hand, the unobservable frailty induces a within-subject correlation among the inter-event times. On the other hand, the informative censoring controls the per subject accumulation of events. There is a rarity of analytical tool for estimating survivor parameters in the presence of correlated recurrent events under informative censoring; such as in instances where the survival time of a subject is censored because of deterioration of their physical condition or due to the accumulation of their events occurrences. In this talk, we approach the parameter estimation problem through a fully parametric baseline hazard model where recurrent events and censoring intensity are reconciled through the generalized Koziol-Green (KG) model. The method is developed for exponential inter-event times. We finally apply our method to a neural activity biomedical data set.

email: akima@mst.edu

ON SEVERAL TEST STATISTICS FOR PAIRED CENSORED DATA

Liang Li*, Cleveland Clinic

We studied several test statistics for comparing the two marginal survival functions of paired censored data. The null distribution of these test statistics was approximated by permutation. These tests do not require explicit modeling or estimation of the within-pair correlation structure, accommodate both paired data and singletons, and the computation is straightforward with most statistical software. Numerical studies showed that these tests have competitive size and power performance. One test statistic has higher power than previously published test statistics against the alternative hypothesis of crossing survival functions. We illustrated the use of these tests in the analysis of a propensity score matched data set.

email: linden.liang.li@gmail.com

89. GENOME-WIDE ASSOCIATION STUDIES

BAYESIAN VARIABLE SELECTION IN A GENETIC ASSOCIATION STUDY

Mengye Guo*, Dana-Farber Cancer Institute
Edward George, University of Pennsylvania
Nandita Mitra, University of Pennsylvania
Daniel Heitjan, University of Pennsylvania

We seek to identify a parsimonious model for predicting a binary outcome when there is a large pool of potential predictors, potentially including both main effects and interactions. The specific example involves predicting patient success in a smoking cessation program from a panel of 144 single-nucleotide polymorphisms (SNPs) located on 9 candidate genes. We adapt Stochastic Search Variable Selection (SSVS; George and McCulloch 1997; Wang and George 2007), an efficient Bayesian model search algorithm, to the logistic regression model. To deal with the hierarchy of main effects and interactions, we incorporate prior constraints (Chipman 1996) that eliminate unreasonable models and thereby reduce the set of possible models to consider. We apply our method to the smoking cessation data and identify some well-fitting models. Simulations suggests that in settings similar to our data example the method performs well in detecting the true model.

e-mail: mengye@jimmy.harvard.edu

SURVIVAL KERNEL MACHINE SNP-SET ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDIES

Xinyi Lin*, Harvard University
Tianxi Cai, Harvard University
Xihong Lin, Harvard University

Genome-wide association studies are increasingly used to identify typed SNPs associated with a disease. Such studies typically use a case-control design. When interest lies in assessing how genetic profiles affect the risk of developing a disease, the phenotype is often time to occurrence of a clinical event. A common approach in identifying SNPs predictive of future risk is to fit a Cox model to each SNP individually. However such a single-SNP approach suffers from low power due to partial linkage disequilibrium between the typed SNP and causal SNP. It also fails to account for the joint effects of multiple causal SNPs. When multiple SNPs relate to the phenotype simultaneously via a complex structure, such marginal analysis may not be effective. To overcome these difficulties, we first group typed SNPs into SNP-sets based on genomic features, and test for the overall joint effects of all SNPs in a SNP-set. Specifically, we employ a kernel machine Cox regression framework and apply an efficient score test to assess the overall effect of a SNP set on the survival outcome. This approach uses genetic information from all SNPs simultaneously and can also capture potentially non-linear effects of the SNPs. We show using simulations that our approach has better performance than the standard single SNP test when the typed SNPs are in linkage disequilibrium with each other and with the true causal SNP.

e-mail: xinyilin@fas.harvard.edu

A GENOME-WIDE ASSOCIATION APPROACH ON DETECTING CNVs FOR SNP GENOTYPING DATA

Yaji Xu*, University of Texas M.D. Anderson Cancer Center
Bo Peng, University of Texas M.D. Anderson Cancer Center
Christopher I. Amos, University of Texas M.D. Anderson Cancer Center

SNP genotyping arrays have been developed to characterize SNPs and DNA copy number variations (CNVs). Nonparametric and model-based statistical algorithms have been developed to detect CNVs from SNP genotyping data. Recent studies have shown the advantages of generalized genotyping approaches that involve both intensity value and genotype information in a Hidden Markov Model (HMM). However, these methods can not differentiate common copy number polymorphisms (CNPs) and rare CNVs effectively, and they are lack of power in detecting small CNVs because the probability changes caused by these CNVs are usually not enough to trigger an emission event. Association tests based on detected CNVs therefore lack power even if these CNVs are common. In this research, we propose a new genome-wide association approach to detect CNVs for case-control studies based on the probabilities, instead of emitted states, of being a particular hidden state given the data at each SNP. Our approach is a genome-wide algorithm on a population level. It is more powerful in studying the association between genetic CNVs and complex diseases because it is more sensitive and is able to capture copy number rare variants that are related to the disease.

e-mail: yajixu@mdanderson.org

KERNEL MACHINE METHODS FOR THE ANALYSIS OF LARGE SCALE GENETIC ASSOCIATION STUDIES

Michael C. Wu*, University of North Carolina-Chapel Hill

Advances in high-throughput genotyping have culminated in the development of large scale genetic association studies for identifying gene variants, epistatic effects, and gene-environment interactions that are related to a clinical phenotype. The high-dimensionality of the feature space, the limited availability of samples, and the complex interactions between genetic features impose a grand challenge in analyzing such studies. To overcome such difficulties, we introduce a flexible non-parametric framework for analyzing high-dimensional genetic data. Specifically, we consider the use of kernel machine regression and propose efficient procedures for modeling genetic data and subsequently identifying important gene variants and interactions. The advantages of our approach will be made evident via theoretical and empirical investigations as well as data applications.

e-mail: mwu@bios.unc.edu

A GENOME IMPRINTING TEST WITH APPLICATION TO WHOLE-GENOME SCANS OF INSULIN RESISTANCE AND GLUCOSE

Rui Feng*, University of Pennsylvania
Donna Arnett, University of Alabama-Birmingham

Imprinting is an epigenetic phenomenon where the same parental copies have unequal transcriptions and thus different contributions to a trait depending on their parent-of-origins. This mechanism has been found to affect a variety of human disorders including common ordinal traits such as cancer, diabetes, and bipolar disease. In a previous study, we developed a latent variable model and a computationally efficient score statistic to test the imprinting effect on ordinal traits while adjusting for non-genetic covariates. The test statistic was calculated based on two different tests of linkage - with and without separate paternal and maternal effects. In this work, we derived a simple yet more robust test statistic skipping the test without parental difference. We evaluated the type I errors and power of our test statistic through simulations under various scenarios. We applied our method to a dataset from Genetics of Lipid Lowering drugs and Diet Network (GOLDN) for genome scans for alcoholism and diabetes-related phenotypes. A paternal imprinting signal was detected around 20cM on chromosome 22 for glucose.

e-mail: ruifeng@upenn.edu

A CROSS-POPULATION COMPARISON SCORE TEST TO DETECT POSITIVE SELECTION IN GENOME-WIDE SCANS

Ming Zhong*, Texas A&M University
Kenneth Lange, University of California-Los Angeles
Ruzong Fan, Texas A&M University

In this project, we developed a cross-population comparison test statistic to detect chromosome regions in which there is no significant excess homozygosity in one population but homozygosity remains high in another population. As in our previous work, we treat extended stretches of homozygosity as a surrogate indicator of recent positive selection. Conditioned on existing linkage disequilibrium, we propose to test the haplotype version of Hardy-Weinberg equilibrium (HWE). Under the assumption of no significant excess homozygosity in a chromosome region, HWE is roughly true in one population; on the other hand, the HWE is hardly true in the region if homozygosity remains high in the other population. For each population, assume that a random sample of unrelated individuals are typed on a large number of single nucleotide polymorphisms (SNPs). A pooled-test statistic is constructed by comparing the measurements of homozygosity of the two samples around a core SNP. In the chromosome regions, in which one population is roughly true in HWE and the other is not, the pooled-test statistic leads significant results to detect the positive selection. We evaluated the test by type I error comparison and power evaluation. Then, we applied the test to HapMap Phase II data.

e-mail: rfan@stat.tamu.edu

OPTIMAL CHOICE OF LATENT ANCESTRY VARIABLES FOR ADJUSTMENT IN GENOME-WIDE ASSOCIATION STUDIES

Gina M. Peloso*, Boston University
L Adrienne Cupples, Boston University
Kathryn L. Lunetta, Boston University

The principal components (PCs) computed from genome-wide genetic markers can be used as covariates in regression models to adjust for population structure within a sample, but the optimal criteria for selecting PCs to include in the model are not known. Using simulation to create structured populations, we explore methods of selecting PCs as covariates for a case-control study under the 4 scenarios that the outcome is/isn't structured and the risk SNP is/isn't structured. For several PC selection schemes, we compare Type I error and power to detect an association between case status and the risk SNP adjusting for the PCs selected as covariates. Type I error is correct for all methods and scenarios. When both the phenotype and the risk SNP are structured, the power for all PC selection methods is similar. When the phenotype is not structured but the risk SNP is, adjusting for a fixed number of top PCs or the PCs associated with the SNP improves power. In contrast, when the phenotype is structured but the risk SNP is not, adjusting for PCs reduces power. Contrary to current practice, our findings suggest the optimal choice of covariate PCs in a genome-wide association study should be SNP-dependent.

e-mail: gpeloso@bu.edu

90. NEW METHODOLOGY FOR LINEAR MIXED MODEL FRAMEWORK

GOODNESS OF FIT TESTS IN LINEAR MIXED MODELS

Min Tang*, University of Maryland-College Park
Eric V. Slud, University of Maryland-College Park

Linear mixed models are widely used for regression analysis in many fields, including small area estimation in surveys and in genetic research, because of their ability to accommodate correlated data. We propose a class of omnibus chi-squared goodness of fit tests for linear mixed models. These tests are based on the discrepancy between observed values and the values expected under the model in question, where a data point falls into one of L mutually exclusive categories. The partition into categories is based on the covariate space. Our test statistic is a quadratic form in the observed minus expected differences with coefficients obtained by substitution of maximum likelihood estimators into a matrix function. We show that under some conditions, this test statistic has an asymptotic chi-square distribution. Two examples are used to illustrate the proposed test.

e-mail: mintang@math.umd.edu

REGULARIZED REML FOR ESTIMATION AND SELECTION OF FIXED AND RANDOM EFFECTS IN LINEAR MIXED-EFFECTS MODELS

Sijian Wang*, University of Wisconsin-Madison
Peter XK Song, University of Michigan
Ji Zhu, University of Michigan

The linear mixed effects model (LMM) is widely used in the analysis of clustered or longitudinal data. In the practice of LMM, inference on the structure of random effects component is of great importance not only to yield proper interpretation of subject-specific effects but also to draw valid statistical conclusions. This task of inference becomes significantly challenging when a large number of fixed effects and random effects are involved in the analysis. The difficulty of variable selection arises from the need of simultaneously regularizing both mean model and covariance structures, with possible parameter constraints between the two. In this paper, we propose a novel method of regularized restricted maximum likelihood to select fixed and random effects simultaneously in the LMM. The Cholesky decomposition is invoked to ensure the positive-definiteness of the selected covariance matrix of random effects, and selected random effects are invariant with respect to the ordering of predictors appearing in the model. We also investigate large sample properties for the proposed estimation, including the oracle property. Both simulation studies and data analysis are included for illustration.

e-mail: swang@biostat.wisc.edu

LIKELIHOOD REFORMULATION METHOD IN NON-NORMAL RANDOM EFFECTS MODELS

Lei Liu*, University of Virginia
Zhangsheng Yu, Indiana University School of Medicine

In this paper we propose a practical computational method to obtain the maximum likelihood estimates (MLE) for mixed models with non-normal random effects. By simply multiplying and dividing a standard normal density, we reformulate the likelihood conditional on the non-normal random effects to that conditional on the normal random effects. Gaussian quadrature technique, conveniently implemented in SAS Proc NLMIXED, can then be used to carry out the estimation process. Our method substantially reduces computational time, while yielding similar estimates to the probability integral transformation method (Nelson et al. 2006). Furthermore, our method can be applied to more general situations, e.g., finite mixture random effects, or correlated random effects from Clayton copula. Simulations and applications are presented to illustrate our method.

e-mail: liulei@virginia.edu

A GOODNESS-OF-FIT TEST FOR THE RANDOM-EFFECTS DISTRIBUTION IN MIXED MODELS

Roula Tsonaka*, Leiden University Medical Center, The Netherlands
Dimitris Rizopoulos, Erasmus University Medical Center, The Netherlands
Geert Verbeke, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium
Geert Molenberghs, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium

In mixed models misspecification of the random-effects distribution can seriously affect inference for the random-effects and possibly fixed-effects parameters. Evaluating the validity of such distributional assumptions has been traditionally based on the Empirical Bayes estimates of the random effects. However, such approaches are rather limited due to the shrinkage effect. In this work we consider an alternative approach and develop a formal testing procedure to check the validity of the assumed random-effects distribution. In particular, this test is based on the properties of the directional derivative of the marginal log-likelihood, as they have been formalized by Lindsay (The Annals of Statistics, 1983; 11:783 - 792). Appealing features of the proposed procedure are: (i) its computational simplicity, since it can be implemented with standard statistical software, (ii) its wide range of applications including linear, non-linear and generalized linear mixed models and high-dimensional random-effects structures, and (iii) the possibility to identify areas of misfit. The performance of our proposal is evaluated for various mixed models and illustrated using a real dataset.

e-mail: s.tsonaka@lumc.nl

NORMALIZATION AND ANALYSIS OF LONGITUDINAL QUANTITATIVE PCR DATA BY LINEAR MIXED MODELS

Xuelin Huang*, University of Texas, M.D. Anderson Cancer Center

An approach of quantitative PCR experiments is to keep amplifying (doubling) the expression levels of genes of interest and house keeping genes, and record the number of cycles it takes for each gene to reach a pre-specified level. Then the original gene expression levels can be calculated. The gene expression levels are measured in this way at different time points after treatments. Often the expression levels of the genes of research interest are normalized based on the assumption that the expression level of house keeping genes are relatively constant over time. However, due to experiment variations, the above normalization procedure may introduce substantial systematic bias for some samples. We will propose and demonstrate a method to detect and correct such bias and then complete the data analysis by using linear mixed models.

e-mail: xlhuang@mdanderson.org

ASYMPTOTIC EQUIVALENCE BETWEEN CROSS-VALIDATIONS AND AKAIKE INFORMATION CRITERIA IN MIXED-EFFECTS MODELS

Yixin Fang*, Georgia State University

For model selection in mixed effects models, Vaida and Blanchard (2005) demonstrated that the marginal Akaike information criterion is appropriate as to the questions regarding the population and the conditional Akaike information criterion is appropriate as to the questions regarding the particular clusters in the data. This presentation shows that the marginal Akaike information criterion is asymptotically equivalent to the leave-one-cluster-out cross-validation and the conditional Akaike information criterion is asymptotically equivalent to the leave-one-observation-out cross-validation.

e-mail: matyxf@langate.gsu.edu

VARIABLE SELECTION IN LINEAR MIXED MODELS FOR LONGITUDINAL DATA

Rain Cui*, Harvard University
Xihong Lin, Harvard University
Victor DeGruttola, Harvard University

We consider variable selection in linear mixed models for longitudinal continuous data. We propose the regularized log-likelihood for variable selection of fixed effects. Several penalty functions are considered, including the LASSO, adaptive LASSO, and SCAD penalties. We show that the maximized regularized likelihood estimator of the regression coefficients can be equivalently obtained by jointly maximizing the penalized likelihood of both random effects and fixed effects. This connection allows us to obtain the parameter estimators by iteratively applying the BLUP method to calculate random effect estimators and use the existing software for variable selection developed for independent data to calculate regression coefficient estimators. We propose a modified REML estimator to estimate variance components by accounting for variable selection of fixed effects. The finite sample performance of the proposed method is evaluated using simulations and illustrated by applying to a HIV codon data set.

e-mail: rcui@hsph.harvard.edu

91. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

ESTIMATION OF POPULATION ABUNDANCE USING A HIERARCHICAL DEPLETION MODEL

Thomas F. Bohrmann*, University of Florida
Mary C. Christman, University of Florida
Xiaobo Li, University of Florida

Depletion experiments (single and multiple-pass) are often used to estimate total abundance of some animal species in a region. Hierarchical models have been proposed which allow inference on both the abundance of the animals and their catchability, which is the proportion of animals caught in a single sweep of the depletion experiment. The abundance of animals at a given site of the experiment has been modeled as a Poisson random variable conditional on some site-specific parameter lambda, where lambda is also a random variable from its own distribution. Randomly chosen experimental sites may provide the most information about this process, but in the case of highly clustered animals, this may prove inefficient because multi-pass depletion experiments would be likely to occur in areas of no animals. Thus multi-pass depletion experiments are sometimes performed only at locations where abundance is known to be high. Using data from nonrandom sites to directly inform the distribution of the lambdas introduces a bias. Therefore, we propose a method for efficiently estimating the overall abundance of a clustered population when single and multi-sweep depletion experiments are performed over the area of interest. We apply our method to data on blue crabs (Callinectes sapidus) in the Chesapeake Bay.

e-mail: bohrmann@ufl.edu

ASSESSING THE EFFICACY OF SLOW SPEED ZONES IN FLORIDA'S WATERWAYS

Kenneth K. Lopiano*, University of Florida
Linda J. Young, University of Florida

Florida's waterways are home to the endangered manatee species, Trichechus manatus latirostris. Florida's Fish and Wildlife Conservation Commission has implemented slow speed zones in an effort to reduce propeller-related manatee fatalities. The efficacy of such speed zones has been questioned recently. We provide a review of the current evidence associated with manatee fatalities and slow speed zones. We also address ways to compile data from multiple sources in order to better understand the effect slow speed zones have on the Florida Manatee.

e-mail: klopiano@ufl.edu

A SINGLE STATE SUPER POPULATION CAPTURE-RECAPTURE MODEL AUGMENTED WITH INFORMATION ON POPULATION OF ORIGIN

Zhi Wen*, North Carolina State University
Kenneth Pollock, North Carolina State University
James Nichols, U.S.Geological Survey Patuxent Research Center
Peter Waser, Purdue University

Ecologists applying capture-recapture models to animal populations sometimes have access to addition information about individuals' populations of origin. We consider a single super population model without age structure, and split the entry probability into separate components due to births in situ and immigration. We show that it is possible to estimate these two probabilities separately. We first consider the case of perfect information about population of origin, where we can distinguish individuals born in situ from immigrants with certainty. Then we consider the more realistic case of imperfect information, where we use genetic or other information to assign probabilities to each individual's origin in situ or outside the population. We use a resampling approach to impute the perfect origination assignment data based on the imperfect assignment tests. The integration of data on population of origin with capture-recapture data allows us to determine the contributions of immigration and in situ reproduction to the growth of the population, an issue of importance to ecologists. Further, the augmentation of capture-recapture data with origination data should improve the precision of parameter estimates. We illustrate our new models with capture-recapture and genetic assignment test data from a population of banner-tailed kangaroo rats Dipodomys spectabilis in Arizona.

e-mail: zwen@ncsu.edu

MODIFICATION BY FRAILTY STATUS OF THE ASSOCIATION BETWEEN AIR POLLUTION AND LUNG FUNCTION IN OLDER ADULTS

Sandrah P. Eckel*, University of Southern California
Thomas A. Louis, Johns Hopkins University
Karen Bandeen-Roche, Johns Hopkins University
Paulo H. Chaves, Johns Hopkins University
Linda P. Fried, Columbia University
Helene Margolis, University of California-Davis

Older adults have been found to be particularly vulnerable to the health effects associated with air pollution. Age may act as an imperfect surrogate for health status, so we aimed to enhance our understanding of this susceptibility by investigating whether frailty (a measure of health status in older adults) modifies the effects of air pollution on lung function. We used longitudinal data on a cohort of older adults from the Cardiovascular Health Study (CHS) and monthly average ambient air pollution levels from the CHS Environmental Factors Ancillary Study, interpolated to participant residence locations. We applied models that examined sub-

acute and chronic air pollution effects by relating a time-varying individual-level exposure (O3 or PM10) to a time-varying health outcome (lung function as measured by FEV1 or FVC), with an emphasis on effect modification by frailty status history. For chronic effects, we used cumulative summaries of exposure (analogous to pack years smoked) and frailty status and found evidence of increased air pollution related decline in FVC amongst participants with a longer history of frailty.

e-mail: eckel@usc.edu

A NEW STOCHASTIC MODEL OF CARCINOGENESIS FOR INITIATION-PROMOTION BIOASSAY

Wai-Yuan Tan, University of Memphis
Xiaowei (Sherry) Yan*, University of Memphis

Based on a generalized two-stage model of carcinogenesis with progression, in this paper we propose a new stochastic model for the generation of papillomas and carcinomas in mouse bioassay via initiation-promotion experiments. In this model, papillomas are generated by clonal expansion from primary first-stage initiated cells whereas carcinomas are generated by clonal expansion from primary second-stage initiated cells. To account for the observation that the initiators can generate simultaneously papillomas and carcinomas, we propose a two-pathways model involving a generalized two-stage model with clonal expansion and a one-stage model with clonal expansion. These models are framed in terms of Markov process. Based on the observed data of number of animals with papillomas and/or carcinomas starting with a fixed number of animals and the observed data of average number of papillomas and/or carcinomas per mouse, we have developed a generalized Bayesain procedure to estimate the mutation rates and the proliferation rates in each pathway. The fitting of some observed data clearly indicates that this multiple pathway model not only fits observed data better than single pathway, but also sheds light on multiple mechanisms of carcinomas induced by carcinogen and quantifies the percentage of carcinomas developing from each pathway.

e-mail: xwyan2001@yahoo.com

ESTIMATING THE ACUTE HEALTH EFFECTS OF COARSE PARTICULATE MATTER ACCOUNTING FOR EXPOSURE MEASUREMENT ERROR

Howard Chang*, Statistical and Applied Mathematical Sciences Institute
Roger D. Peng, Johns Hopkins University
Francesca Dominici, Harvard University

In air pollution epidemiology there is a growing interest in estimating the health effects of coarse particulate matter (PM) with aerodynamic diameter between 2.5 and 10 um. Coarse PM concentrations can exhibit considerable spatial heterogeneity because the particles travel shorter distances and do not remain suspended in the atmosphere for an extended period of time. We develop a modelling approach for estimating the short-term effects of air pollution in time series analysis when the ambient concentrations vary spatially within the study region. Specifically, our approach quantifies the error in the exposure variable by characterizing, on any given day, the disagreement in ambient concentrations measured across monitoring stations. This is accomplished by viewing monitor-level measurements as error-prone repeated measurements of the unobserved true exposure. Inference is carried out in a Bayesian framework to fully account for uncertainty in the estimation of model parameters. Finally, by using different exposure indicators, we investigate the sensitivity of the association between coarse PM and daily hospital admissions based on a recent national multi-site time series analysis. Among Medicare enrollees from 59 U.S. counties between the period 1999 to 2005, we find a consistent positive association between coarse PM and same-day admission for cardiovascular diseases.

e-mail: hhchang@samsi.info

RANK TESTS FOR SELECTIVE PREDATION

Yankun Gong*, Auburn University
Shuxin Yin, Auburn University
Asheber Abebe, Auburn University

This presentation considers nonparametric tests for selective predation. Of particular interest is how prey feature guides the prey selection pattern of predators. General rank tests are given for the case of one predatory species and prey characterized by a binary feature of interest and the case of two predatory species and prey characterized by either a continuous or a categorical feature of interest. The tests are designed to detect simply ordered alternatives because the score functions used to construct the test statistics are monotone. The results based on the asymptotic Gaussian distribution of the test statistics show that the tests retain nominal Type-I error rates. The results also show that the power of the asymptotic test depends on the choice of score function.

e-mail: gongyan@auburn.edu

92. VARIABLE SELECTION AND PENALIZED REGRESSION MODELS

A PERTURBATION METHOD FOR INFERENCE ON ADAPTIVE LASSO REGRESSION ESTIMATES

Jessica Minnier*, Harvard University
Tianxi Cai, Harvard University

Analysis of massive 'omics' data often seeks to identify a subset of genes or proteins that are associated with disease outcomes. Traditional statistical methods for variable selection fail for such high-dimensional data cases. Classification algorithms based on

gene expression levels and protein marker concentrations have been developed for prediction of clinical outcomes. However, many such algorithms are based on heuristics and often yield classification rules that have limited performance in validation studies. Robust regularization methods can achieve an optimal trade-off between the complexity of the model by simultaneously performing variable selection and estimation. Adaptive LASSO, in particular, gives consistent and asymptotically normal estimates. However, in finite samples, it remains difficult to construct an estimate of the covariance matrix of the parameter estimates that gives accurate inference about the regression parameters. We propose a perturbation method to approximate the distribution of the adaptive LASSO parameter estimates, which provides a simple way to estimate the covariance matrix and confidence regions. Through finite sample simulations, we verify the ability of this method to give accurate inference and compare it to other standard methods. We also illustrate our proposals with a study relating HIV drug resistance to genetic mutations.

e-mail: jminnier@hsph.harvard.edu

BOOTSTRAP INCONSISTENCY AND AN ORACLE BOOTSTRAP

Mihai C. Giurcanu*, University of Louisiana-Lafayette
Brett D. Presnell, University of Florida

In many applications, the bootstrap is consistent on all but a small subset of the underlying parameter space. We examine several such cases, involving estimators such as the Hodges, the Stein, and the LASSO estimator, whose limiting distributions are discontinuous as a function of the underlying parameter. We develop a straightforward approach for determining the precise limiting behavior of the nonparametric bootstrap in these problems. We also show that the bootstrap can be repaired by coupling the intentionally-biased bootstrap of Hall and Presnell (1999) with an estimator having an appropriate oracle property. Simulation results examining the performance of the resulting bootstrap are provided.

e-mail: giurcanu@louisiana.edu

A LASSO-TYPE APPROACH FOR ESTIMATION AND VARIABLE SELECTION IN SINGLE INDEX MODELS

Peng Zeng*, Auburn University
Tianhong He, Purdue University
Yu Zhu, Purdue University

The single index model is a natural extension of the linear regression model for applications where linearity between the response variable and the predictor variables may not hold. In this talk, we propose a penalized local linear smoothing method called sim-lasso for estimation and variable selection in the single index model. The sim-lasso method incorporates an L1 penalty of the derivative of the link function into the loss function of local linear smoothing and can be considered as an extension of the usual lasso to the single index model. We develop several algorithms to calculate the sim-lasso estimates and solution paths. The properties of the solution paths are investigated. In simulation study and a real data application, sim-lasso demonstrates excellent performance.

e-mail: zengpen@auburn.edu

COORDINATE DESCENT ALGORITHMS FOR NONCONVEX PENALIZED REGRESSION METHODS

Patrick Breheny*, University of Kentucky
Jian Huang, University of Iowa

A number of variable selection methods have been proposed involving nonconvex penalty functions. These methods, which include SCAD and MCP, have been demonstrated to have attractive theoretical properties, but model fitting is not a straightforward task, and the resulting solutions may be unstable. Here, we demonstrate the potential of coordinate descent algorithms for fitting these models, establishing theoretical convergence properties and demonstrating that they are significantly faster than competing approaches. In addition, we demonstrate the utility of convexity diagnostics to determine regions of the parameter space in which the objective function is locally convex, even though the penalty is not.

e-mail: patrick.breheny@uky.edu

VARIABLE SELECTION FOR PANEL COUNT DATA VIA NONCONCAVE PENALIZED ESTIMATING FUNCTION

Xingwei Tong, Beijing Normal University
Xin He*, University of Maryland-College Park
Liuquan Sun, Chinese Academy of Sciences
Jianguo Sun, University of Missouri-Columbia

Variable selection is an important issue in all regression analyses, and in this talk we discuss this in the context of regression analysis of panel count data. Panel count data often occur in long-term studies that concern occurrence rate of a recurrent event, and their analysis has recently attracted a great deal of attention. However, there does not seem to exist any established approach for variable selection with respect to panel count data. For the problem, we adopt the idea behind the non-concave penalized likelihood approach and develop a non-concave penalized estimating function approach. The proposed methodology selects variables and estimates regression coefficients simultaneously, and an algorithm is presented for this process. We show that the proposed procedure performs as well as the oracle procedure in that it yields the estimates as if the correct submodel were known. Simulation studies are conducted for assessing the performance of the proposed approach and suggest

that it works well for practical situations. An illustrative example from a cancer study is provided.

e-mail: xinhe@umd.edu

VARIABLE SELECTION WITH THE SEAMLESS-L0 PENALTY

Lee Dicker*, Harvard University
Baosheng Huang, Harvard University
Xihong Lin, Harvard University

We propose the seamless-L0 (SELO) penalty for penalized likelihood variable selection methods. The SELO penalty function is symmetric about 0 and non-differentiable at the origin, yet it is smooth, increasing and concave on the positive real numbers. The penalized likelihood procedure with SELO penalty is shown to have the oracle property of Fan and Li (2001). Tuning parameter selection is crucial to the performance of the SELO procedure. Tuning parameter selection procedures which do not require the use of testing data are of particular interest. We propose a BIC-like tuning parameter selection method for SELO and show that it consistently identifies the true model. The SELO method is efficiently implemented using a coordinate descent algorithm. Simulation results and a real data example show that the SELO procedure with BIC tuning parameter selection performs very well, even when the sample size is relatively small.

e-mail: ldicker@hsph.harvard.edu

EEBOOST: A GENERAL FRAMEWORK FOR HIGH-DIMENSIONAL VARIABLE SELECTION BASED ON ESTIMATING EQUATIONS

Julian Wolfson*, University of Minnesota

Most variable selection procedures for handling 'p > n' problems assume that the data follow one of a small class of simple models, typically ignoring any unusual features of the data (eg. correlation, missingness). This rather barren toolbox contrasts sharply with the wide variety of methods available for low-dimensional estimation in these more complex problem setups, methods typically based on solving a set of estimating equations. Here, we describe EEBoost, a procedure for variable selection in high-dimensional problems where low-dimensional estimation would typically be performed by solving a set of estimating equations. We present theoretical results which provide some intuition as to why EEBoost may be expected to outperform more naive variable selection approaches in certain situations, and show the close correspondence between EEBoost and a particular member of the class of L1 penalized methods. We illustrate the use of EEBoost in several simulated scenarios, and apply it to immunological data from the Step HIV vaccine trial.

e-mail: julianw@uw.edu

93. STATISTICAL ANALYSIS OF BRAIN IMAGING DATA

DETERMINING DIFFERENCES IN RESTING-STATE BRAIN CONNECTIVITY BETWEEN PATIENTS WITH DEPRESSION AND HEALTHY CONTROLS: A COMBINED fMRI/DTI ANALYSIS

DuBois Bowman*, Emory University
Gordana Derado, Emory University
Shuo Chen, Emory University

There is strong interest in investigating functional connectivity (FC) of the human brain, which involves a search for correlations between fMRI measures of brain activity from spatially distinct regions. Many such associations between regional brain activity measures are mediated by structural connections along white matter fiber tracts, providing an opportunity to incorporate DTI information into statistical analyses. We develop a novel statistical method for determining FC, called anatomically-weighted FC (awFC), which combines functional information from fMRI and anatomical information from DTI. We demonstrate that our multimodal approach achieves superior accuracy to FC analyses based solely on fMRI data. We also present an inference framework for comparing awFC results between subgroups of subjects. We apply our methodology to resting-state fMRI and DTI data to assess differential patterns of connectivity between patients with major depressive disorder and healthy control subjects.

e-mail: dbowma3@sph.emory.edu

STATISTICAL METHODS FOR EVALUATING CONNECTIVITY IN THE HUMAN BRAIN

Brian S. Caffo*, Johns Hopkins University
Cirpian M. Crainiceanu, Johns Hopkins University

In this talk we discuss statistical methods for the study of human brain connectivity. We will overview different forms of connectivity and potential modalities for measurement. We propose methods for analyzing connectivity measures from DTI tractography and methods for summarizing connectivity from functional neuroimaging. In the latter, we focus in particular on general inter-subject eigenvalue decompositions. The methods will be applied to a study of subjects at-risk for Alzheimer's disease and matched controls as well as a study of multiple sclerosis.

e-mail: bcaffo@jhsph.edu

FRATS: FUNCTIONAL REGRESSION ANALYSIS OF DTI TRACT STATISTICS

Hongtu Zhu*, University of North Carolina-Chapel Hill
Martin G. Styner, University of North Carolina-Chapel Hill
Weili Lin, University of North Carolina-Chapel Hill
Zhexing Liu, University of North Carolina-Chapel Hill
Niansheng Tang, Yunnan University
John H. Gilmore, University of North Carolina-Chapel Hill

This paper presents a functional regression framework, called FRATS, for the analysis of multiple diffusion properties along fiber bundle as functions in an infinite dimensional space and their association with a set of covariates of interest, such as age, diagnostic status and gender, in real applications. The functional regression framework consists of four integrated components: the local polynomial kernel method for smoothing multiple diffusion properties along individual fiber bundles, a functional linear model for characterizing the association between fiber bundle diffusion properties and a set of covariates, a global test statistic for testing hypotheses of interest, and a resampling method for approximating the p-value of the global test statistic. The proposed methodology is applied to characterizing the development of five diffusion properties including fractional anisotropy, mean diffusivity, and the three eigenvalues of diffusion tensor along the splenium of the corpus callosum tract and the right internal capsule tract in a clinical study of neurodevelopment. Significant age and gestational age effects on the five diffusion properties were found in both tracts.

e-mail: hzhu@bios.unc.edu

OVER-CONNECTIVITY OF 3D BRAIN NETWORK USING DIFFUSION TENSOR IMAGING

Moo K. Chung*, University of Wisconsin-Madison

Diffusion tensor imaging offers a unique opportunity to characterize the structural connectivity of the human brain non-invasively by tracing white matter fiber tracts. Whole brain tractography studies routinely generate up to half million tracts per brain, which serves as edges in an extremely large 3D graph with up to half million edges. Currently there is no agreed-upon method for constructing the brain structural network graphs out of large number of white matter tracts. In this paper, we present a scalable iterative framework for building a large graph and apply it to testing for under- and over-connectivity hypothesis in the autistic brain. Our proposed method can localize the abnormal regions in the brain (as identified as nodes in the graph) that are responsible for connectivity difference in autism.

e-mail: mkchung@wisc.edu

94. REGRESSION MODELS WITH COMPLEX COVARIATE INPUTS

REGRESSION ANALYSIS OF GRAPH-STRUCTURED DATA WITH GENOMIC APPLICATIONS

Hongzhe Li*, University of Pennsylvania

In many genomics applications, the genomic data are often supplemented by addition information in the form of a graph. Examples include genes observed on the protein-protein interaction networks, SNPs linked on the weighted linkage disequilibrium graphs and bacteria sequences linked on the phylogenetic trees. The graph structured genomic data induce some covariances among the genomic data observed. In this paper, we present and compare two different approaches to incorporate the graph information into regression analysis, including both a graph-constrained regularization estimation and a kernel-based regression appraoches using the similarity data. We will illustrate these methods using networkd data in gene expression analysis and pyrosequence data from gut microbiome studies.

e-mail: hongzhe@upenn.edu

COVARIATE ADJUSTED ASSOCIATION TESTS FOR ORDINAL TRAITS

Wensheng Zhu, Yale University
Yuan Jiang, Yale University
Heping Zhang*, Yale University

Identifing the risk factors for comorbidity is important in psychiatric research. Empirically, studies have shown that testing multiple, correlated traits simultaneously is more powerful than testing a single trait at a time in association analysis. Furthermore, for complex diseases, especially mental illnesses and behavioral disorders, the traits are often recorded in ordinal scales. In absence of covariates, nonparametric association tests has been developed for multiple (ordinal and/or quantitative) traits to study comorbidity. However, genetic studies generally contain measurements of some covariates that may confound the relationship between the risk factors of major interest (such as genes) and the outcomes. While it is relatively straightforward to include the covariates in the analysis of multipe quantitative traits, it is challenging for multiple ordinal traits. In this article, we propose a weighted test statistic based on a generalized Kendall's tau to adjust for the effects of the covariates. We conducted simulation studies to compare the type I error and power of our proposed test with an existing test. The empirical results suggest that our proposed test increases the power of testing association when adjusting for the covariates. We further demonstrate the advantage of our test by analyzing a real data set.

e-mail: heping.zhang@yale.edu

ADAPTIVE FUNCTIONAL LINEAR MIXED MODELS

Veera Baladandayuthapani*, University of Texas M.D. Anderson Cancer Center
Jeffrey S. Morris, University of Texas M.D. Anderson Cancer Center

We consider a regression setting with a scalar response and a functional covariate. Estimation and inference in such models presents a major challenge, since typically the dimension of the functional measurements far exceeds number of observations, and thus requires some form of regularization on the functional process. We propose an adaptive spline based regularization of the functional covariate that not only serves as a dimension reduction device but also, more importantly, accommodates a wide variety of behavior of the functional covariate, especially in cases where the functions may not be smooth. We also conduct pointwise FDR-based inference to identify which regions of the functional coefficient that are statistically and practically significant. We cast the problem in a Bayesian generalized functional linear mixed model framework and illustrate the method using a mass spectrometry proteomic data set.

e-mail: veera@mdanderson.org

FUNCTIONAL SINGLE INDEX MODELS FOR FUNCTIONAL COVARIATE

Ciren Jiang, University of California-Berkeley
Jane-Ling Wang*, University of California-Davis

Single index models are popular dimension reduction tools to model the nonparametric effect of multivariate covariates for univariate response variables. Although some scattered results exist for multivariate responses, few methods are for longitudinal response data. The goal of this paper is to extend the scope of single index models to longitudinal response data, possibly measured with errors, for both longitudinal and time-invariant covariates. The extension differs from current single index models in two ways: (i) it accommodates longitudinal data, both as response and as covariates; and (ii) the time-dynamic effects of the single-index is reflected in the model. In particular, we extend the approach and algorithm of MAVE to longitudinal data for the purpose of estimating the parametric index. With appropriate initial estimates of the parametric index, the proposed estimator is shown to be root-n consistent and asymptotically normally distributed. We also address the nonparametric estimation of regression functions and provide estimates with optimal convergence rates. One advantage of the new approach is that the same bandwidth is used to estimate both the nonparametric mean function and the parameters in the single-index. The finite sample performance of the proposed procedure is studied through simulations and AIDS CD4 cell count data.

e-mail: wang@wald.ucdavis.edu

95. METHODS FOR COMBINING MATCHED AND UNMATCHED CASE-CONTROL STUDIES

COMBINING MATCHED AND UNMATCHED CASE-CONTROL STUDIES USING STANDARD CONDITIONAL LOGISTIC REGRESSION SOFTWARE

Bryan Langholz*, University of Southern California

Methods are described for combining matched and unmatched case-control studies. We show how one can arrange data into an analytic data set and define appropriate models, such that likelihood inference can be performed using standard conditional logistic regression software. The methods were published in Huberman M, Langholz B American Journal of Epidemiology 1999;150(2):219-220 are easy to implement and allow for tests of homogeneity of the odds ratio across studies. We also show that the recursive algorithm for the unmatched case-control conditional logistic likelihood may also be used for analysis of combined data.

e-mail: langholz@usc.edu

COMBINING MATCHED AND UNMATCHED CONTROL GROUPS IN CASE-CONTROL STUDIES, USING SANDWICH OR BOOTSTRAPPING METHODS

Saskia le Cessie*, Leiden University Medical Center

We discuss an easy way to combine odds ratios of several case-control analyses with the same cases (1). The approach is based upon methods for meta-analysis, but takes into account that the same cases are used, and that the estimated odds ratios are therefore correlated. Two ways to estimate this correlation are discussed, sandwich methodology and the bootstrap. Confidence intervals for the pooled estimates and a test to check if the odds ratios in the separate case-control studies differ significantly are derived. We will demonstrate these methods on data from a large study on risk factors for thrombosis, the MEGA study, where cases of first venous thrombosis were included with a matched control group of partners and an unmatched population-based control group. Reference: Le Cessie S, Nagelkerke N, Rosendaal FR, van Stralen KJ, Pomp ER, van Houwelingen HC. Combining matched and unmatched control groups in case-control studies. Am J Epidemiol 2008;168:1204-10.

e-mail: cessie@lumc.nl

ON COMBINING RELATED AND UNRELATED CONTROLS

Nilanjan Chatterjee*, National Cancer Institute
Bhramar Mukherjee, University of Michigan

The selection of appropriate controls requires thoughtful considerations in genetic association mapping. Family-based designs using unaffected family members of cases as controls protect against the possibility of spurious association due to presence of population stratification. However, family-based controls are often difficult and expensive to recruit. On the other hand, using an additional sample of unrelated controls which may be available from another existing case-control study or a generic publicly available database of usable controls, potentially increases the efficiency of the family-based design but lack robustness under presence of population stratification. In this note, we propose a very simple and general tool to use information on unrelated controls to boost the efficiency (power) of a family-based design in a data adaptive way, without sacrificing much on the desired robustness properties in presence of population stratification. We consider two of the most popular family-based designs using case-sib pairs and case-parent triads to illustrate our methods. Simulation results indicate the newly proposed estimator is able to trade off between bias and efficiency depending on the sample size and the degree of population stratification present in the data.

e-mail: chattern@mail.nih.gov

A POLYTOMOUS CONDITIONAL LIKELIHOOD APPROACH FOR COMBINING MATCHED AND UNMATCHED CASE-CONTROL STUDIES

Mulugeta Gebregziabher*, Medical University of South Carolina
Paulo Guimaraes, University of South Carolina
Wendy Cozen, University of South Carolina
David Conti, University of South Carolina

In genetic association studies it is becoming increasingly imperative to have large sample sizes to identify and replicate genetic effects. To achieve these sample sizes, many research initiatives are encouraging the combination of several existing matched and unmatched case-control studies. Usually, a naive approach of fitting separate models for each case-control comparison is used to make inference about disease-exposure association. But, this approach does not make use of all the observed data and hence could lead to inconsistent results. The problem is compounded when a common case group is used in each case-control comparison. An alternative to fitting separate models is to use a polytomous logistic model but, this model does not combine matched and unmatched case-control data. Thus, we propose a polytomous logistic regression approach based on a latent group indicator and a conditional likelihood to do a combined analysis of matched and unmatched case-control data. We use simulation studies to evaluate the performance of the proposed method and a case-control study of multiple myeloma and Inter-Leukin-6 as an example. Our results indicate that the proposed method leads to a more efficient homogeneity test and a pooled estimate with smaller standard error.

e-mail: gebregz@musc.edu

96. HIGH DIMENSIONAL DATA ANALYSIS: DIMENSION REDUCTION AND VARIABLE SELECTION

GROUPWISE DIMENSION REDUCTION

Lexin Li*, North Carolina State University
Bing Li, Penn State University
Li-Xing Zhu, Hong Kong Baptist University

In many regression applications, the predictors fall naturally into a number of groups or domains, and it is often desirable to establish a domain-specific relation between the predictors and the response. In this article, we consider dimension reduction that incorporates such domain knowledge. The proposed method is based on the derivative of the conditional mean, where the differentiable operator is constrained to the form of a direct sum. This formulation also accommodates the situations where dimension reduction is focused only on part of the predictors; as such it extends Partial Dimension Reduction to cases where the blocked predictors are continuous. The proposed method is shown to achieve greater accuracy and interpretability than the dimension reduction methods that ignore group information.

e-mail: li@stat.ncsu.edu

NON-EUCLIDEAN DIMENSION REDUCTION VIA GRAPH EMBEDDING

Michael W. Trosset*, Indiana University
Minh Tang. Indiana University

Manifold learning techniques construct Euclidean representations of data that lie on manifolds in Euclidean feature spaces. More generally, they produce low-dimensional Euclidean representations of non-Euclidean proximity data. The Isomap algorithm does so by three explicit steps: (1) construct a local graph, (2) measure distance on the graph, and (3) embed the graph distances by classical multidimensional scaling. By varying the particulars of these steps, especially (2), one obtains alternative descriptions of other popular manifold learning techniques, e.g., Laplacian eigenmaps, diffusion maps, and Locally Linear Embedding. These descriptions reveal some interesting connections between these techniques and lead to simple examples of idiosyncratic behavior.

e-mail: mtrosset@indiana.edu

NON-LINEAR PENALIZED VARIABLE SELECTION

Gareth James*, University of Southern California
Peter Radchenko, University of Southern California

We propose a non-linear penalized variable selection approach. Our method not only allows for non-linear main effects but can also automatically select two way interactions. The penalty function we utilize is convex and can be efficiently solved using coordinate descent methods. Another advantage of our penalty function is that it automatically discourages interaction terms if the corresponding main effects are not already present but includes the main effects once the interaction term is selected. We show through simulation studies and real data sets that this approach works well in comparison to competing methods.

e-mail: gareth@usc.edu

BOOSTING FOR HIGH-DIMENSIONAL LINEAR MODELS WITH GROUPED VARIABLES

Lifeng Wang*, Michigan State University
Yihui Luan, Shandong University
Hongzhe Li, University of Pennsylvania

In regression analysis, variables can often be combined into groups based on prior knowledge. Such a group structure of the predictor variables must be effectively utilized in regression analysis in order to improve identification of relevant groups of variables and to improve the prediction performance. To address this issue, we propose a general boosting framework for high-dimensional functional additive models. Under this framework, we investigate the theoretical properties of a groupwise L2-Boosting method which can effectively account for the grouping structures. Its empirical performance will be demonstrated through both simulated and real world data.

e-mail: wang@stt.msu.edu

97. NONPARAMETRIC METHODS

SEMI-PARAMETRIC MEASUREMENT ERROR MODELING IN LOGISTIC REGRESSION

Jianjun Gan*, University of South Carolina
Hongmei Zhang, University of South Carolina

It is widely accepted that both environmental and genetic factors are related to the causation of NTDs (neural tube defects). Many Studies have been contributed to this area and demonstrated the effect of Folic Acid from daily supplement and the effect folate from food on NTD risk reduction. Because survey questionnaires are usually the only instrument used in these studies and responses to survey questions are likely to be biased (i.e. mis-measured). Parametric measurement error modeling is widely used to adjust for or study the impact of measurement errors. In this talk, we discuss a semi-parametric measurement error model based on P-spline with the help of a biomarker. The measurement error model is further incorporated into polytomous logistic regression models to infer interesting factor effects. The advantage of this method is its flexibility and no requirement to gold standard or replicates when adjusting for measurement errors. Simulations are used to demonstrate the methods and compare the proposed methods with other semi-parametric approaches. Finally, we apply this method to the NTD data to study the effect of folate intakes from food and prenatal multivitamins, in which red blood cell folates is used as a biomarker to adjust for measurement errors.

e-mail: ganj@mailbox.sc.edu

ESTIMATION OF THE PROBABILITY THAT TREATMENT IS BETTER THAN CONTROL IN CLINICAL TRIALS WITH CONTINUOUS OUTCOMES

Suporn Sukpraprut*, Boston University
Michael P. LaValley, Boston University

The performance of parametric and non-parametric estimators and their confidence intervals for the probability of treatment better than control with continuous outcomes are investigated. We considered a range of values for the probability of treatment better than control from 0.05 to 0.95 in normal and non-normal data distributions using simulation methods. Non-normal distributions considered include the uniform, contaminated normal and the log-normal. Bias and mean-squared error for the parametric estimator are minimal for the symmetric data distributions, but may become large when the distribution is skewed. Confidence interval coverage for the parametric estimators degraded quickly in the presence of skewness as the probability of treatment greater than control deviated from 0.5. The non-parametric Wilcoxon (or Mann-Whitney) estimator with the Halperin-Gilbert-Lachin confidence interval performed best across the range of conditions considered.

e-mail: mlava@bu.edu

SEMIPARAMETRIC SPLINE REGRESSION FOR LONGITUDINAL/CLUSTERED DATA

Arnab Maity*, Harvard University
Xihong Lin, Harvard University
Raymond J. Carroll, Texas A&M University

We consider the problem of semiparametric spline regression for clustered/longitudinal data with continuous outcomes. We develop profile and backfitting estimators of the model components, investigate their asymptotic properties and derive their asymptotic distribution. It is shown that the profile-spline and backfitting-spline estimators are asymptotically equivalent to their corresponding kernel counterparts when one uses Silverman's kernel. We perform a simulation study to observe the performance of our estimators for different covariance structures. We demonstrate our method by applying it to the data on the time evolution of CD4 cell numbers in HIV seroconverters arising from the Multicenter AIDS Cohort Study (MACS).

e-mail: amaity@hsph.harvard.edu

DIMENSION REDUCTION FOR THE CONDITIONAL KTH MOMENT VIA CENTRAL SOLUTION SPACE

Yuexiao Dong*, Temple University
Zhou Yu, East China Normal University

Various sufficient dimension reduction methods have been proposed to find linear combinations of predictor X, which contain all the regression information of Y versus X. If we are only interested in the partial information contained in the mean function or the kth moment function of Y given X, estimation of the central mean space (CMS) or the central kth moment space (CKMS) becomes our focus. However, existing estimators for CMS and CKMS require a linearity assumption on the predictor distribution. In this paper, we relax this stringent limitation via the notion of central solution space (CSS). Central kth moment solution space is introduced and its estimators are compared with existing methods by simulation.

e-mail: ydong@temple.edu

TESTING FOR CONSTANT NONPARAMETRIC EFFECTS IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH INTERACTIONS

Jiawei Wei*, Texas A&M University
Arnab Maity, Harvard University

We consider the problem of testing for constant nonparametric effect in a general semi-parametric regression model when there is the potential for interaction between the parametrically and non-parametrically modeled variables. The work was originally motivated by a unique testing problem in genetic epidemiology (Chatterjee, et al., 2006) that involved a typical generalized linear model but with an additional term reminiscent of the Tukey one-degree-of-freedom formulation. In this formulation, there are genetic variables, environmental variables, and demographic variables. The interest is in testing for main effects of the genetic variables, while gaining statistical power by allowing for a possible interaction between genes and the environment. Later work (Maity, et al., 2009) involved the possibility of modeling the environmental variable non-parametrically, but they focused on whether there was a parametric main effect for the genetic variables. In this paper, we consider the complementary problem, where the interest is in testing for the main effect of the non-parametrically modeled environmental variable. We derive a generalized likelihood ratio test for this hypothesis, show how to implement it, and give evidence that it can improve statistical power when compared to standard partially linear models. An empirical example involving colorectal adenoma is used to illustrate the methodology.

e-mail: wjw@stat.tamu.edu

NONPARAMETRIC ADDITIVITY TEST UNDER RANDOM DESIGN

Zhi He*, University of Illinois at Urbana-Champaign
Douglas G. Simpson, University of Illinois at Urbana-Champaign

Nonparametric additive model is a powerful approach for high-dimensional data and it is widely used in applied statistics. The advantage of additive model is they can achieve accurate nonparametric estimates while avoiding to some extend the curse of dimensionality. However, estimating the additive components and testing for additivity is more complex than in the classical nonparametric regression problem. In this paper, we propose a semiparametric test statistic based on nonparametric version of Tukey type additivity test. In particular, when the design density is independent, we simplify the test statistics to a more convenient version. The asymptotic consistency of our estimates is also established. Finally we illustrate the methodology with simulated data.

e-mail: zhihe2@illinois.edu

BALANCING THE OPTIMAL MATCH: A CLEVER SWAP

Shoshana R. Daniel, Covance, Inc.

The aim in matching is to associate, by creating strata, individuals who receive the treatment with individuals who receive the control that are similar for a specified set of covariates. In this way, differences between the control and treatment can be made at the strata level and then aggregated across stratum to determine whether the treatment is effective. However, where the overall distribution of covariate(s) differs in treatment versus control, the match that arises is not balanced. We introduce an algorithm, balance match, to modify the optimal match so as to ensure that there is a better balance of covariates, which ultimately guards against the possibility of bias in the estimate of the treatment effect, while maintaining close agreement of treated and control subjects within stratum. Balance match is easy to implement and is flexible. A simulation study shows that in realistic situations, one might sacrifice 5% diminution in the quality of the match and gain by 50% in the extent to which balance is achieved. The algorithm is illustrated in a SEER-Medicare endometrial cancer data set where implementation of the balance match algorithm creates better matches.

e-mail: shoshana.daniel@covance.com

98. NONPARAMETRIC AND SEMIPARAMETRIC SURVIVAL MODELS

SIEVE MAXIMUM LIKELIHOOD ESTIMATION USING B-SPLINES FOR THE ACCELERATED FAILURE TIME MODEL

Ying Ding*, University of Michigan
Bin Nan, University of Michigan

It is well known that the partial likelihood estimator in the Cox proportional hazards model is semiparametric efficient. Despite of previous research efforts, developing an efficient estimator in the semiparametric linear model is not completely satisfactory. In this paper, we propose a different approach to the existing semiparametric estimating equation methods that are known to be statistically inefficient. Specifically, we directly maximize the log likelihood function over a sieve space, in which the log hazard function is approximated by B-splines. The numerical implementation can be achieved through the conventional gradient-based search algorithms such as the Newton-Raphson algorithm. We show that the proposed estimators are consistent and asymptotically normal. Moreover, the limiting covariance matrix of the estimators reaches the semiparametric efficiency bound and can be estimated nicely by inverting either the information matrix based on the efficient score function of the regression parameters or the observed information matrix for all parameters including the "nuisance" parameters for estimating the log hazard function. Simulation studies demonstrate that the proposed method performs well in practical settings and yields more efficient estimates than existing estimating equation based methods. Illustrations with two real data examples are also provided.

e-mail: yingding@umich.edu

EFFICIENT COMPUTATION OF NONPARAMETRIC SURVIVAL FUNCTIONS VIA A HIERARCHICAL MIXTURE FORMULATION

Yong Wang*, University of Auckland, New Zealand
Stephen M. Taylor, Auckland University of Technology, New Zealand

We propose a new algorithm for computing the maximum likelihood estimate of a nonparametric survival function, which extends the constrained Newton method in a hierarchical fashion. By making use of the fact that a mixture distribution can be recursively written as a mixture of mixtures, it takes a divide and conquer approach to break down a large-scale optimization problem into many smaller-scale ones, which can thus be quickly solved. The new algorithm, which we call the hierarchical constrained Newton method, can efficiently reallocate the probability mass both locally and globally among potential support intervals. Results from simulation studies suggest that the new algorithm performs best in virtually all scenarios.

e-mail: yongwang@auckland.ac.nz

SCIENTIFIC IMPLICATIONS OF PARAMETRIC, SEMIPARAMETRIC, AND NONPARAMETRIC STATISTICAL MODELS

Scott S. Emerson*, University of Washington

Historically, many statistical analysis models were derived under a parametric probability model. Those that are most widely used (the general linear model) tend to also have very robust distribution-free interpretations. More recently, there has been much interest in semi-parametric probability models, particularly in the setting of time to event analyses that incorporate censored data. In this talk I focus on the ability of the various semi-parametric models to provide robust distribution-free interpretations. I will then discuss some nonparametric regression approaches that show promise in duplicating the robustness of the general linear model. An ultiimate goal is to describe methods for censored data that cover nearly the whole spectrum of methods available for complete data.

e-mail: semerson@uw.edu

TARGETED MAXIMUM LIKELIHOOD FOR TIME TO EVENT DATA

Ori M. Stitelman*, University of California-Berkeley
Mark J. van der Laan, University of California-Berkeley

Current methods used to analyze time to event data rely on highly parametric assumptions which result in biased estimates of parameters which are purely chosen out of convenience. By using Targeted Maximum Likelihood Estimation one may consistently estimate parameters which directly answer the statistical question of interest. The Targeted Maximum Likelihood Estimator, is a substitution estimator, which relies on estimating the underlying distribution. However, unlike any other substitution estimator, the underlying distribution is estimated specifically to reduce bias in the estimate of the parameter of interest, or feature of the distribution which answers the relevant statistical question. The advantages of these methods will be displayed through the use of a simulation study and results will be presented for a data analysis examining HIV patient outcomes.

e-mail: ostitelman@berkeley.edu

PARTIALLY MONOTONE TENSOR SPLINE ESTIMATION OF JOINT DISTRIBUTION FUNCTION WITH BIVARIATE CURRENT STATUS DATA

Yuan Wu*, University of Iowa
Ying Zhang, University of Iowa

This article develops a tensor spline-based nonparametric sieve estimation method to estimate joint distribution function with bivariate current status data. Asymptotic properties including

consistency and convergence rate of the proposed estimation are derived, and its finite-sample performance is studied via simulation studies. The method is applied to an AIDS study for estimating the joint distribution of the time to CMV shedding and the time to MAC colonization.

e-mail: yuan-wu@uiowa.edu

A NON-PARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION APPROACH TO FRAILTY MODEL

Zhenzhen Xu*, University of Michigan
John D. Kalbfleisch, University of Michigan

Survival data are often clustered and frailty model have been widely used to adjust for population heterogeneity and intraclass correlation. There is much literature dealing with the identification and estimation of frailty models using parametric and semi-parametric approaches. In these, parametric models have been used for the frailty distribution or the baseline hazard or both. We consider a Cox model with a frailty for clustered data with both the frailty distribution and cumulative baseline hazard left nonparametric and propose an approach based on non-parametric maximum likelihood estimation. For implementation, a three-step iterative algorithm is developed. First, we use a fast converging algorithm like the Intra-simplex Direction Method (ISDM) of Lesperance and Kalbfleisch (1995) or the CNM algorithm of Wang (2007) to estimate the empirical frailty distribution; second, baseline hazard is estimated using a variation of Breslow's cumulative baseline hazard estimator; and third, the regression parameter is estimated given the current estimate of the frailty distribution and baseline hazard. A simulation study indicates that the approach performs well for practical situations.

e-mail: zzxu@umich.edu

A NOVEL SEMIPARAMETRIC METHOD FOR MODELING INTERVAL-CENSORED DATA

Seungbong Han*, University of Wisconsin-Madison
Adin-Cristian Andrei, University of Wisconsin-Madison
Kam-Wah Tsui, University of Wisconsin-Madison

Interval-censored (IC) data are frequently encountered in medical studies focusing on time-to-event analyses. For IC data, existing methods for modeling the survival function are computationally intensive and usually require assumptions that sometimes could be difficult to verify in practice. A major obstacle in censored data modeling is that the event time of interest is incompletely observed, IC in this case. We propose a novel, flexible, computationally efficient and easily implementable modeling strategy based on the jackknife pseudo-observations (POs). The POs obtained using nonparametric methods are used as substitute outcomes in regression models that are asymptotically equivalent to the original ones. This way, the IC data modeling problem is translated into the realm of generalized linear models, where inferential and testing options are numerous. Outcome transformations via appropriately chosen link functions lead to familiar modeling contexts such as the proportional hazards, proportional odds or accelerated failure time models. Simulations studies show that the proposed method produces virtually unbiased covariate effect estimates and an example from the International Breast Cancer Study Group Trial VI further illustrates the practical advantages of this new approach.

e-mail: hanseung@stat.wisc.edu

99. RATER AGREEMENT AND SCREENING TESTS

ASSESSING THE "BROAD SENSE AGREEMENT" BETWEEN ORDINAL AND CONTINUOUS MEASUREMENTS

Limin Peng*, Emory University
Ruosha Li, Emory University
Ying Guo, Emory University
Amita Manatunga, Emory University

Conventional agreement studies have been focused on assessing the exchangeability of different instruments, and thus require measurements produced by these instruments to be on the same scale. To accommodate comparison of instruments in different scales as needed in many practical situations, we propose a broad framework of agreement study which addresses the replaceability of instruments instead of their exchangeability, thereby relaxing the identical scale constraint in assessing agreement. In this work, we investigate the case with one continuous measurement and one ordinal measurement. A broad sense agreement measure is developed to define the extent to which a continuous scale (eg. Carrol-D depression score) can be interpreted in an ordinal scale (eg. graded severity of depression). We propose a nonparametric estimator of the new measure, and establish its consistency and asymptotic normality. A well-justified bootstrap procedure is developed to perform inferences. Finally, we extend our proposal to longitudinal settings which involve agreement assessments at multiple time points. A test on the time trend of the proposed broad sense agreement is also developed. Simulation studies have demonstrated good performance of the proposed methods with small and moderate sample sizes. We illustrate our methods via applications to two recent depression studies.

e-mail: lpeng@sph.emory.edu

ESTIMATING THE AGREEMENT AND DIAGNOSTIC ACCURACY OF TWO DIAGNOSTIC TESTS WHEN ONE TEST IS CONDUCTED ON ONLY A SUBSAMPLE OF SPECIMENS

Hormuzd A. Katki*, National Cancer Institute
Yan Li, University of Texas at Arlington
Philip E. Castle, National Cancer Institute

A major research goal is the efficient usage of large specimen repositories for the evaluation of new diagnostic tests and for comparing new tests with existing tests. When no test can be considered a gold standard, inference focuses on agreement statistics and tests of symmetry; otherwise, measures of diagnostic accuracy are computed. Typically, all pre-existing diagnostic tests will already have been conducted on all specimens. However, to minimize study costs and specimen consumption, we propose retesting only a judicious subsample of the specimens with the new diagnostic test. We introduce methods to estimate agreement statistics and conduct symmetry tests when one test is conducted on only a subsample. The methods use inverse-probability weighting (IPW) by treating the subsample as a stratified two-phase sample. The strata use information available on all specimens to retest only the most informative specimens. We also generalize previous IPW-based estimators of diagnostic accuracy, developed under the analogous situation of "verification bias", to handle stratified sampling. We demonstrate that adequate statistical efficiency can be achieved under subsampling while greatly reducing costs and the number of specimens requiring retesting. Naively using standard estimators that ignore subsampling can lead to drastically misleading estimates. R code is available upon request.

e-mail: hkatki@gmail.com

TESTING THE SIGNIFICANCE OF OVERLAPPING SETS IN A VENN DIAGRAM

Aixiang Jiang*, Vanderbilt University

Venn diagrams show all possible logical relationships between a finite collection of sets. Since John Venn introduced the Venn diagram in the 1880s, this tool has been used in many fields, including set theory, probability, logic, statistics, and computer science. Today, the use of the Venn diagram has been extended to display high-dimensional data analysis results. The Venn diagram is a very useful tool to visually compare winner genes or copy number variation region sets from different cell lines, platforms, binding sites, or different experiments; however, no statistical method exists to test whether the number of overlapping sets is significant. Our current research applies a re-sampling procedure to solve this problem. We will show how our procedure works, and illustrate our method with both simulation datasets and real high-dimensional data sets.

e-mail: aixiang.jiang@vanderbilt.edu

ESTIMATION OF CUT-POINTS ON A CONTINUOUS SCALE ACCORDING TO A CATEGORICAL SCALE

Ming Wang*, Emory University
Amita Manatunga, Emory University
Ying Guo, Emory University
Limin Peng, Emory University

Measures of agreement have received increased attention in establishing the accuracy or validity of a new instrument compared to a standard instrument. In this paper, we focus on developing agreement-based methods to find optimal cut-points of a continuous scale according to a categorical scale, assuming that both scales measure the same biologic phenomena. We propose to use the weighted kappa coefficient between the discretized continuous scale and categorical scale as the objective function. The cut-points are estimated by optimizing the objective function. Other objective functions based on Kendall's tau and misclassification rate are considered and compared. We establish analytical advantages of the weighted kappa objective function and investigate its statistical properties through simulation studies with comparisons to other objective functions. This proposed method is illustrated with an application to a mental health study.

e-mail: wm_pku@hotmail.com

COMPARING THE CUMULATIVE FALSE-POSITIVE RISK OF SCREENING MAMMOGRAPHY PROGRAMS USING A DISCRETE TIME SURVIVAL MODEL ALLOWING FOR NON-IGNORABLE DROP-OUT

Rebecca A. Hubbard*, Group Health Research Institute
Diana L. Miglioretti, Group Health Research Institute

Mammography is the only screening modality shown to reduce breast cancer mortality among women 50 and older. To evaluate screening mammography programs we must understand both benefits and harms. The most prevalent harm of screening mammography is the risk of a false-positive recall. However, no existing research has compared the cumulative false-positive risk associated with different screening programs. Estimation in this context is complicated by the presence of a non-ignorable censoring mechanism. Drop-out may be related to factors that influence performance of the screening test, such as family history of breast cancer, or to the event itself. We propose a discrete time survival model for time to the first false-positive exam result, accounting for non-ignorable drop-out via a censoring bias function. We demonstrate that in this context, the censoring bias function is identifiable because the time of drop-out is known even for subjects who have previously experienced the event of interest. We develop a Bayesian estimation method for jointly estimating the censoring bias function and false-positive risk and use this model to compare

the cumulative false-positive risk of programs characterized by initiation at age 40 versus 50 and annual versus biennial screening.

e-mail: hubbard.r@ghc.org

LOGIC FOREST: AN ENSEMBLE CLASSIFIER FOR DISCOVERING LOGICAL COMBINATIONS OF BINARY MARKERS

Bethany J. Wolf*, Medical University of South Carolina
Elizabeth H. Slate, Medical University of South Carolina
Elizabeth G. Hill, Medical University of South Carolina

Adequate screening tools allowing physicians to diagnose diseases in asymptomatic individuals or identify individuals at risk of developing disease can reduce disease related mortality. Diagnostic tests based on multiple biomarkers may lead to enhanced sensitivity and specificity. Statistical methods that can model complex biologic interactions and that are easily interpretable allow for translation of biomarker research into diagnostic tools. Logic regression (Ruczinski et al., 2003), a relatively new multivariable regression method that predicts binary outcomes using logical combinations of binary predictors, can model the complex interactions in biologic systems in easily interpretable models. However the performance of logic regression degrades in noisy data. We implement an extension of logic regression methodology to an ensemble of logic trees (Logic Forest). We conduct several simulation studies comparing the ability of logic regression and Logic Forest to identify interactions among variables predictive of disease status. Our findings indicate Logic Forest is superior to logic regression for identifying important predictors. We also apply our method to SNP data to determine association between genetic and health factors with periodontal disease.

e-mail: wolfb@musc.edu

100. BAYESIAN METHODS: JOINT LONGITUDINAL/SURVIVAL MODELING AND DISEASE MODELING

BAYESIAN JOINT MODELS OF ZERO-INFLATED LONGITUDINAL PATIENT-REPORTED OUTCOMES AND PROGRESSION-FREE SURVIVAL TIMES IN MESOTHELIOMA

Laura A. Hatfield*, University of Minnesota
Mark E. Boye, Eli Lilly and Company
Michelle D. Hackshaw, Eli Lilly and Company
Bradley P. Carlin, University of Minnesota

Malignant pleural mesothelioma (MPM) is a deadly form of lung disease caused by asbestos exposure. Previous studies have established that pemetrexed (Alimta(R)) plus best supportive care is effective in prolonging progression-free survival (PFS) for previously treated patients when compared to best supportive care alone. In this paper, we seek to extend these findings by modeling longitudinal patient-reported outcomes (PROs) jointly with PFS. We build a hierarchical Bayesian model that combines zero-inflated beta distributions for the PROs with a proportional hazards Weibull model for the right-censored PFS values. Correlations among the probability of a non-zero PRO, its severity, and PFS is modeled using bivariate latent random effects. The results reveal significant effects of treatment, including decreased probability of symptom distress and decreased severity of symptom interference with daily life over time. We observe correlation between the individual probability of non-zero PRO and severity of non-zero PROs, and associations between these latent variables and PFS. We also describe novel multivariate extensions of our work to permit simultaneous modeling of PFS and all PROs at once. This paper contributes the understanding of treatment benefits of pemetrexed for second-line MPM, connecting survival improvements to improvements in patient-reported outcomes.

e-mail: hatfield@umn.edu

BAYESIAN SEMIPARAMETRIC MULTIVARIATE JOINT MODELS

Dimitris Rizopoulos*, Erasmus University Medical Center, The Netherlands
Pulak Ghosh, Novartis Pharmaceuticals

Motivated by a real data example on renal graft failure, we propose a new semiparametric multivariate joint model that relates multiple longitudinal outcomes to a time-to-event. To allow for greater flexibility, key components of the model are modelled nonparametrically. In particular, for the subject-specific longitudinal evolutions we use a spline-based approach, the baseline risk function is assumed piecewise constant, and the distribution of the latent terms is modelled using a Dirichlet Process prior formulation. Additionally, we discuss the choice of a suitable parameterization, from a practitioners point of view, to relate the longitudinal process to the survival outcome. Specifically, we present three main families of parameterizations, discuss theirs features, and present tools to choose between them.

e-mail: d.rizopoulos@erasmusmc.nl

SEMIPARAMETRIC BAYESIAN JOINT MODEL WITH VARIABLE SELECTION

Haikun Bao*, University of South Carolina
Bo Cai, University of South Carolina
Pulak Ghosh, Novartis Pharmaceuticals
Nicole Lazar, University of Georgia

In longitudinal studies, a popular model is the linear mixed model that includes fixed effects and subject specific random effects. An important aspect in this kind of studies is the occurrence of dropout due to frequent missing visits. If the missing data is non-ignorable,

this lead to informative dropout of the longitudinal data. Recently, method for jointly modeling longitudinal and survival data (for informative dropout) have gained popularity in the statistical literature. In this paper, we consider the problem of variable selection in a joint modeling framework where a longitudinal data and the informative dropout are modeled jointly. Dirichlet process priors are used to relax the parametric assumption of random effects, which has advantages of robustifying the model against possible misspecifications and nature of clustering of subjects. A fully Bayesian method for subset selection for fixed and random effects in joint models is proposed. Simulation examples and an application are used for evaluation and illustration.

e-mail: haikun.bao@gmail.com

A DYNAMIC PROJECTION MODEL OF THE BURDEN OF DIABETES IN THE U.S. ADULT POPULATION

James P. Boyle*, Centers for Disease Control
Theodore J. Thompson, Centers for Disease Control
Lawrence Barker, Centers for Disease Control

A three state dynamic model consisting of a system of three difference equations in time projecting the future burden of diabetes in the U.S. adult population is described. The states are no diabetes, undiagnosed diabetes, and diagnosed diabetes. Two principal data sources are U.S. Bureau of Census projections of the U.S. resident population characteristics and estimates and standard errors of historical incidence rates of diagnosed diabetes for the U.S. adult population. Census projections were used to constrain the model to produce the Census projections of the adult population. Incidence data produced a Bayesian model projecting future incidence rates. Additional inputs consisted of published estimates of key parameters. Posterior distributions of incidence rate projections determined Bayesian confidence intervals of all quantities of interest. The results suggest that the prevalence of any diabetes in the adult population will increase from 14.5 percent in 2010 with a 95% confidence interval of (14.4,14.7) to 32.6 percent in 2050 with a 95% confidence interval of (24.5,41.2). Also, incidence rates of any diabetes increased from 11.3 cases per thousand in 2010 with a 95% confidence interval of (10.2,12.2) to 17.0 cases per thousand in 2050 with a 95% confidence interval of (10.3,25.3).

e-mail: hzb0@cdc.gov

AN EMPIRICAL, INFORMED PRIOR FOR THE BETWEEN-STUDY HETEROGENEITY IN META-ANALYSES

Eleanor M. Pullenayegum*, McMaster University

It is well known that when a Bayesian meta-analysis includes a small number of studies, inference can be sensitive to the choice of prior for the between-study variance. Choosing a vague prior does not solve the problem, as inferences can be substantially different depending on the degree of vagueness. Moreover, because the data provide little information on between-study heterogeneity, posterior inferences for the between-study variance based on vague priors will tend to be unrealistic. It is thus preferable to adopt a reasonable, informed prior for the between-study variance. However, relatively little is known about what constitutes a realistic distribution. Based on data from the Cochrane Database of Systematic Reviews, this talk will describe the distribution of between-study variance in published meta-analyses, and propose some realistic informed priors for use in meta-analyses of binary outcomes. It is hoped that these priors will improve the calibration of inferences from Bayesian meta-analyses.

e-mail: pullena@mcmaster.ca

BAYESIAN HIERARCHICAL MODELING OF HOST GENETIC CORRELATES OF IMMUNE RESPONSE TO ANTHRAX VACCINE

Nicholas M. Pajewski*, University of Alabama-Birmingham
Purushottam W. Laud, Medical College of Wisconsin
Scott D. Parker, University of Alabama-Birmingham
Robert P. Kimberly, University of Alabama-Birmingham
Richard A. Kaslow, University of Alabama-Birmingham

Weaponized spores from Bacillus anthracis have been used as a lethal agent of bioterrorism. Prevention of Anthrax depends upon the efficacy of the licensed anthrax vaccine (Anthrax Vaccine Adsorbed, AVA), with antibody levels to B. anthracis protective antigen (AbPA) closely predicting survival following lethal spore challenge in animal models. Studies such as Pittman et al. (2002) have demonstrated sizeable inter-individual variation in duration of AbPA titers, suggesting a potential genetic influence on immune response. Here we investigate host genetic correlates of immune response to AVA within an ethnically diverse study population collected as part of a clinical trial investigating the immunogenicity and reactogenicity of a reduced dose schedule and intramuscular injection of AVA. Modeling of the immune response requires accounting for the different regimens used as part of the trial, longitudinal measurements taken over a 43-month follow-up period, left-censoring due to assay lower limits of quantification, and potential heterogeneity from the innate and adaptive immune response. We propose a Bayesian hierarchical framework for characterizing the longitudinal profile of response to AVA accounting for these factors, including shrinkage mechanisms to simultaneous incorporate a large number of genetic effects due to single nucleotide polymorphisms.

e-mail: npajewski@ms.soph.uab.edu

RELATIVE BREADTH OF MOSAIC AND CON-S HIV-1 VACCINE DESIGN STRATEGIES

Sydeaka P. Watson*, Baylor University and Los Alamos National Laboratory
Bette T. Korber, Los Alamos National Laboratory
Mark R. Muldoon, University of Manchester
John W. Seaman, Baylor University
James Stamey, Baylor University

Genetic diversity is a challenge that the scientific community must overcome before the development of a global HIV-1 vaccine is realized. Two vaccine strategies addressing genetic diversity, namely HIV-1 global consensus envelope sequence (CON-S) and polyvalent vaccine antigens (Mosaic), have been investigated. The consensus vaccine strategy aligns available HIV-1 gene sequences and selects the most prevalent amino acid at each position. Mosaic proteins are assembled via a computational method using fragments of HIV-1 protein sequences. The mosaic cocktails are optimized to promote maximal coverage of T-cell epitopes for a given population of viral strains. Preliminary studies of these two vaccines yield promising results; each has been shown to increase the number of positive immune responses in vaccinated monkeys after challenge. We investigate the relative breadth of of the CON-S and Mosaic vaccine immune responses in two related animal studies with a generalized linear model including mixed effects for Poisson counts. We discuss this approach and compare the conclusions to those resulting from a complimentary Bayesian analysis.

e-mail: sydeaka_watson@baylor.edu

101. INTEGRATION OF INFORMATION ACROSS MULTIPLE STUDIES OR MULTIPLE -OMICS PLATFORMS

INTEGRATING DIVERSE GENOMIC DATA USING GENE SETS

Svitlana Tyekucheva*, Dana-Farber Cancer Institute
Rachel Karchin, Johns Hopkins University
Giovanni Parmigiani, Dana-Farber Cancer Institute

Gene set analysis (GSA) considers whether genes that form a set from a specific biological standpoint, also behave in a related way in experimental data. It is commonly used in high throughput genomic experiments to help with interpretability of results. In recent cancer genome projects it has also been used to integrate information provided by multiple genome-wide assays. The diversity of genomic data collected in current cancer research is increasing and calls for scalable statistical methods that allow integrating across data types. GSA provides on such solution, though a rigorous framework is still lacking. We introduce, compare, and systematically evaluate two set-based data integration approaches: computing integrated gene-to-phenotype association score followed by conventional GSA, and using a consensus significance score after all data types were analyzed individually. We use integrated analysis tools to jointly examine gene expression and copy number variation data about glioblastoma multiforme tumor samples, from the TCGA. We show that using integration techniques allows discovering gene sets associated with the differences in the phenotype that would not be discovered when each data type is analyzed individually. In the glioblastoma analysis, when we consider differences survival times, these sets include the WNT, glycolysis, and stress pathways.

e-mail: svitlana@jimmy.harvard.edu

BAYESIAN JOINT MODELING OF MULTIPLE GENE NETWORKS AND DIVERSE GENOMIC DATA TO IDENTIFY TARGET GENES OF A TRANSCRIPTION FACTOR

Peng Wei*, University of Texas
Wei Pan, University of Minnesota

We consider integrative modeling of multiple gene networks and diverse genomic data including protein-DNA binding, gene expression and DNA sequence data to accurately identify the regulatory target genes of a transcription factor (TF). Rather than treating all the genes equally and independently a priori in existing joint modeling approaches, we incorporate the biological prior knowledge that neighboring genes on a gene network tend to be (or not to be) regulated by a TF together. To maximize the use of all existing biological knowledge, we allow the incorporation of multiple gene networks into joint modeling of genomic data by introducing two mixture models based on the use of Gaussian Markov random fields (GMRFs) and Discrete Markov random fields (DMRFs), respectively. Another contribution of our work is to allow different genomic data to be correlated and to examine the validity and effect of the independence assumption as adopted in existing methods. Due to a fully Bayesian approach, inference about model parameters can be carried out based on MCMC samples. Application to an E. coli dataset together with simulation studies demonstrates utility and statistical efficiency gains with the proposed joint models.

e-mail: Peng.Wei@uth.tmc.edu

A LATENT MIXTURE MODEL FOR ANALYZING MULTIPLE GENE EXPRESSION AND CHIP-ChIP DATA SETS

Hongkai Ji*, Johns Hopkins University

Gene expression and genome-wide chromatin immunoprecipitation data from multiple cellular contexts allow one to identify core targets of a transcription factor and characterize context-dependency of gene regulation. A statistical framework is developed to jointly analyze multiple gene expression and ChIP-chip data sets to identify direct and indirect target genes of a transcription factor and define their cell-type dependencies. Our approach involves

systematic modeling of complex correlation structures among multiple experiments. Simulations and real data analyses show that this approach significantly reduces false positive and false negative rates compared to analyzing individual datasets independently.

e-mail: hji@jhsph.edu

EFFECT OF COMBINING STATISTICAL TESTS AND FOLD-CHANGE CRITERIA

Doug Landsittel*, University of Pittsburgh
Nathan Donohue-Babiak, Duquesne University

A key concern for analysis of microarray data is to identify genes which are both statistically and biologically significant. To accomplish this, investigators often combine fold-change cut-offs with t-test p-values (after adjustment for multiple comparisons) to identify significant genes for further study. In an upcoming publication, we use the Benjamini-Hochberg adjustment to the t-test in conjunction with a k-fold change criterion (for k = 1.5, 2, and 3) to show that, under many scenarios, the fold-change cut-off dominates determination of significance and renders the statistical results irrelevant. Further, since the power of statistical tests increase with larger sample sizes, but deterministic criteria may have decreased power with larger sample sizes, the resulting power curve is often unpredictable and neither monotonically increasing nor decreasing. In this study, we address more general cases through further simulations and compare the approach of simply combining t-tests and fold change to newer alternative criteria, such as variable or probability fold change statistics, which incorporate statistical variation into judging the magnitude of the observed fold change. Results are critical to advising laboratory and clinical investigators on appropriate use of fold change criteria, and understanding the impact of related approaches in microarray analysis.

e-mail: landsitteldp@upmc.edu

A COMPARISON OF METHODS FOR INTEGRATED OMICS ANALYSIS

John Barnard*, Cleveland Clinic

Measuring multiple aspects of complex biological systems on the same samples in order to understand their interplay is becoming common. Examples of such aspects include messenger RNA expression, micro RNA expression, copy number variation, SNP polymorphism and DNA methylation. How to analyze and synthesize these multiple aspects is an active area of research. Proposed methods include sparse canonical correlation, Gaussian graphical models, latent variable models and regression-based methods. We compare and contrast some of the proposed methods using multiple real and simulated integrated omics datasets to assess their performance on a number of operating characteristics.

e-mail: barnarj@ccf.org

PATHWAY-DIRECTED WEIGHTED TESTING PROCEDURES FOR THE INTEGRATIVE ANALYSIS OF GENE EXPRESSION AND METABOLOMIC DATA

Laila M. Poisson*, University of Michigan
Debashis Ghosh, Penn State University

In the post-genomic era we have witnessed an explosion of high throughput data measuring global snapshots of the molecular behavior of cells. In cancer, the most common high-throughput assay has been the gene expression microarray. More recently, other omics technologies such as metabolomics are being considered for analysis as well. Metabolomics measure the complement of small molecules in a cell, on the order of 500-1000 molecules measured per experiment. Here we explore the utility of p-value weighting for enhancing the power to detect differential metabolites in a two-sample setting. Related gene expression information is mapped to a metabolite through metabolic pathways. The gene expression information is summarized using gene set enrichment tests. Through simulation we explore four styles of enrichment tests and five different weight functions. We comment on the utility of the different weights in various situations and make recommendations for improving the power of per-metabolite tests of differential intensity.

e-mail: lpoisson@umich.edu

AN EMPIRICAL BAYES APPROACH TO JOINT ANALYSIS OF MULTIPLE MICROARRAY GENE EXPRESSION STUDIES

Lingyan Ruan*, Georgia Institute of Technology
Ming Yuan, Georgia Institute of Technology

With the prevalence of gene expression studies and the relatively low reproducibility caused by insufficient sample sizes, it is natural to consider joint analyses that could combine data from different experiments effectively in order to achieve improved accuracy. In particular, we present in this paper a model-based approach for better identification of differentially expressed genes by incorporating data from different studies. The model can accommodate in a seamless fashion a wide range of studies including those performed at different platforms, and/or under different but overlapping biological conditions. Model-based inferences can be done in an empirical Bayes fashion. Because of the information sharing among studies, the joint analysis dramatically improves inferences based on individual analysis. Simulation studies and real data examples are presented to demonstrate the effectiveness of the proposed approach under a variety of complications that often arise in practice.

e-mail: lruan@gatech.edu

102. ESTIMATING EQUATIONS

ESTIMATING EQUATIONS IN BIASED SAMPLING PROBLEMS

Bin Zhang*, University of Alabama-Birmingham
Jing Qin, National Institute of Allergy and Infectious Diseases

This paper discusses a biased sample problem, where items are observed with probabilities that depend on the outcomes, with the parameter defined by some unbiased estimating equations. Among others, Qin (1993) considered the problem when only one response variable is involved and there exist two sets of biased samples with parameter of interest being the mean of the response variable. Here we consider more general situation with I biased samples and an auxiliary variable in addition to the response variable. Furthermore, the parameter can be any function of the two variables that associated with the unknown distribution. For the analysis, we generate the empirical likelihood approach in Qin (1993) and derive the likelihood ratio statistic. This statistic can be applied to both the hypothesis test and computing the confidence intervals. The likelihood ratio statistic is proved to follow a chi-square distribution asymptotically. Simulation studies show that our approach performs well. The methods are illustrated by application to a real data from cancer study.

e-mail: binzhang@uab.edu

BIAS SAMPLING, NUISANCE PARAMETERS, AND ESTIMATING EQUATIONS

Kunthel By*, University of North Carolina-Chapel Hill

Under random sampling, population parameters are identifiable. Biased sampling, on the other hand, introduces nuisance parameters that makes it hard to identify the population parameters of interest. In this paper, connections between biased sampling, nuisance parameters, and regression models for correlated data are explored. For a particular type of model, an estimating equation is proposed for estimating the parameters of interest when the study design is based on a biased sampling scheme. An example is given that illustrates how the method works. A calculation shows that the method provides an efficient framework for studying the relationship between correlated binary responses and covariates. A comparison of estimates and standard errors is made between the method of this paper and a related likelihood method.

e-mail: kby@bios.unc.edu

ESTIMATION IN LOGISTIC REGRESSION MODELS FOR CLUSTERED/LONGITUDINAL DATA WITH COVARIATE MEASUREMENT ERROR

Jeff Buzas*, University of Vermont

This talk considers estimation and inference in population averaged logistic regression models with covariate measurement error. It is shown that, surprisingly, standardized "residuals" in logistic regression models with covariate measurement error are unbiased, have constant variance, and preserve the correlation structure of the model with no measurement error. These properties are used to define unbiased estimating equations for population averaged logistic regression models when observations are clustered or measured repeatedly over time and covariates are measured with error.

e-mail: buzas@cems.uvm.edu

SEMIPARAMETRIC TRANSFORMATION MODELS FOR PANEL COUNT DATA WITH DEPENDENT OBSERVATION PROCESS

Ni Li*, University of Missouri
Liuquan Sun, Chinese Academy of Sciences
Jianguo Sun, University of Missouri

Panel count data usually occur in longitudinal follow-up studies that concern occurrence rates of certain recurrent events and in which study subjects can be observed only at discrete time points rather than continuously. In these situations, only the numbers of the events that occur between the observation times, not their occurrence times, are observed. Furthermore, the observation times or process may differ from subject to subject and more importantly, it may be related to the underlying recurrent event process. This paper discusses regression analysis of such data and for the problem, a class of semiparametric transformation models is presented. For estimation of regression parameters, some estimating equations are developed and the derived estimators are consistent and asymptotically normal with a covariance matrix that can be estimated consistently. An extensive simulation study was conducted and indicates that the proposed approach works well for practical situations. An illustrative example is provided.

e-mail: lini95@gmail.com

A COMPARISON OF SEVERAL APPROACHES FOR ANALYSIS OF LONGITUDINAL BINARY DATA

Matthew Guerra*, University of Pennsylvania
Justine Shults, University of Pennsylvania
Thomas Ten Have, University of Pennsylvania

In this presentation, we compare several approaches for the analysis of correlated binary measurements from longitudinal trials, including maximum likelihood analysis (ML), generalized estimating equations (GEE), quasi-least squares (QLS), and alternating logistic regressions (ALR). In contrast to the other approaches we consider, each of which models association via correlation, ALR models association via the less severely constrained

odds-ratio. We hypothesized that for longitudinal binary data with an AR(1) correlation structure, the performance of the ML approach would be superior, although QLS and GEE would be similar to each other and to the ML approach. We describe the relative benefits and limitations of each approach via asymptotic comparisons and simulations to compare the methods with respect to mean square error and bias. We also describe functions that we have developed in R for ML analysis of longitudinal binary data that allows for testing and construction of confidence intervals for both the regression and the correlation parameters.

e-mail: guerraw@mail.med.upenn.edu

AUGMENTED ESTIMATING EQUATIONS FOR SEMIPARAMETRIC PANEL COUNT REGRESSION WITH INFORMATIVE OBSERVATION TIMES AND CENSORING TIME

Xiaojing Wang*, University of Connecticut
Jun Yan, University of Connecticut

We propose an augmented estimating equation (AEE) approach for a semiparametric mean regression model with panel count data. On a fine grid, counts in all the subintervals of each observation window are treated as missing values, and are imputed with a robust working model given the observed count in the window. The observation scheme and the event process are allowed to be dependent through covariates and an unobserved frailty, which enters the mean function multiplicatively. The censoring time and the event process can be either conditionally independent or dependent through frailty given covariates. Regression coefficients and unspecified baseline mean function are estimated with an Expectation-Solving (ES) algorithm, which solves the conditionally expected version of the complete-data estimating equations given the observed data; distribution of observation times, censoring time, and frailty are all considered as nuisance. The equivalence of the ES algorithm and solving AEEs resulted from each imputed complete dataset is exploited to provide asymptotic distribution and variance estimator of the proposed estimator. Simulation studies demonstrate that the proposed estimator performs well for moderate sample sizes and appears to be competitive in comparison with existing estimators under a wide range of practical settings. The utility of the proposed methods is illustrated with a bladder tumor study.

e-mail: wangxj03@gmail.com

ANALYZING LENGTH-BIASED DATA WITH ACCELERATED FAILURE TIME MODELS

Jing Ning*, University of Texas Health Science Center at Houston
Jing Qin, National Institute of Allergy and Infectious Diseases
Yu Shen, University of Texas M.D. Anderson Cancer Center

Right-censored length-biased time to event data are often encountered in studies of epidemiologic cohorts, cancer prevention, and labor economy. One difficulty in analyzing this type of data is the informative right censoring due to the length-biased sampling mechanism. In this talk, we evaluate covariate effects on the failure times of the target population under a semiparametric accelerated failure time (AFT) model, given the observed length-biased data. The AFT model structure changes under length-biased sampling, and the techniques for conventional survival analysis are not applicable. We develop two estimating equation approaches to estimate the covariate effects on the unbiased failure times. The asymptotic properties of the new estimators are developed rigorously with the use of martingale theory. An elegant variance-covariance structure between the two estimating functions leads to a simple formula to study the asymptotic efficiency of the two estimators. We evaluate the empirical performance through simulation studies, and apply the method to data from a prevalent cohort study of individuals with dementia.

e-mail: jing.ning@uth.tmc.edu

103. RECENT ADVANCES IN MODELING NONLINEAR MEASUREMENT ERROR

CORRECTION FOR MEASUREMENT ERROR IN COVARIATES FOR INTERACTION MODELS

Havi Murad, Gertner Institute for Epidemiology, Israel
Victor Kipnis, National Cancer Institute
Laurence S. Freedman*, Gertner Institute for Epidemiology, Israel

When explanatory variables in regression models are measured with error, standard regression analyses yield biased estimators of the regression coefficients. If the covariates have non-differential error, regression calibration (RC) is often used to adjust the estimators, and is particularly simple when the covariates are normally distributed. However, when the regression model includes an interaction between two covariates measured with error, the error term of the interaction variable involves products of the covariates' errors and of the error of one covariate with the value of the other. We describe simple methods based on RC and the method of moments (MM) that consistently estimate interactions of normally distributed covariates with classical error. We show that the RC method adapted especially for normally distributed covariates is more efficient, but less robust to departures from normality, than MM. When covariate measurements are based on self-reports, the measurement error model is often non-classical, with error related to the true value of the covariate. We describe generalizations of our RC and MM approaches to a commonly-adopted class of linear non-classical measurement error model and present simulations that evaluate these methods. We illustrate the methods in applications to nutritional and other epidemiological data.

e-mail: lsf@actcom.co.il

CORRECTION FOR MEASUREMENT ERROR IN NUTRITIONAL EPIDEMIOLOGY: ALLOWING FOR NEVER AND EPISODIC-CONSUMERS IN MEASUREMENT ERROR MODELS FOR DIETARY ASSESSMENT INSTRUMENTS

Ruth Keogh*, MRC Centre for Nutritional Epidemiology in Cancer Prevention and Survival and MRC Biostatistics Unit, University of Cambridge
Ian White, MRC Biostatistics Unit, Cambridge, UK

Measures of dietary intake are subject to error with respect to measuring 'usual' intake, giving biased estimated diet-disease associations. Measurements from food records (24-hour recalls, food diaries) are often assumed to follow the classical measurement error model. Using repeated measures, diet-disease associations can be corrected using regression calibration. However, standard measurement error models do not always apply, one case being where the distribution of measurements is zero-inflated. The motivation is foods which some people never consume or consume episodically such that intake is not captured in a food record. A measurement error model which allows for never- and episodic-consumers is outlined, extending Kipnis et al's (2009) episodic-consumers model. Simulation studies are used to assess the proposed model, and results are compared with those from a classical measurement error approach and Kipnis et al's model. It is also shown how the effects of systematic error can be investigated using sensitivity analyses, since evidence suggests food record measurements may not be unbiased. Many studies use food frequency questionnaires (FFQ) as the main dietary assessment, with food records available in a sub-study, and the model is extended to incorporate FFQs and food records. The methods are illustrated using alcohol intake measurements in EPIC-Norfolk.

e-mail: ruth.keogh@mrc-bsu.cam.ac.uk

SIMULTANEOUS MODELING OF MULTIVARIATE DATA WITH EXCESS ZEROS AND MEASUREMENT ERROR WITH APPLICATION TO DIETARY SURVEYS

Victor Kipnis*, National Cancer Institute
Raymond J. Carroll, Texas A&M University
Laurence S. Freedman, Gertner Institute for Epidemiology and Public Health Policy, Israel
Douglas Midthune, National Cancer Institute

In public health surveillance, it is important to estimate the distribution of usual intake, i.e., the average daily intake of a dietary component in a fixed time period, using several short-term reported intakes. When a dietary component is episodically consumed, as occurs with most foods, the distribution of short-term reported intake has a spike at zero making conventional methods of modeling continuous data inappropriate. The National Cancer Institute (NCI) recently published a method for modeling a single episodically consumed food, based on the idea of a logistic mixed model for consumption and a linear mixed model for a transformed amount. However, nutritionists are often interested in estimating distributions of food densities, i.e., usual intakes expressed as percent calories or per 1,000 calories, which requires simultaneous modeling of a food and energy intakes. We modify and extend the NCI method to allow for the probability of food consumption on a certain day to be correlated with energy intake on that day. We discuss two principle approaches to modeling semi-continuous data, the two-part and the sample selection model. We show that the sample selection model includes the two-part model as a special but important case which dies not lead to ill-specified likelihoods.

e-mail: kipnisv@mail.nih.gov

104. USE OF BIOMARKERS IN PERSONALIZED MEDICINE

A PROCEDURE FOR EVALUATING PREDICTIVE ACCURACY OF BIOMARKERS FOR SELECTING OPTIMAL TREATMENTS

Xiao-Hua A. Zhou*, University of Washington
Yunbei Ma, University of Washington

Among patients with the same clinical disease diagnosis, response to the same treatment is often quite heterogeneous. For many diseases this may be due to molecular heterogeneity of the disease itself, which may be measured via a biomarker. Due to this molecular heterogeneity, a molecularly targeted treatment may be effective for only a subset of patients. A biomarker that can predictive which patients would benefit from one particular treatment is called the predictive biomarker. In this talk, we introduce a new procedure to measure the predictive accuracy of a biomarker and discuss how to estimate this measure.

e-mail: azhou@u.washington.edu

CLINICAL TRIAL DESIGNS FOR PREDICTIVE BIOMARKER VALIDATION: THEORETICAL CONSIDERATIONS AND PRACTICAL CHALLENGES

Sumithra J. Mandrekar*, Mayo Clinic
Daniel J. Sargent, Mayo Clinic

Biomarkers can guide patient- specific treatment selection by providing an integrated approach to prediction using the genetic makeup of the tumor and the genotype of the patient. Designs for predictive marker validation are broadly classified as retrospective (i.e., using data from previously well-conducted randomized controlled trials (RCT)) versus prospective (enrichment, all-comers or unselected, hybrid, or adaptive analysis). Well designed retrospective analysis can bring forward effective treatments to marker defined subgroup of patients in a timely manner (e.g.

K-RAS and colorectal cancer). Prospective enrichment designs are appropriate when compelling preliminary evidence suggests that not all patients will benefit from the study treatment, however this may sometimes leave questions unanswered (e.g. Trastuzumab and breast cancer). An unselected design is optimal where preliminary evidence regarding treatment benefit and assay reproducibility is uncertain (e.g. EGFR and lung cancer). Hybrid designs are appropriate when preliminary evidence demonstrate the efficacy of certain treatments for a marker defined subgroup, making it unethical to randomize patients with that marker status to other treatments (e.g. multigene assay and breast cancer). Adaptive analysis designs allow for pre-specified marker defined subgroup analyses. The implementation of these design strategies will lead to a more rapid clinical validation of biomarker guided therapy.

e-mail: mandrekar.sumithra@mayo.edu

ADAPTIVE DESIGNS TO VALIDATE CANCER BIOMARKERS

Liansheng Tang*, George Mason University

Most biomarker validation designs assume that the probabilities of assigning patients to treatment arms are fixed. We will propose adaptive designs which sequentially uses accumulating information about the treatment effect during the study to change the probabilities of assigning incoming patients to the two treatments so that we can put more patients on the better treatment.

e-mail: ltang1@gmu.edu

A THRESHOLD SAMPLE-ENRICHMENT APPROACH IN A CLINICAL TRIAL WITH HETEROGENEOUS SUBPOPULATIONS

Aiyi Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Qizhai Li, Chinese Academy of Sciences
Chunling Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Kai F. Yu, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Vivian Yuan, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Large comparative clinical trials usual target a wide-range of patients population in which subgroups exist according to certain patients' characteristics. Often, scientific knowledge or existing empirical data support the assumption that patients' improvement is larger among certain subgroups than the others. Such information can be used to design a more cost-effective clinical trial. The goal of the article is to use such information to design a more cost-effective clinical trial. A two-stage sample-enrichment design strategy is proposed that begins with enrollment from certain subgroup of patients and allows the trial to be terminated for futility in that subgroup. Simulation studies show that the two-stage sample-enrichment strategy is cost-effective if indeed the null hypothesis of no treatment improvement is true, as also so illustrated with data from a completed trial of calcium to prevent preeclampsia. The two-stage sample-enrichment approach borrows strength from treatment heterogeneity among target patients in a large scale comparative clinical trial. Feasibility of the proposed enrichment design relies on the knowledge prior to the start of the trial that certain patients can benefit more than others from the treatment. We will discuss some adaptive strategies to incorporate the information from an interim analysis.

e-mail: liua@mail.nih.gov

105. ANALYSIS OF RECURRENT EVENTS DATA IN THE PRESENCE OF A TERMINAL EVENT

SEMIPARAMETRIC ADDITIVE RATE MODEL FOR RECURRENT EVENT WITH INFORMATIVE TERMINAL EVENT

Jianwen Cai*, University of North Carolina-Chapel Hill
Donglin Zeng, University of North Carolina-Chapel Hill

We propose a semiparametric additive rate model for modelling recurrent events in the presence of the terminal event. The dependence between recurrent events and terminal event is fully nonparametric and is due to some latent process in the baseline rate function. Additionally, a general transformation model is used to model the terminal event given covariates. We construct an estimating equation for parameter estimation. The asymptotic distributions of the proposed estimators are derived. Simulation studies demonstrate that the proposed inference procedure performs well in realistic settings. Application to a medical study of patients with HIV is presented.

e-mail: cai@bios.unc.edu

MODELS FOR JOINT LONGITUDINAL AND EVENT-TIME OUTCOMES

Elizabeth H. Slate*, Medical University of South Carolina

This talk reviews alternate approaches for jointly modeling longitudinal and event-time outcomes. A shared random effects model and a latent class model are described. Complications associated with a recurrent event outcome, as opposed to a single event-time outcome, when modeled jointly with continuous longitudinal data are discussed. Applications to biomedical research data provide context and motivation.

e-mail: slate@musc.edu

ROBUST ESTIMATION OF MEAN FUNCTIONS AND TREATMENT EFFECTS FOR RECURRENT EVENTS UNDER EVENT-DEPENDENT CENSORING AND TERMINATION

Richard J. Cook*, University of Waterloo
Jerald F. Lawless, University of Waterloo
Lajmi Lakhal-Chaieb, Université Laval, Québec
Ker-Ai Lee, University of Waterloo

In clinical trials featuring recurrent clinical events, the definition and estimation of treatment effects involves a number of interesting issues, especially when loss to follow-up may be event-related and when terminal events such as death preclude the occurrence of further events. In this talk we consider a clinical trial of breast cancer patients with bone metastases where the recurrent events are skeletal complications, and where patients may die during the trial. We argue that treatment effects should be based on marginal rate and mean functions. When recurrent event data are subject to event-dependent censoring, however, ordinary marginal methods may yield inconsistent estimates. Incorporating correctly specified inverse probability of censoring weights into analyses can protect against dependent censoring and yield consistent estimates of marginal features. An alternative approach is to obtain estimates of rate and mean functions from models that involve some conditioning to render censoring conditionally independent. We consider three methods of estimating mean functions of recurrent event processes and examine the bias and efficiency of unweighted and inverse probability weighted versions of the methods with and without a terminating event. We compare the methods via simulation and use them to analyse the data from the breast cancer trial.

e-mail: rjcook@uwaterloo.ca

ANALYZING RECURRENT EVENTS DATA: A BAYESIAN PERSPECTIVE

Debajyoti Sinha*, Florida State University
Bichun Ouyang, Rush Medical Center
Elizabeth Slate, Medical University of South Carolina
Yu Gu, Florida State University

There has been a recent surge of interest in modeling and methods for analyzing recurrent events data with additional complexities. Two major examples of these recurrent events data are when risk of termination dependent on the history of the recurrent events and when the longitudinal measurements are recorded only at recurrent event times. We demonstrate how the modeling strategy for such data may depend on the practical issues related to the data example. We review the state of the art statistical methods and present novel theoretical properties, identifiability results and practical consequences of key modeling assumptions for several fully specified stochastic models for such studies. We also discuss the relationship as well as the major differences between these models in terms of their motivations and physical interpretations. We discuss associated Bayesian methods based on Markov chain Monte Carlo tools, and advantages of these Bayesian methods over competing analysis tools.

e-mail: sinhad@stat.fsu.edu

106. GENETIC STUDIES WITH RELATED INDIVIDUALS

HERITABILITY ESTIMATION USING REGRESSION MODELS FOR CORRELATION: QUANTITATIVE TRAITS FROM EXTENDED FAMILIES

Hye-Seung Lee*, University of South Florida
Myunghee C. Paik, Columbia University
Jefferey P. Krischer, University of South Florida

Heritability was originally developed as a part of quantitative genetics, which is to measure a genetic effect on a trait. Heritability is defined as a ratio of genetic variance to total phenotypic variance, which often employs variance component model to analyze data from extended families. Although the inference for the heritability has been well established based on restricted maximum likelihood under the general framework by Lange et al. (1976) and Hopper and Mathews (1982), it can be biased with higher correlation due to shared non-genetic effect among family members. This study proposes to use regression models for correlation parameter to infer the heritability, which will accommodate both one and multiple trait cases, and compares the performance with variance component model through simulations.

e-mail: leeh@epi.usf.edu

ASSOCIATION ANALYSIS OF ORDINAL TRAITS ON RELATED INDIVIDUALS

Zuoheng Wang*, Yale University

Statistical methods for association mapping of genetic variants are now well established for binary traits and continuous traits with normal distributions. However, many traits in health studies, such as cancer and psychiatric disorders, are recorded on a discrete, ordinal scale. Here we propose a novel method for the association analysis of ordinal traits when some sampled individuals are related, with known relationships. Our association test is a quasi-likelihood score test that accounts for relatedness of individuals. Simulation studies are conducted to evaluate the validity and power of the new method. We also discuss the extension of our method in the presence of population substructure.

e-mail: zuoheng.wang@yale.edu

PENALIZED ESTIMATION OF HAPLOTYPE FREQUENCIES FROM GENERAL PEDIGREES

Kui Zhang*, University of Alabama-Birmingham

Haplotype inference plays an important role in association studies and many EM based methods have been developed. One drawback of these methods is that many rare haplotypes with low explanatory power can be included, especially in the presence of missing data. This problem becomes more severe when haplotypes are estimated from general pedigrees or sibs. For general pedigrees, the genotypes of many founders can be missing. For sibs, the genotypes of parents are missing. To discourage the inclusion of rare haplotypes with low explanatory power, we propose a penalized method for haplotype inference. Specifically, a linear penalty is imposed to haplotypes with low frequency and the penalty levels off for haplotypes with frequency greater than a pre-specified threshold. Then the penalized likelihood is used to infer haplotypes and estimate their frequencies by a general minorize-maximize (MM) algorithm. The partition-ligation technique is also implemented to handle large number of markers. We evaluate its performance and compare it with the EM based method for haplotype inference from general pedigrees and sibs through extensive simulations. Our results indicate that the new proposed method outperforms the EM based methods in most situations.

e-mail: kzhang@ms.soph.uab.edu

AN EM COMPOSITE LIKELIHOOD APPROACH FOR MULTISTAGE SAMPLING OF FAMILY DATA

Yun-Hee Choi*, University of Western Ontario
Laurent Briollais, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto

Multistage sampling of family data is a common design in the field of genetic epidemiology, however appropriate methodologies to analyze data collected under this design are still lacking. We propose here a statistical approach based on the composite likelihood framework. The composite likelihood is a weighted product of individual composite likelihoods corresponding to the sampling strata where the weights are the inverse sampling probabilities of the families in each stratum. Our approach is developed for time to event data and can handle missing genetic covariates by using an EM algorithm. A robust variance estimator is employed to account for the non independence of individuals within families. An application to a family study of early-onset breast cancer demonstrates the interest of our approach. It confirms the important role of the genes BRCA1 and BRCA2 in these families and also shows evidence for a possible additional major gene that still need to be identified.

e-mail: ychoi97@uwo.ca

A LIKELIHOOD APPROACH FOR DETECTION OF IMPRINTING AND MATERNAL EFFECTS USING GENERAL PEDIGREE DATA

Jingyuan Yang*, The Ohio State University
Shili Lin, The Ohio State University

Both imprinting and maternal effects could cause parent-of-origin pattern in complex disease traits. Tests for imprinting effect may report false positives that are actually due to maternal effect. Existing likelihood approaches based on case-parent triads or nuclear families with multiple affected children are able to detect imprinting and maternal effects simultaneously and thus avoid potential confounding; however, none of them could accommodate extended families, which are commonly recruited in family-based association studies. We propose a Likelihood approach for detection of Imprinting and Maternal Effects (LIME) using general pedigrees. LIME formulates the probability of familial genotypes by considering conditional mating types of founders marrying into a pedigree, and models the penetrance of non-founders using a logit link. To utilize pedigrees with missing genotypes, LIME enumerates possible unobserved genotypes and sums over the likelihood. Our simulations have demonstrated that applying LIME to general pedigrees is much more powerful in detection of imprinting and maternal effects than trimming pedigrees to nuclear families, which is required by the existing methods.

e-mail: yj@stat.osu.edu

FUNCTIONAL MAPPING IN HUMAN POPULATION WITH GENETIC DATA STRUCTURE OF PARENTS AND CHILDREN

Jiangtao Luo*, Penn State College of Medicine
William W. Hager, University of Florida
Rongling Wu, Penn State College of Medicine

In this paper we consider the functional mapping in human population with genetic data structure of both parents and children. We study the linkage disequilibrium map of two generations. After giving a statistical model we also talk about its solution and related algorithm. The strategy is to provide some insight for the study inherited diseases in human population.

e-mail: jluo@hes.hmc.psu.edu

107. HYPOTHESIS TESTING, MULTIPLE TESTING, AND COMPUTATIONAL METHODS

FALSE DISCOVERY RATE CONTROL FOR HIGH DIMENSIONAL MULTIVARIATE DATA

Jichun Xie*, University of Pennsylvania
Tianwen Tony Cai, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

We consider the problem of false discovery rate (mFDR) control under high dimensional multivariate normal models. Using a method of compound decision rule, we develop an optimal joint oracle procedure and use a marginal procedure to approximate this optimal procedure. We show that the marginal plug-in procedure is asymptotically optimal under mild conditions of short-ranged dependency. Procedure-wise, the marginal plug-in procedure is the same as the the adaptive compound decision rules developed by Sun and Cai (2007). We show that the multiple testing procedure developed under the independent model is not only valid but also asymptotically optimal for under short-ranged dependency. The simulation studies have shown that the marginal oracle procedure can approximate the joint oracle procedure well, and the marginal plug-in procedure has smaller false non-discovery rate (mFNR) than the Banjamini and Hochberg (BH) step-up procedure. We applied the marginal plug-in and the procedures to the analysis of a case-control genetic association study of high-density lipoprotein and observed that the marginal plug-in procedure identified more single nucleotide polymorphisms than the BH procedure for any given mFDR level, suggesting that the former has a smaller mFNR.

e-mail: jichun@mail.med.upenn.edu

ADAPTIVE MULTIPLE TESTING PROCEDURES UNDER DEPENDENCE

Wenge Guo*, New Jersey Institute of Technology

In the context of multiple hypotheses testing, the proportion of true null hypotheses among all nulls often plays an important role, although it is generally unknown a priori. In adaptive procedures this proportion is estimated and then used to derive more powerful multiple testing procedures. Hochberg and Benjamini (1990) first presented adaptive procedures for controlling familywise error rate (FWER). However, until now, no mathematical proof has been provided to demonstrate that these procedures control the FWER. In this talk, we present new adaptive multiple testing procedures with control of the FWER under various conditions of dependence. First, we introduce a simplified version of Hochberg and Benjamini's adaptive Bonferroni and Holm procedures. In a conditional dependence model we prove that the former procedure controls the FWER in finite samples while the latter controls it approximately. Second, we present a new adaptive Hochberg procedure and prove it can control the FWER under positive regression dependence. Finally, through a small simulation study and a real data analysis, we illustrate that these adaptive procedures are more powerful than the corresponding conventional procedures.

e-mail: wenge.guo@njit.edu

A UNIVARIATE APPROACH TO REPEATED MEASURES AND MANOVA FOR HIGH DIMENSION, LOW SAMPLE SIZE

Yueh-Yun Chi*, University of Florida
Keith E. Muller, University of Florida

High-throughput technology in metabolomics and other fields gives rise to high dimension, low sample size (HDLSS) data when the number of variables exceeds the sample size. The inability of classical multivariate and Analysis of Variance (ANOVA) methods to provide valid analysis of such HDLSS data creates a grand challenge for statistics. To help meet the challenge, we extend the traditional "univariate approach" for Gaussian repeated measures (UNIREP) to HDLSS settings. Analytic and simulation results demonstrate that the proposed extension accurately controls Type I error for any population covariance pattern while the traditional method fails with HDLSS data. The extension also outperforms existing HDLSS methods, especially with small sample size. Without HLDSS, i.e., more subjects than variables, the extension provides a better overall control of Type I error than the traditional UNIREP tests. Free software facilitates implementing the method for a wide range of HDLSS applications, including repeated measures and multivariate ANOVA, discriminant analysis, canonical correlation, and multivariate regression.

e-mail: yychi@biostat.ufl.edu

APPLICATION OF ANBAR'S APPROACH TO HYPOTHESIS TESTING TO DETECT THE DIFFERENCE BETWEEN TWO PROPORTIONS

Julia Soulakova*, University of Nebraska
Ananya Roy, University of Nebraska

Anbar's (1983) approach for estimating a difference between two binomial proportions is discussed with respect to a hypothesis testing problem. Such an approach results in two possible testing strategies. While the results of the tests are expected to agree for a large sample size when two proportions are equal, the tests are shown to perform quite differently in terms of their probabilities of a type I error for selected sample sizes. Moreover, the tests can lead to different conclusions, which are illustrated via a simple example; and the probability of such cases can be relatively large. In an attempt to improve the tests while preserve their relative simplicity feature, a modified test is proposed. The performance of this test and a conventional test based on normal approximation is assessed. It is shown that the modified Anbar's test controls the probability of a type I error better for moderate sample sizes.

e-mail: jsoulakova2@unl.edu

NONPARAMETRIC TEST OF SYMMETRY BASED ON OVERLAPPING COEFFICIENT

Hani M. Samawi*, Georgia Southern University
Amal Helu, University of Jordan
Robert Vegal, Georgia Southern University

In this paper we introduce a new nonparametric test of symmetry based on the empirical overlap coefficient using kernel density estimation. Our investigation reveals that the new test is more powerful than the runs test of symmetry proposed by McWilliams (1990). Intensive simulation is conducted to examine the power of the proposed test. Data from a level I Trauma center are used to illustrate the procedures developed in this paper.

e-mail: hsamawi@georgiasouthern.edu

RAPIDSTAT: A HYBRID OF EXCEL AND GRAPHICAL LANGUAGE TO EXPEDITE USER INTERFACE CREATION

Pallabi Saboo, Harmonia Inc.
Marc Abrams*, Harmonia Inc.

Cancer research, especially in bioinformatics, has spawned work on new statistical techniques. We propose RapidStat, a programming langauge and system to assist statisticians and researchers to create easy-to-use user interfaces (UIs) for new statistical computation engines. RapidStat uses a novel graphical programming approach coupled to a declarative method of describing UIs. To facilitate transition RapidStat is an adjunct to a popular tool already used by researchers, and not a stand-alone tool. The results expected are to reduce the cost of adding UIs by 60%, to reduce training time to learn RapidStat by 80% over conventional languages, and to apply 35% of good UI style rules automatically to UIs created with RapidStat. Our research method is to first design the RapidStat language and system, then prototype the system, next enlist one cancer researcher and one leader in statistical software development to build one statistical and one non-statistical application with RapidStat to test and evaluate; to solicit feedback from focus groups; and to demonstrate at a conference. We leverage past work on graphical programming, UI design tools, the User Interface Markup Language we pioneered for standards group OASIS, and community-based documentation and knowledge sharing of reusable statistical applications through a semantic wiki.

e-mail: psaboo@harmonia.com

A SUMMARY OF GRAPHIC APPROACHES TO MONITOR PERFORMANCE OF LIVER TRANSPLANT CENTERS

Jie (Rena) Sun*, University of Michigan
John D. Kalbfleisch, University of Michigan

We present various graphical approaches to monitoring survival outcomes in medical centers over time using nationwide liver transplant centers as an example. A one-sided risk adjusted CUSUM with a constant control limit and an O-E risk-adjusted CUSUM with a V-mask as a control mechanism are introduced and evaluated theoretically and through simulation. We discuss processes associated with reviewing and reacting to signals and of restarting a CUSUM following such review. We also study the performance of both CUSUMs under different departures from the null distribution, and compare the methods through simulation with more traditional approaches to monitoring survival outcomes. Finally, the use of such charts in a national quality improvement program is discussed.

e-mail: renajsun@umich.edu

108. GENOMICS AND PROTEOMICS

MULTI-GENE DOMAIN CLUSTERS FOUND THROUGHOUT THE MOUSE GENOME VIA HIDDEN MARKOV MODELS

Jessica L. Larson*, Harvard University
Guocheng Yuan, Harvard University

There is evidence that neighboring genes, although not always involved in the same pathways, are still similarly regulated at the level of transcription via various histone modifications. We discovered and characterized the largest of these multi-gene domains through a novel genome-wide analysis of ChIP-seq histone modification data in mouse embryonic stem (ES) cells. We examined the activity of five of these modifications (H3K4me2, H3K4me3, H3K27me3, H3K9me3, H3K36me3) at all known mouse genes. We first obtained a 5-dimenisinal score for each gene based on average modification activity in select gene regions. Then, with hidden Markov models and corresponding algorithms, we were able to determine the most probable domain status of each gene. Our method located the known olfactory receptor and Hox gene clusters. Moreover, certain domains contain genes only found in select Gene Ontology groups. We also noted less gene expression variability within each of our domains when compared to randomly selected boundaries. We thus have evidence of multi-gene domains in mouse stem cells, which are characterized by similar patterns in five histone modifications. As we continue to apply our method to other cell lines, we will provide important insight into the general structure, organization, and regulation of the mammalian genome.

e-mail: jlarson@hsph.harvard.edu

DISSECTION OF ALLELE SPECIFIC COPY NUMBER CHANGES AND ITS APPLICATIONS

Wei Sun*, University of North Carolina-Chapel Hill

We developed a statistical software named genoCN, to simultaneously dissect copy number states and genotypes (i.e., allele-specific copy number) using high-density SNP arrays. Different strategies are employed to dissect Copy Number Variations (CNVs) in germline DNA and Copy Number Aberrations (CNAs) in tumor tissue. In contrast to most existing methods, GenoCN is more flexible since it estimates the parameters needed for the algorithm from the data, and it provides more informative results by outputting the posterior probabilities of allele-specific copy number calls. We will discuss thebackground, the software implementation and its application in association studies.

e-mail: weisun@email.unc.edu

BAYESIAN MODELING OF ChIP-CHIP DATA THROUGH A HIGH-ORDER ISING MODEL

Qianxing Mo*, Memorial Sloan-Kettering Cancer Center
Faming Liang, Texas A&M University

Chromatin immunoprecipitation (ChIP) followed by tiling microarray (Chip) analysis has been widely used for research in molecular biology. The most important feature of ChIP-chip data is that the intensity measurements of probes are spatially correlated due to the fact that the DNA fragments are hybridized to neighboring probes in the experiments. We propose a Bayesian hierarchical model for ChIP-chip data in which the spatial dependency of the data is modeled through a high-order Ising model. The proposed method is illustrated using several publicly available data sets and various simulated data sets, and compared with three alternative Bayesian methods, namely, BAC, HGMM, and Tilemap HMM. The numerical results indicate that the proposed method performs as well as the other three methods for high resolution data, but significantly outperforms the others for low resolution data. Additionally, the proposed method has better operating characteristics in terms of sensitivities and false discovery rates under various simulation scenarios.

e-mail: moq@mskcc.org

SIGNAL EXTRACTION AND BREAKPOINT IDENTIFICATION FOR ARRAY CGH DATA USING STATE SPACE MODEL

Bin Zhu*, University of Michigan
Peter X.K. Song, University of Michigan
Jeremy Taylor, University of Michigan

Motivation: Array comparative genomic hybridization(CGH) is a high solution technique to detect the DNA copy number variation, which plays a key role in the pathogenesis of cancers. Almost all of the CGH profiles contain biological and random errors, which make the position where copy number changes, called breakpoints, difficult to detect. A number of approaches have been proposed, most of which are sensitive to outliers and do not consider the uncertainty of profile estimation. Results: We propose a time-varying state space model for array CGH data analysis. The model consists of two equations: observation equation and state equation, where both the measurement error and evolution error are specified as the t-distribution with small degree of freedom. The CGH profiles are regarded as unknown signals estimated by a Markov Chain Monte Carlo algorithm. The breakpoints and outliers are identified by a backward selection procedure. Our method is robust to outliers and the estimation uncertainties are measured by posterior credible intervals. Glioblastoma Multiforme(GBM) data are used to demonstrate the characteristics of the proposed method. Compared to three other popular methods, our approach presents superior detection ability, which is exemplified through a simulated dataset and a breast tumor dataset.

e-mail: bzhu@umich.edu

A BAYESIAN MODEL FOR ANALYSIS OF COPY NUMBER VARIANTS IN GENETIC STUDIES

Juan R. Gonzalez*, CIBER Epidemiology and Public Health (CIBERESP), Spain Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, Spain
Juan J. Abellan, CIBER Epidemiology and Public Health (CIBERESP), Spain Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, Spain
Carlos Abellan, CIBER Epidemiology and Public Health (CIBERESP), Spain Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, Spain

The main goal of copy number variant (CNV) association studies is to assess the potential relationship between CNVs and disease. In these studies quantitative methods give, for each individual, CNV measurements from which the copy number status is usually inferred. Subsequently the CNV distribution is compared between cases and controls using standard tests. When/if there are different sub-phenotypes (e.g. cases from two or more diseases) one might be interested in assessing CNVs that are specific for each sub-phenotype. CNVs usually outnumber individuals, which makes the application of standard statistical models such as logistic regression infeasible. In this talk we present a novel approach to addressing CNV association studies. We apply a Bayesian shared-component model to differentiate the information that is common to cases and controls from the one that is specific to cases. This allows detecting the CNVs that show the strongest association with the disease(s) and that are specific to (each group of) cases only. We will also show through simulation studies that this method outperforms existing ones. The model will be illustrated using a real data set in which two different groups of cases diagnosed with asthma and atopy are compared to a control group. Acknowledgments: This work has

been partly funded by projects MTM2008-02457 from MICIN, Spain and AP-055/09 from Generalitat Valenciana.

e-mail: jrgonzalez@creal.cat

SEQUENTIAL SAMPLING DESIGNS FOR SMALL-SCALE PROTEIN INTERACTION EXPERIMENTS

Denise Scholtens*, Northwestern University
Bruce Spencer, Northwestern University

Bait-prey technologies that assay cellular protein interactions have recently surged in popularity. The most widely cited applications use steady-state systems such as Saccharomyces cerevisiae under assumedly stable cellular conditions and target global estimation of the cellular 'interactome' (Uetz et al. 2000; Ito et al. 2001; Gavin et al. 2006; Krogan et al. 2006). In contrast to genome-wide models, disease-relevant settings often consist of a small set of starting baits with local connectivity among their neighbors being the estimation goal. We present a collection of sequential experimental design schemes to increase coverage of each bait-prey assay and reduce variability of the inferred topologies for small-scale experimental settings in which local features take precedence over global modeling. Depending on the cost function for each round of experimentation, the size and connectivity of the relevant network, and the expected measurement error of the technology, various weighting schemes can offer distinct advantages to simple random sampling from among all eligible baits for each round of experimentation.

e-mail: dscholtens@northwestern.edu

A MULTI-STEP PROTEIN LYSATE ARRAY QUANTIFICATION METHOD AND ITS STATISTICAL PROPERTIES

Ji-Yeon Yang*, University of Illinois at Urbana-Champaign
Xuming He, University of Illinois at Urbana-Champaign

The protein lysate array is an emerging technology for quantifying the protein concentration ratios in multiple biological samples. Statistical inference for a parametric quantification procedure has been inadequately addressed in the literature, mainly because the appropriate asymptotic theory involves a problem with the number of parameters increasing with the number of observations. In this paper, we develop a multi-step procedure for the Sigmoidal models, ensuring consistent estimation of the concentration level with full asymptotic efficiency. The results obtained in the paper justify inferential procedures based on large-sample approximations. Simulation studies and real data analysis are used in the paper to illustrate the performance of the proposed method in finite-samples. The multi-step procedure is simpler in both theory and computation than the one-step least squares method that has been used in current practice.

e-mail: jiyang@illinois.edu

109. INFECTIOUS DISEASE AND MEDICAL CASE STUDIES

MODELING INFECTIVITY RATES AND ATTACK WINDOWS FOR TWO VIRUSES

Jing Zhang*, Miami University
Douglas Noe, Miami University
Jian Wu, Miami University
A. John Bailer, Miami University
Stephen Wright, Miami University

Cells exist in an environment in which they are simultaneously exposed to the number of viral challenges. In some cases, infection by one virus may preclude infection by other viruses. Under the assumption of independent times until infection by two viruses, a procedure is presented to estimate the infectivity rates along with the time window during which in a cell might be susceptible to infection by multiple viruses. A test for equal infectivity rates is proposed and interval estimates of parameters are derived. The operating characteristics of this test and estimation procedure is explored in a simulation study.

e-mail: zhangj8@muohio.edu

BAYESIAN INFERENCE FOR CONTACT NETWORKS GIVEN EPIDEMIC DATA

Chris Groendyke*, Penn State University
David R. Hunter, Penn State University
David Welch, Penn State University

Networks are now commonly used in the study of epidemics, where it has been shown that different contact network structures lead to different epidemic dynamics and, therefore, require different containment strategies. There has been little work, however, on the problem of inferring the structure of an underlying network having observed features of an epidemic. In this paper we build on work by Britton and O'Neill (2002) and estimate the parameters of a stochastic epidemic on a simple random network using data consisting of recovery times of infected hosts. The SEIR epidemic model we fit has exponentially distributed transmission times with gamma distributed exposed and infective periods on a network where the probability of any tie is p. We employ a Bayesian framework to make estimates of the joint posterior distribution of the model parameters. We discuss the accuracy of the estimates of different parameters and show that it is often possible to accurately recover the network parameter p. We demonstrate some important aspects of our approach by studying a measles outbreak in Hagelloch, Germany in 1861 consisting of 188 individuals. We provide an R package to carry out these analyses, which is available publicly on CRAN.

e-mail: cxg928@psu.edu

OPTIMIZING EXCHANGES IN A KIDNEY PAIRED DONATION (KPD) PROGRAM

Yijiang (John) Li*, University of Michigan
Yan Zhou, University of Michigan
John D. Kalbfleisch, University of Michigan
Peter X.-K. Song, University of Michigan

The old concept of barter exchange extends its application to the modern area of kidney transplantation, where incompatible donor-recipient pairs exchange donor organs to achieve better matches. In a large pool of kidney donor-recipient pairs, we consider planning exchanges so as to achieve maximum mutual benefits. Different from what has been previously considered, we propose to optimize the overall benefit of exchanges rather than just the total number of transplants. Our model explicitly takes into consideration utilities associated with planned transplants as well as the probabilities that a chosen exchange will actually result in a completed transplant. We develop a new strategy that maximizes expected utility in which the expectation takes account of all possible sub-cycles that might be implemented. This strategy takes advantage of possible alternative transplants that might be performed when the planned exchange fails. A graph search algorithm is implemented to evaluate all possible exchange cycles up to a certain length limit. The optimal choices can then be identified using an integer programming framework to maximize overall expected utilities. We will use simulation examples to demonstrate the model, algorithms, and solutions in a close reference to real world KPD programs.

e-mail: yijiang@umich.edu

ESTIMATION IN TYPE I CENSORED VIRAL LOAD ASSAYS UNDER NON-NORMALITY

Evrim Oral, Louisiana State University
Robbie A. Beyl*, Louisiana State University
William T. Robinson, Louisiana State University

Type I censored data are usually encountered in analyzing biomarker data. For example, left censored data are characteristic of many HIV studies due to inherent limit of detection in the assays. Left censored HIV biomarker measurements tend to be positively skewed, thus, in practice, the investigators often log-transform the distribution, and discard the non-detectable values or substitute a small constant value for them. Transforming the data and applying ad hoc practices can lead to bias in estimation. To address this issue, we consider maximum likelihood estimation of skewed Type I censored data. Since the maximum likelihood estimation is computationally very problematic for most non-normal distributions, we utilize the method of modified maximum likelihood methodology. The latter method is computationally straightforward and in situations where applied it yields estimators (MMLEs) essentially as efficient as the maximum likelihood estimators. We compare the efficiencies of derived MMLEs with the estimators of ad hoc models via simulations. We also study the robustness properties of the derived MMLEs under plausible deviations from an assumed model. We demonstrate the effectiveness of the MMLEs on a sample of HIV viral load data obtained from laboratory records reported to the Louisiana Office of Public Health HIV/AIDS Program.

e-mail: eoral@lsuhsc.edu

IMPROVED META ANALYSIS OF RANDOMIZED CONTROLLED TRIALS ON THE COMPARATIVE EFFICACY OF DAILY LOW-INTAKE OF DARK CHOCOLATE AMONG MIDDLE-AGED HYPERTENSIVE PATIENTS AS COMPARED TO PLACEBO

Martin Dunbar*, Georgia Southern University

An improved meta analysis is conducted to reveal, as compared to white chocolate, that daily low-intake of dark chocolate significantly reduces systolic blood pressure (SBP) and diastolic blood pressure (DBP). In this meta analysis study, there were six studies that met the inclusion-exclusion and were combined for analysis. The study designs used in these studies were randomized, investigator blinded clinical trials. The previous meta analysis compared the efficacy of cocoa and tea in relation to the reduction of blood pressure (Tambert, Roesen, Shomig, 2007). The goal of this meta analysis is to improve the meta analysis study that was published previously by Dirk Tambert, et al. (2007). Collectively, there were 141 participants that were randomized in different phases of the clinical trial. The results of the study show that there was a significant decrease in both forms of blood pressure (SBP and DBP) after taking flevanol-rich dark chocolate and after taking flevanol-free white chocolate. Because of heterogeneity among the studies, there were only four studies that were included in the final analysis.

e-mail: m_x_bar@hotmail.com

INFORMATIVE DORFMAN SCREENING WITH RISK THRESHOLDS

Christopher S. McMahan*, University of South Carolina
Joshua M. Tebbs, University of South Carolina
Christopher R. Bilder, University of Nebraska

Since the advent of group testing, there have been many successful methodological applications of pooling to problems in infectious disease screening, drug discovery, and genetics. In many of these applications, the goal is to identify the individual as either positive or negative through initial testing responses of the groups and the subsequent process of decoding positive pools. There have been many decoding procedures suggested, although all of these procedures fail to acknowledge, and to further exploit, the heterogeneous nature of the individuals. In our work, we utilize the individuals' positive probabilistic status to drive an informed Dorfman decoding procedure. We derive closed form expressions for the probability mass function for the number of tests needed to decode the pools, under our informed decoding procedure.

Then, we introduce the idea of thresholding a set of individuals by classifying individuals as either high or low risk, so that class-specific decoding procedures can be employed. Our testing procedure's efficiency is then illustrated through simulation and is further applied to a chlamydia and gonorrhea study. Overall, this work shows that our approach provides significant gains in reducing testing expenditures while providing an easy-to-implement decoding procedure.

e-mail: mcmahanc@mailbox.sc.edu

110. VARIABLE SELECTION METHODS

A PATH FOLLOWING ALGORITHM FOR SPARSE PSEUDO- LIKELIHOOD INVERSE COVARIANCE ESTIMATION (SPLICE)

Guilherme V. Rocha*, Indiana University
Peng Zhao, Citadel Investment Group
Bin Yu, University of California-Berkeley

Given n observations of a p-dimensional random vector, the covariance matrix and its inverse (precision matrix) are needed in a wide range of applications. Sample covariance (e.g. its eigenstructure) can misbehave when p is comparable to the sample size n. Regularization is often used to mitigate the problem. In this paper, we proposed an l1-norm penalized pseudo-likelihood estimate for the inverse covariance matrix. This estimate is sparse due to the l1-norm penalty, and we term this method SPLICE. Its regularization path can be computed via an algorithm based on the homotopy/LARS-Lasso algorithm. Simulation studies are carried out for various inverse covariance structures for p=15 and n=20, 1000. We compare SPLICE with the l1-norm penalized likelihood estimate and a l1-norm penalized Cholesky decomposition based method. SPLICE gives the best overall performance in terms of three metrics on the precision matrix and ROC curve for model selection. Moreover, our simulation results demonstrate that the SPLICE estimates are positive-definite for most of the regularization path even though the restriction is not enforced.

e-mail: gvrocha@indiana.edu

MULTICATEGORY VERTEX DISCRIMINANT ANALYSIS FOR HIGH-DIMENSIONAL DATA

Togntong Wu*, University of Maryland
Kenneth Lange, University of California-Los Angeles

In response to the challenges of data mining, discriminant analysis continues to evolve as a vital branch of statistics. Our recently introduced method of vertex discriminant analysis (VDA) is ideally suited to handle multiple categories and an excess of predictors over training cases. The current paper explores an elaboration of VDA that conducts classification and variable selection simultaneously. Adding lasso (L1-norm) and Euclidean penalties to the VDA loss function eliminates unnecessary predictors. Lasso penalties apply to each predictor coefficient separately; Euclidean penalties group the collective coefficients of a single predictor. With these penalties in place, cyclic coordinate descent accelerates estimation of all coefficients. Our tests on simulated and benchmark real data demonstrate the virtues of penalized VDA in model building and prediction in high-dimensional settings.

e-mail: ttwu@umd.edu

ORDER THRESHOLDING

Min HeeKim*, Penn State University
Michael G. Akritas, Penn State University

A new thresholding method, based on L-statistics and called order thresholding, is proposed as a technique for improving the power when testing against high-dimensional alternatives. The new method allows great flexibility in the choice of the threshold parameter. This results in improved power over the soft and hard thresholding methods. Moreover, order thresholding is not restricted to the normal distribution. An extension of the basic order threshold statistic to high-dimensional ANOVA is presented. The performance of the basic order threshold statistic and its extension is evaluated with extensive simulations.

e-mail: mzk132@psu.edu

VARIABLE SELECTION IN PARTIAL LINEAR ADDITIVE MODEL

Fengrong Wei*, University of West Georgia

We consider the problem of simultaneous variable selection and estimation in partial linear additive models with a large number of grouped variables in the linear part and a large number of nonparametric components. In our problem, the number of grouped variables may be larger than the sample size, but the number of important groups is "small" relative to the sample size. We apply the adaptive group Lasso to select the important variables in the linear part based on the polynomial spline approximation for the nonparametric additive components. We first use the group Lasso to obtain an initial rate consistent estimator and reduce the dimension of the problem. Under appropriate conditions, it can be shown that the group Lasso selects the number of groups which is comparable with the underlying important groups and is estimation consistent, the adaptive group Lasso selects the correct important groups with probability converging to one as the sample size increases and is selection consistent.

e-mail: weifengrong@hotmail.com

CONFIDENCE REGION BASED TUNING FOR FORWARD AND BACKWARD SELECTION

Funda Gunes*, North Carolina State University
Howard Bondell, North Carolina State University

Forward and backward selection are standard and widely used methods for variable selection. However, if the goal is to identify the correct sparse model, these methods can perform poorly. We propose to tune the regression coefficients based on a confidence region level which can be applied to these stepwise procedures. As opposed to sequential testing for addition/deletion of a predictor, a full joint confidence region is used to define the stopping rule. The tuning parameter has a simple interpretation as a confidence level, and thus chosen a priori, as usual by specifying the value, such as a 95% confidence region. The proposed method has the ability to be used with a large variety of statistical methods where confidence regions can be created for model coefficients. Furthermore, the approach can be applied to regions constructed via likelihood-based methods, Wald-type methods, or any other approach. Simulation studies show that the proposed approach generally outperforms the usual forward and backward selection methods in terms of correct selection.

e-mail: fgunes@ncsu.edu

ROBUST PENALIZED LOGISTIC REGRESSION WITH TRUNCATED LOSS FUNCTIONS

Seo Young Park*, University of North Carolina-Chapel Hill
Yufeng Liu, University of North Carolina-Chapel Hill

The Penalized Logistic Regression (PLR) is a powerful statistical tool for classification. It has been commonly used in many practical problems. Despite its success, since the loss function of the PLR is unbounded, resulting classifiers can be sensitive to outliers. To build more robust classifiers, we propose the Robust PLR (RPLR) which uses truncated logistic loss functions, and suggest three schemes to estimate conditional class probabilities. Connections of the RPLR with some other existing work on robust logistic regression have been discussed. Our theoretical results indicate that the RPLR is Fisher-consistent and more robust to outliers. Through numerical examples, we demonstrate that truncating the loss function indeed yields better performance in terms of classification accuracy and class probability estimation.

e-mail: seoyoung@email.unc.edu

ADAPTIVE MODEL SELECTION IN LINEAR MIXED MODELS

Bo Zhang*, National Institute of Child Health and Human Development
Xiaotong Shen, University of Minnesota
Zhen Chen, National Institute of Child Health and Human Development

In biomedical studies, linear mixed models are commonly used when the data are grouped according to one or more clustering factors. It is accepted that the selection of covariates and variance components is crucial to the accuracy of both estimation and prediction in linear mixed models. Most existing information criteria, such as Akaike's information criterion, Bayesian information criterion, and the risk inflation criterion, penalize an increase in the size of a model through a fixed penalization parameter. In this project, we developed a model selection procedure with a data-adaptive model complexity penalty for selecting linear mixed models, based on the derived generalized degrees of freedom of linear mixed models. We studied the asymptotic optimality of the adaptive model selection procedure in linear mixed models over a class of information criteria and evaluated its finite-sample performance with numerical simulations. Our simulation results show that the adaptive model selection procedure outperforms the information criteria in selecting covariates and variance components in linear mixed models. Finally we demonstrated the adaptive model selection procedure by applying it to a real data example. This research was supported in part by the Intramural Research Program of the NIH, Eunice Kennedy Shriver National Institute of Child Health and Human Development.

e-mail: zhangb5@mail.nih.gov

111. GENERALIZED LINEAR MODELS

CONDITIONAL LOGISTIC MIXED EFFECTS MODEL FOR UNBALANCED MATCHED CASE-CONTROL STUDIES

Inyoung Kim*, Virginia Tech University
Feng Guo, Virginia Tech University

In matched case-control studies, the conditional logistic regression is the most commonly used to study association between the relative risk of binary outcome and the interest covariate. A limitation of the conditional logistic regression model is that all stratum has the same effect among all stratum. Another limitation is that the covariates whose values are the same between case and control do not play a role in conditional logistic regression model because any covariates whose values are the same between case and control are removed by conditioning on the fixed number of cases and controls in the stratum. Hence, in this paper, we propose the mixed effects model to overcome these limitations in the conditional logistic regression model. We consider the stratum variable is following random effect with depending on subjects in each stratum. Four different methods are developed: (1) Bias corrected quasi likelihood based approach (2) Monte Carlo Expectation Maximization algorithm (3) Parametric Bayesian method and (4) Semiparametric Bayesian method. We perform simulation to compare these methods. We demonstrate the advantage of our approaches using both balanced and unbalanced matched case-control studies from public health and traffic accident, respectively.

e-mail: inyoungk@vt.edu

ACCURACY AND PRECISION OF ESTIMATES IN COVARIATE- ADJUSTED GENERALIZED LINEAR REGRESSION MODELS WITH OR WITHOUT TREATMENT AND COVARIATE INTERACTION

Junyi Lin*, Penn State University
Lei Nie, U.S. Food and Drug Administration
Runze Li, Penn State University

People debate whether or not a covariate-adjusted approach should be used as the primary analysis. As expected, omitting predictive covariates often leads to misspecified models in which the parameter of interest is difficult to interpret, particularly when omitted covariates interact with the main predictive variable. Under a generalized linear model framework, we derive the analytical relationship between the parameters of interest in the potentially misspecified model and the true model. Meanwhile, we show that for a broad class of generalized linear models, the estimates obtained from a covariate-adjusted model have greater variances compared to those from an unadjusted model. These theoretical results are illustrated and validated through two examples and a simulation study. We allow models to include treatment/covariates interactions, and hence in terms of accuracy, our results include analogue conclusions in Gail et al (1984) as special cases. In terms of precision, we make substantial extensions of results in Robinson and Jewell (1991) to a broad class of generalized linear models including the most frequently used ones.

e-mail: jul216@psu.edu

MCEM-SR AND EM-LA2 FOR FITTING GENERALIZED LINEAR MIXED MODELS

Vadim V. Zipunnikov*, Johns Hopkins University
James G. Booth, Cornell University

The expectation-maximization algorithm has been advocated recently by a number of authors for fitting generalized linear mixed models. However, the E-step typically involves analytically intractable integrals which have to be approximated. We suggest two alternative approaches to solve this problem. The first one, MCEM-SR, approximates the integrals by using a randomized spherical-radial integration which dramatically reduces the computational burden of implementing EM. The other approach, EM-LA2, is based on higher-order Laplace approximation of the integrals. A closed form of the standardized cumulants for generalized linear models which are the higher-order terms of the Laplace approximation is incorporated in the EM algorithm resulting in a fast and efficient procedure. We illustrate both methods by fitting models to two well-known data sets, and compare our results with those of other authors.

e-mail: vzipunni@jhsph.edu

RATIONALE FOR CHOOSING EXPLICIT CORRELATION STRUCTURE IN A MULTIVARIATE

Folefac D. Atem*, University of Pittsburgh
Stewart J. Anderson, University of Pittsburgh

The analysis of multileveled data with bivariate outcomes is very common in the fields of education, health economics and health service research (Rochon 1996, Thiebaut, Jacqmin-Gadda,etc 2002). Modeling bivariate outcomes is very useful in HIV research where the joint evolution of HIV RNA and CD4+t lymphocytes in a cohort of HIV-1 infected patient treated with active antiretroviral treatment. The use of MIXED model method and the Generalized Estimating Equations (GEE) are the most influential recent developments in statistical practice analysis techniques. The GEE model estimates are consistent irrespective of the correlation structure but greater efficiency will be realized by those correlation models closer to true correlation structure (Rochon 1996). The Mixed model procedure uses the Newton-Raphson algorithm known to be faster than the EM algorithm. Nonetheless, the linear mixed model takes into account all available information and deal with both serial and the intra-subject correlation. The efficiency of the model depends on the correlation structure. Hence the difference between a correctly detecting a significant result or notmight be due to a poor correlation structure. In this paper we will come up a rationale in choosing explicit working correlation structure in a multilevel data with bivariate outcome.

e-mail: fda1@pitt.edu

ESTIMATION OF THE STANDARD DEVIATION FOR AN EXPONENTIAL DISTRIBUTION FROM LIMITED DATA

Yvonne M. Zubovic*, Indiana University Purdue University-Fort Wayne
Chand K. Chauhan, Indiana University Purdue University-Fort Wayne

In a variety of applications, such as sample size determination, an estimate of the standard deviation for the underlying population is required. If no estimate is readily available, the experimenter may conduct a pilot study to obtain an estimate of the standard deviation, but this reduces the resources available for the experiment. Another approach is to use information from a previous study published by another investigator. Frequently, the information available from a published study may be limited to various summary statistics rather than the full set of sample data. The objective of this paper is to derive an estimator for the standard deviation of the underlying population when only the sample size and select percentiles are available. In this paper, the authors present estimators of the standard deviation for an exponential distribution based on specified percentiles. Various theoretical properties of the estimators are presented and these properties are compared via simulation to estimators based on more complete information from

the sample. In addition, simulation results are shared to investigate the effect of using these derived estimators in the problem of sample size determination.

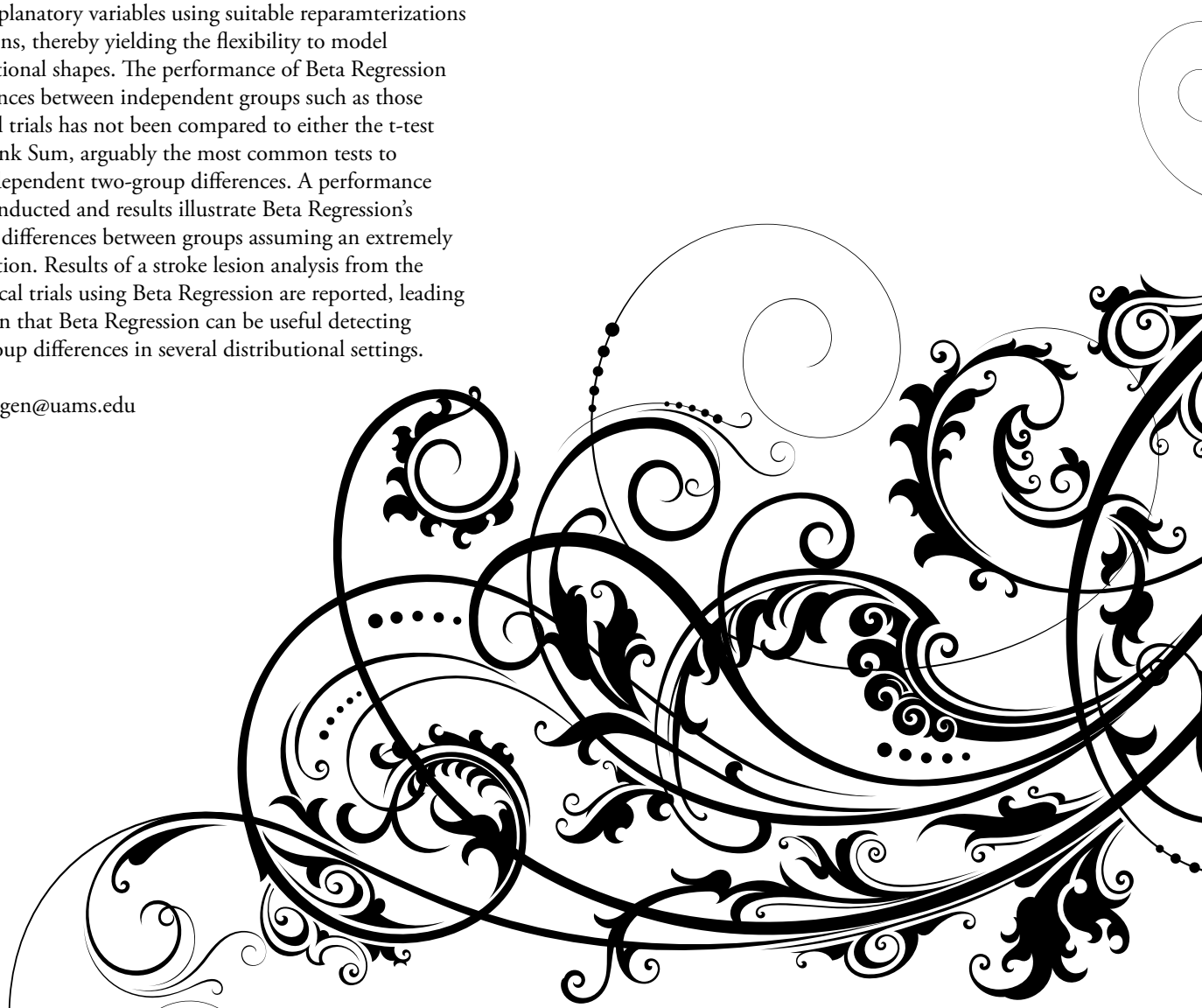
e-mail: zubovic@ipfw.edu

PERFORMANCE OF BETA REGRESSION IN DETECTING INDEPENDENT GROUP DIFFERENCES

Christopher J. Swearingen*, University of Arkansas for Medical Sciences
Dipankar Bandyopadhyay, Medical University of South Carolina
Robert F. Woolson, Medical University of South Carolina
Barbara C. Tilley, University of Texas Health Science Center

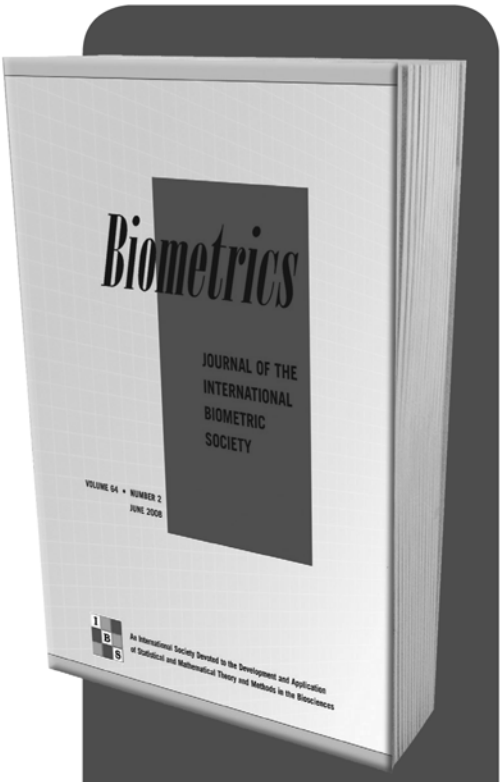
Biomedical measurements can generate skewed distributions consisting of valid measurements at the measure's boundary combined with other observations scattered in the known measurement space. Motivated by an extremely skewed distribution of stroke lesion volume from two clinical trials, we investigate Beta Regression as a possible modeling approach. Beta Regression assumes the dependent variable follows a Beta distribution marginally, estimating the parameters of the Beta distribution by regressing on explanatory variables using suitable reparamterizations and link functions, thereby yielding the flexibility to model various distributional shapes. The performance of Beta Regression to detect differences between independent groups such as those found in clinical trials has not been compared to either the t-test or Wilcoxon Rank Sum, arguably the most common tests to determining independent two-group differences. A performance simulation is conducted and results illustrate Beta Regression's ability to detect differences between groups assuming an extremely skewed distribution. Results of a stroke lesion analysis from the motivating clinical trials using Beta Regression are reported, leading to the conclusion that Beta Regression can be useful detecting independent group differences in several distributional settings.

e-mail: cswearingen@uams.edu



Biometrics

Published on behalf of the International Biometric Society (IBS)



Biometrics emphasizes the role of statistics and mathematics in the biosciences. Its objectives are to promote and extend the use of statistical and mathematical methods in the principal disciplines of biosciences by reporting on the development and application of these methods. A centerpiece of most Biometrics articles is a scientific application that sets scientific or policy objectives, motivates methods development, and demonstrates the operations of new methods.

Have you read these top accessed papers from 2009?

Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression
(Vol. 65, No. 1)

Diagnosis of Random-Effect Model Misspecification in Generalized Linear Mixed Models for Binary Response
(Vol. 65, No. 2)

On Gene Ranking Using Replicated Microarray Time Course Data
(Volume 65, No. 1)

The International Biometric Society (IBS) is an international society devoted to the development and application of statistical and mathematical theory and methods in the biosciences. The IBS is the parent society of ENAR, the Eastern North American Region of the IBS.

Sign up for email alerts, access a free sample copy, submit your manuscript and more...
www.interscience.wiley.com/journal/biometrics

09-10357

Index

Index

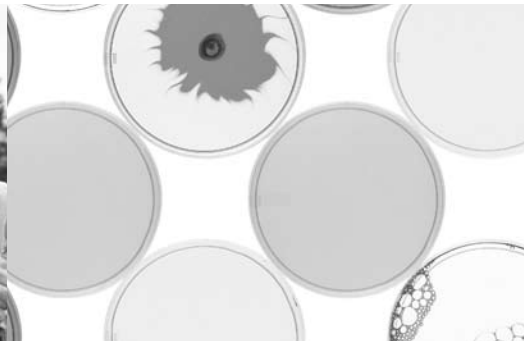
Index

Index

Index

Commitment to Scientific Innovation





At Abbott, we value the diversity of our products, technologies, markets and most importantly, our people. Our global workforce provides the perspectives and experience necessary to translate science into real solutions for patients in over 130 markets worldwide. And through our diverse healthcare businesses, we connect people and potential in ways that no other company can.

Our employees are given the tools to succeed, lead and grow, with challenging and rewarding opportunities and work that makes a difference. Innovative thinkers who are passionate about the work they do, Abbott employees strive to improve the lives of the patients we serve every single day across the globe.

Manager, Statistics
Req #69216BR

You will work on clinical pharmacology/pharmacokinetic studies or on parts of phase II/III studies pertaining to pharmacokinetics and the relationship between response and exposure to drug. Requires PhD in statistics with 5 years' experience or MS in statistics with 8 years' experience in clinical pharmacology/pharmacokinetic area of pharmaceutical development, strong knowledge of statistical theory and methodology and theory of linear models, including repeated measures analysis and mixed effects models.

Senior Research Statistician
Req #71033BR

Effectively partner with other team members in a multidisciplinary setting on phase II-IV clinical programs, including the design of innovative clinical trials, development of statistical analysis strategies, analysis and interpretation of safety/efficacy data, and the preparation of study reports and regulatory submissions. Qualified candidates should have a degree in statistics, biostatistics, or closely related field, with 4-7 years' experience with PhD or 5-8 years' experience with MS in the pharmaceutical industry. Good communication skills are essential. SAS programming experienced preferred.

Manager, Statistics
Req #66335BR

You will work as a project statistician in the immunology therapeutic area. Requires PhD in statistics or biostatistics with at least 5 years' pharmaceutical experience. Experience with clinical trials, including interaction with regulatory agencies required. Good communication skills are essential. SAS programming experience required.

Associate Statistician
Req #67514BR

Effectively work with senior statistical staff on phase II-IV clinical programs, including the design of innovative clinical trials, analysis and interpretation of safety/efficacy data, and the development of statistical analysis plans, study reports and regulatory submissions. Qualified candidates should have a master's in statistics, biostatistics, or closely related field, with 0-2 years experience in the pharmaceutical industry. Good communication skills are essential. SAS programming experienced preferred.

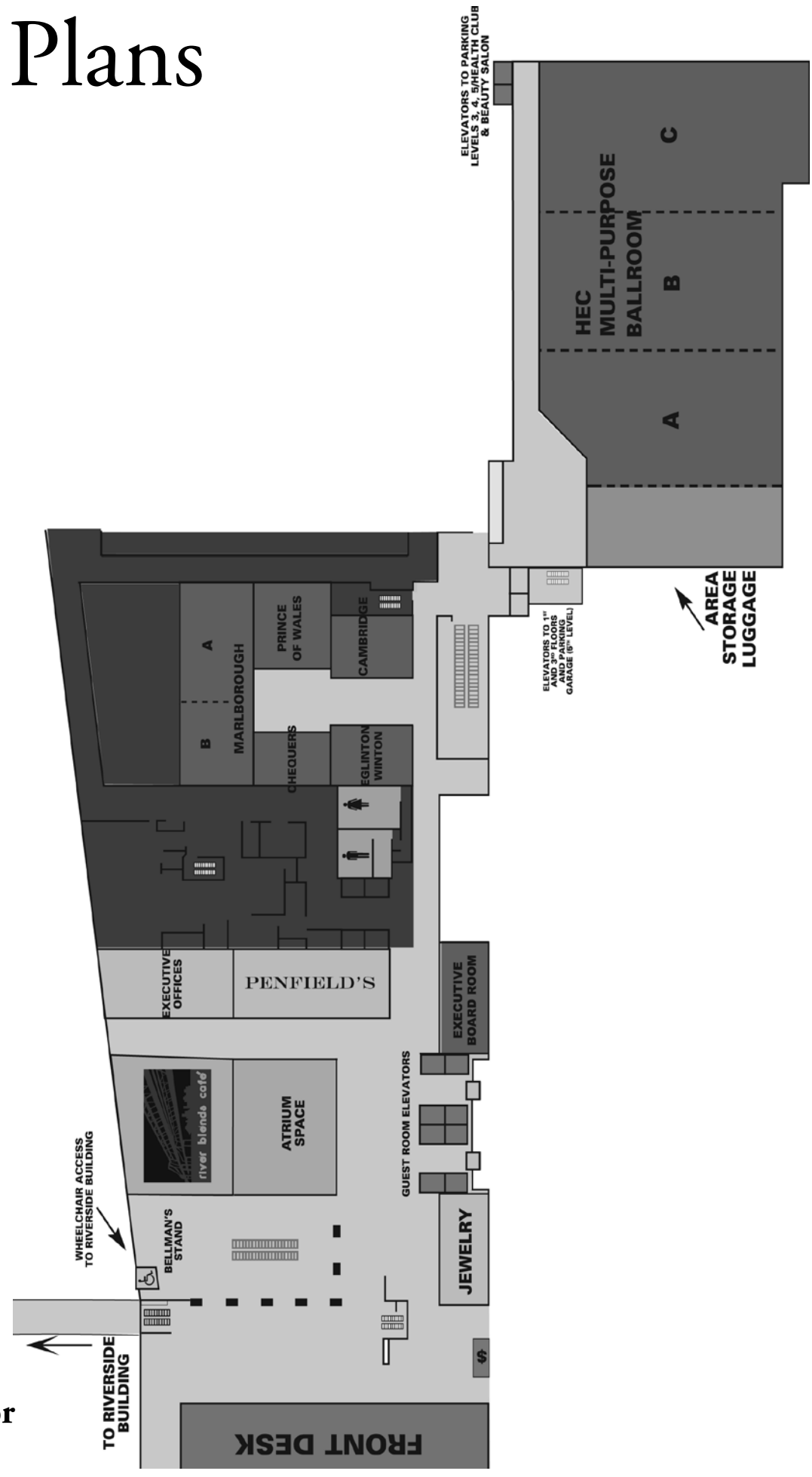
If you want to work with people who share a common desire to improve the quality of people's lives, we invite you to visit www.abbott.com, click on Careers, Search Openings and enter the appropriate Req # into the keyword field.

www.abbott.com

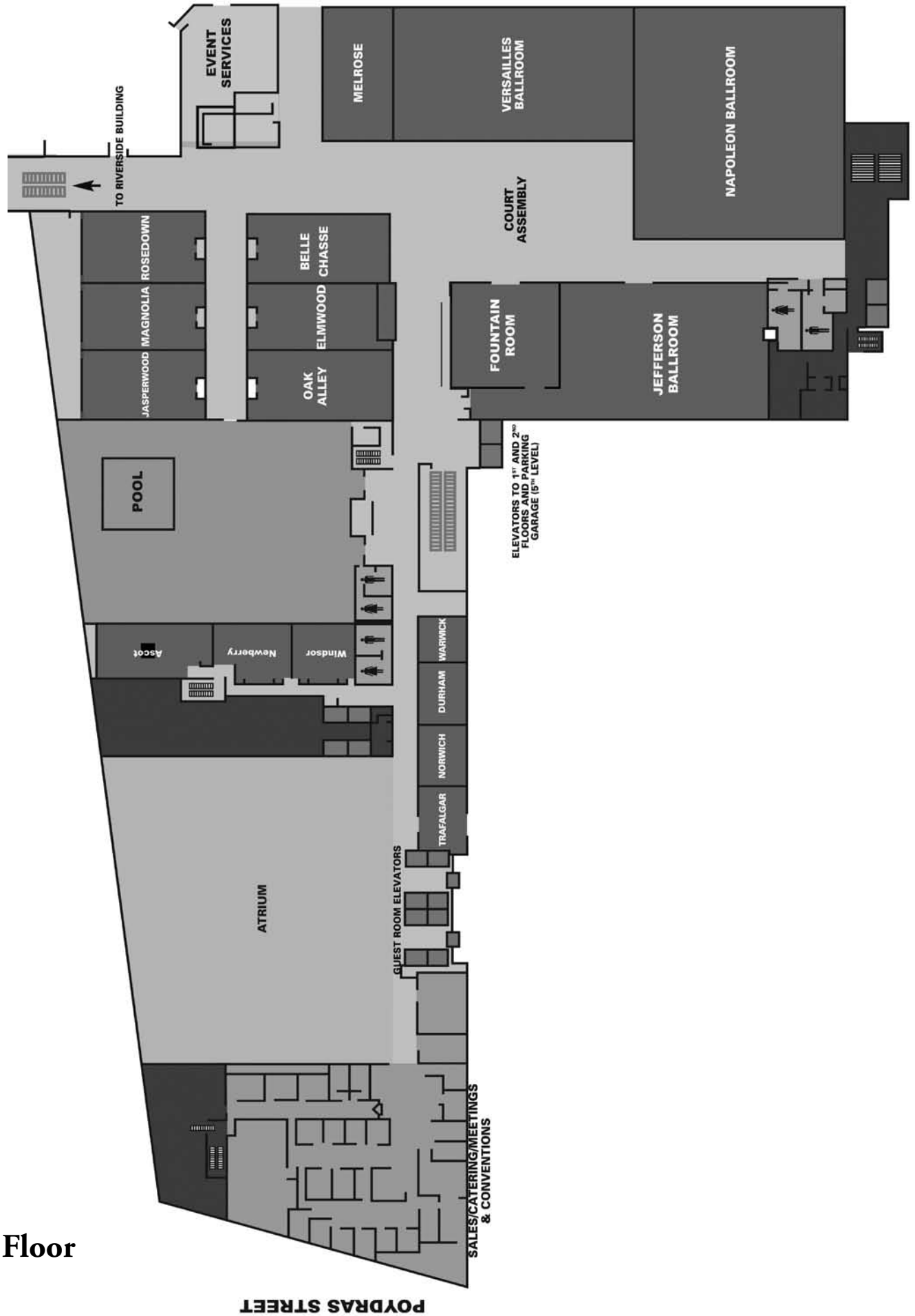
Abbott
A Promise for Life

An equal opportunity employer, Abbott welcomes and encourages diversity in our workforce.

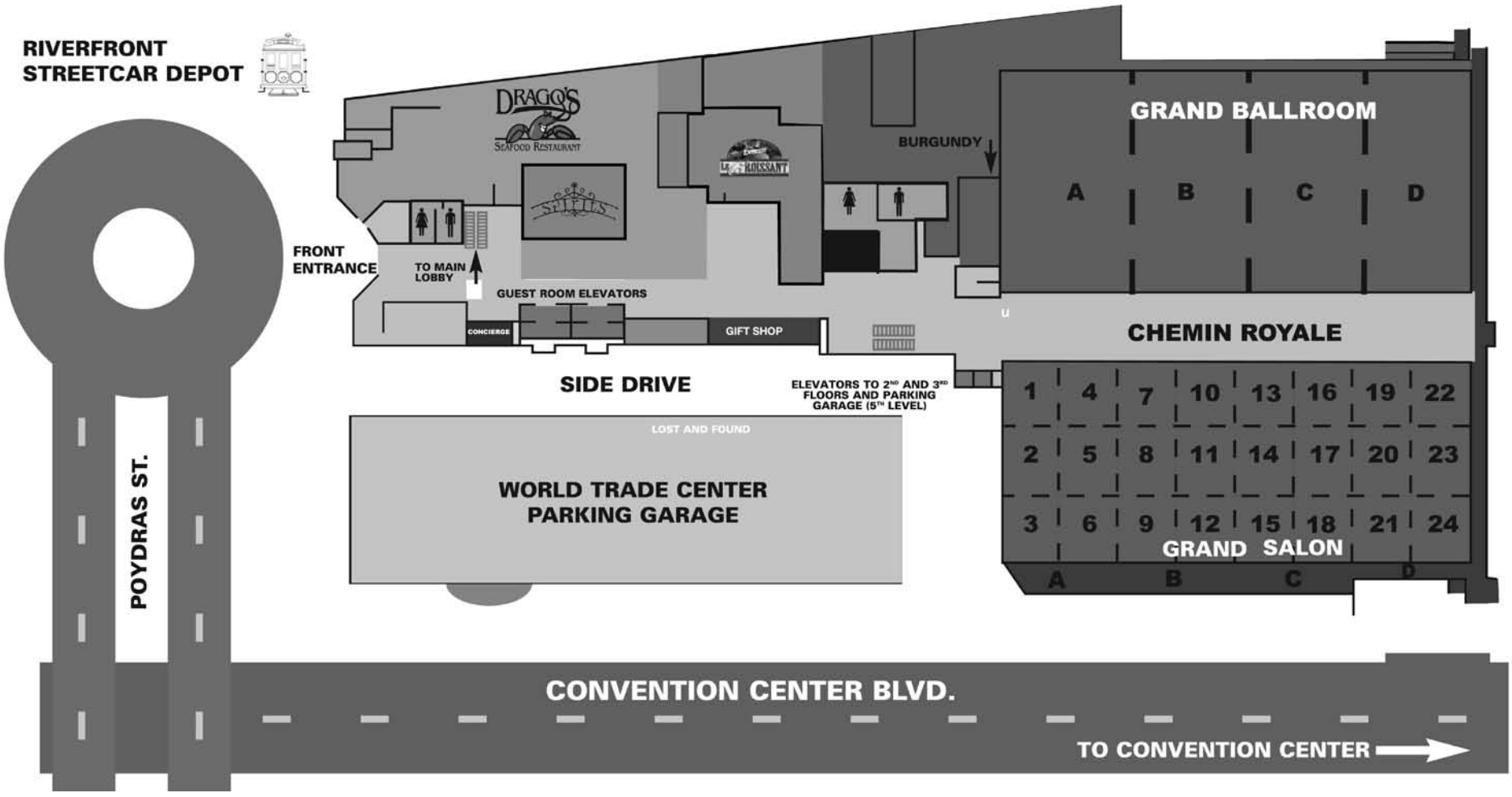
Floor Plans

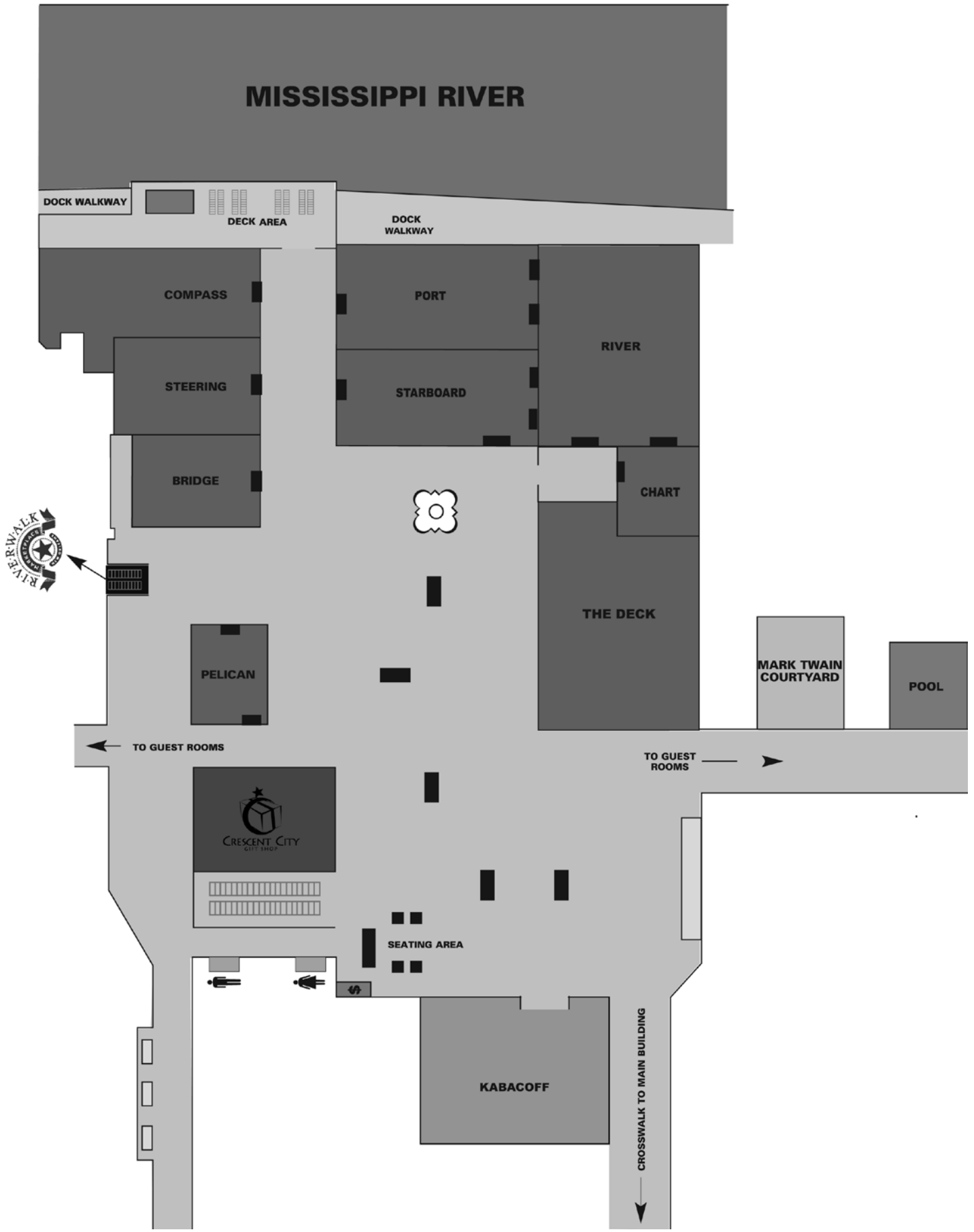


Second Floor


TO RIVERSIDE BUILDING
EVENT SERVICES
MELROSE
VERSAILLES BALLROOM
NAPOLEON BALLROOM
COURT ASSEMBLY
ROSEDOWN
MAGNOLIA
JASPERWOOD
BELLE CHASSE
ELMWOOD
OAK ALLEY
FOUNTAIN ROOM
JEFFERSON BALLROOM
POOL
ELEVATORS TO 1ST AND 2ND FLOORS AND PARKING GARAGE (5TH LEVEL)
Ascot
Newberry
Windsor
WARWICK
DURHAM
NORWICH
TRAFALGAR
ATRIUM
GUEST ROOM ELEVATORS
SALES/CATERING/MEETINGS & CONVENTIONS
POYDRAS STREET

Third Floor


RIVERFRONT
STREETCAR DEPOT
DRAGO'S
SEAFOOD RESTAURANT
LE RESTAURANT
SUITES
BURGUNDY
GRAND BALLROOM
A
B
C
D
FRONT
ENTRANCE
TO MAIN
LOBBY
GUEST ROOM ELEVATORS
CONCIERGE
GIFT SHOP
CHEMIN ROYALE
SIDE DRIVE
ELEVATORS TO 2ND AND 3RD
FLOORS AND PARKING
GARAGE (5TH LEVEL)
1 4 7 10 13 16 19 22
2 5 8 11 14 17 20 23
3 6 9 12 15 18 21 24
GRAND SALON
A
B
C
D
LOST AND FOUND
WORLD TRADE CENTER
PARKING GARAGE
POYDRAS ST.
CONVENTION CENTER BLVD.
TO CONVENTION CENTER


MISSISSIPPI RIVER
DOCK WALKWAY
DECK AREA
DOCK WALKWAY
COMPASS
PORT
RIVER
STEERING
STARBOARD
BRIDGE
CHART
RIVERWALK
THE DECK
PELICAN
MARK TWAIN COURTYARD
POOL
TO GUEST ROOMS
TO GUEST ROOMS
CRESCENT CITY GIFT SHOP
SEATING AREA
KABACOFF
CROSSWALK TO MAIN BUILDING

Notes

Notes