

---

**Tenth Meeting of New Researchers  
in Statistics and Probability**

University of Utah

Salt Lake City

July 24 - July 28, 2007

---

## Welcome

Hello and welcome to the tenth meeting of New Researchers in Statistics and Probability! The purpose of this conference is to provide a comfortable setting for new researchers to share their research and make connections with their peers, as well as provide an opportunity to interact with the invited senior participants in a less formal setting. The conference is kept relatively small so we hope you will get to meet most, if not all, of your fellow participants. We are very excited about our technical and social program, and we hope you enjoy your time at the University of Utah!

## Conference Highlights

The structure of the conference is built around short presentations by the participants. There will be around 3 sessions per day of four to five 12 minute talks, with ample breaks for meals and informal discussions. On Friday evening, just before the conference BBQ, we will have a poster session. Please stop by and interact with the poster presenters.

We are looking forward to several exciting talks by our invited speakers Michael Newton (after the Conference Dinner on Thursday), David Banks (Lunch speaker on Friday), and Deborah Nolan (Saturday morning). We present two panel sessions: the first, on Thursday, comprising editors of major statistical journals, will provide tips and tricks to maximize the probability of getting papers accepted; while the second, on Friday, will have program directors of funding agencies giving advice on how to get funding for your research.

Finally, we are extremely pleased to have Samuel Kou, from the Department of Statistics at Harvard University, give the “Tweedie New Researcher Invited Lecture” at this conference. Richard Tweedie played a significant role throughout his professional career in mentoring young colleagues at work and through professional society activities. Funds donated by his friends and family cover travel for the award winner. We hope you enjoy the meeting!

*Mayetri Gupta*, Chair of the IMS New Researchers’ Committee  
gupta@bios.unc.edu

*Xiaoming Sheng*, Chair of the Tenth Meeting of New Researchers  
Xiaoming.Sheng@hsc.utah.edu

## **A brief history of the NRC**

The first NRC was held in Berkeley, CA in 1993. Many of the characteristics of the conference were initialized at the first meeting— it was held prior to a much larger meeting (JSM), it had a small number of participants (49), and it had esteemed invited speakers (Iain Johnstone and Terry Speed). This first meeting was such a success that another one was planned for 1995, in Kingston, Ontario. This every-two-year pattern continued for the next 8 years, with NRCs being held in Laramie, Wyoming (1997), Baltimore (1999), Atlanta (2001), and Davis, California (2003). One important element of the NRC is the requirement that each participant present their work, in either a short presentation or a poster. Presentations, combined with the small conference size, allows there to be maximal intellectual and social interactions. However, by 2003, the conference had grown so much in popularity that it was impossible to accommodate all applicants, and still hold to the criterion that everyone presents their work; the decision was made to hold the conference every year (Toronto (2004)). Last year's meeting was at Seattle, and next year it will be in Denver. The current plan is to hold the NRC every year prior to JSM, assuming the demand for the conference continues.

## **Acknowledgments**

We thank the following institutions for their generous support and funding of this conference.

- Institute of Mathematical Statistics
- US National Science Foundation (NSF)
- National Institutes of Health (NIH) /National Cancer Institute (NCI)
- National Security Agency (NSA)
- US Office of Naval Research (ONR)
- Department of Family and Preventive Medicine, University of Utah
- Department of Biostatistics, The University of North Carolina at Chapel Hill

### Conference Schedule

	Wednesday (July 25)	Thursday (July 26)	Friday (July 27)	Saturday (July 28)
8:45– 9:00	Introductory Remarks	Introductory Remarks	Introductory Remarks	Introductory Remarks
9:00– 10:15	<b>Session 1:</b> Stochastic modeling and space-time processes	<b>Session 4:</b> Data mining, clustering, and classification	<b>Session 7:</b> Statistical computing and Bayesian methods	Teaching Speaker
10:15– 10:30	Break	Break	Break	Closing remarks
10:30– 11:45	<b>Session 2:</b> Semiparametric and nonparametric inference	Journal Panel [11:15-12:30]	<b>Session 8:</b> Analysis of censored data	
11:45– 2:00	Lunch	Lunch	Invited Speaker Luncheon	
2:00– 3:15	<b>Session 3:</b> Classification/Experimental Design/Causal inference	<b>Session 5:</b> Nonparametric regression and smoothing methods	Grants Panel	
3:15– 3:30	Break	Break	Break	
3:30– 4:45	Tweedie Speaker	<b>Session 6:</b> High dimensional data in biomedical applications	Poster session	
5:00– 7:00	Dinner in dorms (followed by Pub event)	Break	Break	
7:00– 10:00		Invited Speaker Dinner Banquet	BBQ on campus (6:00-9:00)	

Notes:

- **The conference begins Tuesday, July 24, at 6 PM, with an Opening Mixer**
- All talks (except the Conference Dinner) will be in the Health Science Education Building (HSEB)
- **Each talk by participants is 12 minutes in length with 3 minutes for questions.**
- For the pub event, we will meet in front of the Heritage Center at 7.30 PM on the 25th and walk over from there.

## Important information for participants

The conference will begin on Tuesday, 24 July 2007, with an informal mixer from 6pm to 9pm. The main conference sessions begin around 8.45 am on Wednesday 25 July, 2007. The conference ends late in the morning of Saturday 28 July, 2007.

Participants will be staying in the dorms (Sage Point buildings 810-814) at the University of Utah. Accommodation will be fully covered for attendees through grant support.

Participants can check-in on the 24th at the Heritage Center (check-in time is 2 pm, check-out is 12 pm on Saturday, July 28). Please ask about a guest computing account at time of check-in. If you plan to attend JSM, you will need hotel reservations (not paid for by NRC) from Saturday July 28th onwards.

Travel expenses will be **partially reimbursed** by conference grant support. In order to receive reimbursement, participants **must send in all receipts AND boarding passes** after the conference.

(IMPORTANT: U.S. granting agencies prefer any travel to be on U.S. Carriers – to help out please try to book your travel using an U.S. carrier.)

You will find a draft copy of the program and abstracts, as well as other conference-related information on the New Researchers Website at

<http://www.bios.unc.edu/~gupta/NRC/>

## Event locations

Opening mixer: “Commander’s House” on the University of Utah campus

Talks & poster session: HSEB Alumni Hall (2100)

Speaker Luncheon: HSEB Alumni Hall (2100)

Speaker Dinner: Marriott Hotel Research Park

BBQ: outside Heritage Center

## Technical sessions

Note: HSEB  $\equiv$  Health Science Education Building Alumni Hall (2100)

---

**Wednesday, July 25, 2007 (morning)**

---

### Session 1: Stochastic modeling and space-time processes

9:00-10:15 am. Location: HSEB. Chair: Mayetri Gupta (UNC, Chapel Hill)

**Alejandro Veen** (IBM T.J. Watson Research Center)

Point process model assessment using the Mean K-function of Thinned Residuals

**Bo Li** (National Center for Atmospheric Research)

A Nonparametric Assessment of Properties of Space-Time Covariance Functions

**Kevin J. Ross** (Stanford University)

Optimal stopping and free boundary characterizations for some Brownian control problems

**Xiaoqian Sun** (Clemson University)

Bayesian Spatial Model with Application to the Site Index Analysis

**Yuval Nardi** (Carnegie Mellon University)

Maxima of empirical (asymptotically Gaussian) random fields

### Session 2: Semiparametric and nonparametric inference

10:30-11:45 am. Location: HSEB. Chair: Xiaoming Sheng (University of Utah)

**Guang Cheng** (Duke University)

Semiparametric Additive Isotonic Regression

**Huixia Judy Wang** (North Carolina State University)

Quantile Rank Score Tests for Linear Mixed Effects Models

**Tao Huang** (University of Virginia)

Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function

**Weixing Song** (Kansas State University)

Model Checking in Errors-in-Variables Regression

**Xiao Wang** (University of Maryland Baltimore County)

Semiparametric Likelihood Inference of Nonhomogeneous Wiener Process

---

**Wednesday, July 25, 2007 (afternoon)**

---

**Session 3: Classification/Design of Experiments/Causal inference**

*2:00-3:15 pm. Location: HSEB. Chair: Ying Guo (Emory University)*

**Jing Cheng** (University of Florida)

[Efficient nonparametric estimation of causal effects in randomized trials with noncompliance](#)

**Radu Herbei** (Ohio State University)

[Classification with reject option](#)

**Zhiguang (Peter) Qian** (University of Wisconsin-Madison)

[Nested Space-Filling Designs for Experiments with Multiple Levels of Accuracy](#)

**Yan Yang** (Arizona State University)

[Marginal Mixture Analysis of Correlated Bound-Inflated Data](#)

**Tweedie Award session**

*3:30-4:45 pm. Location: HSEB. Chair: Mayetri Gupta (University of North Carolina at Chapel Hill)*

[Stochastic modeling and inference in nano-scale biophysics](#)

**S. C. Samuel Kou**, John L. Loeb Associate Professor of the Natural Sciences, Department of Statistics, Harvard University

Abstract. Recent advances in nanotechnology allow scientists to follow a biological process on a single-molecule basis. These advances also raise many challenging stochastic modeling and statistical inference problems. First, by zooming in on single molecules, recent nano-scale experiments reveal that some classical stochastic models derived from oversimplified assumptions are no longer valid. Second, the stochastic nature of the experimental data and the presence of latent processes much complicate the statistical inference. In this talk we will use the modeling of subdiffusion phenomenon in enzymatic conformational fluctuation and the inference of DNA hairpin kinetics to illustrate the statistical and probabilistic challenges in single-molecule biophysics.

---

**Thursday, July 26, 2007 (morning)**

---

**Session 4: Data mining, clustering, and classification**

*9:00-10:15 am. Location: HSEB. Chair: Joe Blitzstein (Harvard University)*

**Bin Li** (Louisiana State University)

[Additive Regression Trees and Smoothing Splines - Predictive Modeling and Interpretation in Data Mining](#)

**Guangzhe Fan** (University of Waterloo)

[Kernel-Induced Classification Trees and Random Forests](#)

**Lisha Chen** (Yale University)

[Predicting multiple responses with iterative regression](#)

**Samiran Ghosh** (Indiana University Purdue University Indianapolis)

[Scalable Regularized K-Means Clustering with Probabilistic Support for High Dimensional Data](#)

**Yujun Wu** (University of Medicine and Dentistry of New Jersey)

[Controlling Variable Selection By the Addition of Pseudo Variables](#)

**Journal Panel**

*10:30-11:45 am. Location: HSEB. Chair: Xiaoming Sheng (University of Utah)*

**Susan Murphy** (Co-editor, The Annals of Statistics)

**David Banks** (Coordinating Editor and A&CS Editor, Journal of the American Statistical Association)

**David van Dyk** (Editor, Journal of Computational and Graphical Statistics)

**Geert Molenberghs** (Co-editor, Biometrics)

---

**Thursday, July 26, 2007 (afternoon)**

---

**Session 5: Nonparametric regression and smoothing methods**

*2:00-3:15 pm. Location: HSEB. Chair: Kevin Ross (Stanford University)*

**Jiguo Cao** (Yale University)

Generalized Profiling Method and the Applications to Adaptive Penalized Smoothing, Generalized Semiparametric Additive Models and Estimating Differential Equations

**Jing Wang** (University of Illinois at Chicago)

Spline confidence band with mixing error in fixed design

**Timothy L. McMurry** (DePaul University)

Nonparametric regression with automatic bias correction

**Xuerong Meggie Wen** (University of Missouri Rolla)

The generalized inverse k-means regression

**Yehua Li** (University of Georgia)

Functional Data Analysis in Lipoprotein Profile Data

**Session 6: High dimensional data in biomedical applications**

*3:30-4:45 pm. Location: HSEB. Chair: Sanjay Chaudhuri (National University of Singapore)*

**Haiyan Wang** (Kansas State University)

Inference on nonparametric hypothesis testing and its application in microarray and mass spectrometric data

**Qing Zhou** (University of California, Los Angeles)

Statistical Learning on ChIP-chip Data by Bayesian Additive Regression Trees

**Wei Chen** (Wayne State University)

A False Discovery Rate Based Loss Framework for Selection of Interactions

**Ying Guo** (Emory University)

Methods of Group Independent Component Analysis for Multisubject fMRI Data

---

**Thursday, July 26, 2007 (evening)**

---

**Conference dinner**

*7:00-10:30 pm. Location: Marriott Hotel Research Park. Chair: Mayetri Gupta (University of North Carolina at Chapel Hill)*

[Buried treasures: examples of old statistics in new contexts](#)

**Michael Newton**, Professor of Statistics and Biostatistics, University of Wisconsin at Madison.

Abstract. Isaac Newton's hackneyed attribution of his success to "standing on the shoulders of giants" reminds us that our ever-developing field of statistics is much indebted to great thinkers from the past. Though we learn in school about the most important of these scientists, I am intrigued by a particular form of the "past effect" which seems to have recurred in my own work, and which I shall attempt to articulate in this talk. It is simply that in the context of an extant statistical problem (be it applied, methodological, or theoretical), the solution is found by realizing a new interpretation of an uncelebrated but beautiful statistical result from the past. I expect there are numerous examples; I will discuss a few.

---

**Friday, July 27, 2007 (morning)**

---

**Session 7: Statistical computing and Bayesian methods**

*9:00-10:15 am. Location: HSEB. Chair: Qing Zhou (University of California, Los Angeles)*

**Fei Liu** (Duke University)

[Bayesian Functional Data Analysis for Complex Computer Model Validation](#)

**Joseph Blitzstein** (Harvard University)

[Sequential Importance Sampling vs. MCMC for Discrete Structures](#)

**Leming Qu** (Boise State University)

[Bayesian Wavelet Estimation of Copulas for Dependence Modeling](#)

**Taeyoung Park** (University of Pittsburgh)

[Using Incompatibility to Build Fast Gibbs Samplers](#)

**Session 8: Analysis of censored data**

*10:30-11:45 am. Location: HSEB. Chair: Lisha Chen (Yale University)*

**Jialiang Li** (National University of Singapore)

[ROC Analysis with Multiple Classes and Multiple Tests](#)

**Pang Du** (Virginia Polytechnic Institute and State University)

[Nonparametric Hazard Estimation for Gap Time in Recurrent Event Data](#)

**Yu Cheng** (University of Pittsburgh)

[Nonparametric analysis of multivariate competing risks data](#)

**Zhigang Zhang** (Oklahoma State University)

[A class of linear transformation models under interval censoring](#)

---

**Friday, July 27, 2007 (afternoon)**

---

### **Invited speaker Luncheon**

*12:00-2:00 pm. Location: HSEB. Chair: Xiaoming Sheng (University of Utah)*

[Snakes & Ladders: Building a Career in Statistics](#)

**David Banks**, Professor of the Practice of Statistics, Duke University

Abstract. Many years ago, I served on the IMS New Researchers Committee and helped write the first edition of the “New Researcher’s Survival Guide”. Looking back, I realize I have learned a lot more since then about career strategies in statistics. These can make a big difference in improving the chances of achieving your professional goals, no matter which professional path you choose. So, with the benefit of hindsight and engagement in many different kinds of jobs and professional activities, I shall provide an oral update to the Survival Guide.

### **Grants Panel**

*2:00-3:15 pm. Location: HSEB. Chair: Mayetri Gupta (University of North Carolina at Chapel Hill)*

**Grace Yang** (Probability & Statistics Program, National Science Foundation)

**Wendy Martinez** (Office of Naval Research)

**Ram Tiwari** (National Cancer Institute, National Institutes of Health)

### **Session 9: Poster session**

*3:30-4:45 pm. Location: HSEB.*

**Sanjay Chaudhuri** (National University of Singapore)

[On qualitative comparisons of dependence between d-connected vertices of a singly connected Gaussian DAG](#)

**Dongfeng Li** (Fred Hutchinson Cancer Research Center)

[Mixed-Effects Model for Periodic Functional Data with Multiple Cycles](#)

**Yan Li** (University of California, San Diego)

[Mediators of the Association between the Apolipoprotein E4 Allele and Cognitive Function](#)

**Elena Perez-Bernabeu** (Universidad Politecnica de Valencia, Spain)

[Attribute Control Charts: improvement of the u-chart applying J.J. Daudin’s methodology](#)

**Hongkun Wang** (University of Virginia)

[A study on confidence intervals for incremental cost-effectiveness ratios with censored data](#)

**Jianan Peng** (Acadia University)

[Stepwise Confidence Intervals for Monotone Dose-Response Studies](#)

**Qingzhao Yu** (Louisiana State University)

[Bayesian Synthesis](#)

---

**Saturday, July 28, 2007 (morning)**

---

### **Session on Teaching**

*9:00-10:15 am. Location: HSEB. Chair: Xiaoming Sheng (University of Utah)*

#### [The Why and How of Teaching Statistics](#)

**Deborah Nolan**, Professor of Statistics, University of California at Berkeley.

The aim of this talk is two fold. I will first focus on what we teach our students, and then turn to how we teach them. The teaching of statistics, especially in advanced undergraduate courses, typically does not resemble the current practice of statistics. Too often, theory comes before practice, statistics is confused with mathematics, computing is relegated a cook book instruction set, and modern methods are barely mentioned. I hope to encourage you and challenge you to break away from this tradition and bring the teaching of statistics into the modern era. I will describe recent efforts in this direction, and invite you to join them as you embark on your teaching career. In the second part of my talk, I will provide many practical tips that I have found useful in teaching. These will range from general ideas on how to engage students in learning, to discipline specific ideas for activities that help get basic statistical concepts across, to advice on classroom and course management.

## Abstracts: Participants

### Session 1: Stochastic modeling and space-time processes

#### Alejandro Veen

##### [Point process model assessment using the Mean K-function of Thinned Residuals](#)

Abstract. This work presents an alternative derivation of the asymptotic distribution of Ripley's K-function for a homogeneous Poisson process and shows how it can be combined with point process residual analysis in order to assess the fit of point process models. The asymptotic distribution of the Mean K-function of Thinned Residuals ( $K_M$ ) is presented for inhomogeneous Poisson processes and compared to the asymptotic distribution of a weighted analogue called the weighted or inhomogeneous K-function.  $K_M$  is then applied to California earthquake data in order to assess the goodness-of-fit of seismological point process models.

#### Bo Li

##### [A Nonparametric Assessment of Properties of Space-Time Covariance Functions](#)

Abstract. We propose a unified framework for testing a variety of assumptions commonly made for covariance functions of stationary spatio-temporal random fields. The methodology is based on the asymptotic normality of space-time covariance estimators. We focus on tests for full symmetry and separability in this paper, but our framework naturally covers testing for isotropy and Taylor's hypothesis. Our test successfully detects the asymmetric and nonseparable features in two sets of wind speed data. We perform simulation experiments to evaluate our test and conclude that our method is reliable and powerful for assessing common assumptions on space-time covariance functions.

#### Kevin Ross

##### [Optimal stopping and free boundary characterizations for some Brownian control problems](#)

Abstract. We study a singular stochastic control problem with state constraints in two-dimensions. We show that the value function is continuously differentiable and its directional derivatives are the value functions of certain optimal stopping problems. Guided by the optimal stopping problem we then introduce the associated no-action region and the free boundary and show that, under appropriate conditions, an optimally controlled process is a Brownian motion in the no-action region with reflection at the free boundary. This proves a conjecture of Martins, Shreve and Soner (1996) on the form of an optimal control for this class of singular control problems. An important issue in our analysis is that the running cost is Lipschitz but not continuously differentiable. This lack of smoothness is one of the key obstacles in establishing regularity of the free boundary. We show that the free boundary is Lipschitz and if the Lipschitz constant is sufficiently small, a certain oblique derivative problem on the no-action region admits a unique viscosity solution. This uniqueness result is key in characterizing an optimally controlled process as a reflected diffusion in the no-action region. (Based on joint work with Amarjit Budhiraja.)

#### Xiaoqian Sun

##### [Bayesian Spatial Model with Application to the Site Index Analysis](#)

Abstract. In this talk, we propose a Bayesian spatial model for analyzing the site index dataset from the Missouri Ozark Forest Ecosystem Project (MOFEP). Based on ecological background and availability, we choose three variables: aspect class, land type association and soil depth as covariates. To allow greater flexibility of the smoothness

of the random field, we use the Matern family as correlation function. Due to the fact that there is no previous knowledge of the parameters in the model, we choose a non-informative prior for analysis. A new efficient algorithm based on the generalized Ratio-of-Uniforms method is proposed for the posterior simulation. The main advantage of our algorithm is that it will generate independent samples from the joint posterior distribution, a key difference from other MCMC simulation algorithms. Prediction of the site index at unmeasured locations is then easily implemented based on independent samples generated from the joint posterior distribution. Our results show that aspect class and soil depth are both significant while land type association is less significant. Finally we discuss some future work.

**Yuval Nardi**

[Maxima of empirical \(asymptotically Gaussian\) random fields](#)

Abstract. The distribution of the maxima of an empirical random field, which is asymptotically Gaussian, is investigated. We develop an asymptotic expansion for the probability that the maxima of such field exceeds a high threshold. Our method is a direct probabilistic one, which involves a change of measure, and several localization arguments. The expansion relates the terms in the expansion to both the sample size and the threshold level.

## **Session 2: Semiparametric and nonparametric inference**

**Guang Cheng**

[Semiparametric Additive Isotonic Regression](#)

Abstract. This paper is about the efficient estimation of semiparametric additive isotonic regression model, i.e.  $Y = X'\beta + \sum_{j=1}^J h_j(W_j) + \varepsilon$ . Each additive component  $h_j$  is assumed to be a monotone function. It is shown that the least square estimator of the parametric component is asymptotically normal. Moreover, the isotonic estimator for each additive functional component is proved to have the oracle property, which means it can be estimated with the highest asymptotic accuracy, equivalently, as if the other components were known.

**Huixia Judy Wang**

[Quantile Rank Score Tests for Linear Mixed Effects Models](#)

Abstract. In this talk, I will introduce a rank score test for linear quantile models with a random effect. The developed rank score test does not impose any strict parametric assumptions on the error distributions, and is robust to outliers. The test is shown to be a very valuable complement to the usual mixed model analysis based on Gaussian likelihood. The proposed approach is applied to GeneChip microarray studies for detecting differentially expressed genes by modeling and analyzing the quantiles of gene expressions through probe level measurements. I will focus on an enhanced quantile rank score test, which aims to improve the efficiency of the quantile rank score test at small samples through borrowing information across genes. In addition, I will discuss a rank score test for censored quantile regression models with correlated data. A simulation study shows that the proposed test performs almost as well as the omniscient test based on the latent response variable, and it outperforms the naive method that simply ignores censoring. The performance of the proposed test is assessed by analyzing the annual salary of U.S. adult men. The inference on censored quantile regression is an interesting problem with wide applications to biological, clinical and social science studies.

**Tao Huang**

[Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function](#)

Abstract. Improving efficiency for regression coefficients and predicting trajectories of individuals are two important aspects in analysis of longitudinal data. Both involve estimation of the covariance function. Yet, challenges arise in estimating the covariance function of longitudinal data collected at irregular time points. A class of semiparametric models for the covariance function is proposed by imposing a parametric correlation structure while allowing a nonparametric variance function. A kernel estimator is developed for the estimation of the nonparametric variance function. Two methods, a quasi-likelihood approach and a minimum generalized variance method, are proposed for estimating parameters in the correlation structure. We introduce a semiparametric varying coefficient partially linear model for longitudinal data and propose an estimation procedure for model coefficients by using a profile weighted least squares approach. Sampling properties of the proposed estimation procedures are studied and asymptotic normality of the resulting estimators is established. Finite sample performance of the proposed procedures is assessed by Monte Carlo simulation studies. The proposed methodology is illustrated by an analysis of a real data example.

**Weixing Song**

[Model Checking in Errors-in-Variables Regression](#)

Abstract. This paper discusses a class of minimum distance tests for fitting a parametric regression model to a class of regression functions in the errors-in-variables model. These tests are based on certain minimized distances between a nonparametric regression function estimator and a deconvolution kernel estimator of the conditional expectation of the parametric model being fitted. The paper establishes the asymptotic normality of the proposed test statistics under the null hypothesis and that of the corresponding minimum distance estimators. The significant contribution made in this research is the removal of the common assumption in the literature that the density function of the design variable is known. A simulation study shows that the testing procedures are quite satisfactory in the preservation of the finite sample level and in terms of a power comparison.

**Xiao Wang**

[Semiparametric Likelihood Inference of Nonhomogeneous Wiener Process](#)

Abstract. Many systems degrade over time. Degradation data are a rich source of reliability information and offer many advantages over failure time data. We propose a class of models based on nonhomogeneous Wiener process to analyze degradation data. Random effects can also be incorporated into the process to represent heterogeneity in the degradation paths for different subjects. This model is flexible, accommodates a variety of degradation shapes, and has a tractable form for the time-to-failure distribution. Semiparametric methods are developed to estimate the unknown parameters and asymptotic properties of these estimators, such as consistency, convergence rate and asymptotic distributions, will also be established. Simulation study is conducted to validate the method and we illustrate the method by real degradation data such as fatigue crack growth data and bridge beam data.

**Session 3: Classification/Design of Experiments/Causal inference**

**Jing Cheng**

[Efficient nonparametric estimation of causal effects in randomized trials with noncompliance](#)

Abstract. Causal approaches based on the potential outcome framework provide a useful tool for addressing the non-compliance problems in randomized trials. Various estimators, e.g. the instrumental variable (IV) estimator, have

been proposed for causal effects of treatment. In this paper, we propose a new empirical likelihood-based estimator of causal treatment effects in randomized clinical trials with noncompliance. By using the empirical likelihood approach to construct a profile random sieve likelihood and taking into account the mixture structure in outcome distributions, this estimator is robust to parametric distribution assumptions and more efficient than the standard IV estimator. Our method is applied to data from a randomized trial of an encouragement intervention to improve adherence to prescribed depression treatments among depressed elderly patients in primary care practices.

## **Radu Herbei**

### [Classification with reject option](#)

Abstract. Under the classic binary supervised learning framework, the statistician's task is to form a classifier which represents his/her guess of the label of a future observation. However, allowing for the reject option besides taking a hard decision (0 or 1) is of great importance in practice. In a medical setting, for instance, when classifying whether a disease is present or absent, the reject option is sometimes preferred and leads to making the right decision after additional data is collected. We study binary classification that allows for a reject option, in which case no decision is made. This reject option is to be used for those observations for which the conditional class probabilities are close, and as such are hard to classify. We generalize existing theory for both plug-in rules and empirical risk minimizers to this setting. We also extend our approach to the situation when the type I and type II errors have distinct costs.

## **Zhiguang Peter Qian**

### [Nested Space-Filling Designs for Experiments with Multiple Levels of Accuracy](#)

Abstract. Computer experiments with different levels of accuracy have become prevalent in many engineering and scientific applications. Design construction for such experiments is a new issue because traditional methods deal almost exclusively with experiments with one level of accuracy. Some nested space-filling designs are proposed for these experiments. The proposed designs possess some nested structure and space-filling property. The construction is aided by using Galois fields and exploiting nested structure in orthogonal arrays. This is joint work with Boxin Tang at Simon Fraser University and C. F. Jeff Wu at Georgia Tech.

## **Yan Yang**

### [Marginal Mixture Analysis of Correlated Bound-Inflated Data](#)

Abstract. Data with bound-inflated responses are common in many areas of application. Often the data are bounded below by zero with excess zero observations, so ordinary generalized linear models fail. Three methods in the literature for modeling zero-inflated data are left censored regression models, two-part models, and latent mixture models. We develop a general class of bound-inflated mixture models that unifies the three existing methods and extends the zero-inflated exponential family models by Hall and Zhang (2004, *Statistical Modelling*, 4, 161-180) to non-exponential families. A marginal approach is proposed to correlated bound-inflated measures by adopting the theory of generalized estimating equations. Quantitative risk assessment is investigated for bounded adverse outcomes based on the developed mixture models. We illustrate the issues and methodology in the context of an ultrasound safety study of the occurrence and extent of lung hemorrhage in laboratory animals due to focused ultrasound exposure.

## Session 4: Data mining, clustering, and classification

### Bin Li

#### [Additive Regression Trees and Smoothing Splines - Predictive Modeling and Interpretation in Data Mining](#)

Abstract. We suggest a two-phase boosting method, called “additive regression tree and smoothing splines” (ARTSS), which is highly competitive in prediction performance. However, unlike many automated learning procedures, which lack interpretability and operate as a “black box”, ARTSS allows us to (1) estimate the marginal effect smoothly; (2) test the significance of non-additive effects; (3) separate the variable importance on main and interaction effects; (4) select variables and provide a more flexible modeling strategy. We apply ARTSS to two public domain data sets and discuss the understanding developed from the model.

### Guangzhe Fan

#### [Kernel-Induced Classification Trees and Random Forests](#)

Abstract. Motivated by the success of support vector machine (SVM), a recursive-partitioning procedure using kernel functions is proposed for classification problems. We call it KICT-kernel-induced classification trees. Essentially, KICT uses kernel functions to construct CART models. The resulting model could perform significantly better in classification than the original CART model in many situations, especially when the pattern of the data is non-linear. We also introduce KIRF: kernel-induced random forests. KIRF compares favorably to random forests and SVM in many situations. KICT and KIRF also largely retain the computational advantage of CART and random forests, respectively, in contrast to SVM. We use simulated and real world data to illustrate their performances. We conclude that the proposed methods are useful alternatives and competitors to CART, random forests, and SVM.

### Lisha Chen

#### [Predicting multiple responses with iterative regression](#)

Abstract. We consider the variable selection problem in predicting multiple responses from a common set of predictors. Very often the multiple responses are correlated, and it is natural to ask how to borrow strength among responses to improve prediction as compared to doing separate individual regressions. In this paper, we introduce an iterative regression procedure in which the useful predictors found by one response variable in a current step are used in predicting the other response variables in the next step. This procedure can significantly improve the predictive accuracy when the multiple responses are related to the predictors in shared patterns. The power of this algorithm is demonstrated with both simulated and real data examples.

### Samiran Ghosh

#### [Scalable Regularized K-Means Clustering with Probabilistic Support for High Dimensional Data](#)

Abstract. Over the last decade a variety of clustering algorithms have evolved. However one of the simplest (and possibly overused) partition based clustering algorithm is K means. It can be shown that the computational complexity of K-means does not suffer from exponential growth with dimensionality rather it is linearly proportional with the number of observations and number of clusters. The crucial requirements are the knowledge of cluster number and the computation of some suitably chosen similarity measure. For this simplicity and scalability among large data sets, K-means remains an attractive alternative when compared to other competing clustering philosophies especially for high dimensional domain. However being a deterministic algorithm, traditional K-means have several drawbacks. It only offers hard decision rule, with no probabilistic interpretation. In this paper we have developed a

decision theoretic framework by which traditional K-means can be given a probabilistic footstep. This will not only enable us to do a soft clustering, rather the whole optimization problem could be recasted into Bayesian modeling framework, in which the knowledge of cluster number could be treated as an unknown parameter of interest, thus removing a severe constrain of K-means algorithm. Our basic idea is to keep the simplicity and scalability of K-means, while achieving some of the desired properties of the other model based or soft clustering approaches.

**Yujun Wu**

#### Controlling Variable Selection By the Addition of Pseudo Variables

Abstract. We propose a new approach to variable selection designed to control the *false selection rate* (FSR), defined as the proportion of uninformative variables included in selected models. The method works by adding a known number of pseudo variables to the real data set, running a variable selection procedure, and monitoring the proportion of pseudo variables falsely selected. Information obtained from bootstrap-like replications of this process is used to estimate the proportion of falsely-selected real variables and to tune the selection procedure to control the false selection rate.

### Session 5: Nonparametric regression and smoothing methods

**Jiguo Cao**

#### Generalized Profiling Method and the Applications to Adaptive Penalized Smoothing, Generalized Semiparametric Additive Models and Estimating Differential Equations

Abstract. Many statistical models involve three distinct groups of variables: local or nuisance parameters, global or structural parameters, and complexity parameters. In this thesis, we introduce the generalized profiling method to estimate these statistical models, which treats one group of parameters as an explicit or implicit function of other parameters. The dimensionality of the parameter space is reduced, and the optimization surface becomes smoother. The Newton-Raphson algorithm is applied to estimate these three distinct groups of parameters in three levels of optimization, with the gradients and Hessian matrices written out analytically by the Implicit Function Theorem if necessary and allowing for different criteria for each level of optimization. Moreover, variances of global parameters are estimated by the Delta method and include the variation coming from complexity parameters. We also propose three applications of the generalized profiling method.

First, penalized smoothing is extended by allowing for a functional smoothing parameter, which is adaptive to the geometry of the underlying curve, which is called *adaptive penalized smoothing*. In the first level of optimization, the smoothing coefficients are local parameters, estimated by minimizing sum of squared errors, conditional on the functional smoothing parameter. In the second level, the functional smoothing parameter is a complexity parameter, estimated by minimizing generalized cross-validation (GCV), treating the smoothing coefficients as explicit functions of the functional smoothing parameter. Adaptive penalized smoothing is shown to obtain better estimates for fitting functions and their derivatives.

Next, the generalized semiparametric additive models are estimated by three levels of optimization, allowing response variables in any kind of distribution. In the first level, the nonparametric functional parameters are nuisance parameters, estimated by maximizing the regularized likelihood function, conditional on the linear coefficients and the smoothing parameter. In the second level, the linear coefficients are structural parameters, estimated by maximizing the likelihood function with the nonparametric functional parameters treated as implicit functions of linear coefficients and the smoothing parameter. In the third level, the smoothing parameter is a complexity parameter,

estimated by minimizing the approximated GCV with the linear coefficients treated as implicit functions of the smoothing parameter. This method is applied to estimate the generalized semiparametric additive model for the effect of air pollution on the public health.

Finally, parameters in differential equations (DE's) are estimated from noisy data with the generalized profiling method. In the first level of optimization, fitting functions are estimated to approximate DE solutions by penalized smoothing with the penalty term defined by DE's, fixing values of DE parameters. In the second level of optimization, DE parameters are estimated by weighted sum of squared errors, with the smoothing coefficients treated as an implicit function of DE parameters. The effects of the smoothing parameter on DE parameter estimates are explored and the optimization criteria for smoothing parameter selection are discussed. The method is applied to fit the predator-prey dynamic model to biological data, to estimate DE parameters in the HIV dynamic model from clinical trials, and to explore dynamic models for thermal decomposition of alpha-Pinene.

### **Jing Wang**

#### [Spline confidence band with mixing error in fixed design](#)

Abstract. Base on polynomial spline regression, the asymptotically conservative confidence bands of the the strong mixing time series are proposed in the cases of fixed design ( equally-spaced design). The explicit form is provided in the paper. Simulation experiments provide strong evidence that collaborates with the asymptotic theory. The result is applied to the Leaf Area Index data in East Africa to test the trigonometric curve assumption on the Leaf Area Index data which is collected regularly over years to improve the Regional Climate Model System.

### **Timothy McMurry**

#### [Nonparametric regression with automatic bias correction](#)

Abstract. The problem of estimating nonparametric regression with associated confidence intervals will be addressed. It is shown that through appropriate choice of infinite order kernel, it is possible to construct a nonparametric regression estimator which has favorable asymptotic performance (bias, variance, and mean squared error). The proposed class of kernels is characterized by a Fourier transform which is flat near the origin and infinitely differentiable. This property allows the bias of the estimate to decrease at the maximal rate without harming the rate at which the variance decreases, thus leading to a faster rate of convergence. An additional benefit of using these kernels is that it becomes possible to construct bootstrap confidence intervals which do not require either explicit bias correction or suboptimal levels of smoothing at any stage in the estimation. In particular, it is demonstrated that in this setting, consistent confidence intervals are obtained when both the pilot and final smoothings are estimated at the mean square error optimal bandwidth for estimating the regression.

### **Xuerong Meggie Wen**

#### [The generalized inverse k-means regression](#)

Abstract. Few methodologies are available for estimating the central subspace for regressions with multivariate responses due to the difficulties arising from slicing multivariate responses. Setodji and Cook (2004) introduced a new way of performing the slicing. They developed a method called *k-means inverse regression* (KIR), which makes use of the *k-means* algorithm to cluster the observed response vectors. However, their method ingored the intra-cluster information which could be substantial under some circumstances. In this paper, we proposed an improved method by incorporating the intra-cluster information into estimation. Our method outperformed KIR with respect to estimation accuracies of both the central subspace and its dimension. It also allows us to test the predictor effects in a

model-free approach.

**Yehua Li**

#### [Functional Data Analysis in Lipoprotein Profile Data](#)

Abstract. Lipoprotein concentration in human serum is an important risk factor for cardiovascular heart disease. Different species of lipoprotein in a serum sample can be separated by a centrifugation treatment, according to their density differences. A lipoprotein profile for a patient is obtained by taking the image of his centrifuged serum sample, after an application of a lipophilic stain. In this paper, we use functional data analysis techniques to study the lipoprotein profile curves. The goal is to predict the quantity of total cholesterol and each species of lipoprotein from the profile curves. We discuss implementation issues including smoothing the profile curves with P-spline using the Kalman filter technique, functional principal component analysis and functional linear regression. We also study a functional projection pursuit model, which bears the simple functional linear model as a reduced model. A bootstrap method is proposed to test the adequacy of a functional linear model. A simulation study for the proposed method is also presented.

### **Session 6: High dimensional data in biomedical applications**

**Haiyan Wang**

#### [Inference on nonparametric hypothesis testing and its application in microarray and mass spectrometric data](#)

Abstract. Studies for high dimensional data have received a lot of attention recently. In this talk I will present hypotheses testing methods developed for heteroscedastic high dimensional data. Asymptotic distributions of test statistics are obtained for both independent and dependent functional data under the high dimensional low sample size setting. Results based on the original observations and (mid-)ranks are presented separately. Simulation studies reveal that the (mid-)rank procedures outperform those based on the original observations in all non-normal situations while they do not lose much power when normality holds. Two clustering algorithms are developed using results of above hypotheses testings. They have been successfully applied to data collected from microarray experiments to effectively identify differentially expressed genes and detect patterns and clusters for time course gene expression data. Recently, they are also applied to lipidomics data collected from mass spectrometer to identify effective mutants in high throughput screening. Applications to some data sets will be given and potential extensions in several directions will be discussed.

**Qing Zhou**

#### [Statistical Learning on ChIP-chip Data by Bayesian Additive Regression Trees](#)

Abstract. The ChIP-on-chip technology has generated informative data sets for understanding the interaction of transcription factors (TF) and their DNA binding sites, which provide both binding regions and intensities of the target TFs. We extract sequence features, including background words, motif scores, and cross-species conservation, from the binding regions, and utilize the Bayesian additive regression trees (BART, Chipman et al. 2006) to learn the relation between TF-DNA binding intensity and the extracted sequence features. The developed approach is applied to two recently published ChIP-chip data sets of the TFs Oct4 and Sox2 in human embryonic stem cells. Cross validations show that our approach outperforms all the other competing methods based on linear regressions, neural networks, or MARS. The study has demonstrated that (1) the use of background features substantially improves the predictive power of statistical learning, and (2) BART is a robust and flexible method which does not overfit training

data. Interesting biological insights are discussed based on the features selected by BART.

**Wei Chen**

#### [A False Discovery Rate Based Loss Framework for Selection of Interactions](#)

Abstract. Interaction effects have been consistently found important in explaining the variation of outcomes in many scientific research fields. Yet, in practice, variable selection including interactions is complicated due to the limited sample size, conflicting philosophies regarding model interpretability, and accompanying amplified multiple testing problems. Lack of statistically sound algorithms for automatic variable selection with interactions has discouraged activities in exploring important interaction effects. In this article, we investigated issues of selecting interactions from three aspects: (1) What is the model space to be searched? (2) How is the hypothesis-testing performed? (3) How to address the multiple testing issue? We propose loss functions and corresponding decision rules that control FDR in Bayesian context. Properties of the decision rules are discussed and their performance in terms of the power and FDR are compared through simulations. Methods are illustrated on data from a colorectal cancer study assessing the chemotherapy treatments and data from a diffuse large-B-cell lymphoma study assessing the prognostic effect of gene expressions.

**Ying Guo**

#### [Methods of Group Independent Component Analysis for Multisubject fMRI Data](#)

Abstract. Independent component analysis (ICA) is becoming increasingly popular for analyzing functional magnetic resonance imaging (fMRI) data. ICA has been successfully applied for single-subject fMRI analysis. The extension of ICA for group inferences is not straightforward and remains an active research topic. Two promising group ICA methods that have been proposed are the tensor probabilistic ICA (Beckmann and Smith, 2005. NeuroImage, 25: 294-311) and the group ICA approach by Calhoun et al. (2001. Hum Brain Mapp, 14: 140-151). It is important to investigate the relative effectiveness of the two methods to help researchers choose a more appropriate method for a particular data set. We conduct a simulation study to evaluate the performance of these two group ICA approaches under various scenarios. The results show that both group ICA approaches perform best for a trilinear model where the multisubject fMRI data can be decomposed into the outer product of loadings in spatial, temporal and subject domains. When between-subject variability becomes more complicated, each of the two methods performs well in certain scenarios but neither is consistently better than the other. We also propose two hypothesis tests for comparing group differences in the independent components. We evaluate the performance of the tests through simulation studies of group fMRI data.

### **Session 7: Statistical computing and Bayesian methods**

**Fei Liu**

#### [Bayesian Functional Data Analysis for Complex Computer Model Validation](#)

Abstract. Functional data analysis (FDA) inference on curves or functions has wide application in statistics. An example of considerable recent interest arises when considering computer models of processes; the output of such models is a function over the space of inputs of the computer model. The output is functional data in many contexts, such as when the output is a function of time, a surface, etc. In this research, we develop or extend four Bayesian FDA approaches to computer model validation, tailored to interdisciplinary problems in engineering and the environment. The first approach we consider is a nonparametric Bayesian statistics approach, utilizing separable

Gaussian Stochastic Process as the prior distribution for functions. This is a natural choice for smooth functions with a manageable (time) dimension. The methodology is developed in the context of a thermal computer model challenge problem, proposed by the Sandia National Laboratory. Direct use of separable Gaussian stochastic processes is inadequate for irregular functions, and can be computationally infeasible for high dimensional functions. The approach developed for such functions consists of representing the function in the wavelet domain; reducing the number of nonzero coefficients by thresholding; modeling the nonzero coefficients as functions of the associated inputs, using the nonparametric Bayesian method; and reconstructing the functions (with confidence bands) in the original (time) domain. The third approach extends the second in terms of function representation. We represent the functions in the eigen-space whose basis elements are linear combinations of the wavelet basis elements. The number of nonzero coefficients is greatly reduced in this eigen-space, as consequently is the computational expense for the statistical inverse problem. This method is developed in the context of computer modeling of vehicle suspension systems. The fourth approach models functions as multivariate Dynamic Linear Models. This approach is useful when the functions are highly variable and, as opposed to attempting to represent the functions exactly, one seeks primarily to capture relevant stochastic structure of the functions. The method has been tested with a simulated data set, and will be applied to validate the Community Multi-scale Air Quality model, a computer model for air quality involving factors such as tropospheric ozone, fine particles, and visibility degradation. In addition to the basic issue of functional data, all the above approaches must also contend with three other issues associated with computer model validation. First, emulators must typically be constructed for expensive-to-run computer models, by treating them as spatial processes defined on the input space. Second, computer model bias the discrepancy between the computer model output and reality must be taken into account. Third, the computer models typically have unknown parameters, requiring solution of an inverse problem in their estimation. Because these issues must all be addressed simultaneously and with limited data, extensive use is made of Markov Chain Monte Carlo (MCMC) algorithms. Some modular versions of MCMC are introduced to reduce the confounding between some of the elements in the corresponding statistical models.

### **Joseph Blitzstein**

#### [Sequential Importance Sampling vs. MCMC for Discrete Structures](#)

Abstract. Two popular approaches to estimation problems in large discrete state space are sequential importance sampling (SIS) and Markov Chain Monte Carlo (MCMC). We compare the two methods for some network-related problems, and show how well-constructed Markov chain transitions shed can be used to analyze and improve the efficiency of the importance sampling technique.

### **Leming Qu**

#### [Bayesian Wavelet Estimation of Copulas for Dependence Modeling](#)

Abstract. Copulas are full measures of dependence among random variables. It is now widely used in financial engineering for modeling high-dimensional problems, such as value-at-risk or portfolio credit risk. A copular's hidden dependence structure that couples a joint distribution with its marginals makes a parametric copular model non-trivial. We propose a nonparametric estimator using wavelet in the Bayesian framework. A mixture prior of a point mass at zero and a normal distribution is imposed on the wavelet coefficients. The Markov chain Monte Carlo algorithm is used for posterior inference. Performances are evaluated on simulated data and on a real dataset.

## **Taeyoung Park**

### [Using Incompatibility to Build Fast Gibbs Samplers](#)

Abstract. Ever increasing computational power along with ever more sophisticated statistical computing techniques is making it possible to fit ever more complex statistical models. Among the popular, computationally intensive methods, the Gibbs sampler (Geman and Geman, 1984) has been spotlighted because of its simplicity and power to effectively generate samples from a high-dimensional distribution. Despite its simplicity to implement and describe, however, the Gibbs sampler is criticized for its sometimes slow convergence especially when it is used to fit highly structured, complex models. In this talk, we present partially marginalized Gibbs sampling strategies that improve the convergence characteristics of a Gibbs sampler by capitalizing on a set of incompatible conditional distributions. Such incompatibility is generally avoided in the construction of a Gibbs sampler because the resulting convergence properties are not well understood. We, however, introduce three basic tools (marginalization, permutation, and trimming) which allow us to transform a Gibbs sampler into a partially marginalized Gibbs sampler with known stationary distribution and fast convergence.

## **Session 8: Analysis of censored data**

### **Jialiang Li**

#### [ROC Analysis with Multiple Classes and Multiple Tests](#)

Abstract. The accuracy of a single diagnostic test for binary outcome can be summarized by the area under the ROC curve. Volume under the surface and hyper-volume under the manifold have been proposed as extensions for multiple class diagnosis. However, calculating these quantities is not straightforward, even with a single test. The decision rule used to generate the ROC surface requires class probability assessments, which are not provided by the tests. We develop a method based on multinomial logistic regression to combine the tests. Bootstrap inferences are proposed to account for variability in estimating the regression model and perform well in simulations. The ROC measures are compared to the correct classification rate, which depends heavily on class prevalences. An example of tumor classification with microarray data demonstrates that this property may lead to vastly different analyses based on the two diagnostic accuracy measures. The ROC based analysis yields remarkable decreases in model complexity over previous analyses.

### **Pang Du**

#### [Nonparametric Hazard Estimation for Gap Time in Recurrent Event Data](#)

Abstract. Recurrent event data arise in many biomedical and engineering studies when failure events can occur repeatedly over time for each study subject. In this article, we are interested in nonparametric estimation of the hazard function for gap times. A penalized likelihood model is proposed to estimate the hazard as a function of both gap time and covariate. Method for smoothing parameter selection is developed and Bayesian confidence intervals for the hazard function are derived. Asymptotic convergence rates of the estimates are also established by assuming no gap times of a subject are the same. Empirical studies are performed to evaluate various aspects of the method. The exploratory role of the proposed technique is illustrated through an application to the well-known bladder tumor cancer data.

**Yu Cheng**

[Nonparametric analysis of multivariate competing risks data](#)

Abstract. While nonparametric analyses of bivariate failure times have been widely studied, nonparametric analyses of bivariate competing risks data have not been investigated. Such analyses are important in familial association studies, where multiple interacting failure types may invalidate nonparametric analyses for independently censored clustered survival data. We develop nonparametric estimators for the bivariate cause-specific hazards function and the bivariate cumulative incidence function, which are natural extensions of their univariate counterparts and make no assumptions about the dependence of the risks. The estimators are shown to be uniformly consistent and to converge weakly to Gaussian processes. Summary association measures are proposed and yield formal tests of independence in clusters. The estimators and test statistics perform well in simulations with realistic sample sizes. Their practical utility is illustrated in an analysis of dementia in the Cache County Study.

**Zhigang Zhang**

[A class of linear transformation models under interval censoring](#)

Abstract. Regression analysis of interval-censored failure time data has recently attracted a great deal of attention and for the problem, several methods have been proposed. However, most of these methods employ specific models such as the proportional hazards model. The linear transformation models, a class of generalized proportional hazards models, provide great flexibility in fitting survival data. Although the application of these models on right-censored data has been discussed, the only existing inference procedure for the linear transformation models when interval-censored data are observed is limited to multi-sample situations and lacks rigorous technical justification. In this study we present approaches that allow the covariate to be arbitrary and provide a theoretical framework.

**Session 9: Poster session**

**Sanjay Chaudhuri**

[On qualitative comparisons of dependence between d-connected vertices of a singly connected Gaussian DAG](#)

Abstract. Pearl's d-connection criterion identifies the independence relations implied by the global Markov property for a directed acyclic graph (DAG). It allows one to read off the conditional independencies directly from the DAG, by checking for the existence of a certain kind of path. However, the d-connection criterion does not indicate the degree of conditional dependence between a pair of d-connected vertices given a subset disjoint to them. Conditional dependence between the vertices in a DAG can be changed in two ways. The subset to be conditioned on may be fixed and the dependence of different pairs of vertices, conditional on that fixed set may be compared. On the other hand, one may be interested in fixing two vertices and comparing their conditional dependence by conditioning on different subsets. The talk will address both of these issues for singly connected Gaussian DAGs. In particular by conditioning upon a fixed subset we will show that for certain d-connecting paths the squared conditional correlation decreases with an increase in the length of the path. We will also show that the squared conditional correlations between two given vertices, conditioned on different subsets may be ordered by examining the relationship between the d-connecting path and the set of vertices conditioned upon. Some negative results on multiply connected and discrete DAGs will also be mentioned.

## **Dongfeng Li**

### [Mixed-Effects Model for Periodic Functional Data with Multiple Cycles](#)

Abstract. In many experiments data arise as periodic curves with multiple periods. In mixed effects functional models for such data, the group-level fixed effects are usually periodic functions of single period, and subject-level random effects describe subject-specific deviations thus can be of multi-period in nature. We propose smoothing spline models for the functional effects with different patterns of periodicity while taking into account the within-subject serial correlations. We introduce two equivalent algorithms to compute the proposed models, with emphasis in the derivation of an efficient state space algorithm. A simulation study is conducted to evaluate the proposed methods, which are further illustrated using a multi-beat pressure data set for evaluating effects of a heart failure medication on human left ventricular function.

## **Yan Li**

### [Mediators of the Association between the Apolipoprotein E \$\epsilon\$ 4 Allele and Cognitive Function](#)

Abstract. The presence of one or two copies of the  $\epsilon$ 4 allele of the apolipoprotein E gene (ApoE  $\epsilon$ 4) has been shown to be a predictor of cognitive decline and Alzheimer's disease. In a study of deceased subjects from the Religious Orders Study, we use mediation models to examine whether measures of brain pathology are in the causal chain between ApoE  $\epsilon$ 4 and cognitive function measured proximate to death. In addition, we consider the estimation of the mediation effect when the mediator is a binary variable. In our empirical study, the binary mediator is the presence of cerebral infarcts on autopsy. We give a precise definition of the mediation effect for a binary mediator and examine asymptotic properties of five different estimators of the mediation effect. Our theoretical developments, which are supported by a Monte Carlo study, show that the estimators that account for the binary nature of the mediator are consistent for the mediation effect while other estimators are inconsistent. We present empirical results for the model with one binary mediator and the model where two mediators are considered simultaneously. Research supported by National Institute on Aging grants P30 AG10161, R01 AG15819, and R01 AG17917.

## **Elena Perez-Bernabeu**

### [Attribute Control Charts: improvement of the u-chart applying J.J. Daudin's methodology.](#)

Abstract. In this work I show a comparison between the classic u-chart and a proposed new u-chart, based on JJ.Daudin's paper. Daudin modifies the Shewhart chart applying a double sampling, so that way, he improves the behavior of the chart. I apply Daudin's philosophy to the attribute chart, in this case, to the u-chart. In this study, I have used genetic algorithms and I give numerical examples and draw the power curves for some of the numerical examples to illustrate the new chart and to study the properties of this chart and compare to the classic one.

## **Hongkun Wang**

### [A study on confidence intervals for incremental cost-effectiveness ratios with censored data](#)

Abstract. In health economics and outcomes research, the incremental cost-effectiveness ratio (ICER) has long been used to compare the economic consequences relative to the health benefits of therapies. There are a number of approaches to estimating confidence intervals for ICERs in literature. We compare several non-parametric bootstrap methods, Fieller's method, and bootstrap Fieller's method in terms of their coverage probability under certain circumstance. The coverage probabilities of bootstrap methods are low due to their finite length of confidence intervals. We propose a new approach to obtain the confidence interval endpoints for the percentile bootstrap method. Simulation studies show that our new approach outperforms the percentile bootstrap method and the bootstrap Fieller's

method works the best. A real data example from a cardiovascular clinical trial is used to demonstrate the application of the methods. It is followed by some concluding remarks.

### **Jianan Peng**

#### [Stepwise Confidence Intervals for Monotone Dose-Response Studies](#)

Abstract. In dose-response studies, one of the most important issues is the identification of the minimum effective dose (MED), where the MED is defined as the lowest dose such that the mean response is better than the mean response of a zero-dose control by a clinically significant difference. Usually the dose-response curves are monotonic. Various authors have proposed step-down test procedures based on contrasts among the sample means to find the MED. In this paper, we improve Marcus and Peritz's method (1976, Journal of Royal Statistical Society, Series B, Vol 38, 157-165) and combine Hsu and Berger's DR method (1999, Journal of the American Statistical Association, Vol 94, 468-482) to construct the lower confidence bound for the difference between the mean response of any non-zero dose level and that of the control under the monotonicity assumption to identify the MED. The proposed method is illustrated by numerical examples and simulation studies on power comparisons are presented.

### **Qingzhao Yu**

#### [Bayesian Synthesis](#)

Abstract. Bayesian model averaging enables one to combine the disparate predictions of a number of models in a coherent fashion, leading to superior predictive performance. The improvement in performance arises from averaging models that make different predictions. In this work, we tap into perhaps the biggest driver of different predictions—different analysts—in order to gain the full benefits of model averaging. In a standard implementation of our method, several data analysts work independently on portions of a data set, eliciting separate models which are eventually updated and combined through Bayesian synthesis. The methodology helps to alleviate concerns about the sizeable gap between the foundational underpinnings of the Bayesian paradigm and the practice of Bayesian statistics. This paper provides theoretical results that characterize general conditions under which data-splitting results in improved estimation which, in turn, carries over to improved prediction. These results suggest general principles of good modeling practice. In experimental work we show that the method has predictive performance superior to that of many automatic modeling techniques, including AIC, BIC, Smoothing Splines, CART, Bagged CART, Bayes CART, BMA, BART and LARS. Compared to competing modeling methods, the data-splitting approach 1) exhibits superior predictive performance for real data sets and simulations; 2) makes more efficient use of human knowledge; 3) selects sparser models with better explanatory ability and 4) avoids multiple uses of the data in the Bayesian framework.

## Contact information: participants

Elena Perez Bernabeu	Universidad Politecnica de Valencia, Spain	elenapb@eio.upv.es
Joe Blitzstein	Harvard University, USA	blitzstein@stat.harvard.edu
Jiguo Cao	Yale University, USA	caojiguo@gmail.com
Sanjay Chaudhuri	National University of Singapore	sanjay@stat.nus.edu.sg
Lisha Chen	Yale University, USA	lisha.chen@yale.edu
Wei Chen	Wayne State University	chenw@karmanos.org
Guang Cheng	Duke University, USA	chengg@duke.edu
Jing Cheng	University of Florida, USA	jcheng@biostat.ufl.edu
Yu Cheng	University of Pittsburgh, USA	yucheng@pitt.edu
Pang Du	Virginia Polytechnic Institute and State University, USA	pangdu@vt.edu
Guangzhe Fan	University of Waterloo, Canada	gfan@uwaterloo.ca
Samiran Ghosh	Indiana University Purdue University Indianapolis, USA	samiran@math.iupui.edu
Ying Guo	Emory University, USA	yguo2@sph.emory.edu
Radu Herbei	Ohio State University, USA	herbei@stat.osu.edu
Tao Huang	University of Virginia, USA	th8e@virginia.edu
Bin Li	Louisiana State University, USA	bli@lsu.edu
Bo Li	National Center for Atmospheric Research, USA	boli@ucar.edu
Dongfeng Li	Fred Hutchinson Cancer Research Center, USA	dli@scharp.org
Jialiang Li	National University of Singapore	stalj@nus.edu.sg
Yan Li	University of California, San Diego, USA	graceyanli@ucsd.edu
Yehua Li	University of Georgia, USA	yehuali@uga.edu
Fei Liu	Duke University, USA	fei@stat.duke.edu
Timothy L. McMurry	DePaul University	tcmurry@depaul.edu
Yuval Nardi	Carnegie Mellon University, USA	yuval@stat.cmu.edu
Taeyoung Park	University of Pittsburgh, USA	tpark@pitt.edu
Jianan Peng	Acadia University, Canada	jianan.peng@acadiau.ca
Zhiguang Qian	University of Wisconsin-Madison, USA	zhiguang@stat.wisc.edu
Leming Qu	Boise State University, USA	lqu@boisestate.edu
Kevin J. Ross	Stanford University, USA	kjross@stanford.edu
Weixing Song	Kansas State University, USA	weixing@ksu.edu
Xiaoqian Sun	Clemson University, USA	xsun@clemson.edu
Alejandro Veen	IBM T.J. Watson Research Center, USA	aveen@us.ibm.com
Haiyan Wang	Kansas State University, USA	hwang@ksu.edu
Hongkun Wang	University of Virginia, USA	hkwang@virginia.edu
Huixia Judy Wang	North Carolina State University, USA	wang@stat.ncsu.edu
Jing Wang	University of Illinois at Chicago, USA	wangjing@math.uic.edu
Xiao Wang	University of Maryland Baltimore County, USA	wangxiao@umbc.edu
Xuerong Wen	University of Missouri Rolla, USA	wenx@umr.edu
Yujun Wu	University of Medicine and Dentistry of New Jersey, USA	wuy5@umdnj.edu
Yan Yang	Arizona State University, USA	yy@math.asu.edu
Qingzhao Yu	Louisiana State University, USA	qyu@lsuhsc.edu
Zhigang Zhang	Memorial Sloan-Kettering Cancer Center, USA	zhangz@mskcc.org
Qing Zhou	University of California, Los Angeles, USA	zhou@stat.ucla.edu

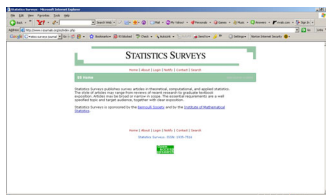
## Contact information: invited speakers

Samuel S. C. Kou	Harvard University, USA	kou@stat.harvard.edu
Michael Newton	University of Wisconsin at Madison, USA	newton@stat.wisc.edu
Deborah Nolan	University of California at Berkeley, USA	nolan@stat.berkeley.edu
David Banks	Duke University, USA	banks@stat.duke.edu
Susan Murphy	Co-editor, The Annals of Statistics	samurphy@umich.edu
David Banks	Coordinating Editor and A&CS Editor, JASA	banks@stat.duke.edu
David van Dyk	Editor, Journal of Computational and Graphical Statistics	dvd@uci.edu
Geert Molenberghs	Co-editor, Biometrics	geert.molenberghs@uhasselt.be
Grace Yang	Probability & Statistics Program, US-NSF	gyang@nsf.gov
Wendy Martinez	US Office of Naval Research	wendy.martinez@navy.mil
Ram Tiwari	National Institutes of Health/ National Cancer Institute	tiwarir@mail.nih.gov



## *Electronic Journals of the IMS*

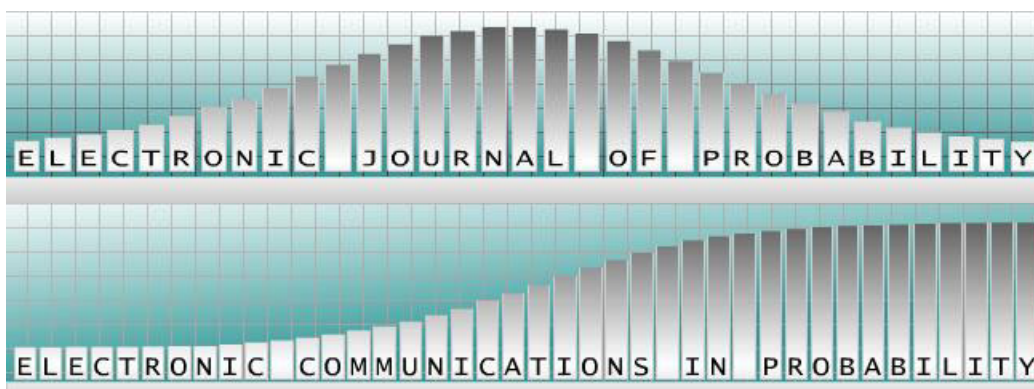
The Institute of Mathematical Statistics (IMS) co-sponsors several electronic, open-access journals. Please consider the following journals when publishing your work.



### ***Statistics Surveys* and *Probability Surveys***

<http://www.i-journals.org/ss/statistics.php>

<http://www.i-journals.org/ps/index.php>



### ***Electronic Journal of Probability* *Electronic Communications in Probability***

<http://www.math.washington.edu/~ejpecp/>



### ***Electronic Journal of Statistics***

<http://www.imstat.org/ejs/>