

# Ninth Meeting of New Researchers in Statistics and Probability

University of Washington  
Seattle, WA

August 1 - August 5, 2006

## Welcome

---

Hello and welcome to the Ninth IMS Conference for New Researchers. The purpose of the conference is to provide a comfortable setting for new researchers to share their research and make connections with their peers in an informal 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 programs, and we hope you enjoy your time at the University of Washington.

## 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 smaller 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 Minnesota, and next year will be in Utah. The current plan is to hold NRC every year prior to JSM, assuming the demand for the conference continues.

## Highlights of 2006 NRC

---

The structure of the conference is built around short presentations by the participants. There will be 3-4 sessions per day of six 12 minute talks, with ample break and lunch times factored in. On Thursday evening, just before the conference dinner, we will have a posters session. Please interact with the poster presenters. We are pleased to have talks by our invited speakers (Roger Berger, Dalene Stangl, and Bruce Weir), and we present two panel sessions. The first panel session will contain members of funding agencies who will give advice on how to get funding for your research. The second panel will contain editors of relevant journals and will provide some tips and tricks to maximize the probability of getting papers accepted.

Finally we are pleased to have Savas Dayanik 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!

*Peter Craigmile*, Chair of the IMS New Researchers Committee  
pfc@stat.ohio-state.edu

*Peter Hoff*, Chair of the Ninth Meeting of New Researchers Conference  
hoff@stat.washington.edu

## Acknowledgments

---

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

- Institute of Mathematical Statistics (IMS)
- National Institutes of Health (NIH) / National Cancer Institute (NCI)
- National Science Foundation (NSF)
- Office of Naval Research (ONR)
- National Security Agency (NSA)
- University of Washington
- The Ohio State University

## Schedule

---

|         | Tuesday<br>Aug 1                            | Wednesday<br>Aug 2                 | Thursday<br>Aug 3   | Friday<br>Aug 4  | Saturday<br>Aug 5                                  |
|---------|---|------------------------------------|---|--|--|
| 8:45    |   | Intro<br>Remarks                   | Intro<br>Remarks  | Intro<br>Remarks   | Intro<br>Remarks                                   |
| 9:00    |   | <b>Session 1</b><br>(6 speakers)   | <b>Session 5</b><br>(6 speakers)                                | <b>Session 8</b><br>(5 speakers)                               | <b>Journal Panel</b>                               |
| 10:15   |   | <b>Break</b>                       | <b>Break</b>  | <b>Break</b>   | <b>Break</b>                                       |
| 10:30   |   | <b>Session 2</b><br>(6 speakers)   | <b>Session 6</b><br>(6 speakers)                                | <b>Session 9</b><br>(5 speakers)<br>(10:30-11:30)              | <b>Session 11</b><br>(5 speakers)<br>(10:30-11:30) |
| 11:45   |   | <b>Lunch</b>                       | <b>Lunch</b>  | <b>Luncheon</b><br><b>Dalene Stangl</b><br><b>(11:45-2:00)</b> | <b>Roger Berger</b><br><b>(11:30-12:30)</b>        |
| 2:00    |   | <b>Session 3</b><br>(5 speakers)   | <b>Grants Panel</b>   | <b>Session 10</b><br>(6 speakers)                              |  |
| 3:15    |   | <b>Break</b>                       | <b>Break</b>  | <b>Break</b>   |  |
| 3:30    |   | <b>Session 4</b><br>(6 speakers)   | <b>Session 7</b><br>(6 speakers)                                | <b>Tweedie Speaker</b><br><b>Savas Dayanik</b>                 |  |
| 6:00    |   | <b>Dinner in Dorms</b>             | <b>Poster Mixer</b>   | <b>Dinner in Dorms</b>   |  |
| Evening | <b>Opening Mixer</b><br><b>(7:00-10:00)</b> | <b>Pub Event</b><br><b>(8:00-)</b> | <b>Conf. Dinner</b><br><b>Bruce Weir</b><br><b>(7:30-10:30)</b> |  |  |

### Notes

The opening mixer will be held from 7:00 - 10:00pm on Tuesday in the Pompeii room in McMahon Hall.

For the pub event, we will meet at the front entrance to McMahon Halls at 7:30pm on the Wednesday and go from there.

All participant talks are **12 minutes long**.

## Wednesday morning

---

### Session 1: Mixed and Random effects models, Robust methods for Repeated Measures and Finite Populations

9:00-10:15am, Location: Miller 301. Chair: Peter Hoff.

#### [Impact of the random effect distribution on inference for mean in linear mixed models](#)

Joshua Rushton: Cornell University, USA

#### [Functional Mixed-Effects Models for Periodic Data](#)

Li Qin: Fred Hutchison Cancer Research Center, USA

#### [Mixed-Effects, Posterior Means and Penalized-Least Squares](#)

Yolanda Munoz Maldonado: University of Texas-Houston, USA

#### [A Random Effects Four-Part Model for Longitudinal Medical Costs](#)

Lei Liu: University of Virginia, USA

#### [Rank-Based Analyses of Repeated Measure Designs Under Exchangeable Errors](#)

John Kloke: Pomona College, USA

#### [Robust model-based predictor of the finite population total](#)

Yan Li: University of Maryland, USA

---

### Session 2: Statistical Genetics, Microarray and Proteomic Analysis

10:30-11:45am, Location: Miller 301. Chair: Shuangge Ma.

#### [Estimation of Gene by Exposure Interactions in Case-Parent Triad Studies](#)

Tracy Bergemann: University of Minnesota, USA

#### [Building gene trees from SNPs data: an ancestral mixture models approach](#)

Shu-Chuan Chen: Arizona State University, USA

#### [MCMC Linkage Analysis for Two Genes and a Polygenic Component on General Pedigrees](#)

Yun Ju Sung: University of Washington, USA

#### [Uncertainty in clustering posterior distributions of gene expression levels using MCMC samples](#)

Tanzy Love: Carnegie Mellon University, USA

#### ~~[Boosting nearest shrunken centroid classifier for microarray data](#)~~

~~Baolin Wu: University of Minnesota, USA~~

#### [Feature identification quantitation and sample size: a comparative analysis of workflows for LC-MS based proteomics](#)

Olga Vitek: Institute for Systems Biology, USA

## Wednesday afternoon

---

### Session 3: Dimension Reduction, Principal components, Central Subspace Regression

2:00-3:15pm, Location: Miller 301. Chair: Jeongyoun Ahn.

#### [Penalized Likelihood Principal Component Rotation](#)

**Trevor Park:** University of Florida, USA

#### [Penalized Spline Models for Functional Principal Component Analysis](#)

**Fang Yao:** Colorado State University, USA

#### [Constrained Dimension Reduction Based on CANCOR](#)

**Jianhui Zhou:** University of Virginia, USA

#### [Optimal sufficient dimension reduction for the conditional mean in multiple-response regression](#)

**Jae Keun Yoo:** University of Louisville, USA

#### [Model Based Approaches for Simultaneous Dimension Reduction and Clustering](#)

**Xiaodong Lin:** University of Cincinnati, USA

#### [Using intra-slice information for improved estimation of the central subspace in regression](#)

**Liqiang Ni:** University of Central Florida, USA

---

### Session 4: Higher order asymptotics, Kurtosis, Probability models on simplexes, Methods for Censored data

3:30-4:45pm, Location: Miller 301. Chair: Yanyuan Ma.

#### [Higher-Order Asymptotic Normality Of Approximations To The Modified Signed Likelihood Ratio Statistic For Regular Models](#)

**Heping He:** The Australian National University, Australia

#### [Conditional Properties of a Parametric Bootstrap](#)

**Russell Zaretzki:** University of Tennessee, USA

#### [Kurtosis: New Theoretical Results and Inference Issues](#)

**Anna Maria Fiori:** Universita degli Studi di Milano - Bicocca, Italy

#### [A class of probability measures on the simplex, with emphasis on the Dirichlet distribution](#)

**Zach Dietz:** Tulane University, USA

#### [Estimation of truncated and censored regression models](#)

**Maria Karlsson:** Umeaa Universitet, Sweden

#### [Estimation of Wood Fibre Length Distributions from Censored Data through an EM Algorithm](#)

**Ingrid Svensson:** Umeaa Universitet, Sweden

## Thursday morning

---

### Session 5: Statistical learning, Kernel density estimation, Design and Process control

9:00-10:15am, Location: Miller 301. Chair: Ming Yuan.

#### [The \$F\_\infty\$ norm Support Vector Machine](#)

**Hui Zou:** University of Minnesota, USA

#### [Model Averaging via Penalized Regression for Tracking Concept Drifts](#)

**Cheolwoo Park:** University of Georgia, USA

#### [High Dimension Low Sample Size Data Analysis](#)

**Jeongyoun Ahn:** University of North Carolina, Chapel Hill, USA

#### [A cross-validation method for choosing the pilot bandwidth in kernel density estimation](#)

**Jose Enrique Chacon:** Universidad de Extremadura, Spain

#### [Robust Prediction and Extrapolation Designs for Censored Data](#)

**Xiaojian Xu:** University of Alberta, Canada

#### [Estimation of process parameters to determine the optimum diagnosis interval for control of defective items](#)

**Abhyuday Mandal:** University of Georgia, USA

---

### Session 6: Causal Inference and Small Area Estimation

10:30-11:45am, Location: Miller 301. Chair: Mayetri Gupta.

#### [Correction of Bias Due to Assay Dilution Effect in Immunogenicity Assessment](#)

**Yue Wang:** Merck, USA

#### [Efficiency of Study Design in Diagnostic Randomized Clinical Trials](#)

**Bo Lu:** The Ohio State University, USA

#### [Causal Inference, Sequential Monte Carlo and Clustering](#)

**Junni Zhang:** Peking University, China

#### [A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance](#)

**Qi Long:** Emory University, USA

#### [Robust estimation of the mean square error of an EBLUP of a small area mean](#)

**Shijie Chen:** Research Triangle Institute, USA

#### [Bayesian methodology which accounts for uncertainty about the commonality of a set of small area parameters](#)

**Guofen Yan:** University of Virginia, USA

## Thursday afternoon

---

### Grant funding panel

2:00-3:15pm, Location: Miller 301.

Members of the panel:

- Michelle Wagner: National Security Agency  
mdwagn4@nsa.gov
- Rong Chen: National Science Foundation, Division of Mathematical Sciences, Statistics  
rchen@nsf.gov
- Patrick Heagerty: University of Washington  
heagerty@u.washington.edu

---

### Session 7: Advanced Bayesian modeling

3:30-4:45pm, Location: Miller 301. Chair: Xiaoming Sheng.

#### **Flexible And Empirical Bayes Estimator For High Dimensional Data: Sparseness And Asymmetry**

**Min Zhang:** Purdue University, USA

#### **Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models**

**David Dahl:** Texas A&M University, USA

#### **High-dimensional Regression Model Search and Uncertainty**

**Christopher Hans:** The Ohio State University, USA

#### **High-Dimensional Sparse Factor Models and Latent Factor Regression**

**Carlos Carvalho:** Duke University, USA

#### **Multiple curve-fitting with BARS**

**Sam Behseta:** California State University, USA

#### ~~**Bayesian analysis of repeated data with many zeros: Application to the longitudinal adolescent substance abuse study**~~

~~**Hyonggin An:** University of Iowa, USA~~

## Thursday evening

---

### Posters

6:00-7:15pm, Location: McMahon Hall, Pompeii Room.

#### [Bayesian Hierarchical Model in Nest Survival Studies](#)

**Jing Cao:** Southern Methodist University, USA

#### [Statistical Comparison of Observed and Multi-Resolution CMAQ Modeled Ozone Concentrations](#)

**Li Chen:** University of Chicago, USA

#### [On Nonlinear Model Selection in Accelerated Failure Time Models](#)

**Chenlei Leng:** National University of Singapore, Singapore

#### [A Study of Multiple Imputation Methods Using Nonparametric Outlier Identification and Relative Accuracy Criteria with Clinical Laboratory Data](#)

**Xin Dang:** University of Mississippi, USA

#### [Applications of Copulas to Improve Covariance Estimation for PLS](#)

**Gina D'Angelo:** University of Pittsburgh, USA

#### [A Sensitivity Analysis for Sliced Inverse Regression](#)

**Ulrike Genschel:** Iowa State University, USA

#### [Functional Clustering by Shape of Elevation Profiles](#)

**Mark Greenwood:** Montana State University, USA

#### [Bootstrap Investigation of the Median Curve of a Functional Data Set](#)

**David Hitchcock:** University of South Carolina, USA

#### [Spatial and temporal models for evaluating IPCC climate model outputs](#)

**Mikyong Jun:** Texas A&M, USA

#### [Nonparametric Approach to Multivariate Massive Data Analysis by Convex Hull Peeling](#)

**Hyunsook Lee:** The Pennsylvania State University, USA

#### [A collaborative research with biomedical engineers in a cervical cancer detection project: some highlights](#)

**Jong Soo Lee:** The University of Texas M. D. Anderson Cancer Center, USA

#### [Locally Efficient Estimators for Semiparametric Models With Measurement Error](#)

**Yanyuan Ma:** Texas A&M University, USA

#### [Search for Level Sets of Functions by Computer Experiments](#)

**Curtis Miller:** University of California at Riverside, USA

## Thursday evening (cont.)

---

### Posters (cont.)

#### **Gaussian Mixture Models based on Frequency Phase Spectra for Efficient Face Authentication**

**Sinjini Mitra:** University of Southern California, USA

#### **Power transformation towards a linear regression quantile**

**Yunming Mu:** Texas A&M University, USA

#### **Use of Geometric Programming in Statistics**

**Xinlei Wang:** Southern Methodist University, USA

#### **Small sample inference for two sample location tests**

**Yu Wang:** Hogskolan Dalarna, Sweden

#### **Estimation of High Dimensional Predictive Densities**

**Xinyi Xu:** The Ohio State University, USA

#### **Efficient polynomial spline estimation of partially linear models for clustered data**

**Lan Xue:** Oregon State University, USA

#### **A Comprehensive Spatial-Temporal Analysis of Breast Cancer: First Primary, Second Primary and Breast Cancer Survival**

**Song Zhang:** The University of Texas M. D. Anderson Cancer Center, USA

#### **Case-Control Studies With Longitudinal Covariates**

**Honghong Zhou:** Indiana University, USA

---

### Conference dinner

7:30-10:30pm, Location: McCarty Hall, Rooms A and B.

#### **A statistician in court**

**Bruce Weir:** Chair, Professor of Biostatistics, University of Washington (Adjunct in Genome Sciences)

`bsweir@u.washington.edu`

**Abstract:** The training we receive as statisticians provides us with knowledge and tools that often have application in legal proceedings. Unfortunately this training does not extend to the issues surrounding the presentation of expert testimony at a trial. My work on statistical analyses of genetic data led to my involvement in the "DNA Wars" of the early to mid 1990s. I will describe some of my experiences from that time and describe some of the current statistical questions arising in the forensic uses of DNA profiling.

## Friday morning

---

### Session 8: Survival and ROC methods

9:00-10:15am, Location: Miller 301. Chair: Tracy Bergemann.

#### Weibull Prediction of Event Times in Randomized Clinical Trials

**Gui-Shuang Ying:** University of Pennsylvania, USA

#### A Hybrid Newton-Type Method for Censored Survival Data Using Double Weights in Linear Models

**Menggang Yu:** Indiana University, USA

#### Nonparametric Estimation of Survival Functions Conditional on Sparsely Observed Covariate Processes

**Ying Zhang:** University of Minnesota, USA

#### Two-Phase Designs for ROC Studies

**Mei-Hsiu Chen:** Brown University, USA

#### Regularized ROC method in disease classification using microarray data

**Shuangge Ma:** University of Washington, USA

---

### Session 9: Time series

10:30-11:45am, Location: Miller 301. Chair: Peter Craigmile.

#### Rank-Based Estimation for Autoregressive Moving Average Time Series Models

**Beth Andrews:** Northwestern, USA

#### A spatial blockwise empirical likelihood

**Dan Nordman:** Iowa State University, USA

#### ~~Inference in linear regression with long memory design and heteroscedastic long memory errors with application to model diagnostics~~

~~**Hongwen Guo:** Michigan State University, USA~~

#### Multivariate Time Series Analysis with Categorical and Continuous Variables in an LSTR model

**Ginger Davis:** University of Virginia, USA

#### Forecasting and Dynamic Updating of Time Series of Curves

**Haipeng Shen:** University of North Carolina, Chapel Hill, USA

## Friday lunchtime and afternoon

---

### Research Luncheon

11:45-2:00pm, Location: McCarty Hall, Rooms A and B.

#### **Who You Know, What You Know and Timing: Merging the Technical and the Personal to Define Your Professional Pathway**

**Dalene Stangl:** Professor and Director, Institute of Statistics and Decision Sciences, Duke University  
dalene@stat.duke.edu

**Abstract:** Having a successful career depends on balancing who you know, what you know, and timing. This talk will highlight a department chair's observations of many pathways by which faculty have had successful careers merging who and what they know both about their technical fields and about themselves personally. A few mistakes that are better avoided will also be shared.

---

### **Session 10: Spatial and Spatio-temporal Statistics, Surface Shape Analysis, Methods for fMRI, Extreme Value Distributions**

2:00-3:15pm, Location: Miller 301. Chair: Mikiyoung Jun.

#### **Gaussian process models for a sphere, with application to Faraday Rotation Measures.**

**Margaret Short:** Los Alamos National Laboratory, USA

#### **Anomaly Detection in Space-Time Point Processes**

**Michael Porter:** University of Virginia, USA

#### **Non Stationary Spatial Processes viewed as Locally Stationary Processes**

**Petrutza Caragea:** Iowa State University, USA

#### **Diagnosis and Exploration of Massively Univariate fMRI Models**

**Wen-Lin Luo:** Merck, USA

#### **Surface shape analysis with an application to brain cortical surface analysis in schizophrenia**

**Christopher Brignell:** University of Nottingham, United Kingdom

#### **Nonparametric estimation of the dependence function for a multivariate extreme value distribution**

**Dabao Zhang:** Purdue University, USA

## Friday afternoon (cont.)

---

### Tweedie Award Speaker

3:30-4:30pm, Location: Miller 301.

#### Poisson disorder problems

**Savas Dayanik:** Department of Operations Research and Financial Engineering, Princeton University

[sdayanik@princeton.edu](mailto:sdayanik@princeton.edu)

**Abstract:** Point processes have important applications in manufacturing, service enterprises, telecommunication systems, insurance and risk management. Much of the modeling efforts with them goes into the specification of their intensities. Sometimes, abrupt changes are expected in the intensity at the bursts of some critical and unpredictable events. These change times (also known as 'disorder times') demarcate different regimes. The difference between regimes can be significant, and it is important to detect the regime change as quickly as possible. In the classical Poisson disorder problem, the rate of a Poisson process changes at an unobservable disorder time from one known constant to another, and the question is how to design a procedure that quickly detects the disorder time. In this talk, the solution of Poisson disorder problem will be described when (i) the detection delay time is penalized by an exponential function, which is a better choice for financial applications, (ii) new arrival rate after the disorder time is unknown and unobservable, (iii) every arrival is accompanied by i.i.d. and observable marks whose identical distribution may also change at the disorder time. We will describe optimal sequential detection rules and accurate numerical methods to calculate their parameters. Finally, we will illustrate them on several numerical examples.

## Saturday morning

---

### Journal Editor Panel

9:00-10:15am, Location: Miller 301.

Members of the panel:

- Stephen Portnoy: Co-editor, JASA Theory and Methods  
sportnoy@uiuc.edu
- Jianqing Fan: Co-editor, The Annals of Statistics  
jqfan@Princeton.edu
- Edward George: Co-editor, Statistical Science  
edgeorge@wharton.upenn.edu
- Naisyin Wang: Co-editor, Biometrics  
nwang@stat.tamu.edu

---

### Session 11: Multivariate statistics, Data Depth, Graphical Models, Bivariate Growth charts

10:30-11:30am, Location: Miller 301. Chair: Hui Zou.

#### **Data Depth: Theory, Computations and Applications**

**Shojaeddin Chenouri:** University of Waterloo, Canada

#### **Estimate of regression coefficients based on the projection depth-weighted scatter matrix**

**Weihua Zhou:** University of North Carolina, Charlotte, USA

#### **Log-density functional ANOVA model estimation and nonparametric graphical model building**

**Yongho Jeon:** University of Wisconsin-Madison, USA

#### **Model Selection and Estimation in the Gaussian Graphical Model**

**Ming Yuan:** Georgia Institute of Technology, USA

#### **Bivariate growth charts**

**Ying Wei:** Columbia University, USA

## Saturday morning (cont.)

---

### Teaching speaker

11:30-12:30pm, Location: Miller 301.

### Software for Intro Stats: Is Excel an Option?

**Roger Berger:** Professor and Chair, Department of Mathematical Sciences and Applied Computing, Arizona State University

Roger.Berger@asu.edu

**Abstract:** Most academic statisticians face the question at some point, "What software (if any) to use in an introductory statistical methods course?" Microsoft Excel is one option. We will describe some of Excel's many shortcomings as a statistical analysis package. Nevertheless, we will describe our experience using Excel and argue that it is a reasonable choice considering the learning objectives of an introductory statistical methods course.

# Abstracts

---

## High Dimension, Low Sample Size Data Analysis

**Jeongyoun Ahn**

University of North Carolina, Chapel Hill, USA

[jyahn@email.unc.edu](mailto:jyahn@email.unc.edu)

Most of the literature regarding the statistical analysis of High Dimension, Low Sample Size (HDLSS) data deals with the situations where both the dimension  $d$  and the sample size  $n$  go to infinity together. In this talk the case where  $d$  tends to infinity while  $n$  is fixed is examined. We show that the sample covariance matrix behaves as if the underlying distribution is spherical if  $d$  is much larger than  $n$ . This result plays a key role in extending to more general settings the asymptotic geometric representation of HDLSS data, which says the randomness of the data only lies in random rotations of a regular  $n$ -simplex. The classification problem with HDLSS data is also considered in this presentation. There exists a one-dimensional direction in the data space (i.e., the  $n$  dimensional subspace generated by the data vectors) such that the projected data have only two distinct values. This direction is uniquely determined in the data space and lies within the affine set of the data. It has a similar formula to the Fisher's linear discrimination direction and is shown to be equivalent in non-HDLSS cases.

## Rank-Based Estimation for Autoregressive Moving Average Time Series Models

**Beth Andrews**

Northwestern, USA

[bandrews@northwestern.edu](mailto:bandrews@northwestern.edu)

A rank-based technique is used to estimate the parameters of autoregressive moving average (ARMA) time series models. The estimators minimize the sum of mean-corrected model residuals weighted by a function of residual rank. They are shown to be consistent and asymptotically normal under very mild conditions on the noise distribution, and so the estimation technique is robust. Because the weight function can be chosen so that rank estimation has the same asymptotic efficiency as maximum likelihood estimation, the estimators are also relatively efficient. The relative efficiency of the estimators extends to the unknown noise distribution case since rank estimation with the Wilcoxon weight function (a linear weight function) is nearly as efficient as maximum likelihood for a large class of noise distributions. We also give a weight function for which rank estimation is asymptotically equivalent to least absolute deviations estimation and a weight function for which rank estimation uniformly dominates Gaussian quasi-maximum likelihood estimation with respect to asymptotic efficiency. The quality of the asymptotic approximations for finite samples is studied via simulation.

## Multiple curve-fitting with BARS

**Sam Behseta**

California State University, USA

[sbehseta@csu.edu](mailto:sbehseta@csu.edu)

In this talk, I will present two methods for fitting curves to a population of histograms. The first method utilizes the properties of Bayesian Adaptive Regression Splines (BARS) to obtain the fits simultaneously. The second procedure updates the fits individually. Consequently, two applications associated with multiple curve-fitting will be discussed. First, a hierarchical model is formed to assess the variability between the curves while accounting for the variability due to the individual curve estimation. Second, a method is proposed to test the hypothesis of the equality of two functions. Both techniques can be studied in the context of Bayesian functional data analysis (Behseta, et al., 2005; Behseta and Kass, 2005). I will explain these ideas along with some interesting applications to the analysis of neuronal data. This is joint work with Rob Kass.

## Surface shape analysis, with an application to brain cortical surface analysis in schizophrenia

**Christopher Brignell**

University of Nottingham, United Kingdom

[pmaxcjb@nottingham.ac.uk](mailto:pmaxcjb@nottingham.ac.uk)

In many application areas it is of interest to compare the shapes and sizes of high-dimensional surfaces, and to investigate symmetry. We focus on a particular application in neuroscience, investigating large scale

cortical shape differences between control and schizophrenia patients. We introduce an automatic maximum likelihood method for brain registration, identifying the inter-hemispherical join, the anterior commissure, and the posterior commissure. Likelihood based inference is considered, and significant differences between the two groups are observed. The model is extended to account for curvature in the inter-hemispherical join. General practical issues in high-dimensional data analysis will be discussed.

### **Estimation of Gene by Exposure Interactions in Case-Parent Triad Studies**

**Tracy Bergemann**

University of Minnesota, USA

berge319@umn.edu

Studies that genotype individuals within nuclear families are now widespread. Generally, samples are drawn from an affected offspring, manifesting a disease or phenotype of interest, as well as from the parents [Ahsan H et al, 2002]. Case-parent triad designs avoid the potential for spurious association results due to admixture that can occur in case-control studies. And, if parents of the offspring are alive and consent to genotyping, the use of nuclear families can be a powerful method in certain disease settings. In my collaborations, we are applying this design to a genetic study of adolescent osteosarcoma patients. We will have 2-6 SNPs genotyped within each of eleven candidate genes, as well as exposure information for three different variables of interest. It is of interest to test for association, not only of single SNPs, but also any possible gene-gene interactions and gene-environment interactions using methods outlined in Umbach DM and Weinberg CR, 2000. The number of potential log-linear models to fit the data is therefore quite large. We suggest a strategy to find optimal models that incorporate both the biological information, e.g., the linkage disequilibrium patterns of the SNP data, as well as traditional methods for model selection such as the Bayesian Information Criterion. Further, we expand upon existing methods to test for haplotype association in case-control and cohort studies for various outcomes [Bergemann TL and Clarkson DB, 2005, Zeng D and Lin DY, 2005], and them to the case-parent triad design.

### **Bayesian Hierarchical Model in Nest Survival Studies**

**Jing Cao**

Southern Methodist University, USA

jcac@smu.edu

Recently, logistic nest survival models (Dinsmore, White, and Knopf 2002; Shaffer, 2004) have been developed to incorporate biological covariates with the assumption that the nest age on the first encounter of nest can be decided accurately. Also, the nest curve is assumed to be a parametric function (linear or quadratic) of nest age. In this paper, we propose a Bayesian hierarchical model with nest-specific covariates to estimate age-specific daily survival rates. The model can handle any mixture of irregular visiting schedules, and it allows a broad variety of covariates and competing models to be evaluated. With the least restrictive assumptions, the model does not require the knowledge of exact nest age when nest is first found. The typical features of nest survival data, truncation and censoring, are accounted for by the likelihood function and the latent variables. An intrinsic auto-regressive (IAR(2)) prior is employed for the nest age effect. This nonparametric prior provides a much more flexible and parsimonious alternative to the parametric specification. Last but not least, the Bayesian computation is efficient because the full conditional distributions either have closed forms or are log-concave. Finally, we present a simulation study and an analysis of a Missouri dickcissel dataset to illustrate the performance of the model.

### **Non Stationary Spatial Processes viewed as Locally Stationary Processes**

**Petrutza Caragea**

Iowa State University

pcaragea@iastate.edu

Environmental networks monitoring air pollutants often span over large geographical areas and include a very large number of locations. Modeling spatial correlation in a classical geostatistical sense could present several challenges, if one relies on the maximum likelihood estimation of the spatial parameters characterizing the 'global' covariance structure: restrictions on the computational capability when the number of spatial locations is very large as well as data collection patterns imposed by the general topography (less monitors located in mountainous regions, more sites located in highly populated areas). We propose a method of constructing a

'global' likelihood function that incorporates 'local' information through a conditional approach. This technique relies on the fact that the spatial process behaves as a stationary process at the local level, but not necessarily over the whole spatial domain. Another advantage of this method is that the computational effort associated with the estimation process is significantly diminished. Estimators obtained through the newly constructed 'global' model are compared to the 'local' estimators via a series of likelihood ratio tests. For data structures that allow the evaluation of the classical likelihood funcperformance of the proposed model with the classical approach.

## **High-Dimensional Sparse Factor Models and Latent Factor Regression**

**Carlos Carvalho**

Duke University, USA

[carlos@stat.duke.edu](mailto:carlos@stat.duke.edu)

We describe latent factor models for multivariate analysis in very high-dimensions, and classes of models that couple this framework with factor regressions for predictive modeling of multivariate response variables. We use sparse factor models relationships between high-dimensional variables and underlying lower-dimensional latent factors are sparse created using sparsity inducing priors. Model search and fitting are addressed through stochastic simulation (MCMC) and a novel evolutionary search. The latter computational approach explores, defines and fits models for higher-dimensional problems through an evolutionary process that gradually expands the dimension of the sample space. Combined/parallel modeling of response variables introduces additional response factor regression models in which the responses are linked components of the latent factor space. The approach extends prior work on factor analysis, sparsity prior modeling and factor regression in a number of key ways, including scalability to very high-dimensions as well as effective computational methods. Examples are drawn from studies in breast cancer genomics where the sparse factor models represent observed relationships in measured mRNA levels of thousands of genes, and where predictive molecular phenotyping is naturally addressed using factor regressions. This case study also exemplifies use of efficient software implementing the methodology.

## **A cross-validation method for choosing the pilot bandwidth in kernel density estimation**

**Jose Enrique Chacon**

Universidad de Extremadura, Spain

[jechacon@unex.es](mailto:jechacon@unex.es)

Bootstrap or plug-in bandwidth selection in kernel density estimation requires the use of a second pilot bandwidth. Typically, this pilot bandwidth is chosen according to some asymptotic criterion, which depends also on the unknown density. The options are to use a reference distribution here or to continue to the next stage, using another kernel density estimator and so on, ending up this process by using a reference distribution at some stage. In contrast, our proposal is based on a non-asymptotic minimum variance unbiased estimator of the error criterion for the pilot bandwidth, providing us with a crossvalidation type of pilot bandwidth selector. We illustrate our method with exact calculations for the Gaussian case and a simulation study.

## **Statistical Comparison of Observed and Multi-Resolution CMAQ Modeled Ozone Concentrations**

**Li Chen**

University of Chicago, USA

[lichen@uchicago.edu](mailto:lichen@uchicago.edu)

To run the Community Multi-scale Air Quality (CMAQ) modeling system at higher resolutions, one has to first run it at lower resolutions. This paper compares observed hourly ozone concentrations to the CMAQ modeled hourly ozone concentrations at three different spatial resolutions, e.g., 36 km, 12 km and 4 km. Bilinear interpolation is used to interpolate CMAQ model output to the locations of monitoring sites. Some performance measures, e.g., fractional bias (FB) and root normalized mean square error (RNMSE), are then calculated. The results show that higher resolution CMAQ model output does not necessarily provide smaller FB and RNMSE than the lower ones. Aggregation is performed to obtain new versions of lower resolution model output based on higher resolution model output. The aggregated lower resolution model output predicts better than either the unaggregated high resolution run or the low resolution run in terms of RNMSE. Variation decomposition is used as a tool to better understand the statistical behavior of CMAQ model output

at different resolutions. The temporal variation is well captured by CMAQ model output, but spatial variation and space-time interactions are not.

### **Two-Phase Designs for ROC Studies**

**Mei-Hsiu Chen**

Brown University, USA

[chenmei@stat.brown.edu](mailto:chenmei@stat.brown.edu)

Receiver Operating Characteristic (ROC) curve analysis has been widely used in evaluating the performance of diagnostic modalities for diagnostic populations in recent years. However, measuring accuracy of diagnostic tests based on ROC curves often proves to be costly and time-consuming when tests are applied to screening populations in which the number of participants is large and among them few diseased cases. Therefore, we propose to study two-phase designs that do not require verifying the disease status for every participant. Such designs can reduce cost, while still ensuring adequate precision for estimates of diagnostic performance in studies with low prevalence of disease. We approach the design and analysis of two-phase screening studies primarily from a Bayesian perspective. We discuss designs for estimating the ROC curve of a single test as well as those for comparing curves of two tests. In each case, we describe a framework for comparing and selecting designs that meets pre-specified statistical precision and cost criteria. To evaluate alternative study designs we simulate data according to available prior information and derive posterior estimates of the quantities of interest. Computations are performed using Markov Chain Monte Carlo (MCMC) and Data Augmentation methods. We compare the results from the Bayesian methods to those obtained from maximum likelihood methods. We conclude that two-phase designs for estimating and comparing ROC curves can be more efficient than single sample designs. The Bayesian approach incorporates available prior information in the design of the study and permits a fuller accounting of the uncertainty in selecting and evaluating alternative designs.

### **Robust estimation of the mean square error of an EBLUP of a small area mean**

**Shijie Chen**

Research Triangle Institute, USA

[schen@rti.org](mailto:schen@rti.org)

In this paper, we present a general method for estimating the mean square error (MSE) of EBLUP for the well-known Fay-Herriot small area model. We first obtain a second-order approximation to the MSE, i.e., an approximation which ignores all terms of the order  $o(m-1)$ , where  $m$  denotes the number of small areas. Unlike the normality-based approximations, our approximation involves the kurtosis (but not the skewness) of both the sampling and model errors and depends on the method of estimating variance component. We note that for the method of moments (MOM) estimator of the variance component, we do not require the estimation of kurtosis of the model error in order to obtain a MSE estimator which is unbiased up to the order  $O(m-1)$ . This extends an earlier result of Lahiri and Rao (1995) to the situation when the sampling errors are non-normal. However, in order to obtain a MSE estimator which is unbiased up to the order  $O(m-1)$ , the estimation of the kurtosis is necessary for other methods of estimating variance components, e.g., the method of estimating equation proposed by Fay and Herriot (1979). We propose a method for estimating the kurtosis which in turn provides a method of MSE estimation for a very general class of variance component estimators for the non-normality of both the sampling and model errors.

### **Building gene trees from SNPs data: an ancestral mixture models approach**

**Shu-Chuan (Grace) Chen**

Arizona State University, USA

[scchen@math.asu.edu](mailto:scchen@math.asu.edu)

An ancestral mixture model is proposed for clustering discrete multivariate sequences. This model has a natural relationship to the coalescent process of population genetics. The sieve parameter in the model plays an important role of time in the evolutionary tree of the sequences. In this talk, I will show how an ancestral mixture model can be used to build up a hierarchical tree from binary sequence data by sliding the sieve parameter. An example of genetic single nucleotide polymorphisms (SNP) data will be used for illustration. Some properties of the ancestral mixture model, such as its nested structure and the relationship to the coalescent process of population genetics, will be also presented.

## **Data Depth: Theory, Computations and Applications**

**Shojaeddin Chenouri**

University of Waterloo, Canada

[schenouri@uwaterloo.ca](mailto:schenouri@uwaterloo.ca)

Multivariate analysis plays an ever-increasing importance role in statistics. Most statistical experiments are multivariate in nature, and analysis of large multivariate data sets is now made possible by the recent advancements in computer technology. In multivariate inference, data often cannot be fitted by normal or, more general, elliptically symmetric distributions. Then the classical multivariate analysis which relies heavily on the assumption of normality or near normality (ellipticity), fails. These assumptions also are often difficult to justify in practice. The goal of this talk is to give an introduction to the recent advancements in multivariate nonparametric inference and data analysis based on the concept of data depth. A data depth is a measure of how deep or outlying a given point is with respect to a data cloud or a distribution. Depth functions introduce center-outward orderings and rankings of multidimensional data. In this talk, we shall review briefly different notions of data depth, the associated multidimensional medians, and their statistical properties and computational complexities. Along with examples we shall discuss some graphical techniques for the multivariate goodness of fit, multivariate dispersion, and skewness, etc. We also construct families of nonparametric multivariate multi-sample location and dispersion tests. To conclude, some other applications of data depth in multivariate data analysis and also research topics related to the notion of data depth will be discussed.

## **Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models**

**David B. Dahl**

Texas A&M University, USA

[dahl@stat.tamu.edu](mailto:dahl@stat.tamu.edu)

This paper proposes a new efficient merge-split sampler for both conjugate and nonconjugate Dirichlet process mixture (DPM) models. These Bayesian nonparametric models are usually fit using Markov chain Monte Carlo (MCMC) or sequential importance sampling (SIS). The latest generation of Gibbs and Gibbs-like samplers for both conjugate and nonconjugate DPM models effectively update the model parameters, but can have difficulty in updating the clustering of the data. To overcome this deficiency, merge-split samplers have been developed, but until now these have been limited to conjugate or conditionally-conjugate DPM models. This paper proposes a new MCMC sampler, called the sequentially-allocated merge-split (SAMS) sampler. The sampler borrows ideas from sequential importance sampling. Splits are proposed by sequentially allocating observations to one of two split components using allocation probabilities that condition on previously allocated data. The SAMS sampler is applicable to general nonconjugate DPM models as well as conjugate models. Further, the proposed sampler is substantially more efficient than existing conjugate and nonconjugate samplers.

## **A Study of Multiple Imputation Methods Using Nonparametric Outlier Identification and Relative Accuracy Criteria with Clinical Laboratory Data**

**Xin Dang**

University of Mississippi, USA

[xdang@olemiss.edu](mailto:xdang@olemiss.edu)

New nonparametric depth-based multivariate outlier identifier methods based on popular statistical depth functions are used as criteria in a study to compare several established methods of multiple imputation of missing data. Specifically, we examine the degree of agreement between the outliers detected in a complete data set and those detected in a modified data set after random removal of a fraction of data and replacement by imputed values. A second criterion based on a relative accuracy measure is also employed. The study utilizes actual clinical laboratory data sets kindly supplied by Dr. Kay Penny and reported in Penny and Jolliffe (1999). We extend their investigation based on the Mahalanobis and a generalized PCA outlier detection method, obtaining results on the relative merits of these methods and outlier detectors based on the spatial and projection depths. Among other results, an interesting finding is that there are optimal pairings of outlier detection method and multiple imputation method. That is, these should be selected together, for best performance.

## **Applications of Copulas to Improve Covariance Estimation for PLS**

**Gina D'Angelo**

University of Pittsburgh, USA

[gmdst17@pitt.edu](mailto:gmdst17@pitt.edu)

Dimension reduction techniques such as partial least squares (PLS) are currently being applied to classification problems in the area of genetics. These methods have also been applied to PET imaging data with the goal of creating summary measures and examining relationships between voxel-level data and covariates of interest. Previously, we have examined the use of standard PLS techniques for the analysis of amyloid deposition in AD and control subjects using PIB PET imaging techniques (Ziolko et al., 2005). The present work extends PLS to accommodate the unique correlation structure of this data set, for which the distribution of PIB voxel intensity values is a mixture of normal distributions while that of FDG PET is a single normal distribution. This extension is implemented by using a copula to estimate the covariance structure and illustrated in the PIB/FDG data set.

## **Multivariate Time Series Analysis with Categorical and Continuous Variables in an LSTR model**

**Ginger M. Davis**

University of Virginia, USA

[gingerdavis@virginia.edu](mailto:gingerdavis@virginia.edu)

We develop a methodology for multivariate time series analysis when our time series has components that are both continuous and categorical. Our specific contribution is a logistic smooth transition regression (LSTR) model whose transition variable is related to a categorical time series (LSTR-C). This methodology is necessary for series that exhibit nonlinear behavior dependent on a categorical time series. The estimation procedure is investigated both with simulation and an economic time series. We obtain superior or equivalent model fits as compared to another smooth transition regression model. Furthermore, even when the nonlinear behavior of the time series is dependent on a continuous time series, we propose a simplification of the modeling process which is the automatic formulation of the transition variable from the categorical time series. We are able to capture this nonlinear dependence on a continuous time series by using regression theory for categorical time series.

## **A class of probability measures on the simplex, with emphasis on the Dirichlet distribution**

**Zach Dietz**

Tulane University, USA

[zdietz@math.tulane.edu](mailto:zdietz@math.tulane.edu)

The  $n$ -dimensional simplex is the collection of  $n$ -dimensional real-valued vectors whose elements are non-negative and sum to 1. Probability distributions on the simplex arise in various mathematical contexts, for example as prior distributions for the weight coefficients of a mixture model in a Bayesian analysis. Unfortunately, there are only a few tractable distributions on the simplex, the most prominent being the Dirichlet distribution. This prominence results for many reasons, a primary one being its conjugacy to the multinomial distribution. However, the Dirichlet distribution also plays a central role in controlling the behavior of certain Generalized Polya Urn schemes, and RAM(residual allocation model) phenomenon. We will demonstrate how the Dirichlet distribution may be elicited from a Markovian reinforcement scheme, and then generalize the procedure to introduce a class of distributions on the simplex.

## **Kurtosis: New Theoretical Results and Inference Issues**

**Anna Maria Fiori**

Universita degli Studi di Milano - Bicocca, Italy

[anna.fiori@unimib.it](mailto:anna.fiori@unimib.it)

What is kurtosis? What sorts of distributions have high and low kurtosis? How does kurtosis change when observations are added to an existing distribution? Zenga (1996, 2006) defines kurtosis to be the average of right- and left-concentration of a random variable. His methodology leads to a new partial ordering by kurtosis (here denoted by  $\prec_k$ ) which applies to both symmetric and asymmetric distributions. This work aims at providing a deeper insight into the implications of Zenga's approach to kurtosis. I first prove that the kurtosis ordering  $\prec_k$  is a weakening of previously defined orderings on continuous distributions, such as Van Zwet's  $\prec_s$  (1964) and Lawrence's  $\prec_r$  (1975). Since a larger set of probability models may be ranked by kurtosis according to  $\prec_k$  it seems reasonable to require that kurtosis measures preserve this ordering. Some counterexamples

based either on skewness or majorization theory demonstrate that the usual coefficient of kurtosis  $\beta_2$  is not consistent with  $\prec_k$ . I then turn to Zenga's coherent measures  $K_1$  and  $K_2$  and explore their applicability in inference about kurtosis. A simulation study is performed to investigate the sample behaviour of  $K_1$  and  $K_2$  for increasingly heavy-tailed parent distributions. Compared with a normalized estimator of  $\beta_2$  Zenga's kurtosis measures exhibit superior empirical properties. This talk is based on my PhD dissertation under the supervision of Professor Michele Zenga (Universita degli Studi di Milano-Bicocca, Italy).

### **A Sensitivity Analysis for Sliced Inverse Regression**

**Ulrike Genschel**

Iowa State University, USA

ulrike@iastate.edu

A typical difficulty with nonparametric regression with a large number of regressor variables is the so-called curse of dimensionality. That is, as the dimension of the regressor space increases, more data are needed to fill the space densely enough to accurately estimate an underlying regression function. As a remedy, various dimension reduction procedures, such as SIR, SIR II (Li, 1991), SAVE (Cook and Weisberg (1991), Cook (2000)), or MAVE (Xia et al. (2002)) have been proposed for identifying an appropriate, smaller subspace of the original regressor space before fitting an underlying regression function. Because ultimately the estimation of a regression curve or link function relies crucially on the correct identification of the linear combinations that span the dimension reduction subspace, the sensitivity of a dimension reduction procedure to outlying observations becomes crucial to understand. Sensitivity of a statistical procedure is often measured by the amount of contaminated data necessary to cause the applied procedure to yield unreliable results, which is generally referred to as 'breakdown.' Our analysis is restricted to the influence of one or more observations on the estimate of the subspace  $\mathcal{B}$  obtained by applying SIR. In particular, we suppose that SIR produces an estimate  $\hat{\mathcal{B}}$  of the e.d.r. subspace  $\mathcal{B} \subset \mathbb{R}^p$  based on a sample  $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$  of size  $n$ . We then study the robustness of SIR when one or more observations in  $(X, Y)^n$  are replaced with arbitrarily contaminated values, producing a contaminated subspace estimate  $\hat{\mathcal{B}}^*$ . In order to assess this influence of contamination, we quantify the discrepancy of  $\hat{\mathcal{B}}$  and  $\hat{\mathcal{B}}^*$  based on a suitable metric. Geometrically, a meaningful distance measure can be defined via a metric based on the Frobenius norm which indicates that a worst case behavior of the SIR procedure is reached if spaces  $\hat{\mathcal{B}}$  and  $\hat{\mathcal{B}}^*$  are as orthogonal to each other as possible.

As a result we show the extent to which SIR is sensitive to outlying observations. Not only is it possible to obtain additional e.d.r. directions under contamination as stated by Cook and Critchley (2000), e.d.r. directions can also become lost under contamination to the extent that none of the true e.d.r. directions of the e.d.r. subspace are recoverable by SIR. Furthermore, we demonstrate that the data contamination scenarios that produce erroneous e.d.r. subspace estimates in SIR depend on the knowledge of the covariance structure of the regressor variables as well as on the knowledge of the dimension  $K$  of the true e.d.r. subspace. Our theoretical findings are supported by a simulation study.

### **Functional Clustering by Shape of Elevation Profiles**

**Mark Greenwood**

Montana State University, USA

greenwood@math.montana.edu

Methods for performing cluster analysis of functional data are discussed. Mountain valley elevation cross-profiles contain information on the shape of the valley at particular locations. It is of interest to explore potential groupings of profile shapes, especially regarding their relative  $U$  or  $V$ -shapedness as that pertains to the amount of glaciation that has occurred at a particular location. Functional data analysis (FDA) techniques are used to estimate the curvature of a set of Himalayan valley profiles. Hierarchical clustering along with a continuous, weighted distance semi-metric is used to explore the grouping in different profile shapes as estimated through the curvature.

## High-dimensional Regression Model Search and Uncertainty

Christopher Hans

The Ohio State University, USA

[hans@stat.ohio-state.edu](mailto:hans@stat.ohio-state.edu)

Model search in regression with very large numbers of candidate predictors raises challenges for both model specification and computation. Standard methods of model space exploration such as Markov chain Monte Carlo (MCMC) and step-wise methods are often infeasible or ineffective when there are thousands of possible predictor variables, as these methods often become stuck in localized regions of model space. New search methods inspired by the increasing availability of large distributed computing environments have dominated existing methods of model space exploration in various multivariate model settings. I describe these 'Shotgun Stochastic Search' (SSS) techniques and demonstrate how they apply to regression modeling. Results from both simulation studies and real-data examples are presented demonstrating the advantages of SSS methods over competing MCMC approaches. Examples arising from gene expression cancer genomics are discussed.

## Higher-Order Asymptotic Normality Of Approximations To The Modified Signed Likelihood Ratio Statistic For Regular Models

Heping He

The Australian National University, Australia

[he@galton.uchicago.edu](mailto:he@galton.uchicago.edu)

Approximations to the modified signed likelihood ratio statistic are asymptotically standard normal with error of order  $n^{-1}$  where  $n$  is sample size. Proofs of this fact generally require that the sufficient statistic of the model be written as  $(\hat{\theta}, a)$  where  $\hat{\theta}$  is the maximum likelihood estimator of parameter  $\theta$  of the model and  $a$  is an ancillary statistic. This condition is very difficult or impossible to verify for many models. However, calculation of the statistics themselves do not require this assumption. This paper is devoted to exploring higher-order asymptotic normality of these statistics under general conditions. It focuses on the case that  $\theta$  may be parameterized as  $\theta = (\psi, \lambda)$ , where  $\psi$  is the scalar parameter of interest and  $\lambda$  is a nuisance parameter vector. Under general assumptions, the asymptotic properties of the statistics are proved. These proofs do not put any requirements of the sufficient statistic, they just assume general conditions which are easy to verify for and satisfied by commonly used models. Therefore this research removes the theoretical obstacle for applying these statistics to commonly used models.

## Bootstrap Investigation of the Median Curve of a Functional Data Set

David Hitchcock

University of South Carolina, USA

[hitchcock@stat.sc.edu](mailto:hitchcock@stat.sc.edu)

A new definition of the median of a set of functional data is given. The sample median curve is defined as the data curve with the minimum average pairwise L1 distance to all other curves, and is a better representative of a 'typical' curve than the pointwise mean or pointwise median curve. The median curve has connections to the sample median of a set of univariate or multivariate observations. An analogous definition of a population median curve is presented, along with methods for approximate inference about the population median curve. These methods in particular, approximate confidence bands for the median curve are based on a bootstrap approximation of the sampling distribution of the sample median curve.

## Log-density functional ANOVA model estimation and nonparametric graphical model building

Yongho Jeon

University of Wisconsin-Madison, USA

[yjeon@stat.wisc.edu](mailto:yjeon@stat.wisc.edu)

The (undirected) graphical models use graphs to compactly display the conditional dependence among random variables and have become popular, but have mostly been studied in the parametric framework. To enhance the scope of applicability of the graphical model, we consider the building of (undirected) nonparametric graphical models through their connection with log-density functional analysis of variance (ANOVA) models. Existing methods for fitting the log-density ANOVA model require repeated numerical integration of high dimensional functions, and are infeasible in problems of dimension larger than four. We propose a new method for fitting the log-density ANOVA model based on a penalized M-estimation formulation with a novel loss function. Solving the penalized M-estimation problem does not require high dimensional integration. When

the smoothing spline type penalty is used in our formulation, the procedure is shown to achieve the optimal rate in nonparametric estimation, and an easily computable approximation to the leave-out-one cross-validation is constructed to facilitate the choice of the tuning parameter. With a sparsity-inducing penalty, we obtain a sparse solution in terms of functional ANOVA components. This provides a practical way to construct and estimate nonparametric graphical models by carrying out both estimation and model selection simultaneously. We also introduce a characterization of the log-density smoothing spline ANOVA model.

### **Spatial and temporal models for evaluating IPCC climate model outputs**

**Mikyoung Jun**

Texas A&M, USA

[mjun@stat.tamu.edu](mailto:mjun@stat.tamu.edu)

There has been extensive efforts to develop climate models to study the climate change. The Intergovernmental Panel on Climate Change (IPCC) is one of the efforts and we have more than 20 climate models from IPCC. The previous works in the literature dealing with the climate model outputs commonly assume that the climate model outputs are random samples from a symmetric distribution centered around the true climate. Now, one of the most interesting problems to climate scientists and climate modelers is to verify the bias of these climate models and how the biases from the models are correlated. We propose statistical models to answer this question: we model the climate model outputs as spatial and temporal processes defined on sphere across time domain and we focus on modeling the spatial and temporal covariance structure of these processes. We propose covariance models not only for each climate model outputs but also cross-covariance models for pairs of climate model outputs. Based on our method, we can quantify the climate model biases in various ways and particularly, we can classify the climate models that have common biases in certain subset of spatial or temporal domain.

### **Estimation of truncated and censored regression models.**

**Maria Karlsson**

Umeaa Universitet, Sweden

[maria.karlsson@stat.umu.se](mailto:maria.karlsson@stat.umu.se)

A vast flora of different standard statistical methods is made available to practitioners in statistical computer software packages. These methods rest upon assumptions about the data used in the analysis and the population from which the data is generated. However, there are many situations when these assumptions are not fulfilled whereby standard methods of analysis may be inappropriate. For these situations modifications of available methods are required as is the development of new methods. In this talk estimation of linear regression models using truncated and/or censored data is considered. Several suggestions for estimators, both likelihood based estimators and estimators of so called semiparametric regression models, i.e., models with only weak regularity conditions placed on the random part of the model, have been made. In the talk semiparametric estimators which can be employed under asymmetric distributions, a feature not shared with many existing semiparametric estimators, are proposed.

### **Rank-Based Analyses of Repeated Measure Designs Under Exchangeable Errors**

**John D. Kloke**

Pomona College, USA

[john.kloke@pomona.edu](mailto:john.kloke@pomona.edu)

Rank-based analyses based on R estimates have been developed over the last twenty years for handling linear models with independently distributed errors. In this paper, we extend this analysis to linear models where the responses consist of repeated measures on subjects. For such designs, the random errors for a subject are generally not independent. We consider the case where their distribution is exchangeable, a case which often occurs in practice. The estimates are the R estimates used in fitting linear models and we obtain their asymptotic distribution for these repeated measure designs. The asymptotic theory for the corresponding Wald type and drop in dispersion type tests for general linear hypotheses is developed under null and local alternative models. Studentized residuals are also derived. Asymptotic relative efficiency of the analysis relative to the traditional least squares analysis are obtained. A simulation study confirms the validity and efficiency of the analysis. The theory developed is for general score functions; hence, the analysis can be optimized depending on the underlying knowledge of the error distributions. Several examples are discussed in detail which illustrate

the robustness of the analysis in coping with outliers in response space.

### **Nonparametric Approach to Multivariate Massive Data Analysis by Convex Hull Peeling**

**Hyunsook Lee**

The Pennsylvania State University, USA

[hlee@stat.psu.edu](mailto:hlee@stat.psu.edu)

An ad hoc device on multidimensional massive data is in demand. However, multivariate data analysis tools not imposing multivariate normal distribution exist rarely. We introduce convex hull peeling algorithms as a such device for the analysis of multidimensional massive data. Only the convexity of data sets is assumed. These convex hull peeling algorithms are designed to estimate quantiles, detect outliers, and measure distribution shapes of multidimensional data. Additionally, the algorithms are exemplified with Monte Carlo simulations and SDSS DR4 Quasars.

### **A collaborative research with biomedical engineers in a cervical cancer detection project: some highlights.**

**Jong Soo Lee**

The University of Texas M. D. Anderson Cancer Center, USA

[jslee@stat.rice.edu](mailto:jslee@stat.rice.edu)

We present some of the findings in the collaborative research between statisticians and biomedical engineers. Both sides work together on the cervical cancer detection project, where the biomedical engineers have proposed the fluorescence spectroscopy medical device for an early detection of cancer. The output from the device is a functional data, so the statisticians are needed for the data analysis aspect of the project. Thus, we consider selected topics in functional data analysis for this problem. First, we propose a procedure for testing pointwise difference of functional data in the two-sample framework. Our proposed method is a generalization of Hotelling's  $T^2$  test, where we utilize an adaptive truncation technique for dimension reduction. Also, we propose a method to detect the significantly different regions between curves. Once we determine that the samples curves from the two or more populations are significantly different overall, we want to look at the local regions of the curves and see where the differences occur. The problems posed by this collaborative work have motivated the development of the methodologies in the present work, and we demonstrate that our techniques work well and that we have been successful in applying statistical methods to other fields.

### **On Nonlinear Model Selection in Accelerated Failure Time Models**

**Chenlei Leng**

National University of Singapore, Singapore

[stalc@nus.edu.sg](mailto:stalc@nus.edu.sg)

We propose a nonlinear model selection method in accelerated failure time models. Formulated in the framework of smoothing spline ANOVA, the technique achieves a sparse representation of functional decomposition, by utilizing a reproducing kernel Hilbert norm penalty. Theoretical properties of the method are investigated. The usefulness of the methodology is demonstrated via simulated and real data sets.

### **Robust model-based predictor of the finite population total**

**Yan Li**

University of Maryland, USA

[yli@survey.umd.edu](mailto:yli@survey.umd.edu)

The prediction approach to inference has received considerable attention in finite population sampling. Under this approach, the finite population is assumed to be a realization from a superpopulation model. Often the standard linear model is assumed on the response variable. However, this assumption has been found to be inadequate in many survey applications. As a remedy, a known transformation, such as the log transformation or the square root transformation, is frequently used before a linear model is applied. Specification of a transformation is not a trivial problem in many applications. This paper introduces the well-known Box-Cox transformation in analyzing survey data from a finite population. The method is adaptive in the sense that the appropriate transformation needed is automatically determined by the data. Thus, robustness with respect to the model misspecification is achieved. I justify the need for selecting a model within the Box-Cox class using a real life data.

## **Model Based Approaches for Simultaneous Dimension Reduction and Clustering**

**Xiaodong Lin**

University of Cincinnati, USA

linxg@email.uc.edu

High dimensional data are regularly generated from various sources. Traditional data analysis performs dimension reduction and clustering separately. In this talk we present a penalized version of the mixture of factor analyzers model to address the problem of simultaneous dimension reduction and clustering. We have proposed a penalty that accounts for the ratio of the total variation explained by the noise component and the factors. We have developed a modified EM algorithm for parameter estimation. In our model, different mixture components are allowed to have different number of factors. This flexibility generates model selection challenges. To overcome the difficulty, we have proposed a two-step model selection procedure during which parameter estimations and model specifications are altered dynamically. We have shown that this procedure converges under both AIC and BIC criteria. Several high dimensional data sets are studied to show the superior performance of our model. Open questions and further research will be discussed.

## **A Random Effects Four-Part Model for Longitudinal Medical Costs**

**Lei Liu**

University of Virginia, USA

liulei@virginia.edu

Increasing interest has focused on the analysis of longitudinal medical cost data. The availability of longitudinal cost data can make health economic analyses more efficient and interpretation more comprehensive, insightful, and useful. Analysis of longitudinal cost data is also essential to understanding the dynamics of medical costs, e.g., the time pattern of medical costs, which should be of interest to clinicians, patients and payers. In this paper we extend the four-part model, which was proposed by Duan et al. (1983) in the cross-sectional medical cost setting, to longitudinal medical cost setting (e.g., monthly medical costs) with correlated random effects. Four joint equations are used to model respectively: (1) the probability of seeking medical treatment, (2) the probability of being hospitalized (conditional on seeking medical treatment), and the actual amount of (3) outpatient and (4) inpatient costs. Our model simultaneously takes account of the inter-temporal correlation of each patient and the cross-equation correlation of the four equations, by means of joint linear mixed models and generalized linear mixed models. The estimation is proceeded by the high-order Laplace approximation technique in Raudenbush et al. (1998) and Olsen and Schafer (2001). Our model is applied to monthly medical costs of 1397 chronic heart failure patients from the clinical data repository (CDR) at the University of Virginia.

## **A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance**

**Qi Long**

Emory University, USA

qlong@sph.emory.edu

We consider the analysis of clinical trials that involve randomization to an active treatment ( $T = 1$ ) or a control treatment ( $T = 0$ ), when the active treatment is subject to all-or-nothing compliance. We compare three approaches to estimating treatment efficacy in this situation: as-treated analysis, per-protocol analysis, and instrumental variable (IV) estimation, where the treatment effect is estimated using the randomization indicator as an instrumental variable. Both model-based and method-of-moment based IV estimators are considered. The assumptions underlying these estimators are assessed, standard errors and mean squared errors of the estimates are compared, and design implications of the three methods are examined. Extensions of the methods to include observed covariates are then discussed, emphasizing the contrasting role of covariates in these extensions. Methods are illustrated on data from the Women Take Pride study, an assessment of behavioral treatments for women with heart disease.

## **Uncertainty in clustering posterior distributions of gene expression levels using MCMC samples** **Tanzy Mae Tallapoosa Paz Love**

Carnegie Mellon University, USA

[tanzy@andrew.cmu.edu](mailto:tanzy@andrew.cmu.edu)

In large time series or multiple treatment microarray experiments, we are interested in locating groups of genes that react together. Subject matter theory designates these groups as coregulated by the same biologic pathways. For example, genes responsible for photosynthetic processes may express together in an experiment covering time periods in light and darkness. The statistical problem is then clustering genes based on their expression values over multiple treatments. However, we don't have values for gene expression, rather replicated estimates for each treatment condition. We have used the mean of the gene expression estimates in clustering genes. However, this ignores the uncertainty we have in the actual values of expression. To incorporate this uncertainty, we have modeled expression estimates using hierarchical models. This provides posterior probability distributions for quantities such as expression value and expression ratio for two treatments. We also can construct the joint posterior probability distribution of all expression values or all pairwise expression ratios for each gene. We use multiple sampling from the posterior distributions of gene expression vectors to cluster genes and estimate the uncertainty in this clustering.

## **Efficiency of Study Design in Diagnostic Randomized Clinical Trials**

**Bo Lu**

The Ohio State University, USA

[blu@sph.osu.edu](mailto:blu@sph.osu.edu)

From the patients' management perspective, a good diagnostic test should contribute to both reflecting the true disease status and improving clinical outcomes. Two study designs for the randomized clinical trial—the two-arm design and the paired design—are compared in the evaluation of diagnostic tests with patient outcomes as the primary endpoint. In the conventional two-arm design, patients are randomized to one of the diagnostic tests. In the paired design, patients undergo both tests and randomization occurs in the patients with discordant test results. Treatment will be applied based on test results. The follow-up clinical outcomes will be measured to determine the prognostic value of the tests. The paired design is shown to be more efficient than the two-arm design when the operating characteristics of the tests are given. The efficiency gain depends on the discordant rate of test results. Estimation of important quantities under the paired design is derived and simulation studies are also conducted to verify the theoretical results.

## **Diagnosis and Exploration of Massively Univariate fMRI Models**

**Wen-Lin Luo**

Merck, USA

[wen\\_lin\\_luo@merck.com](mailto:wen_lin_luo@merck.com)

Statistical thresholds and P-values cannot be trusted unless you assess the assumptions of the model. Linear models assume (1) no systematic lack-of-fit, i.e. mean zero errors, (2) constant variance, (3) independence or a specific dependence structure, and (4) Gaussian distributed errors. In neuroimaging checking assumptions is particularly challenging, as there are 100,000 univariate models fit simultaneously. In this work we reported methods to swiftly find and characterize violations of these assumptions, using either a working correlation model of independence or an arbitrary covariance structure in the functional magnetic resonance imaging (fMRI) data. We demonstrate diagnosis of linear model assumptions, including intra-subject time series analysis as well as inter-subject group analysis, through the statistical parametric mapping diagnosis (SPMd) toolbox.

## **Regularized ROC method in disease classification using microarray data**

**Shuangge Ma**

University of Washington, USA

[shuangge@u.washington.edu](mailto:shuangge@u.washington.edu)

An important application of microarrays is to discover genomic biomarkers, among tens of thousands of genes assayed, for disease classification. It is desirable to develop efficient statistical methods that can simultaneously identify important biomarkers from such highthroughput genomic data and construct appropriate classification rules. The ROC (receiver operating characteristic) technique has been widely used in disease classification with low dimensional biomarkers because (1) it does not assume a parametric form of the class probability as

required for example in the logistic regression method; (2) it accommodates outcome-dependent samplings, for example case-control and case-cohort designs; and (3) it allows different penalty on false positives and false negatives. We propose using the binormal AUC (area under the ROC curve) as the objective function for two-sample classification, and the threshold gradient directed regularization method for regularized estimation and biomarker selection. Tuning parameter selection is based on the V-fold cross validation. We develop Monte Carlo based methods for evaluating the stability and prediction performance of the proposed estimator and individual biomarkers.

### **Locally Efficient Estimators for Semiparametric Models With Measurement Error**

**Yanyuan Ma**

Texas A&M University, USA

[ma@stat.tamu.edu](mailto:ma@stat.tamu.edu)

We derive constructive locally efficient estimators in semiparametric measurement error models. The setting is one where the likelihood function depends on variables measured with and without error, where the variables measured without error can be modelled nonparametrically. The algorithm is based on backfitting. We show that if one adopts a parametric model for the latent variable measured with error and if this model is correct, then the estimator is semiparametric efficient; if the latent variable model is misspecified, our methods lead to a consistent and asymptotically normal estimator. Our method further produces an estimator of the nonparametric function that achieves the standard bias and variance property. We extend the methodology to allow for parameters in the measurement error model to be estimated by additional data in the form of replicates or instrumental variables. The methods are illustrated via a simulation study and a data example, where the putative latent variable distribution is a shifted lognormal, but concerns about the effects of misspecification of this assumption and the linear assumption of another covariate demands a more model-robust approach. A special case of wide interest is the partially linear measurement error model. If one assumes that the model error and the measurement error are both normally distributed, then our estimator has a closed form. When a normal model for the unobservable variable is also posited, our estimator becomes consistent and asymptotically normally distributed for the general partially linear measurement error model, even without any of the normality assumptions under which the estimator is originally derived. We show that the method in fact reduces to a same estimator in Liang et al. (1999), thus showing a previously unknown optimality property of their method.

### **Mixed-Effects, Posterior Means and Penalized-Least Squares**

**Yolanda Munoz Maldonado**

University of Texas-Houston, USA

[Yolanda.M.Munoz@uth.tmc.edu](mailto:Yolanda.M.Munoz@uth.tmc.edu)

A general framework that encompasses mixed-effects model methodology (using frequentist and Bayesian perspectives), and penalized least-squares techniques is developed. This framework allows for extension of classical results on the numerical equivalence between smoothing spline estimators, the best linear unbiased predictor of a certain normal mixed-effects model, and the posterior mean of a Gaussian signal-plus-noise model with diffuse initial conditions. We also show that the methods of generalized cross-validation, generalized maximum likelihood, and unbiased risk prediction can be used to estimate the variance components or smoothing parameters in any of the three settings. Our proposed framework has the implication that, in many cases of practical interest, an efficient, linear time, algorithm can be used to obtain the desired predictors and corresponding Bayesian confidence intervals. This algorithm also permits the evaluation of the exact likelihood function with the same level of computational efficiency. To illustrate the range of applicability of our main results we use examples from three different settings: varying coefficient models, ridge regression and randomized block designs.

## **Estimation of process parameters to determine the optimum diagnosis interval for control of defective items**

**Abhyuday Mandal**

University of Georgia, USA

[amandal@stat.uga.edu](mailto:amandal@stat.uga.edu)

The on-line quality monitoring procedure for attributes proposed by Taguchi has been critically studied and extended by a few researchers. Determination of the optimum diagnosis interval requires estimation of some parameters related to the process failure mechanism. Improper estimates of these parameters may lead to incorrect choice of the diagnosis interval and consequently huge economic penalties. In this paper, we highlight both the theoretical and practical problems associated with the estimation of these parameters, and propose a structured approach to solve them. For the so-called Case II model, two estimation methods, one based on Bayesian procedure and the other on the EM algorithm, are developed and compared using extensive simulations. These two methods are demonstrated using a case study from a hot rolling mill. A Bayesian method is proposed for estimation of parameters in Case III. A systematic way to utilize available engineering knowledge in eliciting the prior for the parameters is also discussed.

## **Search for Level Sets of Functions by Computer Experiments**

**Curtis Miller**

University of California at Riverside, USA

[millercp@ufl.edu](mailto:millercp@ufl.edu)

In engineering and other fields, it is common to use a computer simulation to model a real world process. The inputs to a function  $f$  represent factors that influence the outcome. The output represents a quantity of interest. Often there will be a specified level  $L$ , and the objective is to find the inputs for which output is above  $L$ .  $L$  may be a tolerance level, and the inputs for which response is larger than  $L$  form a tolerance region. We might estimate the tolerance region by evaluating  $f$  on a grid, but even a coarse grid may have thousands of points in four or five dimensions. If the function  $f$  is costly to evaluate, we need to be able to estimate the tolerance region with as few evaluations as possible. We approach this problem with a sequential search. Use data at any stage to fit a spatial process that approximates the function. Fit a Gaussian spatial process, as described in Currin, Mitchell, Morris, and Ylvisaker[1991]. The spatial process gives an estimate of the  $L$ -contour. We can also use the process to estimate how much information would be gained if  $f$  is evaluated at point  $p$ . Choose points where it is estimated that  $f$  takes value  $L$ , but where uncertainty is high. Evaluate  $f$  at the chosen points. This will augment the set of data points and the vector of data values. Repeat the procedure with this augmented data. Calculate convergence criteria after each iteration, and stop when criteria reach predetermined goals. The search process is applied to several functions defined in low dimensional space. Finally, it is applied to an actual simulation function.

## **Gaussian Mixture Models based on Frequency Phase Spectra for Efficient Face Authentication**

**Sinjini Mitra**

University of Southern California, USA

[mitra@isi.edu](mailto:mitra@isi.edu)

The modern world has seen a rapid evolution of the technology of biometric authentication, prompted by an increasing urgency to ensure a systems security. The need for efficient authentication systems has skyrocketed since 9/11, and the proposed inclusion of digitized photos in passports shows the importance of biometrics in homeland security today. Based on a persons essentially unique biological traits, these methods are potentially more reliable than traditional methods like PINs and ID cards. In this talk, I present a novel face authentication system developed in the frequency domain by exploiting the well-known significance of phase in face identification. A Gaussian mixture model (GMM)-based approach is used and we show that our proposed system outperforms an existing state-of-the-art system. Some classification results, inference and associated challenges are discussed.

## **Power transformation towards a linear regression quantile**

**Yunming Mu**

Texas A&M University, USA

[ymu@stat.tamu.edu](mailto:ymu@stat.tamu.edu)

In this paper, we consider the linear quantile regression model with a power transformation on the dependent variable. Like the classical Box-Cox transformation approach, it extends the applicability of linear models without resorting to nonparametric smoothing, but transformations on the quantile models are more natural due to the equivariance property of the quantiles under monotone transformations. We propose an estimation procedure and establish its consistency and asymptotic normality under some regularity conditions. The objective function employed in the estimation can also be used to check inadequacy of a power-transformed linear quantile regression model and to obtain inference on the transformation parameter. The proposed approach is shown to be valuable through illustrative examples.

## **Using intra-slice information for improved estimation of the central subspace in regression**

**Liqiang Ni**

University of Central Florida, USA

[lqi@mail.ucf.edu](mailto:lqi@mail.ucf.edu)

Many methods for estimating the central subspace in regression require slicing a continuous response. However, slicing can result in loss of information and in some cases that loss can be substantial. We use intra-slice covariances to construct improved inference methods for the central subspace. These methods are optimal within a class of quadratic inference functions and permit chi-squared tests of conditional independence hypotheses involving the predictors.

## **A spatial blockwise empirical likelihood**

**Dan Nordman**

Iowa State University, USA

[dnordman@iastate.edu](mailto:dnordman@iastate.edu)

We consider an empirical likelihood for weakly dependent spatial data located on a grid. This formulation of empirical likelihood combines spatial data blocking and general estimating functions to estimate spatial parameters. The method allows likelihood-type inference on spatial parameters, such as means and variograms, without a spatial model or knowledge of the underlying spatial dependence structure. The spatial empirical likelihood results in log-likelihood ratios which have chi-square limits under spatial dependence and maximum empirical likelihood estimators are possible for parameter estimation and testing spatial moment conditions. A practical Bartlett correction also is proposed to improve the coverage accuracy of confidence regions for spatial parameters. The performance of the spatial method is investigated through a simulation study.

## **Model Averaging via Penalized Regression for Tracking Concept Drifts**

**Cheolwoo Park**

University of Georgia, USA

[cpark@stat.uga.edu](mailto:cpark@stat.uga.edu)

A supervised learning algorithm aims to build a prediction model using training examples. This paradigm typically has the intrinsic assumptions that both training examples and an example to predict are drawn from the same distribution and the true input-output dependency does not change. However, these assumptions often fail to hold, especially in sequential data streams. This phenomenon is known as a concept drift, which has recently been treated as one of crucial issues in machine learning and data stream mining community. In this paper, we propose a new model averaging type algorithm based on penalized regression, which is effectively adaptive to drifting concepts. The proposed algorithm combines learners obtained from each batch, and the combining weights are estimated by ridge regression with the constraints of nonnegativity and being sum to one. When a rational measure of concept drifts is defined as the angle between the estimated and the equal weights which is optimal under no concept drift, it is shown that the ridge parameter plays a crucial role of forcing the proposed algorithm adaptive to concept drifts. Main results include that (i) the algorithm can achieve the optimal weights in the case of no concept drifts if the ridge parameter is large, and (ii) the angle is monotonically increasing as the ridge parameter decreases. This implies that if the ridge parameter is well-controlled, the algorithm can produce weights which reflect the extent of concept drifts measured by the angle.

## **Penalized Likelihood Principal Component Rotation**

**Trevor Park**

University of Florida, USA

tpark@stat.ufl.edu

Principal component analysis provides a ready multivariate exploratory tool for high-dimensional data. However, principal components based on limited sample sizes are subject to high sampling variation that can obscure straightforward interpretations. Ad hoc techniques like varimax rotation can enhance interpretability, at the expense of fidelity to the data. We instead use rotation criteria as penalty functions in a maximum penalized likelihood setting. Desirable features of this approach include a smooth continuum of possible rotations, preferential rotation of components that are poorly defined, and a way to measure fidelity of rotated components to the data. Computations are made possible by special algorithms for orthogonality-constrained optimization.

## **Anomaly Detection in Space-Time Point Processes**

**Michael Porter**

University of Virginia, USA

mdp2u@virginia.edu

Anomalous behavior of a process refers to its tendency to deviate beyond what is expected. Real-time anomaly detection in space-time point processes refers to finding a spatial region and time point where the behavior of the process exceeds some threshold. In applications, this can be accomplished by forecasting the conditional intensity of the point process and then examining the process residuals for anomalous behavior. The martingale property of the forecasted residual process is utilized to detect the significant anomalies

## **Functional Mixed-Effects Models for Periodic Data**

**Li Qin**

Fred Hutchison Cancer Research Center, USA

lqin@fhcrc.org

Periodic data are frequently collected in biomedical experiments. We consider the underlying periodic curves giving rise to these data, and account for the periodicity in their functional model to improve estimation and inference. We propose to incorporate the periodic constraint in the functional mixed effects model setting. Both the fixed functional effects and random functional effects are modeled in the same periodic functional space, hence the population-average estimates and subject-specific predictions are all periodic. This model can be rewritten in multivariate state space form and estimated by an  $O(N)$  modified Kalman filtering and smoothing algorithm. The proposed method is evaluated in different scenarios through simulations. Treatments to none-full period data and missing observations along the period are also given. Analysis of a cortisol data set obtained from a study on fibromyalgia is conducted as illustration.

## **Impact of the random effect distribution on inference for mean in linear mixed models**

**Joshua Rushton**

Cornell University, USA

jdr66@cornell.edu

In this paper we study the impact of various random effect distributions on inference for the mean in simple mixed linear models. We establish precise asymptotic behavior of the tail probabilities of the mean and use this to study the Bahadur efficiency, Pitman efficiency, and vanishing shortcomings of tests and estimators. These concepts are then used to bring out various aspects of inference that are impacted by the distributional assumptions of the random effects. The main technical tools involve large deviations for normalized and self-normalized sums. Several examples and extensions are also provided.

## **Forecasting and Dynamic Updating of Time Series of Curves**

**Haipeng Shen**

University of North Carolina, Chapel Hill, USA

haipeng@email.unc.edu

The set of ideas called 'Functional Data Analysis' has been very productive in analyzing a wide array of complex modern data sets, such as populations of curves, images and shapes of human body parts. We extend these ideas to the case of a time series of curves, and develop time series models of functional data and new methods for forecasting and dynamic updating of curves. The motivation comes from a time series of daily customer service call volume profiles in a network of US banking call centers. Both inter-day forecasting and dynamic

intra-day updating of call volumes are needed for staffing and scheduling purposes. Our approach starts with dimension reduction through functional Principal Component Analysis, which is achieved via a regularized low rank approximation technique. The time series of curves is then represented as a linear combination of several functional principal components plus error. One or multistep ahead forecast of a curve can be obtained using the principal components and time series forecasts of their coefficient series. To achieve dynamic updating within a curve, a shrinkage approach is proposed to combine information from the previous curves and the early part of the current curve. A data-driven mechanism for selecting the shrinkage parameter is also developed, and appears to work well empirically. The methods developed are illustrated via the call center application, and show substantial improvements over existing industry standards in an out-of-sample forecast comparison. Other potential applications are discussed as well.

### **Gaussian process models for a sphere, with application to Faraday Rotation Measures.**

**Margaret Short**

Los Alamos National Laboratory, USA

[mbshort@lanl.gov](mailto:mbshort@lanl.gov)

Our primary goal is to obtain a smoothed summary estimate of the magnetic field generated in and near to the Milky Way by using Faraday rotation measures (RMs). The ability to estimate the magnetic field generated locally by our galaxy and its environs will help astronomers distinguish local versus distant properties of the universe. Each RM in our data set provides an integrated measure of the effect of the magnetic field along the entire line of sight to an extragalactic radio source. RMs can be considered prototypical of geostatistical data on a sphere. In order to model such data, we employ a Bayesian process convolution approach which uses Markov chain Monte Carlo (MCMC) for estimation and prediction. Complications arise due to contamination in the RM measurements, and we resolve these by means of a mixture prior on the errors. This represents joint work with D. Higdon and P. Kronberg.

### **MCMC Linkage Analysis for Two Genes and a Polygenic Component on General Pedigrees**

**Yun Ju Sung**

University of Washington, USA

[yunju@u.washington.edu](mailto:yunju@u.washington.edu)

Linkage analysis involves statistical inference about the location of genes influencing a trait, using trait and genetic marker data collected on families. We describe a new approach, implemented in a computer program, for parametric linkage analysis with a quantitative trait model having one or two genes and a polygenic component, which models additional familial correlation from other unlinked genes. Competing programs use simpler models: one gene, one gene plus a polygenic component, or a crude approximation to the two gene model. Using simple models when they are incorrect, as for complex traits that are influenced by multiple genes, can bias estimates and reduce power to detect linkage. Simulated examples, with various sizes of pedigrees, show that two-gene analysis correctly identifies the location of both genes, whereas other analyses based on simpler models fail to identify the location of genes with modest contributions. We compute the likelihood with MCMC realization of segregation indicators at hypothesized gene locations conditional on marker data, summation over phased multilocus genotypes of founders, and peeling of the polygenic component. This is the first program for two genes and a polygenic component. It has no restriction on number of markers or complexity of pedigrees, facilitating use of more complex models with general pedigrees. This is joint work with Elizabeth Thompson and Ellen Wijsman.

### **Estimation of Wood Fibre Length Distributions from Censored Data through an EM Algorithm**

**Ingrid Svensson**

Umeaa Universitet, Sweden

[ingrid.svensson@math.umu.se](mailto:ingrid.svensson@math.umu.se)

An EM algorithm is proposed to find fibre length distributions in standing trees. The available data comes from cylindrical wood samples (increment cores). The sample contains uncut fibres as well as fibres cut once or twice. Moreover the sample contains not only the fibres of interest (tracheids), but also other cells, so called 'fines'. The method proposed to estimate the length distributions is adapted to the situation where the lengths are measured by an automatic optical fibre-analyser. The fibre-analyser is not able to distinguish fines from fibres and cannot tell if a cell has been cut. The resulting data thus come from a censored version of a mixture

of the fine and fibre length distributions in the tree. The parameters of the length distributions are estimated by a stochastic version of the EM algorithm. The method is applied to data from Northern Sweden. To evaluate the performance of the method, a simulation study is presented. The method works well for sample sizes commonly obtained from an increment core.

### **Feature identification, quantitation and sample size: a comparative analysis of workflows for LC-MS based proteomics**

**Olga Vitek**

Institute for Systems Biology, USA

[ovitek@systemsbiology.org](mailto:ovitek@systemsbiology.org)

Liquid chromatography coupled to mass spectrometry (LC-MS) offers great promises for global protein profiling, and for discovery of biomarkers. The approach compares volumes of features in LC-MS spectra to determine elution time, mass and charge of differentially expressed peptide ions. The characteristics are subsequently used for identification of disease-related proteins, and for classification of biological samples. The unknown identities of LC-MS features, and variation in elution profiles present considerable challenges in analysis of these data. At the same time, assessment of signal processing tools is difficult because there is no gold standard measurements of sample composition. I will present a framework for comparing the statistical characteristics of LC-MS workflows, such as the sensitivity and specificity of signal and noise classification, and the bias and variance of quantitation of peptide ion abundance, in the absence of gold standard. I will also discuss aspects of experimental design, in particular calculations of sample sizes for LC-MS experiments, that take into account the unknown number and identity of features.

### **Use of Geometric Programming in Statistics**

**Xinlei Wang**

Southern Methodist University, USA

[swang@mail.smu.edu](mailto:swang@mail.smu.edu)

A geometric program (GP) is a type of nonlinear optimization problem characterized by objective and constraint functions that have a special form. Recently developed methods can solve even large-scale GPs extremely efficiently and reliably; meanwhile a number of practical problems, particularly in circuit design, have been found to be equivalent to (or well approximated by) GPs. These have aroused considerable recent interest in GPs in fields of Electrical Engineering and Operation Research. However, neither of these has been widely recognized in our statistical community. In this research, we will explore use of geometric programming to solve several long-existing estimation problems in Statistics and compare this approach with existing methods. We also search for other potential GP problems in Statistics.

### **Summer jobs and the ensuing labor market achievement: A quasi-experimental evaluation**

**Yu Wang**

Hogskolan Dalarna, Sweden

[iris\\_wangyu@hotmail.com](mailto:iris_wangyu@hotmail.com)

This paper investigates effects of high-school summer job experiences on individuals future labor market outcomes using quasi-experimental data from Sweden. During the middle 1990s, the local government of Falun municipality, a mid-size town in central Sweden, randomly allocated the summer job positions to high school student applicants through a lottery. This practice provided a unique opportunity to examine the treatment effects of summer job experiences on high school students in a quasi-experimental set-up, which provides good control of the self-selection biases in the summer job participation. Both Intention-to-treat (ITT) analysis and On-treatment (OT) analysis are employed to identify the treatment effects of summer job participations on participants earnings when they entered the labor market. Our study finds that, the participations of summer job do help to improve ones earnings at the initial period when the students entered the labor market but such effect vanishes soon as the students stay long time in the labor market. So this paper suggests that there are no persistent effects of summer job participation on participants future earnings. This observation seems to imply that, summer job experiences of high school students in Sweden only provide port advantage to these participants via channels such as early labor market contacts but did not evidently improve participant productivity required by the Swedish labor market. This is joint work with Kenneth Carling and Ola Naas.

## **Correction of Bias Due to Assay Dilution Effect in Immunogenicity Assessment**

**Yue Wang**

Merck, USA

yue\_wang2@merck.com

Many vaccines provide protection against diseases by inducing active immunity. Therefore, immunogenicity is an important factor in the evaluation of a vaccine. The level of antibody (titers) induced by a vaccine is typically measured in a quantitative bioassay. For bioassays that utilize a reference standard against which test sample concentrations are calibrated, a dilution effect (i.e., a dependence of dilution corrected titer on dilution) may be observed. Depending upon an assay's quantifiable range and the range in immune response, a portion of samples may have pre-titers and post-titers obtained at different dilutions. In these cases, a bias in GMT and in fold-rise response (comparing post-titers to pre-titers) would be introduced due to the dilution effect. Due to the inherent biological heterogeneity, a sample-to-sample variation in the magnitude of the dilution effect usually exists. However, when bias correction is performed, typically little attention has been paid to account for the effect of the sample-to-sample variation. This is partly due to the fact that dilution effect characterized in typical assay validation is the average dilution effect while the variability at the sample level is often not characterized and hence often ignored in clinical assessment. We propose to characterize this sample variation in dilution effect using a mixed model approach and to derive the appropriate variance estimates corresponding to the bias corrected estimates of GMT and fold-rise.

## **Bivariate growth charts**

**Ying Wei**

Columbia University, USA

yw2148@columbia.edu

Growth charts have been widely used in clinics and medical centers to monitor an individual subjects growth or health status in context of population values. Current growth charts consider only one measurement at a time. Instead of single measurement, health evaluation very often evolves a pair of measurements, height and weight for example. More informative readings can be obtained by screening multiple measurements simultaneously(jointly). We first propose new definitions of bivariate growth chart that are mathematically formal but clinically sensible. The charts, consisting of a sequence of nested two-dimensional reference contours, are time/age dependent and incorporate with other potential covariate effects (e.g. past growth). Estimation of such charts based upon quantile regression was also provided, and their performance was demonstrated by a Monte-Carlo simulation study, as well as the applications with real height-weight data set.

## **Estimation of High Dimensional Predictive Densities**

**Xinyi Xu**

The Ohio State University, USA

xinyi@stat.ohio-state.edu

Commonly used statistical approaches to prediction provide a single number as a forecast of an unknown future quantity, sometimes attaching an error bound to convey the uncertainty of the prediction. A more comprehensive approach to prediction provides a complete predictive estimate that assigns probabilities to every possible outcome that may occur. Because they are more comprehensive, such descriptions of uncertainty lead to better decision making and sharper assessment of risks. In this talk, the problem of estimating the predictive density of a multivariate normal variable under Kullback-Leibler loss is considered. We show that there exist broad classes of formal Bayes rules, including Bayes rules under superharmonic priors, which dominate the best invariant minimax estimator for this problem. We also show that the class of generalized Bayes estimators is a complete class, and obtain sufficient conditions for the admissibility of formal Bayes rules. Fundamental similarities and differences with the parallel theory of estimating a multivariate normal mean under quadratic loss are described throughout.

## **Robust Prediction and Extrapolation Designs for Censored Data**

**Xiaojian (Jennifer) Xu**

University of Alberta, Canada

xiaojian@ualberta.ca

In this paper we present the construction of designs for both response prediction and extrapolation with a possibly misspecified generalized linear regression model when the data are censored. The minimax gdesigns

are found for maximum likelihood estimation in the context of both prediction and extrapolation problems in case of with or without restraint of design unbiasedness. This paper extends preceding work of robust designs for complete data by incorporating censoring and maximum likelihood estimation. It also broadens former work of robust designs for censored data from others by considering both nonlinearity and much more arbitrary uncertainty in fitted regression response and by dropping all restrictions on the structure of regressors. Solutions are derived by a nonsmooth optimization technique analytically and given in full generality. A typical example in accelerated life testing is also demonstrated.

### **Efficient polynomial spline estimation of partially linear models for clustered data**

**Lan Xue**

Oregon State University, USA

xuel@stat.oregonstate.edu

We consider estimation of the partially linear models for clustered data using polynomial spline smoothing. The estimation procedure characterizes the infinitely dimensional nonparametric function by a slowly growing number of parameters. Thus the computation is comparable to parametric least squares. On the other hand, it incorporates the within cluster correlation properly. The resulting estimators are a ‘polynomial spline version’ of both the profile-kernel (PK) estimators (Lin & Carroll 2001) and backfitting (BF) estimators (Zeger & Diggle 1994), replacing kernel smoothing by polynomial spline smoothing, and have the same asymptotic property as the PK estimators. Simulated example demonstrates that the proposed estimators are computationally efficient and also as accurate as the PK estimators. Application to milk protein content data is described.

### **Bayesian methodology which accounts for uncertainty about the commonality of a set of small area parameters**

**Guofen Yan**

University of Virginia, USA

guofen.yan@virginia.edu

We describe and evaluate Bayesian methodology to improve inference for ‘small areas.’ Inference for each small area will be improved by pooling data from other, like, entities. However, the pooled data must be concordant with that from the small area of direct interest. Our methodology ensures this concordance while the methods in current use may not. We show this using a set of samples.

### **Penalized Spline Models for Functional Principal Component Analysis**

**Fang Yao**

Colorado State University, USA

fyao@stat.colostate.edu

In this talk we propose an iterative estimation procedure for performing functional principal component analysis. The procedure aims at functional or longitudinal data where the repeated measurements from the same subject are correlated. An increasingly popular smoothing approach, penalized spline regression, is used to represent the group mean trends. This allows straightforward incorporation of covariates and simple implementation of inference procedures for coefficients. For the handling of the within-subject correlation, we develop an iterative procedure which would gradually reduce the dependence amongst the repeated measurements made for the same subject. The resulting data after iteration are theoretically shown to be asymptotically independent, which suggests that the general theory of penalized spline regression developed for independent data can also be applied to functional data. The effectiveness of the proposed procedure is demonstrated via a simulation study and an application to yeast cell cycle data.

### **Weibull Prediction of Event Times in Randomized Clinical Trials**

**Gui-Shuang Ying**

University of Pennsylvania, USA

gsying@mail.med.upenn.edu

In clinical trials with planned interim analyses, it can be valuable for a variety of reasons to predict the times of landmark events in advance of their occurrence. Bagiella and Heitjan (Statistics in Medicine 2001; 20:2055-2063) proposed a parametric prediction model for failure-time outcomes assuming exponential survival and Poisson enrollment. There is concern that predictions from their model may be inaccurate if distributional assumptions are wrong. Ying, Heitjan and Chen (Clinical Trials 2004; 1:352-361) proposed a nonparametric method for point and interval prediction that involves sampling survival times from bootstrapped Kaplan-Meier

curves. This approach is robust to distributional assumptions but suffers from potential loss of efficiency. A middle course would be to base predictions on a more flexible family of parametric survival models. This paper describes a generalization of the Bagella-Heitjan exponential survival model to a two-parameter Weibull model. The survival probability in the future is estimated from the available data and prior information on the values of the Weibull parameters. For interval prediction, we use sampling-importance-sampling to approximate the posterior distribution of future event times, from which we generate the predictive distribution of landmark event times. Monte Carlo studies show that the Weibull model provides valid and efficient predictions for Weibull and gamma data, but is generally biased when the underlying survival is lognormal. We demonstrate the methods using data from a trial of immunotherapy for chronic granulomatous disease.

## **A Hybrid Newton-Type Method for Censored Survival Data Using Double Weights in Linear Models**

**Menggang Yu**

Indiana University, USA

meyu@iupui.edu

As an alternative to the Cox model, the rank-based estimating method for censored survival data has been studied intensively since it was proposed by Tsiatis (1990) among others. Due to the discontinuity feature of the estimating function, a significant amount of work in the literature has been focused on numerical issues. In this article, we consider the computational aspect of a family of the doubly weighted rank-based estimating functions. This family is rich enough to include both estimating functions of Tsiatis (1990) for the randomly observed data and of Nan, Yu, and Kalbfleisch (2004) for the case-cohort data as special examples. The later belongs to the biased sampling problems. We show that the doubly weighted rank-based discontinuous estimating functions are monotone, a property established for the randomly observed data in the literature, when the generalized Gehan-type weights are used in addition to the subject-specific weights. Though the estimating problem can be formulated to a linear programming problem as that for the randomly observed data, due to its easily uncontrollable large scale even for a moderate sample size, we instead propose a Newton-type iterated method to search for an approximate solution of the (system of) discontinuous monotone estimating equation(s). Simulation results provide a good demonstration of the proposed method.

## **Model Selection and Estimation in the Gaussian Graphical Model**

**Ming Yuan**

Georgia Institute of Technology, USA

myuan@isye.gatech.edu

We propose a penalized likelihood method for estimating the concentration matrix in the Gaussian graphical model. The method leads to a sparse and shrinkage estimate of the concentration matrix that is positive definite, thus conducts model selection and estimation simultaneously in the graphical model. The implementation of the method is non-trivial due to the positive definite constraint on the concentration matrix, but we show that the computation can be done effectively by taking advantage of the maxdet algorithm developed in convex optimization. We propose a BIC type criterion for the selection of the tuning parameter in the penalized likelihood method. The connection between our method and existing methods is illustrated. Simulations and real examples demonstrate the competitive performance of the new method.

## **Conditional Properties of a Parametric Bootstrap**

**Russell Zaretzki**

University of Tennessee, USA

[rzaretzk@utk.edu](mailto:rzaretzk@utk.edu)

DiCiccio, Martin and Stern(2001) introduced the parametric bootstrap of the signed root statistic as a useful computational alternative to analytic approximations when highly accurate statistical inference is desired. This performance is equivalent to asymptotic techniques such as the  $r$  formula; see Barndorff-Nielson(1986). In addition, simulation examples contained in DiCiccio (2001) suggest that this bootstrap technique can produce extremely accurate conditional inferences. The present work further investigates these conditional properties. In particular, we prove that, in the presence of nuisance parameters, inferences based on a parametric bootstrap of the signed root are conditionally accurate to  $O_p(n^{-1})$ .

## **Nonparametric estimation of the dependence function for a multivariate extreme value distribution**

**Dabao Zhang**

Purdue University, USA

[zhangdb@stat.purdue.edu](mailto:zhangdb@stat.purdue.edu)

Understanding and modeling dependence structures among multivariate extreme values are of interest in a number of application areas. One of well known approaches is to investigate the Pickands dependence function. In the bivariate setting, there exist several estimators for estimating the Pickands dependence function which assume known marginal distributions (Pickands, 1981; Deheuvels, 1991; Hall and Tajvidi, 2000; and Capraua, Fougueres and Genest, 1997). In this paper, we generalize the bivariate results to  $p$ -variate multivariate extreme value distributions with  $p \geq 2$ . We demonstrate that the proposed estimators are consistent and asymptotically normal as well as have excellent small sample behavior. This is a joint work with Martin T. Wells and Liang Peng.

## **Causal Inference, Sequential Monte Carlo and Clustering**

**Junni Zhang**

Peking University, China

[zjn@gsm.pku.edu.cn](mailto:zjn@gsm.pku.edu.cn)

After graduation, my research has been concentrated on causal inference, sequential Monte Carlo (SMC) and clustering. In research on causal inference, I'm mostly interested in making inferences about causal effects when some outcomes are 'truncated by death.' Together with my collaborators, we have developed a principal stratification approach and showed that this approach can help researchers extract more detailed information from the data compared with traditional methods. My research on SMC has been focused on developing new variants of SMC methods that can improve statistical inference. Together with my collaborators, we have developed a new set of SMC methods called the Independent Particle Filters, which can better deal with stochastic dynamic systems in which the current observation provides significant information about the current state whereas the state dynamics is weak. We have also borrowed ideas from both SMC and Markov Chain Monte Carlo to develop lookahead and piloting strategies for variable selection. In research on clustering, I'm interested in clustering in some non-conventional settings, which I called relation-based clustering and hybrid clustering. In relation-based clustering, for each object, various attributes and some response variables are observed, and we can cluster the objects using heterogeneity in the relation between the response variables and the attributes. In hybrid clustering, we can cluster the objects using heterogeneity in both the distribution of attributes, and the relation between the response variables and the attributes.

## **Flexible And Empirical Bayes Estimator For High Dimensional Data: Sparseness And Asymmetry**

**Min Zhang**

Purdue University, USA

[minzhang@purdue.edu](mailto:minzhang@purdue.edu)

By assuming that signals in high dimensional data are symmetric, classical shrinkage estimators have been constructed such that thresholding rules have the same sized thresholds for both positive and negative observed values. However, the high dimensional data from genomic or proteomic research usually present sparse and asymmetric signals. To account for the sparseness and asymmetry of signals in high dimensional data,

we proposed a new class of estimators called generalized shrinkage estimator that can be constructed under a hierarchical Bayes framework. Empirical Bayes estimators are further developed to allow its adaptation to the sparseness of the signals in high dimensional noisy data. Under mild assumptions, the new estimators are shown to have nice theoretical properties. A simulation study shows the superior performance of the newly proposed estimators, and their strengths are demonstrated by an application to microarray data.

### **A Comprehensive Spatial-Temporal Analysis of Breast Cancer: First Primary, Second Primary and Breast Cancer Survival**

**Song Zhang**

The University of Texas M. D. Anderson Cancer Center, USA [yszhang@wotan.mdacc.tmc.edu](mailto:yszhang@wotan.mdacc.tmc.edu)

We propose a comprehensive spatial-temporal analysis of breast cancer, including first primary occurrence, second primary occurrence and breast cancer survival. We assume a Poisson model for first primary breast cancer incidence. We are interested in two competing risks after first primary occurrence: developing second primary occurrence or death from first primary breast cancer. For each risk, we define a semiparametric age-hazard function. Although each subject was observed over a relatively short period, the age-hazard function allows us to piece together information from subjects of different ages and construct a much longer hazard curve. The unknown baseline age-hazard curve is modelled by a Bayesian P-spline prior, with the assumption that the age-hazard rate changes smoothly over time. Random spatial and temporal effects are assumed for the log-rates of the first primary occurrence, and for the two risk-specific age-hazard functions. We model these random effects with multivariate conditionally autoregressive priors to allow borrowing strength. This model is evaluated through a simulation study and applied to a breast cancer dataset from the Iowa SEER program.

### **Nonparametric Estimation of Survival Functions Conditional on Sparsely Observed Covariate Processes**

**Ying Zhang**

University of Minnesota, USA [yingz@biostat.umn.edu](mailto:yingz@biostat.umn.edu)

The problem of estimating a distribution function conditional on a scalar or vector predictor has been studied since the seminal work of Beran (1981). We consider an extension where the covariates are not vector-valued but rather are random trajectories of some longitudinal characteristic of interest and focus on the practically relevant case where such longitudinal trajectories are sampled irregularly and with measurement error. Our method aims at irregularly spaced longitudinal data, where the number of repeated measurements available per subject may be small. Functional principal component analysis (FPCA) is used to recover the underlying trajectory based on the available measurements. The predictor trajectories up to current time are represented by time-varying functional principal component scores, which are continuously updated as time progresses. Distances between two predictor trajectories are defined in terms of their functional principal component scores, and used to define local neighborhoods that form the basis for the proposed conditional Kaplan-Meier estimates. The size of these neighborhoods is determined by maximizing a likelihood. The proposed methods are illustrated with survival data on liver cirrhosis with prothrombin index and also with primary biliary cirrhosis data.

### **Case-Control Studies With Longitudinal Covariates**

**Honghong Zhou**

Indiana University, USA [zhouh@iupui.edu](mailto:zhouh@iupui.edu)

The case-control design is commonly used for studying rare diseases. In a case-control study, subjects are recruited according to their case or control status, and thus the sampling is outcome-based, retrospective and biased. An emerging problem in case-control studies is the presence of a longitudinal covariate, which is collected longitudinally but retrospectively. We consider case-control studies with longitudinal covariates. We propose a logistic model coupled with a linear mixed model to jointly model the binary outcome and the longitudinal covariate. Specifically, we assume that the longitudinal covariates follow a linear mixed model and the primary binary outcome relates to the longitudinal covariate through latent subject specific random effects, for example, individual baselines and slopes. We study several estimation procedures, such as the Two-stage

method, the Best Linear Unbiased Predictor (BLUP) method, the Maximum Likelihood Estimation (MLE) based on the true retrospective likelihood, and the Sufficiency Score method. The asymptotic properties of these methods are also discussed. The proposed methods are illustrated through simulation studies and by application to a case-control study of breast-cancer in postmenopausal women with weight as the longitudinal covariate. This is joint work with Xihong Lin and Bin Nan.

### **Constrained Dimension Reduction Based on CANCOR**

**Jianhui Zhou**

University of Virginia, USA

jz9p@virginia.edu

'The curse of dimensionality' makes high-dimensional data analysis usually challenging. The sliced inverse regression (SIR) and canonical correlation (CANCOR) methods reduce the dimensionality of data by replacing the explanatory variables with a small number of composite directions without losing much information. However, the composite directions estimated by SIR or CANCOR often involve all of the variables, which makes them difficult to interpret and sometimes meaningless in the following up analyses. To simplify the direction estimates, Ni, Cook and Tsai (2005) proposed the shrinkage sliced inverse regression (SSIR) method based on SIR. In this paper, we propose the constrained canonical correlation (CCANCOR) method based on CANCOR. Each direction estimated by CCANCOR consists of only a subset of the variables and is easier to interpret. When there are only a few variables involved in the model, CCANCOR identifies them precisely while SIR and CANCOR usually select more due to the noise in the data or correlations among the variables. The direction estimates by CCANCOR are shown to be consistent. The simulation studies show that the proposed CCANCOR method performs better than the SSIR method. The Boston housing data is analyzed by CCANCOR as an example.

### **Estimate of regression coefficients based on the projection depth-weighted scatter matrix**

**Weihua Zhou**

University of North Carolina, Charlotte, USA

wzhou2@uncc.edu

A new estimator of the regression parameters is introduced in a multivariate regression model. The affine equivariant estimate is based on the projection depth-weighted mean and scatter estimators. The influence function, finite breakdown and asymptotic theorem are developed to consider robustness and limiting efficiencies of this new estimate. The estimate is shown to be consistent with a limiting multivariate normal distribution. The influence function, as a function of the length of the contaminated vector, is shown to be bounded in elliptic cases. The new estimate is highly efficient in the multivariate normal case and it is also highly robust. Simulations are used to consider finite sample efficiencies with similar results.

### **The $F_\infty$ norm Support Vector Machine**

**Hui Zou**

University of Minnesota, USA

hzou@stat.umn.edu

We propose a new support vector machine (SVM), the  $F_\infty$  SVM, to perform automatic factor selection in classification. The  $F_\infty$  SVM methodology is motivated by the feature selection problem in cases where the input features are generated by factors, and the model is best interpreted in terms of significant factors. This type of problem arises naturally when a set of dummy variables are used to represent a categorical factor and/or a set of basis functions of a continuous variable are included in the predictor set. In problems without such obvious group information, we propose to first create groups among features by clustering, and then apply the  $F_\infty$  SVM. We show that the  $F_\infty$  SVM is equivalent to a linear programming problem and can be efficiently solved using the standard linear programming technique. Analysis on simulated and real world data shows that the  $F_\infty$  SVM enjoys competitive performance when compared with the 1-norm and 2-norm SVMs.