**2014 Conference of Texas Statisticians**
Department of Mathematical Sciences
University of Texas at Dallas

**Program**

**Friday March 21, 2014**

- **Noon-1:00 pm**  
  **Registration**  
  Founders Building (FO), 2nd Floor Atrium

- **1:00-1:30**  
  **Welcome & Opening Remarks**  
  Kusch Auditorium, Founders North Building (FN), 2.102

**Invited Talks**
Kusch Auditorium, FN, 2.102

- **1:30-2:20**  
  **Joan Staniswalis**, UTEP  
  “Low-Dimensional Approximations for Functional Data with Covariates”

- **2:30-3:20**  
  **Yulia Gel**, UTD  
  “The Role of Modern Social Media Data in Surveillance and Prediction of Infectious Diseases: from Time Series to Networks”

- **3:30-4:00**  
  **Break**  
  (FO, 2nd Floor Atrium)

- **4:00-4:50**  
  **Peter Thall**, MD Anderson  
  “Bayesian Clinical Trial Design: 23 Years of Theory and Practice”

- **5:00-5:45**  
  **COTS Business Meeting**  
  Kusch Auditorium, FN, 2.102

- **5:30-6:30**  
  **Student Poster Session**  
  FO, 3rd Floor Atrium  
  &  
  **Social Hour**  
  FO, 2nd Floor Atrium & 3rd Floor Atrium

- **6:30-8:30**  
  **Buffet Dinner & Don Owen Award**  
  FO, 2nd Floor Atrium
Saturday, March 22, 2014

- 8:15-8:45 am **Open Poster Session**
  FO, 3rd Floor Atrium

**Invited Talks**
Kusch Auditorium, FN, 2.102

- 8:50-9:10 **Pradeep Ravikumar, UT-Austin**
  “Learning Graphs with a Few Hubs”

- 9:15-9:35 **Michael Schweinberger, Rice**
  “Local Dependence in Random Graph Models: Characterization, Properties, and Statistical Inference”

- 9:40-10:00 **Leif Ellingson, TTU**
  “Nonparametric Estimation of the Extrinsic Mean Shape of Planar Contours”

- 10:05-10:25 **Bhargab Chattophadyay, UTD**
  “Sample Size for Estimation of Coefficient of Variation with Minimum Risk”

- 10:25-10:50 **Break**
  FO, 2nd Floor Atrium

- 10:50-11:10 **Qiongxia Song, UTD**

- 11:15-11:35 **Xinlei (Sherry) Wang, SMU**
  “Bayesian Joint Modeling for Integrative Gene Set Enrichment Analysis Involving RNA-Seq”

- 11:40-12:00 **Swati Biswas, UTD**
  “Detecting Rare Haplotype-Environment Interaction with Logistic Bayesian LASSO”

- 12:05-12:25 pm **Min Chen, UTD**
  “Detecting Epistatic SNPs Associated with Complex Diseases via a Bayesian Classification Tree Search Method”

- 12:30-12:40 **Closing Remarks**
<table>
<thead>
<tr>
<th>Faculty/Professionals (non-UTD)</th>
<th>Students (non-UTD)</th>
<th>Faculty (UTD)</th>
<th>Students (UTD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Das, Kumer (Lamar)</td>
<td>Alshaikh, Ali (SMU)</td>
<td>Ammann, Larry</td>
<td>Acosta-Mejia, Cesar</td>
</tr>
<tr>
<td>Ellingson, Leif (Texas Tech)</td>
<td>Asbury, Taylor (SHSU)</td>
<td>Baron, Michael</td>
<td>Cao, Yuying</td>
</tr>
<tr>
<td>Fuentes, Jesus (Incarnate Word)</td>
<td>Bai, Ou (SMU)</td>
<td>Biswas, Swati</td>
<td>Cao, Rukai</td>
</tr>
<tr>
<td>Hallum, Cecil (SHSU)</td>
<td>Barnard, Ben (Baylor)</td>
<td>Chattopadhyay, Bhargab</td>
<td>Carcea, Marcel</td>
</tr>
<tr>
<td>Holt, Melinda (SHSU)</td>
<td>Boling, Chelsea (Lamar)</td>
<td>Chen, Min</td>
<td>Chen, Ming</td>
</tr>
<tr>
<td>Hu, Jianhua (MD Anderson)</td>
<td>Buros, Amy (Baylor)</td>
<td>Choudhary, Pankaj</td>
<td>Chowdhury, Marzana</td>
</tr>
<tr>
<td>Kahle, David (Baylor)</td>
<td>Burrow, Whitney (Baylor)</td>
<td>Dearing, Ron</td>
<td>Chu, Jufen</td>
</tr>
<tr>
<td>Kalanka, Jayalath (SFASU)</td>
<td>Cheng, Joyce (Baylor)</td>
<td>Gel, Yulia</td>
<td>Darku, Francis</td>
</tr>
<tr>
<td>Ravikumar, Pradeep (UT Austin)</td>
<td>Chowdhury, MD (Lamar)</td>
<td>Serfling, Bob</td>
<td>Datta, Ananda</td>
</tr>
<tr>
<td>Sager, Tom (UT Austin)</td>
<td>Dey, Asim (Lamar)</td>
<td>Song, Qiongxia</td>
<td>Dong, Bo</td>
</tr>
<tr>
<td>Scariano, Stephen M. (SHSU)</td>
<td>Drevets, Madeline (Baylor)</td>
<td></td>
<td>Huang, Xin</td>
</tr>
<tr>
<td>Schoolfield, John (UTSA)</td>
<td>Edwards, Audrene (Lamar)</td>
<td></td>
<td>Kotinkaduwa, Lak</td>
</tr>
<tr>
<td>Schweinberger, Michael (Rice)</td>
<td>Eschmann, Mark (Baylor)</td>
<td></td>
<td>Nawarathna, Lakshika</td>
</tr>
<tr>
<td>Song, Joon (Baylor)</td>
<td>Gates, Amber (SHSU)</td>
<td></td>
<td>Ngo, Hoai Ngoc</td>
</tr>
<tr>
<td>Staniswalis, Joan (UTEP)</td>
<td>Halder, Shaymal (Lamar)</td>
<td></td>
<td>Rathnayake, Lasitha</td>
</tr>
<tr>
<td>Stein, Caleb (UArk for Med Sci)</td>
<td>Hapuwitharana, Janitha (Lamar)</td>
<td></td>
<td>Smirnova, Ekaterina</td>
</tr>
<tr>
<td>Stokes, Lynne (SMU)</td>
<td>Hartnett, Casey (SHSU)</td>
<td></td>
<td>Tian, Yahui</td>
</tr>
<tr>
<td>Sun, Shuying (Texas State)</td>
<td>He, Sha (SMU)</td>
<td></td>
<td>Wang, Cheng</td>
</tr>
<tr>
<td>Thalli, Peter (MD Anderson)</td>
<td>Karunarathna, Charith (SHSU)</td>
<td></td>
<td>Wang, Shanshan</td>
</tr>
<tr>
<td>Wang, Xinlei (Sherry) (SMU)</td>
<td>Liu, Yang (SHSU)</td>
<td></td>
<td>Wang, Tiansong</td>
</tr>
<tr>
<td>Yu, Donghyeon (UTSW)</td>
<td>Liu, Yuhang (SMU)</td>
<td></td>
<td>Wang, Yunfei</td>
</tr>
<tr>
<td>Yun, Jonghyun (UTSW)</td>
<td>Liu, Zhaocd (SHSU)</td>
<td></td>
<td>Wijesuriya, Uditha</td>
</tr>
<tr>
<td></td>
<td>Liyanarachchi, Prasansa (SHSU)</td>
<td></td>
<td>Wu, Jiayi</td>
</tr>
<tr>
<td></td>
<td>Lou, Ying (SMU)</td>
<td></td>
<td>Zhang, Yuan</td>
</tr>
<tr>
<td></td>
<td>Marcovitz, Michelle (Baylor)</td>
<td></td>
<td>Zhang, Lei</td>
</tr>
<tr>
<td></td>
<td>Nath, Gopal (Lamar)</td>
<td></td>
<td>Zhao, Tian</td>
</tr>
<tr>
<td></td>
<td>Nelson, Tyler (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nieuwoudt, Christina (SHSU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nussbaum, Amy (SMU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>O’Brien, Kelly (Incarnate Word)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Odom, Gabriel (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perez, Maria (Incarnate Word)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sadler, Bivin (SMU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sai, Sun (Incarnate Word)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sams, Chris (Lamar)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Silva, Rajitha (Simon Fraser)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sims, Justin (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tecson, Kristen (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thannehelage, Ruwan (SHSU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vallejo, Jonathon (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waken, RJ (Baylor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wang, Tao (UTSW)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Woroszylo, Casper (Rice)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zhong, Rui (UTSW)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zhu, Xiujun (Sylvia) (SMU)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstracts

Invited Talks (1)
Friday, March 21, 1:30pm-4:50pm

Low-Dimensional Approximations for Functional Data with Covariates

Joan Staniswalis
Department of Mathematical Sciences, University of Texas at El Paso
http://faculty.utep.edu/Default.aspx?alias=faculty.utep.edu/jstaniswalis

Ramsay (1996) first proposed the method of principal differential analysis (PDA) for fitting a differential equation to a collection of noisy data curves. Each data curve is modeled as a (possibly noisy observation of a) smooth curve in the null space of a linear differential operator of order m. Smooth estimates of the coefficient functions defining the linear differential operator are obtained by minimizing a penalized sum of the squared norm of the forcing functions: that part of the data curve that is not annihilated by the linear differential operator. Once the linear differential operator is estimated, a nonparametric basis of functions for the null space is computed using iterative methods. The nonparametric basis functions can be used to provide a smooth low dimensional approximation to the data curves. This paper extends PDA to allow for the coefficients in the linear differential equation to smoothly depend upon continuous subject covariates. This is implemented with local smoothing in the Rsoftware and used to explore how the cortical auditory evoked potentials (CAEP) of subjects with normal hearing change with age. Data-based selection of the smoothing parameters is addressed.

The Role of Modern Social Media Data in Surveillance and Prediction of Infectious Diseases: from Time Series to Networks

Yulia Gel
Department of Mathematical Sciences, University of Texas at Dallas
yxg142030@utdallas.edu

The prompt detection and forecasting of infectious diseases with rapid transmission and high virulence are critical in the effective defense against these diseases. Despite many promising approaches in modern surveillance methodology, the lack of observations for near real-time forecasting is still the key challenge obstructing operational prediction and control of disease dynamics. For instance, even CDC data for well monitored areas in USA are two weeks behind, as it takes time to confirm influenza like illness (ILI) as flu, while two weeks is a substantial time in terms of flu transmission. These limitations have ignited the recent interest in searching for alternative near real-time data sources on the current epidemic state and, in particular, in the wealth of health-related information offered by modern social media. For example, Google Flu Trends uses flu-related searches to predict a future epidemiological state at a local level, and more recently, Twitter has also proven to be a very valuable resource for a wide spectrum of public health applications. In this talk we will review capabilities and limitations of such social media data as early warning indicators of influenza dynamics in conjunction with traditional time series epidemiological models and with more recent random network approaches accounting for heterogeneous social interaction patterns.
Bayesian Clinical Trial Design: 23 Years of Theory and Practice

Peter F. Thall
Department of Biostatistics, M.D. Anderson Cancer Center
http://odin.mdacc.tmc.edu/~pfthall/

A clinical trial is a medical experiment with human subjects. Its two goals are to treat the patients in the trial and find out something useful that may benefit future patients. Despite the immense literature on clinical trial design, there still remains a large gap between the designs in clinical protocols and actual trial conduct. There are many reasons for this, most related to ethics, logistics, and the fact that medical practice often is very complicated. Most of the designs that I have published are Bayesian because, whether they realize it or not, nearly all physicians are Bayesians, and moreover Bayesian methods often are more practical for the sort of trials that I must design in the cancer center where I work. In this talk, I will discuss some designs I have published and various practical difficulties I have encountered during the process of implementing them in actual trials. As time permits, these will include single-arm phase II trials with binary or multivariate outcomes, use of hierarchical or regression models to borrow strength between patient subgroups or disease subtypes, dose-finding designs based on multiple toxicities, efficacy-toxicity trade-offs or joint utilities, a group sequential model-adaptive phase III design, randomized studies of dynamic treatment regimes for multi-stage therapies, and a new design to optimize sedative dose for newborn babies who must be intubated in order to treat respiratory distress syndrome.
Invited Talks (2)
Saturday, March 22, 8:50am-12:25am

Learning Graphs with a Few Hubs
Pradeep Ravikumar
Department of Computer Science, University of Texas, Austin
http://www.cs.utexas.edu/~pradeepr/

We consider the problem of recovering the graph structure of a hub-networked Ising model given iid samples, under high-dimensional settings, where number of nodes $p$ could be potentially larger than the number of samples $n$. By a "hub-networked" graph, we mean a graph with a few "hub nodes" with very large degrees. State of the art estimators for Ising models have a sample complexity that scales polynomially with the maximum node-degree, and are thus ill-suited to recovering such graphs with a few hub nodes. Some recent proposals for specifically recovering hub graphical models do not come with theoretical guarantees, and even empirically provide limited improvements over vanilla Ising model estimators. Here, we show that under such low sample settings, instead of estimating "difficult" components such as hub-neighborhoods, we can use quantitative indicators of our inability to do so, and thereby identify hub-nodes. This simple procedure allows us to recover hub-networked graphs with very strong statistical guarantees even under very low sample settings.

Model-Based Clustering of Large Networks
Michael Schweinberger
Department of Statistics, Rice University
http://www.stat.rice.edu/~ms88/

Dependent phenomena, such as relational, spatial, and temporal phenomena, tend to be characterised by local dependence in the sense that units which are close in a well-defined sense are dependent. In contrast to spatial and temporal phenomena, though, relational phenomena tend to lack a natural neighborhood structure in the sense that it is unknown which units are close and thus dependent. Owing to the challenge of characterizing local dependence and constructing random graph models with local dependence, many conventional exponential-family random graph models induce strong dependence and are not amenable to statistical inference. We take first steps to characterize local dependence in random graph models, inspired by the notion of finite neighborhoods in spatial statistics and M-dependence in time series, and show that local dependence endows random graph models with desirable properties which make them amenable to statistical inference. We discuss how random graph models with local dependence can be constructed by exploiting either observed or unobserved neighborhood structure. In the absence of observed neighborhood structure, we take a Bayesian view and express the uncertainty about the neighborhood structure by specifying a prior on a set of suitable neighborhood structures. We present simulation results and applications to two real-world networks with ground truth.
Nonparametric Estimation of the Extrinsic Mean Shape of Planar Contours

Leif Ellingson
Department of Mathematics and Statistics, Texas Tech University
http://www.math.ttu.edu/~lellings/

With the widespread availability of digital imaging data, the analysis of the shape of planar configurations has grown in importance. The analysis of such data typically follows from D. G. Kendall’s definition of direct similarity shapes of planar k-ads. In recent years, however, the focus has shifted to the study of shapes of planar curves as infinite-dimensional objects. While much work has been done in developing mathematical models for these data objects, the development of methodology for statistical inference is still ongoing. In this talk, we approach this problem by embedding the space of shapes of planar contours, which is a Hilbert manifold, into the space of Hilbert-Schmidt operators in order to perform an extrinsic statistical analysis. We will discuss two nonparametric methods for performing inference for the one sample problem. First, we will utilize a neighborhood hypothesis test for the extrinsic mean shape. Secondly, we will present confidence regions for the extrinsic mean shape obtained through the use of the nonparametric bootstrap.

Sample Size for Estimation of Coefficient of Variation with Minimum Risk

Bhargab Chattopadhyay
Department of Mathematical Sciences, University of Texas at Dallas
http://www.utdallas.edu/~bxc126030/

The coefficient of variation has long been a widely used in psychology. In experimental psychology, the coefficient of variation is used for analyzing data sets observed from experiments related to sleep deprivation or performance stability and others (c.f. Babkoff, Kelly, and Naitoh, 2001; Kelley, 2007, 2012; Hayashi, 2000). In a psychiatry setting, the coefficient of variation has been used to compare patterns of homogeneity/heterogeneity in brain metabolism for Alzheimer’s disease patients with those in a control group without Alzheimer’s (c.f. Volkow et al., 2002). Apart from these studies, coefficient of variation is also used in computing educational inequality (c.f. Monchar, 1981). So, the data in the studies related to these biological or psychological or management fields are often non-negative in nature. In order to compute Coefficient of Variation for a study at any given time, a procedure is needed which will minimize both the error of estimation as well as the cost of sampling without assuming any distribution. When estimating a parameter of interest, estimation error is unavoidable. A typical approach to reducing estimation error is to increase sample size. Increasing sample size yields smaller estimation error, but it also results in increasing cost due to sampling every additional observation. In other words, smaller estimation error comes at the price of increased cost, and vice versa. Thus, a method of estimation should be developed such that the cost of sampling and the estimation error are kept as low as possible. No procedure with fixed sample size can achieve a trade-off between sampling cost and estimation error. Hence, this problem falls in the domain of sequential analysis, where it is known as minimum risk point estimation problem. In this presentation, a minimum risk point estimation procedure of Coefficient of Variation will be presented along with several asymptotic properties like convergence results on final sample size and also the ratio-regret which was proved without any distributional assumption. This will be followed up with a simulation study.

This is joint work with Ken Kelley, University of Notre Dame, Notre Dame, USA
Jump Detection in Time Series Nonparametric Regression Models: a Polynomial Spline Approach

Qiongxia Song
Department of Mathematical Sciences, University of Texas at Dallas
http://www.utdallas.edu/~song/

For time series nonparametric regression models with discontinuities, we propose to use polynomial splines to estimate locations and sizes of jumps in the mean function. Under reasonable conditions, test statistics for the existence of jumps are given and their limiting distributions are derived under the null hypothesis that the mean function is smooth. Simulations are provided to check the powers of the tests. A climate data application and an application to the U.S. unemployment rates of men and women are used to illustrate the performance of the proposed method in practice.

Bayesian Joint Modeling for Integrative Gene Set Enrichment Analysis involving RNA-Seq

Xinlei (Sherry) Wang
Department of Statistical Science, Southern Methodist University
http://faculty.smu.edu/swang/

To understand molecular mechanisms underlying complex human diseases, one important task is to identify groups of related genes that are combinatorially involved in such biological processes, mainly through gene set enrichment analysis (GSEA). In the past, many statistical methods have been developed for GSEA. However, there is very limited literature in its integrative analysis, despite a pressing need in an emerging big data era. In this project, we propose a Bayesian joint modeling approach to combine multiple gene set enrichment studies that involve microarray and/or RNA-seq expression data, which can capture isoform-phenotype relationships, gene-phenotype relationships, isoform-gene relationships, gene-gene interactions, (potential) co-expression within the same gene group in one integrated model, while accounting for between-study heterogeneities explicitly.

Detecting Rare Haplotype-Environment Interaction with Logistic Bayesian LASSO

Swati Biswas
Department of Mathematical Sciences, University of Texas at Dallas
http://www.utdallas.edu/~sxb125731/

Two important contributors to missing heritability are believed to be rare variants and gene-environment interaction (GXE). Thus, detecting GXE where G is a rare haplotype variant (rHTV) is a pressing problem. Haplotype analysis is usually the natural second step to follow up on a genomic region that is implicated to be associated through single nucleotide variants (SNV) analysis. Further, rHTV can tag associated rare SNV and provide greater power to detect them than popular collapsing methods. Recently we proposed Logistic Bayesian LASSO (LBL) for detecting rHTV association with case-control data. LBL shrinks the unassociated (especially common) haplotypes towards zero so that an associated rHTV can be identified with greater power. Here we incorporate environmental factors and their interactions with haplotypes in LBL. As LBL is based on retrospective likelihood, this extension is not trivial. We model the joint distribution of haplotypes and covariates given the case-control status. We apply the approach (LBLGXE) to the Michigan, Mayo, AREDS, Pennsylvania Cohort Study on Age-related Macular Degeneration (AMD). LBL-GXE detects interaction of a specific rHTV in CFH gene with smoking. To the best of our knowledge, this is the first time in the AMD literature that an interaction of smoking with a specific (rather than pooled) rHTV has
been implicated. We also carry out simulations and find that LBL-GXE has reasonably good powers for
detecting interactions with rHTV while keeping the type I error rates well-controlled. Thus, we conclude
that LBL-GXE is a useful tool for uncovering missing heritability.

Detecting epistatic SNPs associated with complex diseases via a
Bayesian classification tree search method

Min Chen¹, Judy Cho², Hongyu Zhao³
Department of Mathematical Sciences, University of Texas at Dallas¹
Mount Sinai Hospital²
Yale University³
http://www.utdallas.edu/~mchen

Complex phenotypes are known to be associated with interactions among genetic factors. A growing body
of evidence suggests that gene-gene interactions contribute to many common human diseases. Identifying
potential interactions of multiple polymorphisms thus may be important to understand the biology and
biochemical processes of the disease etiology. However, despite the great success of genome-wide association
studies that mostly focus on single locus analysis, it is challenging to detect these interactions, especially
when the marginal effects of the susceptible loci are weak and/or they involve several genetic factors. Here
we describe a Bayesian classification tree model to detect such interactions in case-control association studies.
We show that this method has the potential to uncover interactions involving polymorphisms showing weak
to moderate marginal effects as well as multi factorial interactions involving more than two loci.
Statistical Completion of Partially Identified Graph with Application to Estimation of Gene Regulatory Network

Donghyeon Yu
Quantitative Biomedical Research Cencer, University of Texas Southwestern Medical Center
Donghyeon.Yu@UTSouthwestern.edu

We study the estimation of a Gaussian graphical model whose dependent structures are partially identified. In a Gaussian graphical model, an off-diagonal zero entry in the concentration matrix (the inverse covariance matrix) implies conditional independence of corresponding two variables given all other variables. A number of methods have been proposed to estimate a sparse large-scale Gaussian graphical model or, equivalently, a sparse large-scale concentration matrix. In practice, the graph structure to be estimated is often partially identified by other sources or a pre-screening. In this paper, we propose a simple modification of existing methods to take into account this information in the estimation. We show that the partially identified dependent structure reduces the error in estimating the dependent structure. We apply the modified procedure to estimating the gene regulatory network from lung cancer data, where protein-protein interactions are partially identified from the human protein reference database.

Identifying hemimethylation and differential methylation patterns using statistical and computational methods

Shuying Sun¹, Xiaoqing Yu², Peng Li³
Department of Mathematics, Texas State University
Department of Epidemiology and Biostatistics, Case Western Reserve University²
Department of Electrical Engineering and Computer Sciences, Case Western Reserve University³
s_s355@txstate.edu

DNA methylation is one of the most common molecular changes in cells. It involves the addition of a methyl group (CH₃) to the 5’ cytosine (C). This epigenetic event plays an important role in regulating gene expression. Therefore, it is important to study DNA methylation patterns such as hemimethylation (i.e., methylation only occurs on one DNA strand) and differential methylation between two groups. With next generation sequencing (NGS) technologies, it is now possible to identify different methylation patterns by considering methylation at the single CG site level in an entire genome. However, it is challenging to identify DNA methylation patterns in large and complex NGS data. In order to address this difficult question, we have developed a computational pipeline (HMPL) to identify hemimethylation patterns and a new hidden Markov model-based statistical approach to identify differentially methylated regions. In this presentation, we will introduce our methods and compare them with currently those available, using both simulated data and real sequencing data.
Spatial Prediction of Functional Data: Application to Porosity Curves

Joon Jin Song, Associate Professor
Department of Statistical Science, Baylor University
Joon_Song@baylor.edu

Functional data analysis has emerged as a new area of statistical research with a wide range of applications. Moreover, presence of spatial correlation among these functional data makes it a challenging problem. We propose novel Bayesian models for spatially correlated functional data. The proposed models enable one to regularize curves as well as to predict curves at unobserved locations. The proposed models are employed to analyze a real porosity data set in petroleum engineering.
The Gini Autocovariance Function Applied to Heavy Tailed Linear Time Series

Marcel Carcea, Robert Serfling
Department of Mathematical Sciences, University of Texas at Dallas
mdc090020@utdallas.edu

Many time series settings in economics, finance, and actuarial science involve heavy tailed distributions and data. The fitting of autoregressive models (AR) plays a central role. With heavy tailed innovations or contaminants, the usual second-order assumptions fail to hold. However, a ‘Gini autocovariance function’ is well-defined under just first-order moment assumptions. Estimators for AR(p) models based on a robustified sample Gini autocovariance function are linear, easily interpreted, and have closed form expressions. This talk presents results on their performance via simulation studies allowing a wide range of typical innovation and outlier scenarios. Comparisons are made with the Least Squares and Robust Least Squares approaches. It is seen that the Gini approach competes very well with standard methods and provides a new reliable tool in time series modeling in heavy tailed settings.

Forecasting Financial Volatility: An Exogenous Log-GARCH Model

Ming Chen, Qiongxia Song
Department of Mathematical Sciences, University of Texas at Dallas
mxc104720@utdallas.edu

In this article, we develop a new model for financial volatility estimation and forecasting by including exogenous variables in a semi-parametric log-GARCH model. With additional information, we expect to gain an increased prediction power. We propose a quasi maximum likelihood procedure via spline smoothing technique. Consistent estimators and asymptotic normality are obtained under mild regularity conditions. Simulation experiments provide strong evidence that corroborates the asymptotic theories. Additionally, an application to S&P 500 index data demonstrates advantages of our model comparing with GARCH(1,1) and log-GARCH(1,1) models.

Efficient Nonparametric Hazard Rate Estimation With Left-Truncated And Right-Censored Data

Jufen Chu
Department of Mathematical Sciences, University of Texas at Dallas
jxc109920@utdallas.edu

Nonparametric estimation of a hazard rate from left truncated and right censored data is a typical situation in applications, and a number of consistent and rate-optimal estimators, under the mean integrated squared error (MISE) criteria, have been proposed. It is known that, under a mild assumption, neither truncation nor censoring affects the rate of the MISE convergence. Hence a sharp constant of the MISE convergence is needed to create a benchmark for an estimator. The paper develops the theory of sharp minimax nonparametric estimation of the hazard rate with left truncated and right censored data. It is shown how the truncation and censoring affect the MISE. The proposed data-driven sharp minimax estimator adapts to smoothness
of an underlying hazard rate and it also adapts to unknown distributions of the truncating and censoring random variables. The performance of the proposed estimator is illustrated in applications to simulated and real data.

**Comparison of Haplotype-based Statistical Tests for Disease Association with Rare Variants**

Ananda Datta  
Department of Mathematical Sciences, University of Texas at Dallas  
asd107020@utdallas.edu

Experimental studies suggest that both common and rare variants contribute to complex diseases. Till date, the effects of common variants have been thoroughly assessed through Genome Wide Association Studies (GWAS). Yet the implicated common variants account for only a small proportion of disease variability leading to the quest for “missing heritability” in complex diseases. Thus, rare variants have been increasingly cited as major contributors in the disease etiology, especially for common diseases, referred as Common Disease, Rare Variant (CDRV) hypothesis. Many approaches have been proposed for analyzing the association of rare variants with diseases. Compared to the SNP-based approaches, relatively lesser attention has been given to the development of haplotype-based approaches for investigating rare variants. However, haplotypes can hold key information in understanding the role of candidate genes in disease etiology as they can provide insights into the underlying linkage disequilibrium structure. Thus, haplotype association test is the natural next step once a genomic region is implicated to be associated in single SNP analysis. Also, rare haplotypes can result from common SNPs (from GWAS), and hence haplotype association allows us to investigate the CDRV hypothesis using the amassed GWAS data in addition to the data being generated from the Next Generation Sequencing technologies. Further, it is also suggested that haplotypes can tag uncommon variants that are not well tagged by common variants. Recognizing these benefits of haplotype-based testing, lately several approaches have been proposed for rare variants. However, there is a lack of evaluations and comparisons among those tests leading to lack of guidance for the practitioners. Our goal here is to fill this gap by comparing various statistical tests through simulations. In particular, we compare eight different tests out of which seven are haplotype-based while one is a popular SNP-based test, SKAT. Some of the haplotype-based tests are global association tests for a given region, i.e., they evaluate whether a haplotype region is associated without testing the effects of specific haplotypes. While others test for each individual haplotype in a given region. We compare the global association methods by constructing ROC-type plots of type I error rate versus power. For the methods that test individual haplotype association, we compare their powers for detecting truly associated haplotypes after adjusting their average type I error rates over non-associated haplotypes to be the same (by adjusting the threshold for declaring significance). We conduct simulations under various combinations of rare and common haplotype frequencies. Our results show that haplotype-based approaches are all more powerful than SKAT. Among the global haplotype-based tests, Haplo Score test and Generalized Linear Model with regularization (rGLM) out performs the other global association tests. While among the approaches that test for individual haplotype effect, Logistic Bayesian LASSO (LBL) is more powerful than other approaches in identifying rare associated haplotypes when the average type I error for all the approaches are kept the same.
Sample size estimation for a two-group comparison of repeated count outcomes using GEE

Ying Lou¹, Jing Cao¹, Song Zhang², Chul Ahn²
Department of Statistical Science, Southern Methodist University¹
Department of Clinical Sciences, UT Southwestern Medical Center²
ylou@mail.smu.edu

Generalized estimating equations (GEE) method (Zeger and Liang, 1986) is been used by many researchers in analyzing correlated data due to its robustness to mis-specification of data correlation structure. In this paper, we derive a closed-form sample size formula for comparing time-averaged difference of repeated count data using GEE method. To investigate the sufficiency of the sample size calculated by the proposed formula, we conduct intensive simulations under various parameter settings. In the simulation, we consider two correlation structures for repeated data within each subject. One is auto-regressive with order of one (AR(1)), and the other is compound symmetry (CS). We also consider two missing patterns, monotonic missing and independent missing. We observe that the empirical powers and type I errors of the estimated sample size are close to their nominal levels. The sample size estimated by the proposed formula is smaller than that estimated by traditional adjustment method when missing data is present.

A Measurement Error Model for Method Comparison Data

Lakshika S. Nawarathna, Pankaj K. Choudhary
Department of Mathematical Sciences, University of Texas at Dallas
lsm082000@utdallas.edu

Common approaches to analyze method comparison data generally make two simplifying assumptions: (a) mean difference between the two methods being compared is constant, and (b) the measurement errors are homoscedastic. However, these assumptions are often violated in practice. In particular, the mean difference as well as the error variability of the methods may depend on the true measurement, which remains unobservable. These data need to be modeled using a heteroscedastic measurement error model. We present such a model for the case when the measurements are replicated, and extend the existing agreement evaluation methodology to work under the proposed model. Fitting of this model presents some computational difficulties because the likelihood function is not available in a closed form. To deal with these difficulties, we suggest three estimation methods to obtain approximate maximum likelihood estimates. Simulations show that these methods work well for moderately large samples. As an application, the proposed methodology is used to analyze a real dataset involving measurement of serum cholesterol.

Transient Error in Personality Assessments

Amy Nussbaum
Department of Statistical Science, Southern Methodist University
anussbaum@mail.smu.edu

As with all latent variable measurement models, a certain amount of measurement error in personality assessments is to be expected. However, some of the error may be attributable to mood and other events affecting a persons state rather than their true trait level. We present a model that partitions total variance into item variance, transient error, and measurement error. We also outline various techniques for estimating the parameters of the model, including two separate approaches using method of moments as well as maximum likelihood implementation. Finally, we summarize simulation results and compare the outcomes of all methodologies.


**IRT and High Stakes Educational Testing**

Bivin Philip Sadler, Lynne Stokes  
Department of Statistical Science, Southern Methodist University  
bsadler@mail.smu.edu

High stakes educational testing has become a focal point of interest at the individual, State and National levels. Colleges and Universities as well as State and National educational budgets rely on these tests as a stable and accurate measure of the ability of their respective target populations. For this reason, tremendous efforts have been taken to develop the theory and tools necessary to help construct, assess and compare tests; the current state of the art is termed Item Response Theory (IRT). One of the principle tools provided by IRT is the information function which is the source of the standard error of the ability estimates and is often used to construct and compare tests of different designs. However, the information function is an asymptotic result and is merely an upper bound for the information a test may provide. The focus of this research is to introduce and assess two methods to better assess the actual amount of information provided (particularly for small tests). The first method is an exact method and is shown to outperform the information function although it is only available for tests with 15 or less items. The second method does not perform as well as the exact method although it fortunately increases its performance as the number of items increases. Both methods provide bias estimation which is briefly shown to aid in bias correcting raw ability estimates. The research concludes with an extension of the methods to an adaptive strategy of testing known as multi-stage testing (MST).

**Large Covariance Matrix Estimation for Functional Data**

Ekaterina Smirnova  
Department of Mathematical Sciences, University of Texas at Dallas  
exs093220@utdallas.edu

The theory, methods and application of statistical analysis of large-p-small-n cross-correlation matrices arising in fMRI studies of neuroplasticity, which is the ability of the brain to recognize neural pathways based on new experience and change in learning. Traditionally these studies are based on averaging images over large areas in right and left hemispheres and then finding a single cross-correlation function. It is proposed to conduct such an analysis based on a voxel-to-voxel level which immediately yields large cross-correlation matrices. Furthermore, the matrices have an interesting property to have both sparse and dense rows and columns. Main steps in solving the problem are: (i) treat observations, available for a single voxel, as a nonparametric regression; (ii) use a wavelet transform and then work with empirical wavelet coefficients; (iii) develop the theory and methods of adaptive simultaneous confidence intervals and adaptive rate-minimax thresholding estimation for the matrices. The developed methods are illustrated via analysis of fMRI experiments and the results allow us not only conclude that during fMRI experiments there is a change in cross-correlation between left and right hemispheres (the fact well known in the literature), but that we can also enrich our understanding how neural pathways are activated on a single voxel-to-voxel level.

**Masking and Swamping Robustness of Mahalanobis Distance Outlier Identifiers for Multivariate Data**

Shanshan Wang, Robert Serfling  
Department of Mathematical Sciences, University of Texas at Dallas  
sxw096320@utdallas.edu

Two key measures for studying robustness of outlier detection procedures are breakdown points associated with masking and swamping, respectively. Here these robustness measures are formulated in the setting of multivariate data and evaluated for the special case of Mahalanobis distance outlyingness, which has many
variants depending on choice of location measure and scatter estimator. Several examples are compared, including the classical location and scatter estimators and the robust, but computationally intensive, Minimum Covariance Determinant (MCD) estimators. We find that with the classical estimators the swamping breakdown point is excellent but the masking breakdown point is unacceptable, whereas with the MCD estimators a suitable trade-off between the masking and swamping breakdown points is achieved, although with a heavier computational burden.

A Model-based Approach to Identify Binding Sites in CLIP-Seq Data

Tao Wang
Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center
tao.wang@utsouthwestern.edu

Cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) has made it possible to identify the targeting sites of RNA-binding proteins in various cell culture systems and tissue types on a genome-wide scale. Here we present a novel model-based approach (MiClip) to identify high-confidence protein-RNA binding sites from CLIP-seq datasets. This approach assigns a probability score for each potential binding site to help prioritize subsequent validation experiments. The MiClip algorithm has been tested in both HITS-CLIP and PAR-CLIP datasets. In the HITS-CLIP dataset, the signal/noise ratios of miRNA seed motif enrichment produced by the MiClip approach are between 17% and 301% higher than those by the ad hoc method for the top 10 most enriched miRNAs. In the PAR-CLIP dataset, the MiClip approach can identify 50% more validated binding targets than the original ad hoc method and two recently published methods. To facilitate the application of the algorithm, we have released an R package, MiClip (http://cran.r-project.org/web/packages/MiClip/index.html), and a public web-based graphical user interface software (http://galaxy.qbrc.org/tool_runner?tool_id=mi_clip) for customized analysis.

On Fast Affine Equivariant Robust Scatter Estimation

Yunfei Wang, Robert Serfling
Department of Mathematical Sciences, University of Texas at Dallas
yxw109320@utdallas.edu

The Minimum Covariance Determinant (MCD) approach is a leading method for constructing multivariate location and scatter estimators that are affine equivariant and highly robust. Although direct computation of the MCD estimator is usually prohibitive, it can be computed approximately and more efficiently by the Fast-MCD algorithm. However, as shown in various experiments, even Fast-MCD becomes computationally prohibitive when the dimension of the data is sufficiently high. Here we introduce and study a fast version, Fast-TSO, of the Trimmed Spatial Outlyingness (TSO) scatter estimator proposed by Mazumder and Serfling (2013). Like Fast-MCD, it is affine equivariant and robust, but it is computationally much faster. Also, it maintains the same efficiency as Fast-MCD relative to estimation of a population scatter matrix, with respect to the Frobenius norm and other measures. This estimator uses the transformation-retransformation spatial outlyingness function to quickly select a middle 50.
Nonparametric outlier detection with functional data using the spatial depth approach

Uditha Wijesuriya, Robert Serfling
Department of Mathematical Sciences, University of Texas at Dallas
uaw090020@utdallas.edu

The spatial depth and outlyingness approach with multivariate data has been very successful for its tractability, computational ease, and convenient asymptotics. Here its extension to the setting of outlier identification in functional data analysis is treated. Computations may be carried out in the Hilbert space of curves or in a corresponding Euclidean space obtained by discretization. For a data set of real-valued curves, methods are described for useful display of the sample median curve, the 50% central region of curves, and sample "outlier" curves, including both location and shape outliers. A spatial functional boxplot approach is used to identify outliers. Here we illustrate with several actual and simulated data sets, comparing the spatial approach with several leading competing methods, with respect to the false positive rate, the false negative rate, and the computational burden as criteria. It is seen that the spatial approach is among the very best in performance.

Detecting Rare Haplotype-Environment Interaction in Presence of Gene-Environment Dependence

Yuan Zhang\textsuperscript{1}, Swati Biswas\textsuperscript{1}, Shili Lin\textsuperscript{2}
University of Texas at Dallas
Department of Mathematical Sciences, University of Texas at Dallas\textsuperscript{1}
The Ohio State University\textsuperscript{2}
yxz112020@utdallas.edu

Rare variants and gene-environment interactions (GXE) have been suggested in the literature as potential causes for "missing heritability" in complex diseases. Thus, detecting GXE where G is a rare haplotype variant (rHTV) is an important research problem. Recently, Biswas et al. (2013; Genetic Epidemiology) proposed a method based on Logistic Bayesian Lasso (LBL) for detecting GXE (LBL-GXE) using case-control sample. However, a key assumption of the method is independence of gene (G) and environmental factor (E) for controls. Here we propose a novel approach to deal with the scenarios when this assumption may not hold. In particular, in the framework of LBL, we model the haplotype frequencies as functions of environment using a multinomial logistic regression model. Simulation studies show that our new approach has reasonably good power for detecting interactions with rHTV while keeping the type I error rates well-controlled whereas LBL-GXE leads to inflated type I error rates in presence of G-E dependence. However, when G-E independence holds, LBL-GXE has higher power. So, to optimize the power without sacrificing type I error in both G-E dependence and independence scenarios, we also explore a two-stage approach. In the first stage, we conduct a chi-square or Fishers exact test to test for G-E independence. In the second stage, LBL-GXE or our proposed new approach is used depending on the result of the test in the first stage. Simulations indicate that the performance of this two-stage approach is close to LBL-GXE in case of G-E independence while it retains the advantage of our novel method in the case of G-E dependence.
Testing Multiple Hypotheses in Truncated Group Sequential Experiments
Tian Zhao, Michael Baron
University of Texas at Dallas
Department of Mathematical Sciences, University of Texas at Dallas
txz102020@utdallas.edu

Statistical procedures are developed for the simultaneous testing of multiple hypotheses in group sequential experiments such as group sequential clinical trials. As a result, a decision has to be made on each tested parameter instead of combining all tests into one hypothesis. Further, the maximum number of sampled groups is restricted. Under these conditions, the proposed multiple testing procedure controls the familywise error rate and the familywise power in the strong sense. Following the Holm approach for multiple comparisons, our truncated sequential schemes require uniformly smaller sample sizes than the commonly used Bonferroni procedure.

Computational Detection and Suppression of Sequence-specific Off-target Phenotypes from Whole Genome RNAi Screens
Rui Zhong
Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center
rui.zhong@utsouthwestern.edu

A challenge for large-scale siRNA loss-of-function studies is the biological pleiotropy resulting from multiple modes of action of siRNA reagents. A major confounding feature of these reagents is the microRNA-like translational quelling resulting from short regions of oligonucleotide complementarity to many different mRNAs. We developed a computational approach, Deconvolution Analysis of RNAi Screening data (DecoRNAi), for automated quantitation of off-target effects in RNAi screening data sets. Substantial reduction of off-target rates was experimentally validated in 5 distinct biological screens across different genome-wide siRNA libraries. A public-access graphical-user-interface has been constructed to facilitate application of this algorithm.

Comparison of Quasi-Alignment Methods for Metagenomic Classification
Xiujun (Sylvia) Zhu, Monnie McGee
Department of Statistical Science, Southern Methodist University
xiujunz@mail.smu.edu

Both number and size of genomic databases have dramatically increased with recent genomic sequencing advances. These sequences can be used to determine evolutionary relationships among the corresponding organisms. First the sequences are aligned and then scores are calculated to determine the distance between the sequences. However, these straightforward methods are very computationally intensive. Previously, we applied a novel quasi-alignment method using the Extensible Markov Model (Dunham, et. al., 2004) as the representations of taxa. This supervised learning technique can rapidly and accurately classify bacterial 16S rRNA sequences into lower-order taxonomic rank, such as species, genus and family level. The standard procedure of Quasi-alignment method includes sequence preprocessing, EMM learning and sequence score. A competitor alignment-free methods to Quasi-alignment is using Abstracted Augmented Markov Models (AAMM). This study aims to evaluate gene classification using EMMs and AAMMs and compare the performance of two models in sequence classification, address statistical inference and construct statistical tests on the sequence analysis using EMMs and AAMMs.
Control Charts Using the Smirnov Statistic

Shon Taylor Asbury
Department of Mathematics and Statistics, Sam Houston State University
sta001@SHSU.EDU

Within the subject of statistical quality control, the Shewhart control chart for the sample mean is a useful tool for detecting departures from the desired mean of a process variable. There has been an increased interest in establishing a similar technique when the distribution of the process variable of interest is unknown. This project examines utilizing the Kolmogorov-Smirnov test for two independent samples in combination with an interesting approach for simulating the use of this nonparametric test within a typical process control setting. The goal of this project is to provide a comparative study of this nonparametric technique versus the established Shewhart technique.

Semantic Similarity of Documents Using LSA

Chelsea Boling, Kumer Das
Lamar University
cmboling@lamar.edu

Nowadays, big data often refers to networks connecting a myriad of information that requires techniques in order to be extracted and analyzed. Ongoing developing methods of extraction are crucial in leveraging constructive information of high dimensional data. Furthermore, Latent Semantic Analysis is a technique that analyzes relationships between documents and its terms, and it discovers a data representation that has a lower dimension than the original semantic space. Essentially, the reduced dimensionality preserves the most crucial aspects of the data since LSA analyzes documents to find latent meaning in those particular documents. The latent semantic space is determined by singular value decomposition (SVD), which enables a powerful process to simplify any given matrix into a product of factorized, unique components. This technique should cover the basis of reducing dimensionality, as using this process on actual queries should provide data that shows the similarities in documents and terms. LSA is further studied to better understand how such semantic models could be used to improve the search of user queries in a particular domain of interest, such as newspaper articles, literature, and social media. Our goal is to experimentally produce a modified framework, which improves the search for similarity in documents, and to verify its performance from several queries.

Quantifying Extreme Hurricane Risk in the United States Gulf Coast

Asim Kumer Dey
Department of Mathematics, Lamar University
adey@lamar.edu

Hurricanes threaten the United States every year. It is important to quantify the risk of these events for emergency managers. Extreme value statistics are used to model hurricane characteristics at different locations. Here, we will analyze the tail behavior of hurricane losses in the United States Gulf of Mexico shoreline and the Atlantic shore of Florida. Economic damage is normalized to adjust for temporal shifts in societal vulnerability. Using over a specified threshold, the risk of extreme hurricanes is estimated. The risk estimates are provided as statistical return periods, or the expected frequency of specific hurricane magnitudes.
A Linear Algebraic/Multivariate Statistical Journey Through Music Using MIDI Data

Amber Gates
Department of Mathematics and Statistics, Sam Houston State University
ALG021@SHSU.EDU

While music has been around for thousands of years, the emergence of Music Instrumentation Digital Interface (MIDI) in 1982 has given us the capability to conduct mathematical and statistical research in music. This presentation references and addresses various linear algebraic and multivariate statistical aspects of music that are retrievable with the use of MIDI data. Future research ideas based off of these studies are presented as well.

Comparing Singular Value Decomposition, Principal Component Analysis, and Non-negative Matrix Factorization

Janitha C. Hapuwitharana, Kumer P. Das, Jennifer Daniel, Stan Young
Lamar University
jchapuwitharana@gmail.com

In statistical theory and application, two-way tables of numeric data are often analyzed using dimension reduction methods like principal component analysis (PCA), singular value decomposition (SVD), and non-negative matrix factorization (NMF). The effect of PCA, SVD and NMF on large data sets can be represented in different ways. This study is designed to compare PCA, SVD, and NMF using compressed mortality data in USA from 1968 to 2010 with breakdown of each year of death by gender and race. The clustering feature of the NMF has mainly been focused in this study.

Application of Feature Detection and Statistics in Target Analysis

Casey Hartnett
Sam Houston State University
CXH016@SHSU.EDU

The focus in on the production of a computer program to convert a paper shooting target into a series of coordinates representing the centers of each bullet hole. This is done through the use of a digital scanner to create an image that is then processed to find the holes in the image. Furthermore, robust methods for redundancy are developed along with demanding tests of program accuracy. Lastly, advanced statistical tools are designed such as the Shewhart control chart to analyze the target data.

Control Charts Using the F-Statistic

Ruwan C. Karunanayaka
Department of Mathematics & Statistics, Sam Houston State University
rxk011@SHSU.EDU

The Shewhart $\bar{X}$–Chart is the most popular control chart for monitoring the mean of the distribution of a quality characteristic produced by a process. In this project we develop a technique that can be used as an alternative to the $\bar{X}$–Chart. In this technique, we developed an F-statistic that is similar to the F-Statistic of the one-way ANOVA. Its performance is measured by statistical simulation and it turns out that the
developed F-Statistic is not only quite robust against deviations from normality, but also perform very well under the Normality Assumptions.

Multivariate Statistical Benefits of Traversing the ESRI Bridge Between SAS and ArcGIS

Charith Bhagya Karunarathna
Department of Mathematics and Statistics, Sam Houston State University
cbk013@SHSU.EDU

The SAS Bridge for ESRI adds the powerful analytic and business intelligence capabilities of SAS to the Geographic Information System (GIS). ArcGIS is used worldwide for spatial data analysis and, increasingly, for spatial statistics. The bridge allows users to exchange spatial attribute data between SAS and ArcGIS and to import SAS programming directly into ArcGIS. The bridge provides ArcGIS users access to a wide range of statistical functions afforded by SAS and enhances the joint use of SAS and ArcGIS to do enhanced spatial statistical analyses. This presentation provides initial evidence regarding multivariate statistical benefits of traversing the ESRI Bridge between SAS and ArcGIS, particularly in regard to multivariate statistical clustering in support of detecting anomalies (e.g., detection of missing persons, assessing the magnitude of aftermaths from episodic events such as earthquakes, tornadoes, etc.) using RGB, HSV and YCbCr color spaces retrievable from digital imagery.

Do the Bowlers Have Higher Economy Rates During the PowerPlay in Twenty20 Cricket? - A Negative Binomial Regression Approach

Yang (Mike) Liu
Department of Mathematics and Statistics, Sam Houston State University
yxl019@SHSU.EDU

Models for count data have been widely applied in variety of sports. The Poisson model and negative binomial model are two basic count response models. Due to over-dispersion of data, the negative binomial regression model is employed in this study to predict the number of runs for the batting team against different types of bowlers at different stages of Twenty20 cricket matches. In particular, economy rate during the PowerPlay is a key interest in this study.

Assessing the Sensitivity of Interval Estimates in Errors-in-Variables Regression

Christina Nieuwoudt
Department of Mathematics and Statistics, Sam Houston State University
CMN006@SHSU.EDU

To aid in the process of modeling the linear relationship between correlated variables, we have at our disposal linear regression techniques. Unfortunately, the assumptions needed to apply these concepts do not always align with the realities of data collection. Specifically, the assumption that we are able to observe the variables under study without incurring measurement error is not always valid. When confronted by this problem, errors-in-variables regression techniques are typically employed to estimate the regression parameters. In what follows, we will address the relative merits and detriments of several errors-in-variables methods by way of a sensitivity analysis that addresses the coverage of the 95% confidence interval for the slope.
Extreme Value Theory

Chris Sams
Lamar University
sdts_man@yahoo.com

Extreme Value Theory (EVT) is a branch of Statistics that seeks to access, from a given ordered sample, the probability of events that are more extreme than usually observed. Under very general conditions, EVT’s main results characterize the distribution of the samples maximum or the distribution of values above a given threshold. Thus, two particular approaches exist, block maxima; which divides a time period into equal sections and the maximum of each section is selected. This usually leads to generalized extreme value distribution (GEV). The second approach is point over threshold (POT) which selects every value that exceeds a certain threshold. POT data generally trends toward generalized pareto distribution (GPD). This study focuses on modeling the tail of a heavy tail distribution by using (POT) approximation method. The GPD parameters were estimated by using maximum likelihood (MLE), method of moments (MOM), and principal of maximum entropy (POME).

A study of the Powerplay in One-Day Cricket

Rajitha Silva\textsuperscript{1}, Ananda B.W. Manage\textsuperscript{2}, Tim B. Swartz\textsuperscript{1}
Simon Fraser University\textsuperscript{1}
Sam Houston State University\textsuperscript{2}
rsilva@sfu.ca

This paper investigates the powerplay in one-day cricket. The form of the analysis takes a ”what if” approach where powerplay outcomes are substituted with what might have happened had there been no powerplay. This leads to a paired comparisons setting consisting of actual matches and hypothetical parallel matches where outcomes are imputed during the powerplay period. Some of our findings include (a) the various forms of the powerplay which have been adopted over the years have different effects, (b) recent versions of the powerplay provide an advantage to the batting side, (c) more wickets also occur during the powerplay than had there been no powerplay and (d) there is some effect in run production due to the over where the powerplay is initiated.
Author Index

Asbury, Shon, 17
Baron, Michael, 15
Biswas, Swati, 5
Boling, Chelsea, 17
Cao, Jing, 11
Carcea, Marcel, 9
Chattopadhyay, Bhargab, 4
Chen, Min, 6
Chen, Ming, 9
Cho, Judy, 6
Chu, Jufen, 9
Daniel, Jennifer, 18
Das, Kumer, 17, 18
Datta, Ananda, 10
Dey, Asim, 17
Ellingson, Leif, 4
Gates, Amber, 18
Gel, Yulia, 1
Hapuwitharana, Janitha, 18
Hartnett, Casey, 18
Karunanayaka Thannehelage, Ruwan, 18
Karumarathna, Charith, 19
Liu, Yang, 19
Lou, Ying, 11
McGee, Monnie, 15
Nawarathn, Lakshika S., 11
Nieuwoudt, Christina, 19
Nussbaum, Amy, 11
Ravikumar, Pradeep, 3
Sadler, Bivin Philip , 12
Sams, Chris, 20
Schweinberger, Michael, 3
Serfling, Robert, 9, 12–14
Silva, Rajitha, 20
Smirnova, Ekaterina, 12
Song, Joon Jin, 8
Song, Qiongxia, 5, 9
Staniswalis, Joan, 1
Stokes, Lynne, 12
Sun, Shuying, 7
Thall, Peter, 2
Wang, Xinlei, 5

Wang, Shanshan, 12
Wang, Tao, 13
Wang, Yunfei , 13
Wijesuriya, Uditha, 14
Young, Stan, 18
Yu, Donghyeon, 7
Zhang, Yuan, 14
Zhao, Hongyu, 6
Zhao, Tian, 15
Zhong, Rui, 15
Zhu, Xiujun, 15