The Conference of Texas Statisticians 2017

March 24-25, 2017
Southern Methodist University
Dallas, Texas

Southern Methodist University and the North Texas Chapter of the American Statistical Association (NTSCHASA) welcome you to Dallas for the 37th edition of the Conference of Texas Statisticians.

ORGANIZING COMMITTEE

H. K. Tony Ng (Southern Methodist University, Dallas, Texas, USA)

Stephen Robertson (Southern Methodist University, Dallas, Texas, USA)

We acknowledge the Department of Statistical Science and Dedman College of Southern Methodist University for their financial support. We wish to thank Sheila Crain and Ginny Diaz (Southern Methodist University) for their organizational support for this conference.

We wish to thank *Digital Matrix Systems* of Dallas and Michael Sogomonian, Director, Decision Science Department for their corporate support of the Poster Awards at this conference. Their support is very much appreciated.

PROGRAM SUMMARY

Time	***Activity***
	Friday, March 24, 2017
12:00-13:00	**Registration**
13:00-13:20	Opening Remarks *Lynne Stokes*, Chair and Professor, Department of Statistical Science, SMU
13:20-14:10	*Michael J. Daniels* (The University of Texas at Austin) A Flexible Bayesian Framework for Missing Data and Casual Inference Problems
14:10-15:00	*Robert Cezeaux* (Capital One Financial Services) Statisticians in Banking: Its Not Just Credit Scoring Anymore
15:00-15:30	**Coffee Break**
15:30-16:20	*Suojin Wang* (Texas A&M University) Efficient Estimation in Partially Linear Single-index Models for Longitudinal Data
16:20-17:10	*Jerome P. Keating* (The University of Texas at San Antonio) An Application of the Cox Regression Model to a NASA's MMS Mission
17:10-17:45	COTS Business Meeting
17:00-17:30	Poster Presentations Set-up
17:30-19:00	Poster Session & Social Hour
19:00-21:00	**Banquet Dinner; Welcome Remarks by Professor Thomas DiPiero (Dean, Dedman College of Humanities and Sciences, SMU); Don Owen Award; Honoring Professors Richard F. Gunst and Wayne A. Woodward; Poster Awards**

Time	***Activity***
	Saturday, March 25, 2017
7:00-8:45	**Breakfast**
8:45-9:15	*Han Hao* (University of North Texas) Modeling the Genetic Architecture of Biological Interactions
9:15-9:45	*Fangyuan Zhang* (Texas Tech University) Testing for Associations of Opposite Directionality in a Heterogeneous Population
9:45-10:15	*Suvra Pal* (The University of Texas at Arlington) EM-Based Likelihood Inference for Destructive COM-Poisson Regression Cure Rate Model with Weibull Lifetime
10:15-10:45	**Coffee Break**
10:45-11:15	*Irina Gaynanova* (Texas A&M University) Integrative Association Analysis of Multiple Heterogeneous Data Sources
11:15-11:45	*Dan Cheng* (Texas Tech University) Multiple Testing of Local Maxima for Detection of Peaks in Random Fields
11:45-12:15	*Kalanka P. Jayalath* (University of Houston at Clear Lake) A Graphical Test for Testing Random Effects
12:15-12:30	Closing Remarks

ABSTRACTS

[March 24, 2017, 13:20 – 14:10]
A Flexible Bayesian Framework for Missing Data and Casual Inference Problems

Michael J. Daniels
Department of Statistics and Data Sciences
The University of Texas at Austin
Austin, Texas
mjdaniels@austin.utexas.edu

We describe a general framework for causal inference and missing data problems using Bayesian nonparametric models for the distribution of the observed data and classes of uncheckable assumptions to allow identification and estimation of parameters of interest. We demonstrate this approach in the setting of nonignorable missing data, causal effects in observational studies, and causal effects in mediation (as time permits).
Note: Joint work with Antonio Linero (FSU), Jason Roy (UPENN), Chanmin Kim (Harvard), and Joe Hogan (Brown)

[March 24, 2017, 14:10 – 15:00]
Statisticians in Banking: Its Not Just Credit Scoring Anymore

Robert Cezeaux
Capital One Financial Services
Plano, Texas
Robert.Cezeaux@capitalone.com

In this talk I will provide an overview of how banks are leveraging statisticians beyond traditional credit scoring applications. Banks have a large amount of data on how customers use and interact with their products from credit card transaction data to call center logs. The advent of big data platforms that can handle these types of large unstructured data sources has enabled new applications of analytics that can benefit our customers. This presentation will review several real-world examples of analytic solutions that our statisticians have developed and how they are helping us serve our customers in new ways.

[March 24, 2017 15:30 – 16:20]
Efficient Estimation in Partially Linear Single-index Models for Longitudinal Data

Suojin Wang
Department of Statistics
Texas A&M University
College Station, Texas
sjwang@stat.tamu.edu

In this talk, we consider the estimation of both the parameters and the nonparametric link function in partially linear single-index models for longitudinal data which may be unbalanced. In particular, a new three-stage approach is proposed to estimate the nonparametric link function using marginal kernel regression and the parametric components with generalized estimating equations. The resulting estimators properly account for the within-subject correlation. We show that the parameter estimators are asymptotically semiparametric efficient. We also show that the asymptotic variance of the link function estimator is minimized when the working error covariance matrices are correctly specified. The new estimators are more efficient than estimators in the existing literature. These asymptotic results are obtained without assuming normality. The finite sample performance of the proposed method is demonstrated by simulation studies. In addition, two real data examples are analyzed to illustrate the methodology.

[March 24, 2017 16:20 – 17:10]
An Application of the Cox Regression Model to a NASA's MMS Mission

Jerome P. Keating
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, Texas
jerome.keating@utsa.edu

We present a simple application of the Cox regression model to analyze failures of a voltage stepper used to power electron spectrometers and ion spectrometers on satellites used in NASA's Magnetospheric Multiscale (MMS) mission. The project examines causes of failures revealed through statistical methods. The input provided was used in engineering decision to modify the operational procedures of the electron and ion spectrometers.
Note: Joint work with Robert Mason

[March 25, 2017 8:45 – 9:15]
Modeling the Genetic Architecture of Biological Interactions

Han Hao
Department of Mathematics
University of North Texas
Denton, Texas
han.hao@unt.edu

Biological interactions have been extensively studied in ecology and evolutionary biology. However, the understanding of how genes regulate biological interactions is still very limited. In this work, we put forth a new perspective to map the genetic control of biological interactions by integrating evolutionary game theory into the genome-wide association study (GWAS) framework. Biological interactions are quantified by a generalized Lotka-Volterra differential equation system, which contains flexible parameters that can discern different types of biological interactions. By estimating these parameters over different genotypes, we detect genetic variants that significantly affect biological interactions. A generalized profiling approach is used for parameter estimation to reduce computation cost. The usefulness of the approach is demonstrated by both simulations and a case study of poplar growth.

[March 25, 2017 9:15 – 9:45]
Testing for Associations of Opposite Directionality in a Heterogeneous Population

Fangyuan Zhang
Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas
fangyuan.zhang@ttu.edu

In gene networks, it is possible that the patterns of gene co-expression may exist only in a subset of the sample. In studies of relationships between genotypes and expressions of genes over multiple tissues, there may be associations in some tissues but not in the others. Despite the importance of the problem in genomic applications, it is challenging to identify relationships between two variables when the correlation may only exist in a subset of the sample. The situation becomes even less tractable when there exist two subsets in which correlations are in opposite directions. By ranking subset relationships according to Kendalls tau, a tau-path can be derived to facilitate the identification of correlated subsets, if such subsets exist. However, the current tau-path methodology only considers the situation in which there is association in a subsample; the more complex scenario depicting the existence of two subsets with opposite directionality of associations was not addressed. Further, existing algorithms for finding tau-paths may be suboptimal given their greedy nature. In this paper, we extend the tau-path methodology to accommodate the situation in which the sample may be drawn from a heterogeneous population composed of subpopulations portraying positive and negative associations. We also propose the use of a cross entropy Monte Carlo procedure to obtain an optimal tau-path, CEMCtp. The algorithm not only can provide simultaneous detection of positive and negative correlations in the same sample, but also can lead to the identification of subsamples that provide evidence for the detected associations. An extensive simulation study shows the aptness of CEMCtp for detecting associations under various scenarios. Compared with two standard tests for detecting associations, CEMCtp is seen to be more powerful when there are indeed complex subset associations with well-controlled type-I error rates. We applied CEMCtp to the NCI-60 gene expression data to illustrate its utility for uncovering network relationships that were missed with standard methods.

[March 25, 2017 9:45 – 10:15]
EM-Based Likelihood Inference for Destructive COM-Poisson Regression Cure Rate Model with Weibull Lifetime

Suvra Pal
Department of Mathematics
The University of Texas at Arlington
Arlington, Texas
suvra.pal@uta.edu

In this talk, I will consider the destructive COM-Poisson cure rate model. This model assumes the initial risk factors in a competitive scenario to undergo a destructive process so that what is recorded is only from the undamaged portion of the original number of risks factors. By assuming a COM-Poisson distribution for the initial risk factors, the steps of the EM algorithm will be discussed in detail to calculate the MLEs of the model parameters. An extensive simulation study will be carried out to demonstrate the performance of the proposed estimation method. The flexibility of the COM-Poisson family will be utilized to carry out a model discrimination using the likelihood ratio test. Finally, a melanoma data will be analyzed for illustrative purpose.

[March 25, 2017 10:45 – 11:15]
Integrative Association Analysis of Multiple Heterogeneous Data Sources

Irina Gaynanova
Department of Statistics
Texas A&M University
College Station, Texas
irinag@stat.tamu.edu

The Growth of Data Collection and Data Sharing Led to Increased Availability of Multiple Types of Data Collected on the Same Set of Objects. As an example, RNASeq, miRNA expression and methylation data for the same tumor samples are publicly available through the Cancer Genome Atlas (TCGA). Due to the scale of the data, as well as its heterogeneity, it is typical to analyze each data type separately. In this work we use penalized

risk minimization framework as a building block for integrative association analysis of multiple heterogeneous data sources. By learning the sparse representation of underlying matrix decompositions, we are able to identify the patterns that are common across the data sources as well as source-specific patterns.

[March 25, 2017 11:15 – 11:45]
Multiple Testing of Local Maxima for Detection of Peaks in Random Fields

Dan Cheng
Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas
dan.cheng@ttu.edu

Detection of sparse localized signals embedded in noise background is an important problem in statistics, with applications in many scientific areas such as neuroimaging, microscopy and astronomy. In this talk, I will present a topological multiple testing scheme for detecting signals (peaks) in images under stationary ergodic Gaussian noise, where tests are performed at local maxima of the smoothed observed signals. Two methods are developed according to two different ways of computing p-values: (i) using the exact height distribution of local maxima, available explicitly when the noise field is isotropic; (ii) using an approximation to the overshoot distribution of local maxima above a pre-threshold, applicable when the exact distribution is unknown, such as when the stationary noise field is non-isotropic. The algorithms, combined with the Benjamini-Hochberg procedure for thresholding p-values, provide asymptotic strong control of the False Discovery Rate (FDR) and power consistency, with specific rates, as the search space and signal strength get large. Simulations show that error levels are maintained for nonasymptotic conditions and that power is maximized when the bandwidth of smoothing kernel is close to the theoretical optimal result. The methods are illustrated in a data example of functional magnetic resonance images of the brain. I will also discuss certain further developments of the methods.

[March 25, 2017 11:45 – 12:15]
A Graphical Test for Testing Random Effects

Kalanka P. Jayalath
Department of Mathematical Sciences
University of Houston at Clear Lake
Houston, Texas
Jayalath@UHCL.edu

Analysis of means (ANOM) is an alternative graphical testing procedure that can be used to test the mean effects in fixed effect models. The procedure becomes increasingly popular among practitioners due to its attractive graphical interpretation. One major drawback of this technique is its inability to deal with random factor effects. We evoke this long lasting issue with a reasonable remedy to broaden its applications in a wide range of situations. Specifically, we utilize the range of the treatment averages to identify the dispersion of the underlying population which can be applied to test random effects. We also discuss the possible extension of this approach in a wide range of common statistical designs with both random and mixed effects. Comparative results of ANOM vs ANOVA also discuss via a Monte Carlo simulation study.

[Poster Presentation 1] Mixtures of Polisicchio's Distribution for Inequality Measures

Scott Smith
Department of Mathematics and Statistics
University of Incarnate Word
San Antonio, Texas
sesmith@ulwtx.edu

As new data sources arrive, novel methods of defining undiscovered distributions are increasingly useful. In this presentation, the Hazard-Product method of generalizing survival functions is introduced. Some limitations and considerations are discussed, and a new Gompertz model is proposed. Finally, basic moment and estimation properties of the model are given, and the new model is compared to a previously-defined Gompertz generalization when fit to a dataset of device lifetimes.

[Poster Presentation 2] A Bayesian Latent Variable Approach to Aggregation of Partial and Top Ranked Lists

Xue Li
Department of Statistical Science
Southern Methodist University
Dallas, Texas
xuel@smu.edu

Rank aggregation, a meta-analysis method, combines different individual rank lists into one single rank list which is ideally more reliable. It has a rich history in the field of information retrieval, with applications to text mining, webpage ranking, meta-search engine building, etc. However, methods developed in such contexts are often ill suited for genomic applications, in which gene lists generated from individual studies are inherently noisy, due to various sources of heterogeneity. Further, because of missing or zero-count data, a portion of genes are not analyzed in all component studies, leading to partially ranked lists; and for some lists, only top-ranked genes are reported. In this study, we develop Bayesian latent variable approaches to rank aggregation that formally deals with top and partial preference lists. The performance of the proposed method is shown to be an improvement over many popular methods based on a simulation study.

Note: Joint work with Xinlei Wang (Southern Methodist University).

[Poster Presentation 3] Bayesian Hidden Potts Models for Pathological Image Analysis

Qiwei Li
Department of Clinical Sciences
UT Southwestern Medical Center
Dallas, Texas
qiwei.li@UTSouthwestern.edu

Technological advances in tumor pathological imaging analyses have created unprecedented opportunities to study tumor morphology using computational methods. Tumor tissue slide scanning is becoming a routine clinical procedure and can produce massive digital pathological images that capture histological details in high resolution. The spatial patterns and interactions among different types of cells may reveal important information about patient prognosis and response to therapy. Reliable computational methods to analyze tumor pathological slides will have an immediate impact on patient care in cancer. We consider the problem of modeling a pathological image with irregular cell locations and propose a novel Bayesian hierarchical model. The model incorporates a hidden Potts model to project the image to a square lattice and a Markov random field (MRF) prior model to identify the two subpopulations in the image. The model allows us to quantify the interactions between different types of cells, while selecting clinically meaningful patterns from the background area. We use Markov chain Monte Carlo (MCMC) sampling techniques that combine the double Metropolis-Hastings (DMH) algorithm, which is able to simulate samples approximately from a distribution with an intractable normalizing constant. For the lung cancer pathological imaging data of 205 patients, we find that the interaction strength between tumor and stromal cells are significantly related to the patients survival time. This is encouraging, as the result indicates a new biomarker for precise prognosis and treatments for lung cancer patients. In addition, this statistical methodology provides a new perspective to understand the role of the cell-cell interactions in cancer progression.

Note: Joint work with Faliu Yi, Faming Liang, Xinlei Wang, Guanghua Xiao.

[Poster Presentation 4] Nonparametric Methods for General ANCOVA Designs

Cong Cao
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
cxc106920@utdallas.edu

In many biological, ecological, psychological, or medical studies, data is collected in terms of a factorial design. The aim of such studies is making inferences among the treatment effects involved in the trial. Hereby, other variables that are called covariates can obscure the factor effects, i.e. the weight gain may be associated with the original weight (baseline) of the animal in the comparison of different feeds. Analysis of Covariance (ANCOVA) can adjust the treatment effects for the impact of the covariates on the response variable. The classical ANCOVA model is regularly based on the assumptions of multivariate normality and equal variances. In many experiments, however, these assumptions are hard justify, e.g., when reaction times or count data are observed. Inference with violation of assumptions may lead to conservative or liberal test decisions. We develop statistical inferential method for general ANCOVA designs, which neither assume homogeneous variances across the treatment groups nor a normal distribution. Simulation studies show that the new test controls the type-1 error rate for moderately large sample sizes. It furthermore controls the rate in unbalanced designs with non-normality and variance heteroscedasticity.

Note: Joint work with Frank Konietschke.

[Poster Presentation 5] Modeling Weather-Induced Home Insurance Risks: Machine Learning Approach

Asim Dey
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
adey@utdallas.edu

Insurance industry is one of the most vulnerable sectors to climate change. Assessment of future number of claims and incurred losses is critical for disaster preparedness and risk management. In this project, we study the effect of precipitation on a joint dynamics of weather-induced home insurance claims and losses. We discuss utility and limitations of such machine learning procedures as Support Vector Machines and Artificial Neural Networks, in forecasting future claim dynamics and evaluating associated uncertainties. We illustrate our approach by application to attribution analysis and forecasting of weather-induced home insurance claims in a middle-sized city in the Canadian Prairies.

Note: Joint work with Yulia R. Gel and Vyacheslav Lyubchich.

[Poster Presentation 6] On the Tsallis Statistics in the Reliability Analysis and Lifetime Testing

Fode Zhang
Department of Statistical Science
Southern Methodist University
Dallas, Texas
fodez@mail.smu.edu

Tsallis statistics, which is based on a non-additive entropy characterized by an index q, is a very useful tool in statistical mechanics, information geometry, mathematics and statistics. In statistical mechanics, it has been proved that the Tsallis statistics in the grand canonical ensemble satisfies the requirements of the equilibrium thermodynamics in the thermodynamic limit under certain conditions. In information geometric, the geometry originating from Tsallis q-entropy is equivalent to the alpha-geometry. In mathematics

and statistics, the Kullerback-Leibler (KL) divergence, one of the important measures for the distance between two probability distributions, can be q-generalized in form of Tsallis statistics. We can also q-generalize the Fourier transform and inverse Fourier transform from the point of view of the Tsallis statistics. This paper presents an application of Tsallis statistics in reliability analysis. We first show that the q-gamma and incomplete q-gamma functions are q-generalized. Then, three commonly used statistical distributions in reliability analysis are introduced in Tsallis statistics, and the corresponding reliability characteristics including the reliability function, hazard function, cumulative hazard function and mean time to failure are investigated. In addition, we study the statistical inference based on censored reliability data. Specifically, we investigate the point and interval estimation of the model parameters of the q-exponential distribution based on the maximum likelihood method. The demonstration is operated on the progressively Type-II censored data. Simulated and real-life datasets are used to illustrate the methodologies discussed in this paper. Finally, some concluding remarks are provided.

Note: Joint work with Yimin Shi, Hon Keung Tony Ng, Ruibing Wang.

[Poster Presentation 7] A Model for Predicting Individualized Absolute Risk of Contralateral Breast Cancer

Marzana Chowdhury
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
mxc122330@utdallas.edu

Purpose: Patients diagnosed with invasive breast cancer (BC) or ductal carcinoma in situ are increasingly choosing to undergo contralateral prophylactic mastectomy (CPM) to reduce their risk of contralateral BC (CBC). This is a particularly disturbing trend as a large proportion of these CPMs are believed to be medically unnecessary. Many BC patients tend to substantially overestimate their CBC risk. Thus, there is a pressing need to educate patients effectively on their CBC risk. We develop a CBC risk prediction model to aid physicians in this task. Methods: We used data from two sources: Breast

Cancer Surveillance Consortium and Surveillance, Epidemiology, and End Results to build the model. The model building steps are similar to those used in developing the BC risk assessment tool (popularly known as Gail model) for counseling women on their BC risk. Our model, named CBCRisk, is exclusively designed for counseling women diagnosed with unilateral BC on the risk of developing CBC. Results We identified eight factors to be significantly associated with CBCage at first BC diagnosis, antiestrogen therapy, family history of BC, high-risk pre-neoplasia status, estrogen receptor status, breast density, type of first BC, and age at first birth. Combining the relative risk estimates with the relevant hazard rates, CBCRisk projects absolute risk of developing CBC over a given period. Conclusions: By providing individualized CBC risk estimates, CBCRisk may help in counseling of BC patients. In turn, this may potentially help alleviate the rate of medically unnecessary CPMs.

Note: Joint work with David Euhus, Tracy Onega, Swati Biswas, Pankaj K. Choudhary.

[Poster Presentation 8] Bounded-Width Confidence Interval for Gini Index Under Complex Survey

Francis Bilson Darku
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
Francis.BilsonDarku@utdallas.edu

Gini index is an income inequality measure in determining the inequality in the distribution of income or assets among individuals or groups within a society or region. For any geographical location, the computation of Gini index is helpful in evaluating the performance of different economic policies. However, Gini index computed based on census data is available once in every 10 years or more, since many countries cannot afford to collect data from all households annually. In order to estimate Gini index for periods in-between two census years for a region with large number of households, complex sampling designs involving stratification and clustering are often used to ensure

adequate representation of groups of interest. Fixed-sample size methodologies exist for constructing confidence intervals for Gini index under complex household survey scenarios but it cannot be used to find bounded-width confidence interval for Gini index. This article therefore develops a two-stage sequential procedure for estimating and constructing a bounded-width confidence interval for Gini index under a complex survey design using the smallest possible cluster sizes.

Note: Joint work with Bhargab Chattopadhyay.

[Poster Presentation 9] On Employing Multi-Resolution Weather Data in Crop Insurance

Azar Ghahari
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
axg124131@utdallas.edu

In agriculture, changes in climate may lead to increased insurance costs. Previous works have created forecasting models that incorporate weather variables and crop production information. We consider two weather data sets for crop yield forecasting: historical observation data and climate reanalysis. We utilize screening regression, cross-validation, principal component analysis, and suggest a new trend-based clustering technique for spatial data. We apply the methods to forecast crop yields in Manitoba, Canada. The results unveil the differences obtained by using multi-resolution input weather data and provide a background for further development of data fusion techniques.

Note: Joint work with Yulia R. Gel, V. Lyubchichy, Yongwan Chun, Daniel Uribe.

[Poster Presentation 10] Wavelet Analysis of Large-P-Small-N Cross-Correlation Matrices in an fMRI Study of Neuroplasticity

Jiayi Wu
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
jxw133130@utdallas.edu

Functional magnetic resonance imaging (fMRI) allows researchers to analyze brain activity on a voxel level, but using this ability is complicated by dealing with Big Data and large noise. A novel wavelet approach, which takes into account multiresolution components of the noise, has allowed us to deal with these two complications. The approach is illustrated via a study of the brain motor cortex plasticity, whose aim was to recognize changes in connections between left and right motor cortices after button clicking training sessions. The proposed wavelet analysis allowed us to analyze pathways between left and right hemispheres on a voxel-to-voxel level via estimation of cross-correlations, and this immediately necessitated statistical analysis of large-p-small-n cross-correlation matrices. The paper explains how this problem was solved, and presents results of the dynamic analysis of neuroplasticity for 24 healthy adults. In particular, the results indicate that neuroplasiticity is individual rather than class characteristic; this understanding may help in healing head injuries and diseases.

Note: Joint work with Sam Efromovich.

[Poster Presentation 11] A Model for Sequential Refinement and Coagulation of Random Partitions

Carlo Tadue Pagani Zanini
Department of Mathematical Sciences
University of Texas at Austin
Austin, Texas
carlostpz@utexas.edu

We analyze protein activation data to find subsets of proteins which have their expression levels changed (directly or indirectly) after being exposed to

related inhibitors. To implement such inference, we develop a time-varying random partition model. Starting with an initial partition of the recorded proteins, we allow for a nested refinement of the initial partition and eventual coagulation. In the application, the changes of the initial partition reflect how the treatment effects subsets of the proteins differently and how the treatment effect eventually subsides. The proposed model builds on a Dirichlet multinomial model and on a constructive definition of the refinement.

it Note: Joint work with Peter Mueller, Yuan Ji, Fernando Quintana.

[Poster Presentation 12] Understanding Placebo Responders in the EMBARC Trial

Charles South
Department of Psychiatry/Clinical Sciences
UT Southwestern Medical Center
Dallas, Texas
csouth2@juno.com

Depression is a common, heterogeneous mental disorder that is difficult to treat. However, some patients with depression are able to improve (and potentially even go into remission) when assigned to the placebo group in a randomized control trial. A better understanding of this sub-population could allow clinicians to make more informed treatment decisions. The EMBARC study was a multi-site, randomized clinical trial comparing drug to placebo; it was the first study of its kind to also measure biological characteristics of the patients in the trial. Using the elastic net in combination with Bayesian multiple linear regression, we attempt to isolate those variables most predictive of depression level at the end of the acute phase of the trial. Further, we study the trajectories of those in the placebo group using a mixed Weibull-linear function and attempt to cluster patients on the basis of these trajectories.

Note: Joint work with Madhukar Trivedi, Jing Cao.

[Poster Presentation 13] A Bayesian Hierarachical Model for Pathway Analysis with Simultaneous Inference on Pathway-Gene-SNP Structure

Lei Zhang
Department of Mathematical Sciences
University of Texas at Dallas
Richardson Texas
lxz096120@utdallas.edu

Pathway analysis jointly tests the combined effects of all single nucleotide polymorphisms (SNPs) in all genes belonging to a molecular pathway. It is usually more powerful than single-SNP analyses if there are multiple associated variants in a given genomic region, each with a modest effect. We develop a Bayesian hierarchical model that fully models the natural three level hierarchy, namely SNPgenepathway, unlike many other methods that use ad hoc ways of combining such information for analysis. The joint modeling allows detecting not only the associated pathways but also testing for association with genes and SNPs within significant pathways and genes in a hierarchical manner, which can be useful for follow-up studies. Appropriate priors are assigned to regularize the effects and hierarchical FDR is used for multiplicity adjustment of the entire inference procedure. To study the proposed approach, we conducted simulations with samples generated under realistic linkage disequilibrium patterns obtained from the HapMap project. We find that our method has higher power than some standard approaches for identifying pathways that have multiple modest-sized variants. Moreover, in some settings, it has reasonable power to detect associated genes, a feature unavailable in other methods.

Note: Joint work with Dr. Swati Biswas, Dr. Pankaj Choudhary.

[Poster Presentation 14] Detecting Anomalies in Higher Order Structures of Dynamic Networks Using Generalized Tensor Spectrum

Ruikai Cao
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas
ruikai.cao@utdallas.edu

In the analysis of dynamic networks, one of the key tasks is anomaly detection. Its applications range from new gang formation to brain damages to money laundering. However, most currently available anomaly and change point methods are based on two dimensional structure, that is, links connecting pairs of nodes, and disregard the temporal dependence structure among consecutively observed network snapshots. In this paper, we address these challenges by introducing a new anomaly detection method based on tensor spectral characteristics. The new data-driven approach is distribution-free and allows to detect change points in higher-order network motifs. We evaluate our new anomaly detection procedure on synthetic networks, Enron email networks and the stock correlation networks. Both simulations and real data indicate competitive performance of the new method for identifying higher order structural changes in time evolving networks.

Note: Joint work with Yulia R. Gel.

[Poster Presentation 15] Statistical Analysis of Binocular Eye Gaze Trajectories

Pansujee Dissanayaka
Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas
pansujee.dissanayaka@ttu.edu

Eye gaze trajectories explain how humans would search the visual space during natural exploration. Typically, eyes move to capture objects in the visual space and head follows towards objects in focus. Subsequently eyes move to

stabilize focused objects against head movement, while occasionally exploring additional targets introduced in the visual space. It follows that the gaze trajectory alternates between saccades and fixation points. Using segmentation and clustering algorithm on trajectories of points on a 2-sphere, we analyze binocular eye movement trajectory to understand how the gaze shifts between saccades and focused regions, marked by a dense set of sub-trajectories. We propose a new clustering algorithm on eye sub-trajectories using a well-known geodesic distance metric on the sphere. We combine the eye gaze trajectory data with the head rotation velocity data to estimate when and where the gaze is compensating for the head movement. The sub-trajectory based clustered fixation regions, we obtain, has also been compared with other point based clusters (K-mean clusters) that are in the literature, and superiority of our proposed algorithm has been demonstrated. Key words: Binocular Eye gaze trajectories, Smoothing splines, Segmentation, Clustering

Note: Joint work with Jingyong Su, Bijoy K. Ghosh.

[Poster Presentation 16] Functional Data Analysis Methods in Classifying Caenorhabditis Elegans

Manjari Dissanayake
Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas
manjari.dissanayake@ttu.edu

Caenorhabditis Elegans (C. Elegans) are the simplest organisms with a nervous system. Scientists in Chemical Engineering, study muscle strength of the C. Elegans to detect the genes that affect the lifespan of beings, in order to seek for methods to enhance the lifespan of humans. In this study, data from four classes of worms with two gene types and two treatment methods are analyzed with the aim of finding a classifier using functional data analysis tools. Data for each worm are observed in the form of force against time, which are in functional form, to classify by gene type and treatment method. Cubic splines are fitted to each function and smoothed using regression approach. The smoothing parameter for splines is obtained using the Roughness-Penalty approach and the Generalized Cross Validation Criterion.

Phase-plane plots, Functional Principal Component Analysis (FPCA), and Functional F-tests are used to seek for a candidate classifier.

Note: Joint work with Adao Alexandre Trindade.

[Poster Presentation 17] A Neighborhood Hypothesis Test for Functional Data with an Application to Biological Data

Dhanamalee Bandara
Department of Mathematics and Statistics
Texas Tech University
Lubbock, Texas
dhanamalee.bandara@ttu.edu

A common problem arising when analyzing high-dimensional or functional data is that estimates of the covariance are not of full rank, resulting in the inverse being degenerate. Munk et al. (2008) applied the idea of a neighborhood hypothesis test to the one- and multi-sample problems for functional data by deriving a test statistic to determine whether a group of means are approximately equal. More precisely, they tested whether the means were within a predetermined distance to each other. Unfortunately, in many applications, this pre-determined distance is difficult to both specify and interpret. In this presentation, we present a modified test for determining whether the distance between a mean and a hypothesized function is less than a proportion of the total population variance. We will derive a test statistic that is asymptotically normal, and present both simulation studies of the power of the procedure and an application to a data set arising from biology.

Note: Joint work with Souparno Ghosh and Leif Ellingson.

[Poster Presentation 18] Predicting the Potential Economic Cost of a Car Accident Under Different Circumstances

Yihan Xu
Department of Statistical Science
Southern Methodist University
Dallas, Texas
xiaohanx@smu.edu

In this study, we aim to identify variables that contribute to the injury severity level of victims when car accidents happen. Predictive models for injury severity level are developed. We also exploit outside data source such as expenditure of vehicle accidents by states to determine the average economic cost of accidents. Our ultimate goal is to develop a paradigm to envision the maximum injury severity level and potential economic cost under different circumstances if an accident happens.

[Poster Presentation 19] Hemimethylation Analysis Using Wilcoxon Signed-Rank Tests and Bioinformatics Approaches

Shuying Sun
Department of Mathematics
Texas State University
San Marcos, Texas
s-s355@txstate.edu

DNA methylation is an epigenetic event that involves the addition of a methyl-group added to the cytosine (C) site that pairs with a guanine (G) site (i.e., CG site). This event plays an important role in both cancerous and normal cell development. Previous studies often assume symmetric methylation on both DNA strands. However, asymmetric methylation, or hemimethylation (methylation that occurs only on one DNA strand), does exist and has been reported in several cancer studies. Due to the limitation of previous DNA methylation sequencing technologies, researchers could only study hemimethylation on specific genes, but the overall genomic hemimethylation landscape remains relatively unexplored. With the development of advanced

next generation sequencing techniques, we can now measure methylation levels on both the forward and reverse strands at single CG sites in an entire genome. Analyzing hemimethylation patterns may potentially reveal regions related to undergoing tumor growth. For our research, we identify hemimethylated sites in breast cancer cell lines using Wilcoxon signed-rank tests. Meanwhile, we also identify hemimethylation patterns by grouping consecutive hemimethylated sites based on their methylation states, methylation M or un-methylation U. These patterns include regular hemimethylation clusters (e.g., MMM on one strand and UUU on another strand) and polarity (or reverse) clusters (e.g., MU on one stand and UM on another strand). We then map these hemimethylation clusters and sites to corresponding genes and study the functions of these genes. Our results reveal that hemimethylation does occur across the entire genome with notably higher numbers in the breast cancer cell lines. In addition, several of the highly hemimethylated genes may influence tumor growth or suppression. These genes may also indicate a progressing transition to a new tumor stage.

Note: Joint with Yu-Ri Lee.

[Poster Presentation 20] Predicting Parasite Counts of Apis Mellifera Using Poisson Regression

Kristina Yount
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
kry001@shsu.edu

There has been a decline in the European honey bee (Apis mellifera) that is believed to be caused by the presence of parasites, microsporidians, or parasitic arthropods. One particular parasite directly affecting honey bee colonies in Texas is varroa destructor. The damage this pest can cause ranges from weakening or shortening the lifespan of adult honey bees to causing deformities. In this study, 100 bees were selected from 58 honey bee apiaries located in 10 different ecoregions in Texas and the varroa destructor counts were recorded. Due to the count nature of the response variable and other

assumptions met, a Poisson regression model was fit in order to make predictions of varroa destructor counts under certain conditions. The covariates used to model the data include the ecoregion in which the apiary is located and the subspecies of the honey bees sampled from a particular apiary. To account for any possible interaction between the ecoregion and subspecies, the one interaction term will be included as well in the modeling. For this research, we will explore the different possible models with the given covariates and discuss any recommendations to improve the reliability of the model.

[Poster Presentation 21] Statistical Quality Control Methods for Testing the Limit of Equine Racing Excellence

Jillian Parker
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
jep030@shsu.edu

Quality control is a statistical method used to monitor the quality of products and services to ensure that a process maintains a desired performance level. In this study, the American Triple Crown races are analyzed using quality control methods to determine whether equines have reached the pinnacle of horse racing excellence. Individuals/Moving Range, Cumulative Sum and Exponentially Weighted Moving Average control charts are used to identify potential changes in the mean winning times for the American Triple Crown races from 1919 to 2016. Further research addresses the issue of high correlation between winning race times using a multivariate control charting technique, since many of the same thoroughbreds compete in all three of these major races. Some interesting and unexpected results are realized from this study as well.

[Poster Presentation 22] Hot Spots and Cluster Analysis of a Mosquito-Born Disease Spread

Heranga Rathnasekara
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
hkr009@shsu.edu

Spatial Statistics provide a variety of useful analytical techniques for studying spatial characteristics, relationships, patterns and trends in geographical phenomena. Mapping of clusters plays a fundamental role in identifying the event clusters that are statistically significant. This study employs two different methods, the Getis-Ord Gi* statistic and the Local Morans I statistic, to detect the spatial pattern of disease spread for a particular mosquito-born disease. It focuses on disease incident counts identified by the blood IgG (Immunoglobulin G) index and obtained from forty-nine census block groups. The positive and negative IgG indexes, respectively, indicate places where the disease occurred and places that a similar but different fever occurred. All disease and collection-site identification information has been suppressed at the request of those collecting the data, as it is highly proprietary. Statistically significant hotspots, cold spots clusters and special outliers are identified using the ArcMap component of the ArcGIS system. Results provide a vital source of information for the development of methodologies for future research and disease prevention.

Note: Joint work with Heranga Kalpani Rathnasekara and Melinda Holt.

[Poster Presentation 23] Sensitivity Analysis of Errors in Variables Multiple Regression

Dholamulla Preethika
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
dnp007@shsu.edu

Multiple linear regression using Ordinary Least Squares (OLS) assumes that the regressors are measured without error. When this assumption does not

hold, OLS estimators of the regression coefficients may be inconsistent and biased. Model II, or errors-in-variables, regression relies heavily on an accurate knowledge of the reliability of the regressor measurements. This study examines the sensitivity of the measurement error model to the assignment of reliability values. Three datasets are examined: one with low reliability, one with moderate reliability and one with high reliability in independent variable measurement. The analysis examined the sensitivity of the model with respect to R2, Root Mean Square Error (RMSE) and the regression coefficients. The models were fit using the eivreg command in STATA.

Note: Joint work with M. M. Holt.

[Poster Presentation 24] Empirical Insight into Vector Length Versus Information Content

Mohammad Bhuyian
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
mab184@shsu.edu

This paper empirically investigates the proposition which says: Any item can be described as accurately as one wishes with a sufficiently long vector. A specific example is utilized in a discriminant and classification analysis environment to provide insight into a better understanding of this proposition. Of course, a number of specific determinations come into play beginning with what the specific measured variables are that comprise the vector and what discriminant information they possess, the accuracy of measurement, etc. Moreover, the Box and Cox transformation is applied to the data collected to attain near normality and the additional positive aspects of this transformation are seen in terms of improvements of the classification results. Finally, cross-validation results are presented as well. Again the results of this poster are empirically-based and are used to attain additional insight into the theoretical structure of discriminant and classification analyses and the applications of such.

Note: Joint work with M.M. Holt.

[Poster Presentation 25] Longitudinal Study of the Efficacy of Smoking Cessation Treatment in Cocaine/Meth Dependent Patients

Anita Bhandari Sharma
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
axb085@shsu.edu

Smoking cigarettes remains to be a leading cause of deaths and productivity losses in general population. Its prevalence in illicit drug users is higher as compared to the general population. Our main goal of this project is to evaluate the impact of smoking cessation treatment plus treatment as usual (SCT + TAU) compared to treatment as usual (TAU) from a longitudinal study in cocaine/meth dependent patients. The data set is obtained from Clinical Trials Network (CTN) of the National Drug Abuse Treatment (NIDA) program. Average carbon monoxide (CO) level (1 if CO = 8ppm and 0 if CO¿ 8ppm) is measured over time from 532 patients randomized to either SCT+TAU or TAU. The missing values were imputed using MCMC method. We applied generalized estimating equation (GEE) and generalized linear mixed model (GLMM) considering CO level as response with other applicable covariates. The results showed that SCT + TAU is effective in achieving smoking abstinence for cocaine/meth dependent patients.

Note: Joint work with Dr. Ram C. Kafle

[Poster Presentation 26] Modeling Bowling Effectiveness in T20I Cricket: A Quantile Regression Approach

Sulaitha Bowala
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
smb098@shsu.edu

Bowling effectiveness is a key factor in winning cricket matches. The captain of the team should decide to use the right bowler at the right time, so that

the team can optimize the outcome of the game. In this study, we investigate the effectiveness of different types of bowlers at different stages of the game based on conceded runs in each over. Bowlers are categorized into three types, namely fast bowlers, medium fast bowlers and spinners. A match is divided into four stages, namely, stage 1: over 1-6 (PowerPlay), stage 2: over 7-10, stage 3: over 11-15, and stage 4: over 16-20. Quantile regression methodology is used with statistical analysis.

[Poster Presentation 27] Quality Control Techniques to Monitor On-Time Performance of Houston Airport

Meghan Sealey
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
mds010@shsu.edu

Airports have seen an increase in poor timing and delays. As one of the top 10 busiest airports in the nation, Houstons George Bush Intercontinental Airport (IAH) is held to a high standard of getting its passengers to their destination on time. To monitor whether Bush Intercontinental Airport has held up to this standard, several on-time and delayed variables were taken from the FAA database of the last 13 years to see if their arrival and departure times maintained a decent rate. Univariate quality control techniques were used on these time variables to determine the pattern of on-time flights, as well as the pattern of delays pertaining to the Houston airport. The data shows that Bush Intercontinental Airport reached a peak in performance in September of 2004, yet has strayed away from this accomplishment throughout recent years. Currently, Houstons busiest airport has not shown improvement in minimizing their delay times, but has shown potential in increasing their on-time performance within the last year. The scope of this study monitors the performance of the top ten national airports, with Houstons George Bush Intercontinental Airport being the main interest. When compared to the busiest airport in the nation, Atlantas Hartsfield-Jackson Airport, Bush Intercontinental Airport definitely has room to improve in the on-time performance category.

[Poster Presentation 28] Using the Statistical Approach to Model Crude Oil Data

Audrene Edwards
Department of Mathematics
Lamar University
Beaumont, Texas
audreneedwards@yahoo.com

Extreme value analysis is an area of statistical analysis that can be used in many disciplines, such as engineering, science, business and statistics. Extreme Value Theory (EVT) deals with the extreme deviations from the median probability distribution and is used to study rare but extreme events. When considering the use of EVT to model data where extremes exist, one must consider whether extreme events are stationary or non-stationary. There are two methods that can be used within EVT for effective modeling of data; the block maxima method, which follows a generalized extreme value (GEV) distribution, and the peaks over threshold method which follows a generalized Pareto distribution (GPD). For this study, EVT will be used to model spot prices for West Texas Intermediate (WTI) crude oil data from January, 1986 to February, 2016. With the spot prices for crude oil data, descriptive statistics will be used to model and interpret the characteristics of the data set, while determining whether the data contain extreme data. Next, hypothesis testing will also be used to determine whether the data is stationary or nonstationary. After determining if the data is stationary or nonstationary, a goodness of fit test will be used to determine if the data can be modeled by GEV distribution nonstationary model. With the conclusion that the data are nonstationary, the block maxima method for non-stationary extreme events will be used to analyze return levels. The return levels provide insight about the cost of future crude oil prices.

[Poster Presentation 29] Tracking Number of HIV Infections Using Poisson Process

Destiny Allain
Department of Mathematics
Lamar University
Beaumont, Texas
allain.destiny@gmail.com

The Poisson process is a stochastic process used to count the number of events that take place over a certain period of time. The probability of an event taking place is a Poisson distributed random variable and the amount of time between events occurring is exponentially distributed. HIV infections are difficult to track because of the lag in time between contracting the disease and when the symptoms of the disease actually emerge. A simulation of HIV infections can be produced in R and the Poisson process can be used to more reliably track the number of HIV infections.

Note: Joint work with Dr. Kumer Das.

[Poster Presentation 30] Statistical Quality Control to Test Human Limits in Golf

Feiyang Gao
Department of Mathematics and Statistics
Sam Houston State University
Huntsville, Texas
fxg009@shsu.edu

Golf is a sport that relies on human judgment and athleticism to compete for the lowest stroke score. Golf equipment has seen more improvement than any other sport over the last 40 years. The purpose of this project is to find out if the human performance in golf has reached its excellence by using statistical quality control methods. The data is collect from four recognized professional Major Championships within the mens category played on an annual basis from 1895 to 2016. These data are prestigious within the sport; represent differing conditions of the game, and the best of human performance in golf throughout the world. Univariate process control methods (I-MR, EWMA

and CUSUM) applied on the total score in each Championship to catch the trend of data over time. In addition, multivariate process control methods (TGV AND MEW) used on the round score to detect changes throughout the same game. The research shows the human limit of excellence in sport of golf have not changed over the past 40 years.

Note: Joint with Dr. Scariano

[Poster Presentation 31] Dimension Reduction Techniques for Analyzing High Dimensional Microarray Data

Mithun Acharjee
Department of Mathematics
Lamar University
Beaumont, Texas
mithunacharjee@du.ac.bd

Microarray data are being currently used for the expression levels of thousands of genes simultaneously. They present new analytical challenges because they have high dimensionality. Thus, the dimension reduction techniques are usually needed to reduce the variable space before the subsequent analysis is carried out. The goal of this study is to reduce the dimension of the gene expression data using Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). For the aspiration of the analysis, we consider the dataset which consists of 240 samples from patients with diffuse large B-cell lymphoma (DLBCL), with gene expression measurements for 7,399 genes.
Keywords: Microarray data, Principal component analysis, Singular value decomposition, Gene expression measurement.

Note: Joint work with Kumer Pail Das.

[Poster Presentation 32] A New R Package Snowboot for Non-parametric Boostrap Inference on Random Network

Yuzhou Chen
School of Natural Sciences and Mathematics
University of Texas at Dallas
Richardson, Texas
yxc154630@utdallas.edu

Complex networks are used to describe a variety of modern disparate social systems and natural phenomena, for power grids to customer segmentation to human brain connectome. Challenges of parametric model specification and validation inspire a search for more data-driven complex networks. In this poster we discuss methodology and R impelementation of the two bootstrap procedures on random network, that is, patchwork bootstrap of Thompson et al. (2016) and Gel et al. (2016) and vertex bootstrap Snijders and Borgatti (1999). To our knowledge, the new R package Snowboot is the first implementation of bootstrap inference on networks in R. Our new package is accompanied with the detailed documentation and users manual and is compiled with the popular R package on network studies igraph. We provide pseudo-codes and showcase the realization of bootstraps algorithms in R. In addition, we evaluate the patchwork bootstrap and vertex bootstrap with extensive simulation studies and illustrate their utility in application to airline networks and power grid resilience studies.

Note: Joint work with Yulia R. Gel, Vyacheslav Lyubchich, Kusha Nezafati

[Poster Presentation 33] Improved Tests for Monotonic Trend in Time Series Data

Xiaojie Zhu
Department of Statistical Science
Southern Methodist University
Dallas, Texas
xioajiez@smu.edu

For testing the monotonic trend, Brillinger (1989) proposed a test statistic, which is a ratio of a linear combination of the time series values to an estimate

of the standard error of the linear combination. However, when there is highly correlated residuals or short records, the procedure proposed by Brillinger (1989) tends to a problem that the observed significance level is higher than the nominal level. We found that the reason could be discrepancies between the empirical distribution of the test statistic and the theoretical asymptotic standard normal distribution. Hence, based on the Brillingers method, we introduced bootstrap procedures to the test for monotonic trend, which shows improved significance level and comparable power. The proposed procedures are further applied to global mean temperature anomaly from 1880 to 2016, which shows a significant monotonic trend.

Note: Joint with Hon Keung Tony Ng and Wayne Woodward