Oral Presentation Abstracts – Friday

1. Presenter: Daniel F. Heitjan, Ph.D. (dheitjan@smu.edu)

Affiliation: Department of Statistical Science, Southern Methodist University and Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas

Collaborators: Yu Lan (Southern Methodist University)

Title: Projecting survival and lifetime costs from short-term smoking cessation trials

Abstract: Cigarette smoking is a leading potentially preventable cause of morbidity and mortality. Therefore, there has been considerable interest in the development of treatments - both behavioral and pharmacologic - to assist smokers to conquer their nicotine addiction. As the number of smokers is large - roughly 20% of the US population, and more in many other countries - it is critical to evaluate not only the effectiveness of such treatments, but also their costs. The fact that many smokers will not experience the major health effects of their addiction for many decades substantially complicates this evaluation. Moreover, clinical trials of smoking cessation treatments commonly take as their primary endpoint a short-term outcome such as six-month quit rate. Thus the evaluation of the cost-effectiveness of smoking cessation treatment, in its usual form of cost per life-year saved, involves a considerable extrapolation. In this talk, I will describe a micro-simulation model that projects early quit rates and treatment costs into survival and long-term cessation-related costs for a general smoking population.
2. Presenter: Daijin Ko, Ph.D. (daijin.ko@utsa.edu)

Affiliation: Department of Management Science and Statistics, University of Texas at San Antonio, Texas

Collaborators: Myung Ko, Nicole Beebe (University of Texas at San Antonio)

Title: Detection of hostile insiders

Abstract: Outliers are points with unusual or dissimilar patterns which appear to be inconsistent with the main body of the data. Outlier detection methods are used to detect intrusions on a network, fraudulent activity on financial accounts, malware on computer systems, and abnormal patterns in medical images. In this work, we consider methods most useful in detecting hostile insiders—specifically insiders that are collecting information on their work computers either for personal use that is inconsistent with organizational norms (as is the case with pornography), or collecting information with the intent of exfiltration (as is the case with those stealing sensitive information). The data we use (aggregated storage profile data) contain many attributes of different types including numeric variables, categorical variables, and directional variables. Many classic outlier detection methods that rely on distance measures struggle from the lack of appropriate distance measure on such mixture of variables. Here, we used supervised learning tools such as Multivariate Adaptive Regression Spline (MARS) method to build a predictive model that predicts the value of that variable from the all other variables including numeric, categorical and directional variables. Appropriate model building tools and subset selection are used and the resulting residuals are used in measuring the outlying-ness of the point. We show that our approach well differentiate suspected malicious behavior from normal behavior by comparing the accumulation of sensitive information against the user’s historical norms and against peers within the organization (i.e. with the same job role).
3. **Presenter:** Joel E. Michalek, Ph.D. (michalekj@uthscsa.edu)

**Affiliation:** Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, Texas

**Title:** Agent Orange exposure and monoclonal gammopathy of uncertain significance: an Operation Ranch Hand veteran cohort study

**Abstract:** Multiple myeloma has been classified as “exhibiting limited or suggestive evidence” of an association with exposure to herbicides in Vietnam War veterans. Occupational studies have shown that other pesticides (i.e., insecticides, herbicides, fungicides) are associated with excess risk of multiple myeloma and its precursor state, monoclonal gammopathy of uncertain significance (MGUS), however, until now no studies have uncovered such an association in Vietnam War veterans. Our purpose was to examine the relationship between MGUS and exposure to Agent Orange, including its contaminant 2,3,7,8 tetrachlorodibenzo-p-dioxin (TCDD) in Vietnam War veterans. This was a prospective cohort study conducted in 2013 to 2014, testing for MGUS in serum specimens collected and stored in 2002 by the Air Force Health Study (AFHS). The relevant exposure data collected by the AFHS was also used. We tested all specimens in 2013 without knowledge of the exposure status. The AFHS included former US Air Force personnel who participated in Operation Ranch Hand (Ranch Hand veterans) and other US Air Force personnel who had similar duties in Southeast Asia during the same time period (1962 to 1971) but were not involved in herbicide spray missions (comparison veterans). This talk summarizes the findings of this study and the history of the Ranch Hand spray operations in Vietnam and the AFHS.
Oral Presentation Abstracts – Saturday

1. **Presenter:** Khairul Islam, Ph.D. (khairul.islam@tamuk.edu)

   **Affiliation:** Mathematics Department, Texas A&M University at Kingsville, Texas

   **Title:** On communicating the concept of sampling distribution: an alternative approach

   **Abstract:** Sampling distribution plays a significant role in the process of statistical inference. It is well-known that this topic can be challenging for many students. Many students find it overwhelming, which leads to loss of concentration and other adverse consequences, including dropout. Often, the topic is introduced superficially or skipped while addressing estimation and test of hypothesis problems, which makes students suffer further. Students need a thorough understanding of sampling distribution techniques in order to adequately solve problems. In this presentation, we address an alternative approach of sampling distribution via multiple examples and simulation. We utilize an open source software R as a simulation tool. It is expected that instructors would benefit from adapting this approach.

2. **Presenter:** Amy Nussbaum, Ph.D. (anussbaum@smu.edu)

   **Affiliation:** Department of Statistical Science, Southern Methodist University, Dallas, Texas

   **Collaborators:** Cornelis Potgieter, Mike Chmielewski (Southern Methodist University)

   **Title:** Discretized longitudinal data with two sources of measurement error

   **Abstract:** Personality traits are latent variables, and as such, are impossible to measure without the use of an assessment. Responses on the assessments can be influenced by both state-related error and measurement error, obscuring the true trait levels. These assessments typically utilize Likert scales, which yield only discrete data. The loss of information due to the discrete nature of the data represents an additional challenge in estimating an individual’s true trait level. We will explore a latent variable model relating personality traits, state error and measurement error. Two methods
for parameter estimation are detailed: correlation reconstruction, which is based on polychoric correlations, and a Monte Carlo EM algorithm for maximum likelihood implementation. These methods are applied to a motivating dataset of 440 college students taking the Big Five inventory twice in a two-month period.

3. **Presenter:** Tahir Ekin, Ph.D. (t_e18@txstate.edu)
   
   **Affiliation:** Department of Computer Information Systems and Quantitative Methods, Texas State University, San Marcos, Texas
   
   **Collaborators:** Muzaffer Musal, Greg Lakomski (Texas State University)
   
   **Title:** Unsupervised Bayesian approaches for medical fraud assessment
   
   **Abstract:** U.S. governmental agencies report that three to ten percent of the annual health care spending is lost to fraud, waste and abuse. These fraudulent transactions have direct cost implications to the tax-payers and diminish the overall quality of medical services. This talk discusses the use of unsupervised Bayesian approaches as pre-screening tools to aid in medical fraud assessment. These can help identify the hidden patterns among providers and medical procedures via outlier detection and similarity assessment. We provide an illustration with a focus on Latent Dirichlet allocation and U.S. Medicare Part B data.

4. **Presenter:** Thierry Chekouo, Ph.D. (tchekouo@mdanderson.org)
   
   **Affiliation:** Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas
   
   **Collaborators:** Francesco Stingo, James Doecke, Kim-Anh Do (University of Texas MD Anderson Cancer Center)
   
   **Title:** A Bayesian approach for the integrative analysis of omics data: a kidney cancer case study
**Abstract:** Integration of genomic data from multiple platforms has the capability to increase precision, accuracy, and statistical power in the identification of prognostic biomarkers. A fundamental problem faced in many multi-platform studies is unbalanced sample sizes due to the inability to obtain measurements from all the platforms for all the patients in the study. We have developed a novel Bayesian approach that integrates multi-regression models to identify a small set of biomarkers that can accurately predict time-to-event outcomes. This method fully exploits the amount of available information across platforms and does not exclude any of the subjects from the analysis. Moreover, interactions between platforms can be incorporated through prior distributions to inform the selection of biomarkers and additionally improve biomarker selection accuracy. Through simulations, we demonstrate the utility of our method and compare its performance to that of methods that do not borrow information across regression models. Motivated by The Cancer Genome Atlas kidney renal cell carcinoma dataset, our methodology provides novel insights missed by non-integrative models.

5. **Presenter:** Matthias Katzfuss, Ph.D. ([katzfuss@stat.tamu.edu](mailto:katzfuss@stat.tamu.edu))

**Affiliation:** Department of Statistics, Texas A&M University, College Station, Texas

**Title:** Multi-resolution approaches for big spatial data

**Abstract:** Remote-sensing instruments have enabled the collection of big spatial data over large spatial domains such as entire continents or the globe. Basis-function representations are well suited to big spatial data, as they can enable fast computations for large datasets and they provide flexibility to deal with the complicated dependence structures often encountered over large domains. We propose two related multi-resolution approximations (MRAs) that use basis functions at multiple resolutions to (approximately) represent any covariance structure. The first MRA results in a multi-resolution taper that can deal with large spatial datasets. The second MRA is based on a multi-resolution partitioning of the spatial domain and can deal with truly massive datasets, as it is highly scalable and amenable to parallel computations on distributed computing systems.
Title: Hidden Markov model-based methods for identifying differential methylation

Abstract: DNA methylation is an epigenetic event that mainly occurs at the CG sites, where a cytosine (C) is paired with a guanine (G), in a mammalian genome. This event plays a key role in regulating gene expression. It is important to study differential methylation (DM) patterns between two groups of samples, which can help to study genomic regions associated with diseases. With next generation sequencing (NGS) technologies, it is now possible to identify DM patterns by considering methylation at the single CG site level in an entire genome. However, it is challenging to analyze large and complex NGS data. In order to address this difficult question, we have developed two new statistical methods using hidden Markov models (HMM). These two methods are named HMM-Fisher and HMM-DM. For the HMM-Fisher, we first use a hidden Markov chain to model methylation signals at each CG site and infer the methylation state as Not methylated (N), Partially methylated (P), and Fully methylated (F) for each sample. We then use Fisher’s exact test to identify DM CG sites. For the HMM-DM, we use a hidden Markov model to identify DM CG sites accounting for spatial correlation across CG sites and variation across samples. For both methods, identified sites are summarized as DM singleton CG sites and DM regions (including at least two consecutive CG sites). In this presentation, these two methods will be introduced in detail and are then compared with other available methods using both simulated data and real sequencing data.
Poster Presentation Abstracts

1. Presenter: Francis Bilson Darku (xb130230@utdallas.edu)
   
   Affiliation: University of Texas at Dallas
   
   Collaborators: Ken Kelly (University of Notre Dame) and Bhargab Chattopadhyay (University of Texas at Dallas)
   
   Title: Accuracy in effect size estimation for i.i.d. observations
   
   Abstract: Effect sizes are widely used quantitative measures of the strength of a phenomenon, with many potential uses in psychology and related disciplines. In this article, we propose a general theory for a sequential procedure for constructing sufficiently narrow confidence intervals for effect sizes (such as correlation coefficient, coefficient of variation, etc.) using smallest possible sample sizes, importantly without specific distributional assumptions. Fixed sample size planning methods, which are commonly used in psychology, economics and related fields, cannot always give a sufficiently narrow width with high coverage probability. The sequential procedure we develop is the first sequential sampling procedure developed for constructing confidence intervals for effect sizes with a prespecified width and coverage probability. As a general overview, we first present a method of planning a pilot sample size after the research goals are specified by the researcher. Then, after collecting a sample size as large as the estimated pilot sample size, a check is performed to assess if the conditions to stop the data collection have been satisfied. If not, an additional observation is collected and the check is performed again. This process continues, sequentially, until the specified conditions are satisfied. Our method ensures a sufficiently narrow confidence interval width along with a specified coverage probability.

2. Presenter: Zifei Han (zifei.han@utsa.edu)
   
   Affiliation: University of Texas at San Antonio
   
   Collaborator: Victor De Oliveira (University of Texas at San Antonio)
   
   Title: On the correlation structure of Gaussian copula models for geostatistical count data
Abstract: We describe a class of random field models for geostatistical count data based on Gaussian copulas. Unlike hierarchical Poisson models often used to describe this type of data, Gaussian copula models allow a more direct modeling of the marginal distributions and association structure of the count data. We study in detail the correlation structure of these random fields when the family of marginal distributions is either negative binomial or zero-inflated Poisson; these represent two types of overdispersion often encountered in geostatistical count data. We also contrast the correlation structure of one of these Gaussian copula models with that of a hierarchical Poisson model having the same family of marginal distributions, and show that the former is more flexible than the latter in terms of range of feasible correlation, sensitivity to the mean function and modeling of isotropy. An exploratory analysis of a dataset of Japanese beetle larvae counts illustrate some of the findings. All of these show that Gaussian copula models are a useful alternative to hierarchical Poisson models, especially for geostatistical count data that display substantial correlation and small overdispersion.

3. Presenter: Chelsea McCarty Allen (cdallen@smu.edu)

Affiliation: Southern Methodist University

Collaborator: Daniel F. Heitjan (Southern Methodist University)

Title: Comparing models for the heaping mechanism with cigarette count data

Abstract: A problem sometimes seen in self-reported data is that of “heaping”, where certain numerals (often multiples of two, five, ten or twelve) appear with higher frequency than would be expected. In smoking cessation studies, for example, it is common to see daily cigarette count data having excess counts of 10, 20, 30, and so on. A smoking cessation dataset where both time-line follow-back (TLFB-delayed recall) and ecological momentary assessment (EMA-recording events electronically as they occur) were used to record subjects’ cigarettes smoked per day presents a unique opportunity to examine and model the mechanism of heaping. This double-coded dataset allows us to see the impact of heaping, which is notoriously an issue with the TLFB method, and to attempt to model the heaped data (TLFB) based on the “real” data (EMA). In the novel “proximity and gravity” model, the conditional probability of reporting a TLFB value is based on its proximity to the EMA value and the gravity or “attractiveness” of the TLFB value. One parameterizes the model in terms of a measure of proximity (of the truth to the reported value) and a vector of gravities (designating the intrinsic attractiveness of
each possible reported value). In this project, we explore a range of possible variants of the proximity-gravity model to determine which provides the best fit to the TLFB-EMA data.

4. Presenter: Xiao Wang (xwang@stat.tamu.edu)

Affiliation: Texas A&M University

Collaborator: Mohsen Pourahmadi (Texas A&M University)

Title: Multivariate time series forecasting by grouping and time-varying dynamic orthogonal component

Abstract: An innovative algorithm is introduced to analyze high-dimensional non-stationary multivariate time series and to make multi-step ahead forecasting.

5. Presenter: Tianyu Bai (tianyu.bai24@gmail.com)

Affiliation: University of Texas at San Antonio

Collaborator: David Han (University of Texas at San Antonio)

Title: Inference for the progressively Type-I censored step-stress accelerated life test under interval monitoring

Abstract: In this work, we considered a step-stress accelerated life test under progressive Type-I censoring when a continuous monitoring of failures is infeasible but inspections at particular time points is possible. In addition to the accelerated failure time model to explain the effect of stress changes, a general scale family of distributions was considered for flexible modeling by allowing different lifetime distributions at different stress levels. When the in-section points align with the stress-change time points, the maximum likelihood estimators of the scale parameters and their conditional density functions could be derived explicitly. If the inspection points do not align with the stress-change time points, the parameter estimates can be obtained numerically. Using the exact distributions of the estimators, confidence intervals for the respective parameters were obtained along with the asymptotic
and bootstrap confidence intervals for comparison of their performances.

6. **Presenter:** Dylan Dimock (dddimock@hotmail.com)

**Title:** Does Big Data have a place in psychological research?

**Abstract:** It is irrefutable that sampling plays an enormous role in psychological research. Sampling is known to be a reasonably accurate, cost effective, time efficient, and simple way of collecting information from a population. But what happens when all the data for an entire population is collected? Does sampling become redundant? On the surface, the idea of having all data that’s needed is intriguing but theoretically, as well as practically, it’s flawed. The concept of big data is based on the premise that all of the data from a population has been collected: N=All. Despite the intuitive paradox, having all desired data from a population does not allow greater more accurate results to be produced than traditional sampling methods. For psychological research, big data doesn’t solve any real-world issues despite proclamations of its superiority over traditional sampling. In addition, using N=All places more restrictions on methods of statistical analysis that are commonly used in psychological research because of the inability to factor in sample size (n).

7. **Presenter:** Cynthia Garza (cindy_ann2123@yahoo.com)

**Collaborator:** Dylan Dimock

**Title:** The ethical issues of Big Data

**Abstract:** The present day technological advancements in smart phones and computer systems has made it possible to collect and store enormous amounts of information from a population with minimal time and effort. With these capabilities, organizations and businesses have extracted and used all generated data from an entire selected population for their own research using data from an entire population is referred to as Big Data: N=All. Although the luxury of not having to exert much time, effort, and money to recruit and collect data from all of the participants in a population is alluring, the
participants’ informed consent to use the data is often overlooked or given with vague understanding. Informed consent is defined as full disclosure of the nature of the research and the participant’s involvement, adequate comprehension on the part of the potential participant, and the participant’s voluntary choice to participate? Since the only feasible way to obtain big data is through data generated by smart phones and web activity, and every user of this technology is typically unaware of what’s being done with the data that they generate, the ethics of using Big Data need to be called into question.

8. Presenter: Ryan Patrick Alexander McShane (rmcshane@smu.edu)
Affiliation: Southern Methodist University

Collaborators: Ian R. Harris, Ronald W. Butler, Andrew T.A. Wood
(Southern Methodist University)

Title: Infinity-calibrated Laplace approximations
Abstract: Laplace approximation is typically used to find integrals of functions which are difficult to evaluate. Using a second-order Taylor expansion, the integral can be re-expressed as the kernel of a normal density. Ultimately, expressing the integral in this way leads to faster calculations as well as fairly accurate values. Laplace approximations are particularly useful in the multivariate setting, for example hypergeometric functions with matrix arguments. The hypergeometric function is an infinite sum with no known closed form. In Butler and Wood (2002), Laplace approximations were calibrated to 0. We give a second calibration option by calibrating the approximations to infinity. We compare the values obtained in Butler and Wood (2002) and the infinity-calibrated values to their true values (obtained by simulation). An example concerning the moments of the non-central Wilk’s is given.

9. Presenter: Raanju Sundararajan (raanch316@gmail.com)
Affiliation: Texas A&M University
Collaborator: Mohsen Pourahmadi (Texas A&M University)

Title: Nonparametric change point detection in multivariate nonstationary time series

Abstract: Detecting change points in multivariate time series is an important problem with numerous applications. Much of change point literature is on tackling this problem in the univariate case or is parametric in nature. We develop a nonparametric method to detect multiple change points in multivariate piecewise stationary processes when the locations and number of change points are unknown. Based on a test statistic that measures differences in the spectral density matrices through the $L_2$ norm, we propose a two stage procedure. The first stage identifies potential change points using a sequential test and the second stage tests for the significance of each of the potential change points. The asymptotic properties of the test for significant change points under the null and alternative hypothesis are derived. Monte Carlo simulation of values of a stationary process given its spectral density matrix is used to obtain critical values of the test statistic under the null. We illustrate the better performance of our method in comparison to some of the recent methods through a few simulation examples and discuss an application of our method in seismology.

10. Presenter: Audrene Edwards (audrenee12003@yahoo.com)

Affiliation: Lamar University

Title: Using statistical approaches to model natural disasters

Abstract: The study of extremes has attracted the attention of scientists, engineers, actuaries, policy makers, and statisticians for many years. Extreme value theory (EVT) deals with the extreme deviations from the median of probability distributions and is used to study rare but extreme events. EVT’s main results characterize the distribution of the sample maximum or the distribution of values above a given threshold. In this study, EVT has been used to construct a model on the extreme and rare earthquakes that have happened in the United States from 1700 to 2011. The primary goal of fitting such a model is to estimate the amount of losses due to those extreme events and the probabilities of such events. Several diagnostic methods (for example, QQ plot and Mean Excess Plot) have been used to justify that the data set follows generalized Pareto distribution (GPD). Three estimation techniques have been employed to estimate parameters. The consistency and reliability of estimated parameters have
been observed for different threshold values. The purpose of this study is manifold: first, we investigate whether the data set follows GPD, by using graphical interpretation and hypothesis testing. Second, we estimate GPD parameters using three different estimation techniques. Third, we compare consistency and reliability of estimated parameters for different threshold values. Last, we investigate the bias of estimated parameters using a simulation study. The result is particularly useful because it can be used in many applications (for example, disaster management, engineering design, insurance industry, hydrology, ocean engineering, and traffic management) with a minimal set of assumptions about the true underlying distribution of a dataset.

11. Presenter: Jian Wang (jianwang@mdanderson.org)

Affiliation: University of Texas MD Anderson Cancer Center

Collaborator: Sanjay Shete (University of Texas MD Anderson Cancer Center)

Title: Estimation of indirect effect when the mediator is a censored variable

Abstract: A mediation model is a statistical approach that explores the direct and indirect effects of an initial variable on an outcome variable by including a mediator variable. In many realistic scenarios, investigators observe censored data instead of the complete data. Current research in mediation analysis for censored data focuses mainly on censored outcome variables, but not when the mediator is a censored time-to-event variable. Few approaches to this area, with limited applicability, have been suggested in the literature. In this study, we first established the bias in the estimations of different paths (i.e., the effect of the initial variable on mediator [a], the effect of the mediator on outcome [b] and the effect of the initial variable on the outcome given mediator [c]) and indirect effects when analyzing the data using existing approaches, Mplus, complete-case analysis, and the Tobit mediation model. We then proposed a strategy that can accurately estimate the coefficients of different paths based on the accelerated failure time model and a multiple imputation approach. We adapted a measure of indirect effect from the study of Li et al. for the mediation model with a censored mediator, which can assess the indirect effect at both the group and individual levels which are useful for intervention. We conducted simulation studies to investigate the performance of the proposed
strategy compared to that of the existing approaches. The proposed strategy accurately estimates the coefficients of different paths with accurate coverage probabilities.

12. Presenter: Maria Ivette Barraza-Rios (mbrios@utep.edu)

Affiliation: University of Texas at El Paso

Collaborator: Joan G. Staniswalis (University of Texas at El Paso)

Title: Estimating the coefficients of a linear differential operator

Abstract: Principal Differential Analysis (PDA; Ramsay, 1996) is used to obtain low-dimensional representations of functional data, where each observation may be represented as a curve. PDA seeks to identify a Linear Differential Operator (LDO) denoted by L that satisfies as closely as possible that \( Lf = 0 \) for each functional observation \( f \). A theorem from analysis establishes that the coefficients of the LDO are in the Sobolev space, and thus can be approximated by B-splines. Current PDA software used to estimate the LDO assumes that the leading coefficient is 1. We present a method that eliminates this restriction, and ensures that the coefficients of the LDO are in the Sobolev space, and that their approximation by B-splines is mathematically valid. The proposed method is inspired by results in linear regression (Frees, 1991 and Wu, 1986) that show that the weighted average of pairwise slopes between data points is equivalent to the least squares estimator of the regression line slope. We address numerical complications arising from eliminating the assumption that the leading coefficient of the LDO is 1. By analyzing data, the proposed method is compared with pda.fd (R library fda).

13. Presenter: Kristina Yount (kry001@shsu.edu)

Affiliation: Sam Houston State University

Title: Using weight least products regression with heteroscedasticity

Abstract: Errors-in-variables (EIV) regression is used when both the X and Y variables are measured with error. The most widely used EIV methods include major axis, standardized major axis, ranged major
axis, ordinary least squares bi-sector and a non-informative Bayesian approach. Like ordinary least squares regression, these methods may not perform well if the assumption of homoscedasticity is not met. Weighted Least Products regression is one EIV method that has been proposed specifically for data containing heteroscedasticity. We will apply this method to a data set that is known to have heteroscedasticity and compare the results against traditional EIV techniques.

14. Presenter: Lochana Kanishka Palayangoda (lkp012@shsu.edu)
Affiliation: Sam Houston State University
Collaborator: Ananda B Manage
Title: Methods of evaluating the consistency of cricket players
Abstract: Consistency is an essential measurement to estimate players’ performance in any kind of game. In game of cricket, consistency can be measured for both batting and bowling. Proper consistency measurement will provide the selectors and evaluators to identify the current form and performance trend of cricket players. In this study, three methods of are introduced to measure the consistency. They are Gini Index, Coefficient of Variation, and Proportion Method. With all the approaches batting average, batting strike-rate, bowling economy rate, bowling strike rate and bowling average are evaluated for selected players and ranked them accordingly. Cumulative average of performance measurements is applied to segregate the positive consistent and negative consistent innings. Negative consistent innings are used to measure the consistency of the player. To estimate the not-out runs of batsmen, estimators, $e_2$, $e_6$ and $e_{26}$ have been used from Lemmer (2008a). Gini Index is applied in economics and related fields as a measurement of equality. In this paper, it is introduced as a measurement for consistency for cricket players. Gini Index has been determined in two ways. One way is performance measurement of each inning and the other way is moving average of the performance measurement. Coefficient of Variation is measured for the moving average of the performance measurement. It appears that, player ranking through Gini Index and Coefficient of Variation have very similar results. Proportion method also introduced to identify the percentage of positive consistent innings for both batsmen and bowlers.

15. Presenter: Christopher Casement (chris_casement@baylor.edu)
Affiliation: Baylor University

Collaborator: David Kahle (Baylor University)

Title: **Lineup-style methods for graphical prior elicitation**

Abstract: Prior specification is fundamental to the Bayesian paradigm. Informative priors allow analysts to inject expert opinion directly into the modeling process. However, such priors can strongly influence an analysis, so using them demands a principled approach to prior specification. This approach is prior elicitation, which refers to the process of quantifying an expert’s belief into a probability distribution. Common elicitation schemes ask experts to quantify their beliefs in the form of two or more distribution summaries, such as means, modes, or percentiles, which are then converted into the standard parameters of a given family. While software exists that assist experts in the process, it is still often difficult for experts to reliably quantify distribution summaries, and eliciting a prior distribution that accurately reflects expert opinion is still a challenging process. Drawing from recent developments in the statistical graphics literature on visual hypothesis testing, we propose two graphical strategies for prior elicitation, along with web-based implementations, that function without the expert needing to explicitly quantify their beliefs. Instead, they move through a series of lineup-style tests where they are tasked with selecting datasets they believe to be most likely from a collection of candidate datasets. The algorithms then convert these specifications into a prior on the parameter of interest. The first method uses a deterministic framework, while the second is stochastic in nature and utilizes Markov Chain Monte Carlo. To illustrate the methods’ abilities to accurately quantify expert opinion, we consider the prior elicitation of a population proportion.

16. Presenter: **Dave Le Hoang** (dlh047@shsu.edu)

Affiliation: Sam Houston State University

Title: **Performance comparison study of bootstrap methods for OLS-bisector regression**

Abstract: Linear regression is commonly used to model the relationship between a dependent variable and one or more independent variables. The method that is most widely used to determine the linear regression relationship is Ordinary Least Squares (OLS). When we observe both the dependent and
independent variables as random, we may or may not assume that there are associated measurement errors. If the errors are associated with the dependent and independent variables, then OLS will underestimate the slope. In this situation, a method that biologists call Model II regression should be utilized (Isobe et al., 1990). This simulation will concentrate on the OLS-Bisector regression from the Model II regression family. The only available interval estimate for OLS-Bisector is asymptotic. Therefore, we will focus on the bootstrap performance of OLS-Bisector as an alternative estimate. Bootstrapping is a nonparametric resampling method that allows the researcher to estimate the sampling distribution and thus properties of an estimator. We employ pairwise resampling. We then compare normal, BCa, and percentile intervals to the asymptotic confidence interval. We provide a simulation study of coverage and average width of the confidence intervals to determine which bootstrapping method is a viable alternative.

17. Presenter: **Segun Ahmed (seg3k@yahoo.com)**

   **Affiliation:** Texas A&M University – Kingsville

   **Collaborator:** Stephen Sedory and Sarjinder Singh (Texas A&M University – Kingsville)

   **Title:** Simultaneous estimation of means of two sensitive quantitative variables

   **Abstract:** In this paper, we introduce a new problem of simultaneous estimation of means of two quantitative sensitive variables by using only one randomized response and another pseudo response from a respondent in a sample. The proposed estimators are extended to stratified random sampling, and the relative efficiency values are computed for equal, proportional and optimum allocation with respect to the newly introduced naive estimators.

18. Presenter: **Dilini Sandamali Katukoliha Gamage (dsk009@shsu.edu)**

   **Affiliation:** Sam Houston State University

   **Title:** Assessing the effectiveness of smoking cessation treatment using survival analysis models
Abstract: The occurrence of cigarette smoking among drug abusers is high causing serious health issues, high mortality rate, and severe addiction to drug and alcohol. In this study, we assess the effect of implementation of smoking cessation treatment and smoke-free policies in drug and alcohol rehabilitation clinics to control or cease this problem. The Data from a randomized trial of nicotine replacement therapy (NRT) plus group cognitive-behavioral counseling as compared with standard substance abuse treatment-as-usual (TAU) across multiple community-based outpatient substance abuse treatment programs within the National Drug Abuse Treatment (NIDA), Clinical Trials Network (CTN) is used to get results. 217 cigarette smokers were taken from two Outpatient drug and alcohol rehabilitation clinics and five methadone/LAAM maintenance treatment programs with the ratio of 2:1 from substance abuse TAU with Smoking Cessation (SC) treatment and substance abuse TAU respectively. Several Survival methods such as Kaplan-Meier, Cox proportional hazards, parametric, and frailty models are applied to assess the effectiveness of SC treatment.

19. Presenter: Pubudu Thilan Abeysiri Wickrama Liyanaarachchige (pta003@shsu.edu)

Affiliation: Sam Houston State University

Title: The impact of color to grayscale conversion and image resolution in face recognition

Abstract: In face recognition, with reliance on principle component analysis (PCA), often-times grayscale images are utilized because of their simplicity and achieved reduction in data load requirements especially in comparison to that of color images. This presentation compares eight different color-to-grayscale conversion algorithms and demonstrates the accuracy of each. This presentation details the specifics of an approach that uses PCA as a feature extraction method in face recognition and provides results that confirm that the PCA approach works well even for considerable reduction in image resolution which, for searches across large populations, can greatly minimize data loads and, consequently, be considerably more timely.
20. Presenter: Shalima Zalsha (sxz017@shsu.edu)

Affiliation: Sam Houston State University

Title: Confidence interval guidelines for ranged major axis regression

Abstract: Ranged major axis (RMA) is a statistical method of studying the relationship between variables measured with error. This method was introduced by Pierre Legendre as an enhancement of the major axis (MA) regression approach and is gaining popularity. One vital, but often unrealistic, assumption of MA is that the variables are expressed in the same physical units or that the error variances of the variables are approximately equal. RMA is designed to loosen those assumptions, so that it is more generally applicable. (Legendre, 2013). Sokal and Rohlf (1995) mentioned that MA will likely fail to produce the interval when either the correlation is low or the sample size is small. RMA suffers a similar fate, as mentioned by Nieuwoudt (2014). It is still unclear, however, exactly what parametric conditions cause the interval calculation to fail within MA and RMA. This research will investigate the conditions that cause the failure of RMA confidence interval calculation both analytically and with simulations, providing improved guidelines for the use of RMA.

21. Presenter: Benjamin White (bw2121@gmail.com)

Affiliation: Rice University, Houston, Texas

Collaborator: Aaron Kelly (Baylor College of Medicine), Tsz-Kwong Man (Baylor College of Medicine), Rudy Guerra (Rice University)

Title: Genomic and proteomic feature selection using Bayesian penalized regression

Abstract: One of the main goals of cancer bioinformatics is to analyze large data sets in genomic, proteomic, transcriptomic, and other macromolecular spaces to determine factors that are associated with the growth or inhibition of different types of human cancer cells. The identification of new factors not only aids in improving our understanding of cancer biology, but can also aid in decision-making regarding patient treatments (i.e. “personalized medicine”) through the detection of novel biomarkers. Osteosarcoma is a malignant bone tumor most commonly found in children and young
adults. It has high treatment resistance and low patient survival rate upon metastasis. We have implemented a probit latent variable Bayesian penalized regression model with Gibbs sampling for detecting genomic features associated with high and low metastatic risk in a set of osteosarcoma tumor cell lines. Bayesian penalized regression achieves sparsity (feature selection) through the shrinkage of weak effects to zero in the model, even for data bases which have many more features than observations ($p >> n$). Through simulation we verified our model’s efficacy for small sample sizes ($n < 10$) and applied the model to Affymetrix mRNA array and reverse-phase protein array data to detect candidate biomarkers for metastatic risk in our osteosarcoma tumor cell lines.

22. **Presenter:** Matthew Tichenor (mwtichenor@gmail.com)

**Affiliation:** Texas Tech University

**Collaborator:** Jingyong Su

**Title:** Elastic functional data analysis with applications in stock market and Alzheimer’s data

**Abstract:** This research focuses on applications for a method of registration based on the Fisher-Rao metric. First a simulation introduces function registration, then a brief review of the Fisher-Rao metric is given. It’s shown that square root velocity functions are invariant to time-warping, which forms the basis for this registration method. Two applications of registration follow. The first application is the weekly closing stock price for Urban Outfitters common stock, and registration is used in exploratory analysis. Outliers are removed from the data set and a mean price curve is found for Urban Outfitters’ common stock. The second application involves tremor curves from subjects with Alzheimer’s disease. For the tremor curves before and after registration the resulting $p$-values from functional hypothesis tests for the equality of covariance functions and equality of function means are compared and interpreted.

23. **Presenter:** Sara Biesiadny (txsarab@gmail.com)

**Affiliation:** Texas Tech University
**Title:** Modeling the relationship between MLB ballparks and home team performance using shape analysis

**Abstract:** In the last few decades, there has been a sharp rise in the amount of work done in analyzing statistics related to the sport of baseball. These studies typically use previous performance figures to predict a team or player’s production in future games. Here, we approach baseball statistics in a different manner. Interestingly, baseball is the only major American sport whose primary professional league does not standardize the dimensions of its playing arenas. This thesis plays on this idea by using various shape analysis techniques to model how home team performance and ballpark shape are related. Particularly, this piece implements the techniques of Bookstein Registration and Principal Component Analysis to investigate a possible relationship between home team performance and ballpark shape for the 2015 Major League Baseball regular season. Notably, two common performance aspects are considered: home runs and on-base percentage. The preliminary results of this study show that the ideal ballpark to maximize home team on-base percentage would have a deeper-than-average left field wall and a shallower-than-average center field wall, whereas the ideal ballpark to maximize home team home runs would have a shape very close to that of the mean Major League Baseball ballpark.

24. **Presenter:** Nhieu Bo (nhieu.bo@tamucc.edu)

**Affiliation:** Texas A&M University - Corpus Christi

**Collaborator:** Swint Friday, Lei Jin (Texas A&M University - Corpus Christi)

**Title:** The market pricing of anomalous weather: evidence from emerging markets

**Abstract:** The paper examines impacts of the weather on emerging stock markets. Research by Friday and Hoang (2015) [1] documented a significant Halloween Effect in the Vietnam market. Additional analysis indicated that the lower performance during the May through October period was correlated with abnormally high rainfall levels during their rainy season. There was a significant negative correlation between average monthly rainfall measures and mean monthly stock returns indicating that rain could be an explanation. This research is expanded to emerging markets to determine if there is a cause and effect relationship between rainfall levels and stock returns during observed periods,
which gives wealth managers and investors opportunities to hedge weather related risks with weather derivatives. Using an ARIMA and GARCH model specification and regression analysis, this paper aims to determine whether there exists any irregular component in such as weather related factors in monthly stock returns in emerging markets.

25. Presenter: Shuying Sun (s_s355@txstate.edu)
Affiliation: Texas State University
Title: A comparative analysis of statistical methods for differential methylation identification

Abstract: DNA methylation plays an important role in regulating gene expression in cancerous cells. It is important to study differential methylation (DM) patterns between two groups of samples (e.g. cancerous vs. normal individuals). With next generation sequencing (NGS) technologies and the bisulfite technique, it is now possible to identify DM patterns by considering methylation at the single CG site level in an entire genome. However, it is challenging to analyze large and complex NGS data. Even though a number of methods have been developed for DM identification, there is still not a consensus on statistical and computational approaches. In this poster, we will show the comprehensive comparative analysis results of five DM identification methods: methylKit, BSsmooth, BiSeq, HMM-DM, and HMM-Fisher. Our comparison results are summarized below. First, parameter settings may largely affect the accuracy of DM identification. Second, all five methods show more accurate results when identifying simulated DM regions that are long and have small within-group variation, but they have low concordance, probably due to the different approaches they have used for DM identification. Third, HMM-DM and HMM-Fisher yield relatively higher sensitivity and lower false positive rates than others, especially in DM regions with large variation. With the comprehensive comparative analysis, we aim to share new perspectives about the challenges and opportunities of DM identification, which may pave the way for novel and better methodology development in the near future.

26. Presenter: Oluwaseun Lekan Olanipekun (seunolanipekun88@gmail.com)
Affiliation: Texas A&M University – Kingsville
Collaborator: Rajan Lamichhane

Title: Impact of local economic conditions in sleep habits: understanding of the pathways linking local economic conditions to food insecurity and sleep habits

Abstract: Structure Equation Modelling (SEM) is one of the least used techniques in Statistics but it is widely used in social and behavior sciences, business, education, and other fields. Structure Equation Modeling (SEM) was implemented to understand and validate the pathways between food insecurity and sleep habits. Food insecurity and sleep habits are directly affected by household income. Low income populations are vulnerable to both food insecurity and inadequate sleep. US job market and economic conditions had undergone significant changes during recession period (2007-2009), resulting in significant change in food security status and sleep habits. We focus in understanding the three-way relationship between local economic conditions, food insecurity, and sleep habits. Other statistical analyses such as simple linear regression, multivariate regression, and logistic regression were implemented by incorporating the complex survey design in order to generalize the result. We investigated the relationship between food security status and sleep duration/sleep considering the indirect effect of stress and compare the relative change due to economic conditions. Factors such as physical activity, body mass index (BMI), smoking behavior, socio economic status (SES), diet, and other demographic variables such as age, sex, and educational background were also considered in the analyses.

27. Presenter: Jasdeep Pannu (jpannu@lamar.edu)
Affiliation: Lamar University
Collaborator: Nedret Billor (Auburn University)

Title: Robust weighted LAD-group-LASSO for functional regression model

Abstract: Functional data has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities, in the last two decades. Data recorded densely over time is called functional data. Variable selection is an important problem in functional regression analysis, just as in ordinary regression analysis. In this study, our focus is on a functional linear regression
model with functional predictors and scalar responses where functional coefficients of functional predictors are to be estimated. Since a functional dataset has issues of high dimensionality and multicollinearity inherently, estimation of functional parameters is complicated. Further, the presence of outlying curves adds extra complication to the estimation of the functional regression coefficients. The objective of this research is to develop a robust functional variable selection method based on $L_1$ regularization. Recently, a method called functional LAD-group LASSO (LAD-gLASSO) has been developed which uses the combination of the well-known robust loss function least absolute deviation (LAD) and penalty function group LASSO, where the functional parameters are estimated and selected through the minimization of the sum of the absolute value of the errors and penalizing the parameter functions. However, it is well-known that the LAD based method is only resistant to the outlier in the response variable, but not resistant to the outliers in the explanatory variables (leverage points). To tackle this problem, we propose a new method called functional Weighted LAD-group LASSO (WLAD-gLASSO). This method is not only resistant to outliers in the response variable but also minimizes the effect of the leverage points by introducing weights which are only dependent on the functional predictors. In order to show the performance of the proposed method, we conduct an extensive simulation and use a real dataset.

28. Presenter: **Indika Wickramasinghe** (iprathnathungalage@pvamu.edu)

**Affiliation:** Prairie View A&M University

**Collaborator:** Harsha K. Kalutarage, Dumindu Samanwickrama

**Title:** **Comparison of methods to identify malicious android apps**

**Abstract:** We compare three machine learning techniques to classify whether an Android app is benign or malicious, based on its communication features. In this study, Principal Component Analysis (PCA), Support Vector Machine (SVM) and a mixture of PCA-SVM techniques are utilized, and different performance measures such as sensitivity, specificity and precision are computed for the comparison. Finally, the behavior of above measurements is observed by changing the parameters and the type of kernel of the SVM. Through experimental validation, we show that the mixture of PCA-SVM outperforms other two classifiers.
29. Presenter: **Sonali Saxena (s.saxena@analyticfocus.com)**

**Title:** The application of statistics in predictive coding

**Abstract:** Researchers have been studying ways to automate document or text classification for decades. With the abundance of text-based materials being converted to electronic formats for today’s ease of use, the need for automatic classification methods is greater than ever. Innovative advancements are now coming to light by combining search technologies, machine learning, and statistical methodologies to create best practices for any categorization needs. This poster outlines five statistical methodologies i.e. Bayesian, Regression, K-Nearest Neighbor, Decision Tree and Data/Text Mining Hybrid Consulting, that when appropriately applied to text classification, will provide researchers the most efficient methods of sorting when needed. Predictive Coding, or Computer Assisted Review, is a tool that combines human guidance with computer-piloted concept/keyword searching in order to “train” document review software to recognize relevant documents within a document universe. This technique is revolutionizing information extraction as users seek to spend less time wading through vast amounts of data. Accuracy and efficiency are also improved for those who understand the hidden statistics inside the “black box” of complex software routines. The analytical methodologies used in predictive coding bring value to those who handle a myriad of information and identify the relevant items deserving of users’ time and priority.

30. Presenter: **Gabriel J. Odom (gabriel_odom@baylor.edu)**

**Affiliation:** Baylor University

**Collaborator:** Dean Young

**Title:** Reducing misclassification rates via iterative dimension reduction

**Abstract:** In poorly-posed classification problems, class covariance estimates are often highly unstable. For k classes in p dimensions, full location and scale parameter estimation requires k(p^2 + 3p)/2 distinct observations. Because of this unrealistically high minimum-instance threshold, classifiers often have
structural inaccuracies and high instability. To counter this, we propose performing preliminary linear dimension reduction via orthogonal subspace projection, then recalculating the location and scale parameters in the new, smaller dimension. Linear reductions in dimensionality result in quadratic reductions in observation cost, causing increased stability of parameter estimates and more stable and accurate classifiers. Classifiers built in parameter subspaces are more stable, and yield smaller misclassification errors. We demonstrate this fact by simulation study over a suite of parameter spaces.

31. Presenter: Mohan D. Pant (mpant@uta.edu)
Affiliation: University of Texas at Arlington
Title: Simulating non-normal distributions using Fleishman’s third-order polynomials based on t distribution
Abstract: This paper introduces a new family of non-normal distributions using a doubling technique (Morgenthaler & Tukey, 2000) on Fleishman’s (1978) third-order polynomials based on t-distribution. The parametric forms of probability density function (p.d.f.) and cumulative distribution function (c.d.f.) for this new family of distributions are derived. Also derived is a system of equations for mean, standard deviation, skew, and kurtosis for this new family of distributions. A methodology based on method of matching moments is described to solve for the values of scale and shape parameters for specified values of standard deviation, skew, and kurtosis. It is also demonstrated how to compute other measures of central tendency such as the mode, median, and trimmed mean. Subsequently, a methodology is developed to simulate multivariate non-normal distributions with specified values of cross-product moments (e.g., skew, kurtosis and Pearson’s correlation). Examples are provided to demonstrate how this family can be used in the context of distribution fitting using real data sets. The results of a Monte Carlo simulation are provided to confirm that the proposed method generates distributions with user specified values of skew, kurtosis, and intercorrelation matrix.

32. Presenter: Leif Ellingson (leif.ellingson@ttu.edu)
Affiliation: Texas Tech University
Collaborator: G.A.I.C. Premarathna (Texas Tech University)

Title: Classification of protein binding ligands using their structural dispersion

Abstract: It is known that a protein’s biological function is in some way related to its physical structure. Many researchers have studied this relationship both for the entire backbone structures of proteins as well as their binding sites, which are where binding activity occurs. However, despite this research, it remains an open challenge to predict a protein’s function from its structure. Main purpose of this research is to gain a better understanding of how structure relates to binding activity and to classify proteins according to function via structural information. We approach the problem from the dataset compiled by Kahraman et al. (2007) and extended Kahraman dataset. There we compared the performance of model based technique using the distance to principal axes with alignment-based techniques.

33. Presenter: Joshua Mayer (joshua.mayer@ttu.edu)

Affiliation: Texas Tech University

Title: Multivariate random forests using copulas

Abstract: In recent years, ensemble based approaches such as Random forests have been shown to perform well in modeling drug sensitivities. However, the Random Forest generates a predictive model for each drug individually, and ignores the relationship between different drug sensitivities in the creation of the model. To utilize the dependence in sensitivities across different drugs, a multivariate framework is more appropriate. We proceed with a tree based model where the covariates to be split are chosen via a conditional inference frame-work. Splits are chosen by maximizing the node costs that are decomposed into an L1 copula distance that preserves the dependence structure, and a sum of squares distance. From the conditional inference structure, we use the distribution of P-values to develop a new measure of variable importance. We then assess our model and the variable importance measure using a simulation study. Furthermore, we use a real dataset to compare the prediction accuracies obtained from our methodology to the traditional multivariate random forest.

34. Presenter: Dissanayaka M. Pansujee Dissanayaka (pansujee.dissanayaka@ttu.edu)
Affiliation: Texas Tech University

Collaborator: Jingyong Su, Bijoy Ghosh

Title: Fitting smoothing splines to eye movement trajectories

Abstract: Eye movement refers to the voluntary or involuntary movement of the eyes. Eye movements are an important behavioral resource to understand human development and learning. Recent research has revealed that variations of these trajectories reflect the dynamics of the oculomotor system. Here we address the problem of estimating a smoothed curve to an eye movement trajectory given by a set of time-indexed points on the sphere. So that the estimated curve would approximate the data points best at given time instants while being regular as possible. This optimization problem is carried according to the method discussed by J. Su, I.L. Dryden, E. Klassen, H.Le and A. Srivastava in 2011. Furthermore, the estimated curves as a set of time indexed points are used in interpolation and classification. Two data sets obtained by two experiments are used in the study. Eye movement data is represented as points on the unit sphere as described by Ghosh, B.K; Wijayasinghe, I.B. and Kahagalage, S.D. in 2014. The data sets are divided into segments prior to smoothing using attributes such as velocity and location data.

35. Presenter: Zakry Zapata (doublez711@gmail.com)

Affiliation: Texas A&M - Kingsville

Title: Multiple mean hypothesis testing - parachute example
Abstract: The purpose of this experiment is to give an educational example where a one-way ANOVA can be used. The one-way ANOVA is an extension of a two-sample t-test, where we can compare the means of independent populations by testing the null hypothesis that all means are equal. In our experiment we will test the fall-rate of a light weight toy soldier, with parachutes created from common materials including a trash bag, cardboard, construction paper, cloth, and paper. The figure will be dropped from the top of a ladder ten times for each parachute type, and the hypothesis that all means are equal will be tested at 0.05 level of significance. To perform the ANOVA, the SAS software suite will be used.

36. Presenter: Niharika Yennum
   (niharika.hcu@gmail.com)

   Affiliation: Texas A&M University – Kingsville

   Title: Bayesian analytics

   Abstract: In my work, I would like to use the borehole temperature data for past 30 years over Indian shield that is already available, to reconstruct the history of subsurface temperature profile using Bayesian inference Inversion scheme. This technique uses Bayes’ theorem to map a priori probability distribution function, via the data likelihood into a posterior probability distribution function. Rather than trying to find the best parameter i.e., m*, it asks for a probability density σ(m) for the parameters ‘m’ that updates the prior information we have on m, which is also given as a probability density, in light of the actual measurements that we will denote by d. This probability density function (p.d.f.) is computed taking into account the mathematical model, information about the measurement process, and the measured data, as well as any prior knowledge we may have about the parameters we seek; it is consequently a posterior probability density as we can only determine it after making actual measurements; it is conditioned on d.

37. Presenter: Chalani Prematilake (chalani.prematilake@ttu.edu)

   Affiliation: Texas Tech University
Collaborator: Leif Ellingson (Texas Tech University)

Title: Prediction and comparison of approximations of planar contours for two parameterizations

Abstract: As a result of rapid advancements of technology, high dimensional data can be easily found almost everywhere. These types of data arise in medical imaging as well as in computer vision can be considered as data on manifolds and studied using methods of Statistical Shape analysis. In statistical shape analysis, each observation is considered as an object in the corresponding ambient space. These observations are represented by matrices of dimension where \( m \) is the dimension of the ambient space and \( k \) is the number of points chosen to represent the object. Even though high dimensional data increase the accuracy of the results, it is computationally expensive and time consuming. Closer approximations can be made much faster, using comparably smaller number of carefully chosen sampling points. Shape of a configuration that sometimes considered as the outline of the image is mostly a contour in 2D plane? In this study the shape space of discretized observation is considered as the Kendall’s Shape Space’s Polygon approximation (k-gon), which reduces the error due to discretization of digital images was used over the traditional landmark selection methods (k-ad). This study reveals two decision criteria for obtaining a lower bound for number of sampling points needed to reach at a predefined error threshold. Two different parameterizations are used for choosing the sampling points. Models are then developed for predicting these bounds based on simple characteristics of the curves. Choosing sampling points to reduce dimensionality helps to keep the computational cost low.

38. Presenter: Qiwei Li (ql6@rice.edu)

Affiliation: Rice University

Collaborator: Marina Vannucci, Michele Giundani

Title: Bayesian models of high-dimensional count data

Abstract: Modern big data analytics often involves high-dimensional count data sets such as bag-of-
words data, ecological survey data, and biological sequence data. A main goal in the analysis of such data is to uncover the group structure in the samples as well as identify the discriminating features. We propose a Bayesian nonparametric hierarchical Poisson mixture model that accounts for the overdispersion observed across samples as well as across multiple features. Our model formulation incorporates a feature section mechanism and prior distributions that appropriately account for identifiability constraints on the model parameters. Our strategy for posterior inference results in a unified approach to achieve the goal. We demonstrate the performance of our method on simulated data and present applications to document clustering, based on a bag-of-words benchmark dataset.

39. Presenter: Erik John Harwell
(erikharwell1980@yahoo.com)
Affiliation: Texas A&M University - Kingsville
Collaborator: Reza Ahangar

Title: Complex matter space with an introduction to hyperbolic geometry
Abstract: In this poster we will introduce a new concept we call Complex Matter Space (CMS), and its postulates. Complex Matter particles and CMS are new fundamental views of matter that we will present here and the paradigm will be shifted from pure real or pure imaginary particles to Complex Matter particles. Initially, we will assume that matter has two intrinsic components: mass and charge and will be denoted by \( M = m + iq \), where \( i = \sqrt{-1} \). Complex numbers and Hyperbolic Geometry are very important to visualizing and analyzing CMS. We will thus introduce complex numbers, the complex plane, vectors, and the basics of Hyperbolic Geometry. We will look at momentum and energy in CMS, the Quantum Mechanical view of CMS, the Relativistic view of CMS, and the derivation of the Einstein Equation in CMS.