2020 Florida Chapter Annual Meeting

American Statistical Association

March 6-7, 2020

Talks and Poster Abstracts

Welcome **University of West Florida, March 6-7, 2020**

We are pleased to welcome you to the City of Pensacola, the University of West Florida, and the 2020 ASA FL Chapter Annual Meeting.

We are very pleased and honored to have a great panel of speakers and attendees this year. We would like to thank our Keynote, Invited, and Workshop speakers for agreeing to take time out of their busy schedules to give us their perspectives on a broad-ranging set of topics.

We would like to thank the UWF Academic Affairs, Hal Marcus College of Science and Engineering, the Department of Mathematics and Statistics, Office of Undergraduate Research, Usha Kundu - MD College of Health, and Florida Chapter of the ASA for sponsoring this event. Finally, kudos to the members of the Organizing Committee for making this event a Success!

Pensacola is the site of the first Spanish settlement within the borders of the continental United States in 1559, predating the establishment of St. Augustine by 6 years, although the settlement was abandoned due to a hurricane and not re-established until 1698. Pensacola is a seaport on Pensacola Bay, which is protected by the barrier island of Santa Rosa and connects to the Gulf of Mexico. A large United States Naval Air Station, the first in the United States, is located southwest of Pensacola near Warrington; it is the base of the Blue Angels flight demonstration team and the National Naval Aviation Museum. The main campus of the University of West Florida is situated north of the city center. It is nicknamed "The City of Five Flags", due to the five governments that have ruled it during its history: the flags of Spain (Castile), France, Great Britain, the United States of America, and the Confederate States of America. Other nicknames include "World's Whitest Beaches" (due to the white sand of Florida panhandle beaches), "Cradle of Naval Aviation", "Western Gate to the Sunshine State", "America's First Settlement", "Emerald Coast", "Red Snapper Capital of the World", and "P-Cola".

We hope you will all have enough time and opportunity during your stay here to sample the fantastic things that Pensacola and the Northwest Florida to offer.

Once again, welcome to UWF - Pensacola!

The Organizing Committee
The 2020 ASA Florida Chapter Meeting

The Statistical Assessment of Bioequivalence and Biosimilarity

Workshop

Thomas Mathew, UMBC Presidential Research Professor
mathew@umbc.edu

The topic of bioequivalence deals with procedures for testing the equivalence of two drug products: typically, a generic drug and a brand name drug. Biological availability or bioavailability of a drug is the rate and extent to which the active drug ingredient is absorbed into the blood, and becomes available at the site of drug action. Two drug products are bioequivalent if they have similar rate and extent of absorption into the blood. In the workshop, the bioequivalence problem will be introduced using examples, and its historical development will be described. The data for bioequivalence assessment are generated using cross-over designs, and the data are obtained on three variables: Area under the time-concentration curve (AUC), the maximum blood concentration (Cmax) of the active ingredient, and the time to reach the maximum concentration (Tmax). Univariate bioequivalence consists of the separate modeling and analysis of the data on each of these variables. Multivariate bioequivalence consists of the joint modeling and analysis of the data on the three variables. Various statistical criteria used for bioequivalence assessment will be explained, and the statistical methodology for testing the corresponding hypotheses will be addressed. A well known and widely used procedure for testing average bioequivalence (i.e., equivalence of population means) is the two one-sided test or TOST. The emerging area of equivalence testing in the context of biosimilars will be addressed. Some recent developments in the area of equivalence testing will also be mentioned

---

2020 Census, Lagrange's Identity, and Apportionment of the U.S. House of Representatives

Keynote Speaker

Tommy Wright, US Census Bureau
Tommy.Wright@census.gov

Given the impracticality of a pure democracy, the U.S. Constitution (1787) calls for a representative form of democracy where the people elect persons to represent them for governing. Each state gets a number of representatives in the U.S. House of Representatives "...according to their respective numbers..." as recorded in a census of the nation to be conducted every ten years starting in 1790. We make use of an elementary result known as Lagrange's Identity to provide a bridge between an insightful motivation and an elementary derivation of the method of equal proportions. The method of equal proportions is the current method for apportioning the 435 seats in the U.S. House of Representatives among the 50 states, following each decennial census. We highlight why the numbers from the census matter and affect our condition and behavior. We also present some historical comments about the first two methods of apportionment.

TreeScan Datamining for Drug and Vaccine Safety Surveillance

Keynote Speaker

Martin Kulldorff, Harvard Medical School and Brigham and Women's Hospital
mkulldorff@bwh.harvard.edu

Post-market drug and vaccine safety surveillance is important to detect serious adverse events too rare to find in clinical trials. Data mining methods are useful to screen many diagnostic codes for unexpected safety issues in the absence of specific safety concerns. Adjusting for multiple testing, the tree-based scan statistic (TreeScan) evaluates thousands of potential adverse reactions and groups of related reactions. Using electronic data from health insurance companies, it has been used to evaluate the safety of both vaccines and pharmaceutical drugs, using a variety of statistical designs, including Poisson models, binomial models and self-control designs. Several examples are provided.

A Biostatistical Potpourri

Keynote Speaker

Martin Kulldorff, Harvard Medical School and Brigham and Women's Hospital
mkulldorff@bwh.harvard.edu

It is truly fascinating to be a statistician. One encounters many interesting problems in a wide variety of scientific fields, requiring different statistical methods and models. In this talk, we discuss how to make statistically independent analyses from the same data; how to dissect meat from different animals; how to safely return from Las Vegas without upsetting your spouse; how to suck-up to sharks; and how to make people upset at their neighbors.

Statistical Analysis of Noise Multiplied Data Using Multiple Imputation

Keynote Speaker

Bimal Sinha, UMBC
sinha@umbc.edu

When statistical agencies release data to the public, a major concern is the control of disclosure risk, while ensuring utility in the released data. SDC methodology aims to suitably modify a dataset prior to its release - there are several ways this can be achieved. We present here a data perturbation method based on noise multiplication and discuss statistical analysis of the noise multiplied data. This is based on the notion of multiple imputation.

Monitoring and Improving Surgical Quality

Invited Speaker

William H. Woodall, Virginia Tech
bwoodall@vt.edu

Some statistical issues related to the monitoring of surgical quality will be reviewed in this presentation. The important role of risk-adjustment in healthcare, used to account for variations in the condition of patients, will be described. Some of the methods for monitoring quality over time will be outlined and illustrated with examples. The National Surgical Quality Improvement Program (NSQIP) will be described, along with a case study demonstrating significant improvements in surgical infection rates and mortality.

Bayesian inference in high-dimensional vector autoregressive models

Invited Speaker

Kshitij Khare, University of Florida
kdkhare@stat.ufl.edu

Vector autoregressive (VAR) models aim to capture linear temporal interdependencies among multiple time series. They have been widely used in macroeconomics and financial econometrics and more recently have found novel applications in functional genomics and neuroscience. These applications have also accentuated the need to investigate the behavior of the VAR model in a high dimensional regime, which will provide novel insights into the role of temporal dependence for regularized estimates of the model parameters. However, hardly anything is known regarding posterior consistency for Bayesian VAR models in such regimes. In this talk, we will consider methods using hierarchical matrix normal priors on transition matrices as well as methods inducing exact sparsity in the transition matrices. We will establish posterior consistency under high-dimensional scaling where the dimension p of the VAR system grows with the sample size n. We will also present novel methodology for the related problem of mixed frequency regression.

Laplace Deconvolution with dependent errors: an Application to the Analysis of Dynamic Contrast-Enhanced imaging

Contributed

Rida Benhaddou, Ohio University
benhaddo@ohio.edu

We investigate the problem of estimating a function f based on observations from its noisy convolution when the noise exhibits long-range dependence (LRD). We consider both Gaussian and sub-Gaussian errors. We construct an adaptive estimator based on the kernel method, with the optimal selection of the bandwidths performed via Lepski's Method. We derive a minimax lower bound for the -risk when f belongs to a Sobolev ball and show that such estimator attains optimal or near-optimal rates that deteriorate as the LRD worsens. We carry out a limited simulations study which confirms our conclusions from theoretical results

---

A Nonparametric CUSUM Chart for Multiple Stream Processes Based on the Extend Median Test

Contributed

Austin Brown, Kennesaw State University
abrow708@kennesaw.edu

In statistical process control applications, situations may arise in which several presumably identical processes or "streams" are desired to be simultaneously monitored. Such a monitoring scenario is commonly referred to as a "Multiple Stream Process (MSP)." Traditional MSP charting techniques rely on the assumption of normality, which may or may not be met in practice. Thus, a cumulative summation nonparametric MSP control charting technique, based on a modification of the classical extended median test was developed and is referred to as the "Nonparametric Extended Median Test Cumulative Summation (NEMT-CUSUM) chart." Chart development, including calculation of control limits, is given. Through simulation, the NEMT-CUSUM is shown to perform consistently in the presence of normal and non-normal data. Moreover, it is shown to perform more optimally than parametric alternatives in certain circumstances. Results suggest the NEMT-CUSUM may be an attractive alternative to existing parametric MSP monitoring techniques in the case when distributional assumptions about the underlying monitored process cannot reasonably be made.

Interpretation of Regression Models: Using Word Clouds for Visualization

Contributed

Heather Rollins, University of West Florida
hlr10@students.uwf.edu

Have you ever just stared at all the numbers that your beautiful regression method gave you, and wondered what they mean? We are familiar with word clouds as a visualization, showing us the relative frequency of word use in a text. This new application for word clouds gives us a visualization to aid in interpreting a regression model, showing intuitively the intersection of the qualitative and quantitative aspects of our work.

A Statistical Analysis of Coastal Beach Morphology by Various Citizen Scientists

Contributed

Tyler Watson, University of West Florida
tjw38@students.uwf.edu

This project assesses the accuracy of an inexpensive beach surveying tool as compared to expensive professional electronic survey tools including an engineering level, total station, and RTK-GPS. The surveying tool under consideration is comprised of a wooden A-frame with a digital level attached to its crossbar. If shown to be accurate, this inexpensive and simple tool will allow citizen scientists to easily participate in beach profiling. To test the tool, we engaged various citizen science user groups (high school students, math/stat college students, environmental science college students, and local adult residents) in conducting beach profile surveys. A statistical model was constructed to assess the difference between the proposed tool and professional tools. Overall, our study shows that the inexpensive A-frame device is an acceptable tool to produce accurate beach profile data. A-frames that were previously constructed used gray levels and newly constructed A-frames used orange levels. We found that A-frames with the gray levels produced more accurate results compared to the orange levels. Math/Stat and EES students produced accurate results compared to the professional tools while the high school students did not. We hypothesize that the increased error by the high school group was due to the low number of participants (n=2) and their use of the less accurate orange level, not ability due to age/education. Finally, there was statistical significance between group category (p= 0.0320) and A-frame numbers (p= 0.0319) but not between levels(p= 0.1017).

Likelihood-based finite-sample inference for synthetic data from the Pareto model

Contributed

Sandip Barui, University of South Alabama
baruis@southalabama.edu

Statistical agencies often publish microdata or synthetic data to protect the confidentiality of survey respondents. This is most prevalent in the case of income data. In this paper, we develop a likelihood-based finite sample inferential methods for a singly imputed synthetic data using plug-in sampling and posterior predictive sampling techniques under Pareto distribution, a well-known income distribution. The estimators are constructed based on sufficient statistics and the estimation methods possess desirable properties. For example, the estimators are unbiased and confidence intervals developed are exact. An extensive simulation study is carried out to analyze the performance of the proposed methods.

Generalization of Sample Size and Power Computations Methods for Two-Stage Randomized Trial

Contributed

Rouba Chahine, University of Alabama at Birmingham
chahine@uab.edu

Background: A standard component of informed consent in randomized clinical trials is participants' awareness of all procedures or treatments that might benefit them. This knowledge may decrease compliance or reduce tolerance for inconveniences or difficulties when participants receive non-preferred treatments. A two-stage randomized design (two-stage RCT) can be used to incorporate participants' preference. The first stage randomizes participants to either a random group or a choice group. Next, participants in the random group are randomized to one of two treatments, while participants in the choice group can choose between treatments. Motivation Methods: Sample size computation is an essential part of clinical trials. An inadequate sample size can lead to an underpowered study that fails to detect clinically relevant effects, while an unnecessarily large sample size can waste resources. Power analysis methods to calculate appropriate sample size to estimate treatment, selection, and preference effects based on the ANOVA approach have been proposed for Normal and Binary outcomes. In our work, we show that this approach can be generalized for any outcome whose distribution satisfy the regulatory conditions for the central limit theorem. We also use simulations to evaluate the performance of these methods for time-to-event outcomes following Exponential distributions for complete-case and right-truncated data, and Weibull distributions for the increased and decreased hazard cases. Simulation Findings: Type I error results indicate that the ANOVA approach performs well with non-censored and right-truncated Exponential data. Type I error for the Weibull model with a decreased hazard was conservative particularly with balanced treatment, while the increased hazard model performed well in most cases. Power analysis for truncated Exponential and Weibull with decreased hazard data were underpowered, implying that the sample size was underestimated. Weibull with increased hazard was overpowered, while the non-censored Exponential performed fairly well. Conclusions: In current medical practices, patient's treatment choice is taken into consideration as it affects patient's compliance. New clinical trials designs are needed to account for these factors as well as the development of new statistical methods for these designs. Our work showed that the current ANOVA based method for two-stage RCT can be generalized to any distribution that satisfy the regulatory conditions, and that it performs well for non-censored exponential model, but it has several limitations for other type of time-to-event models and do not allow for the inclusion of censoring.

Finite Sample Properties of an Exponential-Compound Symmetric Covariance Structure

Contributed

Amber K. Weydert, University of West Florida
akw27@students.uwf.edu

This project is a simulation study of model misspecification when analyzing cardiovascular MRI data observed post myocardial infarction. A covariance structure was designed specifically for left ventricular (LV) data based on clinical observations. Using the American Heart Association's segmentation model, LV data can be segmented into three levels, based on the location of the segments. Observations from the same level were assumed to have correlation depending on the distance between segments while observations from different levels share a common covariance. Rotation of the LV was simulated using the multivariate normal; a fixed-effects model was constructed to compare the rotation of diabetics and non-diabetics after adjusting for the level of LV. We modeled the simulated data using a simple fixed-effects model specifying the LV level and diabetic status as predictors. We examined bias, relative efficiency, power, and choice of working covariance structure via fit indices. Briefly, bias is close to zero for all structures, even when the covariance structure is misspecified and relative efficiency showed that our proposed structure resulted in the smallest standard error in most cases. However, type I error and power are inflated for the exponential and unstructured working structures, thus should not be specified when analyzing data of this type. When examining model fit indices, the proposed structure was chosen 99.90% of the time by both the AIC and BIC while the other structures where chosen less than 1% of the time. Specifying the proposed structure as the working structure, unsurprisingly, give the best results in terms of type I error and power and is chosen as the best fit by fit indices.

Multivariate Outlier Screening for First-order Assessment of Flight Test Data.

Poster

James Bryan Mackey, US Air Force and (ASA Member)
james.mackey.14@us.af.mil

Introduction/Background: Developmental flight testing follows a strict adherence to approved test plans that seek a balance between safe conduct of testing and gathering of appropriate data to answer test objectives. Given costs and time constraints, test data review timeliness is pivotal in successfully maintaining a test schedule. Consistency within measures during test point execution enables success; recognizing the need to re-accomplish a test point facilitates test scheduling. Objectives: Provide flight test data analysts with a flexible tool to quickly highlight potential data outliers using a combination of dimensional scaling and clustering. Provide graphical and quantitative means to recognize test measurement validity. Background: Typical flight testing for a test platform produces well over 1,000 engineering unit measurements with 20 to 50 typically being relevant to specific measures of performance. The ability to quickly screen these measures without overloading

an analyst is challenging. Using an adaptation of an approach proposed by Leland Wilkinson (Wilkinson, 2017) for large data sets, the Matlab®-based Multivariate Outlier Screening (MVOS) procedure enables the analyst to quickly screen data parameters for a series of test points and identify potential outlier observations within the relevant time of the test point. Potential outliers are identified via timestamp, thereby enabling the data analysts to quickly focus their attention on the appropriate portion of the data set and identify which parameters are suspect. Early identification of invalid test points promotes test efficiency, ultimately saving time and subsequently expediting delivery to the end user. Description: The MVOS procedure is applied to a collection data grouped by test objective, for instance level flight performance. Data parameters are generated at a consistent sample rate (nominally 20Hz). The MVOS procedure expedites test data review by highlighting potential outlier data points. Identification of Outlier Candidates consists of: construction of a unit normalized data matrix, calculating Mahalanobis distances of each observation from the matrix mean, clustering the data into discrete collections, fitting the Mahalanobis distances to an exponential distribution, identifying any cluster that exceeds an established probability threshold, and evaluating the original measurements graphically using the timestamp where the outlier cluster occurred. Conclusion: Application of a first-order screening procedure that simultaneously evaluates 20 to 50 test parameters for a given test point has enabled a 30% reduction in analyst efforts in identifying questionable data and enhanced visualization of data through clustering. Better visualization and data identification in concert with enhanced communication between the analysts and engineering teams, will ultimately lead to more efficient use of test resources and schedules. Wilkinson, L. (2017). Visualizing Big Data Outliers through Distributed Aggregation. IEEE Transactions on Visualization and Computer Graphics (Volume: 24 , Issue: 1 , Jan. 2018), 256 - 266.

A Bayesian approach to assessing publication bias in meta-analysis of a binary outcome with controlled false positive rate

Contributed

Linyu Shi, Florida State University

ls16d@my.fsu.edu

Publication bias (PB) is a major threat to the validity of meta-analysis. Egger's regression test is the most widely-used tool to assess PB. It can be easily implemented and generally has satisfactory statistical power. It examines the association between the observed effect sizes and their sample standard errors (SEs) among the collected studies; a strong association indicates the presence of PB. However, Egger's regression may have a seriously inflated false positive rate caused by this association even when no PB appears, particularly in meta-analysis of a binary outcome. Although various alternative methods are available to deal with this problem, they are powerful in specific cases and are less intuitive than Egger's regression. This article proposes a new approach to assessing PB in meta-analyses of ORs via Bayesian hierarchical models. It reduces false positive rates by using the latent "true" SEs, rather

than the observed sample SEs, to perform Egger-type regression; those "true" SEs can be feasibly estimated with Markov chain Monte Carlo algorithm. We present extensive simulations and three case studies to illustrate the performance of the proposed method.

The Impact of Undergraduate Research Programs on Scientific Outcomes

Poster

Mariam Khachatryan, Auburn University
mzk0070@auburn.edu

Over the past decade, there has been significant concern addressing the current shortage of PhDs in Science, Technology, Engineering, and Mathematics (STEM) disciplines. In an effort to increase the number and quality of STEM PhDs, the National Science Foundation (NSF) funds a large number of sites to provide research experience for undergraduates - REU sites. To assess the impact of such REU programs a study was conducted (Wilson, et al., 2018, p. 529) where the scientific outcomes (presentations, publications and scientific awards) of 88 participants and applicants were compared. And it was found that REU participants have better scientific outcomes in terms of presentations, publications, and scientific awards. However, the article states that selection process may bias the finding in favor of students who are more likely to be successful in science which may explain their finding. To remove the selection bias, propensity score matched analysis is conducted in this project and Friedman test p-values comparing the scientific outcomes of REU participants to non-participants using newly matched pairs are 0.0000721315, 0.2040239, 0.00192954 for presentations, publications and scientific awards, respectively. Hence, we have a compelling evidence to assert the positive impact of REU programs on presentations and awards but not for publications which is different from the original findings of the above-mentioned paper.

Fragility index of network meta-analysis with application to smoking cessation data

Contributed

Aiwen Xing, Florida State University
ax17@my.fsu.edu

Objectives: Network meta-analysis (NMA) is frequently used to synthesize evidence for multiple treatment comparisons; however, some comparisons' statistical significance can be influenced by changing the event status of a few subjects. The fragility index (FI) is recently proposed to assess the robustness (or fragility) of results from clinical studies and from pairwise meta-analyses. We extend the FI to NMAs with binary outcomes. Methods: We define the FI for each treatment comparison in NMAs. It quantifies the minimal number of events necessary to be modified for altering the comparison's statistical significance. We introduce an algorithm to derive the FI and methods to visualize the process. A worked example of

smoking cessation data is used to illustrate the derivation and interpretation of the FI. Results: Some treatment comparisons had small FIs; their significance (or non-significance) could be altered by modifying a few events' status. They were possibly related to various factors, such as the P-values, event counts and sample sizes, etc., in the original NMA. After modifying events' status, treatment ranking measures also changed to different extents. Conclusion: Many NMAs include insufficiently-compared treatments, small event counts, or small sample sizes; their results are potentially fragile. The FI offers a useful tool to evaluate treatment comparisons' robustness and reliability.

A Review of Equivalence and Noninferiority Tests for Longitudinal Designs in Clinical Trials

Poster

Joseph Ficek, Yuanyuan Lu, Yangxin Huang, Henian Chen
College of Public Health, University of South Florida
jficek@usf.edu

Rather than establishing superiority of one treatment over another, many clinical trials seek to establish equivalence or noninferiority in their effects. Standard statistical approaches, e.g., the two one-sided test (TOST), assess equivalence or noninferiority at a single point of comparison. Questions remain, however, as to how to characterize equivalence/noninferiority in longitudinal designs, where study endpoints are measured repeatedly over time. A review of the literature was conducted to assess relevant statistical techniques. For continuous outcomes, one suggestion is to compare time-averaged statistics from linear mixed models that account for the within-subject covariance. A suggestion for functional data is to construct bootstrap confidence intervals for pointwise treatment differences and assess whether all intervals lie within a specified equivalence band. No literature was found concerning equivalence/noninferiority tests for repeated binary measurements collected at more than two time points. Given the paucity of literature in this area, further research is warranted.

Modern Principal Component Analysis for Genomic Data with pathway PCA

Poster

Gabriel J. Odom, James Ban, Lizhong Liu, Lily Wang, and Steven Chen, Florida International University
gabriel.odom@fiu.edu

To process the petabytes of clinically-relevant genomic data generated every year, medical researchers need sophisticated statistical methods and matching computational tools. This genomic data includes copy-number aberration, site methylation, gene expression, protein abundance, and many other levels of measurement. Unfortunately, the advancement of statistical methods to match the complexity and size of this data has fallen behind. One of the most common challenges facing statistical methods applied to genomic data is the curse of dimensionality: the number

of features, p, is vastly greater than the number of samples, n. However, while the quantitative samples may be scarce, pathway analysis seeks to alleviate the curse of dimensionality via incorporating additional qualitative data, grouping features by related biological function. This additional layer of information enables estimation of pathway-specific principal components with reduced variance. However, most currently available pathway analysis tools neither provide subject-specific estimates of pathway activities (important for precision medicine), nor data analysis results across multiple levels of measurement—for example concurrent DNA methylation and gene expression. To address these challenges, we present the open-source software tool pathwayPCA, an R/Bioconductor package for pathway analysis that utilizes modern statistical methodology, including Supervised Principal Component Analysis and Adaptive, Elastic-net, Sparse Principal Component Analysis. Further, pathwayPCA can be used to predict continuous, binary, or survival outcomes from design matrices incorporating genomic measurements with multiple covariate and/or interaction effects. We apply pathwayPCA to detect pathway effects and to predict clinical responses in human kidney, colon, and ovarian cancers.

A Comparison of Neural Decoding Methods and Population Coding Across Thalamo-Cortical Head Direction Cells

Poster

Zishen Xu, Florida State University
zx16@my.fsu.edu

Animals can navigate by monitoring an online record of their spatial orientation in an environment and using this information to produce direct trajectories to hidden goals. Head direction (HD) cells, which fire action potentials whenever an animal points its head in a particular direction, are thought to subserve the animal's sense of spatial orientation. HD cells are found prominently in several thalamo-cortical regions including anterior thalamic nuclei (ATN), postsubiculum (PoS), medial entorhinal cortex (MEC), parasubiculum (PaS), and the parietal cortex (PC). While a number of methods in neural decoding have been developed to assess the dynamics of spatial signals within thalamo-cortical regions, studies conducting a quantitative comparison of machine learning and statistical model-based decoding methods on HD cell activity are currently lacking. Here, we compare statistical model-based and machine learning approaches by assessing decoding accuracy and evaluate variables that contribute to population coding across thalamo-cortical HD cells.

Dirichlet Depths for Point Process

Contributed

Yang Chen, Florida State University
cy16e@my.fsu.edu

Statistical depths have been well studied for multivariate and functional data over the past few decades, but remain under-explored for point processes. A first attempt on the notion of point process depth was conducted recently where the depth was defined as a weighted product of two terms: (1) the probability of the number of events in each process and (2) the depth of the event times conditioned on the number of events by using a Mahalanobis depth. We point out that multivariate depths such as the Mahalanobis depth cannot be directly used because they often neglect the important ordering property in the point process events. To deal with this problem, we propose a model-based approach for point process systematically. In particular, we develop a Dirichlet-distribution-based framework on the conditional depth term, where the new methods are referred to as Dirichlet depths. We examine mathematical properties of the new depths and conduct asymptotic analysis. In addition, we illustrate the new methods using various simulated and real experiment data. It is found that the proposed framework provides a reasonable center-outward rank and the new methods have accurate decoding in one neural spike train dataset.

ProDCoNN: Protein Design using a Convolutional Neural Network

Poster

Chenran Wang, Florida State University
chenran.wang@stat.fsu.edu

Designing protein sequences that fold to a given three-dimensional (3D) structure has long been a challenging problem in computational structural biology with significant theoretical and practical implications. In this study, we first formulated this problem as predicting the residue type given the 3D structural environment around the C atom of a residue, which is repeated for each residue of a protein. We designed a nine-layer 3D deep convolutional neural network (CNN) that takes as input a gridded box with the atomic coordinates and types around a residue. Several CNN layers were designed to capture structure information at different scales, such as bond lengths, bond angles, torsion angles, and secondary structures. Trained on a very large number of protein structures, the method, called ProDCoNN (protein design with CNN), achieved state-of-the-art performance when tested on large numbers of test proteins and benchmark datasets.

Integrating brain imaging GWAS results to identify genes associated with Alzheimer's disease

Contributed

Shengjie Jiang, Florida State University
shengjie.jiang@stat.fsu.edu

The genetic architecture of Alzheimer's disease is largely unknown. Imaging-wide association study (IWAS) that integrates brain imaging information with genome-wide association studies (GWAS) results have successfully enhanced the discovery for genetic risk loci for late-onset Alzheimer's disease (AD). However, IWAS requires an individual-level reference panel that has matched brain imaging and genetic data. This reference panel can be limited in sample size and hard to acquire, which creates a challenge for IWAS. Here, we propose an integrative analysis method that integrates an arbitrary number of brain imaging GWAS summary results with the trait of interest GWAS results directly. As an illustration, we integrated GWAS summary results of several brain volume related regions of interest from 8,428 individuals in UK Biobank with several late-onset AD GWAS results. For example, when reanalyzing to-date the largest AD GWAS results, our proposed method identified 274 AD-associated genes, 209 of which have been ignored by standard gene-based tests and transcriptome-wide association studies (TWAS). These newly identified genes provide a potential new understanding of genetic regulation in AD.

Bayesian Variable Selection Through Wavelet Neural Network For Sparse Spatio-Temporal Data

Contributed

Jaehui Lim, Florida State University
jl15j@my.fsu.edu

It has become commonplace to observe big spatio-temporal datasets, which has led to a rich development of machine learning algorithms (e.g. deep learning). A standard approach to analyzing big data is to use a neural network (NN), and in this article, we introduce a new type of NN. We are motivated by a well-known limitation of NNs. Specifically, a standard implementation requires a large number of "inputs" (i.e. features, or basis functions and covariates), and it is a still open problem that some features may not be necessary. This limitation implies that NNs often need subjective tuning. Hence, we propose a new type of neural network that we call a Bayesian Wavelet Neural Network (BWNN), which utilizes wavelet bases and a Bayesian variable selection technique to address the limitations of standard NNs. In particular, the BWNN includes a node that is modeled using a Bayesian variable selection method. BWNN is a 5-layer neural network constructed by using two Spatial Mixed Effects (SME) models and a Spike-and-Slab (SS) model, where the BWNN uses the selected features from SS in the third node. This is another motivating component of our model because the SS (by itself) does not use a single set of selected features. We demonstrate how to construct the BWNN and show its high prediction performance and accurate feature selection in multi-dimensional spatio-temporal data settings.

Implementation of clusterability testing prior to clustering

Contributed

Naomi Brownstein, Moffitt Cancer Center
naomi.brownstein@moffitt.org

Cluster analysis is utilized in numerous biometrics applications, such as genomics and cancer to find and study subpopulations of interest. Thus, clustering is useful when the population under study is known to contain multiple distinct subgroups. On the other hand, the interpretation and properties of clustering methods are less clear when the population consists of a single homogeneous population. Clusterability testing enables the user to measure evidence of multiple inherent clusters and signals when such evidence is lacking, potentially rendering cluster analysis inappropriate. There is a need for user-friendly software with clusterability testing to serve as a sanity check before subsequent clustering. In this talk, we first provide a brief introduction to clusterability. Then, we discuss a new package to implement clusterability tests, including a brief sketch of the package requirements and setup. We conclude with example applications.

A Data Scientist goes to the Museum

Contributed

Bernhard Klingenberg, New College Florida
bklingenberg@ncf.edu

This talk is about a large-scale data science study of artist diversity in U.S. art museums, and the collection of the National Gallery of Art in particular. Through scraping the public online catalogs of 18 major U.S. art museums, deploying a sample of 10,000 artist records to crowdsourcing, and analyzing 45,000 responses, we infer artist genders, ethnicities, geographic origins, and birth decades. Our results for the first time provide information on gender and ethnic diversity of collections, and overall we find that 85% of artists are white and 87% are men. Our analysis also identifies museums that are outliers, having significantly higher or lower representation of certain demographic groups than the rest of the pool. Based on this work, we were granted exclusive access to the entire permanent collection of the National Gallery of Art, and we analyze which artists are on view and how their demographics have changed over time. An online app accompanies the talk.

Using Convolutional Neural Networks for Live Classification of Dolphins Whistles in Sarasota Bay

Contributed

Austin Anderson, New College Florida
austin.anderson18@ncf.edu

Bottlenose Dolphin bioacoustics research currently requires either temporary capture or close following of individual dolphins by boat. One fundamental limitation of this work is that boat noises and human interaction strongly disrupt the dolphin's natural behavior. Over the past few years, in order to obtain undisturbed dolphin recordings, researchers have installed passive acoustic monitoring stations throughout Sarasota Bay. Over one-hundred gigabytes of sparse and noisy audio data is produced every day and researchers cannot realistically parse through this much data to locate relevant dolphin events. For this project, a dataset of approximately 6,000 dolphin whistles was curated and then utilized to train convolutional neural networks (CNNs), the most popular computer vision models currently in use. These CNNs were designed to identify the presence of unique dolphins by their signature whistles. The CNNs are now computationally inexpensive enough to function on field devices for deployment in the Sarasota bay and are currently being used at a select site for live classification and data curating. In this talk, we will present the development, current results, and future directions of this project.

Canonical Decomposition and Wavelet Natural Vectors in High Dimensional Classification Applications

Contributed

Senthil B. Girimurugan, Florida Gulf Coast University
sgirimurugan@fgcu.edu

High dimensional data analysis continues to be a challenge while novel methods are being developed in the scientific community. Especially the $n << p$ is of importance from a statistical perspective; adding to this challenge is the size of the data itself. Recently, data sizes have become significantly larger (varying in multiple Gigabytes) with a need for efficient computation. In this talk, a method is proposed using Canonical Decomposition (CANDECOMP) to reduce the dimension of large data (in terms of size as well as $n << p$ situations) to obtain feature matrices. These matrix features are then used in the computation of Wavelet Natural Vectors (WNV) using Discrete Wavelet Packet Transform (DWPT) to perform classification. CANDECOMP offers a sound approach to shrink the dimension without losing information in the original data. In order to achieve computational efficacy in classification with the lower dimensional feature matrices, matrix Quadratic Discriminant Analysis (mQDA) is employed. The method has shown good classification accuracy on a well known dataset. This talk will elaborate on the details of the method in performing a classification task in higher dimensions and demonstrating the efficiency on CANDECOMP, and its variants, in shrinking high dimensional data.

Genetic Connection to Drug Induced Liver Injury (DILI) through Statistical Learning Methods

Poster

Roland Moore, Florida State University
rm16n@my.fsu.edu

Drug Induced Liver Injury (DILI) is the major cause of drug development failure or drug withdrawal from the market after development. The US government puts a lot of funds into investigating major causes and remedies of DILI. Therefore, investigating factors associated with DILI is of paramount importance. Environmental factors that contribute to DILI have been investigated and are, by and large, known. However, recent genomic studies have indicated that genetic diversity can lead to inter-individual differences in drug response. Consequently, it has become necessary to also investigate how genes contribute to DILI in the presence of environmental factors. Thus, our aim is to find appropriate statistical methods to investigate gene-gene and/or gene-environment interactions that are associated with DILI. This is an initial study that only explores statistical learning methods to find gen-gene interactions (epistasis). We introduce Multifactor Dimensionality Reduction (MDR), Random Forest (plus logistic regression), and Multivariate Adaptive Regression Splines (MARS), as the few potential methodological approaches that we found. Next, we improved the MARS method by combining it with a variable selection method.

Robust estimation and selection for single-index regression model

Contributed

Huybrechts Bindele, University of South Alabama
hbindele@southalabama.edu

In this talk, we will consider a single-index regression model for which we will discuss a robust estimation procedure for the model parameters and an efficient variable selection of relevant predictors. The proposed method is known as the penalized generalized signed-rank procedure. Asymptotic properties of the proposed estimator are established under mild regularity conditions. Extensive Monte Carlo simulation experiments are carried out to study the finite sample performance of the proposed approach. The simulation results demonstrate that the proposed method dominates many of the existing ones in terms of the robustness of estimation and efficiency of variable selection. Finally, a real data example is given to illustrate the method.

Determining the Impact of Prerequisite Math Courses on Calculus 1 Pass Rate: A Naïve Bayes Classification Approach

Contributed Talk

Jay Kim Sparks and Anthony Okafor, University of West Florida

The importance of foundational math courses among first time in college STEM majors is well recognized. Calculus 1 is a keystone course tied to retention and graduation. Many students enter college unprepared for calculus, and so there are multiple paths towards preparing for and taking the course. The effectiveness of these paths versus a student's performance in them is unclear. This research asks, "Based on a student's prerequisite math courses and grade earned, demographic variables and GPA, how likely are they to pass Calculus 1?" Data were analyzed using Naïve Bayes classification. Overall, the different courses preparing students for Calculus 1 were found to be less important than students' performance in those courses. Generally, the higher the grade received in the prerequisite courses, the more likely a student was to pass Calculus. This suggests that in terms of improving STEM retention and graduation, improving student performance would be more effective than devising new paths towards preparation.

Geospatial Analysis of 1949-2018 Oscillations in Rainfall Patterns in the Gulf Coast: Impacts for Climatological Modeling

Poster

Cody Goins, Cooper Corey, Samuel Parmer, Anthony Okafor and Jason Ortegren - University of West Florida

Oscillations in springtime (April and May) rainfall between the years 1949 and 2018 were observed at the Pensacola International Airport weather station. Our goals were two-fold: 1) identify other stations that exhibited similar patterns as Pensacola International Airport in order to determine the geographical extent of this oscillation, and 2) find and model factors that are believed to affect these observations. The first step was to examine the oscillation observed at Pensacola International Airport (PNS). Trend-lines and seasonal indices were used to characterize the data. These were then compared with data from 37 other stations spread throughout the southeast. Comparison of trend-lines showed that 15 stations, including PNS, exhibited positive trends in rainfall. Coefficients of variation were also used to compare the station data. These comparisons show that stations along the Gulf Coast exhibited similar patterns in rainfall from 1949-2018. A principal components analysis (PCA) was performed in order to regionalize the data. To determine the factors that caused these observations, a thorough literature review was performed. Our research suggests that this oscillation extends further into the southeastern US, mainly along the Gulf Coast. Our research also suggests that, although there are several influences, the main cause of these oscillations observed is the Bermuda High Index. These findings are important due to the implications rainfall patterns can have for industries such as agriculture and tourism. Identifying

and understanding these oscillations could be important to increasing our ability to predict rainfall patterns. Being able to predict these patterns could also allow for better preparation for rainfall events such as floods and droughts.

Data Exploration and Engagement Strategies for Just-in-time Tutoring and Promoting Active Learning

Workshop

Melanie A. Sutton, Anthony Okafor, Justice Mbizo, Brian Le, Kimberly Rogers, Logan Goodson, Naomi Semaan – University of West Florida

This workshop will provide hands-on training on how to quickly capture and edit MP4 videos of your screen explaining concepts to your students. Next, we'll demonstrate how to showcase these videos on a YouTube channel for rapid dissemination and promoting student-to-student engagement. Finally, we will create shareable maps and data tour movies using Excel 3-D Maps. You will plot data with locations (e.g., addresses, ZIP Codes, or locations) on globes or images and digitally fly through and over data scenes and data layers. Demonstrations of additional animation options when temporal information is added will showcase the power of data visualizations in space and time.
