American Statistical Association
Virginia Chapter 2019 Annual Meeting

LOCATION and DESCRIPTION

The Virginia Chapter of the American Statistical Association (ASA) will hold its annual meeting at the University of Virginia, Charlottesville, VA, on Friday, October 25, 2019. We will meet in Monroe Hall, Room 130, which can be found here (Building 26 on the map included below). We are pleased to announce that Dr. Susan Ellenberg, Professor of Biostatistics at the Pennsylvania University Department of Biostatistics and Epidemiology, will deliver the Honorary Guest Lecture. ASA President Dr. Karen Kafadar will also delivery a guest lecture. There will be additional sessions of student and faculty presentations covering a wide range of important topics in fields such as adaptive clinical trial designs, geospatial clustering, machine learning, and genomics. A detailed agenda is listed below.

PARKING

Parking is available at the Culbreth Road Parking Garage (130 Culbreth Road, Charlottesville, VA 22903, Building 16 on Map), which is 0.5 miles from Monroe Hall. Parking at the Culbreth Road Parking Garage will be free with your registration; no need to pay at kiosk. Additional parking is available at the Central Grounds Parking Garage (400 Emmet St, Charlottesville, VA 22903, Building 9 on Map), which is 0.2 miles from Monroe Hall. The rate at the Central Grounds Parking Garage is $2 per hour and payment can be made with cash or credit card.

LODGING

There are several lodging options near the UVA Campus, which can be explored through the following link.

REGISTRATION

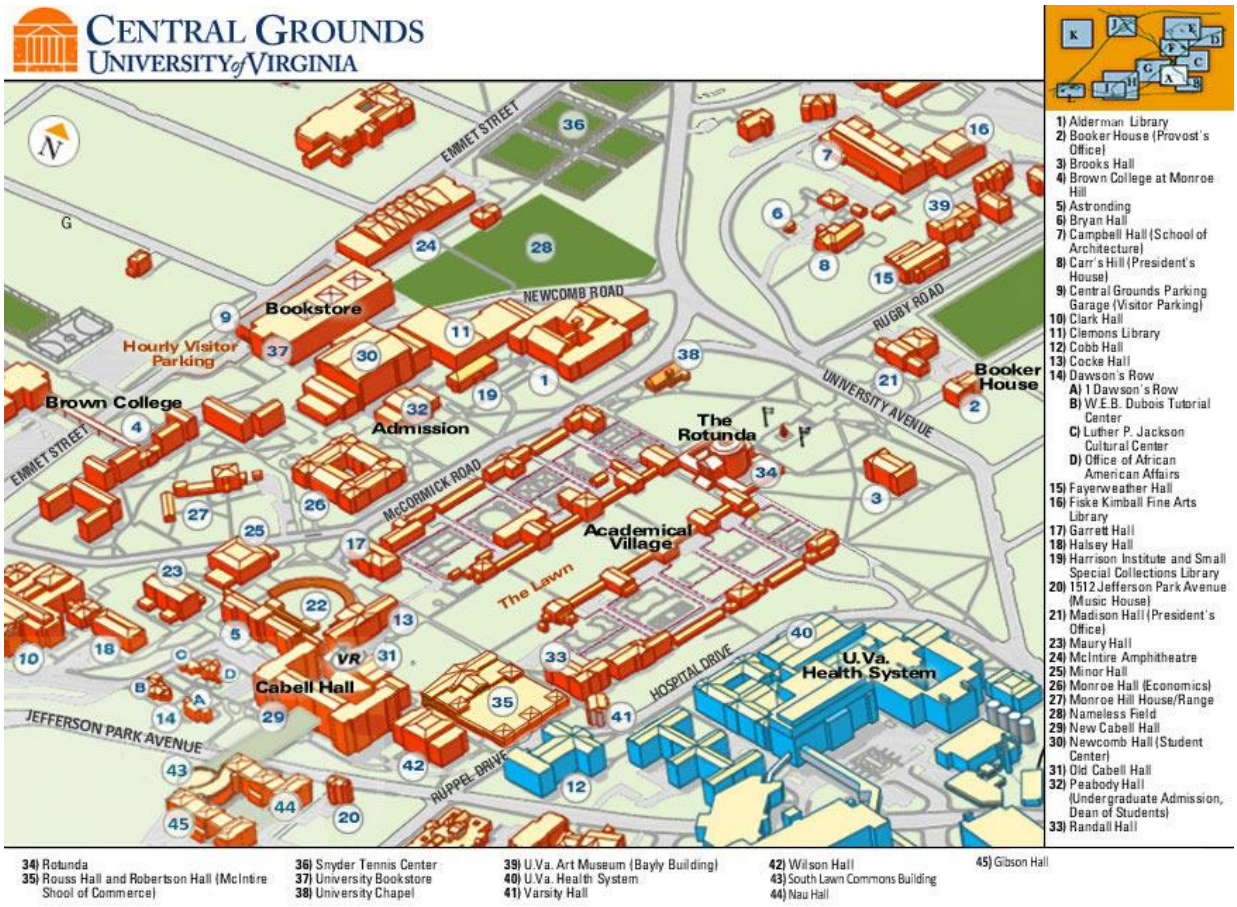
All attendees and presenters must register ($10 registration fee for students; $20 registration fee for professionals and faculty) to help support the Chapter provide light refreshments. You can register in advance via check or on site with cash or personal check; we cannot accept credit or debit cards at this time. For advance registration, please make checks payable to the *American Statistical Association* and mail to: Roy T. Sabo, Ph.D., Department of Biostatistics, Virginia

Commonwealth University, 830 East Main Street, Richmond, VA 23298. Feel free to direct questions or inquiries to roy.sabo@vcuhealth.org.

AGENDA (All events in Monroe Hall, Room 130)

Time	Topic
9:00 – 9:40am	Registration & Introductions
9:40 – 9:45 am	Welcome
9:45 – 10:30am	Opening Address • Karen Kafadar, Ph.D., University of Virginia
10:30 – 10:45am	Break
10:45 – 11:30am	Late Morning Faculty Presentations
11:30 – 12:00 noon	Business Meeting
12:00 – 1:00pm	Lunch
1:00 – 2:00pm	Honorary Guest Lecture • Susan Ellenberg, Ph.D., Pennsylvania University
2:00 – 2:15pm	Break
2:15 – 3:45pm	Afternoon Student Poster Session
3:45 – 4:00pm	Awards and Concluding Remarks

University of Virginia Campus Map



Abstracts for e-POSTERS

- Chenlu Ke, Statistical Sciences and Operations Research, Virginia Commonwealth University

Title: Sufficient Variable Selection via Expected Conditional Hilbert-Schmidt Independence Criterion

Variable selection plays a significant role in modeling modern statistical problems associated with ultrahigh dimensional data. We develop a novel model-free sufficient variable selection procedure based on a powerful independence measure proposed recently. Compared with sure independence screening methods which only consider marginal dependence between the response and each predictor, our approach takes advantages of the new measure and incorporates joint information additionally to achieve sufficient variable selection. As a result, our method is more capable of selecting all the truly active variables. Furthermore, our procedure can handle both continuous and discrete responses with mixed-type predictors, which is unfeasible for most existing methods built upon independence measures. The sure screening property of the proposed approach is established under mild conditions. Simulation studies are presented to illustrate the advantages of our method

- Lawrence Leemis, Department of Mathematics & Statistics, College of William & Mary

Title: Confidence Intervals for the Binomial Parameter

We propose two measures of performance for a confidence interval for a binomial proportion p: the root mean squared error and the mean absolute deviation. We also devise a confidence interval for p based on the actual coverage function that combines several existing approximate confidence intervals. This "Ensemble" confidence interval has improved statistical properties over the constituent confidence intervals. Software in an R package which can be used in devising and assessing these confidence intervals is available on CRAN. In addition, ongoing research to design algorithms to calculate confidence interval bounds that minimize the root mean squared error is outlined

Abstracts for POSTERS

- Kristine Gierz, Department of Mathematics & Statistics, Old Dominion University

Title: A Non-parametric Approach to Estimating the Point of Treatment Time-lag Effect

In general, the change point problem considers inference of a change in distribution for a set of time-ordered observations. This has applications in a large variety of fields, and can also apply to survival data. In survival analysis, most existing methods compare two treatment groups for the entirety of the study period. Some treatments may take a length of time to show effects in subjects. This has been called the time-lag effect in the literature, and in cases where time-lag effect is considerable, such methods may not be appropriate to detect significant differences between two groups. In this paper, we propose a novel non-parametric approach for estimating the point of treatment time-lag effect by using an empirical divergence measure. Theoretical properties of the estimator are studied. The results from the simulated data and real data example support our proposed method.

- Camille J Hochheimer, Public Health Sciences, University of Virginia

Title: cpsurvsim: An R package for simulating data from change-point hazard distributions

Change-point hazard models have several practical applications, including modeling processes such as cancer mortality rates and disease progression. When simulating data from change-point hazard distributions with more than one parameter, however, we begin to see biased parameter estimates when simulation data using the inverse CDF method, a common frequentist method for simulating data. Thus, we propose an alternative method for simulating data that exploits the memoryless property of survival data, compare its performance with the inverse CDF method, and introduce the R package cpsurvsim which implements both methods.

- Huihui Lin, Department of Mathematics & Statistics, Old Dominion University

Title: D-vine copula model for dependent binary data

In real world application, high-dimensional binary response datasets have been available in a wide range of scientific areas. We proposed pair copula models to analysis the marginal mean and the dependence of response, assumed first order autoregressive or equicorrelated structured. Copula is very useful to separate dependency relationship from multivariate distribution. Bivariate copula is sufficient because a multivariate copula, which is basically a joint cumulative distribution function (cdf), can be decomposed into pair vine copula with corresponding margins. In the dependent binary data case, the correlation structure is assumed as autoregressive or equicorrelated structured, which consists of correlation coefficient of the binary variable. High dimension pmf of binary variables can be constructed using discrete D-vine copula, which is preferred if there is no prior information about the correlation. Discrete C-vine can be used as well, if one variable is known to be dominantly related to the rest than the others. The copula parameter can be solved uniquely given binary variables correlation, so the relationships are obtained. We also prove that D-vine pair copula will give the same joint probability mass function, even different copulas are used, including Clayton, Frank and

Gumbel copula. The Multivariate Probit model is one of the most popular model to analyze the dependence relationships of longitudinal binary data. But the MP model sometimes fails even within a feasible range of binary variables correlations, because the MP model requires a positive definite correlation matrix.The D-Vine pair copula we presented works when the MP model fails. We have also shown that MP model is different from vine Gaussian pair copula starting three dimensions.

- Alice Liu, Department of Statistics, University of Virginia

Title: Eyewitness Identification Accuracy: Rethinking the Framework

Eyewitness play a critical role in the investigation of crimes and the subsequent legal proceedings. However, law enforcement do not have the time and resources available to conduct the much-needed research for the development and validation of more reliable practices. Research in the effectiveness of law enforcement practices for eyewitness identification (EWI) procedures remains incomplete. Eyewitnesses make errors, which can result in grievous consequences. Currently, there are a few options for eyewitness identification analysis, including ROC curve analysis, which only considers the positive predictive value and one procedural factor at a time, Bayesian prior-posterior plots, and decision utility. However, all of these methods lack a fundamental way to include variability, as well as consider the complex and interactive relationships of the variables affecting eyewitness identification accuracy. The reviewed statistical methods will be used to evaluate data from eyewitness identification studies, but will also be used to develop new approaches that include additional information for the proper assessment of eyewitness identification accuracy. In order to supplement the model strength, the individual probability of choosing, which is a factor of accuracy, is assessed and estimated using classification models. Previously, researchers generally treat the probability of choosing as deterministic, but it is probabilistic. The probability of choosing forms the fundamental components in the estimation of the probability of accuracy. Further, in a secondary step, a framework is proposed to estimate the probability of target presence in a lineup, based on the decomposition of the probability of eyewitness accuracy. This two-tiered approach represents a new framework of modeling eyewitness identification accuracy in order to understand and augment the efficacy of eyewitness identification procedures. This framework utilizes the entirety of the data available and takes into consideration the fundamental effects of variability, uncertainty, and interactive relationships.

- Xiaoyuan Ma, Department of Statistics, University of Virginia

Title: Locally Hierarchical Graphical Model

Abstract: Detecting potential gene associations in order to provide statistical evidence for oncogenic pathway interpretations is a challenging problem. For data consisting of RNA sequence counts, variations of Log Linear Graphical Models (LLGM) are the predominant method used for analyzing such associations. However, these models encourage overly conservative graphical models with fewer edges because of the model selection mechanism, i.e. graphical lasso combined with instability selection. To address this problem, we propose the

Locally Hierarchical Graphical Model (LHGM), which (1) allows incorporating prior knowledge on the sparsity of the graphical model; and (2) avoids the graphical lasso, promoting a more informative graphical model under the same level of instability. In brief, our method considers the union of all local graphical models, as in LLGM, but allows for flexible penalties on local graphical models. For each local graphical model, the penalty coefficient is controlled by a prior distribution on the activated links. A comprehensive analysis on both simulated and real-world RNA-seq data sets shows that our method discovers a graphical model with richer information and closer approximation to the truth.

- Christine M. Orndahl, Department of Biostatistics, Virginia Commonwealth University

Title: Integrated Multiple Adaptive Clinical Trial Design Involving Sample Size Re-Estimation and Response-Adaptive Randomization for Continuous Outcomes.

Currently, single adaptive clinical trial designs are utilized most often, where only one adaptive design is used within the clinical trial. Consequently, only one pitfall of a fixed clinical trial design is addressed. Recently, increased interest has been developed in the area of multiple adaptive designs, incorporating more than one adaptive design within a single clinical trial. However, these multiple adaptive designs are typically performed in succession and information is not shared across the different designs. The goal of this project is to integrate multiple adaptive designs, specifically sample size re-estimation and response-adaptive randomization, into a clinical trial with a continuous outcome. In order to accomplish this, the weighted sum method for multi-objective optimization with a constraint to maintain statistical power is used to combine two objective functions. The first minimizes the sample size required while the second minimizes the total expected treatment response. These objective functions serve to adaptively adjust the sample size and the allocation ratio; this ensures that the fewest number of patients are enrolled in the trial while still maintaining adequate statistical power, and for the patients enrolled, the expected response of the sample is minimized which seeks to maximize the overall benefit of the trial. Preliminary results for applying these new methods to a clinical trial are presented.

- Kayoung Park, Department of Mathematics & Statistics, Old Dominion University

Title: Evaluation of the treatment time-lag effect for survival data

Medical treatments often take a period of time to reveal their impact on subjects, which is the so-called time-lag effect in the literature. In the survival data analysis literature, most existing methods compare two treatments in the entire study period. In cases when there is a substantial time-lag effect, these methods would not be effective in detecting the difference between the two treatments, because the similarity between the treatments during the time-lag period would diminish their effectiveness. In this paper, we develop a novel modeling approach for estimating the time-lag period and for comparing the two treatments properly after the time-lag effect is accommodated. Theoretical arguments and numerical examples show that it is effective in practice.

- Reuben Retnam, Department of Biostatistics, Virginia Commonwealth University

Title: Matrix-Variate Skew-t Models for Big Data with the Distributed EM Algorithm

The Matrix-Variate Skew-t (MVSt) distribution, introduced by Gallaugher and McNicholas in 2017, was one of the first attempts at creating a distribution that allows researchers to model skewed and heavy-tailed data in a matrix-variate setting. The distribution permits a convenient hierarchical representation, allowing it to be fit relatively easily via the Expectation-Maximization (EM) algorithm. However, the EM algorithm used to fit models that utilize this distribution can become cumbersome when applied to large matrix-variate datasets. In our work, we develop regression models for matrix-variate data that utilize the matrix-variate skew-t distribution while scaling to tackle today's massive amounts of data. This scaling is achieved via the implementation of divide-and-conquer techniques that utilize the distributed expectation-maximization algorithm. Specifically, the E-step of the EM algorithm is run in parallel on multiple worker processes, while manager processes perform the M-step with a fraction of the results from the local expectation steps. Further work will extend these models to other complex data structures, such as irregularly observed longitudinal data.

- Spiro Stilianoudakis, Department of Biostatistics, Virginia Commonwealth University

Title: Developing a computational framework for precise TAD boundary prediction using genomic elements

The advent of chromosome conformation capture combined with high-throughput sequencing experiments (Hi-C) have revealed important 3-dimensional (3D) spatial constructs within the human genome. Among these, Topologically Associated Domains (TADs) represent fundamental building blocks that lead to the organization and regulation of the genome. The enrichment of several epigenetic marks at TAD boundaries suggests a strong epigenetic component of TAD boundary formation. Several methods have been developed to leverage the growing amount of (epi)genomic annotation data for predicting TAD boundaries. However, they overlooked key characteristics of 3D genomic data including Hi-C resolution, severe imbalance in TAD-boundary vs non-TAD boundary defined genomic regions, and spatial associations between genomic annotations and boundaries. Furthermore, Hi-C data resolution (hundreds of thousands of bases) remains well below the resolution of epigenomic assays (hundredths of bases) affecting the ability of conventional TAD-calling algorithms in precisely delineating the boundaries of TADs. Here, we proposed a computational framework to establish the optimal combination of Hi-C data resolution, re-sampling technique, and feature engineering procedure necessary for TAD boundary prediction by leveraging high resolution ChIP-seq data in an ensemble of random forest classification models. Using our optimally built ensemble classifiers we then aimed to more precisely predict the location of chromosome-specific TAD boundaries at base pair resolution. We demonstrate that ensemble random forest classifiers built on 10 kb resolution Hi-C data, with distance-type predictors, using SMOTE (Synthetic Minority Oversampling Technique) re-sampling yield optimal predictive performances, and outperform existing methods. Transcription factor binding sites (TFBS) were the best type of genomic annotations for predicting TAD boundaries when directly compared with histone modifications and chromatin states. We identified the well-known CTCF, SMC3,

RAD21, and ZNF143 TFBS known to be enriched at TAD boundaries as having high predictive importance. Additionally, TAD boundary regions predicted at base pair resolution using our optimally built models were found to be significantly closer to known molecular drivers of the 3D genomic architecture compared to existing TAD calling algorithms, indicating more precise identification of TAD boundaries. Our results outline a useful guide toward the exploration of chromatin organization, while highlighting the impact of data resolution, class imbalance, and feature engineering in TAD boundary prediction. This improved precision in TAD boundary location offers the potential for further exploration into the 3D organization of the human genome.

- Lucia Tabacu, Department of Mathematics & Statistics, Old Dominion University

Title: The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: NHANES 2003-2006

We study the 5-year all-cause mortality predictive performance of measures of physical activity collected from accelerometers worn by older participants in NHANES 2003-2004. A total of 20 measures of objective physical activity and other mortality predictors were compared using univariate and multivariate logistic regression. The accelerometry derived physical activity measures outperformed the traditional predictors of mortality, including age in terms of cross-validated AUC. This is joint work with E. Smirnova, A. Leroux, Q. Cao, V. Zipunnikov, C. Crainiceanu, J. Urbanek.

- Katarzyna M Tyc, Department of Biostatistics, Virginia Commonwealth University

Title: TADcompare: an R package for differential analysis of Topologically Associated Domains

Interphase chromatin folds into highly conserved three-dimensional (3D) structures with ascribed regulatory functions. Development of chromatin conformation capture technologies, such as Hi-C, revealed an existence of chromatin domains characterized by higher chromatin interactions within them than between them. These Topologically Associated Domains (TADs) are fundamental to guiding gene expression regulation and determining a cell fate. Although TADs are considered to be relatively stable 3D genomic structures, many dynamically reorganize during development or disease, and exhibit cell- and condition-specific differences. Quantification of the dynamic behavior of TADs is challenging and methods devoted to this task remain at their infancy. Here, we propose a robust statistical method for accurate differential analysis of TAD boundaries between Hi-C datasets. `TADCompare` employs a spectral clustering-derived measure that enables a highly sensitive loci-by-loci comparison of TAD boundary differences between Hi-C datasets. Based on this measure, we introduce a strategy that allows for differential and consensus TAD boundaries detection and tracking of temporal TAD boundary changes. Taken together, we present a comprehensive framework for a systematic classification of TAD boundary changes. We demonstrate that different types of TAD boundary changes are associated with distinct biological functions, as seen through the enrichment analysis of the affected genes. Simulated datasets and known biological markers are

used to further demonstrate the efficiency and validity of our method. `TADCompare` is available on https://github.com/dozmorovlab/TADCompare.

- Jonathan W. Yu, Department of Biostatistics, Virginia Commonwealth University

Title: A new estimation method for the semiparametric accelerated mixture cure model

In clustered data such as the United Network of Organ Sharing (UNOS) database, the center size can affect patient survival time after transplant with its access to medical resources. Improper statistical procedures to handle informative cluster size can lead to biased results and misleading inferences. While the accelerated failure time (AFT) mixture cure model and the Cox proportional hazards (PH) mixture cure model are two classic models to analyze clustered survival data, the AFT has attracted less attention than its semiparametric counterpart due to the complexity of the estimation method. However, its direct physical interpretation and developments to the rank-based generalized estimating equations (GEE) provides an incentive to use for censored failure time data. We propose a new estimation method for the semiparametric AFT mixture cure model that employs a faster expectation-maximization (EM) algorithm, the SQUAREM, that can accelerate any fixed-point and smooth mapping with linear convergence rate and an induced smoothing inverse cluster size reweighting procedure to handle the informative cluster size. To evaluate the performance of the proposed method, we conducted a simulation study. The results of the simulation study demonstrate that the proposed method performs better than the existing estimation method. We apply the proposed method to UNOS data of failure times from kidney transplant patients to demonstrate that this approach has better numerical performance than existing methods in literature.

- Jing Zhang, Systems Modeling and Analysis, Virginia Commonwealth University

Title: Sufficient Dimension Reduction and Outlier Detection

High dimensionality has been a significant feature in modern statistical modeling. Sufficient dimension reduction (SDR) approach is an efficient tool to explore the low dimensional projection subspace without losing full regression information between the response and the high dimensional predictors. Minimum average variance estimation (MAVE) is a popular method for dimension reduction. However, it is not robust to the outliers in the response due to the use of least squares. In this study, we proposed a new robust SDR method based on MAVE that introduces a mean-shift parameter as an indicator of the influence of each observation in the data. A penalty term on these mean shift parameters can help identify outliers and thus achieve robust estimation. Simulation studies show that our method has high prediction accuracy on estimating the effective dimension reduction directions in the presence of outliers in the response. We also show that our method is not sensitive to the choice of the initial values.