Invited Keynote Session

Title: From Synthesis to Trust: Advanced Statistical Methods for Trustworthy Generative AI in Biomedical Imaging Studies

Speaker: Lily Wang, Professor of Statistics, George Mason University

Abstract: Generative AI has rapidly transformed the biomedical imaging field by enabling image synthesis, helping address challenges of limited data availability, privacy, and diversity in biomedical research. Yet, the adoption of AI-generated images in biomedical studies requires rigorous methods to ensure their reliability for downstream analysis. In this talk, I will introduce novel and rigorous nonparametric approaches that strengthen the trustworthiness and statistical validity of synthetic biomedical imaging data. We develop simultaneous confidence regions to rigorously quantify uncertainty and detect meaningful differences between synthetic and original imaging data. To further enhance fidelity and utility, we propose a transformation that aligns the mean and covariance structures of synthetic images with those of the originals. I will also discuss methods for imputing missing imaging phenotypes using generative models and demonstrate how joint analysis of observed and imputed traits enhances inference while accounting for imputation error. Extensive simulations and applications to brain imaging data validate the proposed framework, demonstrating how these methods empower rigorous statistical inference and promote trustworthy advances in biomedical imaging.

Abstracts for oral presentations

1. **Title: Scalable and robust functional region detection in basis space**

Authors: Ruijin Lu, Hongxiao Zhu*
Affiliation: Virginia Tech

Abstract: Detecting local regions in functional data is critical for many applications. Existing approaches, such as functional regression or pointwise testing, may be applied for this purpose but often lack scalability and suffer from reduced statistical power as the dimensionality or resolution of measurement points increases. In this paper, we introduce a basis-space testing framework to identify local regions that are significantly different across groups of samples. The framework employs compactly supported, potentially multiresolution bases to effectively capture local features of functional data. Efficient region detection is achieved through performing simultaneous testing combined with a p-value-guided compression in basis space. This basis-space testing approach reduces the dimensionality of functional data and results in improved power of region detection. It also supports distributed computation on manycore CPUs or multicore GPUs. Detected regions can be conveniently visualized by inverse-transforming significant components back into the data domain, facilitating intuitive interpretation. We focus on two types of multiplicity adjustment in basis-space testing: the Westfall-Young randomization test, which controls the family-wise error rate, and the Benjamini-Hochberg procedure, which controls the false discovery rate. Our theoretical results demonstrate that, under mild regularity conditions, the Westfall-Young randomization test in basis space achieves strong control of family-wise error rate. Moreover, applying appropriate compression in basis space enhances power compared to pointwise testing in data domain or basis-space testing without compression. The empirical performance of the proposed procedure is evaluated using simulation studies and real-world biomedical and brain imaging datasets defined on 1-D and 3-D domains.

2. **Title: BOOOM: Loss-Function-Agnostic Black-Box Optimization over Orthonormal Manifolds for Machine Learning and Statistical Inference**

Authors: Priyam Das*, Beomchang Kim, Subhrajyoty Roy
Affiliations: Dept. of Biostatistics, Virginia Commonwealth University.

Abstract: Optimization over the Stiefel manifold St(p, d), the set of p × d column-orthonormal matrices, arises in statistics, machine learning, and engineering, and often involves objectives which are non-convex/multi-modal, discontinuous, or of non-closed form. Most existing approaches depend on either convex relaxation techniques or derivative-based algorithms tailored to specific objective functions, which restricts their applicability to a broader class of problems. We present BOOOM (Black-box Optimization Over Orthonormal Manifolds), a general-purpose global optimization framework for St(p, d) based on Recursive Modified Pattern Search. BOOOM reparameterizes the manifold via a complete Givens rotation decomposition, proving that any point on St(p, d) can be obtained from a fixed orthonormal base through a finite sequence of rotations, extending the classical Hurwitz result. This mapping transforms constrained manifold optimization into an unconstrained Euclidean search over rotation angles, enabling derivative-free exploration while preserving feasibility exactly at each iteration. We establish global convergence properties under mild conditions, guaranteeing that all accumulation points are globally optimal for the black-box problem. In benchmark studies, BOOOM is shown to outperform state-of-the-art optimization toolboxes for the Stiefel manifold across a range of synthetic and real-world problems, including maximum heterogeneous quadratic form optimization. We further apply BOOOM to a novel supervised PCA problem to identify key metabolites associated with colorectal cancer while balancing out their classification power across healthy and colorectal cancer patients. These results highlight BOOOM as a versatile and powerful tool for orthogonally constrained learning and statistical estimation.

3. **Title: Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions**
   Author: Youngjin Cho
   Affliation: Virginia Tech (Soon to be at University of Nevada, Las Vegas).

Abstract: In a Cox model, the partial likelihood, as the product of a series of conditional probabilities, is used to estimate the regression coefficients. In practice, those conditional probabilities are approximated by risk score ratios based on a continuous time model, and thus result in parameter estimates from only an approximate partial likelihood. Through a revisit to the original partial likelihood idea, an accurate partial likelihood computing method for the Cox model is proposed, which calculates the exact conditional probability using the Poisson-binomial distribution. New estimating and inference
procedures are developed, and theoretical results are established for the proposed computational procedure. Although ties are common in real studies, current theories for the Cox model mostly do not consider cases for tied data. In contrast, the new approach includes the theory for grouped data,
which allows ties, and also includes the theory for continuous data without ties, providing a unified framework for computing partial likelihood for data with or without ties. Numerical results show that the proposed method outperforms current methods in reducing bias and mean squared error, while
achieving improved confidence interval coverage rates, especially when there are many ties or when the variability in risk scores is large. Comparisons between methods in real applications have been made.
--

4. **Title: Patient facial expressions correlate with cancer genetic test uptake: an application of image-on-scalar regression**

Authors: Pavel Chernyavskiy, PhD[1], John Quillin, PhD[2], Nadia Harika[2], Anaya Gorham[2], Nao Hagiwara, PhD[1]
Affiliations: [1]University of Virginia School of Medicine, Department of Public Health Sciences, [2]Virginia Commonwealth University School of Medicine, Department of Pediatrics

Abstract: In cancer care, emotions have a profound effect on patient decisions and subjective care experience. Emotional reactions, for example to familial risk data presented by the genetic counselor, are among the first reactions a patient will have during genetic counseling, and these reactions can have a profound impact on all subsequent information processing, attention, and reasoning. For many patients with cancer, genetic counseling is often the first step in their continuum of care. Thus, emotional reactions to genetic counseling are a fundamental driver of patients' subsequent behaviors, such as genetic test uptake. Here, we encoded patient facial expressions during 42 virtual genetic counselling sessions using a validated automatic facial coding software (FaceReader 9.1). FaceReader generated numeric values for the patients' valence (spectrum of positive to negative emotions) and arousal (alertness) for each video frame following the circumplex model of affect, which places valence on the x-axis and arousal on the y-axis. These valence-arousal scatterplots for each patient were binned spatially into quadrats, the counts of frames in each quadrat recorded and converted to images to serve as covariates in an image-on-scalar regression predicting genetic test uptake. We describe how to process valence and arousal such that these data can be included in an image-on-scalar regression and show that the resultant model outperforms models using demographic and appointment data alone.

--

5. Title: Use of the Max-Min distance as a novel optimal biomarker cuto: selection method for tree-ordered data

Authors: Benjamin C. Brewer (PhD)[1],*, Leonidas E. Bantis (PhD)[2]
Affiliations: [1]Center for Biostatistics & Health Data Science, Department of Statistics, Virginia Tech, Roanoke, VA, [2]Department of Biostatistics & Data Science, Kansas University Medical Center, Kansas City, KS

Abstract: The use of ROC curves for evaluating the discriminatory ability of continuous markers when there exist two classes of the underlying disease - typically healthy and diseased - is commonplace. However, real-world data are often much more complicated and may feature more than two disease states, some of which may be ambiguously ordered in regard to their biomarker values. Recent research has addressed the analysis of these so-called tree-ordered data. Within this framework, multiple cuto: selection strategies have been proposed; namely, the Tree Youden Index and the Euclidean distance. In this presentation, we discuss and compare these two existing cuto: selection strategies to a new proposed measure, the Max-Min distance, that seeks to find the operating point on the TROC curve furthest from uninformativeness. We provide theoretical results regarding the optimality of our proposed measure that relate to uniqueness and optimal ROCs. We further illustrate the collapse of known measures under certain scenarios. We provide an inferential framework for the derivation of confidence intervals in a parametric and nonparametric fashion. We then evaluate our approaches through simulations and provide an application involving patients with tuberculosis.

--

6. Title: Covariate-Adjusted Placement Value-based ROC Models for Racial Heterogeneity in Birthweight Screening

Author: Soutik Ghosal.
Affiliations: Division of Biostatistics, Public Health Sciences, School of Medicine, University of Virginia

Abstract: Small-for-gestational-age (SGA) and large-for-gestational-age (LGA) births are clinically important outcomes linked to heightened neonatal complications and long-term health risks. Ultrasound-based estimated fetal weight (EFW) is widely used as a screening biomarker for both conditions, yet its diagnostic performance may vary across demographic subgroups. Prior findings show accuracy differences by maternal BMI, and emerging

evidence suggests fetal growth trajectories also differ by patient-reported race, motivating the need for race- and context-specific screening assessments.
We address this problem by developing a nonparametric placement value (PV)-based ROC modeling framework to evaluate diagnostic accuracy of EFW while accounting for race-specific heterogeneity across gestation. Unlike traditional ROC approaches that separately model healthy and diseased biomarker distributions, PV-based methods exploit the standardization of healthy scores, allowing the ROC curve to be expressed as the distribution function of PVs. This representation enables covariate adjustment directly on the ROC function, thereby facilitating the joint incorporation of categorical (e.g., race) and continuous (e.g., gestational age) factors.
Our proposed class of nonparametric PV-ROC models provides subgroup-specific estimates of diagnostic capacity measures, with the capacity to identify race-specific optimal cutoffs if warranted. This framework offers a flexible and principled approach to quantify heterogeneity in screening performance that would be obscured under pooled analyses. Using data from the NICHD Fetal Growth Studies, we illustrate how failing to account for race-specific variability in diagnostic accuracy may perpetuate disparities in prenatal care.

Abstracts for poster presentation

1. Title: Topological Data Analysis Methods As a Localized Prediction Tool for High-Dimensional Data.

Authors: Md Moinul Ahsan*, Nitai Mukhopadhyay.
Affiliations: Virginia Commonwealth University.

Abstract: High-dimensional data often contain intricate nonlinear structures that challenge the assumptions of traditional parametric modeling techniques. In this work, we introduce an alternative prediction framework based on Topological Data Analysis (TDA) using the Mapper algorithm, designed to capture localized data geometry without prespecified global functional forms. Our approach constructs a topological graph with hierarchical clustering that segments the data via adaptive balanced cover. This enables predictions to be made by aggregating information from structurally similar regions of the data. We extend this framework to support both continuous and binary outcomes and propose a variable selection method based on permutation-based predictive loss, providing interpretable insights about the importance of covariates in terms of prediction. Our simulation study showed a better predictive performance compared to LOESS, GLM, and GLM-Q, but comparable performance with GLM-F. We further apply our method to a real-world clinical dataset from the Parkinson's Progression Markers Initiative (PPMI), identifying key clinical, imaging, and CSF biomarkers that distinguish Parkinson's patients with mild cognitive impairment from those with normal cognition at baseline. Our approach achieved comparable predictive performance (87.2% accuracy; AUC = 0.797) to GLM (85.4%, AUC = 0.796), Random Forests (85.4%, AUC = 0.759), but better than LOESS (72.7%, AUC = 0.736) while offering interpretable topological summaries of the disease phenotype. This work establishes Mapper-based TDA as a scalable, interpretable, and powerful alternative to black-box or traditional parametric models for prediction and feature selection in complex, high-dimensional data.

2. Title: Reference-free genotype imputation using a family-aware autoencoder method

Authors: Sichang Wang[1], Xiaowei Wu[2]
Affiliations: [1.] Department of Computer Science, Virginia Tech, [2.] Department of Statistics, Virginia Tech

Abstract: Genotype imputation is crucial in GWAS, enabling the recovery of missing or invalid genotypes and thereby enhancing the scope and accuracy of downstream analyses. Autoencoders, a class of neural networks that compress inputs into latent dimensions and then reconstruct them, can learn data structures and impute missing values. Denoising autoencoders (DAEs), a type of autoencoder, have been applied to genotype imputation.

However, most existing studies using DAEs focus on unrelated individuals and do not account for family structure. Here, we present a family-aware genotype imputation method that leverages trio relationships (father–mother–child). Instead of inputting each individual's SNPs independently, our approach concatenates the SNPs of all three family members and passes them as one input unit into the DAE. Trio-wise imputation yields significant performance improvements compared to individual-wise. Across both simulated datasets generated with Python's msprime library and real data from the Framingham Heart Study, our method consistently improves accuracy, F1 score, and Cohen's Kappa. In both cases, related individuals—originally scrambled in the data matrix—were reorganized into trios before imputation, enabling nontrivial performance improvements. These improvements are consistent across different simulation data at different recombination rates, as well as across different chromosomes in real data. Overall, we demonstrate how DAEs can effectively exploit family structure and linkage disequilibrium to improve genotype imputation, thereby improving their practicality and effectiveness in GWAS applications.

3. Title: Fast and Scalable Computations for the Dynamic ICAR Spatiotemporal Factor Model

Author: Michael Osei Kumi.
Affiliation: Virginia Tech.

Abstract: This study focuses on enhancing the simulation process of Dynamic ICAR Spatiotemporal Factor Models (DIFM) by proposing a new approach for the simulation of the factor loading matrix. This new approach involves sampling from the full conditional distribution of each column of the factor loading matrix, which offers greater flexibility to explore the sparsity of the precision matrix and related structures. Our findings demonstrate that, compared to the existing approach, our new approach reduces the computational complexity from cubic to quadratic time while yielding similar estimates of the components of the factor loadings matrix. We apply our proposed approach to spatio-temporal data of temperature deviations for the various climate divisions in the contiguous United States to compare the computational efficiency of our new approach.

4. Title: Big shells, bigger data: cohort analysis of Chesapeake Bay *Crassostrea virginica* reefs

Authors: Madison Griffin[1], Grace Chiu[1], Roger Mann[1], Melissa Southworth[1]
Affiliations: [1]William & Mary's Batten School of Coastal and Marine Sciences at the Virginia Institute of Marine Science

Abstract: Oysters in Virginia Chesapeake Bay oyster reefs are "age-truncated", possibly due to a combination of historical overfishing, disease epizootics, environmental degradation, and climate change. Oysters may display resilience to environmental stressors, however; the current understanding of oyster lifespan is limited. The Virginia Oyster Stock Assessment and Replenishment Archive (VOSARA), a spatially (222 reefs) and temporally (2003-2023) expansive (more than 2,000,000 individual measurements) dataset of shell lengths (SL, mm), has yet to be examined comprehensively in the context of resilience. We developed a novel method using Gaussian mixture modeling (GMM) to estimate the age groups in each reef using yearly SL data and then link age groups over time to estimate cohorts and their lifespan. Sixty-four reefs (29%) had sufficient data (at least 300 oysters sampled for a minimum of 8 consecutive years) to be considered for this analysis. We fit univariate GMMs for each year (t) and reef (r) to estimate 1) the mean and 80th quantile of shell length for each (r,t)th age group, and 2) the percentage of the (r,t)th population in each age group. We compared the (r,t)th estimates to river-level GMM estimates. We linked age groups across time to infer age cohorts by developing an algorithm that prevents the shrinking of shell length when an (r,t)th group becomes an (r,t+1)th group. Their final lifespan equals the number of years the cohort was found in the data plus its estimated starting age. This method shows promise in identifying oyster cohorts and estimating lifespan solely using SL data. Results show signals of resiliency in almost all river systems: oyster cohorts

live longer and grow larger in the mid-to-late 2010s compared to the early 2000s. Future work includes investigating how climate change and management influence oyster resiliency in Chesapeake Bay.

5. Title: Deep Gaussian Process Poisson Modeling of Large-Scale Satellite Count Data for Computer Model Calibration

Authors: Stephen Barnett.
Affiliations: Virginia Tech

Abstract: The National Aeronautics and Space Administration (NASA) launched the Interstellar Boundary Explorer (IBEX) satellite in 2008 to learn more about the region at the boundary between our solar system and interstellar space, called the heliopause. IBEX detects energetic neutral atoms (ENAs) originating from the heliopause and enables space scientists to estimate their rate of emission throughout the sky. Space scientists have developed a variety of competing theories (with corresponding computer simulations) to explain ENA generation and propagation, and hope to use data collected by IBEX to evaluate the relative strength of evidence for each proposed model. Statistical computer model calibration allows data from IBEX to be paired with computer simulations to accomplish this task. However, the count observations, limited but data-intense simulation runs and field measurements, and non-stationary response surface present some unique challenges. We propose a Vecchia-approximated deep Gaussian process Poisson model for the rate of ENAs. Our model has shown a greater ability to learn the complex, non-stationary mean response surface and provide appropriate uncertainty quantification in different regions of the input space, while maintaining a simpler covariance function. We integrate this model into a novel Markov chain Monte Carlo computer model calibration framework to understand the distribution of the parameters that govern ENA generation.

6. Title: Multicategory Linear Log Odds Calibration Assessment and Recalibration of Probability Predictions

Authors: Amy Vennos, Christopher T. Franck, and Xin Xing
Affiliations: Virginia Tech

Abstract: Machine-generated probability predictions have become increasingly popular in aiding decision-making tasks such as image classification and natural language processing. Calibration measures how well these predictions align with the observed rate of events, and recalibration maps probability predictions to those more in line with observed data. We propose the Multicategory Linear Log Odds (MCLLO) Recalibration function, a recalibration mapping which fills in the gap of many current multicategory recalibrators due to its ability to assess calibration status through hypothesis testing. We demonstrate the effectiveness of our recalibration function through simulations and by applying recalibration on image classification and alligator diet case studies. We compare MCLLO to two comparator recalibration techniques utilizing both our hypothesis test and the existing calibration metric Expected Calibration Error to show how our method works well alone, and in concert with other methods.

7. Title: Optimal Sparse Projection Design for Systems with Treatment Cardinality Constraint

Authors: Kexin Xie[1], Ryan Lekivetz[2], and Xinwei Deng[1]
Affiliation: [1]Department of Statistics, Virginia Tech, [2] JMP Statistical Discovery LLC

Abstract: Modern experimental designs often face the so-called treatment cardinality constraint, which is the constraint on the number of included factors in each treatment. Experiments with such constraints are commonly encountered in engineering simulation, AI system tuning, and large-scale system verification. This calls for the

development of adequate designs to enable statistical efficiency for modeling and analysis within feasible constraints. In this work, we propose an optimal sparse projection (OSP) design for systems with treatment cardinality constraints. We introduce a tailored optimal projection (TOP) criterion that ensures a good space-filling property in subspaces and promotes orthogonality or near-orthogonality among factors. To construct the proposed OSP design, we develop an efficient construction algorithm based on orthogonal arrays and employ parallel-level permutation and expansion techniques to efficiently explore the design space with treatment cardinality constraints. Numerical examples demonstrate the merits of the proposed method.

Keywords: Experimental designs; Space-filling design; Orthogonal arrays; Constraint space; Treatment constraint.

8. Title: Vecchia Approximated Bayesian Heteroskedastic Gaussian Processes

Authors: Parul Patil, Robert B. Gramacy, Cayelan C. Carey, R. Quinn Thomas.
Affiliation: [1]Department of Statistics, Virginia Tech

Abstract: Many computer simulations are stochastic and exhibit input dependent noise. In such situations, heteroskedastic Gaussian processes ({\tt hetGP}s) make ideal surrogates as they estimate a latent, non-constant variance. However, existing hetGP implementations are unable to deal with large simulation campaigns and use point-estimates for all unknown quantities, including latent variances. This limits applicability to small experiments and undercuts uncertainty. We propose a Bayesian hetGP using elliptical slice sampling (ESS) for posterior variance integration, and the Vecchia approximation to circumvent computational bottlenecks. We show good performance for our upgraded hetGP capability, compared to alternatives, on a benchmark example and a motivating corpus of more than 9-million lake temperature simulations.

9. Title: Covariate Selection for RNA-seq Differential Expression Analysis with Hidden Factor Adjustment

Authors: Farzana Noorzahan, Dr. Hyeongseon Jeon, Dr. Yet Nguyen
Affiliation: Department of Mathematics and Statistics, Old dominion University and Department of Mathematics, University of Houston

Abstract: In RNA-seq data analysis, a primary objective is the identification of differentially expressed genes, which are genes that exhibit varying expression levels across different conditions of interest. It is widely known that hidden factors, such as batch effects, can substantially influence the differential expression analysis. Furthermore, apart from the primary factor of interest and unforeseen artifacts, an RNA-seq experiment typically contains multiple measured covariates, some of which may significantly affect gene expression levels, while others may not. Existing methods either address the covariate selection or the unknown artifacts separately. In this study, we will investigate several strategies for dealing with both covariate selection and hidden factors via simulation using a real RNA-seq dataset.

--

10. Title: BOOOM: Loss-Function-Agnostic Black-Box Optimization over Orthonormal Manifolds for Machine Learning and Statistical Inference

Author: Beomchang Kim, Priyam Das.
Affiliations: Department of Biostatistics, Virginia Commonwealth University

Abstract: Optimization over the Stiefel manifold St(p, d), the set of p×d columno rthonormal matrices, arises in statistics, machine learning, and engineering, and often involves objectives which are non-convex/multi-modal, discontinuous, or of non-closed form. Most existing approaches depend on either convex relaxation techniques or derivative-based algorithms tailored to specific objective functions, which restricts their applicability to a broader class of problems. We present BOOOM (Black-box Optimization Over Orthonormal Manifolds), a general-purpose global optimization framework for St(p, d) based on Recursive Modified Pattern Search. BOOOM reparameterizes the manifold via a complete Givens rotation decomposition, proving that any point on St(p, d) can be obtained from a fixed orthonormal base through a finite sequence of rotations, extending the classical Hurwitz result. This mapping transforms constrained manifold optimization into an unconstrained Euclidean search over rotation angles, enabling derivative-free exploration while preserving feasibility exactly at each iteration. We establish global convergence properties under mild conditions, guaranteeing that all accumulation points are globally optimal for the black-box problem. In benchmark studies, BOOOM is shown to outperform state-of-the-art optimization toolboxes for the Stiefel manifold across a range of synthetic and real-world problems, including maximum heterogeneous quadratic form optimization. We further apply BOOOM to a novel supervised PCA problem to identify key metabolites associated with colorectal cancer while balancing out their classification power across healthy and colorectal cancer patients. These results highlight BOOOM as a versatile and powerful tool for orthogonally constrained learning and statistical estimation.

--

11. Title: Anatomical Position and Seat Belt Fit of Pregnant Automobile Occupants: An Exploratory Study

Authors: Lillian Dorathy[1], Keri-Anne Lue[1,] Rachel Jin[2], MD, Michelle Oyen[3], PhD, John Paul Donlon[1], Corina Espelien[1], Jason Forman, PhD[1*], Pavel Chernyavskiy[2], PhD*

Affiliations: [1]Department of Mechanical and Aerospace Engineering, Center for Applied Biomechanics, University of Virginia [2]Department of Public Health Sciences, University of Virginia School of Medicine [3]Department of Biomedical Engineering and Department of Obstetrics and Gynecology, Washington University in St. Louis *Equal Contribution

Abstract: Wearing a seat belt while riding in a vehicle is recommended for everyone, including pregnant individuals. The American College of Obstetricians and Gynecologists (ACOG, 2024) specifically encourages seat belt use as an effective means to protect both the pregnant individual and their fetus during a crash. Both the National Highway Traffic Safety Administration (NHTSA) and ACOG provide guidance on seat belt fit specific to pregnant individuals; however, observations and measurements of actual fit are limited. To address these knowledge gaps, an exploratory study was conducted involving 77 pregnant volunteers who are primarily drivers, where 8 anthropometric measurements (e.g., abdominal circumference) and 22 in-vehicle environment measurements (e.g., distance between the navel and steering wheel) were collected while the volunteers were seated in their own vehicle. A principal component analysis was performed based on 31 variables, comprised of anthropometric, vehicle environment, pregnancy trimester, and body size. Optimal fit was coded by three coders for observed seat belt fit adherence to NHSTA guidelines for pregnant drivers. Here, we report: 1) the number of principal components retained; 2) their biomechanical interpretations; and 3) which principal components are strong predictors of adherence to optimal belt fit. Our results will aid in identifying potential risk factors for poor body position and seat belt fit and suggest strategies for improving belt fit and routine usage through future educational and/or design interventions.

---

12. Title: Predicting Psychopathology Symptom Dimensions via Random Forest Models

Authors: Claudia Clinchard [1], Ashlyn Murphy [1], Brooks Casas [1, 2], and Jungmeen Kim-Spoon [1]
Affiliations: Department of Psychology, Virginia Tech [1], Fralin Biomedical Institute at VTC [2]

Abstract: Childhood adversity (including abuse, neglect, poverty, parental substance use, and more) is a major risk factor for the development of psychopathology. There are many factors that can act as vulnerability or protective factors, such as emotion regulation skills or secure attachment with parents (protective factors) or feelings of guilt and behavioral inhibition (vulnerability factors). The current study utilized 167 participants across eight years (ages 14 to 22). Random forest models were used to account for the random effects between participants and the longitudinal data. The models aimed at predicting internalizing and externalizing symptoms, two distinct psychopathology symptom dimensions. The results demonstrated unique predictors in internalizing (feelings of guilt and behavioral inhibition) and externalizing symptoms (behavioral activation and emotional dysregulation when controlling impulses), with negative affect being a top predictor of both internalizing and externalizing symptoms.

---

13. Title: Modular Jump Gaussian Processes
Author: Anna Flowers
Affiliation: Virginia Tech Department of Statistics

Abstract: Gaussian processes (GPs) furnish accurate nonlinear predictions with well-calibrated uncertainty. However, the typical GP setup has a built-in stationarity assumption, making it ill-suited for modeling data from processes with sudden changes, or "jumps" in the output variable. The "jump GP" (JGP) was developed for modeling data from such processes, combining local GPs and latent "level" variables under a joint inferential framework. But joint modeling can be fraught with difficulty. We aim to simplify by suggesting a more modular setup, eschewing joint inference but retaining the main JGP themes: (a) learning optimal neighborhood sizes that locally respect manifolds of discontinuity; and (b) a new cluster-based (latent) feature to capture regions of distinct output levels on both sides of the manifold. We show that each of (a) and (b) separately leads to dramatic improvements when modeling processes with jumps. In tandem (but without requiring joint inference) that benefit is compounded, as illustrated on real and synthetic benchmark examples from the recent literature.

--

14. **Title: Predicting Health Insurance Claim Denials: A Machine Learning Approach to Identifying Structural Inequities**

Author: Aishwarya Sivasubramanian.
Affiliation: University of Virginia.

Abstract: Health insurance claim denials represent an often overlooked structural barrier to accessing healthcare in the United States. While coverage expansion through the Affordable Care Act has improved insurance rates, quality of coverage remains inequitable, with denial rates ranging from 5% to over 40% across insurers. This study applies interpretable machine learning methods to classify U.S. health insurance plans by their denial risk, leveraging 2024 Public Use Files from the Centers for Medicare and Medicaid Services (CMS). Using Extreme Gradient Boosting (XGBoost), we modeled nearly 2.8 million plan records, integrating features such as business rules, benefit scope, geographic service area, network breadth, and transparency measures. The model achieved a 98.6% accuracy with perfect sensitivity in identifying high-denial plans, ensuring that no risky plans were missed. Interpretability was

achieved using SHAP analysis which highlighted administrative burden, narrow geographic coverage, low appeal overturn rates, and reduced benefit scope as the strongest predictors of denial risk.

Findings also revealed that denials are not random anomalies but predictable structural outcomes of plan design. Restrictive plan features were observed across all metal tiers including Gold and Platinum, challenging prior assumptions that higher cost plans guarantee better access. Subgroup analysis also revealed geographic disparities, with model performance variability pointing to weakness in state-level reporting standards. This research demonstrates the feasibility of using machine learning to uncover systemic inequities in health insurance design. By shifting focus from coverage quantity to coverage quality, the study provides an evidence based framework for regulators and policy makers to ultimately reduce structural barriers to care.

--

15. Title: Geographic Heterogeneities in Group Difference Studies

Author: Yumiao Hui[1*]
Affiliation: [1*]Department of Statistics, University of Virginia.

Abstract: Examining and explaining differences among subpopulations, such as sex and ethnic groups, is a topic of interest in many research fields, including disease risk prediction and prevention. The observed difference in the expectation of the health-related outcome among the two subpopulations can be explained by the different distributions of covariates associated with the outcome. Besides the individual-level covariates, our current research shows that the group difference in the outcome can also be explained by the different spatial distribution of the considered subpopulations. In this paper, we propose an innovative geo-additive model-based Peter-Belson (GGAM-PB) method to consider heterogeneous spatial effects when examining group differences in health, using nationally representative health surveys. We utilize bivariate splines over triangulations to capture the nonparametric spatial effect over the irregular two-dimensional target domain of the population. Our proposed GGAM-PB method accounts for the correlation among geographic areas through spline smoothing. It provides estimates of the group difference that can be explained by the individual-level covariates, spatial effects, and the remainder of the unexplained group difference. We proof the design consistency of the proposed GGAM-PB estimate of the unexplained group difference and provide the corresponding Taylor-Linearization variance estimation under complex survey sample designs. Simulation studies show that the proposed GGAM-PB method overperforms the GLM-PB method, which accounts for the fixed effects of residence locations of the subpopulation members. We apply the proposed method to estimate the difference in the proportion of obtaining women cancer screening among non-Hispanic White and non-Hispanic Black females in the US using data from the Behavioral Risk Factor Surveillance System.

16. Title: Supervised Variational Autoencoder with Mixture-of-Experts Prediction

Authors: Jaeyoung Lee (Presenter), Meimei Liu, Hongxiao Zhu.
Affiliation: Department of Statistics, Virginia Tech.

Abstract: Large-scale datasets, such as images and texts, often exhibit complex heterogeneous structures caused by diverse data sources, intricate experimental designs, or latent subpopulations. Supervised learning from such data is challenging as it requires capturing relevant information from ultra-high-dimensional data while accounting for structural heterogeneity. We propose a unified framework that addresses both challenges simultaneously, facilitating effective feature extraction, structural learning, and robust prediction. The proposed framework employs a supervised Variational Autoencoder (VAE) for both learning and prediction. Specifically, two types of latent variables are learned through the VAE: low-dimensional latent features and a latent stick-breaking process that

characterizes the heterogeneous structure of samples. The latent features reduce the dimensionality of the input data, and the latent stick-breaking process serves as a gating function for mixture-of-experts prediction. This general framework reduces to a supervised VAE when the number of latent clusters is set to one, and to a stick-breaking VAE when both the latent features and response variables are omitted. We demonstrate advantages of the proposed framework by comparing it with supervised VAE and principal component regression in a simulation study and a real data application involving brain tumor images.

--

17. Time-Varying Prior-Inspired Penalization for Multiple Changepoint Detection

Authors: Jonathon Jacobs, Shanshan Chen
Affiliation: Department of Biostatistics, Virginia Commonwealth University

Abstract: As intensive longitudinal time series become increasingly common in health data science, efficient multiple change point (MCP) detection methods are essential in processing the data. Existing methods either locate a single changepoint in pre-defined windows or globally determine the best fitting set of change points given a pre-specified cost assumption. However, few existing methods incorporate prior information regarding the locations of changepoints, especially for the MCP problem. To remedy this, we introduce a time-varying prior-like penalty to incorporate prior information on locations of MCP to an existing MCP detection algorithm – PELT, which approximates the benefits of a Bayesian Monte Carlo framework without its prohibitive computational cost. This penalty enforces structured regularization on the PELT algorithm to discourage spurious changes that do not reflect meaningful shifts in the underlying data-generating process. We demonstrate the robustness of our proposed method by applying it to the sleep-wake cycle detection problem in both simulated and real-world actigraphy data. We show that the time-varying penalty boosted the detection accuracy by further pruning undesired changepoints. Implemented in Rcpp, the proposed method is robust, scalable, and can leverage previously untapped domain knowledge. This algorithm is well-suited for applications with partial prior knowledge on potential change locations, such as biomedical signals, financial market analysis, and environmental sensing.

--

18. Title: Divergence-Minimization for Latent-Structure Models: Theory, Algorithms, and Robust Inference

Authors: Lei Li, Anand N. Vidyashankar
Affiliation: LunarAI LLC, George Mason University

Abstract: We develop a divergence-minimization (DM) framework for estimation in latent-mixture models. By optimizing a residual-adjusted divergence, the DM algorithm recovers EM as a special case and yields robust alternatives through different divergence choices. We establish that the sample objective decreases monotonically along the iterates, leading the DM sequence to stationary points under standard conditions, and that at the population level the operator exhibits local contractivity near the minimizer. Additionally, we verify consistency and asymptotic normality of minimum-divergence estimators, analyze robustness via the residual-adjustment function, and provide a breakdown bound. To address the challenge of determining the number of mixture components, we introduce a penalized divergence criterion with repeated sample splitting. Empirically, DM instantiations based on Hellinger and negative-exponential disparities deliver accurate estimation and remain stable under contamination in mixtures and image-segmentation tasks. The framework clarifies connections to MM and proximal-point methods and offers practical defaults, making DM a drop-in alternative to EM for robust latent-structure inference.

19. **Title: Methods for addressing confounding in the presence of semi-competing risks for CER analysis using multilevel observational data**

Authors: Hyunjae Cho, Hong Zhu
Affiliation: Department of Statistics, University of Virginia & Department of Biostatistics, University of Virginia.

Abstract: Population-based comparative effectiveness research (CER) using observational healthcare data, such as registry, claims, and electronic health record data, are primary research tools for assessing treatment effectiveness in real-world settings. Large observational data, such as registry, claims, and electronic health record data, however, often manifest multiple statistical complexities, necessitating substantial novel methodology development to obtain valid results. Particularly, a "semicompeting risks" situation arises when the observation of non-terminal event (e.g., cancer recurrence) is subject to terminal event (e.g., death), but not vice versa. The analysis of such data is further complicated with multilevel data structure and confounding bias. To address these complexities, we propose a copula-based, hierarchical semi-competing risks regression model framework and utilize propensity score (PS) weighting to estimate average treatment effects (ATEs) on non-terminal and terminal events, adjusting for both observed and potentially unobserved differences in individuals and clusters. We consider three PS models to construct PS weights: a marginal model, a fixed-effect model, and a random-effect model, and examine covariate balance both within clusters and overall by using the standardized mean difference (SMD). The PS weights are then incorporated into a two-stage estimation procedure: In Stage I, we estimate the copula model association parameter and random-effect parameters by a penalized pseudo log-likelihood function, and in Stage II, we estimate regression parameters by non-linear estimating equations. We evaluate the finite-sample performance of the proposed method through extensive simulations, and establish asymptotic properties of proposed estimators. We will apply the proposed method to SEER-Medicare Oropharygeal Cancer data for illustration.

20. Title: Black–Litterman portfolio optimization based on GARCH–EVT–Copula and LSTM models

Authors: Vu Huynh[1] and Bao Quoc Ta[2, 3]

Affiliations: Department of Statistics, Virginia Tech[1]; Department of Mathematics, International University, Ho Chi Minh City, Vietnam[2]; Vietnam National University, Ho Chi Minh City, Vietnam[3]

Abstract: In constructing diversified portfolios, investors may be interested in incorporating some quantifiable views or opinions. The Black–Litterman model is a useful approach to integrate investors' views into the Markowitz allocation model. In this paper, we utilize a deep learning model to estimate the investors' views and use GARCH–EVT–Copula to model the dependence structure between stock market returns in a large portfolio. The findings show that the Black–Litterman model for portfolio optimization based on GARCH–EVT–Copula and LSTM (Long Short-Term Memory) models gives better performance compared with the traditional max-Sharpe and the original Black–Litterman portfolio problems.

--

21. Title : Regression modeling of zero-inflated functional data

Authors : Anbin Rhee and Pang Du
Affiliation: Department of Statistics, Virginia Tech

Abstract: Zero-inflated functional data appear when an excessive number of zeros are recorded for some functional variables due to the threshold of detection limits. To analyze this kind of data we propose a two-part mixed-effects functional regression model. The first part models the probability function of the functional response taking nonzero

values via a mixed-effects functional logistic regression model. The second part models the log-transformed true response function by a mixed-effects functional linear model. We use smoothing splines to estimate both the fixed and random effect functions. The estimation procedures for the two parts are respectively penalized quasi-likelihood and a REML-based EM algorithm. Extensive simulations are presented to evaluate the numerical performance of our method. We also apply the method to a Northwestern ICU study to investigate the relationship between total calcium and albumin measurements in repeated blood tests during each of the multiple ICU visits of a patient. Results show that the proposed approach effectively handles zero inflation while recovering the functional relationship between the variables of interest.

--

22. Title: Analyzing bots' data with a truncated and inflated Poisson model

Authors: Sagnik Chanda*[1] and N. Rao Chaganty[2]
Affiliation: [1]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, USA; [2]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, USA.

Abstract: The Poisson distribution is a well-known model for count data, but in practice, it rarely fits real-world datasets, particularly count data generated by bots. To address this, we examine a truncated and inflated version of the Poisson distribution and develop a new, data-driven method to estimate its parameters. Our estimation method is based on an adjusted method of moments, and in comparisons, it performs better than the usual approaches, such as the method of moments and maximum likelihood. Using a real bot dataset, we demonstrate that our approach yields a significantly better fit. Since this is an empirical study, we also outline the assumptions made in building our model.

--

23. Title: Phase Angle Modeling with Wrapped Gaussian Processes

Authors: Andrew Cooper
Affiliation: Department of Statistics, Virginia Tech

Abstract: Angular data is common in fields where there is a directional component present. One popular method of modeling an angular response is with a “wrapped'' model, in which a common distribution in linear space is wrapped around the unit circle. Recently, wrapped modeling has been extended to the Gaussian Process (GP). Inference on the latent “unwrapped'' GP is difficult as it is unknown whether each observation wrapped around the unit circle and in what direction. These observation-specific “wrapping numbers'' must be estimated, along with GP mean and covariance parameters, which can be tricky due to their large size and infinite support. We propose an Elliptical Slice Sampling (ESS) algorithm for generating posterior samples of wrapping numbers. We utilize this algorithm in a Gibbs scheme to perform fully Bayesian inference for the wrapped GP (WGP) setting. We apply our methods to an RFID tag localization example, in modeling the phase-frequency relationship of radio frequency signals can help to infer tag distance.

--

24. Title: A Comparative Review of Some Internal Validation Techniques for Determining the Number of Clusters in Numeric Data

Authors: Emmanuel Nartey[1*], Carl Lee2, Felix Famoye[2]
Affiliations: [1]Center for Biostatistics and Health Data Science, Virginia Tech, [2]Department of Statistics, Actuarial and Data Sciences; Central Michigan University

Abstract: Determining the optimal number of clusters is a persistent challenge in cluster analysis. Internal cluster validation indices (iCVIs) provide a data-driven approach, yet the large number of available indices and their variable performance across clustering algorithms complicate selection. This study systematically compared iCVIs across diverse data conditions to identify effective algorithm-index combinations. Using a 7-factor, 2-level factorial design with simulated datasets, we evaluated how noise, dimensionality, cluster separation, outliers, and distributional form affect performance. Results show that iCVIs perform most reliably when noise is minimized, clusters are well-separated, and distributions approximate Gaussian structure. Among all combinations tested, the Cubic Clustering Criterion (CCC) with Ward's D2 algorithm consistently yielded the most accurate and computationally efficient solutions. These findings emphasize the need for thorough data preprocessing and highlight CCC combined with Ward's D2 as a robust choice for numeric datasets. More broadly, our factorial design framework offers a structured way to assess clustering strategies and can guide future development of validation methods.

--

25. Title: A Graph-Fused LASSO Regression Technique for Analyzing Real Estate Markets with Spatial Heterogeneity Effects

Authors: Obed Amo.

Affiliations: Old Dominion University

Abstract: Residential real estate prices are influenced not only by observable property characteristics but also by unobserved neighborhood heterogeneity. Traditional regression models, including penalized approaches such as LASSO, often fail to capture this type of spatial dependence. We introduce a graph-fused LASSO regression model that combines property-level covariates with the spatial structure represented through an adjacency graph. By penalizing the differences in estimated neighborhood effects, the method recovers piecewise-constant spatial patterns while preserving flexibility in feature contributions. Applied to housing transaction data from Hampton Roads, the model uncovers meaningful spatial clusters and delivers improved predictive accuracy. Beyond the real estate domain, this approach offers a general framework for modeling spatially structured heterogeneity, providing a useful tool for urban analysis, policy evaluation, and other domains where geographic context is of essence. This is joint work with Michael Pokojovy, Simon Stevenson and Lei Zhang.

--

26. Title: On the Convergence Rate of the Least Trimmed Squares Algorithm

Authors: Samit Ghosh.
Affiliations: Old Dominion University

Abstract: The least trimmed squares (LTS) approach is a popular alternative to ordinary least squares (OLS) commonly used to robustly fit linear regression models in the presence of outliers or other model violations. While a closed-form solution exists for the quadratic optimization problem underlying the OLS estimator, the mixed integer-continuous optimization problem associated with the LTS estimator is typically solved by iteratively applying the so-called concentration or C-step to a set of properly selected warm starts. Although the C-step is known to monotonically reduce the objective function leading to convergence in a finite number of steps, no error estimation mechanisms presently exist in the literature. By studying an equivalent formulation of the problem over a convex domain, we prove that the conventional LTS iteration is a special case of the well-known Frank-Wolfe gradient descent method. This furnishes an O(1/t) convergence of the LTS algorithm to a local minimum of the trimmed sum of squares objective function with respect to a suitable Bregman divergence and provides a convenient error quantification tool. This is joint work with Michael Pokojovy and Guohui Song.

--

27. Title: A Deep Learning Approach to Optimal Control for Runge–Kutta Discretizations of Dynamical Systems

Authors: Lokanshu Malur
Affiliations: Old Dominion University

Abstract: Abstract: Runge–Kutta schemes are widely employed in numerical optimal control of dynamical systems governed by ordinary differential equations, due to their attractive convergence properties. Algorithmic implementations, however, remain computationally expensive. We adopt the discretize-then-optimize framework and represent the control policy for the discrete system as a feedback control given by an artificial neural network (ANN). This reformulation reduces the objective to a function that solely depends on (typically) low-dimensional ANN weights. Leveraging automatic differentiation for efficient gradient computation, the search for the optimal policy amounts then to training the ANN using stochastic gradient descent. Under suitable smoothness assumptions, the integration of a Runge–Kutta scheme into the training loop results in improved convergence rates for the combined method. Using the inverted pendulum as an archetypal Hamiltonian system commonly encountered in physics and engineering, we showcase our approach and benchmark it against the standard first-order explicit Euler scheme. This is joint work with Nadav Azran, Yaacov Kopeliovich and Michael Pokojovy.

--

28. Title: Bayesian Multilevel Network Recovery Selection

Authors: Mohamed Salem[1] and Inyoung Kim[1*]
Affiliations: [1] Department of Statistics, Virginia Polytechnic Institute and State University.

Abstract: Variable selection and network estimation have been popular tools for identifying key variables associated with a response variable of interest in settings involving non-negligible dependency structures among variables. However, the ability to identify relevant variables in a high-dimensional setting while accounting for conditional dependencies within a multilevel structure under a nonadditive model is still limited. Hence, in this paper, we examine multilevel network recovery selection under a two-level structure in which higher-level variables contain lower-level variables nested within them. Due to the dependency structure, variables work together to accomplish certain tasks at both levels. Our main interest is to simultaneously explore variable selection and identify dependency structures among both higher and lower-level variables under a nonadditive model framework. We develop a multi-level nonparametric kernel machine approach with a newly proposed multilevel Ising spike-slab prior, utilizing Markov-chain Monte Carlo and variational Bayes inference to identify multi-level variables and jointly build the network. The variational inference approach is novel in utilizing the sampled dependency structure as the observed variable rather than the response. In addition to the variable selection and network recovery capabilities, our approach can produce both mean and quantile estimations of the original response variable of interest. We demonstrate the advantages of our approach using simulation studies and a genetic pathway-based analysis.

--

29. Title: Exploratory analysis of breast cancer genomics data

Authors: Sabikunnahar Jashe[1], Sinjini Sikdar[1]

Affiliations: [1]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529

Abstract: Breast cancer remains one of the most diagnosed cancers worldwide. Gene expression analysis provides valuable insights into the biological processes within tissue quantifying gene activity levels. By applying machine learning algorithms to genomic data, it might be possible to identify novel biomarkers that can improve our understanding of disease mechanisms and enhance the prediction of patient outcomes. This study utilized a publicly available dataset comprising 1,904 breast cancer patients, of whom 1103 were deceased and the remaining alive. For each patient, the dataset includes expression values for 489 genes along with several clinical features. Our objective was to identify genes that are differentially expressed between deceased and surviving breast cancer patients. Using an 80-20 training-testing split, we applied logistic Lasso regression model on the training set to identify genes associated with breast cancer status (deceased or alive) and validated the findings in the testing set. Pathway enrichment analysis of these genes revealed several biologically relevant pathways, including PI3k-Akt signaling pathway, MAPK signaling pathway, and cell cycle. This study implemented an optimization-based approach to identify the optimal clustering algorithm and the number of clusters to group the testing samples using the Lasso-identified genes. The resulting clusters showed meaningful separation and demonstrated substantial overlap with the patients' breast cancer status. Finally, a Lasso score was constructed for each patient in the testing set using the selected genes and their coefficients from the Lasso regression model. The score was used to identify potential interactions with clinical features in relation to breast cancer status. Notably, significant interactions were observed with the patients' tumor stage and breast cancer subtype, suggesting that the Lasso score may capture clinically relevant variations in disease progression and type.

--

30. Title: Longitudinal Modeling of Brain Volume and Cognitive Decline in Older Adults using OASIS Data

Authors: Dinesha Arachchige, Sandipan Dutta
Affiliation: Old Dominion University

Abstract: Over the past several decades, cognitive decline and dementia diagnoses have been on a rising trend among older adults. A key step toward developing effective treatments is understanding how these conditions progress over time. Longitudinal neuroimaging studies, which track changes within individuals, are well-suited for this purpose. This project uses data from the Open Access Series of Imaging Studies (OASIS), which includes brain scans and related information from 150 participants aged 60 to 96. We investigated the longitudinal trajectories of two key variables: normalized whole brain volume (nWBV) and Mini-Mental State Examination (MMSE) scores. Our goal was to determine if these variables showed different patterns of change between individuals with dementia and those without. We used both marginal and conditional modeling approaches, adjusting for important factors like age and gender. Several correlation structures were compared during the modeling process, and necessary adjustments were made to account for the highly skewed nature of the MMSE scores. Final models were selected using popular criteria such as AIC, BIC, and QIC. Our results indicate that individuals with dementia have a significantly faster decline in nWBV and lower MMSE scores compared to non-demented individuals. Overall, this study demonstrates that advanced longitudinal modeling can reveal important patterns in brain changes and cognitive decline, providing clearer insights into the progression of dementia.

--

31. Title: Copula-Based Models for Multivariate Count Time Series with Applications to Hurricane Data.

Authors: Md Iqbal Hossain[1], Norou Diawara[2.]

Affiliations: Department of Mathematics and Statistics, Old Dominion University (ODU)[1,2]

Abstract: In research involving the environment, health, and economy where events take place over time, count time series (CTS) data naturally emerge. While the CTS is analyzed using Poisson or negative binomial distributions, models must account for both within and between correlation in the multivariate setting Copulas offer an effective

way to handle non-Gaussian data and nonlinear interactions. To verify asymptotic efficiency and robustness under different sample sizes, we use Gaussian copula families to create likelihood-based inference for the suggested models and assess performance using comprehensive simulations. Our methodology is applied to annual hurricane counts from six ocean basins (1980–2023). The copula approach reveals dependences in the basin's associations. These results demonstrate how copula models can improve multivariate understanding and provide fresh perspectives on intricate dependence patterns in real CTS data with sometimes positive and negative correlations.

32. **Title: Incorporating Spatial Structure Improves Violent Crime Prediction in Austin: A Comparison of Bayesian Spatial and Machine Learning Models**

Authors: Dylan Steberg, Grant Drawve, Jyotishka Datta
Affiliations: Department of Statistics, Virginia Tech.

Abstract: Understanding the factors behind violent crime is critical for developing effective prevention and intervention strategies. In criminology, machine learning methods such as random forests and gradient boosting have been widely applied to predict crime patterns (e.g., Wheeler et al., 2021). However, these methods often neglect the spatial structure inherent in crime data, where events cluster geographically rather than occurring independently. We examine violent crimes (aggravated assaults and homicides) in Austin, Texas from 2021 to 2023 within 330 by 330 feet (0.1 km) grid cells using a combination of built environment data (e.g., number of bars, grocery stores, laundromats) and sociodemographic indicators from the American Community Survey (e.g., population density, percent male under 18). We first assess local indicators of spatial association (LISA) to detect hotspots and then compare and contrast the performance of traditional machine learning models with Bayesian spatial models incorporating the conditional autoregressive prior (Besag et al., 1991). Finally, we explore variable importance across models to identify factors most strongly associated with violent crimes. Our findings highlight the advantages of explicitly incorporating spatial dependence into crime prediction models and provide insights into the drivers of violent crime in urban settings.

33. Title: Parameter Estimation and Influence Analysis in Matrix Variate Non-Gaussian Models

Authors: Samuel Soon[1] (Presenter), Dipankar Bandyopadhyay[1,] Qingyang Liu[2], Victor Lachos[3]
Affiliations: [1]Virginia Commonwealth University, [2]University of Wisconson-Madison, [3]University of Connecticut.

Abstract: Clinical attachment loss (CAL) and probing pocket depth (PPD), the two most important endpoints assessing periodontal disease (PD) status and progression, are collected at multiple tooth sites within subjects. The two measurements are known to be correlated and exhibit non-Gaussian behavior characterized by skewness and heavy tails. Despite this, analysis of PD data is often performed using Gaussian random effects without accounting for response-wise covariance. Although parametric regression models exist to handle the apparent non-Gaussianity of the vectorized multivariate responses, we consider the computationally elegant alternative of matrix-variate non-Gaussian (MVNG) regression, which generalizes a multivariate density to matrix-valued random variables and accommodates skewness through additional parameters. We explore an MVNG linear regression framework for the variance-gamma and normal-inverse Gaussian distributions, which reduces the need for numerical approximations and allows for accurate representations of skewness, kurtosis, and covariance parameters. We circumvent the computational inefficiencies of classical maximum likelihood estimation through an expectation-conditional maximization (ECM) algorithm, and provide model-specific case-deletion influence diagnostics to identify influential subjects and illustrate the robustness of MVNG models. We present simulation studies to study the finite-sample behavior of our model. Furthermore, application to a dataset generated from a clinical PD study reveals the advantages of our proposal, in light of existing alternatives.