New Frontiers in Reliability and Risk Analysis:
A Tribute to Nozer D. Singpurwalla
October 13-14, 2023

Book of Abstracts

Sponsored By:

Research Symposium

The research symposium "New Frontiers in Reliability and Risk Analysis: A Tribute to Nozer D. Singpurwalla" will be held at the George Washington University during October 13-14, 2023.

The objective of the symposium is to honor the life and work of Nozer D. Singpurwalla by bringing together researchers and scholars to present and discuss current and emerging trends in a variety of disciplines, including reliability, risk analysis, Bayesian methods, time series analysis, decision theory, and foundations of statistics.

The symposium will feature plenary talks, invited talks as well as contributed talks and poster presentations. Limited travel support will be available for students participating in the poster session. More information about the symposium can be found at the web site: https://statistics.columbian.gwu.edu/nds2023.

Organizing Committee

Kimberly Sellers, Department of Statistics, North Carolina State University.

Refik Soyer, Department of Decision Sciences, George Washington University.

Thomas A. Mazzuchi, Department of Engineering Management and Systems Engineering, George Washington University.

Subrata Kundu, Department of Statistics, George Washington University.

Alyson Wilson, Department of Statistics, North Carolina State University.

Alan Wan, Department of Management Sciences, City University of Hong Kong.

Simon Wilson, Department of Statistics, Trinity College.

Joshua Landon, Department of Statistics, George Washington University.

Xiaoke Zhang, Department of Statistics, George Washington University.

Plenary Sessions

Encounters with Imprecise Probability (IP)

James Berger
Duke University

Almost all real-world probabilities are imprecise; for instance, it would be ridiculous for a weather forecaster to say "the probability of rain tomorrow is 0.364927401482906398710 4;" but, logically, it is then also ridiculous for the weather forecaster to say that the probability of rain tomorrow is 0.4. (The weather forecaster is not being real-world ridiculous here, of course.) The Society for Imprecise Probability consists of many scientists from a variety of disciplines that try to address IP. Four applications of IP reasoning, in which I was involved, will be discussed in this talk: (i) The problem of interval probabilities, which is of great interest in reliability and high-energy physics; (ii) Correcting the p-value problem in science; (iii) Providing a sound analysis for any normal hierarchical model; (iv) Tackling Uncertainty Quantification, the intersection of mathematical modeling of real-world processes and data. Needless to say, Nozer was very involved with this issue.

Start-Over-After-Preempts (SOAP) Repair

Jayaram Sethuraman
Florida State University

We have experienced situations when somebody higher up interrupts the service of our job and we have to start all over again after that interruption has been serviced. This leads us to the **S**tart-**O**ver-**A**fter-**P**reempts (**SOAP**) repair model. Here a repair job has entered the queue for service but a preempt comes in and stops the repair. It has to wait for the preempt service to finish and service on the original job has to start afresh. We study the time it takes to complete repair on a job under these circumstances when the repair could have finished in **b** units of un-interrupted time. We also study the optimal order of placing two jobs that require **b**, **a** units of un-interrupted time to complete, when $\mathbf{b < a}$. The answers will depend on the distributions of the repair times interlaced by preempts and the duration of preempts. A new stochastic process with independent increments appears in this study. There are many open questions related to this model which will be of interest to graduate students.

Invited Sessions

Adversarial Risk Analysis

David Banks
Duke University

Adversarial Risk Analysis (ARA) is a decision-theoretic alternative to game theory, applicable to corporate competition, auctions, and counterterrorism. Nozer played an essential role in getting the first paper on the topic published. In ARA, one builds a model for the strategic decision making of one's opponent(s), and then places subjective Bayesian distributions over unknown quantities. This structure enables the analyst to compartmentalize distinct kinds of uncertainty. Within this framework one can use standard Bayesian techniques to develop a probability distribution over the actions of the opponent. Given this distribution, the decision theorist chooses the action that maximizes expected utility.

Error Bounds in Applied Probability Models: Exponential and Geometric Approximations

Mark Brown
Columbia University

Frequently in probability work simple approximations are sought for mathematically intractable probability distributions. Limit theorems often supply the approximating distribution, but what is really needed are error bounds for fixed n or t. In this talk I'll discuss some of my work over the years in error bounds for exponential and geometric distribution approximations. Points of interest include:

1) The waiting time for patterns in multinomial trials.
2) The first passage time to a set, A, for time reversible Markov chains.
3) The approximate exponentiality of geometric convolutions, with various applications.
4) The reliability of repairable systems.
5) Hazard function based bounds and inequalities.

How Can Subjective Bayesians Avoid Solipsism?

Joseph Kadane
Carnegie Mellon University

Solipsism is an extreme version of subjectivity. It holds that the only thing I can be sure of is my own mind. This line of thinking was articulated by Descartes' famous "cogito ergo sum" (I think, therefore I am). He did not propose "cogitamus ergo sumus" (we think, therefore we are). Solipsism is unpopular among philosophers, but they have not been able to refute it. However, why would a solipsist seek to speak or write, since (by hypothesis) there's nobody out there to address? A subjective Bayesian who is content only to state a personal model (likelihood and prior) veers close to solipsism. Instead, I propose that subjectivists give reasons for the choices made, reasons other than "that's what I think". By giving arguments, such a Bayesian is seeking to persuade others that the views expressed are interesting and perhaps worthy of study.

Prediction with Confidence – General Framework for Predictive Inference

Regina Liu
Rutgers University

We present a general framework for prediction in which a prediction is in the form of a distribution function, called 'predictive distribution function'. This predictive distribution function is well suited for prescribing the notion of confidence under the frequentist interpretation and providing meaningful answers for prediction-related questions. Its very form of a distribution function also makes itself a useful tool for quantifying uncertainty in prediction. This general framework is formulated and illustrated using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity to serve as a common platform for directly connecting the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. We discuss the theory underlying the CD-based predictive distribution and related efficiency and optimality. We also propose a simple yet broadly applicable Monte-Carlo algorithm for implementing the proposed approach. This concrete algorithm together with the proposed definition and associated theoretical development provide a comprehensive statistical inference framework for prediction. Finally, the approach is illustrated by simulation studies and a real project on predicting the application submissions to a government agency. The latter shows the applicability of the proposed approach to even dependent data settings.

Adaptive Multi-Arm Multi-Stage Designs: A Comparison of Methods

Cyrus R. Mehta
Cytel Inc.

A multi-arm multi-stage (MAMS) design is a type of clinical trial design that involves testing multiple treatment arms against a common control in a single trial in multiple stages, separated by unblinded looks at the data with the option to stop early for efficacy or futility, drop non-performing arms, or increase the sample size. Two different approaches for MAMS designs can be distinguished; stage-wise and cumulative MAMS. Both methods allow strong control of the family wise error rate. In stage-wise MAMS designs this is achieved by converging the raw p-values observed at each stage into a single multiplicity adjusted p-value, and combining these stage wise multiplicity adjusted p-values into a single test statistic via the inverse normal combination function. In the more recently developed cumulative MAMS design, a separate cumulative test statistic is constructed for each treatment vs control comparison and efficacy can be claimed if at least one of these test statistics crosses a multiplicity adjusted group sequential boundary. The operating characteristics of the two methods are compared under different distributional assumptions and rules for dropping arms.

An Adversarial Risk Analysis Framework for the Software Release Problem

Fabrizio Ruggeri
CNR-IMATI

A major issue in software engineering is the decision of when to release a software product to the market. This problem is complex due to, among other things, the uncertainty surrounding the software quality and its faults, the various costs involved, and the presence of competitors. A general adversarial risk analysis framework is proposed to support a software developer in deciding when to release a product and showcased with an example.

The Information and Inferences under Different Parameterization of a Likelihood Model

Ehsan S. Soofi
University of Wisconsin-Milwaukee

In his critique of Bernardo's reference analysis, Lindley (1997) states his objection to default priors because they are based on the model. He points out that default priors lead to different priors that lead to different inferences about the same parameter for two different statisticians who have the same prior knowledge, but that they have different data sets leading to different models. We consider the utility of learning about a parameter from the same data but using different parametrizations of the same likelihood model. Following Abel and Singpurwalla (1992) we measure utility by the observed sample information and show that conjugate priors of different parametrizations of the same likelihood model may also lead to different information and inferences. The maximum entropy models provide the setting for this study. We focus on comparing the observed information about the moment parameters that determine the maximum entropy likelihood models, the implied maximum entropy model parameters, and other parameters of interest. Probes of important maximum entropy exponential family models illustrate that the conjugate priors for different parametrizations can lead to the same or different information, Bayes estimates, and posterior precision depending on the parameter of interest and hyperparameters.

Joint Modeling of Geometric Features of Longitudinal Process and Discrete Survival Time Measured on Nested Timescales with Dynamic Risk Prediction: An Application to Fecundity Studies

Rajeshwari Sundaram
National Institutes of Health

In biomedical studies, longitudinal processes are collected till time-to-event, sometimes on nested timescales (example, days within months). Most of the literature in joint modeling of longitudinal and time-to-event data has focused on modeling the mean or dispersion of the longitudinal process with the hazard for time-to-event. However, based on the motivating studies, it may be of interest to investigate how the cycle-level geometric features (such as the curvature, location and height of a peak), of a cyclical longitudinal process is associated with the time-to-event being studied. We propose a shared parameter joint model for a cyclical longitudinal process and a discrete survival time, measured on nested timescales, where the cycle-varying geometric feature is modeled through a linear mixed effects model and a proportional hazards model for the discrete survival time. The proposed approach allows for prediction of survival probabilities for future subjects based on their available longitudinal measurements. This allows us to do dynamic risk prediction for time-to-event. Our proposed model and approach are illustrated through simulation and analysis of Stress and Time-to-Pregnancy, a component of Oxford Conception Study. A joint modeling approach was used to assess whether the cycle-specific geometric features of the lutenizing hormone measurements, such as its peak or its curvature, are associated with time-to-pregnancy (TTP). Time permitting, we will discuss further extensions to incorporate multivariate geometric features.

Using "Expert Opinion" and Other Forms of Predictive Information in Personal Inferences and Decisions

Mike West
Duke University

Nozer Singpurwalla's philosophical evolution in the 1980s was partly linked to his interactions with Dennis Lindley and, through Lindley, with others in the mainstream Bayesian statistical community. My own personal interactions with Singpurwalla began in 1986 at a time when he and Lindley (and others) were concerned with questions of how a Bayesian integrates information arising from expressed, elicited or otherwise observed beliefs/opinions of others. Singpurwalla's work with Lindley, and in extensions into applied methodology in reliability (e.g., Lindley & NDS, JASA 1986; NDS JASA 1988; NDS & Son IEEE Reliability 1988; Campódonica & NDS, IEEE Trans Software Eng, 1993; among others) helped to define the core foundations of formal Bayesian approaches to such problems. Fast-forward to October 2023: I will trace one key theme in what has emerged since to define an encompassing, foundational subjective Bayesian framework for evaluating, calibrating, comparing and combining varied sources of information arising from multiple agents, agencies or models—linking into core questions of model uncertainty analysis. Bayesian predictive synthesis (BPS) builds on some of those earlier 1980s developments to provide a Bayesian with formal models and methods for conditioning on information from such varied sources for their own inference and forecasting problems. BPS expands traditional Bayesian thinking and, in particular, includes Bayesian model averaging for pooling multiple posterior or predictive distributions as a special case. The more recent development of Bayesian predictive decision synthesis (BPDS) takes this further to complete the inference-decision circle. BPDS allows a Bayesian to explicitly integrate decision goals into analysis and weighting of information from multiple models, forecasters, forecast agencies, and other-- often competing-- sources. This talk will touch on the ideas of BPDS at a high-level, and note applied foci in areas including optimal design for prediction and control, forecasting for financial portfolio decisions, and forecasting to advise macro-economic policy.

Special Sessions

Bayesian Modeling of Power Outages at Multiple Locations under a Common Environment

Atilla Ay
James Madison University

In this paper, our objective is to model power outages and their consequences in multiple locations by measuring the total number of households affected in each location. In so doing, we assume that the power distribution systems in these locations operate under a common environment. For each location, we consider a compound Poisson process to model the number of affected households, whose jump rate (outage rate) and jump size (the number of households affected by each outage) change with the evolving state of the common environment, which is modeled as a latent Markov process. Therefore, we refer to this model as a Markov modulated multiple compound Poisson processes model. We use both simulated and actual power outage data from five counties in northern Virginia to demonstrate the proposed model and the methodology.

Robust Estimation and Variable Selection for Linear Mixed Models

Olivia Atutey
University of South Alabama

A robust variable selection procedure based on BIC is proposed to select and estimate fixed effects in linear mixed models simultaneously. Robustness in the presence of outliers and sparse estimation of fixed effects are achieved by adapting Jaeckel's dispersion function based on the absolute values of the ranks of the residuals for a given choice of score function with hyperbolic tangent penalty function. The finite sample behavior of the newly proposed rank-based penalty estimator for variable selection and fixed effects estimation is evaluated via simulation studies under heavy-tailed error distributions and outlying observations. In addition, a dataset on Medicare spending on a cohort of colon cancer patients is used to illustrate the proposed methodology further.

Waiting Time Distributions of Exchangeable Events with Application to Polya Process

Srinivasan Balaji
The George Washington University

In this talk we will discuss De Finetti's theorem for exchangeable events and the distribution and moments of waiting times for exchangeable sequences. As an application, we will derive the mean and variance of waiting times generated by the Polya-Eggenberger process.

Reliability Model for Systems with Clusters of Degrading Components Based on a Superposition of Gamma Processes

David Coit
Rutgers University

Two system reliability models are presented for complex multi-component systems when the component degradation paths are stochastically dependent, and components degrade as clusters that behave probabilistically similarly. Degradation paths of components within a complex system are often assumed to be independent, which is a common assumption, and it is often a realistic assumption. However, these new models pertain to the cases when degradation paths of different components are dependent in such a way that they can be classified into specific groups, called clusters. Degradation paths of different components can be considered as different groupings that correspond to different operational conditions and/or stresses. Extensions to gamma process models are used to model the stochastic processes of individual component deterioration. The first model is based on random effects gamma processes for each component. The scale parameter for each component is a function of a set of random variables, one for each cluster, which are common for all components within the system. Sensitivity factors then control the relative degree to which components belong to a certain cluster. The second model begins with a set of different independent gamma processes for each component and each cluster, and then the component degradation model is a sum of the individual gamma process for each component and a linear combination of the cluster gamma processes. This creates a system model with dependent component degradation paths that can be grouped into clusters. Both models are demonstrated with maintenance optimization examples.

Bayesian Quantile Regression with Subset Selection: A Posterior Summarization Perspective

Joseph Feldman
Duke University

Quantile regression is a powerful tool for inferring how covariates affect specific percentiles of the response distribution. Existing methods either separately estimate conditional quantiles for each quantile of interest or estimate the entire conditional distribution using semi- or non-parametric models. The former often produce severely inadequate models for real data and do not share information across quantiles, while the latter are characterized by complex and constrained models that are difficult to interpret and computationally inefficient. Further, neither approach is well-suited for quantile-specific subset selection. Instead, we pose the fundamental problems of linear quantile estimation, uncertainty quantification, and subset selection from a Bayesian decision analysis perspective. Our approach is compatible with any Bayesian regression model, which should be curated to capture salient features of the data-generating process. We then derive optimal (in a decision theory sense) and interpretable linear estimates and uncertainty quantification for each model-based conditional quantile. Most uniquely, our approach introduces a quantile-focused squared error loss, which enables efficient, closed-form computing and maintains a close relationship with Wasserstein distance between (posterior) densities. This feature also unlocks quantile-specific subset search and selection procedures, providing parsimonious yet predictive explanations for any quantile of interest, along with novel measures of variable importance. In an extensive simulation study, our method demonstrates substantial gains over frequentist and Bayesian alternatives in terms of prediction and selection, while also providing coherent estimates which are robust to the quantile. Deploying our techniques in a quantile regression analysis of North Carolina children end of grade test scores, we uncover several predictors with heterogeneous effects across the response distribution which highlight inequities in childhood educational achievement.

Bayesian Inference for Two Nonstandard Flexible Families of Bivariate Kumaraswamy Models: Theory and Applications

Indranil Ghosh
University of North Carolina, Wilmington

Arnold and Ghosh (2017a, 2017b) have proposed a broad spectrum of bivariate Kumaraswamy (henceforth, KW, in short) distributions involving conditional specification, conditional survival specification, and starting from the Arnold-Ng (2011) eight parameter bivariate beta model. In addition, copula-based construction of bivariate KW models was considered. Included among the models that they dis- cussed were the Olkin-Trikalinos (henceforth, in short, OT-BK) and the Ghosh-BK (henceforth, in short, G-BK) model. These two models can accommodate both positive and negative correlation under certain parametric restrictions. However, this comes at the expense of dealing with a density that is mathematically intractable. We focus our attention on estimation in the 4 and 5 parameter OT-BK and the G-BK models respectively using a Bayesian approach. A general framework based on approximate Bayesian computation methodology is proposed and studied in this article. In particular, the choice of priors must be such that they satisfy the parameter constraints for these models. We conduct simulation studies for both the models under a wide selection of priors. For illustrative purposes, a real data set has been re-analyzed.

What Information Theory Brings to Decision Theory and Inference with Complex Data

Amos Golan
American University

The available information is usually too complex, insufficient and imperfect to deliver a unique solution for most economic modeling and inference problems. Problems with multiple solutions are called under-determined, or partially identified. Information Theory within a constrained optimization setup provides a way to deal with complexity under such deep uncertainty. It provides us with a way to sort and rank solutions and then choose the one that satisfies our desired properties. As such, it provides us with a different way of thinking about solving complex problems and a way to nest models in terms of the information and decision criteria they use. It also provides new insights into basic modeling and allows us to solve inference problems that cannot be solved with conventional methods without imposing additional structure or assumptions. Though Information-Theoretic inference provides us with a general framework for inference (I call it, info-metrics), the exact specification is problem-specific. In this talk I will briefly summarize the basic idea of info-metrics via a number of graphical representations of the theory and will then provide a number of examples. I will also discuss the way some other traditional approaches fit within that framework (such as partial identification and misspecification) and some of the benefits of combining classical and information-theoretic econometric modeling.

A Bayesian Approach to Network Classification

Shramista Guha
Texas A&M University

We propose a novel Bayesian binary classification framework for networks with labeled nodes. Our approach is motivated by applications in brain connectome studies, where the overarching goal is to identify both regions of interest (ROIs) in the brain and connections between ROIs that influence how study subjects are classified. We develop a binary logistic regression framework with the network as the predictor, and model the associated network coefficient using a novel class of global-local network shrinkage priors. We perform a theoretical analysis of a member of this class of priors (which we call the Network Lasso Prior) and show asymptotically correct classification of networks even when the number of network edges grows faster than the sample size. Two representative members from this class of priors, the Network Lasso prior and the Network Horseshoe prior, are implemented using an efficient Markov Chain Monte Carlo algorithm, and empirically evaluated through simulation studies and the analysis of a real brain connectome dataset.

Resale-as-a-Service: The Impact of Second-Hand Platforms on Fashion Goods Retailer

Yuan Guo
The George Washington University

We examine how a retailer selling fashion apparel should account for a second-hand market amid changing fashion trends and uncertain customer taste. We propose a customer decision model that incorporates both customer purchase decisions (i.e., purchasing a new product, a used product, or no product) as well as their selling decisions (i.e., selling a used product). Customer taste for product design is uncertain and hence products released by the retailer have random popularity in the market. We first discuss how uncertainty in customer taste can affect the retailer's product design and production decisions when competing with the second-hand platforms. Then, we examine the potential benefit for the retailer from cooperating with a platform under Resale-as-a-Service.

Imputing Missing Values in [0,1] Data Using Endpoint-Heterogenous Beta Regression Modeling

Eugene D. Hahn
Salisbury University

Endpoint-heterogenous beta regression is a new method for modeling [0,1] data. If we know a priori that the endpoints are not missing, we illustrate that endpoint-heterogenous beta regression can have attractive characteristics for imputing missing values versus alternative options for modeling $y\in [0,1]$ data. We describe this process using a combination of frequentist and Bayesian methods in the context of a data set that is widely used in the fields of international business, marketing and organizational behavior.

An Overview of the NIST AI Risk Management Framework

Patrick Hall
The George Washington University

In response to the broad adoption of artificial intelligence and machine learning (AI/ML) technologies across the private and public sectors in the United States (US), the US National Institute of Standards and Technology (NIST) released version 1.0 of the AI Risk Management Framework (RMF) in January 2023. The NIST AI RMF defines an AI lifecycle, the human actors, the desiderata for trustworthiness in AI/ML systems, and the organizational functions necessary to manage risk for sophisticated AI/ML systems. This presentation will overview these aspects of the AI RMF, with a focus on the characteristics of trustworthiness — validity, reliability, safety, security, resiliency, accountability, transparency, explainability, interpretability, privacy, and bias management — and the four risk management functions — govern, map, measure and manage. This presentation will also address common risk-aware approaches in the current practice of AI/ML, such as interpretable models, post hoc explanation, bias testing, and red teaming. Attendees will leave this talk with a high-level understanding of the AI RMF and a few technical pointers for using the RMF to manage risk in contemporary AI/ML systems.

Bernoulli Convolution of the Depth of Nodes in Recursive trees with Generalized Affinities

Hosam Mahmoud
The George Washington University

It has been recently reported that the depth of insertion of nodes in recursive trees grown with certain node affinities can be represented as a sum of independent Bernoulli random variables. In this investigation, we demonstrate that the presence of a Bernoulli convolution covers a much broader range of affinities beyond the known ones. Further, we pursue sufficient conditions to have associated normal regimes and compare them to approximations via Poisson distributions. We give several illustrative examples, where we look at the rates of convergence too. The highlight of our examples are the regularly varying sequences. This is a joint work with Toshio Nakata, Japan.

PASTA: Pessimistic Assortment Optimization

Zhengling Qi
The George Washington University

We consider a class of assortment optimization problems in an offline data-driven setting. A firm does not know the underlying customer choice model but has access to an offline dataset consisting of the historically offered assortment set, customer choice, and revenue. The objective is to use the offline dataset to find an optimal assortment. Due to the combinatorial nature of assortment optimization, the problem of insufficient data coverage is likely to occur in the offline dataset. Therefore, designing a provably efficient offline learning algorithm becomes a significant challenge. To this end, we propose an algorithm referred to as Pessimistic ASsortment opTimizAtion (PASTA for short) designed based on the principle of pessimism, that can correctly identify the optimal assortment by only requiring the offline data to cover the optimal assortment under general settings. In particular, we establish a regret bound for the offline assortment optimization problem under the celebrated multinomial logit model. We also propose an efficient computational procedure to solve our pessimistic assortment optimization problem. Numerical studies demonstrate the superiority of the proposed method over the existing baseline method.

Bounds for Joint Probabilities of Multi-State Systems Using Preservation of Log-Concavity

Sanjeev Sabnis
Indian Institute Technology Bombay

Log-concavity of multivariate distributions is an important concept in general and has a very special place in the field of Reliability Theory. An attempt has been made in this paper to study preservation results for (i) the discrete version of multivariate log-concavity for multi-state series and multi-state parallel systems consisting of n independent components, states of both components and systems being represented by elements in a subset $S_2 = \{0, 1, 2\}$ of $S_M = \{0, 1, 2, \ldots, M\}$, and (ii) the continuous version of multivariate log-concavity under multi-state series and multi-state parallel systems made up of n independent components and states of both, systems and components, taking values in the set SM. These results for discrete and continuous versions of log-concavity have also been extended to systems that are formed using both multi-state series and multi-state-parallel systems. Further, the results in (ii) have been used to obtain important and useful bounds on joint probabilities related to times spent by multi-state components, multi-state series, multi-state parallel systems, and the combinations thereof.

Probabilistic Prediction for Spatial Processes through Deep Learning

Huixia Judy Wang
The George Washington University

In spatial statistics, the kriging predictor is the best linear predictor at unsampled locations, but not the optimal predictor for non-Gaussian or nonstationary processes. In this talk, I will introduce an indicator deep kriging method for univariate and bivariate spatial processes. The method is based on thresholding the spatial observations at a given set of quantile values and a deep neural network framework. The developed method does not require any parametric assumptions on the marginal distribution and, thus, is more flexible than existing methods. The method can provide the entire predictive distribution function at a new location, allowing for both point and interval predictions. I will present some numerical results to demonstrate the method's efficacy compared to existing approaches. This is a joint work with Pratik Nag and Ying Sun from KAUST.

Prognostics of Lithium-ion Batteries by Combining Discharge Voltage and Capacity Fade Information

Mengmeng Zhu
North Carolina State University

Existing studies have shown that multiple-stage models have a better performance in describing Lithium-ion battery degradation dynamics than single-stage models. Change point identification is a critical step to use multiple-stage degradation models. Current studies mainly used capacity fade data to identify change points. However, it is often difficult to do so because battery capacity fade is typically non-monotonic, and the data requires a long time to collect. In addition, voltage data from battery discharging processes collected in our lab shows great potential to detect change point. Therefore, in this study, we develop a hybrid method that combines a similarity measure and a classification method to detect the change points of batteries using voltage discharge profiles. Then, we propose a new two-stage Wiener process model incorporating change point detection to describe battery capacity degradation. The proposed model shows accurate battery prognostics results (e.g., remaining useful life) by using combined information from voltage and capacity data collected in our lab.

Poster Presentations

Bayesian Covariance Estimation for Multi-Group Matrix-Variate Data

Elizabeth Bersson
Duke University

Multi-group covariance estimation for matrix-variate data with small within-group sample sizes is a key part of many data analysis tasks in modern applications. To obtain accurate group-specific covariance estimates, shrinkage estimation methods which shrink an unstructured, group-specific covariance either across groups towards a pooled covariance or within each group towards a Kronecker structure have been developed. However, in many applications, it is unclear which approach will result in more accurate covariance estimates. In this article, we present a hierarchical prior distribution which flexibly allows for both types of shrinkage. The prior linearly combines shrinkage across groups towards a shared pooled covariance and shrinkage within groups towards a group-specific Kronecker covariance, flexibly providing robustness to misspecification of structural assumptions. We present a Metropolis-Hastings algorithm, and, in simulation, we show the proposed prior provides improved accuracy over alternative approaches based on inaccurate structural assumptions. We illustrate the utility of the proposed prior in speech recognition and an analysis of chemical exposure data.

Forecasting Risk through Artificial Intelligence Based on Machine Learning Algorithms

Dila Bhandari
Tribhuvan University
Kathmandu, Nepal

Statistical risk scores can make a useful contribution to decision making under uncertainty. Artificial Intelligence is a significant technological advancement that has everyone buzzing about its incredible potential. The current research study evaluates the influence of supervised artificial intelligence techniques which machine learning techniques on the nonfinancial firms of Nepal and focuses on the practical application of AI techniques for the accurate prediction of corporate risks which in turn will lead to the automation of corporate risk management. So, in this study, we used financial ratios for accurate risk assessment and for the automation of corporate risk management by developing machine learning algorithms using techniques, namely, random forest, decision tree, naïve Bayes, and KNN for secondary data. Statistical modelling is the key to the development of such risk scores. For this purpose, we collected annual data of nonfinancial companies in Nepal for the period June 2005-June 2019, and the data are analyzed and tested through R studio. The results prove that AI techniques can accurately predict risk with minimum error values, and among all the techniques used, the random forest technique outperforms as compared to the rest of the techniques.

Subsampling Based Model Selection for Network Data

Sayan Chakrabarty
University of Illinois at Urbana Champaign

Complex networks are widely prevalent in scientific applications across various domains. Accurate statistical inference in such networks heavily relies on selecting an appropriate underlying model. This paper presents a novel subsampling-based algorithm for model selection in complex networks. The algorithm effectively splits the network into subnetworks, fits candidate models to each subnetwork, and cross-validates the models based on inter-subnetwork edges using a suitable loss function. The proposed method is demonstrated by applying to detect the number of communities in a blockmodel and estimate the rank of a random dot product graph model. Results demonstrate that the algorithm accurately determines the correct model in diverse scenarios and exhibits significantly faster computational performance compared to existing methods. Thus, this subsampling-based algorithm provides an efficient solution for model selection in complex networks, enabling statistical inference in very large networks as well.

Predicting Dengue Incidence in Central Argentina Using Google Trends Data

Sahil Chindal
Virginia Commonwealth University

Dengue is a mosquito-borne disease prominent in tropical and subtropical regions of the world but has recently been emerging in temperate areas. In Cordoba, a city in temperate central Argentina, there have been several dengue outbreaks in the last decade. Prior to 2009, dengue cases had not been reported in Cordoba. However, major outbreaks occurred in 2009, 2013, 2015, 2016, and 2020. Internet data, such as social media and search engine data, have proven to be useful for predicting the spread of infectious diseases. With the rapid growth of and increased accessibility to the internet, its relevance for epidemiological research has increased in recent years. We have developed a predictive model of dengue incidence in Cordoba using Google Trends data. Specifically, using relevant search terms as predictors and dengue case data as the response, our training model determines which search terms are significant for predicting dengue cases. We employ several methods to test the significance of search terms. We discuss the development of a model utilizing these search terms and how this model will be useful to provide warning of outbreaks to local public health and mosquito control agencies.

Nonparametric Hazard Rate Estimation for Missing Right Censored Lifetimes

Lirit Fuksman
University of Texas at Dallas

Estimation of the hazard rate of lifetime of interest is a classical problem in survival analysis. The problem and its solution are well known for the case of direct observations of the lifetime of interest. Less is known when the data is missing right censored. For the model of lifetimes missing at random, MISE criterion is used to propose an efficient estimator for the hazard rate. Then, its performance is demonstrated using simulation studies and real-life data. Scenario with missing not at random mechanism is also explored.

Reinforcement Learning for Pricing and Inventory Control Under Censored Demand

Korel Gundem
The George Washington University

Learning optimal inventory and pricing strategies from the censored batch data posits a unique challenge in practice. In this paper, we adopt the framework of offline reinforcement learning (RL), adapting the idea of a pessimistic Fitted Q-Iteration (FQI) algorithm to learn the optimal inventory control and pricing policy in the censored demand scenario. Specifically, we propose to learn two different optimal Q-functions, which address the violation of Markovian assumption due to the censored demand. We then show that our algorithm can find the optimal with a finite sample regret guarantee under some mild conditions. Lastly, we conduct a simulation study to showcase the effective performance of our method. Overall, our work offers a new solution and insight for the inventory-pricing territory.

Predicting Conflict Using Machine Learning

Josephine Hughes
Mercer University

The commonplace use of defense treaties between allied nations has interwoven the national security of the participating countries, meaning that if one country engages in a conflict, its allies could also be forced to engage. This possibility forces nations to prepare for unexpected conflicts and their consequences. In the United States, the current method (used for decades) of monitoring global disturbances requires highly trained and often senior-level intelligence analysts to perform independent threat and warning analyses so decision-makers are given time to prepare a course of action for a potential future conflict. The retirement of senior analysts who can perform the level of analysis required is beginning to cause a problem for the intelligence sector. In order to compensate for this loss of personnel, we have created a model to aid the remaining capable, but inexperienced intelligence analysts in predicting future conflicts. Starting with a variety of data gathered by the United States and European Union, including an array of economic and institutional variables, which have been statistically proven to have an impact on political stability, we developed an ensemble algorithm using Targeted Maximum Likelihood Estimation (TMLE). This model was able to, with strong likelihood, estimate the probability of a conflict in the future over a 1 and 5 year time horizon. Our model indicates that political stability can be used to predict future conflict, meaning that the model can be used as a tool to aid intelligence analysts in identifying countries that are at an increased risk of conflict.

Adaptive Finite Element Type Decomposition of Gaussian Random Fields

Jaehoan Kim
Duke University

In this paper, we investigate a general class of approximate Gaussian processes (GP) obtained by taking a linear combination of compactly supported basis functions with the basis coefficients endowed with a dependence structure. This general class includes two highly scalable approximate GP methods: the finite element approximation of the stochastic partial differential equation (SPDE) associated with Matern GP and a linear approximation of a general GP on a regular lattice. We propose prior distributions for the number of basis functions to yield the optimal rate of posterior convergence of the underlying function, adaptively over a large class of smooth functions. We also provide two scalable algorithms and numerics to illustrate the methodology.

A Novel Two-Stage Degradation Model Incorporating Time-to-Event Distributions and Stochastic Processes

Zhanhang Li
Rutgers University

This study pioneers a novel two-stage degradation model, introducing a time-to-event distribution to depict the initial degradation initiation stage, which is different from prevailing stochastic process-based approaches. Our method initiates a vital shift and advancement in degradation modeling. It considers an important, yet frequently ignored, early stage before noticeable degradation is observed. This approach allows for a more precise analysis in various scenarios, such as monitoring corrosion level of steel bars inside the reinforced concrete or evaluating immune responses following vaccination. The first stage of our model utilizes a time-to-event distribution, e.g., Weibull distribution, which concludes when it reaches an "alarm-threshold". This alarm-threshold also indicates the beginning of the second stage, where degradation is modeled using a stochastic process. This innovative strategy discussed in three different scenarios: degradation with no alarm-threshold, degradation with a deterministic alarm-threshold, and degradation with a random alarm-threshold. The utilization of a time-to-event distribution in the initial stage presents a remarkable advantage, facilitating a more precise depiction of various degradation mechanism scenarios and thereby promising to improve predictive analysis in numerous fields.

Estimation for Quantile Spatially Varying Coefficient Models over Complex Domains

Jilei Lin
The George Washington University

The focus of this talk is the estimation for quantile spatially varying coefficient models over complex domains. This model is highly flexible, allowing for exploration of spatial non-stationarity of regression relationships for heterogeneous spatial data that are distributed over a domain with a complex or irregular shape. We will describe an estimation method that employs bivariate penalized spline technique to estimate the unknown functional coefficients. However, due to the non-differentiability of the quantile loss function, existing optimization methods can suffer from slower convergence speeds. To address this issue and improve computational efficiency, we will present a smoothed alternative method that involves smoothing the non-differentiable quantile loss function using convolution smoothing. This process transforms the non-differentiable quantile check function into a twice-differentiable and convex surrogate. Through numerical studies, we will demonstrate the advantages of the smoothed procedure regarding both computational and estimation efficiency in various scenarios.

Bayesian High-dimensional Modeling with Sparse Projection-posterior

Samhita Pal
North Carolina State University

We consider a new Bayesian approach to estimation, uncertainty quantification and variable selection for high-dimensional models including sparse linear regression, sparse precision estimation and low-rank tensor regression. Instead of the traditional method of putting priors that incorporate the sparsity, our prior initially disregards it. However, to make inference, we use the posterior distribution induced by a sparse projection map given by the minimizer of an objective function containing an additional suitably scaled penalty term. We show that the resulting projection-posterior distributions concentrate around the true value of the parameters at the optimal rate adapted to the level of sparsity, that an appropriately re-centered projection-posterior credible ball gives the correct frequentist coverage and we also obtain a selection-consistency that recovers the true sparsity structure with large projection-posterior probability. Finally, we describe how sampling from the projection-posterior can be distributed to a large number of machines without needing to load the entire dataset in any single machine, allowing efficient computation in the big- data setting.

Absolute Risk Prediction for Cannabis Use Disorder

Tingfang Wang
University of Texas at Dallas

Substance use has emerged as a pressing public crisis in the United States, particularly among adolescents. Effective strategies for steering adolescents away from the path leading to substance use disorder are needed. To address this, we developed a Bayesian absolute risk prediction model that predicts the personalized risk of developing cannabis use disorder (CUD) for an adolescent or young adult cannabis user over a given time period. Data from a nationally representative longitudinal study, Add Health, were used to train the model. The proposed model has five risk factors: gender, a measure of delinquency, and scores on personality traits of conscientiousness, neuroticism, and openness. For predicting the risk of developing CUD within 5 years from the age of first cannabis use, the area under the receiver operating characteristic curve (AUC) of the model and the ratio of expected and observed number of cases (E/O), computed via 5-fold cross-validation, were 0.68 and 0.95, respectively. This indicates good discrimination and calibration performance of the model. Furthermore, external validation of the model was conducted using independent test data from Add Health. The AUC and E/O for the same prediction were 0.64 and 0.98, respectively, affirming the model's robust performance in both discrimination and calibration. The proposed model is the first absolute risk prediction model for a substance use disorder. It can aid clinicians in identifying adolescent/youth substance users with a high risk of developing CUD in future for providing early interventions.

High Dimensional Classification and Clustering with Feature Selection

Yong Wang
The George Washington University

High-dimensional data are usually accompanied by a great number of noise features that contain limited information, which could undermine the classification process. We propose marginal screening and pairwise screening of the variables to select the relevant features. The marginal screening process uses tests of equality of marginal distribution functions and pairwise screening tests the equality of joint distribution functions to select variables that carry discriminating information. The tests use the dissimilarity indices MADD and MADMD, which take advantage of the distance concentration phenomenon in high-dimensional space. We show classification based on these indices have misclassification rates that tend to zero as the number of features (p) diverges. Popular clustering algorithms such as k-means clustering tends to deteriorate in high dimension, low sample size (HDLSS) situations with misclassification rate of 1/2 as p diverges. We propose several clustering algorithms based on the dissimilarity indices MADD and MADMD, such as k-Means, test-based algorithms, and minimal spanning tree. We present several simulations as well as analysis of a real data set to demonstrate the suitability of these techniques in the HDLSS setup.

Maximal Coverage Problem for a Naval Task Group with Random Threat

Zeyu Wang
The George Washington University

We present new stochastic programming models with decision-dependent uncertainties for designing the air defense formation of a naval task group. The problem determines the defensive resource allocation strategy that maximizes the probabilistic safety level of a naval task group subject to uncertainties. We use a Boolean modeling method to reformulate the problem into deterministic models. Linearization techniques are introduced to reformulate the non-convex problem into a convex one, and valid inequalities are derived to reduce the complexity of the relaxation problem.

Prototype in New Product Development

Gaoyu Xie
The George Washington University

We adapt the binomial option pricing model to investigate (i) conditions when it is optimal to develop a prototype and (ii) what is the optimal timing to develop a prototype. The analysis is conducted under uncertainty in the outcomes of product development efforts when the decision-makers (DMs) maximize their mean-variance utility. We show that prototyping is highly beneficial when both the development uncertainty and the DMs' risk aversion are at the medium level. Furthermore, we prove that there exists a boundary in terms of initial revenue estimate above which prototyping is optimal for DMs. The model and the derived results can be useful for DMs to support the optimal use of prototyping.