Nevada Chapter of the American Statistical Association

2019 Fall Symposium Speakers

Dr. Petra LeBeau - Rho

Bio: Petra LeBeau received a Doctor of Science degree in Environmental Health Sciences from the University of Maastricht in the Netherlands in 1996. After completion of her degree, she joined the Center for Data and Information Management at the University of Maastricht. As a Research Assistant in Statistics, she supported investigators of various departments within the academic hospital. In 1999, Petra was awarded a NASA funded visiting fellowship with the Laboratory of Computational Biology and Risk Analysis at the National Institute of Environmental Health Sciences. She researched the impact of air pollution and temperature on human health and also developed physiologically-based pharmacokinetic models for several environmental compounds. Petra currently is a Senior Biostatistician at Rho, Inc where she has been the statistical lead of many clinical trials since her start in 2005. For the past 5 years, Petra has been leading the Analytics team of the Bioinformatics group at Rho. Her team applies novel analysis and visualization methods to data from experiments that use high-dimensional sequencing technologies to study the mechanisms of disease.

Keynote: Machine Learning: Insights and Examples From Clinical Research

Abstract: Machine learning (ML) has been making its way into clinical research and the healthcare system due in part to an exponential growth in data stemming from new developments in technology like genomics. At Rho, we support many studies with large datasets including microbiome, proteome, metabolome, and transcriptome. The rapid growth of health-related data will continue along with the development of new methodologies like systems biology (i.e. the computational and mathematical modeling of interactions within biological systems) that leverage these data. Due to the ever-increasing amounts of computational power, improvements in data storage devices, and falling computational costs clinical trial centers are now given the opportunity to apply ML techniques to large and complex data which would not have been possible a decade ago.

During this talk, we will address the basic concepts of ML, how it is changing medicine and provide examples of the application of ML in different areas of clinical research. We will also walk through a workflow for building an accurate model including data pre-processing, feature engineering, feature selection, algorithms and tuning, ensemble learning, feature importance, and model validation. Predictive modeling and software examples using R and H2O will be included from clinical studies that were developed to predict concurrent or future events based on patterns within a set of patient-specific clinical and/or mechanistic characteristics.

Finally, we will discuss methods that aid in the interpretability of these often black-box machine learning models. Although many researchers are embracing machine learning

algorithms, challenges with the interpretation of the more complex models still present a barrier for the widespread practical use of these techniques. While developing an accurate model is often the end goal, model interpretability is also important and highly desired. Understanding why a ML algorithm made the decisions it made and which features were most important in that decision supports a researcher in deciding if this decision makes intuitive sense. Hence, model interpretability not only leads to trust but also to understanding and transparency of the model development process and the resulting predictions and decisions.

Dr. Anna Panorska - University of Nevada, Reno

Bio: Dr. Panorska obtained her Ph.D in Mathematics from the University of California at Santa Barbara. She is currently a tenured Professor at the University of Nevada, Reno. Her research was supported by grants from the National Science Foundation, Department of Defense, Nevada Department of Environmental Protection and the European Union funding. Her research interests include:

- Risk modeling in economics, finance and environmental sciences; modeling and forecasting heavy-tailed phenomena and extreme events.
- Climate/its extremes and change: Modeling, predictability, and prediction of weather extremes; modeling the influence of low frequency climate forcings like El Niño on climate variability and extremes; climate change, weather and climate extremes and connections to public health.
- Hydrology and Environmental Science: Statistical and probabilistic models for describing and forecasting physical and chemical processes in watersheds. Total Maximum Daily Loads modeling. Statistical models for air quality.
- Biostatistics: Collaborative work researching the effects of new clinical therapies including joint work with researchers in the Warsaw Lung Cancer Institute and Warsaw School of Medicine – Clinical Imaging (Poland). Collaboration with faculty in Speech Pathology & Audiology, UNR, School of Medicine.

Anna previously served as the Northern Vice President for the NVASA. She also currently serves as the faculty advisor for the UNR Data Science Club.

Talk: Stochastic Episodes with Light and Heavy Tails: Models, Properties, and Testing.

Anna Panorska*, Tomasz Kozubowski*, Marek Arendarczyk** and Fares Qeadan***
*Department of Mathematics and Statistics
University of Nevada Reno, USA
** Marek Arendarczyk
Mathematical Institute
University of Wroclaw, Poland
***Fares Qeadan
Department of Internal Medicine
University of New Mexico, USA

Abstract: We discuss the problem of modeling the joint distribution of duration (N), maximum (Y) and magnitude (X) of stochastic episodes (events). An event is defined as consecutive observations of a process above (or below) a threshold. Examples of events include growth (or decline) periods of a financial series or climatic or hydrologic episodes, e.g. flood, draught, heat wave, cold spell, etc. The distribution of the vector (N, X, Y) is of direct interest to water management, energy management companies,

disaster management, health departments, investors, actuaries, as well as state and federal regulatory agencies. We present exponentially and heavy tailed models and a likelihood ratio test for deciding between them. We illustrate the modeling potential of these distributions using questions and data from climate, hydrology and finance.

Dr. Colin Grudzien - University of Nevada, Reno
Bio: Colin Grudzien is an Assistant Professor of Statistics at the University of Nevada, Reno. Colin received his Ph.D. in Mathematics in 2016 from the University of North Carolina at Chapel Hill with a focus on applied dynamical systems. Prior to joining the faculty at UNR, Colin was a research assistant at Los Alamos National Laboratory and a postdoc at the Nansen Environmental and Remote Sensing Center in Norway. During fall 2018 Colin was a visiting researcher at CEREA, a Joint Laboratory of École des Ponts ParisTech/EDF R&D. Colin's research interests include Bayesian inference and uncertainty quantification in physical systems, and applications in modeling climate and the electric grid.

Talk: On the Numerical Integration of the Lorenz-96 model, With Scalar Additive Noise, for Benchmark Twin Experiments

Abstract: Relatively little attention has been given to the impact of discretization error on twin experiments in the stochastic form of the Lorenz-96 equations when the dynamics are fully resolved but random. We study a simple form of the stochastically forced Lorenz-96 equations that is amenable to higher order time-discretization schemes in order to investigate these effects. We provide numerical benchmarks for the overall discretization error, in the strong and weak sense, for several commonly used integration schemes and compare these methods for biases introduced into ensemble-based statistics and filtering performance. Focus is given to the distinction between strong and weak convergence of the numerical schemes, highlighting which of the two concepts is relevant based on the problem at hand. Using the above analysis, we suggest a mathematically consistent framework for the treatment of these discretization errors in ensemble forecasting and data assimilation twin experiments for unbiased and computationally efficient benchmark studies. Pursuant to this, we provide a novel derivation of the order 2.0 strong Taylor scheme for numerically generating the truth-twin in the stochastically perturbed Lorenz-96 equations.

Dr. Andrey Sarantsev
Bio: University of Nevada in Reno: Tenure-track assistant professor, 2018-now
University of California in Santa Barbara: Visiting assistant professor, 2015-18
University of Washington in Seattle: Ph.D. in Mathematics, 2010-15
Lomonosov Moscow State University: M.S. in Mathematics, 2005-10
Research Interests: Stochastic Analysis, Actuarial Science, Quantitative Finance

Talk: The Size Effect Revisited

Abstract: We compare returns of portfolios consisting of USA stocks based on their market capitalizations. We find that small stocks have, on average, greater return but greater risk than large stocks. More precisely, the excess return (alpha) for small

stocks is zero, but market exposure (beta) is greater than one. We discuss existing problems, including non-normality of residuals in some cases.

Student Presentations

Heyang Qin - University of Nevada, Reno

Bio: Heyang Qin is a PhD student in the Department of Computer Science and Engineering at University of Nevada, Reno. He conducts research in areas of Deep Learning and Reinforcement Learning under the supervision of Dr. Feng Yan and Dr. Lei Yang. He got his bachelor's degree in University of Electronic Science and Technology of China in 2017.

Talk: Swift Machine Learning Model Serving Scheduling: A Region Based Reinforcement Learning Approach

Abstract: The success of machine learning has prospered Machine-Learning-as-a-Service (MLaaS) -- deploying trained machine learning (ML) models in cloud to provide low latency inference services at scale. To meet latency Service-Level-Objective (SLO), judicious parallelization at both request and operation levels is utterly important. However, existing ML systems (e.g., Tensorflow) and cloud ML serving platforms (e.g., SageMaker) are SLO-agnostic and rely on users to manually configure the parallelism. This talk introduces a swift machine learning serving scheduling framework with a novel Region-based Reinforcement Learning (RRL) approach. RRL can efficiently identify the optimal parallelism configuration under different workloads by estimating performance of similar configurations with that of the known ones. Both theoretical analysis and experiment show that the RRL approach can outperform state-of-the-art approaches by finding near optimal solutions over 8 times faster while reducing inference latency up to 79.0% and reducing SLO violation up to 49.9%. This work will appear at International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2019).

Karla Henricksen - University of Nevada, Reno

Bio: Karla Henricksen is a student at the University of Nevada, Reno. She earned a Bachelor's degree in Mathematics and is now working towards a Masters degree in Statistics and Data Science. She is a recent recipient of the Dean's Merit Scholarship. When she isn't studying for her courses, she enjoys hiking and swimming in the nearby Lake Tahoe.

Talk: Hyperbolic Property of Earthquake Networks

Abstract: We examine the geometry of earthquakes in time-space-magnitude domain using the Gromov hyperbolic property of metric spaces. Gromov delta-hyperbolicity quantifies the curvature of a metric space via four point condition, which is a computationally convenient analog of the famous slim triangle property. We estimate the delta-hyperbolicity for the observed earthquakes in Southern California during 1981-2017. A set of earthquakes is quantified by the Baiesi-Paczuski proximity n that has been shown efficient in applied cluster analyses of natural and human-induced seismicity and acoustic emission experiments. The Gromov delta is estimated in the earthquake space (D,n) and in proximity graphs obtained by connecting pairs of

earthquakes within proximity n0. All experiments result in the values of delta that are bounded from above and do not tend to increase as the examined region expands. This suggests that the earthquake field has hyperbolic geometry. We discuss the properties naturally associated with hyperbolicity in terms of the examined field. The results improve the understanding of dynamics of seismicity and further expand the list of natural processes characterized by the underlying hyperbolic geometry.

Zach Kellar - University of Nevada, Reno
Bio: Zachary Kellar is a senior majoring in Mathematics with a specialization in Statistics, with a minor in Information Systems at the University of Nevada, Reno. He have grown up in Northern Nevada and has always had a strong passion for school, the outdoors, and especially sports. This passion has driven him to pursue a greater understanding of the innerworkings of his favorite sport, baseball, and with new technology providing statistics to analyze, he has undergone a thorough investigation to give useful information regarding just how closely math and baseball are interwoven.

Talk: The Effect of Launch Angle and Exit Velocity on Hit Probability in Major League Baseball

Abstract: Beginning in 2015, a new camera-based tracking system called Statcast was installed in all 30 Major League Baseball stadiums allowing advanced recording of 79 different measurements/variables for each pitch thrown in a game. Two such variables that could not previously be measured are “launch angle,” which is the angle at which the baseball leaves the bat when hit, and “exit velocity,” which is the speed at which the ball leaves the bat. Initial research on launch angle has indicated that larger launch angles or exit velocities may be associated with more successful outcomes. Our goal is to use statistical techniques to develop a quantitative model for the association of the launch angle and exit velocity with the outcome of the batted ball.

Hussain Abbas - University of Nevada, Reno
Bio: Hussain is CEO & Chief Data Scientist at StatsAI and a PhD student in Statistics and Data Science at UNR. He earned his BBA in Actuarial Science and MS in Financial Engineering from Temple University's Fox School of Business and has passed the first 3 Society of Actuaries exams P, FM, and MFE.
Prior to joining UNR, Hussain was a Senior Data Scientist at Oracle in San Francisco where he invented AutoML patents for Electricity Fraud Detection and Pipe Leak Prediction in order to solve client problems. Hussain won the Oracle Innovation Award for his work with the State of California Department of Conservation on the application of his patents to Oil Fraud Detection.
Hussain's 8+ years of experience as a Data Scientist spans 15 different industries ranging from Healthcare and National Security to E-commerce and Online Advertising. Some of his most interesting research projects have been:
1) Advising the US military on how to use AI to catch spies/double agents - Oracle
2) Predicting train delays/early departures for the BART train system in San Francisco - Oracle
3) Electricity Fraud Detection - Oracle
4) Pipe Leak Prediction - Oracle

5) Identifying Fraudulent Oil Drillers for the State of California Department of Conservation- Oracle
6) Daily revenue forecasting - Touch of Modern
7) Internet demand forecasting for 22 million customers - Comcast
8) Algorithmic bidding for online reservations for 1.5 million hotels - Priceline.com
9) Predicting who will get Type II Diabetes for the Medicaid Population - Amerihealth Caritas
10) Predicting the win time for the Kentucky Derby winner 3 hours in advance with a median error of 0.5 seconds.
Hussain's research interests include (but are not limited to): AutoML, Autonomous Inferential Systems, Algorithmic Trading

Talk: The Impact of Bayesian Methods on AutoML

Abstract: AutoML is one of the hottest trends in Artificial Intelligence research and is poised to completely upend the standard model currently used in the Data Science industry. Ironically, few understand what it is, how it works, and its relation to Bayesian Statistics.
In this talk, we will discuss the impact of Bayesian Methods on AutoML. Specifically, we will discuss what AutoML is, how Bayesian Methods play a critical role in AutoML, how AutoML is poised to completely upend the standard model currently used in the Data Science industry, and future areas of research & application.

Omar Kamal - University of Nevada, Reno
Bio: Not available

Talk: TBI Mortality Prediction Model for Pediatric Patients (TMP3) with a Clinical Importance-Based Weighting Approach

Abstract: Traumatic Brain Injury (TBI) causes about 30% of all injury-related deaths. However, there exist few TBI outcome prediction and triage methods for pediatric patients. Common practices used to evaluate TBIs in adults are often difficult to implement (e.g. trauma center exit interviews) or even potentially harmful (e.g. computed tomography scans) for children whose brains are not yet fully developed. Prominent outcome prediction tools such as the Trauma and Injury Severity Score (TRISS) are not designed for head injuries or use within the pediatric cohort. Using pediatric patients entered into the National Trauma Data Bank with head injuries, we constructed a new TBI mortality prediction model for pediatric patients using a logistic regression coupled with a clinical importance-based weight (CIW) approach. The proposed procedure allows us to better predict patients who died as non-survivors. Model performance was evaluated with a variety of measurements. Results showed our proposed model to have higher sensitivity and diagnostic odds ratio and a lower false negative rate when compared with the existing methods. Furthermore, we developed user-friendly software to predict the mortality of incoming TBI pediatric patients. Therefore, the proposed method will help medical practitioners to make timely diagnoses for high risk pediatric patients and provide them with early treatment.