Spring Research Conference

Abstracts

May 22-24, 2019

Virginia Tech
New Classroom Building
1455 Perry St, Blacksburg, VA 24061

ABSTRACTS

Wednesday, May 22

Keynote Address I:

Title: Adversarial Risk Analysis
Keynote speaker: David Banks, Duke University

Adversarial Risk Analysis (ARA) is a Bayesian alternative to classical game theory. Rooted in decision theory, one builds a model for the decision-making of one's opponent, placing subjective distributions over all unknown quantities. Then one chooses the action that maximizes expected utility. This approach aligns with some perspectives in modern behavioral economics, and enables principled analysis of novel problems, such as a multiparty auction in which there is no common knowledge and different bidders have different opinions about each other.

Technometrics: Technometrics Invited Session

Organizer: Dan Apley, Northwestern University
Chair: Dan Apley, Northwestern University

Title: Assessing tuning parameter selection variability in penalized regression
Speaker: Leonard Stefanski, North Carolina State University
Co-author: Eric Laber, Wenhao Hu, Clayton Barker

Penalized regression methods that perform simultaneous model selection and estimation are ubiquitous in statistical modeling. The use of such methods is often unavoidable as manual inspection of all possible models quickly becomes intractable when there are more than a handful of predictors. However, automated methods usually fail to incorporate domain-knowledge, exploratory analyses, or other factors that might guide a more interactive model-building approach. A hybrid approach is to use penalized regression to identify a set of candidate models and then to use interactive model-building to examine this candidate set more closely. To identify a set of candidate models, we derive point and interval estimators of the probability that each model along a solution path will minimize a given model selection criterion, e.g., AIC, BIC, etc., conditional on the observed solution path. Then models with a high probability of selection are considered for further examination. Thus, the proposed methodology attempts to strike a balance between algorithmic modeling approaches that are computationally efficient but fail to incorporate expert knowledge, and interactive modeling approaches that are labor intensive but informed by experience, intuition, and domain knowledge.

Title: Detection of hidden additivity and inference under model uncertainty for unreplicated factorial studies via Bayesian model selection and averaging
Speaker: Christopher Franck, Virginia Tech

The two-way unreplicated layout remains a popular study design in the physical sciences. However, detection of statistical interaction and subsequent inference has

been problematic in this class of designs. First, lack of replication precludes inclusion of standard interaction parameters. Second, while several restricted forms of interaction have been considered, existing approaches focus primarily on accept/reject decisions with respect to the presence of interaction. Approaches to estimate cell means and error variance are lacking when the possibility of interaction exists. For these reasons I propose model selection and averaging-based approaches to facilitate statistical inference when the presence of interaction is uncertain. Hidden additivity, a recently proposed and intuitive form of interaction, is used to accommodate latent group-based non-additive effects. The approaches are fully Bayesian and use the Zellner-Siow formulation of the mixture g-prior. The method is illustrated on empirical and simulated data, and estimates from the model averaging approach are compared with a customized regularization approach. I show that Bayesian model selection is a fruitful approach to detect hidden additivity, and that model averaging allows for inference on quantities of interest under model uncertainty with respect to interaction effects within the two-way unreplicated design. I conclude with some current work that generalizes the notion of latent groupings in categorical levels to the broader class of linear models.

Title: Material Degradation Modeling and Failure Prediction Using Microstructure Images
Speaker: Wujun Si, Wichita State University
Co-authors: Qingyu Yang, Xin Wu

The microstructure of a material has been shown to strongly impact the material's degradation behavior and failure. In this talk, we introduce a two-stage statistical method for degradation and failure prediction of dual-phase advanced high strength steel by using the material's microstructure image information. In Stage 1, we show that the microstructure image covariate can be reduced to a functional covariate while statistical properties of the image are preserved up to second order. In Stage 2, a novel functional covariate degradation model is proposed for the material's failure prediction. A penalized least squares estimation method is developed to obtain the model parameters. Analytical inferences on the model are also developed. Simulation studies and physical experiments are designed and conducted to demonstrate the proposed model.

Computer Experiments 1: Recent Advances in Analysis of Stochastic Computer Experiments

Organizer: Xi Chen, Virginia Tech
Chair: Xi Chen, Virginia Tech

Title: Sensitivity of Stochastic Simulation Outputs to Input Model Variance
Speaker: Barry L. Nelson, Northwestern University
Co-authors: Xi Jiang

Many computer experiments, including discrete-event system simulations, are driven by stochastic inputs that are specified via parametric probability distributions, including arrival processes, machine failure and repair times, and patient treatment times. We first provide a family of answers to the question, “how should we define the sensitivity of the simulation output’s mean or variance to the variance of the input distributions?” We then provide a method for estimating any member of this family

from the nominal simulation experiment with no additional simulation runs required.

Title: Polynomial Control Variates for Asian Options
Speaker: David Goldsman, Georgia Tech
Co-authors: Kemal Dinçer Dingeç, Wolfgang Hörmann

Asian options have payoffs that depend on the average of a time series of underlying stock prices. Unfortunately, there are no closed forms for the probability distributions of the average price or the option price. Although the distribution of the average is intractable, its moments are analytically available under most stochastic models; and these moments are often used in the literature to produce approximations of the option price. This work combines these moments with a simulation-based control variates technique to obtain a reduced-variance estimator for the Asian option price. Our method uses an orthogonal polynomial expansion of the option payoff, of which the expected value is given by the weighted sum of the moments; and it is applicable for any stochastic model allowing the exact moments calculation. We give a number of examples showing the efficacy of the new method. The expected value of the new control variate also serves as an accurate approximation of the Asian option price.

Title: Sequential Probability Ratio Test for Multiple-Objective Ranking and Selection
Speaker: Hong Wan, NC State
Co-authors: Wenyu Wang

We introduce a sequential procedure for the Multi-Objective Ranking and Selection (MOR&S) problems that identifies the Pareto front with a guaranteed probability of correct selection (PCS). In particular, the proposed procedure is fully sequential using the test statistics built upon the generalized sequential probability ratio test (GSPRT). The main features of the new proposed procedure are: 1) a unified framework, the new procedure treats the multi-objective problems in the same way as the single-objective problems; 2) an indifference-zone-free formulation, the new procedure eliminates the necessity of indifference-zone parameter; 3) asymptotically optimality, the GSPRT achieves asymptotically the shortest expected sample size among all sequential tests; 4) general distribution, the procedure uses the empirical likelihood for generally distributed observation. A numerical evaluation demonstrates the efficiency of the new procedure.

Emerging Area 1: Novel Methods for Structured and Relational Data

Organizer: Harry Crane, Rutgers University
Chair: Harry Crane, Rutgers University

Title: Hierarchical network models for structured exchangeable interaction processes
Speaker: Walter Dempsey, Harvard University

Network data often arises via a series of structured interactions among a population of constituent elements. E-mail exchanges, for example, have a single sender followed by potentially multiple receivers. Scientific articles, on the other hand, may have multiple subject areas and multiple authors. We introduce hierarchical interaction exchangeable models for the study of these structured interaction networks. In particular, we introduce the hierarchical vertex components model as a canonical example, which partially pools information via a latent, shared population-level

distribution. Theoretical analysis and supporting simulations provide clear model interpretation, and establish global sparsity and power-law degree distribution. A computationally tractable Gibbs algorithm is derived for inferring sparsity and power law properties of complex networks. We demonstrate the model on both the Enron e-mail dataset and an ArXiv dataset, showing goodness of fit of the model via posterior predictive validation.

Title: Matrix Completion for Network Analysis
Speaker: Ji Zhu, University of Michigan

Matrix completion is an active area of research in itself, and a natural tool to apply to network data, since many real networks are observed incompletely and/or with noise. However, developing matrix completion algorithms for networks requires taking into account the network structure. This talk will discuss two examples of matrix completion used for network tasks. First, we discuss the use of matrix completion for cross-validation or non-parametric bootstrap on network data, a long-standing problem in network analysis. The second example focuses on reconstructing incompletely observed networks, with structured missingness resulting from the egocentric sampling mechanism, where a set of nodes is selected first and then their connections to the entire network are observed. We show that matrix completion can generally be very helpful in solving network problems, as long as the network structure is taken into account. This talk is based on joint work with Elizaveta Levina, Tianxi Li and Yun-Jhong Wu.

Title: Is it possible to motivate tree-based measures of variable importance?
Speaker: Jason Klusowski, Rutgers University

Decision trees with binary splits are popularly constructed using Classification and Regression Trees (CART) methodology. For regression models, at each node of the tree, the data is divided into two child nodes according to a split point that maximizes the decrease in variance (impurity) along a particular variable. An attractive feature of tree-based ensembles is that one can compute, essentially for free, various measures of variable importance using the optimal split points and their corresponding impurities. For random forests, one classic and widely used measure is the mean decrease in impurity (MDI), defined as the sum of weighted impurity decreases over all non-terminal nodes in the tree, averaged over all trees in the forest. Unfortunately, its original motivation was based on simple heuristics and so one is left guessing as to why it actually works. In this talk, we will derive an interesting connection between the bias of a random forest and MDI, in an asymptotic data setting. This relationship theoretically motivates MDI in the following sense: the size of a terminal subnode for a variable is small when the MDI for that variable is large. In other words, terminal subnodes are narrower in directions with strong signals than in directions with weak signals---a phenomenon which also partially explains why random forests are locally adaptive.

Machine Learning 1: Statistical Learning from Survival, Graph and Functional Data

Organizer: Kassie Fronczyk, Lawrence Livermore NL
Chair: Xinwei Deng, Virginia Tech

Title: Scientific Modeling with Functional Data
Speaker: Joanne Wendelberger, Los Alamos NL
Co-authors: Claire McKay Bowen

Functional data analysis typically involves the application of statistical analysis procedures to data associated with underlying functional behavior. In scientific applications, functional data may arise in the context of scientific modeling, where a model may have functional inputs and/or outputs. The modeling process requires both methods for representing the functional data and approaches for understanding and working with functional data to achieve desired scientific goals. A variety of mathematical representations, analyses, and visualizations are being explored to enable the development of a Human in the Loop system for scientific modeling involving functional data.

Title: A Bayesian Nonparametric Regression Model for Analyzing Reliability and Machine Failure Data
Speaker: Valerie Poynor, California State University Fullerton

In reliability analysis, inferences for the hazard rate and mean residual life functions are of key interest. The mean residual life function, in particular, has received limited attention in terms of inference methods under a probabilistic modeling framework. We seek to provide general modeling approach to achieve flexible inferences for both functionals. Reliability data often include a set of predictor variables for the survival response distribution, and in many cases, it is natural to include the covariates as random variables into the modeling. We thus employ Dirichlet process mixture modeling for the joint stochastic mechanism of the covariates and survival responses. This approach implies a model structure for the mean residual life of the conditional response distribution, allowing general shapes for the hazard and mean residual life as a function of covariates given a specific time point, as well as a function of time given particular values of the covariate vector. To expand the scope of the modeling framework, we extend the mixture model to incorporate dependence across fixed groups. This extension is built from a dependent Dirichlet process prior for the group-specific mixing distributions, with common locations and weights that vary across groups through latent bivariate Beta distributed random variables. We illustrate the different components of the model through simulated data examples as well as an analysis on data describing the time to machine tool failure

Title: Learning the Nodes of a Graph from Interactions
Speaker: Jose Cadena, Lawrence Livermore NL

A graph is a natural representation for interaction data in many domains. However, often, the graph is not observed, and researchers need to infer the network implicitly from interactions, which involves making several assumptions. The correct construction of the graph is important, since it has effects on the subsequent analysis, but, surprisingly, methods for node inference have received little attention in previous

work. We pose the task of learning a graph from a set of continuous interaction data. We view this task as a clustering problem where the goal is to aggregate interactions into a set of discrete nodes, such that points in the same node 1) are spatially close, and 2) have similar interaction patterns with points elsewhere. We propose the two-stage distance-dependent Chinese Restaurant Process (ddCRP), a model to infer nodes of a graph starting directly from individual interactions with virtually no preprocessing. Inference of the model is performed via Markov Chain Monte Carlo sampling, and we observe convergence with as few as 20 iterations. We also propose an optimization algorithm based on spectral clustering that has much better scalability at the cost of accuracy and some initial preprocessing of the data. We validate our methods on a set of synthetic instances, and then we apply them to model taxi usage in a dataset of taxi trips in New York City.

Industry & Engineering 1: Recent Advances in Bayesian Computation and Optimization

Organizer: Andrew Brown, Clemson University
Chair: Qiong Zhang, Clemson University

Title: Combining game theory and Bayesian optimization to solve many-objective problems
Speaker: Victor Picheny, INRA

Optimization problems with many objectives (more than three) are notoriously difficult, as the notions of Pareto dominance and optimality, which drive most optimization algorithms, are not relevant to guide the search in the design space or perform decisions. In this talk, we focus on an attractive concept originated from game theory: the Kalai-Smorodinsky solution. We first illustrate how those concepts may find a broad range of applications in engineering and machine learning. However, in a derivative-free, expensive black-box context, very few algorithmic solutions are available to solve game problems. Hence, we propose a novel Gaussian-press based approach for solving games. We follow a classical Bayesian optimization framework, with sequential sampling decisions based on acquisition functions. The relevance of our approach is demonstrated on a challenging ten-objective problem, the calibration of a stochastic, expensive agent-based model.

Title: Low Rank Independence Samplers in Hierarchical Bayesian Inverse Problems
Speaker: Andrew Brown, Clemson University

In Bayesian inverse problems, the posterior distribution is used to quantify uncertainty about the reconstructed solution. In fully Bayesian approaches in which prior parameters are assigned hyperpriors, Markov chain Monte Carlo algorithms often are used to draw samples from the posterior distribution. However, implementations of such algorithms can be computationally expensive. We present a computationally efficient scheme for sampling high-dimensional Gaussian distributions in ill-posed Bayesian linear inverse problems. Our approach uses Metropolis-Hastings independence sampling with a proposal distribution based on a low-rank approximation of the prior-preconditioned Hessian. We show the dependence of the acceptance rate on the number of eigenvalues retained and discuss conditions under which the acceptance rate is high. We demonstrate our proposed sampler by

using it with Metropolis-Hastings-within-Gibbs sampling in numerical experiments in image deblurring and computerized tomography.

Title: Adaptive step-size selection for state-space probabilistic differential equation solvers
Speaker: Oksana Chkrebtii, the Ohio State University

When models are defined implicitly by systems of differential equations with no closed-form solution, small local errors in finite-dimensional solution approximations can propagate into deviations from the true underlying model trajectory. Some recent perspectives in quantifying this uncertainty are based on Bayesian probability modeling: a prior is defined over the unknown solution and updated by conditioning on interrogations of the forward model. Improvement in accuracy via grid refinement must be considered in order for such Bayesian numerical methods to compete with state of the art numerical techniques. We apply principles of Bayesian statistical design to develop an adaptive probabilistic method to sequentially select time-steps for state-space probabilistic ODE solvers. We investigate the behaviour of local error under the adaptive scheme which underlies numerical variable step-size methods. Numerical experiments are used to illustrate the performance of such adaptive schemes, showing improved accuracy in terms of global error over uniform designs when small step lengths are considered.

Agent-based Modeling: Statistical Analysis of Agent Based Models

Organizer: Vadim Sokolov, George Mason University
Chair: Vadim Sokolov, George Mason University

Title: Uncertainty Quantification and Bayesian Model Calibration Applied to Stochastic Systems
Speaker: Dave Higdon, Virginia Tech

Agent-based models (ABMs) use rules at the individual (agent) level to simulate a social, ecologic, or social-technical system, producing structured behavior when viewed at an aggregated level. For example, dynamic network simulation models commonly evolve a very large collection of agents interacting over a network that evolves with time. Such models are often used simulate animal populations, epidemics or transportation, typically producing random trajectories, even when model parameters and initial conditions are identical. While Approximate Bayesian Computation has been used with such models to carry out statistical inference, an alternative is to consider the approaches commonly used in UQ and model calibration. Adapting to the inherent randomness in these simulations is necessary before applying the standard tools of UQ. This talk shows some approaches for adapting Bayesian model calibration to these stochastic systems. We'll consider an example in epidemiology and possibly one in civic planning.

Title: Hierarchical Dynamic Spatio-Temporal Agent-Based Models for Collective Animal Movement
Speaker: Christopher K. Wikle, University of Missouri

Modeling complex collective animal movement presents distinct challenges. In particular, modeling the interactions between animals and the nonlinear behaviors

associated with these interactions, while accounting for uncertainty in data, model, and parameters, requires a flexible modeling framework. Traditional dynamic spatio-temporal models are difficult to parameterize for such complex interacting processes. An alternative is to consider an agent-based collective movement model, such as the self-propelled particle (SPP) model that is well established in physics and biology literature. However, the study and application of these models has primarily focused on simulation studies without uncertainty quantification. To address these challenges, we propose a general hierarchical framework for modeling collective movement behavior with multiple stages. Each of these stages can be thought of as processes that are flexible enough to model a variety of complex behaviors. In particular, we demonstrate our general framework with a hierarchical version of the SPP model applied to collective animal movement. This structure allows us to make inference on potential covariates (e.g., habitat) that describe the behavior of agents and rigorously quantify uncertainty. Further, this framework allows for the discrete time prediction of animal locations in the presence of missing observations. Due to the computational challenges associated with the proposed model, we develop an approximate Bayesian computation algorithm for estimation. We illustrate the hierarchical SPP methodology with a simulation study and by modeling the movement of guppies.

Title: Model Exploration of Agent-based Models at HPC-scales
Speaker: Jonathan Ozik, Argonne National Lab & University of Chicago

Advances in high performance agent-based by developing increasingly realistic models, intricate aspects of the modeled systems can be encapsulated, including agent activities and decision-making, agent interactions over social networks, demographic and geographic heterogeneity, and agent adaptation and learning. For ABMs to be trusted as research and policy-relevant electronic laboratories, it is necessary to robustly characterize their possible model behaviors, a process we have referred to as model exploration, through heuristic methods. As ABMs have become more complex and their parameter spaces have grown in size, brute force methods for model exploration have necessarily given way to adaptive and dynamic approaches, that strategically explore the large model behavior spaces. Implementing dynamic model exploration at the requisite scale is a challenge, and this has led to a proliferation of smaller scale and ad hoc approaches due to the expertise impedance mismatch that exists between general computational science and the HPC realm. In many cases, this mismatch creates off-limit areas of research which are abandoned due to the perceived prohibitive time, effort, or computational expenditures required. In this talk I will present examples of complex/large-scale ABMs and the use of large-scale model exploration on HPC resources to characterize their parameter spaces. The examples are based on two free and open source software projects that our group has produced, the Repast Suite of ABM toolkits (https://repast.github.io) and the Extreme-scale Model Exploration with Swift (EMEWS) framework (https://emews.github.io). We demonstrate how sophisticated heuristic algorithms developed in R or Python can be used to directly run large and complex model exploration workflows, thereby allowing the exploitation of the many freely available libraries currently being developed across statistical and machine learning domains.

Computer Experiments 2: Computer to Physical Experimentation and Analysis

Organizer: Matt Plumlee, Northwestern University
Chair: Bruce Ankenman, Northwestern University

Title: Robust Experimental Designs for Model Calibration
Speaker: Bill Myers, Procter and Gamble

A physics-based model can be used for predicting an output only after specifying the values of some unknown physical constants known as calibration parameters. The unknown calibration parameters can be estimated from real data by conducting physical experiments. This paper presents an approach to optimally design such a physical experiment. The problem is more challenging than the existing work on nonlinear optimal design because of the possibility of model discrepancy, that is, the physics-based model may not be an accurate representation of the true underlying model. Therefore, we propose an optimal design approach that is robust to potential model biases. We show that our designs are better than the commonly used physical experimental designs that do not make use of the information contained in the physics-based model and other nonlinear optimal designs that ignore potential model biases. We illustrate our approach using several toy examples and a real example from Procter & Gamble.

Title: Bayesian design for physical models using computer experiments
Speaker: Antony Overstall, University of Southampton

Design of experiments is an "a-priori" activity making the Bayesian approach particularly attractive. Bayesian inference allows any available prior information to be included in both design and analysis, yielding posterior distributions for quantities of interest that are more interpretable by researchers. Furthermore, the design process allows the precise aim of the experiment to be incorporated. Mathematically, a Bayesian design is given by the maximisation of an expected utility function over all unknown quantities. While straightforward in principle, finding a Bayesian design in practice is difficult. The utility and expected utility functions are rarely available in closed form and require approximation, and the space of all designs can be high dimensional. These problems are compounded when the data-generating process is thought to depend on an analytically intractable physical model, i.e. an intractable likelihood model. This talk will review a recent research programme that has developed methodology to find practically-relevant Bayesian designs for a range of physical models. The methodology uses several layers of computer experiments to both approximate quantities of interest (utilities and expected utilities) and to use these approximations to find Bayesian designs.

Title: Uncertainty Quantification for Binary Black Hole Formation
Speaker: Derek Bingham, Simon Fraser University

Computer experiments are often used to simulate complex physical systems to gain insights of their behavior. In our work, we aim to investigate the formation of binary black holes by means of computer models and explore the initial conditions and physical parameters governing the models. Knowledge of the physical parameters can greatly improve the efficiency of the simulator (the success rate of producing binary

black holes, characterized by a "chirp mass", is quite low). A first stage of our study is to construct an emulator of the computer model. A unique feature of this setting is that, under certain initial conditions, binary black holes do not form. Thus, no chirp mass is observed. In this a talk, methodology for emulating computer models where outputs are limited to only a subset of the input space is presented. Our approach combines a Gaussian process classifier and local Gaussian process models to emulate the simulator and provide estimated of uncertainty.

Quality Monitoring: Monitoring with Social Network and Social Media Data

Organizer: Bill Woodall, Virginia Tech
Chair: Anne Driscoll, Virginia Tech

Title: Exploring and assessing the landscape of network monitoring techniques
Speaker: James Wilson, University of San Francisco

There are a variety of approaches to monitor and identify changes in dynamic networks. Unfortunately, these monitoring methods have not been systematically compared. Moreover, many monitoring methods are specifically designed and suitable only to a particular application. In this talk, we propose a family of synthetic dynamic social networks that can be used to systematically compare the performance of network monitoring methods. We compare several families of state-of-the-art network monitoring methods, including those based on parametric models, sufficient statistics, as well as scan statistics. Through our simulations, we show that there is no free lunch when it comes to monitoring networks and conclude that instead a variety of methods should be used and chosen according to the change of interest.

Title: Anomaly detection in static networks using egonets
Speaker: Srijan Sengupta, Virginia Tech

Network data has rapidly emerged as an important and active area of statistical methodology. In this paper we consider the problem of anomaly detection in networks. Given a large background network, we seek to detect whether there is a small anomalous subgraph present in the network, and if such a subgraph is present, which nodes constitute the subgraph. We propose an inferential tool based on egonets to answer this question. The proposed method is computationally efficient and naturally amenable to parallel computing, and easily extends to a wide variety of network models. We demonstrate through simulation studies that the egonet method works well under a wide variety of network models. We obtain some fascinating empirical results by applying the egonet method on several well-studied benchmark datasets.

Title: A volume and sentiment alerting algorithm incorporating statistical process control methods
Speaker: James Brandon Haynie, Babel Street

Monitoring online subjects or themes for changes in behavior continues to prove difficult for operational application. Increases in communication or the shift in the aggregate sentiment of the discussion may tip or cue an analyst into a change in the environment. While machine learning can perform these tasks, the training

requirement and amount of data can present a challenge as new events occur. Constructing an algorithm using SPC methods as its core coupled with a moving window and automated reset functionality allows for generic application across multiple scenarios. This discussion will cover the collection of data using a multilingual ontology, the construction of the alerting parameters, and the continual validation of the parameters once the data triggers an alert.

Spatial Statistics: Recent Advances in Spatial Statistics

Organizer: Brian Reich, NC State University
Chair: Joe Guinness, Cornell University

Title: Statistics for Ocean Heat Content Estimation with Argo Profiling Floats
Speaker: Mikael Kuusela, Carnegie Mellon University

Argo floats measure seawater temperature and salinity in the upper 2000 meters of the global ocean. These floats are uniquely capable of measuring the heat content of the global ocean, a quantity that is of central importance for understanding changes in the Earth's climate system. But providing detailed spatio-temporal estimates of the heat content is statistically challenging due to the irregular and complex structure of Argo data. We have previously demonstrated (Kuusela and Stein, 2018) that locally stationary Gaussian process regression leads to improved and computationally efficient interpolation of Argo data. Here we build upon those findings to produce improved Argo-based global ocean heat content estimates. We study the sensitivity of these estimates to the underlying statistical assumptions and present results indicating that the magnitude of the overall warming trend may depend on the modeling of the climatological time trend in the mean field estimate. We also investigate the benefits of including time in the interpolation and propose a method for uncertainty quantification that yields appropriate spatial correlations without the need for a global covariance model.

Title: Estimating Atmospheric Motion Winds from Satellite Image Data using Space-time Drift Models
Speaker: Indranil Sahoo, Wake Forest University

Geostationary satellites collect high-resolution weather data comprising a series of images which can be used to estimate wind speed and direction at different altitudes. The Derived Motion Winds (DMW) Algorithm is commonly used to process these data and estimate atmospheric winds by tracking features in images taken by the GOES-R series of the NOAA geostationary meteorological satellites. However, the wind estimates from the DMW Algorithm are sparse and do not come with uncertainty measures. This motivates us to statistically model wind motions as a spatial process drifting in time. We propose a covariance function that depends on spatial and temporal lags and a drift parameter to capture the wind speed and wind direction. We estimate the parameters by local maximum likelihood. Our method allows us to compute standard errors of the estimates, enabling spatial smoothing of the estimates using a Gaussian kernel weighted by the inverses of the estimated variances. We conduct extensive simulation studies to determine the situations where our method should perform well. The proposed method is applied to the GOES-15 brightness temperature data over Colorado and reduces prediction error of brightness temperature compared to the DMW Algorithm.

Title: Statistical Modeling of Tropical Cyclone Intensity Change Using Satellite Imagery
Speaker: Irwin (Trey) McNeely, Carnegie Mellon University
Co-authors: Ann B. Lee, Kimberly Wood, Dorit Hammerling

Tropical cyclones are highly organized, rotating storms which are among the most costly natural disasters in the United States. Operational forecasting of the short-term intensity change of such storms has seen little improvement, particularly when intensity changes rapidly. Traditional statistical approaches to intensity forecasting use area-averaged values for physical and environmental predictors, which are inadequate to characterize the evolution of a storm during occurrences of rapid intensity change, such as rapid weakening events. In order to better understand and predict such events, we are developing models of intensity change that can handle high-dimensional inputs from geostationary satellite images of cloud top temperatures, a proxy for convective strength available at high spatial and temporal resolutions. One facet of our approach is a suite of scientifically-motivated features for convective structure based on image gradients for Organization, Radial profiles, and level sets for Bulk morphology (ORB). These features provide interpretable functional and non-functional inputs to otherwise hard-to-interpret statistical models of storm evolution over time, thus enabling a better understanding of the physical processes behind rapid intensity change in tropical cyclones as well as improvements in forecasting such changes. In this work, we examine the utility of these high-dimensional models and ORB features in identifying convective patterns associated with rapid weakening events and compare predictions based on our satellite-derived features with corresponding predictions generated from traditional area-averaged features.

Thursday May 23

Keynote Address II

Title: Data Competitions: Statisticians as More than Competitors
Speaker: Christine M. Anderson--Cook, Los Alamos NL

Crowdsourcing expertise from many disciplines using a variety of approaches has become increasingly popular as a way of obtaining innovative solutions for difficult problems. Statisticians have long excelled in data competitions, like those hosted by Kaggle and Topcoder, as competitors. But there are other aspects where we can make important contributions: namely, as hosts of the competitions. In this talk, I describe how designing the data set to present to the competitors and creating the right metric to rank the winners on the leaderboard are foundational to driving improvement in the areas of most importance. A new kind of design, non-uniform space-filling designs, are shown and how they can help tailor the competition training and test sets. A detailed post-competition analysis allows deeper understanding about specific aspects of the solutions, as well as comparisons of the strengths and weaknesses of different solutions. Multiple competitions allow sequential learning to drive strategic improvement. The methods are illustrated with two competitions recently hosted by LANL to improve detection, identification and location of radiation threats in an urban environment.

DoE 1: Design of Experiments for Modern Applications

Organizer: Dave Woods, University of Southampton
Chair: Dave Woods, University of Southampton

Title: Designing for low-rank matrix recovery: a maximum entropy approach
Speaker: Simon Mak, Georgia Tech

Low-rank matrices play a fundamental role in modeling a variety of statistical and machine learning problems. In many such problems, however, these matrices cannot be fully observed as data, due to expensive costs or massive matrix sizes. It is therefore of interest to design the data collection procedure, in order to maximize matrix recovery from incomplete data. We propose a new design method for (linear) matrix measurements, using a novel Singular Matrix-variate Gaussian (SMG) model on matrix. Fundamental to our method is the “maximum entropy sampling” principle (Shewry & Wynn, 1987), which states that measurements with maximum entropy can in turn maximize information on matrixHigh. For initial design, this principle provides a way to construct designs using well-packed subspaces. For sequential design, the same principle yields a closed-form design construction, which adaptively incorporates learned subspace information from data. We demonstrate the usefulness of the proposed design method in several real-world applications, including solar imaging, database compression, and building recommendation systems for e-commerce

Title: Closed-loop automatic experimentation for optimisation
Speaker: Tim Waite, University of Manchester
Co-author: Dave Woods, University of Southampton

Automated experimental systems, involving minimal human intervention, are becoming more popular and common, providing economical and fast data collection. We discuss some statistical issues around the design of experiments and data modelling for such systems. Our application is to "closed-loop" optimisation of chemical processes, where automation of reaction synthesis, chemical analysis and statistical design and modelling increases lab efficiency and allows 24/7 use of equipment. Our approach uses nonparametric regression modelling, specifically Gaussian process regression, to allow flexible and robust modelling of potentially complex relationships between reaction conditions and measured responses. A Bayesian approach is adopted to uncertainty quantification, facilitated through computationally efficient Sequential Monte Carlo algorithms for the approximation of the posterior predictive distribution. We propose a new criterion, Expected Gain in Utility (EGU), for optimisation of a noisy response via fully-sequential design of experiments, and we compare the performance of EGU to extensions of the Expected Improvement criterion, which is popular for optimisation of deterministic functions. We also show how the modelling and design can be adapted to identify, and then down-weight, potentially outlying observations to obtain a more robust analysis.

Title: Designing and conducting discrete choice experiments with the R-package idefix.
Speaker: Martina Vandebroek, KU Leuven
Co-author: Frits Traets

Discrete choice experiments are widely used in a broad area of research fields to capture the preference structure of respondents. The design of such experiments will determine to a large extent the accuracy with which the preference parameters can be estimated. This presentation presents a new R-package, called idefix, which enables users to generate optimal designs for discrete choice experiments based on the multinomial logit model. In addition, the package provides the necessary tools to set up online surveys with the possibility of making use of the individual adaptive sequential Bayesian design approach for estimating the mixed logit model. After data collection the package can be used to transform the data into the necessary format in order to use existing estimation software in R.

Reliability: Modeling Complex Reliability and Degradation Data

Organizer: Richard Warr, NC State
Chair: Richard Warr, NC State

Title: Maximum Likelihood Estimation for the Poly-Weibull Distribution
Speaker: Jason Freels, Air Force Institute of Technology

Modified Weibull distributions have been introduced to model data for which the hazard function is bathtub-shaped. Each modified distribution's performance is assessed by its ability to fit reference dataset known to produce a bathtub-shaped hazard rate function. This paper compares the performance of modified Weibull distributions in the literature to that of the generalized poly-Weibull distribution. Numerical and analytical procedures are developed for obtaining the maximum likelihood parameter estimates, standard errors, and moments for the generalized poly-Weibull distribution. The results show that the poly-Weibull distribution fits the reference dataset better than the current best-fit models.

Title: Failure time distributions for complex equipment
Speaker: Richard Warr, NC State University

The exponential distribution is inadequate as a failure time model for most components; however, under certain conditions (in particular, that component failure rates are small and mutually independent, and failed components are immediately replaced or perfectly repaired), it is applicable to complex repairable systems with large numbers of components in series, regardless of component distributions, as shown by Drenick in 1960. This result implies that system behavior may become simpler as more components are added. The necessary conditions for this result are reviewed and some simulation studies are presented to assess how well it holds in systems with finite numbers of components. It is also noted that Drenick's result is analogous to similar results in other systems disciplines, again resulting in simpler behavior as the number of entities in the system increases.

Title: General Path Models for Degradation Data with Multiple Characteristics and Covariates
Speaker: Lu Lu, University of South Florida

Degradation data have been broadly used for assessing product and system reliability. Most existing work focuses on modeling and analysis of degradation data with a single characteristic. In some degradation tests, interest lies in measuring multiple characteristics of the product degradation process to understand different aspects of the reliability performance, resulting in degradation data with multiple characteristics. The literature on modeling such data is scarce. Motivated by the photodegradation process of polymeric materials, we propose a multivariate general path model for analyzing degradation data with multiple degradation characteristics (DCs). The model incorporates covariates for modeling the nonlinear degradation path, and it also includes random effects that are correlated among the multiple DCs to capture the unit-to-unit variation in the individual degradation paths and model the interdependency between the multivariate measurements. An expectation-maximization algorithm combined with the Markov Chain Monte Carlo simulation is developed for estimating the model parameters and predicting system reliability with quantified uncertainty. The performance of the developed method is evaluated and compared with existing methods through a simulation study. The implementation of the method is also illustrated through an example for analyzing the NIST coating degradation data.

VT Alumni: Virginia Tech Alumni Session

Organizer: Robert B. Gramacy, Virginia Tech
Chair: David Banks, Duke University

Title: The Isle of Misfit Designs: A Look at Optimal Designs That Break the Mold
Speaker: Caleb King, JMP

We’re all familiar with the typical screening designs: two-level factors, orthogonal main effects, minimal aliasing (where possible), and run sizes that are powers of two or at least some multiple of 4. Well what if I told you that there existed optimal designs that defied convention? What if I told you that, yes, you can directly construct an optimal screening design with, say, 5 factors in 9 runs? Or 10. Or even 11 runs! All without

the need for long computer searches that leave you crossing your fingers and hoping for the best! Join me as I show you a strange new world of optimal designs that are the best at what they do, even though they might not look very nice.

Title: The Application and Impact of Statistical Engineering in Aerospace Research and Development at NASA
Speaker: Peter A. Parker, National Aeronautics and Space Administration

While statistical methods have been employed in industrial product and process improvement to increase productivity, quality, and profitability for almost 70 years, their routine application at NASA is more recent and still growing. Aerospace research and development applications benefit from statistical engineering's application-focused strategy that leverages statistical tools to accelerate learning, maximize knowledge, ensure strategic resource investment, and inform data-driven decisions. In this presentation, a 20-year survey of statistical engineering applications at NASA is presented through a series of micro case studies spanning work in aeronautics, space exploration, and atmospheric science that highlight significant milestones in technical and organizational impact, educational resources, methodological extensions, and acceptance by the aerospace community. Building on these advancements, steps are proposed to achieve broader application and greater impact of statistical engineering in aerospace research and development.

Title: Functional Variable Selection for EMG Control of a Robotic Hand Prosthetic
Speaker: Jon Stallrich, NC State

Robotic hand prostheses translate multiple forearm electromyography (EMG) signals into finger and wrist movement through a prosthesis controller (PC). Training the PC requires an analysis of concurrent movement and EMG data collected across many forearm muscles, producing correlated EMG signals. To improve the prosthetic's prediction accuracy and stability, we want to develop a PC requiring as few EMG signals as possible. We develop a PC based on a functional linear model that accounts for the underlying biomechanics of hand movement, leading to natural, continuous movement of the prosthetic. The model is made parsimonious and interpretable through the proposed Sequential Adaptive Functional Empirical (SAFE) variable selection procedure. SAFE selection is shown to identify clinically important EMG signals with negligible false positive rates for an able-bodied subject across different postures. A simulation study shows decisive variable selection performance in the presence of correlated covariates and misspecified covariance structure.

JQT: Journal of Quality Technology Invited Session

Organizer: Bianca Maria Colosimo, Politecnico Milano
Chair: Bianca Maria Colosimo, Politecnico Milano

Title: Big Data and Reliability Applications: The Complexity Dimension
Speaker: Yili Hong, Virginia Tech

Big data features not only large volumes of data but also data with complicated structures. Complexity imposes unique challenges in big data analytics. Meeker and Hong (2014, Quality Engineering, pp. 102-116) provided an extensive discussion of the opportunities and challenges in big data and reliability, and described engineering

systems that can generate big data that can be used in reliability analysis. Meeker and Hong (2014) focused on large scale system operating and environment data (i.e., high-frequency multivariate time series data), and provided examples on how to link such data as covariates to traditional reliability responses such as time to failure, time to recurrence of events, and degradation measurements. This paper intends to extend that discussion by focusing on how to use data with complicated structures to do reliability analysis. Such data types include high-dimensional sensor data, functional curve data, and image streams. We first provide a review of recent development in those directions, and then we provide a discussion on how analytical methods can be developed to tackle the challenging aspects that arise from the complexity feature of big data in reliability applications. The use of modern statistical methods such as variable selection, functional data analysis, scalar-on-image regression, spatio-temporal data models, and machine learning techniques will also be discussed.

Title: Designing Computer Experiments with Multiple Types of Factors: The MaxPro Approach
Speaker: V. Roshan Joseph, Georgia Tech
Co-authors: Evren Gul, and Shan Ba

Computer experiments may involve not only continuous input factors but also nominal factors, discrete numeric factors, and ordinal factors. Most existing literature in designing computer experiments focus only on continuous factors. Some works have further considered nominal factors, but the cases that also contain discrete numeric or ordinal factors are almost overlooked. In this work, we propose a new optimal design criterion that can accommodate all these types of factors. The proposed design is flexible in run size and number of factors, and can also achieve good space-filling properties in the full design space and in all possible low-dimensional projections.

Title: Monitoring for changes in the nature of stochastic textured surfaces
Speaker: Daniel W. Apley, Northwestern University
Co-authors: Anh Bui

This work develops an approach for monitoring general global changes in the nature of stochastic textured surfaces using streams of high-dimensional images or related profile data. Stochastic textured surfaces are fundamentally different than the profiles and images that are the focus of most prior profile monitoring work. We represent normal in-control behavior by using supervised learning algorithms to implicitly characterize the joint distribution of the stochastic textured surface pixels. Based on this characterization, we develop a control chart monitoring statistic using likelihood-ratio principles to quantify and detect changes in the stochastic nature of the surfaces, relative to the in-control surfaces. Unlike methods that look for changes in specific predefined features, our approach can detect very general changes in the nature of the textured surfaces. We demonstrate the implementation and effectiveness of the approach with a real textile example and a simulation example.

DoE 2: Novel Applications of DoE in Emerging Areas

Organizer: Lu Lu, University of South Florida
Chair: Lu Lu, University of South Florida

Title: Strategic Subdata Selection for Linear Regression Modeling with Big Data
Speaker: Abigael Nachtsheim, Arizona State University

The need to handle datasets with millions or tens of millions of observations is not uncommon today. However, when datasets contain a very large number of observations, standard statistical methods can become infeasible due to computational limitations. A solution is to select a subset of the full data and conduct the statistical analysis using this subdata alone. One approach is to select the subdata to optimize a specified criterion, with the goal of improving properties of the resulting model estimators. Several related methods have been proposed in the context of linear regression, including leverage-based subsampling methods and the Information-Based Optimal Subdata Selection method (IBOSS) (Ma and Sun, 2015; Ma, Mahoney, and Yu, 2015; Wang, Yang, and Stufken, 2018). IBOSS borrows from the field of experimental design, selecting subdata to maximize the determinant of the information matrix for the model parameters, given a fixed subdata size. This strategy provides several advantages over both simple random sampling and leverage-based subsampling methods. We outline these advantages, providing selected results and comparisons from Wang, Yang, and Stufken (2018). Finally, we conclude with a discussion of remaining challenges in this area of research, and extensions to existing subdata selection methods.

Title: Supersaturated Designs: Research-Based Best Practices and the Future
Speaker: Maria Weese, Miami University, Ohio

The history of supersaturated designs is full of unrealized promise. From the first suggestion in the literature, to their mainstream introduction into the design literature, to the impressive body of work that composes the research area today, the volume of research alone would lead an observer to believe that the practical impact of these designs has been considerable. Alas, this is not the case. In fact, in the literature there is a scant record of classic supersaturated experiments being performed. The striking disparity between research and practice calls for some reflection. Why haven't these designs, which promise such resource-efficiency, been more widely used in industry which so prizes efficiency? We present results using an informal survey of the design community as well as our own analysis of the situation. For practitioners intrigued but nervous about the prospect of using supersaturated experiments, we consolidate some recent, promising research and present practical advice for both the design and analysis of these experiments.

Title: The Analysis of A/B Tests with Comparative Probability Metrics
Speaker: Nathaniel T. Stevens, University of Waterloo

A/B testing and online experimentation has become commonplace in the field of data science. Such experimentation is typically used to formally determine whether one product variant performs at least as well as one or more alternatives. Traditionally, such comparisons have been made via various data-dependent hypothesis tests which all require the correct interpretation of a p-value. However, the widespread

misunderstanding and misuse of p-values often hinders the efficacy of such tests. In this talk I propose the use of comparative probability metrics (CPMs) as an estimation-based alternative to traditional hypothesis testing as a means to determine which among a collection of variants is optimal. CPMs provide a flexible and intuitive means of drawing such conclusions by directly calculating, for example, the probability that one variant is superior to another or the probability that two variants are practically equivalent. The methodology will be illustrated on a real example, and its good properties will be demonstrated with simulated data.

Machine Learning 2: Combining Statistical Design with Learning

Organizer: Lulu Kang, Illinois Institute of Technology
Chair: Lulu Kang, Illinois Institute of Technology

Title: Approximation Algorithms for D-optimal Design
Speaker: Weijun Xie, Virginia Tech

Experimental design is a classical problem in statistics. The goal of the optimal design is to pick the most informative subset across all the possible experiments. There are many criteria to select such a design, among which D-optimality is one of the most popular. In this work, we prove a 1/e-approximation for the D-optimal design problem, providing the first constant factor approximation for the problem. We also analyze and provide the first approximation guarantee for the widely-used coordinate exchange algorithm. Finally, we show that our proposed algorithms are asymptotically optimal when the number of experiments chosen is much larger than the dimension of the input variables.

Title: A Simple Active Learning Approach for Estimating Social Intercorrelations in Large-Scale Social Networks
Speaker: Peter Chien, University of Wisconsin-Madison

The increasing access to large social network data has generated substantial interest in the IT industry. However, due to its large scale, traditional analysis methods often become inadequate. We propose a simple active learning approach for efficient estimation of social intercorrelations in large-scale networks with a spatial composite likelihood model. The proposed approach sequentially takes small samples from the network, and adaptively improves model parameter estimates through learnings obtained from previous samples. Through simulation studies based on simulated networks and real networks, we demonstrate significant advantages of the proposed approach over benchmark estimation methods in terms of both computing time and accuracy in parameter estimation.

Title: Data-Driven Electron Microscopic Image Enhancement for Nanomaterial Analysis
Speaker: Yanjun Qian, Virginia Commonwealth University

In this work, we propose a data-driven super-resolution (SR) algorithm to recover a high-resolution (HR) electron microscopic (EM) image from its low-resolution (LR) counterpart, saving time and efforts for high-quality imaging of nanomaterial. To recover the information loss between the resolution levels, a patch-based library is learned from a captured HR image of a partial area and its corresponding LR region.

Our work is focusing on the properties of the EM image data and its differences from natural optical images. The proposed method improves the quality of input EM images, while other generic SR methods fail to do so. Our method also increases the accuracy of further nanomaterial analysis of the processed images.

Emerging Area 2: Spatial Models for Weather and Computer Simulations

Organizer: Joe Guinness, Cornell University
Chair: Chris Wikle, University of Missouri

Title: Sequential Learning of Active Subspaces for Simulation Experiments
Speaker: Mickael Binois, Argonne National Lab
Co-authors: Nathan Wycoff

In recent years, the Active Subspace Method (ASM) has exploded in popularity as a means of performing subspace sensitivity analysis on black-box functions. However, the ASM, naively applied, requires gradient evaluations of the target function. In the event of noisy, expensive, or stochastic simulators, it may be infeasible to evaluate gradients via finite differencing. In such cases, a surrogate model is employed, upon which finite differencing is performed. When the surrogate model is a Gaussian process, we show that the ASM estimator can have a closed form, rendering the finite difference approximation unnecessary. Further, we use our closed form solution to develop acquisition functions focused on the ASM allowing sequential learning tailored to sensitivity analysis. We also show that the traditional ASM estimator may be viewed as a Method of Moments estimator for a certain class of Gaussian processes. Finally, we demonstrate how uncertainty on Gaussian process hyperparameters may be propagated to uncertainty on the sensitivity analysis, allowing model-based confidence intervals on the active subspace. Our methodological developments are illustrated on several examples.

Title: Max-Infinitely Divisible Models for Spatial Extremes Using Random Effects
Speaker: Ben Shaby, Penn State University

Rare events can have crippling effects on economies, infrastructure, and human health and wellbeing. Their outsized impacts make extreme events critical to understand, yet their defining characteristic, rareness, means that precious little information is available to study them. Extremes of environmental processes are inherently spatial in structure, as a given event necessarily occurs over a particular spatial extent at a particular collection of locations. Characterizing their probabilistic structure therefore requires moving well beyond the well-understood models that describe marginal extremal behavior at a single location. Rather, stochastic process models are needed to describe joint tail event across space. Distinguishing between the subtly different dependence characteristics implied by current families of stochastic process models for spatial extremes is difficult or impossible based on exploratory analysis of data that is by definition scarce. Furthermore, different choices of extremal dependence classes have large consequences in the analysis they produce. I will present stochastic models for extreme events in space that are 1) flexible enough to transition across different classes of extremal dependence, and 2) permit inference through likelihood functions that can be computed for large datasets. These modeling goals are accomplished by representing stochastic dependence relationships conditionally, which will induce desirable tail dependence properties and allow

efficient inference through Markov chain Monte Carlo. I will describe models for spatial extremes using max-infinitely divisible processes, a generalization of the limiting max-stable class of processes which has received a great deal of attention. This work extends previous family of max-stable models based on a conditional hierarchical representation to the more flexible max-id class, thus accommodating a wider variety of extremal dependence characteristics while retaining the structure that makes it computationally attractive.

Title: Nonparametric Spectral Methods for Multivariate Spatial and Spatial-Temporal Data
Speaker: Joe Guinness, Cornell University

We propose computationally efficient methods for estimating stationary multivariate spatial and spatial-temporal spectra from incomplete gridded data. The methods are iterative and rely on successive imputation of data and updating of model estimates. Imputations are done according to a periodic model on an expanded domain. The periodicity of the imputations is a key feature that reduces edge effects in the periodogram and is facilitated by efficient circulant embedding techniques. In addition, we describe efficient methods for decomposing the estimated cross spectral density function into a linear model of coregionalization plus a residual process. The methods are applied to two storm datasets, one of which is from Hurricane Florence, which struck the souteastern United States in September 2018. The application demonstrates how fitted models from different datasets can be compared, and how the methods are computationally feasible on datasets with more than 200,000 total observations.

Computer Experiments 3: Engineering-driven Development in Statistical Modeling and Estimation

Organizer: Qiong Zhang, Clemson University
Chair: Qiong Zhang, Clemson University

Title Dimension Reduction for Gaussian Process Models via Convex Combination of Kernels
Speaker: Lulu Kang, Illinois Institute of Technology

Some engineering and scientific computer models that have high dimensional input space are actually only affected by a few essential input variables. If these active variables are identified, it would reduce the computation in the estimation of the Gaussian process (GP) model and help researchers understand the system modeled by the computer simulation. More importantly, reducing the input dimensions would also increase the prediction accuracy, as it alleviates the ``curse of dimensionality" problem. In this talk, we propose a new approach to reduce the input dimension of the Gaussian process model. Specifically, we develop an optimization method to identify a convex combination of a subset of kernels of lower dimensions from a large candidate set of kernels, as the correlation function for the GP model. To make sure a sparse subset is selected, we add a penalty on the weights of kernels. Several numerical examples are shown to show the advantages of the method. The proposed method has many connections with the existing methods including active subspace, additive GP, and composite GP models in the Uncertainty Quantification literature.

Title: Statistical Modeling for Spatio-Temporal Convection-Diffusion Processes and Its Applications
Speaker: Xiao Liu, University of Arkansas

The digitization of physical space by sensing technologies has led to an explosive growth of spatio-temporal data arising from physical convection-diffusion processes. Examples span multidisciplinary areas including environmental science, radar image analysis, geology, material deterioration, etc. This research extends the existing results in the literature and proposes a non-stationary statistical model for spatio-temporal data arising from a generic class of convection-diffusion processes with spatially-varying convection-diffusion, and a nonzero-mean spatio-temporal source-sink process. Unlike existing results assuming constant convection-diffusion and zero-mean source-sink (which leads to a stationary solution of the SPDE in the limiting case), the proposed spatio-temporal model has a non-stationary and non-separate covariance structure. We adopt the idea of spectrum decomposition to approximate the SPDE-based spatio-temporal processes. However, due to the spatially-varying convection-diffusion, the temporal evolution of individual spectrum coefficients are coupled with each other. We provide a systematic approach to integrate the results established by our investigation into the Hierarchical Dynamical Spatio-Temporal Models (DSTM) framework in the literature, and investigate the statistical inference using computationally efficient algorithms. Numerical examples and a case study are presented to demonstrate the advantages of the proposed modeling approach.

Title: Approximate Bayesian Inference for Sequential Accelerated Life Testing
Speaker: Qiong Zhang, Clemson University

Approximate Bayesian inference (Chen and Ryzhov, 2019) is proposed to construct computationally tractable statistical learning procedures for incomplete or censored data. In this talk, I will discuss a sequential model updating procedure via approximate Bayesian inference for the Log-normal model with censored observations. We are able to show that the proposed procedure leads to consistent model parameter estimation. The proposed procedure is applied to accelerated life testing to determine the material alternative with optimal reliability performance in a sequential manner.

Topics on Design and Analysis: Recent Advances in Design and Analysis

Organizer: Devon Lin, Queen's University
Chair: Joanne Wendleberger, Los Alamos National Laboratory

Title: Designing Experiments with Unstructured Treatments for General Network Structures
Speaker: Frederick Phoa, Academia Sinica Taiwan
Co-authors: Ming-Chung Chang, Jin-Wen Huang

Experiments on connected units are commonly conducted in many scientific fields. An experimental unit in these applications may connect with some others, and the treatment applied to a unit has an effect, called a network effect, towards the responses resulted in the neighboring units. Designing such experiments was rarely discussed in the literature. Parker, Gilmour, and Schormans (2017) initiated a study of As-optimal designs on connected experimental units with unstructured treatments, assuming that the network effects are unknown constants. This work investigates in a

similar design problem but the network effects are assumed to be random effects, which lead to a property that the responses of two units are correlated if some neighbors of one unit and those of the other receive the same treatment. Alphabetical optimality criteria are considered for selecting good de- signs with high efficiency of estimating the treatment effects and/or high accuracy of predicting the network effects. We provide theoretical conditions for designs to be optimal and illustrate our theory with some numerical examples.

Title: Vecchia-Laplace approximations of generalized Gaussian processes for big, non-Gaussian spatial data
Speaker: Daniel Zibler, Texax A&M University

Generalized Gaussian processes (GGPs) are highly flexible models that combine a latent GP with a potentially non-Gaussian likelihood from the exponential family, to perform GP classification and non-Gaussian spatial regression. However, inference for GGPs can be analytically intractable, and large datasets pose computational challenges due to the inversion of the GP covariance matrix. To achieve computational feasibility even for very large spatial datasets, we propose a Vecchia-Laplace approximation for GGPs, which combines a Laplace approximation to the non-Gaussian likelihood with a sparse Vecchia approximation to the GP. We outline the properties of the algorithm, provide numerical studies, and show an application to satellite data.

Title: Convergence rates for Gaussian process modeling under boundary conditions
Speaker: Liang Ding, Georgia Tech

Gaussian processes (GPs) are often used as surrogate models for approximating expensive computer code, which in turn simulate complex phenomenon of interest. In many such problems, boundary information (i.e., how the phenomenon behaves on the boundaries of an input variable) is readily available from physical principles or scientific knowledge. While recent work suggests that incorporating boundary information can greatly improve the predictive performance of GPs, such an improvement is not well understood theoretically. To this end, we propose a new Brownian Bridge Gaussian Process (BBGP) model for incorporating boundary information, and show that the BBGP predictor enjoys quicker convergence rates (both in Lp and in probability) over standard GPs. We also show that, by accounting for boundaries, our BBGP model becomes much more resistant to the ``curse-of-dimensionality'' in nonparametric regression. As a by-product, our results also reveal a novel connection between GP modeling and finite element modeling.

Banquet Speech

Speaker: Oliver Schabenberger (SAS)

Friday May 27

JUQ: Journal of Uncertainty Quantification Invited Session

Organizer: David Higdon, Virginia Tech
Chair: David Higdon, Virginia Tech

Title: Calibrating a Stochastic, Agent-Based Model Using Quantile-Based Emulation
Speaker: Arindam Fadikar, Virginia Tech

In a number of cases, the quantile Gaussian process has proven effective in emulating stochastic, univariate computer model output [M. Plumlee and R. Tuo, (2014)]. In this talk, we describe an approach that uses this emulation approach within a Bayesian model calibration framework to calibrate an agent-based model of an epidemic. In addition, we extend this approach to handle the multivariate nature of the model output, which gives a time series of the count of infected individuals. The basic modeling apparatus is adapted from [D. Higdon et al., (2008)], using a basis representation to capture the multivariate model output. The approach is motivated with an example taken from the 2015 Ebola challenge workshop which simulated an Ebola epidemic to evaluate methodology.

Title: Calibration of mathematical computer models
Speaker: Ksenia N. Kyzyurova, Brown University

The problem of calibration of mathematical computer models with respect to collected data occasionally occurs in contemporary research. Calibration task is analogous to identifying the preimage of a set of experimental or observational data under a certain computer model --- function over an input space of parameters to that model. Identified preimage is formed as a subset of the input space to the computer model. However, collected data is typically described by means of probability distributions; thus, leading to performing calibration within the Bayesian inverse framework (using the Bayes' theorem). Bayesian inversion is considered advantageous over other non-Bayesian or pseudo-Bayesian approaches because of interpretability of its results. In practice, calibration quickly runs into computational obstacles: the shape of the posterior distribution resulting from the Bayesian inversion may be ``ugly", such that standard approaches to its estimate (including Markov chain Monte-Carlo (MCMC) approximation) are prohibitive. Instructional examples are provided for illustration. We conclude that calibration must be performed together with investigation of its issues in scientific computing and software/hardware numerics.

Title: Prediction Uncertainties Beyond the Range of Experience: A case study in inertial confinement fusion implosion experiments
Speaker: Scott vander Wiel, Los Alamos National Lab

Scientists often predict physical outcomes, e.g. experimental results, with the assistance of computer codes that, at their best, only coarsely approximate reality. Coarse predictions are challenging in large part due to the multitude of seemingly arbitrary yet consequential decisions that must be made such as choice of relevant data, calibration of code parameters, and construction of empirical discrepancy forms. In this paper, we present a case study in the context of inertial confinement fusion

(ICF) implosion experiments where extrapolative predictions are needed with quantified uncertainties. The purpose of this case study is to illustrate relevant statistical methods, as applied to ICF model fitting and prediction, to document the numerous decisions that must be made in the prediction pipeline, to extend a complex example in extrapolation to the UQ community, and to reflect on the challenges we encountered supporting extrapolations with imperfect models and thereby recommend several future research directions. We end with a discussion about the uncertainty quantification community's role in less than ideal predictive scenarios like our ICF exercise.

Industry & Engineering 2: Data Science for Material Science-Some New Perspectives

Organizer: Tirthankar Dasgupta, Rutgers University
Chair: Tirthankar Dasgupta, Rutgers University

Title: A two-stage Gaussian process modeling approach for constrained optimization of ReaxFF parameters
Speaker: Ying Hung, Rutgers University

ReaxFF is a widely used reactive force field method in materials science to simulate bond formation and dissociation in large atomistic systems. In order to obtain the correct physics from the ReaxFF simulations, a set of force field parameters must be optimized for each materials system. This problem is challenging because some combinations of the force field parameters lead to infeasible results. To address this problem, we propose a framework based on a two-stage procedure. A generalized Gaussian process model is fitted in the first stage to identify the feasible regions. In the second stage, a Gaussian process model is fitted within the feasible region is used to optimize the force field parameters.

Title: Sequential Bayesian Design for Accelerated Life Tests
Speaker: I-Chen Lee, National Cheng Kung University

To ensure a material can last long enough in the field, accelerated life tests are commonly used to collect data and then make prediction for the field performance. For evaluating the property of the material, a good testing strategy is desirable. While there has been a lot development in optimum test planning, most of the methods assume that the true parameter values are known. However, in reality, the true model parameters may depart from the planning values. In this paper, we propose a sequential strategy for test planning, and use Bayesian framework for the sequential model updating. We apply the proposed procedure to the accelerated cyclic fatigue tests of Polymer composite materials, which are widely used in areas such as aerospace and alternative energy industries. We also use extensive simulation to evaluate the properties of the proposed sequential test planning strategy. Finally, we compare the proposed method to the traditional optimum design. Our results show that the proposed strategy is more robust and efficient when true values of parameters are unknown.

Title: Design, analysis and optimization of response surfaces in the presence of internal noise
Speaker: Tirthankar Dasgupta, Rutgers University
Motivated by applications in material science, a unified Bayesian approach to response surface optimization when the input factors are noisy is proposed, and the effect of ignoring available information related to the noise on optimization is investigated. The problem of designing optimal experiments, under the extra uncertainty generated by noisy inputs, is explored. An optimality criterion for generating optimal experimental data is proposed.

HPC: Statistical Modeling with High Performance Computing
Organizer: Ana Kupresanin, LLNL
Chair: Jose Cadena, LLNL

Title: Climate-model Ensemble Consistency Testing
Speaker: Dorit Hammerling, Colorado School of Mines

Constant evolution and improvement of large scale climate simulation codes such as the Community Earth System Model (CESM) necessitate quality checks to verify the constitution of new climate simulations. The CESM Ensemble Consistency Test (CESM-ECT) was developed as a flexible but objective method for checking the statistical consistency between an accepted ensemble of climate simulations and new simulations created with updated code or within a new computational infrastructure. CESM-ECT utilizes a testing framework based on the popular technique of Principal Component Analysis (PCA) to determine whether a set of new simulations is statistically distinguishable from the established ensemble of climate simulations. The test has proven of high practical use and it is now part of the official CESM release. One shortcoming of PCA is that accurate estimation requires the computational expense of a large ensemble. In recent work, we have analyzed theoretical properties of the testing framework and are investigating various ways of maintaining or improving the properties of the test while reducing the required ensemble size.

Title: Some Pieces of Exascale Uncertainty Quantification
Speaker: Earl Lawrence, Los Alamos National Lab

Exascale computing is coming to the Department of Energy. The new prefix brings some new challenges. Raw computing power will overwhelm the I/O capability of these machines. This is a particular difficulty for the practice of statistical uncertainty quantification (UQ) which relies on ensembles of simulations, each of which will necessarily be only partially available. This talk will present two separate piece of work that may be useful parts of the exascale UQ workflow. The first part of the talk will cover in situ analysis of simulations in order to summarize simulations as they are running. The second part of the talk will cover statistical emulation with incomplete data.

Title: Methodological Reconstruction of Historical Seismic Events using Modern Statistical Inversion: A Case Study for the 1852 Banda Arc Earthquake and Tsunami
Speaker: Justin Krometis, Virginia Tech
Co-authors: Jared Whitehead, Ron Harris, Nathan Glatt-Holtz

Most understanding of seismic hazards is based modern instrumental data from the last 75 years even though seismic events can occur on timescales of hundreds or thousands of years. A key priority in emergency preparedness for seismic hazards is understanding events that occurred over longer time periods; unfortunately, data sources describing these events are commonly sparse in detail and laced with high levels of uncertainty. To address this need, we present a Bayesian framework for leveraging nontraditional datasets for the reconstruction of historical seismic events. As a proof of concept, we reconstruct the location and magnitude of the great 1852 earthquake in Indonesia from recently-translated records from Dutch settlements at the time. We model anecdotal or otherwise noisy observations of the resultant tsunami via probability distributions, which are then combined with data on modern seismic events and numerical modeling of tsunami propagation to compute a statistical model of parameters describing the 1852 event. The results of the inversion imply that the event was a massive mega-thrust earthquake on the same scale as the December 2004 Sumatra event and located along a narrow arc in the eastern or southeastern portion of the study region. The approach outlined here can readily be extended to other historical events with tsunami observations and eventually to other types of observations (e.g., shaking) as well.

DoE 3: New Developments in Design of Experiments

Organizer: David Edwards, Virginia Commonwealth University
Chair: David Edwards, Virginia Commonwealth University

Title: Model-Robust Mixture Designs
Speaker: Byran J. Smucker, University of Miami, Ohio

Optimal designs are often used for constrained mixture experiments due to their irregular experimental region. For large mixture experiments, the number of mixtures needed to fit standard linear models for mixture experiments may be too large, especially when consideration of second- or third-order blending terms is desired. Model-robust designs are optimal for a set of models, and a carefully chosen model space provides the ability to estimate a user-specified number of higher-order terms to take advantage of anticipated sparsity. We present a computationally tractable algorithm for generating model-robust mixture designs based on both the D- and I-criterion. The algorithm's engine uses a heuristic which has been implemented in commercial software but to our knowledge has not previously been described in the literature. We compare the model-robust designs with optimal and Bayesian-optimal designs, and the model-robust designs show improved estimation and prediction ability for mixture experiments when the number of blends is less than the number of model parameters in the full mixture model.

Title: Enumeration and multi-criteria selection of orthogonal minimally aliased response surface designs
Speaker: José Núñez Ares, KU Leuven
Co-authors: Peter Goos

We define a new family of orthogonal RSDs, for which there is no aliasing between the main effects and the second-order effects. This family includes not only the classical RSDs, such as the Central Composite Designs or Box-Behnken Designs, but also the more modern Definitive Screening Designs. We name these designs orthogonal

minimally aliased RSDs (or OMARS designs). Using integer programming techniques, we constructed a database of OMARS designs for 3 to 8 factors. Each design in the catalog is extensively characterized in terms of efficiency, power, fourth-order correlations, FDS plots, projection capabilities, etc. We identify interesting designs and investigate trade-offs between different quality criteria. Finally, we present a multi-attribute decision algorithm to select designs from the catalog. An important result of our study is that we discovered some novel and interesting designs that challenge standard RSDs in terms of the number of runs, projection capabilities and other criteria. Another contribution is our multi-attribute decision algorithm itself, which allows the selection of an experimental design for a wide range of experimental circumstances, while considering statistical indicators that cannot be included in current DoE software. In this talk, we will also present our preliminary results on OMARS designs with 2-level categorical factors.

Title: Utilizing the Block Diagonal Information Matrix for Nonregular Two-Level Designs
Speaker: Robert W. Mee, University of Tennessee
Co-authors: David J. Edwards

Two-level fractional factorial designs are often used in screening scenarios to identify active factors. This presentation investigates the block diagonal structure of the information matrix of certain nonregular two-level designs. We connect the block diagonal information matrix to the parallel flats design literature and gain insights into the structure of what is estimable and/or aliased using the concept of minimal dependent sets. Three parallel flat designs represent one common example, but we show how the block diagonal structure arises in other contexts. Recognizing this structure helps with understanding the advantages of alternative designs.

Machine Learning 3: Non-Parametric Tree Partition Models for Modern Statistical Learning

Organizer: Matt Pratola, The Ohio State University
Chair: Matt Pratola, The Ohio State University

Title: Finding and Leveraging Structural Information via Bayesian Regression Tree Ensembles.
Speaker: Antonio Linero, Florida State University

Analysis of complex or high-dimensional datasets is often aided when the data is structured; such structures include sparsity and low-order interaction structures, as well as a-priori known graphical structures obtained from external data sources. In this talk, we present strategies for finding and leveraging structural information in order to boost the performance of Bayesian decision tree ensembles. We focus on three problems: (i) detection of low-order interactions in datasets; (ii) use of grouping information, similar to the group and overlapping-group lasso; and (iii) graphical structures which encode a-priori known relationships between predictors. The methods we develop provide powerful, nonparametric, alternatives to existing frequentist and Bayesian approaches which have focused almost exclusively on linear models, with implementation of our approaches requiring minor modifications to

existing algorithms. We provide simulation evidence for the benefits of our proposed approaches, and apply the methodology to several datasets.

Title: Sparse Additive Gaussian Processes via Recursive Partitioning
Speaker: Matthew Pratola, Ohio State University

We introduce a novel model for Gaussian process (GP) regression in the fully Bayesian setting. Motivated by the idea of sparsification, localization and Bayesian additive modeling, our model is built around a recursive partitioning (RP) scheme. Within each partition, a sparse GP regression model is fitted. A Bayesian additive framework combines the partitions, allowing the model to admit both global trends and local refinements. Our approach mitigates the issue of pseudo-input selection and avoids the need for complex inter-block correlations. The crucial trade-off becomes choosing between many simpler local models or fewer complex global models can be easily and sensibly tuned by the practitioner. We compare our model against popular alternatives with simulated and real datasets, and find the performance is competitive, while the fully Bayesian procedure enables appropriate quantification of modeling uncertainties.

Title: Fully Nonparametric Bayesian Additive Regression Trees
Speaker: Rob McCulloch, Arizona State University
Co-authors: Edward George, Prakash Laud, Brent Logan, Robert McCulloch and Rodney Sparapani

Bayesian Additive regression trees, or BART (Chipman, George, and McCulloch (2010)), has proven to be an effective tool for obtaining inference in the general model $Y = f(x) + e$, where e is normal with zero mean and unknown variance. As in the ensemble approach boosting, the function f is modeled as a sum of many regression trees. Unlike boosting, a fully Bayesian approach it taken in which a prior on the function f is specified through a prior on the underlying trees and the full posterior is computed using an effective Markov Chain Monte Carlo algorithm. Key to the success of BART is a relatively simple specification for the prior on f. However, the assumption of normal errors and the standard inverted chi-squared specification of the prior on the error variance is unsatisfying. In this paper we replace the normal errors with Dirichlet process mixtures (DPM) of normal components with both uknown means and variances. While this DPM modeling approach is a well established methodology, choosing the hyperparamters can be difficult. These difficulties are even more pronounced when the DPM model is coupled with the flexible nature of our BART prior on f. We adapt the BART default priors to the DPM case so that users can obtain a fully flexible nonparametric inference for a wide variety of problems with minimal tuning. When prior information is available, the user has the option of inputing the information in a reasonably uncomplicated manner.

Computer Experiments 4: Advances in Computer Experiments

Organizer: Bill Myers, Procter and Gamble
Chair: Ying Hung, Rutgers University

Title: Approaches to the Emulation of Chains of Computer Models with Application to Epidemic Policy Making
Speaker: Samuel Jackson, University of Southampton

We have developed novel Bayesian emulation methodology to analyse chains of computer models, where the outputs of one model feed into the next model. Such computer models, with inputs and outputs representing quantities of interest, are frequently developed to aid the understanding of the general behaviour of real-world processes. In particular, the motivation for this work comes for epidemic disease modelling, linking, for example, atmospheric dispersion, dose-response and epidemiological models. Emulation techniques are well discussed in the literature as an approach to efficiently understanding computationally intensive models, however, for the purposes of analysing chains of computer models, tend to focus on approximating the entire chain using a single emulator. Our work focusses on several methods to link Bayes linear emulators of each component model of a chain. We have developed emulators for models where the input is uncertain (as the inputs to all but the first model are the uncertain emulated outputs of another model). The first method proposes analysing each emulator's behaviour for a sample of inputs arising from a probabilistic distribution commensurate with our beliefs about the output of the previous emulator. The second method extends the field of emulation to directly incorporate uncertain inputs within each emulator itself. We demonstrate the potential of these novel emulation approaches using intuitive examples, before demonstrating their advantage over the single emulator approach in our application of modelling of epidemic diseases. Application of our techniques to models of such epidemics permits detailed uncertainty quantification via, for example, thorough sensitivity analysis into the effect of unknown quantities, thus aiding online policy decision making in the event of an epidemic.

Title: Predicting the Output from a Stochastic Computer Model When a Deterministic Approximation is Available
Speaker: Evan Baker, University of Exeter

Computer models are increasingly being used to learn things about the real world instead of, or as well as, real life experiments. These computer models can take a long time to run, which can make using them for their intended purpose difficult. To help with this, we can create a statistical replacement of the computer model that is much much quicker to run (an emulator), based on a few runs of the computer model. Emulating stochastic computer models can be difficult; requiring a large number of runs from the computer model before the mean and variance can be accurately estimated. I will discuss how deterministic approximations of stochastic computer models can be leveraged to improve the emulation of the stochastic models. Specific implementation details involve the summation of Gaussian processes, which is a popular idea in many areas of emulation. Two example stochastic computer models will be used to showcase this method: one is an agent based model which can be used for modelling epidemics; and the other is a model for the expected energy usage of a building.

Title: Gaussian Process with Input Location Error and Applications to the Composite Parts Assembly Process

Speaker: Wenjia Wang, SAMSI

Co-authors: Xiaowei Yue, Benjamin Haaland, and C.F. Jeff Wu

In this paper, we investigate Gaussian process regression with input location error, where the inputs are corrupted by noise. Here, we consider the best linear unbiased predictor for two cases, according to whether there is noise at the target untried location or not. We show that the mean squared prediction error does not converge to zero in either case. We investigate the use of stochastic Kriging in the prediction of Gaussian processes with input location error. We show that stochastic Kriging is a good approximation when the sample size is large. Several numerical examples are given to illustrate the results, and a case study on the assembly of composite parts is presented. Technical proofs are provided in the Appendix.