Spring Research Conference

Poster Abstracts

May 22-24, 2019

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

ABSTRACTS

Wednesday, May 22
18:00-20:00, Reception and Poster Session

Poster abstracts are sorted alphabetically by presenter's last name.

Presenter: Samuel Anyaso-Samuel, Boise State
Title: Dynamic Sampling Versions of Popular SPC Charts for Big Data Analysis

The advancement of science and technology has elicited large volumes of data across several domains. Conventional statistical process control (SPC) charts have shown to be suboptimal for monitoring these datasets. In this study, we propose computationally efficient methods and modifications of some existing process control charts for detection of distributional shifts in sequential processes that generate big data. For effective monitoring of these processes, we design control charts with dynamic sampling schemes rather than the conventional fixed sampling scheme used in popular SPC charts. The dynamic schemes which are an increasing function of the p-value of the charting statistic have variable sampling intervals. Results show that the new SPC charts with dynamic sampling schemes are more efficient, yield quicker detection of distributional shifts and provide ample interpretation of the monitoring process

Presenter: Shane Bookhultz, Virginia Tech
Title: Measuring Polarity From News Sources: A Topic Modeling Approach

The world is a polarized place due to individuals that hold firm oppositional opinions. In recent years, this subject has come to light due to the emergence of social media and divisive journalistic campaigns. To demonstrate the increasing level of polarity, we analyze daily news articles by expanding on a high-dimensional, temporal variant of the Latent Dirichlet Allocation (LDA) algorithm. This algorithm allows us to discern evolving daily topics in addition to the intensity and accelerated rate that these topics enter and leave the media. Furthermore, while the temporal LDA algorithm can be computationally expensive, we implement techniques to improve the computational efficiency of the algorithm. Since the veracity of many news sources are at times questionable, we address polarizing topics to determine the levels of agreement between these sources. In order to measure the polarization, we introduce a new metric for assessing polarity between and within news sources and articles, based on entropy metrics. The performance of our temporal LDA algorithm and polarity metrics are compared with other competing methods. Through this work, we assess the nature and origins of various polarizing sources and evaluate their impacts on individuals and society.

Presenter: Claire McKay Bowen, Los Alamos National Lab
Title: Telling the Visual Story within Big Data: Case Studies on Interactive Visualizations for Supercomputer Data

Scientific communication is key in any successful collaboration. For statisticians, data visualization is an important communication technique to explain certain features and gain intuition about the data. However, creating an appropriate data visualization that will provide unknown insights into the data can be difficult, especially with the volume and density of big data. In this talk, I present the challenges of analyzing big spatial and temporal data from Los Alamos National Laboratory supercomputers (e.g. data density) and the various visualizations I created to address these problems. Specifically, I developed R Shiny Apps using Plotly to provide unique and interactive visualizations that revealed interesting data stories for further analysis. Additionally, I applied the latest data privacy methods called differentially private data synthesis on the supercomputer data to satisfy non-disclosure agreement requirements. This kind of data privacy technique is currently being implemented by several entities such as the US Census Bureau, Google, Apple, Microsoft, and Uber.

Presenter: Austin Cole, Virginia Tech
Title: Inducing Point Methods for Gaussian Process Surrogates of Large-Scale Simulations

Gaussian processes (GPs) provide a flexible methodology for modeling complex surfaces. One challenge with GPs is the computational burden with an increasing sample size or number of dimensions. The machine learning community has turned to pseudo-inputs or inducing points to reduce the computational burden in such contexts. We seek to port this family of methods to build GP surrogates for noisy and even heteroskedastic stochastic processes, with extensions to sequential design and Bayesian optimization. We show that using inducing points extends the reach of GP surrogates in big simulation contexts and makes for efficient design and meta-modeling of large scale computer simulation experiments. Examples are provided for epidemiological, industrial, and financial applications.

Presenter: Arnab Chakraborty, NC State
Title: A statistical analysis of noisy crowdsourced weather data

Spatial prediction of weather elements like temperature, precipitation, and barometric pressure are generally based on satellite imagery or data collected at ground stations. None of these data provide information at a more granular or “hyper-local” resolution. On the other hand, crowdsourced weather data, which are captured by sensors installed on mobile devices and gathered by weather-related mobile apps like WeatherSignal and AccuWeather, can serve as potential data sources for analyzing environmental processes at a hyper-local resolution. However, due to the low quality of the sensors and the non-laboratory environment, the quality of the observations in crowdsourced data is compromised. This paper describes methods to improve hyper-local spatial prediction using this varying-quality noisy crowdsourced data. We introduce the Veracity Score (VS) to assess the reliability of the crowdsourced observations with the help of a coarser, but high-quality, reference data. A VS-based methodology to analyze noisy spatial data is proposed and evaluated through theoretical analysis and extensive simulation studies. The merits of the proposed approach are illustrated using several case studies analyzing crowdsourced daily-average ambient temperature readings for a particular day in the continental United States. Moreover, the performance of the proposed methodology has been compared with the existing robust geostatistical methods using the famous 'coalash' dataset.

Presenter: Adam Edwards, Virginia Tech
Title: Precision Aggregated Local Models

Gaussian Process (GP) models have long been used as a flexible method for nonparametric regression. Despite their accuracy, they are infeasible for larger data sets due to the scaling of the computational burden (ON3) and the storage requirements (ON2). Typical methods to combat intractability have focused on splitting the larger GP into smaller problems that can be treated as functionally independent either by partitioning the domain space [Rushdi et al. (2016)], or partitioning the data themselves into discrete sets [Tresp (2000)]. These methods get around the functional dependence of the individual GP models they use by assigning no weight to most of the models during prediction, or exploiting global models learning theoretically the same surface respectively, however each solution has its drawbacks. Partition models maintain accuracy, but lose continuity at the boundaries [Park et al. (2011), Park and Huang (2016)]. Typical averaged models, on the other hand, maintain absolute continuity in both the mean and variance surface while over-smoothing the function as a whole. Using Local Approximate Gaussian Processes (LAGP) [Gramacy and Apley (2013)] as a method to build local experts, the PALM model bridges the gap between these two approaches to create a locally accurate global model that maintains absolute continuity. We address the functional dependence of individual GP models, creating a unified surface with drastically reduced computational burden O(N2C + NC * n3C), and storage requirement O(N2C + NC * n2C), where NC, nc << N.

Presenter: Olga Egorova, University of Southampton
Title: Optimal planning of factorial experiments accounting for multiple objectives

In the framework of response surface factorial experiments with fitted polynomial models, we consider the case of potential model contamination presence and develop compound optimality criteria which components would correspond to the various and, in most cases, conflicting objectives: maximising the precision of the fitted model estimates, minimising the joint effect of potentially missed terms and minimising the bias of the primary model parameters. We adapt the replicate-based approach to error estimate ('pure-error'), and its extension to the framework of experiments with restricted randomisation. Examples of a blocked experiments are considered, and we also discuss aspects of the developed methodology related to the optimal design search.

Presenter: Carl Ehrett, Clemson University
Title: Coupling material and mechanical design processes via computer model calibration

In traditional engineering design, material selection is a matter of choosing a material with appropriate properties for the project at hand from a database of known materials, often as a matter of ad-hoc satisficing. Material design usually occurs separately, and without an eye to specific end-uses. It is desirable to wed these design processes, selecting a material design by modeling its performance outcomes in a particular engineering application. Therefore, here we offer an example of calibrating material design parameters to desired performance targets for a wind turbine blade. We show that existing techniques for model calibration can be profitably reconceptualized as a method for optimization and applied to solve this material design problem. Rather than calibrating a model to find a posterior distribution of unknown parameters in order to bring the model maximally into agreement with reality, we calibrate to find a posterior distribution on controllable model inputs in order to bring the predicted system behavior into agreement with pre-determined performance targets. In essence, we treat performance targets as “desired observations” and use them as the data in the calibration problem. We demonstrate our proposed methodology in both an artificial case and in the case of a finite element model of wind turbine blade performance and cost. In the latter case, we demonstrate how to estimate the Pareto front with uncertainty bands.

Presenter: Marco Ferreira, Virginia Tech
Title: Objective Bayesian Analysis for Gaussian Hierarchical Models with ICAR spatial random effects

Bayesian hierarchical models are commonly used for modeling spatially correlated areal data. However, choosing appropriate prior distributions for the parameters in these models is necessary and sometimes challenging. In particular, an intrinsic conditional autoregressive (CAR) hierarchical component is often used to account for spatial association. Vague proper prior distributions have frequently been used for this type of model, but this requires the careful selection of suitable hyperparameters. We derive several objective priors for the Gaussian hierarchical model with an intrinsic CAR component and discuss their properties. We show that the independence Jeffreys and Jeffreys-rule priors result in improper posterior distributions, while the reference prior results in a proper posterior distribution. We present results from a simulation study that compares frequentist properties of Bayesian procedures that use several competing priors, including the derived reference prior. We demonstrate that using the reference prior results in favorable coverage, interval length, and mean squared error. Finally, we illustrate our methodology with an application to 2012 housing foreclosure rates in the 88 counties of Ohio.

Presenter: Rahul Ghoshal, NC State
Title: Variable Selection in Functional Linear Concurrent Regression

We propose a novel method for variable selection in functional linear concurrent regression. Our research is motivated by a fisheries footprint study where one of the goal is to identify important time varying socio-structural drivers influencing patterns of seafood consumption and hence fisheries footprint over time. We develop a variable selection method in functional linear concurrent regression extending the classically used scalar variable selection methods like LASSO, SCAD and MCP. We show in functional linear concurrent regression the variable selection problem can be addressed as a group LASSO, and their natural extension; group SCAD or a group MCP problem. Through simulations, we illustrate our method, particularly with group SCAD or group MCP penalty can pick out the relevant variables with high accuracy and has minimal false positive and false negative rate even when data is observed sparsely, is contaminated with noise and the error process is highly non stationary. We also demonstrate two real data applications of our method in study of dietary calcium absorption and fisheries footprint, in selection of influential time varying covariates.

Presenter: Chris Grubb, Virginia Tech
Title: Extending Nearest-Neighbor GPs for Non-gridded Data Imputation

Solving spatiotemporal missing data problems will always be of interest; due to data-gathering methods like satellite images, we are constantly collecting data with missingness due to cloud cover, precipitation, or other environmental forces. In spatiotemporal problems, there are many well-studied methods for imputing this missing data, the most popular of which is the Gaussian process. This work is motivated by a real-world problem: downscaling two large surface temperature datasets, one with high spatial resolution and one with high temporal resolution, in order to obtain more accurate temperature data for use in modeling. This problem is complicated by the fact that the (lat, long) pairs are not common to both datasets; only one source is gridded. With large amounts of data, fitting a standard GP can be very computationally expensive, which has led to many alternatives, one of which is NNGPs, the nearest-neighbor variant. The computational advantage of nearest-neighbor algorithms is that they create covariance matrices with block structure, which we exploit to obtain computationally efficient algorithms.

Presenter: Andrew Hoegh, University of Montana
Title: An Assessment of Uncertainty in Unconstrained Ordination

Species composition data collected at sampling locations is projected into lower dimensional settings to identify patterns associated with sampling locations by a process called ordination. Ordination was historically implemented with distance-based methods, but recently there has been a shift toward model-based approaches. Model-based unconstrained ordination is commonly formulated using a Bayesian latent factor model that permits uncertainty assessment for parameters, including the latent factors that correspond to an environmental gradient related to community composition. While model-based methods have an additional benefit of addressing uncertainty in the estimated gradients, the current practice is to report point estimates without summarizing uncertainty. Accurate reporting of uncertainty is an important part of transparency in the scientific process; thus, a model-based approach that accounts for uncertainty is necessary. An R package, *UncertainOrd*, contains visualization tools that accurately represent estimates of the gradients in community composition in the presence of uncertainty.

Presenter: Peter Hovey, University of Dayton

Title: Evaluation of Safety Effectiveness of Median Cable Barriers Installed on Freeways in Ohio

The Ohio Department of Transportation (ODOT) began installing median cable barriers in 2003 along highway medians for all roadways that were narrower than 59 ft. The central goal of this work was to prevent cross-median crashes (CMCs) that raised a concern due to their frequencies and severe injuries they caused when they occurred. Cross-median crashes occur when a vehicle leaves its travel way, enters or crosses the dividing median, and collides with vehicles moving in the opposite direction. This study analyzed data from 41 locations covering 201 miles of installed median cable barriers in the years 2009-2014. The study found that median cable barriers were effective in stopping vehicles from breaching the barrier; 95.4 percent of all cable median barrier crashes had no penetration of the cable barrier, i.e., the vehicles where stopped or bounced by the cables.

Presenter: Jiangeng Huang, Virginia Tech
Title: On-site surrogates for large-scale calibration

Motivated by a challenging computer model calibration problem from the oil and gas industry, involving the design of a so-called honeycomb seal, we develop a new Bayesian calibration methodology to cope with limitations in the canonical apparatus stemming from several factors. We propose a new strategy of on-site experiment design and surrogate modeling to emulate a computer simulator acting on a high-dimensional input space that, although relatively speedy, is prone to numerical instabilities, missing data, and nonstationary dynamics. Our aim is to strike a balance between data-faithful modeling and computational tractability within an overarching calibration framework---tuning the computer model to the outcome of a limited field experiment. Situating our on-site surrogates within the canonical calibration apparatus requires updates to that framework. In particular, we describe a novel yet intuitive Bayesian setup that carefully decomposes otherwise prohibitively large matrices by exploiting the sparse blockwise structure thus obtained. We illustrate empirically that this approach outperforms the canonical, stationary analog, and we summarize calibration results on a toy problem and on our motivating honeycomb example.

Presenter: Tzu Hsiang Hung, University of Wisconsin
Title: A Random Fourier Feature Method for Modeling Computer Experiments with Gradient Information

Computer experiments with gradient information are increasingly conducted in many fields in engineering and science. The gradient-enhanced Gaussian process emulator is often used to model data from such experiments. We propose a random Fourier feature method to mitigate potential numerical issues of this type of emulator. The proposed method employs random Fourier features to obtain an easily computable, low-dimensional feature representation for shift-invariant kernels involving gradients. The effectiveness of the proposed method is illustrated by several examples.

Presenter: Tao Jiang, NC State
Title: Higher Criticism Tuned Sparse Group Lasso for Weak and Sparse Signals in GWAS

In the current study, we propose an extension of least absolute shrinkage and selection operator (LASSO) regression to address variable selection and modeling when sample sizes are limited compared to the data dimension. Our method is motivated by high-throughput biological data, such as genome-wide association studies (GWAS). We propose a new upper bound of the regularization parameter in sparse group LASSO based on an estimated lower bound of the proportion of false null hypotheses with confidence. The bound is estimated by applying the empirical distribution of dependent or independent p-values from single marker/variable analysis, where a second-level significance testing, the higher criticism statistic is used. An upper bound of tuning parameter in LASSO, is decided corresponding to the lower bound of the proportion of false null hypotheses. Thus, the tuning range is narrow since the upper bound is lower. The final decision of non-zero estimates (e.g., significant loci in GWAS) will contain more variables so that the power of modified GWAS is higher than or equal to the original sparse group Lasso. Different correlation levels among variables in true regression models are also studied. We demonstrate the performance of our method using both simulation experiments and a real data application in lipid trait genetics from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial.

Presenter: Zhongnan Jin, Virginia Tech
Title: Multivariate Functional Data Clustering with Variable Selection and an Application to Sensory Data

In this paper, we conduct multivariate functional data clustering in an unsupervised manner. Dimensions of high volume data are significantly reduced by using only partial multivariate functional principal components. The multivariate functional principal component analysis (MFPCA) enables us to transform multi-dimensional and continuously measured data into an orthonormal matrix, where each dimension in the original data can be expressed by corresponding columns in the transformed matrix. In this way, clustering techniques can be applied to the transformed matrix. In our study, we assume this transformed data follow a Gaussian mixture model with K distinct centers, and the covariance matrix is the same across all clusters. We use penalty based maximum likelihood to conduct clustering with automatic variable selection. In addition, grouped variable selection is considered in this study. Because principal scores belonging to the same original function should have the same tendency to be in or out for the variable selection procedure. An application of engineer system sensory data is studied. Model performances are evaluated in the simulation study.

Presenter: Marcin P. Jurek, Texas A&M University
Title: Multi-resolution filters for massive spatio-temporal data

Spatio-temporal data sets are rapidly growing in size. For example, environmental variables are measured with ever-higher resolution by increasing numbers of automated sensors mounted on satellites and aircraft. Using such data, which are typically noisy and incomplete, the goal is to obtain complete maps of the spatio-temporal process, together with proper uncertainty quantification. We focus here on real-time filtering inference in linear Gaussian state-space models. At each time point, the state is a spatial field evaluated on a very large spatial grid, making exact inference using the Kalman filter computationally infeasible. Instead, we propose a multi-resolution filter (MRF), a highly scalable and fully probabilistic filtering method that resolves spatial features at all scales. We prove that the MRF matrices exhibit a particular block-sparse multi-resolution structure that is preserved under filtering operations through time. We also discuss inference on time-varying parameters using an approximate Rao-Blackwellized particle filter, in which the integrated likelihood is computed using the MRF. We compare the MRF to existing approaches in a simulation study and a real satellite-data application.

Presenter: Lata Kodali, Virginia Tech
Title: Anomaly Detection in Time-varying Networks

Network data has emerged as an active research area in statistics. However, much of the focus of ongoing research has been on static networks which are invariant over time. Monitoring time-varying networks to detect anomalous changes has applications in both social and physical sciences. In this work, we propose a general framework for anomaly detection in time-varying networks by incorporating principles from statistical process monitoring. We show that our method works in a variety of well-studied networks models (e.g., a dynamic latent space model and a dynamic degree-corrected stochastic blockmodel) in the context of a simulation study as well as in applications involving well-known network datasets.

Presenter: Arvind Krishna, Georgia Tech
Title: Distributional clustering: A distribution-preserving clustering method

One of the key uses of clustering is to identify representative points from a dataset of interest via cluster centers. However, a drawback of k-means clustering is that it induces a distortion between the distribution of its cluster centers and that of the underlying data. This can be disadvantageous in problems where cluster centers are subsequently used to gain insights on the data, such as density estimation or pattern recognition, as the accuracy of the analysis method in these cases depends on how well the cluster centers mimic the distribution of the data. To address this shortcoming, we propose a new clustering method called ‘distributional clustering’, where cluster centers capture the distribution of the underlying data. We first prove the asymptotic convergence of the proposed cluster centers to the data generating distribution (which addresses the aforementioned distortion problem of k-means), then propose an efficient algorithm for computing these cluster centers in practice. Finally, we demonstrate the effectiveness of our method on synthetic and real datasets.

Presenter: Li-Hsiang Lin, GA Tech
Title: Transformation and Additivity in Gaussian Process

We discuss the problem of approximating a deterministic function using Gaussian Processes (GP). The role of response transformation in GP modeling is not well understood. We argue that transformations on the responses can be used for making the deterministic function approximately additive, which can then be easily estimated using an additive GP. We call such a GP a Transformed Additive Gaussian (TAG) process. To capture possible interactions which are unaccounted for in an additive model, we propose an extension of TAG process called Transformed Approximately Additive Gaussian (TAAG) process. We develop efficient techniques for fitting a TAAG process. In fact, we show that it can be fitted to high-dimensional and big data much more efficiently than the usual GP. Furthermore, we show that the use of TAAG process leads to better estimation, interpretation, visualization, and prediction.

Presenter: Ruijin Lu, Virginia Tech
Title: Scalable, powerful and robust basis space testing for high-dimensional data

With modern high-throughput technologies, scientists can now collect high-dimensional data of various forms, including brain images, medical spectrum curves, and engineering signals. These data provide a rich source of information on disease development, cell evolvement, engineering systems, and many other scientific phenomena. However, these data are featured with high dimension and high correlations in measurement points, making it desirable to find a fast reliable and powerful approach to extract useful information from the wealth of data. This work focuses on improving the power in the testing of high-dimensional functional data. We consider Westfall-Young Randomization Tests in basis-space via lossless or near-lossless compression. We show that these tests satisfy several nice theoretical properties, including the successful control of family-wise error rate, the improving of power with appropriate truncation, and the asymptotic optimality. The effectiveness of this testing approach is demonstrated using two applications - the detection of regions of the spectrum that are related to pre-cancer using fluorescence spectroscopy data and the detection of disease-related regions using Tensor-Based Morphometry data derived from structural magnetic resonance imaging.

Presenter: Christian Lucero, Virginia Tech
Title: The Illustration of Statistical Learning Concepts Using Shiny

There are many concepts in Statistical/Machine Learning that are difficult for students and professionals to fully comprehend. One of the ways to effectively illustrate these concepts is through the use of interactive applications (apps). These apps provide clear examples that focus on visual and numerical summaries and minimize the required knowledge of deeper mathematical and statistical details. Our work focuses upon the development of Shiny apps for use in educational and consulting platforms. The tools designed for the educational platform are intended to accompany lecture slides with the goal of helping students to better visualize the data and understand statistical learning procedures with clear educational objectives. The consulting platform is intended to be used by statistical consultants who often work with outside professionals who may not be well versed in statistical methodology. Our library of apps are useful for demonstrating clear and consistent examples of the techniques that are most applicable to serve the clients needs. The Statistical

Applications and Innovations Group at Virginia Tech is particularly interested in the use of Shiny apps as part of its focus to promote Appropriate Analytics and Intelligent Interpretation.

Presenter: Pulong Ma, SAMSI
Title: Spatial Statistical Downscaling for Constructing High-Resolution Nature Runs in Global Observing System Simulation Experiments

Observing system simulation experiments (OSSEs) have been widely used as a rigorous and cost-effective way to guide development of new observing systems, and to evaluate the performance of new data assimilation algorithms. Nature runs (NRs), which are outputs from deterministic models, play an essential role in building OSSE systems for global atmospheric processes because they are used both to create synthetic observations at high spatial resolution, and to represent the ``true'' atmosphere against which the forecasts are verified. However, most NRs are generated at resolutions coarser than actual observations. Here, we propose a principled statistical downscaling framework to construct high-resolution NRs via conditional simulation from coarse-resolution numerical model output. We use nonstationary spatial covariance function models that have basis function representations. This approach not only explicitly addresses the change-of-support problem, but also allows fast computation with large volumes of numerical model output. We also propose a data-driven algorithm to select the required basis functions adaptively, in order to increase the flexibility of our nonstationary covariance function models. In this article we demonstrate these techniques by downscaling a coarse-resolution physical NR at a native resolution of 1 latitude times 1.25 longitude of global surface CO2 concentrations to 655,362 equal-area hexagons.

Presenter: Suman Majumder, NC State
Title: Statistical Downscaling with Spatial Misalignment: Application to Wildland Fire PM2.5 Emissions Forecasting

Fine particulate matter PM2.5 has been documented to have adverse long term health effects on human and wildfires send out tons of such particles into the air affecting the nearby areas. Forecasters use numerical models to predict PM2.5 concentrations in different areas for the next 24 to 72 hours to warn the public of impending health risk. Statistical methods are needed to calibrate numerical model forecast using monitor data and statistical challenges such as spatial misalignment and potential model bias often come up. Typical model calibration techniques do not allow correction of errors due to misalignment of geographic locations. We propose a spatiotemporal downscaling methodology that, using image registration techniques, identifies the spatial misalignments and accounts for and corrects the bias produced by such warping. Our model is fitted in the Bayesian framework to provide uncertainty of the space warping function as well as the forecasts. Two sources of data: a short term forecasts of PM2.5 concentration from a deterministic model and spatially sparse monitor data is used in our model. The monitor data is assimilated for correcting and re-calibrating the forecasts from the deterministic model. We apply this method to simulated datasets as well as on a real dataset to demonstrate the utility of the proposed method and its applicability as a real-time forecast method.

Presenter: Sierra Nicole Merkes, Virginia Tech
Title: Robust Anomaly Detection in Large-Scale Multi-Type Sensor Systems

Modern wind tunnel systems are constructed of various instruments containing hundreds and thousands of individual sensors. Each sensor requires continual quality monitoring, as corrupted data will increase experimental costs, wasted time, and even worse, lead to spurious scientific and engineering conclusions. Monitoring methodologies such as Gaussian Processes and Principal Component Analysis have been examined for anomaly detection in wind tunnel experiments. However, these methodologies are limited due to continual instrumental movement, wind tunnel re-configurations, or low-levels of adverse observations in training datasets. Along with the challenges of monitoring hundreds to thousands of multi-type sensors, wind tunnel experiments are subject to low number of analyzable experimental runs because of proprietorship or because of the multitude of possible tunnel configurations with low replication size. In this presentation, we compare two robust system monitoring methodologies: (1) a method based on a robust Principal Component Analysis technique, and (2) a proposed method based on a Bayesian, heavy-tailed posterior distribution with Principal Component Analysis. Through the simulation and real wind tunnel experiments, we exemplify the need for anomaly detection methodologies using a robust, correlated, multi-type sensor approach.

Presenter: Tom Metzger, Virginia Tech
Title: Detection of latent heteroscedasticity and group-based regression effects in linear models with the R package slgf

Categorical predictors are widely used in linear models in a variety of applications. Standard modeling approaches make potentially simplistic assumptions regarding the structure of categorical model effects that may obfuscate more complex relationships governing the data. Specifically, we investigate the scenario where the levels of a categorical predictor fall into two latent groups. We propose a fully Bayesian model selection approach of clustering the data according to the levels of a categorical predictor to reveal latent group-based fixed effects, heteroscedasticity, and/or hidden interactions. Through the use of mixture g-priors and fractional Bayes factors, we test for both the presence and structure of such clustering. We illustrate our method through the new R package "slgf" in the context of ANCOVA and unreplicated two-way layouts, although the method we describe is broadly applicable to the class of linear models that include categorical predictors.

Presenter: Matt Slifko, Virginia Tech
Title: The Cauchy-Net Mixture Model for Clustering with Anomalous Data

Clustering is a popular exploratory task for identifying patterns in multivariate data. Unfortunately, many clustering techniques suffer from two major limitations: (1) the need to specify the number of clusters a priori and (2) the detrimental impact of anomalous data on the ability to discover structure. We present the Cauchy-Net Mixture Model (CNMM) for overcoming these limitations. The CNMM extends the flexibility of a Dirichlet Process Mixture Model (DPMM) by creating a mixture of a DPMM with an additional Cauchy distributed component, which we refer to as the Cauchy-Net. The intuition is to leverage the heavy tails of the Cauchy-Net for capturing observations that do not fit into the well-defined clusters and remove their influence on cluster formation. We explore the performance of the CNMM in a variety of situations.

Presenter: John Smith, Virginia Tech
Title: Bayesian Parameter Estimation for Ecosystem State Space Models with Linear Autoregressive Process Models

Our society depends greatly on services provided by ecosystems. Trees, for example, help to clean our air and provide oxygen, store CO2, and provide timber and other building materials. Our understanding of ecosystems is instrumental in planning for the future. With the high cost of planning and executing experiments, the primary method of understanding ecosystems has shifted towards a model driven approach that is calibrated by experimental data rather than purely driven by experimental data. The questions that we ask about ecosystems are intrinsically about the future and therefore it is important that we are able to predict how these ecosystems will evolve over time, with uncertainty, contingent on scenarios that we expect to see. This is the premise of the field of Ecological Forecasting (EF). Latent variables are ubiquitous in EF problems. This makes the State Space model one of the most powerful techniques in EF, for its flexibility in handling many forms of latent variables - such as random and systematic observation errors, missing data, unobserved variables, and proxy data. However, the Ecology community has not fully adopted the state space framework, due to its computationally intensive nature and the inflexibility of pre-packaged MCMC software like JAGS and STAN to handle large ecosystem models. Here we will focus on carbon exchange models for forests. A general method is outlined for fitting these state space models in the Bayesian framework using MCMC, with tractable latent state full conditional distributions derived for a large class of process models, namely autoregressive models that are linear at each time step. Performance of these methods is benchmarked vs standard approaches for several different models, for both computational time and accuracy of estimated parameters.

Presenter: Sumin Shen, Virginia Tech
Title: Dynamic Variable Selection for Generalized Linear Models

In this paper, we consider the dynamic effects of variables in the generalized linear model such as logistic regression. This work is motivated by the engineering problem with varying effects of process variables on product quality caused by equipment degradation. To address such a challenge, we propose a penalized dynamic regression model which is flexible to estimate the dynamic coefficient structure. The proposed method considers modeling the functional coefficient parameter as piecewise constant functions. Specifically, under the penalized regression framework, the fused lasso penalty is adopted for detecting the changes in the dynamic coefficients. The group lasso penalty is applied to enable a sparse selection of variables. Moreover, an efficient parameter estimation algorithm is developed based on the alternating direction method of multipliers. The performance of the dynamic regression model is evaluated in numerical studies and three real-world examples.

Presenter: Eva Marie Tourangeau, Los Alamos National Lab
Title: Evaluating Sensitivity of Beryllium Flyer Plate Simulations to Strength Parameterization

Sensitivity analyses allow model users to assess the dependency of model output on variation in parameter inputs and to assess which parameters are most important for determining model prediction. There is considerable interest in understanding the strength properties of Beryllium, a material strategically important to aerospace and defense research. We apply sensitivity analysis to test and refine Beryllium strength models. Preston-Tonks-Wallace (PTW) is an example of a constitutive strength (plasticity) model. Strength models relate a material's plastic deformation, or permanent deformation after applying stress past the elastic limit. Various PTW strength models, with distinct model settings, have been calibrated at low strain rates using Hopkinson bar and quasi-static experiment data, but have vastly different output when extrapolated to high strain rates. Other constitutive models, such as Mechanical Threshold Stress models, may offer yet more predictions for Beryllium's behavior under high strain. Flyer plate experiments measure the free surface velocity of a target material when struck forcefully by an impactor, subjecting the target to high strain. We run the LANL multi-physics code FLAG to simulate the free surface velocity of Beryllium targets in flyer plate experiments. Our research assesses the sensitivity of the simulated free surface velocity to perturbing the simulation's parameter settings, many of which come from the chosen material strength model - specifically, studying how variation in our simulator response can be apportioned to the various code inputs. This investigation in a high strain rate regime will add insight to which strength model parameter settings offer the best fit to experimental data across a range of strain rates, and by extension, and in combination with damage models and EOS information, inform a general material model for Beryllium.

Presenter: Hung-Ping Tung, Virginia Tech
Title: Planning Gamma Accelerated Degradation Tests with Two Accelerating Variables

Gamma accelerated degradation tests (ADT) are widely used to assess timely lifetime information of highly reliable products whose degradation path follows a gamma process. In the literature, several papers attempted to address the decision problem of how to conduct an efficient accelerated degradation test which includes the determinations of higher testing-stress levels and their corresponding sample size allocations. The results mainly focused on the case of a single accelerating variable, however, may not practically applicable when the degradation rate of the quality characteristics of the product is slow. To overcome the difficulty, this paper proposes an analytical approach to address this decision problem under the case of two accelerating variables. Specifically, based on the criterion of minimizing the asymptotic variance of the estimated q quantile of product’s lifetime distribution, we analytically show that the optimum stress levels and the optimum sample size allocations can be simultaneously obtained via general equivalence theorem.

Presenter: Stephen Walsh, Virginia Tech
Title: Validation and Uncertainty Quantification of Forecast Rainfall from Hurricanes and Tropical Storms

Hurricanes and tropical storms significantly impact coastal and inland communities. As of 2018, the National Oceanic and Atmospheric Administration (NOAA) reported three of the top five most costly hurricanes made landfall in 2017. In this study, we examine the error fields of hurricane precipitation forecasts and use spatial analysis to quantify uncertainty in those forecasts. The study includes 48 storms of tropical storm strength or greater that made landfall in the contiguous U.S. between 2004 and 2017. For observations, we use the Stage IV dataset (~4-km, hourly resolution), and for the forecast, we use the ~12km North American Mesoscale (NAM) forecast model. Both datasets are available through NOAA/National Centers for Environmental Prediction (NCEP). After interpolation, we study the spatial correlation structure generated by all precipitation within a 700km buffer from the eye of the storm, as well as precipitation constrained above a threshold. Upon examining disparities between forecasted and observed fields, we shall explore the relationship of variogram parameter estimates based upon storm characteristics including landfall location and intensity.

Presenter: Yueyao Wang, Virginia Tech
Title: Design of Experiments for High-Performance Computing Variability Management

The management of performance variability is an important research area in high-performance computing (HPC). The performance variability is affected by complicated interactions of numerous factors, such as CPU frequency, the number of I/O threads, file size, and record size. In this paper, we focus on the I/O variability, which is measured by the I/O throughputs. To study the performance variability, computer scientists often use full factorial designs to collect I/O throughput data under various system configurations, which becomes prohibitive when the number of factors is not small. In this paper, we propose to use a space filling design to collect I/O throughput data. The advantage of using a space filling design is that it provides good coverage of the design space, many levels for each variable, and good projection properties. We then compare the effectiveness of different design strategies in the setting of HPC performance management. We finally provide some guidelines for computer scientists to collect data using the design of experiments techniques.

Presenter: Qiuyi Wu, Rochester Institute of Technology
Title: Exploratory analysis of Hurricane Storm Surge

Properly assessing the storm surge risk is important for coastal areas. Typically, we run statistical models to obtain the simulation about water level response towards a collection of synthetic hurricane storms, which we also called “input distribution”. In this scenario, it is crucial to have a good estimate of the input distribution for the simulation. We attempt to get better storm surge input distribution for measuring the model of storm evolution. However, the current practice from FEMA’s approach is somewhat ad-hoc and does not account for the uncertainty of the estimation. In this research project, we investigate various methods (e.g., hierarchical model to incorporate different data sources, spatial-temporal structure) to improve the estimation efficiency and to properly quantify the uncertainty. We demonstrated the exploratory analysis of synthetic storm tracks, and compared the simulated tracks with the real storm data. One of the main discussions in this project is how to improve the current practice with the technique of spatial statistics, such as using hierarchical model to improve the estimation of input distribution, or spatial-temporal point process modeling for the storm occurrence rate. We also fusion different source of data to offset the limitation of insufficient real hurricane observations.

Presenter: Nathan Wycoff, Virginia Tech
Title: Sequential Learning of Active Subspaces

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. We show that in the event that the surrogate model is a Gaussian process, the ASM estimator has a closed form, rendering finite differencing 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.

Presenter: Yunnan Xu, Virginia Tech
Title: A novel approach for Raman spectrum baseline correction and its application on real data

Raman spectroscopy plays a crucial role in many studies, from fundamental science to nano-materials and biological systems, and becomes an essential analytical tool. However, background signals generated by the uninterested molecule structures can heavily interfere the analysis of Raman spectra. Therefore, one critical procedure is to perform baseline correction on the raw spectra before analyses in order to eliminate the unwanted background signals and to enhance the discriminating features. Traditional baseline correction methods can have their performance much affected by the choice of cost function and parameters. For example, the popular polynomial fitting is limited by its small number of tuning parameters, and thus it may not sufficiently capture complicated smooth trends and can be easily distorted by a few influential points. We propose a novel approach, iterative smoothing spline with square root error adjustment (ISSA as in Iterative Smoothing spline with Square root error Adjustment), which employs an asymmetric square root error function. ISSA is simple (semi-automatic), efficient and stable. Its implementation is carried out through R, which is free, accessible and easy to use. The algorithm uses smoothing spline and converges very quickly. Meanwhile, it gives relatively good results compared with another baseline correction method, Goldindec, which claims to be superior than polynomial fitting. Our method and Goldindec are performed on both simulated data and real data, which are from complex dialysis samples and simple mineral samples. ISSA shows to have lower computational cost, more accuracy and better stability.

Presenter: Kai Yan, University of Florida
Title: Online Sequential Monitoring Of Spatio-Temporal Disease Incidence Rates

Online sequential monitoring of the incidence rates of chronic or infectious diseases is critically important for public health and stability of our society. Governments around the world have invested a great amount of resource in building global, national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charts, are usually included for disease surveillance purpose. However, these charts require many assumptions on the observed data, including the ones that the observed data are independent and identically normally distributed when no disease outbreaks are present. These assumptions can hardly be valid in practice, making the results from the conventional control charts unreliable. Motivated by an application to monitor the Florida influenza-like illness data, we develop a new sequential monitoring approach in this paper, which can accommodate the dynamic nature of the disease incidence rates (i.e., the disease incidence rates change over time due to seasonality and other reasons), spatio-temporal data correlation, and non-normality. It is shown that the new method is much more reliable to use in practice than the commonly used conventional charts for sequential monitoring of disease incidence rates. The proposed method should be useful for many other applications, such as spatio-temporal monitoring of air quality in a region and monitoring of sea-level pressure in oceanography.

Presenter: Henry Shaowu Yuchi, Georgia Tech
Title: Uncertainty Quantification in Low-rank Matrix Completion

Low-rank matrices are widely used in modelling scientific and engineering problems. In many such problems, however, only a small portion of the corresponding matrix information can be observed. There are two major reasons for this phenomenon. The first is the cost of obtaining a whole set of observations could be high. It could be computationally expensive and challenging to observe all matrix entries due to experimental and communication costs. The other is that there can be missing observations in the problem procedure itself, for example sensor malfunction, experiment failure and communication error. Carrying out inference on the matrix X with limited observations, it becomes increasingly important not only to provide point estimates of unobserved entries, but also a measure of uncertainty (or uncertainty quantification, UQ) for the estimates. To address the uncertainty quantification task along with the matrix completion problem, BayeSMG model is proposed. It constructs a Bayesian model to achieve the aforementioned goals while factoring in matrix subspace information.

Presenter: Boya Zhang, Virginia Tech
Title: Distance-distributed design for Gaussian process surrogates

A common challenge in computer experiments and related fields is to efficiently explore the input space using a small number of samples, i.e., the experimental design problem. Much of the recent focus in the computer experiment literature, where modeling is often via Gaussian process (GP) surrogates, has been on space-filling designs, via maximin distance, Latin hypercube, etc. However, it is easy to demonstrate empirically that such designs disappoint when the model hyperparameterization is unknown, and must be estimated from data observed at the chosen design sites. This is true even when the performance metric is prediction-based, or when the target of interest is inherently or eventually sequential in nature, such as in blackbox (Bayesian) optimization. Here we expose such inefficiencies, showing that in many cases purely random design is superior to higher-powered alternatives. We then propose a family of new schemes by reverse engineering the qualities of the random designs which give the best estimates of GP lengthscales. Specifically, we study the distribution of pairwise distances between design elements, and develop a numerical scheme to optimize those distances for a given sample size and dimension. We illustrate how our distance-based designs, and their hybrids with more conventional space-filling schemes, outperform in both static (one-shot design) and sequential settings.