IMS/ASA Spring Research Online Meet

May 19-20, 2022 (Thursday and Friday)

DAY 1 (May 19)
----

12 PM to 12:05 PM
Opening remarks by Xinwei Deng, Virginia Tech, Co-Chair, Organizing Committee

12:05 PM to 1:10 PM
Keynote speech by **C. F. Jeff Wu**, Georgia Institute of Technology
Analysis-of-Marginal-Tail-Means (ATM): A Robust Method for Discrete Black-Box Optimization

Chair: Ying Hung, Rutgers University

1:15 PM to 2:45 PM
Technical Session I: Computer Experiments, Gaussian Processes and Bayesian models
Chair: Nathaniel Stevens, University of Waterloo

- **Annie Sauer**, Virginia Tech
  Vecchia-approximated Deep Gaussian Processes for Computer Experiments

- **John Yanotty**, Ohio State University
  Model Mixing with Bayesian Additive Regression Trees

- **Irene Ji**, Duke University
  A graphical multi-fidelity Gaussian process model, with application to emulation of expensive computer simulations

- **Ruda Zhang**, Duke University
  Gaussian Process Subspace Prediction for Dimension Reduction of Computational Models

- **Cheoljoon Jeong**, University of Michigan, Ann Arbor
  Multi-block Parameter Calibration in Computer Models

3 PM to 4:30 PM
Panel Discussion I: Careers in the academia and industry after obtaining a doctoral degree in Statistics
Panelists: **Kimberly Ann Kaufeld** (Los Alamos), **Nicole Pashley** (Rutgers University), **Simon Mak** (Duke University), **Jean Pouget-Abadie** (Google)
Moderator: Tirthankar Dasgupta, Rutgers University

DAY 2 (May 20)
----

12 PM to 1:20 PM
Panel Discussion II: Future of the three traditional pillars of industrial statistics: Statistical Process Monitoring and Control, Reliability and Design of Experiments.

Panelists: Allison Jones Farmer (Miami University, Oxford), **Bill Meeker** (Iowa State University), **Bradley Jones** (JMP).

Moderator: Arman Sabbaghi, Purdue University

1:30 PM to 3:20 PM
Technical Session II: Statistical methodology and applications in today's science, industry engineering and sports
Chair: **David Stenning**, Simon Fraser University

- **Vincent M. Geels**, Ohio State University
  A Tree-Based Transformation Approach for Modeling Counts

- **Chaofan Huang,** Georgia Tech
  Constrained Minimum Energy Designs

- **Shubhajit Sen**, NCSU
  A Flexible Bayesian Regression Approach for Modeling Interval Data.

- **Jie Min**, Virginia Tech
  Reliability Analysis of Artificial Intelligence Systems Using Recurrent Events Data from Autonomous Vehicles

- **Nirodha Epasinghege Dona**, Simon Fraser University
  Expected Economy Rate

- **Zhengzhi Lin**, Virginia Tech
  The Poisson Multinomial Distribution and Its Applications in Voting Theory, Ecological Inference, and Machine Learning

Wrap up: 3:20 - 3:30 PM
Closing Remarks: Tirthankar Dasgupta, Rutgers University, Chair, Program Committee

TALK ABSTRACTS:

- **Annie Sauer**, Virginia Tech
  Vecchia-approximated Deep Gaussian Processes for Computer Experiments

  Deep Gaussian processes (DGPs) upgrade ordinary GPs through functional composition, in which intermediate GP layers warp the original inputs, providing flexibility to model non-stationary dynamics. Two DGP regimes have emerged in recent literature. A "big data" regime, prevalent in machine learning, favors approximate, optimization-based inference for fast, high-fidelity prediction. A "small data" regime, preferred for computer surrogate modeling, deploys posterior integration for enhanced uncertainty quantification (UQ). We aim to bridge this gap by expanding the capabilities of Bayesian DGP posterior inference through the incorporation of the Vecchia approximation, allowing linear computational scaling without compromising accuracy or UQ. We are motivated by surrogate modeling of simulation campaigns with upwards of 100,000 runs - a size too large for previous fully-Bayesian implementations – and demonstrate prediction and UQ superior to that of ``big data" competitors. All methods are implemented in the "deepgp" package on CRAN.

- **John Yanotty**, Ohio State University
  Model Mixing with Bayesian Additive Regression Trees

  In modern computer experiments applications, one often encounters the situation where many plausible models of a physical system are considered, each implemented as a simulator on a computer. An important question in such a setting is determining the best simulator, or the best combination of simulators, to use for prediction or inference. Bayesian model averaging (BMA) and stacking are relevant statistical modeling approaches typically used to account for model uncertainty. In the context of multi-simulator computer experiments, BMA or stacking can be used to combine a set of plausible simulators using a weighted average, where the weights are independent of the input space. Bayesian model mixing (BMM) extends these ideas to capture the localized behavior of each simulator by allowing the weights to depend on the inputs. One possibility is to define this relationship using a flexible nonparametric model that learns the local strengths and weaknesses of each simulator. In our approach, we propose a BMM model using Bayesian Additive Regression Trees (BART). Here, the mean function is defined as a linear combination of the simulators, where each weight is a function of the inputs. To further increase this model's flexibility, the weights are unconstrained rather than defined on the simplex. With BART, the vector of weights is designed as a sum-of-trees, where each tree is a weak learner that slightly adjusts the value of each weight. We demonstrate our method on simulators from a motivating nuclear physics application.

- **Irene Ji**, Duke University
  A graphical multi-fidelity Gaussian process model, with application to emulation of expensive computer simulations

With advances in scientific computing and mathematical modeling, complex phenomena can now be reliably simulated. Such simulations can however be very time-intensive, requiring millions of CPU hours to perform. One solution is multi-fidelity emulation, which uses data of varying accuracies (or fidelities) to train an efficient predictive model (or emulator) for the expensive simulator. In complex problems, simulation data with different fidelities are often connected scientifically via a directed acyclic graph (DAG), which is difficult to integrate within existing multi-fidelity emulator models. We thus propose a new Graphical Multi-fidelity Gaussian process (GMGP) model, which embeds this DAG (capturing scientific dependencies) within a Gaussian process framework. We show that the GMGP has desirable modeling traits via two Markov properties, and admits a scalable formulation for recursive computation of the posterior predictive distribution along sub-graphs. The advantages of the GMGP model are then demonstrated via a suite of numerical experiments and an application to emulation of heavy-ion collisions, which can be used to study the conditions of matter in the Universe shortly after the Big Bang.

- **Ruda Zhang**, Duke University
  Gaussian Process Subspace Prediction for Dimension Reduction of Computational Models

  Subspace-valued functions arise in a wide range of problems including parametric reduced order modeling (PROM). In PROM, each parameter point can be associated with a subspace, which is used for Petrov-Galerkin projections of system matrices. Previous efforts to approximate such functions use interpolations on manifolds, which can be inaccurate and slow. We propose a novel Bayesian nonparametric model for subspace prediction: the Gaussian Process Subspace (GPS) model. This method is extrinsic and intrinsic at the same time: with multivariate Gaussian distributions on the Euclidean space, it induces a joint probability model on the Grassmann manifold, the set of fixed-dimensional subspaces. The GPS adopts a simple yet general correlation structure, and a principled approach for model selection. Its predictive distribution admits an analytical form, which allows for efficient subspace prediction over the parameter space. For PROM, the GPS provides a probabilistic prediction at a new parameter point that retains the accuracy of local reduced models, at a computational complexity that does not depend on system dimension, and thus is suitable for online computation. We give four numerical examples to compare our method to subspace interpolation, as well as two methods that interpolate local reduced models. Overall, GPS is the most data efficient, more computationally efficient than subspace interpolation, and gives smooth predictions with uncertainty quantification.

- **Cheoljoon Jeong**, University of Michigan, Ann Arbor
  Multi-block Parameter Calibration in Computer Models

  Parameter calibration aims to estimate unobservable parameters employed in a computer model by utilizing physical process responses and computer model outputs. Existing studies calibrate all parameters simultaneously using an entire dataset. However, in certain applications, some parameters are associated with only a subset of data. This study provides a multi-block calibration approach that considers such heterogeneity. Unlike existing studies that build emulators for the computer model response, we consider multiple loss functions, each for a block of parameters that use the corresponding dataset and estimate the parameters using a nonlinear optimization technique. The superiority of our approach is demonstrated through numerical studies and a building energy simulation case study.

- **Vincent M. Geels**, Ohio State University
  A Tree-Based Transformation Approach for Modeling Counts

  A common statistical modeling strategy lies in leveraging advantageous model factorizations in order to fit a collection of simpler submodels. Such a strategy is employed in Bayesian Additive Regression Tree (BART) models, which use collections of individual tree models to estimate functions of interest via an additive representation. While the BART framework has enjoyed success as a method for modeling continuous and binary data, there are relatively few Bayesian regression tree methodologies designed to handle count data. We propose a novel framework for modeling counts wherein the count response is transformed from base-10 representation into a base-q vector representation. These base-q vectors are then fit using a collection of Bayesian classification tree submodels, and we describe model formulations for handling within vector independence and correlations in the setting where q=2. The resulting models are highly flexible in their ability to handle features commonly expressed in count distributions, including zero-inflation and overdispersion, but also less common features such as multimodality. We showcase advantages of the proposed base-2 model in modeling real datasets that fail to be adequately described by models using typical count likelihood functions such as the Poisson and Negative Binomial.

- **Chaofan Huang,** Georgia Tech
  Constrained Minimum Energy Designs

  Space-filling designs are important in computer experiments, which are critical for building a cheap surrogate model that adequately approximates an expensive computer code. Many design construction techniques in the existing literature are only applicable for rectangular bounded space, but in real world applications, the input space can often be non-rectangular because of constraints on the input variables. One solution to generate designs in a constrained space is to first generate uniformly distributed samples in the feasible region, and then use them as the candidate set to construct the designs. Sequentially Constrained Monte Carlo (SCMC) is the state-of-the-art technique for candidate generation, but it still requires large number of constraint evaluations, which is problematic especially when the constraints are expensive to evaluate. Thus, to reduce constraint evaluations and improve efficiency, we propose the Constrained Minimum Energy Design (CoMinED) that utilizes recent advances in deterministic sampling methods. Extensive simulation results on 15 benchmark problems with dimensions ranging from 2 to 13 are provided for demonstrating the improved performance of CoMinED over the existing methods.

- **Shubhajit Sen**, NCSU
  A Flexible Bayesian Regression Approach for Modeling Interval Data.

  We propose a novel method for modeling the interval data. In particular, the relationship between an interval-valued response and a set of interval-valued predictors is investigated by considering a joint regression model, one for the centers (the locations of the intervals) of the response and predictors, and the other one for the radii (the imprecision). Previous works on this problem either cannot obtain different regression coefficients for the center and the radii, or they do not consider the joint structure for modeling. Our model overcomes these drawbacks since both the centers and the radii of the predictors are used to model both the center and the radius of the response with the flexibility of identifying the different effects of the center and radius of a predictor on the response, along with accounting for the dependence between the center and the radius. We develop a Bayesian estimation method, with an automated feature screening for selecting the most important predictors in the model using ”slab and spike” and ”local-global shrinkage” priors. We assess the accuracy, precision, and the predictive power of the proposed model through extensive simulation studies and analysis of a dataset obtained from a clinical trial conducted for estimating the efficacy of two standard drugs on the children suffering from acute lymphocytic leukemia (ALL) in the eastern part of India

- **Jie Min**, Virginia Tech
  Reliability Analysis of Artificial Intelligence Systems Using Recurrent Events Data from Autonomous Vehicles

  Artificial intelligence (AI) systems have become increasingly common and the trend will continue. Examples of AI systems include autonomous vehicles (AV), computer vision, natural language processing, and AI medical experts. To allow for safe and effective deployment of AI systems, the reliability of such systems needs to be assessed. Traditionally, reliability assessment is based on reliability test data and the subsequent statistical modeling and analysis. The availability of reliability data for AI systems, however, is limited because such data are typically sensitive and proprietary. The California Department of Motor Vehicles (DMV) oversees and regulates an AV testing program, in which many AV manufacturers are conducting AV road tests. Manufacturers participating in the program are required to report recurrent disengagement events to California DMV. This information is being made available to the public. In this paper, we use recurrent disengagement events as a representation of the reliability of the AI system in AV, and propose a statistical framework for modeling and analyzing the recurrent events data from AV driving tests. We use traditional parametric models in software reliability and propose a new nonparametric model based on monotonic splines to describe the event process and to estimate the cumulative baseline intensity function of the event process. We develop inference procedures for selecting the best models, quantifying uncertainty, and testing heterogeneity in the event process. We then analyze the recurrent events data from four AV manufacturers, and make inferences on the reliability of the AI systems in AV. We also describe how the proposed analysis can be applied to assess the reliability of other AI systems. This paper has online supplementary materials.

- **Nirodha Epasinghege Dona**, Simon Fraser University
  Expected Economy Rate

  We introduce the expected goals concept to limited overs cricket where ideas are illustrated using the economy rate statistic. The approach is primarily explored as a proof of concept since the detailed data that are required for full adoption of the proposed methods are not currently widely available. The approach is based on the estimation of batting outcome probabilities given detailed data on each ball that is bowled in a match. Machine learning techniques are used for the estimation procedure. Some differences between men's and women's T20 cricket are subsequently identified including the greater rate at which 6's occur in the men's game.

- **Zhengzhi Lin**, Virginia Tech
  The Poisson Multinomial Distribution and Its Applications in Voting Theory, Ecological Inference, and Machine Learning

  The Poisson multinomial distribution (PMD) describes the distribution of the sum of n independent but non-identically distributed random vectors, in which each random vector is of length m with 0/1 valued elements and only one of its elements can take value 1 with a certain probability. Those probabilities are different for the m elements across the n random vectors, and form an n times m matrix with row sum equals to 1. We call this n times m matrix the success probability matrix (SPM). Each SPM uniquely defines a PMD. The PMD is useful in many areas such as voting theory, ecological inference, and machine learning. The distribution functions of PMD, however, are usually difficult to compute. In this paper, we develop efficient methods to compute the probability mass function (pmf) for the PMD using multivariate Fourier transform, normal approximation, and simulations. We study the accuracy and efficiency of those methods and give recommendations for which methods to use under various scenarios. We also illustrate the use of the PMD via three applications, namely, in voting probability calculation, aggregated data inference, and uncertainty quantification in classification. We build an R package that implements the proposed methods and illustrates the package with examples.