SAMSI Numerical Analysis in Data Science Transition Virtual Workshop
May 24-26, 2021
Titles & Abstracts

A. TUTORIALS (Alphabetical by last name)

Speaker: Alen Alexanderian
Department of Mathematics
North Carolina State University
Title: Optimal Design of Experiments for Large-scale Bayesian Inverse Problems
Abstract: In this presentation, we will review methods and theory for optimal experimental design for Bayesian inverse problems. The focus will be on large-scale inverse problems governed by PDEs with infinite-dimensional (high-dimensional when discretized) parameters. We will discuss some of the challenges and methods for tackling such problems in the context of Bayesian linear inverse problems and will outline the challenges and opportunities for design of large-scale nonlinear inverse problems. Illustrative numerical results will also be provided in the context of model inverse problems.

Speaker: Pierre Gremaud
Department of Mathematics
North Carolina State University
Title: A Biased Introduction to Sensitivity Analysis
Abstract: Sensitivity analysis (SA) methods can provide lots of answers; it is however not always obvious what the questions were. In this presentation, we will introduce variance based methods for SA and use of them to illustrate the reach and limitations of sensitivity analysis. In particular, we will consider the issues of robustness, dimension reduction, and use of surrogate models. These points will be illustrated through concrete examples.

Speaker: Ilse C. F. Ipsen
Department of Mathematics
North Carolina State University
Title: The Ideas behind Randomized Algorithms for Least Squares Problems
Abstract: We review randomized algorithms for the solution of least squares/regression problems, based on row sketching from the left, and column sketching from the right; as well as a brief discussion of coordinate descent (Kaczmarz) methods. These algorithms tend to be efficient and accurate for matrices that have many more rows than columns.
We present probabilistic bounds for sampling, given a user-specified error tolerance. Along the way we examine the effect of sampling on statistical model uncertainty, how to do randomized sampling, and illustrate important concepts from numerical analysis (conditioning and pre-conditioning), and probability (coherence, concentration inequalities). Numerical experiments illustrate that the probabilistic bounds are informative even for small problem dimensions and stringent success probabilities.

Speaker: Ralph Smith
Department of Mathematics
North Carolina State University
Title: Bayesian Inference and Uncertainty Propagation for Physical and Biological Models
Abstract: The quantification of uncertainties inherent to models, parameters, and experimental data is critical to assess the accuracy of model-based predictions. In this tutorial, we will discuss issues that must be addressed when quantifying uncertainties inherent to parameters and data and propagating these uncertainties through models to determine the accuracy of responses. We will first discuss the use of Bayesian techniques to infer parameter distributions. For uncertainty propagation, we will discuss the role of surrogate models for complex simulation codes. We will also note the manner in which sensitivity analysis can be employed for parameter selection prior to inference and uncertainty analysis. Current research directions will be noted throughout the tutorial.

B. PRESENTATIONS BY WORKING GROUPS (Alphabetical by last name)

Speakers: Eric Acquesta & Mike Smith
Sandia National Laboratories
Title: All Explanations are Wrong, but Some are Useful: Using Global Sensitivity Analysis to Identify Gaps in Machine Learning Explainability
Abstract: Machine learning (ML) explainability (MLE) has been proposed as a means of generating trust in a learned model and as an important indicator in determining its reliability for responsible and ethical artificial intelligence. However, little analysis has been performed to determine *if* the explanations accurately represent the target model and *should* be trusted beyond subjective inspection. We examine the problem of verifying the fidelity of explanations through global sensitivity analysis (GSA), specifically Sobol' indices and Shapley values. We map GSA onto the ML process and identify three main gaps in directly applying GSA to ML: (1) how to sample non-Gaussian, discrete, correlated and sparse data distributions, (2) identifying an appropriate quantity of interest (QoI) for which GSA will provide insight for MLE, and (3) methods that apportion the influence of sources of input uncertainty across output uncertainty, accounting for higher-order interactions in a model and input correlations. We demonstrate that current state-of-the-art MLE methods often give misleading explanations due to these same limitations. In closing, we provide some possible direction for future research to address these gaps. *Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2021-5991 A*

Speakers: Priyan de Alwis and S. Yaser Samdi
School of Mathematical and Statistical Sciences
Southern Illinois University, Carbondale, Illinois, USA
Title: Fourier Method on Sufficient Dimension Reduction in Time Series
Abstract: Dimensionality reduction has always been one of the most important and challenging problems in high-dimensional data analysis. In the context of time series analysis, we are

interested in estimating and making inference about the conditional mean and variance functions. Using the central mean and variance dimension reduction subspaces, that preserve sufficient information about the response, one can estimate the unknown mean and variance functions of the time series. There are a few approaches in the literature to estimate the time series central mean subspace (TS-CMS). However, those methods are computationally intensive and not feasible in practice. Using the Fourier transform, an explicit estimate of the time series central mean subspace is obtained. The proposed estimators are shown to be consistent, asymptotically normal and efficient. Simulation studies are conducted to evaluate the performance of the proposed method. The results show that our method is significantly more accurate, and computationally substantially faster than the existing method. The method is applied to the Canadian Lynx dataset.

Speaker: Grey Ballard
Department of Computer Science
Wake Forest University
Title: Randomized Algorithms for Rounding the Tensor Train Format
Abstract: The Tensor Train (TT) format is a highly compact low-rank representation for high-dimensional tensors. TT is useful in particular in representing approximations to the solution of certain types of parametrized partial differential equations. For many of these problems, computing the solution explicitly would require an infeasible amount of memory and computational time. While the TT format makes these problems tractable, iterative techniques for solving the PDEs must be adapted to perform arithmetic while maintaining the implicit structure. The fundamental operation used to maintain feasible memory and computational time is called rounding, which truncates the internal ranks of a tensor already in TT format. We propose several randomized algorithms for this task that are generalizations of randomized low-rank matrix approximation algorithms and provide significant reduction in computation compared to deterministic TT rounding algorithms. Randomization is particularly effective in the case of rounding a sum of TT tensors, which is the bottleneck computation in the adaptation of GMRES to vectors in TT format. In this talk, we will present the randomized algorithms and compare their empirical accuracy and computational time with deterministic alternatives.

Speaker: Taewon Cho
Department of Mathematics
Virginia Tech
Title: Hybrid Projection Method for Large-scale Inverse Problems with Mean Estimation in Hierarchical Gaussian Priors
Abstract: In many inverse problems, Bayesian methods are used for computing solution approximations and for uncertainty quantification. A common assumption is to treat the unknowns as random variables from a multivariate Gaussian distribution, which is defined by a prior mean vector and a prior covariance matrix. These prior assumptions on the random variables are used in the Bayesian framework to obtain point estimates (e.g., the maximum a posterior (MAP) estimate) and to measure uncertainties (e.g., variance estimates or sampling). A major challenge arises when the prior itself is not known precisely. In this talk, we consider the scenario where the prior is Gaussian with a potentially complicated covariance matrix structure,

and additionally, the mean vector is unknown and must be estimated from the data. We assume that some subspace of basis columns for the prior mean vector is available and approximates the unknown coefficient vector for that basis. We treat these unknowns as Gaussian random variables and use a hierarchical model to simultaneously estimate the desired solution vector and the mean coefficient vector. Using a change of variables, we show that the MAP estimate can be obtained by solving the linear inverse problem with a modified covariance matrix, and we use generalized hybrid projection methods to solve the optimization problem efficiently and with automatic regularization parameter selection.

Speaker: Matthias Chung
Department of Mathematics
Virginia Tech
Title: Learning Regularization Parameters of Inverse Problems via Deep Neural Networks
Abstract: In this work, we describe a new approach that uses deep neural networks (DNN) to obtain regularization parameters for solving inverse problems. We consider a supervised learning approach, where a network is trained to approximate the mapping from observation data to regularization parameters. Once the network is trained, regularization parameters for newly obtained data can be computed by efficient forward propagation of the DNN. We show that a wide variety of regularization functionals, forward models, and noise models may be considered. The network-obtained regularization parameters can be computed more efficiently and may even lead to more accurate solutions compared to existing regularization parameter selection methods. We emphasize that the key advantage of using DNNs for learning regularization parameters, compared to previous works on learning via optimal experimental design or empirical Bayes risk minimization, is greater generalizability. That is, rather than computing one set of parameters that is optimal with respect to one particular design objective, DNN-computed regularization parameters are tailored to the specific features or properties of the newly observed data. Thus, our approach may better handle cases where the observation is not a close representation of the training set. Furthermore, we avoid the need for expensive and challenging bilevel optimization methods as utilized in other existing training approaches. Numerical results demonstrate the potential of using DNNs to learn regularization parameters.
Authors: Babak Maboudi Afkham, Julianne Chung, and Matthias Chung

Speaker: Silvia Gazzola
Department of Mathematics
University of Bath
Title: Learning Methods for Inverse Problems
Abstract: In this talk we will present various approaches that use learning from training data, following an optimal experimental design framework, to solve inverse problems. We consider a general framework, where training data can be used to obtain superior regularization methods. This may refer to optimal regularization parameters, optimal data fidelity terms, and optimal regularizers. Our investigation also includes optimal diagonal transformations for $\ell_p-\ell_q$ regularization, and scenarios where noise comes from non-Gaussian distributions or where some inexactness affects the forward operator. We consider efficient approaches to perform optimal estimation, which is suitable for large-scale problems.

Speakers: Wiranthe Herath and S. Yaser Samdi
School of Mathematical and Statistical Sciences
Southern Illinois University, Carbondale, Illinois, USA
Title: Dimension Reduction for Vector Autoregressive Models
Abstract: The classical vector autoregressive (VAR) models have been widely used to model multivariate time series data, because of their flexibility and ease of use. However, the VAR model suffers from overparameterization particularly when the number of lags and number of time series get large. There are several statistical methods of achieving dimension reduction of the parameter space inVAR models. Reduced-rank VAR model (Velu et al., 1986; Reinsel and Velu, 2013) can be used to restrict the rank of the parameter matrix in one direction. Multilinear low-rank VAR modeling with tensor decomposition (Wang et al., 2021) is a new method that can be used to limit the parameter space across three directions. Another effective approach is to consider the minimal reducing subspace when constructing a relation between the mean function of VAR and the covariance matrix of the VAR (Wang and Ding, 2018) that arises from the envelope method developed by Cook et. al (2010). In this talk we review and compare these methods.

Speakers: Hadi Safari Katesari and S. Yaser Samdi
School of Mathematical and Statistical Sciences
Southern Illinois University, Carbondale, Illinois, USA
Title: Bayesian Copula Factor Autoregressive Models for Time Series Mixed Data
Abstract: In this talk, we propose Bayesian copula factor autoregressive models for time series mixed data. This is a novel model that assumes conditional independence and applies latent factors in both response time series and in high-dimensional mixed-type covariates of the time series with the quadratic regression model. The framework of the model gives an efficient dimension reduction and characterizes the main effects and interactions of the covariates by including the latent variables in the response time series. We apply a semiparametric time series extended rank likelihood for the margins of explanatory variables which reduces the number of estimated parameters and provides a fast computational algorithm. A flexible Bayesian algorithm is proposed to compute the posterior distribution of latent factors and model parameters with Metropolis-Hastings and Forward Filtering Backward methods within Gibbs sampling. The theoretical results and MCMC computations are evaluated with simulation studies. Moreover, the proposed model is applied to the quarterly U.S. economic dataset.

Speakers: Agnes Lagnoux & Thierry Klein
Université de Toulouse
Title: A Novel Generation of Mighty Estimators based on Rank Statistics
Abstract: We propose a new statistical estimation framework for a large family of global sensitivity analysis indices. Our approach is based on rank statistics and uses an empirical correlation coefficient recently introduced by Chatterjee in 2020. We show how to apply this approach to compute not only the Cramér-von-Mises indices, which are directly related to Chatterjee's notion of correlation, but also first-order Sobol indices, general metric space indices and higher-order moment indices. We establish consistency of the resulting estimators and

demonstrate their numerical efficiency, especially for small sample sizes. In addition, we prove a central limit theorem for the estimators of the first-order Sobol indices.

Speaker: **Mike Merritt**
Department of Mathematics
North Carolina State University
Title: Efficient Global Sensitivity Analysis for Rare Event Simulation
Abstract: We consider the task of performing global sensitivity analysis (GSA) for both high-dimensional and expensive-to-evaluate models. While surrogate models have proven to be useful in this context, they typically fail to capture the tail behavior of the QoI. Our goal is to characterize the sensitivity of a rare event probability to the hyperparameters that define the distribution law of the uncertain parameters. This so-called "second level" sensitivity analysis requires additional evaluations of the rare event probability. In this talk, we begin by discussing the Subset Simulation method, which can significantly reduce the cost of estimating the rare event probabilities. Numerical results from porous media flow will be presented to illustrate this point. We will then discuss additional challenges in performing second level GSA using surrogate models.

Speaker: Agnieszka Miedlar
Department of Mathematics
University of Kansas
Title: Randomized FEAST Algorithm for Generalized Hermitian Eigenvalue Problems with Probabilistic Error Analysis
Abstract: Randomized NLA methods have recently gained popularity because of their easy implementation, computational efficiency, and numerical robustness. We propose a randomized version of a well-established FEAST eigenvalue algorithm that
enables computing the eigenvalues of the Hermitian matrix pencil $(\textbf{A},\textbf{B})$ located in the given real interval $\mathcal{I} \subset [\lambda_{min}, \lambda_{max}]$. In this talk we will present deterministic as well as probabilistic error analysis of the accuracy of approximate eigenpair and subspaces obtained using the randomized FEAST algorithm.
First, we derive bounds for the canonical angles between the exact and the approximate eigenspaces corresponding to the eigenvalues contained in the interval $\mathcal{I}$. Then, we present bounds for the accuracy of the eigenvalues and the corresponding eigenvectors. This part of the analysis is independent of the particular distribution of an initial subspace, therefore we denote it as deterministic or structural. In the case of the starting guess being a Gaussian random matrix, we provide more informative, probabilistic error bounds. Finally, we will illustrate, numerically, the effectiveness of all the proposed error bounds.
This is a joint work with E. de Sturler, N. Kapur and A. K. Saibaba.

Speaker: **Houssam Nassif**
Amazon Inc
Title: Solving Inverse Reinforcement Learning, Bootstrapping Bandits,
and Adaptive Recommendation

Abstract: This talk discusses three different ways we leveraged reward signals to inform recommendation. In Deep PQR: Solving Inverse Reinforcement Learning using Anchor Actions (ICML’20), we use deep energy-based policies to recover the true reward function in an Inverse Reinforcement Learning setting. We uniquely identify the reward function by assuming the existence of an anchor action with known reward, for example a do-nothing action with zero reward. In Decoupling Learning Rates Using Empirical Bayes (under review, arXiv), we devise an Empirical Bayes formulation that extracts an unbiased prior in hindsight from an experiment’s early reward signals. We apply this empirical prior to warm-start bandit recommendations and speed up convergence. In Seeker: Real-Time Interactive Search (KDD’19), we introduce a recommender system that adaptively refines search rankings in real time, through user interactions in the form of likes and dislikes. We extend Boltzmann bandit exploration to adapt to the interactively changing embedding space, and to factor-in the uncertainty of the reward estimates.

Speaker: Elizabeth Newman
Department of Mathematics and Computer Science
Emory University
Title: Accelerating Stochastic Optimization of Separable Deep Neural Networks via Iterative Sampling Methods
Abstract: Deep neural networks (DNNs) are versatile machine learning tools, but can be difficult to train to sufficient accuracy with first-order methods. In this talk, we accelerate training by exploiting the separability of most DNN architectures; that is, we separate the DNN into a nonlinear feature extractor followed by a linear model. We use the fact that the weights of the linear model can be obtained using convex optimization, which reduces the training problem. However, this approach has a proclivity to overfit, and hence requires training on sufficiently large samples of data. This can be prohibitively expensive storage-wise, particularly on GPUs. To overcome this limitation and the expensive task of hyperparameter tuning, we use an iterative sampled limited-memory Tikhonov method, a memory-efficient stochastic approximation approach with automatic regularization parameter tuning. We show that our method can train DNNs to high accuracy with minimal computational overhead through several numerical experiments.

Speaker: Mirjeta Pasha
Department of Mathematics
Arizona State University
Title: Efficient Edge-preserving and Sparsity Promoting Methods for Large-scale, Time-dependent, Dynamic Inverse Problems
Abstract: In this talk we consider efficient methods for computing solutions to and estimating the maximum a posteriori (MAP) for time dependent dynamic inverse problems, where the target of interest changes during the measurement process as well as the operator that models the forward problem may change at different time instances. For example, the data and the forward operator may be obtained by limited angle tomography that make the problem to solve challenging from both it's dynamic nature and the limited amount of data available. Moreover, incorporating spatial and temporal information about the prior and providing properties like

edge-preserving can be computationally not attractive. Hence, we propose methods that consider spatial and temporal information and provide solutions with edge-preserving and sparsity promoting properties at low cost and enhanced accuracy. The problems of interest for us are ill-posed inverse problems, whose solution if it exists is unstable to the perturbations in the data. To remedy this difficulty regularization methods are used to replace the original ill-posed with a nearby well posed problem whose solution is less sensitive to the perturbations in the data. We consider regularizers that enforce simultaneous regularization in space and time, and that typically enhance edges (at each time instant) and enforce proximity (at consecutive time instants) by total variation (TV) and group sparsity. The methods that we develop here are iterative methods based on majorization minimization strategy with quadratic tangent majorant that allow the resulting least squares problem to be solved with a generalized Krylov subspace method for large scale problems. The regularization parameter can be defined automatically and at a low cost in the projected subspaces of a relatively small dimension. Numerical examples from a wide range of applications like limited angle computerized tomography (CT), space-time deblurring, and photoacoustic tomography (PAT) illustrate the effectiveness of the described approaches.

Speaker: **Kate Pearce**
Department. of Mathematics
North Carolina State University
Title: A Tailored Parameter Identifiability Approach for Polymerization Models in Wound Healing Applications
Abstract: We consider parameter estimation for an enzyme kinetics model of in vitro fibrin matrix polymerization, accounting for dynamic interactions among fibrinogen, thrombin, fibrin, and intermediate complexes. A tailored parameter subset selection technique is presented to evaluate parameter identifiability with a representative data curve for fibrin accumulation in a short-duration in vitro polymerization experiment. Robustness of this approach will be discussed in the context of objective function cost, parameter sensitivities, model reduction, and the high degree of information within a single fibrin accumulation curve. Our approach incorporates underlying characteristic time scales of the model into a systematic exploration of identifiable (and unidentifiable) parameter subsets across the parameter space, improving robustness of the parameter estimation problem via a significant reduction in objective cost.

Speaker: **William Reese**
Department of Mathematics
North Carolina State University
Title: Bedrock Inversion and Hyper Differential Sensitivity Analysis for the Shallow Ice Model
Abstract: Developing high fidelity ice sheet models is important for modeling global climate and predicting sea-level rise. Modeling large ice sheets such as Antartica or Greenland, require extensive computational resources to solve a system of nonlinear PDEs on fine scale meshes. Additionally, there are uncertainties associated with spatially distributed model parameters such as basal sliding, bedrock topography, and source terms. In this work we utilize Hyper Differential Sensitivity Analysis (HDSA) to provide insight into the sensitivity of estimated bedrock topography with respect to other model parameters. This is accomplished by solving an inverse

problem for bedrock topography constrained by the shallow ice model and then differentiating through the optimality system to determine the influence of the other model parameters on the estimated bedrock topography. Our analysis is performed on a 550 x 450 km region of Greenland over a time interval of 10 years.

Speakers: **Abdolnasser Sadeghkhani**, ITAM-CDMX-Mexico, nasser@itam.mx and **S. Yaser Samdi**, Southern Illinois University,
Title: Compressed Bayesian Predictive Inference for Time Series Count Data
Abstract: In this talk we extend the results of Guhaniyogi and Dunson (2015) regarding the compressed Bayesian regression approach to the time series count data in a high dimensional setting where p>>n. We use the Conway-Maxwell-Poisson distribution which is demonstrated as a viable alternative for real count data that express data over- or under-dispersion. We derive the Bayesian predictive distribution of the future count data corresponding to the huge p computationally fast by avoiding the MCMC methods thanks to the exact posterior distribution, in comparison to other variable selection or dimension reduction in high dimensional time series. We also show it outshines more specifically, for dense models.

Speaker: **Zhaoran Wang**
Northwestern University
Title: Is Pessimism Provably Efficient for Offline RL?
Abstract: Coupled with powerful function approximators such as deep neural networks, reinforcement learning (RL) achieves tremendous empirical successes. However, its theoretical understandings lag behind. In particular, it remains unclear how to provably attain the optimal policy with a finite regret or sample complexity.
In the offline setting, we aim to learn the optimal policy based on a dataset collected a priori. Due to a lack of active interactions with the environment, we suffer from the insufficient coverage of the dataset. To maximally exploit the dataset, we propose a pessimistic least-squares value iteration algorithm, which achieves a minimax-optimal sample complexity.

Speaker: **Zhuoran Yang**
Department of Operations Research and Financial Engineering
Princeton University
Title: On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces
Abstract: The classical theory of reinforcement learning (RL) has focused on tabular and linear representations of value functions. Further progress hinges on combining RL with modern function approximators such as kernel functions and deep neural networks, and indeed there have been many empirical successes that have exploited such combinations in large-scale applications. There are profound challenges, however, in developing a theory to support this enterprise, most notably the need to take into consideration the exploration-exploitation tradeoff at the core of RL in conjunction with the computational and statistical tradeoffs that arise in modern function-approximation-based learning systems. We approach these challenges by studying an optimistic modification of the least-squares value iteration algorithm, in the context of the action-value function represented by a kernel function or an overparameterized neural

network. We establish both polynomial runtime complexity and polynomial sample complexity for this algorithm, without additional assumptions on the data-generating model. In particular, we prove that the algorithm incurs a sublinear regret which is independent of the number of states, a result which exhibits clearly the benefit of function approximation in RL.