samsi
NSF·Duke·NCSU·UNC

DPDA Workshop
September 21-23, 2016

SPEAKER TITLES/ABSTRACTS

Guang Cheng
Purdue University

“Bayesian Aggregation for Extraordinarily Large Dataset”

In this talk, a set of scalable Bayesian inference procedures is developed for a general class of nonparametric regression models. Specifically, nonparametric Bayesian inferences are separately performed on each subset randomly split from a massive dataset, and then the obtained local results are aggregated into global counterparts. This aggregation step is explicit without involving any additional computation cost. By a careful partition, we show that our aggregated inference results obtain an oracle rule in the sense that they are equivalent to those obtained directly from the entire data (which are computationally prohibitive). For example, an aggregated credible ball achieves desirable credibility level and also frequentist coverage while possessing the same radius as the oracle ball. This oracle matching phenomenon occurs due to a delicate but explicit geometric structure of the in finite-dimensional parameter space in consideration.

David Dunson
Duke University

“Scalable Probabilistic Inference from Big and Complex Data”

In modern science and industry applications, it has become routine to collect massive dimensional data. There is a very rich literature defining algorithms for exploiting distributed computing systems for rapid point estimation based on huge data sets, but a clear lack of methods that allow for uncertainty quantification in inferences based on large data sets. In this talk, I provide an overall of some recent developments and directions in algorithms for Bayesian inference in complex data settings. These include: (i) embarrassingly parallel Markov chain Monte Carlo (MCMC); (ii) approximate MCMC; (iii) hybrid optimization-sampling algorithms; and (iv) modular Bayes. The main concepts in these directions will be highlighted by focusing on simple algorithms and particular motivating examples.

Jianqing Fan
Princeton University

“Distributed Estimation and Inference with Statistical Guarantees”

This talk is on hypothesis testing and parameter estimation in the context of the divide and conquer algorithm. In a unified likelihood based framework, we propose new test statistics and point estimators obtained by aggregating various statistics from $k$ subsamples of size $n/k$. In both low dimensional and high dimensional settings, we address the important question of how to choose $k$

as n grows large, providing a theoretical upper bound on the number of subsamples that guarantees the errors due to insufficient use of full sample by the divide and conquer algorithms are statistically negligible. In other words, the resulting estimators have the same inferential efficiencies and estimation rates as a practically infeasible oracle with access to the full sample. For parameter estimation, we show that the error incurred through the divide and conquer estimator is negligible relative to the minimax estimation rate of the full sample procedure. Thorough numerical results are provided to back up the theory.

(Joint work with Heather Battey, Han Liu, Junwei Lu and Ziwei Zhu)

Sam Franklin
360i.com

“HPDA Growth Constraints in Digital Marketing”

The fast-paced growth of digital advertising budgets and the corresponding explosion of data tied to digital marketing activities should lead to the rapid expansion of high performance data analysis throughout the industry. But, that hasn’t happened yet. This presentation will address several key barriers limiting HPDA’s adoption.

Rajarshi Guhaniyogi
University of California, Santa Cruz

“Some Recent Development in Spatial Statistics for Large Datasets”

Spatial process models for analyzing geostatistical data entail computations that / become prohibitive as the number of spatial locations becomes large. We propose a few strategies to facilitate analysis of large point referenced datasets. Each of these strategies proposes partitioning large data into smaller subsets (not necessarily mutually exclusive), and facilitates full analysis based on subsets stored in different processors. The resulting procedure eliminates the need to store the entire data in one processor and offers / parallelizable inferential techniques that assume linear computation complexity in the number of spatial locations, thereby delivering substantial scalability. We illustrate the computational and inferential benefits of the proposed strategies using simulation experiments and a real data example.

Xiaomo Jiang
General Electric Company

“DPDA Application in Predix Ecosystem for Real-time Monitoring and Diagnostics of Energy Assets”

Given over 100 million hours of operating data and 500+ deep data analytics and 30k hours of data increasing every day, distributed and parallel data analytics (DPDA) technology has to be applied exclusively to drive the daily business values for the real-time monitoring and diagnostics of nearly 2000 energy assets. This talk presents the application of DPDA in industrial internet ecosystem - Predix asset performance management (APM) - to turn big data and predictive analytics into power. The DPDA application is explained in terms of big data, predictive analytics, architecture, and real-world case studies. The technology has been validated in the Predix digital ecosystem to improve equipment efficiency and reliability and drive significant value for customers by lowering operating fuel cost and improving productivity, potentially increasing dispatch competitiveness.

Kimon Fountoulakis
University of California, Berkeley

“Parallel Local Graph Clustering”

Locally-biased graph algorithms are algorithms that attempt to find local or small-scale structure in a typically large data graph. In some cases, this can be accomplished by adding some sort of locality constraint and calling a traditional graph algorithm; but more interesting are locally-biased graph algorithms that compute answers by running a procedure that does not even look at most of the graph. While local clustering algorithms are already faster than traditional algorithms that touch the entire graph, they are sequential and there is an opportunity to make them even more efficient via parallelization. In this paper, we show how to parallelize many of these algorithms in the shared-memory multicore setting, and we analyze the parallel complexity of these algorithms.

Fanming Liang
University of Florida

“Bayesian Neural Networks for High Dimensionsl Nonlinear Variable Selection”

Variable selection plays an important role in data mining for high-dimensional nonlinear systems. However, the current variable selection methods are either developed for linear systems or computationally infeasible, rendering imprecise selection of relevant variables. In general, variable selection for high-dimensional nonlinear systems faces three challenges: (i) an unknown functional form, (ii) consistency, and (iii) highly-demanding computation. To circumvent the first difficulty, we employ a feed-forward neural network to approximate the unknown nonlinear function motivated by its universal approximation ability. To circumvent the second difficulty, we conduct structure selection (with variable selection being induced) for the neural network by choosing appropriate prior distributions that lead to the consistency of variable selection. We propose to resolve the computational issue by implementing the population stochastic approximation Monte Carlo algorithm, a parallel adaptive Markov Chain Monte Carlo (MCMC) algorithm, on the OpenMP platform which provides a linear speedup for the simulation. The numerical results indicate that the proposed method can execute very fast on a multicore computer and work very well for identification of relevant variables for general high-dimensional nonlinear system. The proposed method is successfully applied to personalized medicine and selection of anticancer drug response genes for the cancer cell line encyclopedia (CCLE) data.

Han Liu
Princeton University

“Blessing of Massive Scale”

We consider the problem of estimating high dimensional spatial graphical models with a total cardinality constraint (i.e., the L0-constraint). Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this ‘blessing of massive scale’ phenomenon. Motivated by this result, we propose an efficient parallel algorithm to solve the dual problem (which is concave) and prove that the solution achieves optimal statistical properties. Extensive numerical results are also provided (Joint work with Ethan X Fang and Mengdi Wang).

Qi Long
Emory University

“Privacy-Preserving Methods for Handling Missing Data in Distributed Health Data Networks”

Missing data are ubiquitous and present analytical challenges in distributed health data networks that leverage EHRs from multiple institutions/sites, e.g., eMerge, pSCANNER, PEDsnet, the latter two of which are partner networks in PCORnet. The existing methods for handling missing data require pooling patient-level data into a centralized repository and hence sharing of such data across institutions/sites. This approach, however, may not be appropriate or practical due to institutional policies (e.g., Veteran’s Health Administration policies for EHRs require them to be analyzed at VA’s facilities), cost of moving large data, and most importantly, privacy concerns. In particular, a large body of research has demonstrated that given some background information about an individual such as data from EHRs, an adversary can learn sensitive information about the individual from de-identified data. In this talk, I will first describe the issue of missing data in distributed health data networks including 1) horizontally partitioned data - different data custodians such as hospitals and healthcare service providers have the same type of data for different sets of patients; and 2) vertically partitioned data - different data custodians such as hospitals, insurance companies, and sequencing centers have different pieces of patient information (i.e., data from the same patient are distributed across different institutions). I will then present our preliminary work on developing privacy-preserving methods for handling missing data in distributed health data networks that do not require pooling patient-level data into a centralized repository. Lastly, I will present the conceptual architecture for our system for handling missing data in distributed health data networks including different modules for communication, storage, and algorithms. This is joint work with Xiaoqian Jiang, PhD at the University of California, San Diego and Yi Deng at Emory University.

Stanislav Minsker
University of Southern California

“Scalable and Robust Statistical Estimation: a tale of the geometric median”

Contemporary high-dimensional data analysis problems pose several general challenges. One is related to resource limitations: massive data require computer clusters for storage and processing. Another problem occurs when available observations are contaminated by noise and "outliers" that are not easily identified and removed. An attempt to address these challenges raises natural question: can we design estimation techniques that (i) admit strong performance guarantees under weak assumptions on the noise and (ii) can be implemented in parallel while preserving the quality of estimation? A "typical" approach to parallel estimation is based on averaging estimators obtained using independent subsets of data. We propose an alternative solution based on the properties of a geometric median, which is one of the possible extensions of a univariate median to higher dimensions. We will discuss several examples, starting with estimation of the mean on a real line, and continuing with sparse linear regression and matrix completion problems.

Flavio Villanustre
LexisNexis Risk Solutions

“Distributed Data Analysis on the Open Source HPCC Systems Platform Stack”

There are numerous challenges when it comes to large-scale data analysis. Although certain data problems can be categorized as embarrassingly parallel and are easily tractable through data partitioning and SIMD execution across a distributed architecture, others are far harder to approach and have traditionally required custom tailored implementations of algorithms, which can be inefficient, burdensome and prone to errors. Examples of the latter are non-convex optimization problems, network/graph traversal problems and probabilistic generative models, etc. The open source HPCC Systems platform stack takes a different approach, leveraging generalizable hierarchical declarative programming abstractions to address classes of problems with common approaches, pushing the burden of the specific implementation into code generators and compilers, rather than users. While the underlying distributed data processing system is implemented in C++, the data programming layer is implemented in the open declarative dataflow ECL language, which compiles into C++ prior to execution. The programming abstraction to deal with probabilistic record linkage and entity resolution employs a declarative meta-programming language called SALT, which compiles to ECL. In a similar manner, the programming abstraction for network/graph processing and querying utilizes a declarative programming language called KEL, which also compiles to ECL. Above these programming abstraction layers, user-friendly visual "drag'n drop" programming metaphors are available, to automatically generate the SALT, KEL and ECL code. During this discussion we'll introduce the audience to this approach and compare it to more traditional monolithic programming models for distributed data analysis.

Min-ge Xie
Rutgers University

“A Sequential Split-Conquer-Combine Approach for Gaussian Process Model in Analysis of Big Spatial Data”

The task of analyzing massive spatial data is extremely challenging. In this talk, we propose a sequential-split-conquer-combine (SSCC) approach for analysis of dependent big spatial data using a Gaussian process model, along with a theoretical support. This SSCC approach can substantially reduce computing time and computer memory requirements. We also show that the SSCC approach is oracle in the sense that the result obtained using the approach is asymptotically equivalent to the one obtained from performing the analysis on the entire data in a super-super computer. A related prediction problem is also considered. The methodology is illustrated numerically using both simulation and a real data example of a computer experiment on modeling room temperatures. (Joint work with Ying Hung and Chengrui Li)

Keywords: Confidence distribution; Gaussian process model; Kriging; Uncertainty quantification

Eric Xing
Carnegie Mellon University

“Strategies & Principles for Distributed Machine Learning”

The rise of Big Data has led to new demands for Machine Learning (ML) systems to learn complex models with millions to billions of parameters that promise adequate capacity to digest massive datasets and offer powerful predictive analytics (such as high-dimensional latent features, intermediate representations, and decision functions) thereupon. In order to run ML algorithms at such scales, on a distributed cluster with 10s to 1000s of machines, it is often the case that significant engineering efforts are required --- and one might fairly ask if such engineering truly falls within the domain of ML research or not. Taking the view that Big ML systems can indeed benefit greatly from ML-rooted statistical and algorithmic insights --- and that ML researchers should therefore not shy away from such systems design --- we discuss a series of principles and strategies distilled from our resent effort on industrial-scale ML solutions that involve a continuum from application, to engineering, and to theoretical research and development of Big ML system and architecture, on how to make them efficient, general, and with convergence and scaling guarantees. These principles concern four key questions which traditionally receive little attention in ML research: How to distribute an ML program over a cluster? How to bridge ML computation with inter-machine communication? How to perform such communication? What should be communicated between machines? By exposing underlying statistical and algorithmic characteristics unique to ML programs but not typical in traditional computer programs, and by dissecting successful cases of how we harness these principles to design both high-performance distributed ML software and general-purpose ML framework, we present opportunities for ML researchers and practitioners to further shape and grow the area that lies between ML and systems.

Yan Xu
SAS

“Distributed Hyper-Parameter Optimization for Machine Learning”

Machine learning algorithms are sensitive to their hyper-parameter settings, lacking good universal rule-of-thumb defaults. In this talk, we discuss the use of distributed local search optimization (LSO) to tune the hyper-parameters of machine learning. Viewed as a black-box objective function of hyper-parameters, machine learning algorithms create a difficult class of optimization problems. The corresponding objective functions involved tend to be non-smooth, discontinuous, unpredictably computationally expensive, requiring support for both continuous, categorical, and integer variables. Further evaluations can fail for a variety of reasons such as early exits due to node failure or hitting max time. In this context, we developed a novel distributed LSO framework that can make progress despite these difficulties providing significantly improved results over default settings with minimal user interaction.

Wotao Yin
University of California, Los Angeles

“Asynchronous Parallel Coordinate Update Algorithms”

This talk focuses on a class of algorithms, called coordinate update algorithms, which are useful at solving large-sized problems involving linear and nonlinear mappings, and smooth and nonsmooth functions. They decompose a problem to simple subproblems, where each subproblem updates one, or a small block of, variables each time. They have found applications throughout signal/imaging processing, differential equations, and machine learning. We abstract many problems to the fixed-point problem x^{k+1}=Tx^k. This talk discusses the favorable structures of the operator T that enable highly efficient asynchronous parallel iterations running on multi-core platforms. We introduce new scalable coordinate-update algorithms to many problems with global linear constraints Ax=b, nonsmooth nonseparable functions, and large-scale data. We will present a software package and its numerical examples.

Bo Zhang
IBM

“Uncover Customer Insights with Apache Spark and ML”

Better customer insights will lead to better customer experiences and better customer engagement.

To uncover customer insights from all possible touchpoints, implementing statistical models on big volume data presents numerous challenges. Apache Spark and its ML package can be undoubtedly a force to overcome such challenges in a distributed fashion. In this talk, we'll describe our data science pipeline, the uncovered customer insights, and the advantages of using Spark.

Helen Zhang
University of Arizona

"Interaction Selection and Screening for High Dimensional Data"

We investigate the topic of identifying interaction effects in high dimensional regression. A variety of problems are considered including interaction selection and interaction screening, hierarchy-preserving and hierachy-free model structures, linear models and nonlinear models. One common challenge shared by these problems for high dimensional data is the expensive computational cost. We propose a number of new methods which are scalable and can employ parallel computation techniques to reduce the cost to a feasible level. Both theories and numerical examples are presented.

Yilu Zhang	**Wei Tong**
GM	GM

"Challenges and Opportunities in Automated Driving and Connected Vehicles"

Automated driving and connected vehicles are two of the mega trends that will shape the automotive industry in the next 10 to 20 years. They promise to greatly enhance personal mobility and provide improved driving experiences for the consumers. At the same time, new technologies are called for to address many interesting challenges. For example, how to automatically maintain and manage the health of vehicle systems to ensure the missions of transportation are accomplished; how to utilize the on-board processing power from a fleet of automated and connected vehicles as well as the resources from the cloud to enhance the perception capability of each individual vehicles. In this talk, we will present recent progresses that GM R&D has made in these areas, and share some open research problems that may be worked on collaboratively by industrial and academia research communities.