Statistical Learning

THE TENTH EUGENE LUKACS SYMPOSIUM

DEPARTMENT OF MATHEMATICS AND STATISTICS

BOWLING GREEN STATE UNIVERSITY

MAY 14, 2016

The 10th Eugene Lukacs Symposium Program Schedule
(All symposium talks are given in Room 112 ,Life Science Building, BGSU Campus)

May 13 6:00 pm -- 8:00 P Registration and **Mixer** in Mileti Room of Alumni Center (BGSU Campus)

May 14, 2016 -- Morning Session (Room 112, Life Science Building BGSU Campus)
Session chair: Dr. Junfeng Shang, BGSU

8:30 – 9:00 A Registration and breakfast

9:00 – 9:15 A Dr. Hanfeng Chen, Chair, Dept. of Mathematics and Statistics, BGSU
Welcome Remarks and Tribute to Eugene Lukacs

9:15 – 10:15 A Dr. Peter Song, University of Michigan
Method of Divide-and-Combine in Regularized Generalized Linear Models for Big Data

10:15 – 11:00 A Dr. Herman Rubin, Purdue University
Reuse of Random Variables

11:00 – 11: 15 A **Tea break**

11:15 -- 11:45 A Dr. Jim Albert, BGSU
An Undergraduate Data Science Program

11:45 A – 12:15 P Dr. Arunasalam Rahunanthan, Central State University
A Bayesian Predictive Simulation for Aquifer Contamination

12:15 – 1:30 P **Lunch break**

May 14, 2016 -- Afternoon Session (Room 112, Life Science Building BGSU Campus)
Session chair: Dr. Craig Zirbel, BGSU

1:30 -- 2:30 P Dr. Nabanita Mukherjee, Duke University
The Big Data Revolution: Challenges, Hopes & Reality

2:30 – 3:00 P Dr. Igor Melnykov, Colorado State University, Pueblo
Accommodating Positive and Negative Constraints in Model-Based Clustering

3:00 -- 3:15 P **Tea break**

3:15 – 3:45 P Dr. Ying-Ju (Tessa) Chen, Miami University
Adjusted Jackknife Empirical Likelihood

4:45 -- 4:15 P Dr. Ibrahim Capar, BGSU
Online and Open Vehicle Routing Problem with Split Delivery

4:15 – 4:45 P Dr. Rida Benhaddou, Ohio University
Deconvolution Model with Fractional Gaussian Noise: A Minimax Study

4:45 – 5:00 P Concluding Remarks

May 14 6:00 -- 8:00 pm **Symposium Dinner** in Mileti Room of Alumni Center (BGSU Campus)

List of Talk Abstracts

(in presentation order)

Method of Divide-and-Combine in Regularized Generalized Linear Models for Big Data

Peter SONG, University of Michigan

Abstract: When a data set is too big to be analyzed entirely once by one computer, the strategy of divide-and-combine (or division-and-conquer, MODAC) has been the method of choice to overcome the computational hurdle. Although random data partition has been widely adopted, there is lack of clear theoretical justification and practical guidelines to combine results obtained from separately analyzed sub-datasets, especially when a regularization method such as LASSO is utilized for variable selection in the generalized linear model regression. We develop a new strategy to combine separately regularized estimates of regression parameters by means of the confidence distributions of biased corrected estimators. We first establish the theory for the construction of the confidence distribution and then show that the resulting MODAC estimator enjoys the Fisher's efficiency, the efficiency of the maximum likelihood estimator obtained from the analysis of entire data once. Furthermore, using the MODAC estimator we propose a variable selection procedure, which is compared analytically and numerically via extensive simulations with the existing majority-voting method and the gold standard of one-time entire data analysis. This is a joint work with Lu Tang and Ling Zhou.

Reuse of Random Variables

Herman RUBIN, Purdue University

Many procedures for generating random variables use test variables, or in some cases auxiliary random variables. Frequently there can be reuse of the test random variable; this can be somewhat complicated.

About 50 years ago, using the well-known tail probabilities of the exponential distribution, I observed that when the exponential variable was greater than the cutoff point, the difference could be "put back" as an exponential random variable. This can even lead to what I consider to be the cheapest practical way to generate exponential random variables from uniform random input; the errors are within "rounding" limits. This does not lead to any practical reuse if rejection occurs.

Recently, I observed that this method works for testing with uniform random variables if the test cutoff point, rather than the random variable, is changed. A somewhat different approach works if there is rejection. A slightly different version, which involves no errors with the random variable, is actually a realization of the explicit formulation of the situation, using from the random standpoint, only adding bits as needed. For efficiency, either in this or in the simplified version, one can bring in bytes or large units, but there are limitations here. The expected number of bits lost is precisely the entropy of the test, or other discrete procedure for which it can be used. What should be used in practice depends on programming and efficiency constraints.

An Undergraduate Data Science Program

Jim ALBERT, Bowling Green State University

Abstract: Bowling Green State University has created a new undergraduate data science program within a department of mathematics and statistics. This program is a synthesis of courses in mathematics, statistics and computer science designed to prepare the students for opportunities in data science. We describe the new data science and statistical learning courses and discuss challenges in the development of this program.

Title: A Bayesian Predictive Simulation for Aquifer Contamination

Arunasalam RAHUNANTHAN, Central State University

Abstract: In contaminant transport in subsurface we often need to forecast flow patterns. In the flow forecasting, subsurface characterization is an important step. To characterize subsurface properties we establish a statistical description of the subsurface properties that are conditioned to existing dynamic (and static) data. We use a Markov Chain Monte Carlo (MCMC) algorithm in a Bayesian statistical description to reconstruct the spatial distribution of two important subsurface properties: permeability and porosity. By using reconstructed permeability and porosity distributions, we predict subsurface flows. In this talk, we develop a Bayesian framework for predictive simulation of contaminants in an aquifer.

The Big Data Revolution: Challenges, Hopes & Reality

Nabanita MUKHERJEE, Duke University

Abstract: Big data is arriving from different sources at an alarming velocity, volume and variety. With this proliferation of data comes the potential for new insights in fields as diverse as astronomy, social networking, and healthcare management. The real challenge is processing this huge amount of information into some meaningful and interpretable format that can be used to drive decision-making in these fields. Thus, the revolution lies not in data storage, but in advanced statistical and computational methods. This talk will address these issues in an applied context by focusing on some of my recent research experience in network analysis and healthcare management. In network analysis, our motivation is drawn from social science studies involving dynamic multilayer networks. The talk will focus on the computational complexities and challenges we faced with our nonparametric dynamic multilayer model in real world networks. In healthcare management, our goal is to assess how well we can process the large volume of Duke electronic health records (EHRs) data to develop more robust risk model(s) for patient deterioration - defined as RRT (Rapid Response Team) emergency or in-hospital mortality or transfer to the ICU. Until now, Duke University Hospital approached this problem in a pure deterministic framework, by developing an alert risk score for acute care wards: the National Early Warning Score (NEWS). However, it is questionable whether this approach optimally uses the data and accounts for sparsity in the features for each individual (feature-poor data). We discuss here the challenges we are facing in using real time EHR data to develop a time updating risk score for patient deterioration using varying degrees of model complexity.

Accommodating Positive and Negative Constraints in Model-based Clustering

Igor MELNYKOV, Colorado State University, Pueblo

Abstract: The objective of cluster analysis is to locate groups of similar observations in a data set. Without any restrictions on the membership of points, unsupervised clustering takes place. At the same time, when some information is available regarding the placement of points in classes, this leads to a semi-supervised clustering scenario. We consider two specific types of constraints that can be imposed on the solution. Under positive constraints, certain points are joined together so that they must belong to the same cluster, while with negative constraints in place, the points are prevented from being in the same class. We work on an approach that accommodates both negative and positive constraints in the setting of model-based clustering and consider the changes that need to be made in the implementation of the EM algorithm compared to the unsupervised case when finite mixtures are employed.

Adjusted Jackknife Empirical Likelihood

Tessa CHEN, Miami University

Abstract: Jackknife empirical likelihood (JEL) is an effective modified version of empirical likelihood method (EL). Through the construction of the jackknife pseudo-values, JEL overcomes the computational difficulty of EL method when its constraints are nonlinear while maintaining the same asymptotic results for one sample and two-sample U statistics. In this paper, we propose an adjusted version of JEL to guarantee that the adjusted jackknife empirical likelihood (AJEL) statistic is well defined for all the values of the parameter, instead of restricting on the convex hull of the estimation equation. The properties of JEL have been preserved for AJEL. This is joint work with Wei Ning.

Online and Open Vehicle Routing Problem with Split Delivery

Ibrahim CAPAR, Bowling Green State University

Abstract: Online and open VRP with split deliveries is a common problem for shippers that use common carriers. We develop an asymptotical-optimality-based reduction technique and solve a real life problem within reasonable times. Combining this technique with various dispatch policies, we demonstrate over eight percent savings compared to the real-life benchmark.

Deconvolution Model with Fractional Gaussian Noise: A Minimax Study

Rida BENHADDOU, Ohio University

Abstract: We consider estimating the response function in a deconvolution model under fractional Gaussian noise (fGn), derive minimax lower and upper bounds and implement WaveletVaguelette-Decomposition (WVD) to de-correlate fGn. We consider both the regular-smooth and super-smooth convolutions. Our estimator is adaptive and asymptotically optimal or near-optimal. LRD affects convergence rates only in the regular-smooth convolution.