Poster Abstracts

ISBIS 2017
June 7-9

IBM T. J. Watson Research Center in Yorktown Heights, NY

Sponsored by

IBM Research

THE CLIMATE CORPORATION

facebook

NESS | New England Statistical Society

New England Statistics Symposium since 1987 & New England Statistical Society since 2017

Google

Full program and abstracts can be found online at www.isbis2017.org/program.

Poster Sessions are on Wed June 7 and Thu Jun 8, 1-2pm in CR3 and Mezzanine.

"Geospatial analysis of seasonal and unseasonal demand"
Flora Babongo Bosombo, University of Lausanne

Abstract: Accurate demand forecasts are essential input into the decision making process and play an important role in supply chain management. In this study, we focus on spatially analysing seasonal and unseasonal sport goods, and on improving their demand forecasts' accuracy. We analyse real orders lines data from a well-known company. The main purpose of our study is to explain, using model based geostatistics, how considered external information, namely socio-economic characteristics and weather conditions affect the spatial variation of demand. We found that the socio-economic features impact the demand of both seasonal and unseasonal products but much more the unseasonal ones, whereas weather conditions affect only seasonal products. The cross-validation analyses show that the incorporation of the considered external information improve the demand forecasting accuracy. These results can be used in the decision making process, such as deciding of the location of the new retail shop.

"Combined Analysis of Numerical and Text Data in Surveys: Supervised Latent Dirichlet Allocation (sLDA)"
Christine P. Chai, Duke University

Abstract: Many surveys contain both numerical and text data, so it is important to analyze both parts to utilize the whole dataset. The numerical data are rating scores, and text data are answers to free response questions, which often contain useful information. The supervised latent Dirichlet allocation (sLDA) is a commonly used method to perform topic modeling on hybrid datasets, and the model determines which topics are associated with which ratings. For example, in an employee satisfaction dataset, I discovered that higher scores are associated with positive words (e.g. "opportunity" and "challenge"), while responses which mention work-life balance are usually associated with lower score. In addition, a higher number of "nots" per comment also indicates a lower rating because people say they are "not happy", but they do not say they are "not sad". Last but not least, sLDA can also be used to predict the rating score given the text comment, and one application is error correction – some people may get confused by the scale and provide a low score to indicate a positive response. By analyzing the text answers, this kind of error can be discovered and corrected.

"High-dimensional adaptive function-on-scalar regression"
Zhaohu (Jonathan) Fan, Penn State University

Abstract: Applications of functional data with large numbers of predictors have grown precipitously in recent years, driven, in part, by rapid advances in genotyping technologies. Given the large numbers of genetic mutations encountered in genetic association studies, statistical methods which more fully exploit the underlying structure of the data are imperative for maximizing statistical power. However, there is currently very limited work in functional data with large numbers of predictors. Tools are presented for simultaneous variable selection and parameter estimation in a functional linear model with a functional outcome and a large number of scalar predictors; the technique is called AFSL for Adaptive Function-on- Scalar Lasso . It is demonstrated how techniques from convex analysis over

Hilbert spaces can be used to establish a functional version of the oracle property for AFSL over any real separable Hilbert space, even when the number of predictors, I , is exponentially large com- pared to the sample size, N . AFSL is illustrated via a simulation study and data from the Childhood Asthma Management Program, CAMP, selecting those genetic mutations which are important for lung growth.

”Gaussian Graphical Models with Bayesian Regularization”
Lingrui Gan, University of Illinois at Urbana-Champaign

Abstract: We consider a Bayesian framework for estimating the precision matrix of a high dimensional vector, in which adaptive shrinkage and sparsity are induced by using spike and slab Laplace priors. Besides discussing our formulation from the Bayesian standpoint, we investigate the resultant posterior from a penalized likelihood perspective that gives rise to a new non-convex penalty approximating the L_0 penalty. Optimal error rates for estimation consistency in terms of l infinity , Frobenius and spectral norms along with selection consistency for sparsity structure recovery are shown under mild conditions. For fast and efficient computation, an EM algorithm is proposed. Through extensive simulation studies and a real application to a call center data, we have demonstrated the fine performance of our method compared with the existing alternatives.

”Analytics for Service-Estimation in Inventory Systems with Unknown Input Models”
Canan Gunes Corlu, Boston University

Abstract: Stochastic simulation is a commonly used tool by practitioners for evaluating the performance of inventory policies. A typical inventory simulation starts with the determination of the best-fit input models (e.g., probability distribution function of the demand random variable) and then obtains a performance-measure estimate under these input models. However, this sequential approach ignores the uncertainty around the input models, leading to inaccurate performance measures especially when there is limited historical input data. In this study, we take an alternative approach and propose a simulation replication algorithm that jointly estimates the input models and the performance measure, leading to a credible interval for the performance measure under input-model uncertainty. Our approach builds on a nonparametric Bayesian input model and frees the inventory manager from making any restrictive assumptions on the functional form of the input models. Focusing on a single-product inventory simulation, we show that the proposed method improves the estimation of the service levels when compared to the traditional practice of using the best-fit or the empirical distribution as the unknown demand distribution.

”Multivariate Process Monitoring using Phase I Limits and Principal Component Analysis Biplots: A Modern Developing Integrated Approach”
Chisimkwuo John, Michael Okpara University of Agriculture, Nigeria

Abstract: In multivariate Statistical Process Control (SPC) charts, robustness to basic statistical distribution assumptions, interpretability, and real-time (online) charting ability are known to be some major properties of a good charting technique. It is shown here that a new multidimensional integrated framework that uses the traditional SPC and the Principal Component Analysis (PCA) Biplots can be exploited to preserve these properties in the multivariate SPC monitoring setting. The fundamental algorithm starts with the preliminary limits using the classical SPC methods before the incorporation of the PCA Biplots to explore the representation of the relationships between different objects and their corresponding axes on the same plot. The resulting configuration, which is constrained by the Phase I

limits within the grid, becomes the basis for a user defined predictive multivariate monitoring regions with p(p-1)+2 total regions obtainable within a configuration of p variables. Results obtained from case studies on both the tobacco manufacturing process datasets and simulated datasets revealed promising schemes that foster quality decision making and this showcases the viability of the new approach.

”Singular Kalman Smoothing ”
Jonathan Jonker, University of Washington

Abstract: Time series are used to make decisions under uncertainty for a wide range of business analytics. Forecasts and estimates can be obtained by building state space models for time series and then analyzing them using filtering and smoothing techniques. We present a new smoothing method that can fit degenerate multiple source of error models, a richer class than nondegenerate dynamic models or single source of error models (such as Holt-Winters). We show how to incorporate expert domain knowledge via constraints, how to robustify the approach to outliers in observations, and how to model classically difficult situations such as auto-correlated noise. We illustrate using synthetic and real-world examples, in particular by fitting non-stationary time series corrupted with outliers.

”Coupling Data-Mining Analysis and Computer Simulation to Improve the Supply Chain of Supermarkets”
Yue Li, Texas State University

Abstract: Order picking is the most labor-intensive function of distribution centers (DC) in the food and beverage store industry. An efficient order picking process supports this industry’s supply chain to move high volumes of products between the DC and the retail stores. This presentation focuses on the Storage Location Assignment problem to deciding via an algorithm based on Association Rules the most adequate location of incoming products. The algorithm analyzes hundreds of orders received by the DC to find correlated products -the products that are ordered frequently together by retail stores. The algorithm then assigns correlated products to storage locations that are closed to each other in order to minimize order picking times. The results of computer simulation experiments using data from a real distribution center will be presented to evaluate the performance of the DC layout resulting from association rules.

”International Lifestyle Segmentation: Solving the High Dimensionality Dilemma”
Atreyee Majumder, Michigan State University

Abstract: This paper looks at 4 different world cultures to model market research data. The authors show that the models vary according to cultures and depend crucially on their psychographic segmentation. The data is high-dimensional so this paper tackles high-dimensional market research data with modern statistical methodologies. Penalized regression has rarely been explored with marketing data. The paper extensively shows how Ridge, LASSO and elastic net can be used to model market survey data. Additionally, the paper runs a simulation study which illustrates the criticality of the choice of tuning parameters in penalized regression. The simulation study finds an optimal method for tuning parameter selection based on information criterion. Following the simulation study, a market survey questionnaire data is used to build models for different countries elaborating on effective resource allocation by such an approach and establishing the requirement of culture influenced campaigning strategies.

”Making Personalized Skill Recommendations using Bayesian Member-Job Matching”
Abhinav Maurya, Carnegie Mellon University

Abstract: Inefficiencies in the labor market such as friction in matching members to jobs and the existence of skill gaps in various sectors of the economy are considered to be major problems facing economies today. The central premise of our work is that increasing the productivity of a member of the workforce (and thereby of the economy as a whole) crucially depends on identifying skills whose acquisition will yield the highest utility gains for that member. To this end, we develop a novel attribute-based Bayesian matching model BayesMatch to match members to other similar members as well as relevant jobs. The matching step is followed by a skill recommendation step SkillR which makes demand-based skill recommendations to members. Our extensive quantitative evaluation using a rich dataset comprised of professional profiles and job postings from LinkedIn suggests that skill recommendations made by our algorithm are highly correlated with skills demanded in heldout future jobs compared to those made by traditional collaborative filtering algorithms that do not utilize information about skill demand. This indicates that either members of the workforce do not have skills demanded by jobs or do not have enough information about which are the best skills to signal for competing in the labor market.

”Information Collection Optimization in Designing Marketing Campaigns for Market Entry”
Somayeh Moazeni, Stevens Institute of Technology

Abstract: Developing marketing strategies for a new product or a new target population is challenging, due to the scarcity of relevant historical data. Building on dynamic Bayesian learning, a sequential information collection optimization assists in creating new data points, within a finite number of learning phases. This procedure identifies effective advertisement design elements as well as customer segments that maximize the expected outcome of the final marketing campaign. In this paper, the marketing campaign performance is modeled by a multiplicative advertising exposure model with Poisson jumps. The intensity of the Poisson process is a function of the marketing campaign features. A forward-looking measurement policy is formulated to maximize the expected improvement in the value of information in each learning phase. Solving this information collection optimization is reduced to a mixed-integer second-order cone programming problem. A computationally efficient approach is proposed that consists in solving a sequence of mixed-integer linear optimization problems. The performance of the optimal learning policy over commonly used benchmark policies is evaluated using examples from the property and casualty insurance industry.

”Sampling discrete parameters with Hamiltonian dynamics: discontinuous Hamiltonian Monte Carlo”
Aki Nishimura, Duke University

Abstract: Hamiltonian Monte Carlo (HMC) is a powerful sampling algorithm employed by popular probabilistic programming languages. Its fully automatic implementations have made HMC a standard tool for applied Bayesian modeling. Though its performance is often superior to alternatives under a wide range of models, one weakness of HMC is the inability to handle discrete parameters. In this article, we present discontinuous HMC, an extension that can efficiently explore discrete parameter spaces as well as continuous ones. The proposed algorithm is based on two key ideas: embedding of discrete parameters into a continuous space and simulation of Hamiltonian dynamics on a piece-wise

continuous density function. The latter idea has been explored under special cases in the literature, but the extensions introduced here are critical in turning the idea into a general and practical sampling algorithm for discrete parameters. It is additionally shown that discontinuous HMC dominates a random walk Metropolis algorithm in terms of computational efficiency. We apply our algorithm to posterior inference problems in ecology and system reliability to demonstrate its superior performance over alternatives.

”Profile Monitoring of Poisson Data Using Non-parametric methods ”
Sepehr Piri , VCU

Abstract: Profile monitoring is a relatively new technique used to monitor the functional relationship between a response variable and one or more explanatory variables through time. For instance, profile monitoring can be used to monitor the changes in customers’ profiles where the profile is defined as the relationship between the number of times the customer has been contacted and the number of times the customer visited the website or the store. Although many studies have been conducted in this field, most of them assume the distribution of the response variable is to be normal which is not always appropriate. Given the ubiquitous nature of count data in real world situations, we decided to compare parametric and nonparametric methods via a large simulation in monitoring the changes in profiles where the response variable is defined by Poisson distribution. Our results proved that the data driven nature of nonparametric approach played in its favor and it can be a better substitute in detecting the changes in a product or process compared to the parametric approach.

”How Mega is the Mega? Measuring the Spillover Effects of WeChat by Bayesian Network and Econometrics ”
Zhengling Qi, UNC, Chapel Hill

Abstract: WeChat, an instant messaging app, is considered a mega app due to its dominance in terms of usage among Chinese smartphone users. Nevertheless, little is known about it externality in regard to the broader app market. Our work estimates the spillover effects of WeChat on the other top-50 most frequently used apps in China through data on users weekly app usage. Given the challenge of determining causal inference from observational data, we apply a graphical model and econometrics to estimates the spillover effects through two steps: (1) we determine the causal structure by estimating a partially ancestral diagram, using a Fast Causal Inference (FCI) algorithm; (2) given the causal structure, we find a valid adjustment set and estimate the causal effects by an econometric model with the adjustment set as controlling non-causal effects. Our findings show that the spillover effects of WeChat are limited; in fact, only two other apps, Tencent News and Taobao, receive positive spillover effects from WeChat. In addition, we show that, if researchers fail to account for the causal structure that we determined from the graphical model, it is easy to fall into the trap of confounding bias and selection bias when estimating causal effects.

”Statistical Modeling and Analysis of Chronic Disease Progression using Electronic Health Record Data ”
Vijaya Priya Rama Vijayasarathy, Carnegie Mellon University

Abstract: According to the Centers for Disease Control and Prevention, chronic diseases account for approximately 75% of the aggregate healthcare spending per year in the US. Among the most complex, costly, and high mortality chronic illnesses, Chronic Kidney Disease (CKD) poses significant

concerns in disease monitoring and management for both patients and providers. Tracking disease progression and identifying high-risk patients is a serious challenge that needs to be addressed for improving care delivery for this population. Leveraging the availability of a rich and unique 22-year clinical dataset extracted from the Electronic Health Record (EHR) of a community nephrology practice, we develop innovative, data-driven, statistical approaches to identify risk groups and analyze their CKD progression over many years. Early results using group-based trajectory models, applied to a key clinical marker of the disease, indicate five distinct trajectories of disease progression in this population, enabling risk stratification for targeted interventions. Furthermore, we extend the analysis using additional laboratory markers of multiple comorbidities and complications of CKD, adjusting for nonrandom attrition across the different groups, to profile important group characteristics and predict critical outcomes. These insights may empower patients and clinicians to better manage CKD progression, reduce costs, and improve quality of care delivery.

”Multi-Resolution Functional ANOVA for Large-Scale, Many-Input Non-Linear Regression, Estimation, and Inference ”

Chih-Li Sung, Georgia Institute of Technology

Abstract: The Gaussian process is a standard tool for building emulators for both deterministic and stochastic computer experiments. However, application of GP models is greatly limited in practice, particularly for large-scale and many-input computer experiments that have become typical. We propose a multi-resolution functional ANOVA model as an accurate and computationally feasible emulation alternative. More generally, this model can be used for large-scale and many-input non-linear regression problems. An overlapping group lasso approach is used for estimation, ensuring computational feasibility in a large-scale and many-input setting. New results on consistency and inference for the (potentially overlapping) group lasso in a high-dimensional setting are developed and applied to the proposed multi-resolution functional ANOVA model. Numeric examples demonstrate that the proposed model enjoys marked computational advantages. Data capabilities, both in terms of sample size and dimension, meet or exceed best available emulation tools while meeting or exceeding emulation accuracy.

”Modeling Power Laws in Directed Social Networks”

Phyllis Wan, Columbia University

Abstract: Preferential attachment is an appealing mechanism for modeling the widely observed power-law behavior of the degree distributions in directed social networks. In this presentation, we consider fitting a 5-parameter linear preferential model to network data under two data scenarios. In the case where full history of the network formation is available, we derive the maximum likelihood estimators of the parameters and show that they are strongly consistent and asymptotically normal. In the case where only a single-time snapshot of the network is available, we propose an estimation method that combines method of moments with an approximation to the likelihood. The resulting estimators are also strongly consistent and performs well compared to the MLE estimator based on the full history of the network. We illustrate both estimation procedures through simulated data, and explore the usage of this model in a real data example.

”Analysis of Familial Aggregation with Complex Survey Design”

Cong Wang , The George Washington University

Abstract: Familial aggregation is considered as an important aetiology of disease so that many studies focus on the analysis of familial aggregation. In recent studies, marginalization approach is considered as a better method to analyze the familiar aggregation of varying family sizes. The purpose of our research is to combine the marginalization approach with complex survey design, analyzing the familial aggregation for the families with different familial relationships and varying family sizes. Network sampling method is used to obtain the parameter estimates and the robust variance estimators. The recurrence risk is what we use to represent the familial aggregation. We apply our model to diabetes disease data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) in 1976. Propensity score weighting is also applied to adjust for the compounding variables. Also, simulation studies are conducted to examine the parameter estimates and variance estimators in the weighted model.

”Supervised learning on the progression of Alzheimer’s disease using a multi-state Markov model ”

Liangliang Zhang, Michigan State University

Abstract: With the rapid aging of the world population, Alzheimer’s disease is becoming a leading cause of death after cardiovascular disease and cancer. Nearly 10% of people 65 years of age and older are affected by Alzheimer’s disease. The causes of Alzheimer’s disease are currently being researched, but no definitive answers exist as yet. Genetic predisposition, abnormal protein deposits in the brain and environmental factors are suspected to play a role in the development of the disease. Our main focus in this paper is to model the progression of Alzheimer’s disease by applying multi-state Markov model, to investigate the significance of known risk factors like Age, ApoE4 and some brain structural volumetric variables getting from MRI like hippocampus, and to predict the transitions between different clinical diagnosis states using supervised learning. We found that the model with age is not significant (p-value is 0.1733) according to the likelihood ratio test, while ApoE4 is a significant risk factor in our Markov model. Predictions based on transition rates and transition probabilities were made and validated with the AUC as high as 0.8583. Therefore, we established a useful tool to help doctors to decide the best time for a proper treatment.

”A moment-based estimation procedure for fitting deeply nested hierarchical model ”

Ningshan Zhang , New York University

Abstract: In this project, we consider the problem of using a database of book reviews to inform user-targeted recommendations. In our dataset, books are categorized into genres and sub-genres. We use a hierarchical model, which is able to exploit this nested structure to pool information across similar items at many levels within the genre hierarchy simultaneously. Our main challenge is that fitting our model at scale using off-the-shelf maximum likelihood procedures is prohibitive due to the large data sizes involved. To get around this, we extend a moment-based fitting procedure proposed by Perry in 2016 for fitting two-level hierarchical models, which is an order of magnetite faster than maximum likelihood. Our procedure can be used in other contexts for fitting deeply-nested hierarchical generalized linear mixed models efficiently.

”Modeling Financial Durations using Estimating Functions ”

Yaohua Zhang, University of Connecticut

Abstract: Accurate modeling of patterns in inter-event durations is of considerable interest in high-frequency financial data analysis. The class of logarithmic autoregressive conditional duration (Log ACD) models provides a rich framework for analyzing durations, and recent research is focused on developing fast and accurate methods for fitting these models to long time series of durations under least restrictive assumptions. This article describes an optimal semi-parametric modeling approach using martingale estimating functions. This approach only requires assumptions on the first few conditional moments of the durations and does not require specification of the probability distribution of the process. Methodology and computing is described and compared for three approaches for parameter estimation, i.e., solution of nonlinear estimating equations, recursive formulas for the vector-valued parameter estimates, and iterated component-wise scalar recursions. Effective starting values from an approximating time series model increase the accuracy of the final estimates. We demonstrate our approach via a simulation study and a real data illustration based on high-frequency transaction level data on several stocks.