Conference Program

May 10, 2020

The 2020 Meeting of

The International Society for Data Science and Analytics

May 26-27, 2020

University of Notre Dame

Notre Dame, IN, USA

Sponsors

南京郵電大學
Nanjing University of Posts and Telecommunications

師大管理學院
College of Management
National Taiwan Normal University

UNIVERSITY of NOTRE DAME
Institute for Scholarship in the Liberal Arts

UNIVERSITY of NOTRE DAME
Department of Psychology

LIU INSTITUTE
FOR ASIA & ASIAN STUDIES

ISDSA
INTERNATIONAL SOCIETY FOR DATA SCIENCE AND ANALYTICS

Please contact the organizing committee at meeting@isdsa.org for any feedback.

Schedule	
May 26, 2020	
Time	Speaker and Title
	Morning
9:00-9:15	Welcome
9:15-9:45 Zoom	Reading China: Predicting Policy Change with Machine Learning Julian TszKin Chan, Bates White Economic Consulting Weifeng Zhong, Mercatus Center at George Mason University
9:45-10:15 Zoom	A data science approach for integrating water-related social media, population, and administrative data to reduce health disparities Cheng Wang, Wayne State University Richard Smith, Wayne State University Shawn McElmurry, Wayne State University Paul Kilgore, Wayne State University
10:15-10:45 Zoom	TUBE: Embedding Behavior Outcomes for Predicting Success Daheng Wang, University of Notre Dame Tianwen Jiang, Harbin Institute of Technology Nitesh Chawla, University of Notre Dame Meng Jiang, University of Notre Dame
10:45-11:00	Coffee break
11:00-11:30 Zoom	Hybrid test for publication bias in meta-analysis Lifeng Lin, Florida State University
11:30-12:00 Zoom	Capitalist Accumulation and Structure of Cryptocurrencies Ethan Fridmanski, Department of Sociology University of Notre Dame
	Afternoon
1:30-2:00 Zoom	An improved stochastic EM algorithm for large-scale full-information item factor analysis Siliang Zhang, London School of Economics and Political Science
2:00-2:30 Zoom	A Structural Equation Modeling approach to Multilevel Reliability Analysis Laura Lu, University of Georgia Minju Hong, University of Georgia Seohyun Kim, University of Georgia
2:30-3:00 Zoom	Balancing exploratory feature selection, computational limitations, and biological knowledge in computational genetics: The data science “Venn diagram” in action

	Justin Luningham, Georgia State University
3:00-3:15	Coffee break
3:15-3:45 Zoom	Predicting Authoritarian Crackdowns: A Machine Learning Approach Weifeng Zhong, Mercatus Center at George Mason University Julian TszKin Chan, Bates White Economic Consulting
3:45-4:15 Zoom	Modeling relationships from themes in text and covariates with an outcome: A Bayesian supervised topic model with covariates Kenneth Wilcox, University of Notre Dame Ross Jacobucci, University of Notre Dame Zhiyong Zhang, University of Notre Dame
4:15-4:45 Zoom	Imputing missing data with machine learning algorithms: A word of caution Justin Luningham, Georgia State University
4:45-5:15 Zoom	A dynamic and automated content analysis of the depression concept among Chinese netizens: from 2012 to 2019 Mengxin He, Beijing Normal University Hongyun Liu, Beijing Normal University
5:15-5:45 Zoom	Amending a Popular Dataset and Improving Scientific Entity Recognition with No-Schema Distant Supervision Qingkai Zeng, University of Notre Dame
May 27, 2020	
Time	Speaker and Title
	Morning
9:00-9:30 Zoom	Estimation of contextual effect and the impact of ICC in multilevel modeling: Does it matter for estimation methods? Hawjeng Chiou, National Taiwan Normal University Department of Business Administration Department of Educational Psychology and Counseling
9:30-10:00 Zoom	Out-of-bag prediction error estimators for extended redundancy analysis Sunmee Kim, McGill University Heungsun Hwang, McGill University
10:00-10:15	Coffee break
10:15-10:45 Zoom	Treatment effects on an outcome under nonlinear modeling

	Kai Wang, University of Iowa
10:45-11:15 Zoom	Robust Bayesian Growth Curve Modeling using Double Robust methods, growth curve modeling, conditional medians, asymmetric Laplace distribution Conditional Medians Tonghao Zhang, Department of Statistics, University of Virginia Xin Tong, Department of Psychology, University of Virginia Jianhui Zhou, Department of Statistics, University of Virginia
11:15-11:45 Zoom	A confidence interval of noncentrality compatible with test of a point null Hao Wu, Vanderbilt University
11:45-12:15 Zoom	Distributionally-Weighted Least Squares Han Du, University of California, Los Angeles Peter Bentler, University of California, Los Angeles
	Afternoon
	Recorded video presentations
1.	Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist Daniel Eck, Department of Statistics University of Illinois Charles Geyer, Department of Statistics University of Minnesota
2.	Estimation of Multilevel Time Series Longitudinal Data Laura Lu, University of Georgia Zhiyong Zhang, University of Notre Dame
3.	A Monte Carlo confidence interval method for testing measurement invariance Hui Li, Faculty of Psychology, Beijing Normal University Hongyun Liu, Faculty of Psychology, Beijing Normal University
4.	A comparative study on predictability of component-based approaches to structural equation modeling Gyeongcheol Cho, Department of Psychology, McGill University Heungsun Hwang, Department of Psychology, McGill University
5.	Evaluation of the Unsupervised Latent Dirichlet Allocation Model though Simulation Chang Che, University of Notre Dame Kenneth Wilcox, University of Notre Dame Zhiyong Zhang, University of Notre Dame

6.	Propensity score estimation with latent variables: data mining alternatives to logistic regression Ge Jiang, University of Illinois at Urbana-Champaign
7.	Multivariate Feedback Particle Filter and the Well-posedness of its Admissible Control Input Xue Luo, Beihang University
8.	Iterative Least-squares Regression with Censored Data: A Survival Ensemble of Learning Machine Md Hasinur Khan, ISRT, University of Dhaka
9.	Elaboration of economic cost-efficiency analyses based on equilibrium approach Oleksandr Ocheredko, Vinnytsya National Medical University Anastasiia Akhmedova, Vinnytsya National Medical University
	End of the conference

All Accepted Abstracts
(In alphabetical order)

A novel method for scRNA-seq data precise simulation (Cancelled)

Guoshuai Cai, Fei Qin, & Feifei Xiao
University of South Carolina

In recent years, efficient single cell RNA sequencing methods have been developed, enabling the transcriptome profiling of each single cell massively in parallel. Methods tailored for scRNA-seq data analysis are highly demanded. For either developing new methods or comparing existing methods, it is necessary to test the performance of methods and a common and effective way is through a simulation. Simulation provides a known truth to test against, which is usually difficult with real biological data. For scRNA-seq data analysis, several simulators are currently available but showing poor or not demonstrated similarity with real data, due to the missing of important effects in the modeling. To provide more precise simulation for evaluating scRNA-seq data analysis methodology, we developed a new scRNA-seq data simulation method by modeling zero-inflated counts based on a gamma-Poisson hierarchical model, taking all significant factors into consideration. The new simulator shows significantly improved similarity with multiple real datasets. It empowers the community to rapidly and rigorously develop, evaluate and compare scRNA-seq analysis methods.

Reading China: Predicting policy change with machine learning

Julian TszKin Chan
Senior Economist in the Finance
Bates White Economic Consulting

Weifeng Zhong
Senior Research Fellow
Mercatus Center
George Mason University

For the first time in the literature, we develop a quantitative indicator of the Chinese government's policy priorities over a long period of time, which we call the Policy Change Index for China. The PCI is a leading indicator of policy changes that covers the period from 1951 to the first quarter of 2019, and it can be updated in the future. It is designed with two building blocks: the full text of the People's Daily --- the official newspaper of the Communist Party of China --- as input data and a set of machine learning techniques to detect changes in how this newspaper prioritizes policy issues. Due to the unique role of the People's Daily in China's propaganda system, detecting changes in this newspaper allows us to predict changes in China's policies. The construction of the PCI does not require the understanding of the Chinese text, which suggests a wide range of applications in other contexts.

Evaluation of the unsupervised latent Dirichlet allocation model though simulation

Chang Che, Kenneth Wilcox, & Zhiyong Zhang
Department of Psychology
University of Notre Dame

Topic modeling using latent Dirichlet allocation has been increasingly endorsed as a popular procedure in text-mining. Unsupervised topic modeling focuses on the identification of the correct number of latent topics and clustering words into latent topics in the text mining area. This method is critical for latent topic modeling but understudied. Although an enormously wide range of applications emerges in empirical researches, evaluation of the performance of LDA has not covered all variates of text corpses with idiosyncrasies. For instance, there can be limited word counts in the text corpses of interest such as short interview transcripts, social media posts, and online reviews on venues with no more than one-hundred words in practice, while we find that the performance of topic modeling is not completely examined for short text. In this paper, we develop a systematic strategy to simulate data for evaluation of the performance of LDA. Based on the unsupervised analysis results, we demonstrate the effectiveness of LDA under various simulated conditions and provide our recommendations concerning the parameter choice for simulation and empirical analysis. In addition, we illustrate the label switching issues in simulation and provide adequate methods to deal with the specific situation encountered in massive simulation for practical methodology research.

Estimation of contextual effect and the impact of ICC in multilevel modeling: Does it matter for estimation methods?

Hawjeng Chiou
Department of Business Administration
Department of Educational Psychology and Counseling
National Taiwan Normal University

In social science, a collective variable is frequently defined by a contextual variable which is aggregated score of an observed measure from a group of people at the individual level. The cluster-mean of the individual-level variable is treated as a level-2 explanatory variable to predict an outcome in a multilevel model. Contextual effect is defined as the partial effect of the contextual variable on the outcome after removing the impact of the explanatory variable at individual level. The present study introduces the Bayesian estimate into the multilevel structural equation modeling for dealing with the estimation of the contextual effect with bias due to the influence of the magnitude of intra-class correlation on X_{ij} and Y_{ij} with different sample size. A simple Monte Carlo simulation shown that while the sample size at level-1 and level-2 was large, the performance of Bayesian and ML estimates were similar if the ICC of measured variables were huge. In contract, if the ICC was small, Bayesian approach was superior to ML estimates in terms of lower mean square error and higher coverage rate. An empirical dataset contained 38 companies and 1200 employees were adopted to explore the efficiency of Bayesian MSEM in

the estimation of contextual effect. The procedures of Bayesian MSEM with methodological implications were discussed in this study.

A comparative study on predictability of component-based approaches to structural equation modeling

Gyeongcheol Cho & Heungsun Hwang
Department of Psychology
McGill University

Partial least squares path modeling and generalized structural component analysis are two full-fledged component-based approaches to structural equation modeling, where components or weighted composites of observed variables are adopted as statistical representations for unobserved conceptual variables. These approaches are well-suited to predictive analytics because they obtain unique component scores in such a way that the scores minimize prediction errors of dependent variables. To date, however, no study has actually investigated how the approaches perform and compare in terms of predictive accuracy. The purpose of this study is to examine their similarities and differences in model specification and parameter estimation from the prediction perspectives and to evaluate the relative predictability of their different model specifications - PLSPM with mode A, PLSPM with mode B, GSCA with reflective indicators, and GSCA with formative indicators - under a wide range of simulation conditions, while using regression with sum scores as a naive benchmark.

Distributionally-weighted least squares

Han Du & Peter Bentler
University of California, Los Angeles

In real data analysis, data are unlikely to be exactly normally distributed. If we ignore the non-normality reality, the parameter estimates, standard error estimates, and model fit statistics from normal theory based methods such as maximum likelihood and normal theory based generalized least squares estimation are unreliable. On the other hand, the asymptotically distribution free estimator does not rely on any distribution assumption but cannot demonstrate its efficiency advantage with small and modest sample sizes. The methods which adopt misspecified loss functions including ridge GLS can provide better estimates and inferences than the normal theory based methods and the ADF estimator in some cases. We propose a distributionally-weighted least squares estimator, and expect that it can perform better than the existing generalized least squares, because it combines normal theory based and ADF based generalized least squares estimation. Computer simulation results suggest that model-implied covariance based DLS provided relatively accurate and efficient estimates in terms of RMSE. In addition, the relative biases of standard error estimates and the Type I error rates of the Jiang-Yuan rank adjusted model fit test statistic in DLS_M were competitive to the classical methods including ML, GLS, and RGLS. The performance of DLS_M depends on its tuning parameter a. We illustrate how to implement DLS_M and select the optimal a by a bootstrap procedure in the real data example.

Computationally efficient likelihood inference in exponential families when the maximum likelihood estimator does not exist

Daniel Eck
Department of Statistics
University of Illinois

Charles Geyer
Department of Statistics
University of Minnesota

In a regular full exponential family, the maximum likelihood estimator (MLE) need not exist in the traditional sense. However, the MLE may exist in the completion of the exponential family. Existing algorithms for finding the MLE in the completion solve many linear programs; they are slow in small problems and too slow for large problems. We provide new, fast, and scalable methodology for finding the MLE in the completion of the exponential family. This methodology is based on conventional maximum likelihood computations which come close, in a sense, to finding the MLE in the completion of the exponential family. These conventional computations construct a likelihood maximizing sequence of canonical parameter values which goes uphill on the likelihood function until they meet a convergence criteria. Nonexistence of the MLE in this context results from a degeneracy of the canonical statistic of the exponential family, the canonical statistic is on the boundary of its support. There is a correspondence between this boundary and the null eigenvectors of the Fisher information matrix. Convergence of Fisher information along a likelihood maximizing sequences follows from cumulant generating function (CGF) convergence along a likelihood maximizing sequence, conditions for which are given. This allows for the construction of necessarily one-sided confidence intervals for mean value parameters when the MLE exists in the completion. We demonstrate our methodology on three examples in the main text and three additional examples in the supplementary materials. We show that when the MLE exists in the completion of the exponential family, our methodology provides statistical inference that is much faster than existing techniques.

Capitalist accumulation and structure of cryptocurrencies

Ethan Fridmanski
Department of Sociology
University of Notre Dame

To understand capital and its relationship to cryptocurrencies this project looks at three main aspects of this social relation using two--bitcoin and litecoin--cryptocurrency systems: the market, accumulation processes, and the literal structure (network) of transactions within the system. By looking at how these three representations--or dimensions--of capital interact one can gain a better understanding of cryptocurrencies while also generating knowledge that can be applied to the study of the economy overall. Blockchain technology, and the data it generates, provides researchers a unique opportunity to study an arm of the total economic system with complete endogenous structural information as well as a plethora of exogenous market and system level data. With a bit of creativity and computational power one is able to bypass the anonymity of the transactions and identify important organizational actors within these systems. Both cryptocurrency miners and cryptocurrency exchanges provide important infrastructure to these digital systems and are responsible for the accumulation of capital and structural power.

The implications of the findings in this project are two-fold. First, this is a novel and comprehensive sociological analysis of cryptocurrency systems that takes into account markets, accumulation, and structures. Second, this is the first time researchers have a complete record of transactions for such a large trading platform. This means that the findings will provide scholars of capitalism and the economy a thorough model for how the structure of economic relations are related to market outcomes and accumulation processes. Ultimately--using computational social science, multivariate time series models, and theories of capital accumulation--I show that accumulation is occurring within these systems and their infrastructural and organizational contexts--cryptocurrency miners and exchanges--and that this accumulation is related to and in some part preceded by changes in the underlying social (network) structure.

A dynamic and automated content analysis of the depression concept among Chinese netizens: From 2012 to 2019

Mengxin He & Hongyun Liu
Faculty of Psychology
Beijing Normal University

Depression is a kind of mental disorder characterized by a low mood. It is often accompanied by the loss of interest in daily activities and pain without a clear cause. In past decades, depression suffered severe stigma in China. But this situation has changed now because of the development of both psychology and the internet. More and more people are willing to learn and talk about depression through search engines and social websites. Besides, most of their discussion left online are still accessible for us so that we can trace back how netizens' understandings and interests in depression disorder changed over time.

To capture the variation tendency of the concept of depression in the minds of the Chinese netizens, we collected all questions related to depression posted on Zhihu website from 2012 to 2019. Zhihu website is a Quora-like website in China in which users can post and answer questions with each other to share knowledge. It was created in 2011. Now it has more than 220 million users. Our dataset contains 61719 question texts with post time, the number of followers and answers, and topics labeled by their authors.

Most questions in our dataset are short texts which have less than 30 words. Due to all those questions are related to depression, their word usage can be very similar. Our exploring experiments showed that the conventional Latent Dirichlet Allocation topic model, which was based on word frequency, was not adaptable to this kind of task. In this article, we will develop new unsupervised text classifiers based on word embedding, which can take advantage of word order and external corpus to improve itself accuracy. Through those classifiers, we can distinguish those questions into different topics automatically. After naming those topics by experts according to their contents and observing how the frequency of those topics changed over time, we can understand the changes of depression concept among Chinese citizens from 2012 to 2019.

Propensity score estimation with latent variables: Data mining alternatives to logistic regression

Ge Jiang
Assistant Professor of Educational Psychology
University of Illinois at Urbana-Champaign

When random assignment is not feasible, the difference between treatment and control groups on an outcome cannot be fully attributed to the treatment due to covariates and pre-existing differences between the two groups. Propensity score methods are commonly used to balance the two groups and thus reduce the bias in treatment effect estimation. The propensity score is the probability of group assignment Z conditional on a set of covariates is typically used in four types of applications: matching, weighting, stratification, and covariance adjustment. For the propensity scores to be effective in reducing the bias, two assumptions need to be satisfied. The first one is that all covariates must be measured reliably. However, many characteristics in social and behavioral sciences cannot be perfectly measured, moreover, they are latent variables that need to be measured by multiple indicators. In those circumstances, structural equation modeling can be used to account for measurement errors and test the measurement invariance across groups. A limited number of studies have shown that multigroup SEM models in conjunction with the PS methods reduced the bias and increased precision of the treatment effect estimates although the SEM approach has not received widespread attention. The second assumption is that the relationship between Z and the covariates is correctly specified. Existing studies almost exclusively use a logistic regression that predicts Z based on the main effects of covariates. This logistic regression model apparently overlooks the possibilities of higher-order effects, interaction effects, or more complicated patterns. To address the limitations of logistic regression, data mining methods like Generalized Boosted Models and Neural Networks have been considered as alternatives for modeling nonlinear relationships but to my best knowledge, they have not been evaluated for estimating propensity scores with latent variables. ?? Therefore, the goal of the current paper is to propose two new approaches that adapt GBM and NN for estimating propensity scores in SEM models, respectively. The proposed approaches will be evaluated using a simulation study that considers a broad range of scenarios and their performances will be compared to those of the traditional logistic regression. The outcome measures are the bias and efficiency of treatment effect estimates as well as the average standardized absolute mean difference. Preliminary simulation results showed that the two new approaches perform slightly worse than logistic regression when the relationship is linear but can significantly reduce the bias with non-linear relationships. The benefits of the two new approaches are more salient for small samples than large samples. I will also illustrate the new approaches using a real dataset that examines the effect of participating in the teachers' association on teachers' workload manageability.

Analyzing the competition results of team Taiwan in international mathematical Olympiad (Cancelled)

Chu Lan Kao
National Chiao-Tung University

Taiwan participated in the 60th International Mathematical Olympiad in 2019, and obtained its lowest national ranking in history. Does this represent a decrease in Team Taiwan's competition capability? How show we evaluate this? In this paper, we propose a model that both consider the differences in competitors and in exam problems of each year, and an inference method combining Bayesian ordinal regression with EM algorithm. Under this model, we show that the result of Team Taiwan this year does not provide significant evidence for decay in capability. Our proposed method further provides a new statistical tool for future studies in similar competitions and exams.

Iterative least-squares regression with censored data: A survival ensemble of learning machine

Md Hasinur Khan
ISRT, University of Dhaka

Dealing with modeling for high-dimensional censored data is challenging because of the complexities in data structure. Many variable selection methods have been proposed for high-dimensional survival data for accelerated failure time model. The study attempts to focus on extending variable selection procedure for censored high-dimensional data with AFT models using survival ensemble of popular machine learning techniques. Particularly, we modified the iterative least squares estimation technique as proposed by Jin et al. for AFT models by a survival ensemble of random forest and boosting machine learning techniques for obtaining precise estimation and variable selection. The implementation of these machine learning tools has been developed in light with a recent work by Khan and Shaw. The performance of proposed methods has been demonstrated with high-dimensional censored data through a number of simulation examples and with a microarray data known as Diuse Large-B-cell Lymphoma where selection of genes that are linked with the survival time of DLBCL patients are studied. The proposed methods were compared with two similar methods in literature known as the modified resampling-based Buckley James method and buckley (James Dantzig selector both developed by Khan and Shaw (2006)). The simulation studies demonstrate very satisfactory variable selection performances for the proposed methods. The proposed boosting and random forest based methods outperform existing methods for most of the cases. The DLBCL data analysis also suggests that both proposed methods are able to find the important genes that are related to survival of patients and also can predict the survival time of future patients with small prediction error. The proposed methods are easy to understand and they perform estimation and variable selection simultaneously.

Out-of-bag prediction error estimators for extended redundancy analysis

Sunmee Kim & Heungsun Hwang
McGill University

Extended redundancy analysis is a prediction model that investigates the relationship between multiple sets of predictors and one or more response variables. In ERA, a component is extracted from each set of predictors in such a way that it accounts for the maximum variation of the response. In this regard, ERA aims to perform dimension reduction and linear regression simultaneously, providing a simpler description of predictor-response relationships and still achieving good predictive performance. ERA has been extended to improve its data-analytic flexibility, including generalized ERA for the analysis of a response variable that arises from an exponential-family distribution, functional ERA for the analysis of smooth functions or curves, multivariate ERA for the analysis of multiple correlated responses, and Bayesian ERA. In this talk, we introduce several new model evaluation metrics for ERA based on various resampling methods, each of which aims to assess the performance of the model on so-called out-of-bag data. This way of estimating prediction error on an independent set of data is not common in much of psychology and the social sciences, often leading to overly optimistic estimates of model performance. Considerable work has been done in the fields of statistics and machine learning on the use of cross-validation and bootstrap methods for OOB prediction error estimation, but to date, no research has applied these general tools to the ERA setting. Thus, in this work, we formulate five different OOB error estimators of ERA based on five-fold, ten-fold, leave-one-out CV, .632 estimate, and +.632 estimate, and carry out two simulation studies to empirically evaluate their performances. In the first simulation, we graphically examine the trade-off between bias and variance of the five error estimators, where in the second simulation, we investigate which one is the best to find the true model when mis-specified models are considered.

Fit difference between nonnested SEM models given categorical data (Cancelled)

Keke Lai
University of California at Merced

In SEM studies with categorical data, in addition to using RMSEA, CFI, and SRMR to evaluate the fit of a single model, often researchers also use these fit indices to compare rival models. Model selection based on ΔRMSEA, ΔCFI, or ΔSRMR is meaningful because (a) showing one model is better than another is more scientific and easier than establishing a “good” model; (b) it avoids the problems with cutoffs for fit indices; (c) one is less likely to overlook other equally

substantively plausible models; (d) information criteria (e.g., AIC, BIC) are not applicable to categorical data SEM. In this paper we propose point estimators and confidence intervals (CIs) for ΔRMSEA, ΔCFI, and ΔSRMR under categorical data. Our methods are applicable to nonnested models and do not need a true model. Simulation results show our point estimators and CIs are all trustworthy, whereas the bias is large when estimating ΔRMSEA (ΔCFI, ΔSRMR) based on the common estimators in the current literature for RMSEA (CFI, SRMR).

A Monte Carlo confidence interval method for testing measurement invariance

Hui Li & Hongyun Liu
Faculty of Psychology
Beijing Normal University

The multiple-group confirmatory factor analysis is commonly used for measurement invariance testing. However, it usually examines the scale-level invariance first and only conduct item-level tests when the scale-level invariance is rejected, which may cover up the actual existence of non-invariance. In addition, the MG-CFA is typically based on the comparison of the overall fit of nested models. With this kind of overall comparison method, it may lack enough power when there are only a few noninvariant items. Moreover, this method is cumbersome, because a separate model estimation is needed for each item. This study proposed a new method for MI testing, which can directly examine the invariance of all the items at one time. In this method, the Monte Carlo method is used to construct the approximate distribution and the confidence interval of the differences in various parameters across groups for each of the items. If the estimated difference in the parameter based on the empirical data falls outside of the constructed confidence interval, then the null hypothesis of invariance will be rejected. This method allows researchers to conduct item-level tests for all the items in one model estimation, and it abandoned scale-level test to avoid masking actual existence of noninvariant items. Using simulated data, this study further compared the new Monte Carlo confidence interval method and the MG-CFA in different conditions for testing the invariance of various parameters. Specifically, sample size, number of items, percentage of noninvariant items, and magnitude of non-invariance were manipulated. Power and Type I error rate were examined to evaluate the accuracy of detecting a lack of invariance. Results indicated that both our new method and the MG-CFA adequately controlled the Type I error rate in nearly all conditions, but our method provided relatively higher power than the MG-CFA, especially when the percentage of noninvariant items was low and the magnitude of non-invariance was small.

Hybrid test for publication bias in meta-analysis

Lifeng Lin
Florida State University

Assessing publication bias is a critical procedure in meta-analyses for rating the synthesized overall evidence. Many statistical tests have been proposed to detect publication bias. However, they often make dramatically different assumptions about the cause of publication bias;

therefore, they are usually powerful only in certain cases that support their particular assumptions, while their powers may be fairly low in many other cases. Although several simulation studies have been conducted to compare different tests' powers under various situations, it is infeasible to justify the exact mechanism of publication bias in a real-world meta-analysis and thus select the optimal publication bias test. We propose a hybrid test for publication bias by synthesizing various tests and incorporating their benefits, so that it maintains relatively high powers across various mechanisms of publication bias. The superior performance of the proposed hybrid test is illustrated using simulation studies and three real-world meta-analyses with different effect sizes. It is compared with many existing methods.

A Bayesian imputation-based sensitivity analysis procedure for unmeasured confounding in mediation analysis (Cancelled)

Xiao Liu
Department of Psychology
University of Notre Dame

Unmeasured confounding is a common threat to validity of mediation studies. To establish validity of mediation analysis when there are potential unmeasured confounders, an effective approach is to conduct sensitivity analysis. In the context of regression analysis, Bayesian methods have been used for evaluating the impact of unmeasured confounders on statistical inferences of regression coefficients. However, limited research has been done regarding the use of Bayesian methods for sensitivity analysis of mediation models. In the current study, we proposed a Bayesian imputation-based sensitivity analysis procedure for assessing robustness of empirical mediation inferences to potential unmeasured confounders. Specially, for a single-mediator model, we used a latent proxy confounder C to encapsulate the confounding effects, and treated the values of C for all individuals in the sample as missing data that are missing at random. In the proposed Bayesian procedure, the posterior simulation proceeds via two Gibbs sampling steps: an imputation step (I-step) and a posterior step (P-step). In the I-step, values of the latent confounder C are imputed from the conditional distribution conditioning on the observed data and all model parameters. In the P-Step, given the imputed values of C in the I-step and the observed data, samples from the posterior distributions of model parameters are drawn. By treating unmeasured confounding as a missing data problem, the proposed two-step Bayesian procedure accounts for uncertainty associated with the unmeasured confounding effects through both the prior distributions and the sampling of confounder values, which would make a difference especially when the sample size is small. We hope the proposed Bayesian imputation-based sensitivity analysis approach can serve as an alternative tool for researchers to evaluate robustness of empirical mediation analysis to unmeasured confounding.

A structural equation modeling approach to multilevel reliability analysis

Laura Lu & Minju Hong
University of Georgia

Seohyun Kim
University of Virginia

This study aims to propose a structural equation modeling approach to multilevel reliability analysis with multilevel data structure. Green and Yang proposed a nonlinear SEM reliability coefficient for a test with ordinal categorical items within an SEM framework, and it has been found to be more accurate than the linear SEM reliability that treats categorical scores as continuous. But they only considered the items with the same number of categories. Kim, Lu and Cohen extended their research to broader situations and proposed a general formula for reliabilities. But they only considered the single level data structure. This study extends Kim, Lu and Cohen work from single level reliabilities to multilevel reliabilities. We provide a numerical calculation to derive the multilevel nonlinear SEM reliabilities, and then conduct a simulation study to evaluate the performance of nonlinear multilevel SEM reliabilities. An example using real data with different numbers of ordered categories will be provided to illustrate the application of the different reliabilities. Discussion will be provided.

Estimation of multilevel time series longitudinal data

Laura Lu
University of Georgia

Zhiyong Zhang
University of Notre Dame

When longitudinal Data are collected from multiple subjects, there are intra-individual changes and inter-individual changes. The intra-individual changes describe the changes within a person, and the inter-individual changes are to study the change between persons. The analysis of such data can be conducted by adopting the multilevel random-coefficient models, by assuming within-individual measurements, level 1, are nested within the individuals, level 2. But a standard assumption is the within-individual residuals are uncorrelated. In some cases in reality, this assumption will be violated. And then the additional correlation structure should also be modeled. A time series model for such cases is proposed which consists of a standard multilevel model for repeated measures data augmented by an autocorrelation model for the level 1 residuals. First-order autoregressive models, AR, are considered in detail. There are two MLE estimation methods: exact MLE estimation method and conditional MLE estimation method. Both methods obtain the parameter estimates through maximizing the multiple subjects' likelihood function. This study investigates different estimation methods and compares results under different scenarios, such as different lengths of series, different numbers of subjects, and different data types. Simulations are conducted and discussion are provided.

Balancing exploratory feature selection, computational limitations, and biological knowledge in computational genetics: The data science “venn diagram” in action

Justin Luningham
Research Assistant Professor of Biostatistics
Department of Population Health Sciences
School of Public Health
Georgia State University

Data science is often characterized by three domains: statistics, computer science, and subject matter expertise. Data science is also often conflated with business analytics or non-traditional research. In this talk, I present the development of a novel Bayesian machine learning method for genetics research, revealing how the three domains of data science are actually foundational for modern scientific research. Genetic data analysis involves searching through millions of genetic markers across the genome for associations between markers and a trait or disease. Technological advances have allowed for the rapid collection of genetic data on thousands of individuals, resulting in massive data sets. However, true signals between genes and many common diseases have been extremely difficult to detect, necessitating the development of methodology for analyzing genetic data. We present a novel Bayesian machine learning approach for conducting a particular type of genetic analysis called a transcriptome-wide association study. TWASs integrate gene expression data into association studies through the following steps: 1) build a model predicting gene expression scores for each gene from selected individual genetic markers in a reference set; 2) predict genetically regulated gene expression in a larger set containing marker data, and 3) test associations between GReX and the outcome. TWAS methods show promise for interpreting association tests because they reduce the association tests from millions of markers to only 15,000-20,000 genes. Existing TWAS methods only use a very small proportion of available genetic data, utilizing markers that are near the target gene as predictors in the GReX model. Since markers outside of the immediate proximity of the target gene explain a significant amount of variation in gene expression, using genome-wide markers as predictors is expected to increase the prediction accuracy of GReX and then increase TWAS power. However, enormous computation power is required to fit prediction models for ~15,000 genes, with each gene predicted by millions of genome-wide makers. Here, we propose a novel TWAS approach that accounts for genome-wide genotype data in a Bayesian variable selection regression model to identify true signals. Our method uses summary statistics from single-marker regression models and an EM-MCMC algorithm to speed up computation time and enable practical usage. We also incorporate knowledge of the human genome to segment markers into blocks that are informed by population-level correlations between regions of the genome, allowing for massive parallelization of the process. We demonstrate the proof-of-concept for our method through a series of simulations, in which we evaluate the prediction accuracy of GReX with simulated gene expression values, and we test the statistical power to detect an association between GReX and an outcome in subsequent TWAS. Comparing our method to the status quo TWAS approach, our approach has better predictive performance and more power in subsequent TWAS, as seen in Figure 1. Our method has promise for extending TWAS, and this project also demonstrates how all three domains of data science can contribute synergistically to methodological development.

Imputing missing data with machine learning algorithms: A word of caution

Justin Luningham
Research Assistant Professor of Biostatistics
Department of Population Health Sciences
School of Public Health
Georgia State University

Missing data are pervasive in all areas of data analytics. Many widely-used feature selection algorithms handle missing data by listwise deletion; however, when there is a high-dimensional feature space, deleting cases with any missingness across the predictors can drastically reduce the amount of information available. Therefore, data scientists could benefit from methods for handling missing data when conducting large feature selection analyses. This lightning talk will bring to the forefront some of the unique challenges faced when imputing missing data in data science applications. Recent advances in missing data methods reveal model-based Bayesian methods for imputing missing data to be most effective . The challenge in exploratory feature selection, however, is that the form of the analysis model is not known prior to imputation ?€? a necessary requirement for model-based Bayesian imputation. An intuitive idea is to use data-driven prediction algorithms in the imputation model . We apply multiple imputation via chained equations using gradient boosting machines with recursive partitioning to simulated data with moderate-to-high levels of missingness. Preliminary findings demonstrate that the boosted tree approach does not result in more accurate imputation analysis, in particular because the standard errors are far too small. In short, the known objective of many statistical learning prediction models is to introduce a little bias in exchange for reduction in prediction variance . This property may actually be detrimental in imputation models, in which we seek accurate point predictions of the missing values, but we must allow for additional variance due to missing data uncertainty. Future methodological inquiry into imputation for exploratory analyses is proposed.

Multivariate feedback particle filter and the well-posedness of its admissible control input

Xue Luo
School of Mathematical Sciences
Beihang University

In this talk, we shall first derive the admissible control input of the multivariate feedback particle filter by minimizing the f-divergence of the posterior conditional density function and the empirical conditional density of the controlled particles. On the contrast, in the original derivation, a special f-divergence, Kullback-Leibler divergence, is used in the 1-dimensional nonlinear filtering problems. We show that the control input is invariant under the f-divergence class. That is, the control input satisfies exactly the same equations as those obtained by minimizing K-L divergence, no matter what f divergence in use. In the second-half of this talk, we show that if we restrict the control input to be the gradient of certain potential, then the existence and uniqueness of the control input is proven in some suitable functional space under certain regularity conditions. We confirm that the explicit expression of the control input given in is indeed the unique one in some trivial situation.

Elaboration of economic cost-efficiency analyses based on equilibrium approach

Oleksandr Ocheredko
Chairman of social medicine and organization of health services department
Vinnytsya National Medical University

Anastasiia Akhmedova
Vinnytsya National Medical University

Data analysts and clinical researchers applying cost-efficiency, cost-benefit, and cost-utility triad are still in dire need of reliable methodological tools. Some are too simplistic (classical cost-efficiency derivatives), some are cumbersome (Grossman model derivatives), but most are required theoretical substantiation. Promising is equilibrium approach capturing rational trade-off between clinical and economic efficiency dimensions that is pursued in the paper.

The basis for our development is the model suggested by Zweifel & Breyer [1], and Zweifel & Manning [2]. The theoretical foundations of their model are the classical theory of moral hazard and the theory of consumer utilities. The main disadvantage of the model is lack of practicality. We slightly modified the theoretical prepositions and its derivatives. Further, we showed that model given additional conditions streamlines to the classical types of economic analysis, i.e. cost-utilities, cost-effectiveness, and cost-benefit. Analytically derived structural equation suffers from optimization predicament, so simplified version is regarded. Notes for practical use are supplied.

We presented the studies [3, 4] that implemented the development in cost-effectiveness and cost-benefit terms.

A data science approach for integrating water-related social media, population, and administrative data to reduce health disparities

Cheng Wang, Richard Smith, Shawn McElmurry, & Paul Kilgore
Wayne State University

The purpose of this study is to integrate Twitter, water quality and water system, public health, and census data covering the population living Michigan and other states in the U.S. in order to (1) formulate an innovative research framework that can efficiently link multiple sources of digitalized data in large scale, (2) design an effective platform for storing, analyzing and visualizing the data, and (3) develop new models and real-world solutions for major health disparities. Social network analysis, geospatial analysis, data mining and machine learning, and other advanced statistical methods will be applied to analyze the integrated data. This study will enable us to answer fundamental questions about the role of social factors on health outcomes. How do individuals' access to safe drinking water and health care vary across neighborhoods and other geospatial units? What are the pathways and mechanisms through which social advantages influence health? When, where, and how should intervention programs be adopted to improve health and reduce health disparities? The answers to these questions provide the foundation for developing robust evidence to inform policies that improve access and reduce social disadvantages and enhancing resilience across coupled water and health systems.

TUBE: Embedding behavior outcomes for predicting success

Daheng Wang
University of Notre Dame

Tianwen Jiang
Harbin Institute of Technology

Nitesh Chawla & Meng Jiang
University of Notre Dame

Given a project plan and the goal, can we predict the plan's success rate? The key challenge is to learn the feature vectors of billions of the plan's components for effective prediction. However, existing methods did not model the behavior outcomes but component proximities. In this work, we define a measurement of behavior outcomes, which forms a test tube-shaped region to represent success, in a vector space. We propose a novel representation learning method to learn the embeddings of behavior components by preserving the behavior outcome information. Experiments on real datasets show that our proposed method significantly improves the performance of goal prediction as well as context recommendation over the state-of-the-art.

Treatment effects on an outcome under nonlinear modeling

Kai Wang
Professor, Department of Biostatistics
University of Iowa

Exact formulae relating parameters in conditional and reduced generalized linear models are introduced where the reduced model omits a continuous mediator from the conditional model. These formulae are obtained from the maximum likelihood equations as they give consistent estimates of model parameters. For certain link functions including logit, the natural direct effect and the natural indirect effect of the counterfactual method are smaller in magnitude than, respectively, the direct effect used for the difference method and the indirect effect by the product method while for some other links they are larger. Contrary to what is implicitly assumed in Jiang and VanderWeele for logit link, the total effect of the counterfactual method and the TE used for the difference method are generally not the same. They are equal to each other only under special situations. For accelerated failure time model the difference method and the product method are equivalent regardless of censoring or not. This result was stated previously in VanderWeele in the absence of censorship but the proof given there is misleading. For proportional hazards model, maximum likelihood analysis indicates that these two methods can be equivalent in the absence of censorship. In the case of logit link, one can focus on the

treatment effect on the marginal odds instead of the odds of the marginal event and have simple interpretation of model parameters. Similarly, for proportional hazards model, one can focus on the treatment effect on the marginal hazards instead of the hazards for the marginal survival time.

Modeling relationships from themes in text and covariates with an outcome: A Bayesian supervised topic model with covariates

Kenneth Wilcox, Ross Jacobucci, & Zhiyong Zhang
University of Notre Dame

While text analysis has long been of interest in psychology, quantitative approaches tend to rely on predefined dictionaries of words designed to correspond to constructs of interest. However, construction of dictionaries can be time-consuming and dictionary-based methods cannot accommodate relevant words that fall outside the dictionary. Furthermore, dictionary methods typically assign a word to only one construct and fail to account for polysemy and homonymy. One alternative approach, topic modeling, eschews predefined dictionaries and instead models patterns of word co-occurrences using latent categories. These topics have multiple uses and are often incorporated as components of a subsequent model. In psychological research, unsupervised topic models are typically used to obtain estimated topic proportions as a summary of the text. Associations between these estimated topic proportions and an outcome of interest may be modeled as a second stage. However, it is well-known that a two-stage procedure that uses estimated latent variables can be problematic. We propose an extension of the supervised topic model that jointly estimates a topic model and a regression model to predict an outcome using both the latent topic proportions and other manifest predictors. Our model, supervised latent Dirichlet allocation with covariates, can be fit in a single stage rather than two stages, allows for evaluation of the incremental validity of the topics given other established measures, provides concise summarization of the text, and models relationships between the topics and the outcome. To estimate the SLDAX model, we derived a Markov Chain Monte Carlo sampling algorithm. We also developed an R package, psychtm, that implements SLDAX. Performance of the model for different data characteristics is evaluated in a simulation study. These results are used to make recommendations regarding suggested data requirements for the use of the SLDAX model in applications. Finally, we demonstrate the application of SLDAX on an empirical data set.

A confidence interval of noncentrality compatible with test of a point null

Hao Wu
Vanderbilt University

Important fit and effect size measures such as RMSEA and ω2 are based on the noncentrality parameter in an F or χ^2 distribution. The CI of these quantities are usually constructed by inverting an equal tail rejection region of a test, so the upper and lower limits are also confidence bounds. However, such CIs are not compatible with a test of a point null hypothesis. For example, a 95% CI may contain zero but the test of exact fit or zero effect size at level α=0.05

rejects the null. This is because the test of a point null becomes a one sided test when the point null is located at the boundary value. In this work a CI is constructed by inverting a test based on the likelihood ratio of the noncentral F or χ^2 family. This CI will always correspond to a test of point null hypothesis.

Experimental evidence extraction system in data science with hybrid table features and ensemble learning (Cancelled)

Wenhao Yu, Qingkai Zeng, & Meng Jiang
University of Notre Dame

Data Science has been one of the most popular fields in higher education and research activities. It takes tons of time to read the experimental section of thousands of papers and figure out the performance of the data science techniques. In this work, we build an experimental evidence extraction system to automate the integration of tables into a database of experimental results. First, it crops the tables and recognizes the templates. Second, it classifies the column names and row names into "method", "dataset", or "evaluation metric", and then unified all the table cells into -quadruples. We propose hybrid features including structural and semantic table features as well as an ensemble learning approach for column/row name classification and table unification. SQL statements can be used to answer questions such as whether a method is the state-of-the-art or whether the reported numbers are conflicting.

Amending a popular dataset and improving scientific entity recognition with no-schema distant supervision

Qingkai Zeng
University of Notre Dame

Scientific entity recognition is an important task of information extraction. With specialized knowledge needed, scientific corpus annotation is expensive and has low reliability issue. We propose an empirical, visual method to evaluate the consistency of annotation on training and test sets. The idea is that if consistent, the training set and test set are predictive of each other. We observe anomalies that indicate inconsistency on SCIERC, a popular dataset in AI domain. After re-annotate more than 26% of sentences in the test set, we do not see anomalies in amended data. Our second contribution is to propose a new distant supervision method to learn from external dictionaries with no type schema. Our method improves the F1 score of scientific entity recognition and typing by 3.1% on original SCIERC and 4.5% on amended SCIERC.

An improved stochastic EM algorithm for large-scale full-information item factor analysis

Siliang Zhang
Department of Statistics

London School of Economics and Political Science

In this paper, we explore the use of the stochastic EM algorithm for large-scale full-information item factor analysis. Innovations have been made on its implementation, including an adaptive-rejection-based Gibbs sampler for the stochastic E step, a proximal gradient descent algorithm for the optimization in the M step, and diagnostic procedures for determining the burn-in size and the stopping of the algorithm. These developments are based on the theoretical results of Nielsen, as well as advanced sampling and optimization techniques. The proposed algorithm is computationally efficient and virtually tuning-free, making it scalable to large-scale data with many latent traits and easy to use for practitioners. Standard errors of parameter estimation are also obtained based on the missing information identity. The performance of the algorithm is evaluated through simulation studies and an application to the analysis of the IPIP-NEO personality inventory. Extensions of the proposed algorithm to other latent variable models are discussed.

Robust Bayesian growth curve modeling using double robust methods, growth curve modeling, conditional medians, asymmetric Laplace distribution conditional medians

Tonghao Zhang, Xin Tong, & Jianhui Zhou
Department of Statistics
University of Virginia

Growth curve models are widely used to analyze longitudinal data in social and behavioral sciences. While the estimation of Gaussian based growth curve model is relatively easy, the normality assumption is often violated in practice. Failing to account for non-normality may lead to unreliable model estimation and misleading statistical inference. Although a robust growth curve model using conditional medians was recently proposed and outperformed the traditional growth curve modeling using conditional means, this robust approach still does not perform well when the random effects contain extreme values. In this work, we propose a robust growth curve model against the presence of both outliers and leverage observations by employing two conditional medians separately for the measurement errors and random effects. Model estimation and inferences are conducted in the Bayesian framework. Laplace distributions are used to convert the problem of median estimation into a problem of obtaining the maximum likelihood estimator for a transformed model. Monte Carlo simulation studies have been conducted to evaluate the numerical performance of the proposed approach with contaminated data. The results show that the proposed approach yields more accurate and efficient parameter estimates. We illustrate the application of the developed robust approach using a real dataset to study the child development of academic achievement, executive function, social understanding and mastery orientation in different school systems.

Predicting authoritarian crackdowns: A machine learning approach

Weifeng Zhong
Senior Research Fellow

Mercatus Center

Julian TszKin Chan
Senior Economist in the Finance
Bates White Economic Consulting

We develop a quantitative indicator to predict if and when a series of protests in China, such as the one in Hong Kong in 2019, will be met with a Tiananmen-like crackdown. The indicator takes as input protest-related articles published in the People's Daily---the official newspaper of the Communist Party of China. We use a set of machine learning techniques to detect the buildup of negative propaganda in the text against the protesters, and the method generates a daily mapping between the current date in the Hong Kong protest timeline and the ``as-if" date in the Tiananmen protest timeline. We call this counterfactual date the Policy Change Index for Crackdown for the 2019 Hong Kong protests, showing how close in time it is to the June 4, 1989 crackdown on the Tiananmen Square.