# Conference Program

## The 2019 Meeting of

## The International Society for Data Science and Analytics

**July 6-8, 2019**

**Nanjing, China**

**Sponsored by**



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

| Schedule | |
|---|---|
| **July 6, 2019** | |
| Time | Speaker and Title |
| *Morning* | |
| 7:00-9:00 | Registration and badge pickup |
| 9:00-9:15 | Welcome |
| 9:15-9:45 | **Prospect Theory and Stock Returns: A Test for Trading Behavior of Individual VS. Institutional Investors**<br><br>*Xiaoling Zhong[1*], Junbo Wang[2]*<br>*[1]International Institute of Finance, School of Management, University of Science and Technology of China*<br>*[2]Deparment of Economics and Finance, the City University of Hong Kong* |
| 9:45-10:15 | **A Nonparametric Multivariate Statistical Process Control Chart Based on Change Point Model**<br><br>*Yafei Xu*<br>*Beijing AI Lab Vivo Communication Technology Co. Ltd.,*<br>*China* |
| 10:15-10:45 | **WeibullR an R Package for Weibull Analysis for Reliability Engineering**<br><br>*David Silkworth*<br>*Managing Director*<br>*OpenReliability.org*<br>*United States* |
| 10:45-11:00 | Coffee break |
| 11:00-11:30 | **Evaluating informative hypotheses using the Bayes factor**<br><br>*Dr. Xin Gu*<br>*Department of Educational Psychology*<br>*East China Normal University* |
| 11:30-12:00 | **A Slice Inverse Regression Algorithm Based on k-Medoids Clustering**<br><br>*Jiashan Tang*, Michael Ng and Zhichao Xie*<br>*Professor and Vice Dean*<br>*College of Science*<br>*Nanjing University of Posts and Telecommunications* |

| | |
|---|---|
| | *China* |
| 12:15-2:00 | Lunch |
| *Afternoon* | |
| 2:00-2:30 | **Improving Teaching Evaluation using Text Mining**<br><br>*Zhiyong Zhang*<br>*Associate Professor*<br>*University of Notre Dame*<br>*United States* |
| 2:30-3:00 | **Comparing three MASEM approaches to quantifying or explaining between-study heterogeneity in SEM parameters**<br><br>*Zijun Ke*<br>*Assistant Professor*<br>*Sun Yat-Sen University*<br>*China* |
| 3:00-3:30 [Zoom] | **Model Uncertainty in the Comparison of Two Single Dengue Outbreaks**<br><br>*Carlos Rafael Sebrango Rodríguez [1], Lizet Sánchez Valdés [2], Ziv Shkedy [3], Vivian Sistachs Vega [4]*<br>*[1] Universidad de Sancti Spiritus "José Martí Pérez", Cuba*<br>*[2] Centro de Inmunología Molecular, Cuba*<br>*[3] Center for Statistics, Hasselt University, Belgium*<br>*[4] Universidad de La Habana, Cuba* |
| 3:30-3:45 | Coffee break |
| 3:45-4:15 | **An evaluation of statistical differential analysis methods in single-cell RNA-seq data**<br><br>*Dongmei Li*<br>*University of Rochester*<br>*United States* |
| 4:15-4:45 | **Exploring Spatio-Temporal Patterns of Air Quality Index Data in China**<br><br>*Haokun Tang, Yulin Xie, Binbin Lu*<br>*School of Remote Sensing and Information Engineering*<br>*Wuhan University*<br>*China* |

| | |
|---|---|
| 4:45-5:15 | **Pivot Analysis in Weighted Linear Regression**<br><br>*Yuancheng Si*<br>*School of Mathematics University of Manchester*<br>*UK* |
| July 7, 2019 | |
| Time | |
| Morning | |
| 8:00-9:00 | Registration and badge pickup |
| 9:00-9:30 | **On the Estimation with Factor Analysis in High Dimensional Settings**<br><br>*Kentaro Hayashi*<br>*Associate Professor*<br>*University of Hawaii at Manoa*<br>*United States* |
| 9:30-10:00 | **Comparing Regression-Based and Structural Equation Modeling Based Approaches for Conditional Process Analysis**<br><br>*Wai Chan*<br>*Associate Professor*<br>*The Chinese University of Hong Kong*<br>*Hong Kong* |
| 10:00-10:30 | **Latent growth curve models with VAR residuals for longitudinal mediation analysis**<br><br>*Xiao Liu*<br>*University of Notre Dame*<br>*United States* |
| 10:30-10:45 | Coffee break |
| 10:45-11:15 | **A Structural Equation Modeling approach to Reliability Analysis When Data Are Items Having Different Numbers of Ordered Categories**<br><br>*Seohyun Kim, Laura Lu*, Allan Cohen*<br>*Associate Professor*<br>*University of Georgia*<br>*USA* |
| 11:15-11:45 | **VBTree Tutorial: Automatic Completion of Group Operation on Structural Dataset** |

| | |
|---|---|
| | *Chen Zhang*<br>*Department of Materials Processing, Graduate School of Engineering*<br>*Chiba Lab, Institute for Materials Research*<br>*Tohoku University*<br>*Japan* |
| 11:45-12:15 | **Latent Variable Regression by Partial Least Squares and Four Other Composite Scores: Consistency, Bias and Correction**<br><br>*Ke-Hai Yuan*, Yong Wen, and Jiashan Tang*<br>*University of Notre Dame and Nanjing University of Posts and Telecommunications*<br>*USA and China* |
| 12:15-2:00 | Lunch |
| Afternoon | |
| 2:00-2:30 | **A Regularized Method for Parameters Selection in Structural Equation Model**<br><br>*Chengfang Zhu*<br>*Nanjing University of Posts and Telecommunications*<br>*China* |
| 2:30-3:00 [zoom] | **Extension of Biplot Methodology to Multivariate Regression Analysis**<br><br>*Opeoluwa F Oyedele*<br>*University of Namibia*<br>*Namibia* |
| 3:00-3:30 | **A Bayesian Estimation of Deaths and Damages in Major Cyclones, Famine and Floods in Bangladesh**<br><br>*Hasinur Rahaman Khan*, Mahmuda Jahan*<br>*Associate Professor of Applied Statistics*<br>*ISRT*<br>*University of Dhaka*<br>*Bangladesh* |
| 3:30-3:45 | Coffee break |
| 3:45-4:15 | **On the number of components in mixture model based on EM algorithm**<br><br>*Yanglu Zhao*, Dandan Duan, Raomin Hu, Jiashan Tang, Yong Wen, Ke-hai Yuan* |

| | |
|---|---|
| | *College of Science*<br>*Nanjing University of Posts and Telecommunications*<br>*China* |
| 4:15-4:45 | **New Social Media Sampling Strategy: An Exploratory Study using Non-API methods for Social Data Extraction**<br><br>*Karl Ho*<br>*Clinical Associate Professor, Social Data Analytics Program, School of Economic, Political and Policy Sciences*<br>*University of Texas at Dallas*<br>*United States* |
| 4:45-5:15 | **A Review of Aspect-Based Sentiment Analysis with an Application on Teaching Evaluation**<br><br>*Wen Qu*<br>*University of Notre Dame*<br>*United States* |
| 5:30-8:00 | Banquet |
| July 8, 2019 | |
| Time | |
| Morning | |
| 9:00-9:30 [Zoom] | **A Non-Parametric Model to Address Overdispersed Count Response in a Longitudinal Data Setting with Missingness**<br><br>*Hui Zhang*<br>*Associate Member, Department of Biostatistics*<br>*St. Jude Children's Research Hospital*<br>*United States* |
| 9:30-10:00 [Zoom] | **Mediation Analysis for Complex Surveys with Balanced Repeated Replications**<br><br>*Yujiao Mai*<br>*St. Jude Children's Research Hospital*<br>*United States* |
| 10:00-10:30 [Zoom] | **MCMC Bootstrap Based Approach to Power/Sample Size Evaluation**<br><br>*Oleksandr Mykolayovich Ocheredko* |

| | |
|---|---|
| | *Professor and chairman of social medicine and organization of health services*<br>*Vinnytsya National Medical University*<br>*Ukraine* |
| 10:30-10:45 | **A More Accurate Estimator of Multiple Correlation Coefficient**<br><br>*Lu Peng and Bingjiang Li*<br>*Nanjing University of Posts and Telecommunications*<br>*China* |
| 10:45-11:15 [Zoom] | **Advances of Social Network Analysis in Psychological Sciences**<br><br>*Haiyan Liu*<br>*Assistant Professor*<br>*University of California, Merced*<br>*United States* |
| 11:15-11:45 | **A General Bayesian Model-Based Imputation Approach for Multilevel Models with Non-linear Effects: A Sequential Approach**<br><br>*Han Du*, Craig Enders*<br>*Assistant Professor*<br>*University of California, Los Angeles*<br>*United States* |
| 11:45-12:15 | **Correlation Analysis between Tourism Development, Economic Growth and Carbon Emissions: A Comparative Analysis Based on Six Provinces in the Central China**<br><br>*Zhibiao Wang*[1,2], Peibo Yao[2]*<br>*1.Research Centre of Exploiting and Utilizing Featured Resources in Wuling Mountainous Area, Yangtze Normal University*<br>*2. Research Centre of Innovation and Development of Cultural Industries in the Central Plains, Henan University* |
| | End of the conference |

# ABSTRACTS

## Comparing Regression-Based and Structural Equation Modeling Based Approaches for Conditional Process Analysis

*Wai Chan*
*Associate Professor*
*The Chinese University of Hong Kong*

In social science research, the analysis of moderation, mediation, and conditional process has become increasingly popular. Generally speaking, moderation occurs when the relationship between the independent variable (X) and the dependent variable (Y) varies as a function of a third variable or moderator (W). Mediation occurs when the effect of X on Y is transmitted through an intervening variable or mediator (M). Conditional process, collectively, refers to models that combine both mediation and moderation processes. These models are important because they allow researchers to understand, describe, and explain complex social phenomenon and human behavior in a more accurate way. Traditionally, conditional process models are analyzed by the regression-based approach (e.g., Hayes, 2018). This approach, however, does not provide the goodness-of-fit test and therefore there is no way to tell if a hypothesized model fits the data adequately. Until recently, a structural equation modeling (SEM)-based approach was proposed as an alternative (Kwan & Chan, 2018), which entails a likelihood ratio test for model goodness-of-fit assessment. To compare these two approaches empirically, a simulation study is conducted. Four different methods are used to (i) estimate the model parameters and (ii) evaluate the model goodness-of-fit (if applicable). Specifically, the first method is based on regression approach, and the other three are SEM-based methods that subsume three different working models as discussed in the literature. In addition, we examine the performance of these four methods when the hypothesized model is misspecified as some similar alternative structures. Based on the simulation results, we recommend an optimal modeling strategy that provides accurate parameter estimates and is sensitive to model misspecification.

## A General Bayesian Model-Based Imputation Approach for Multilevel Models with Non-linear Effects: A Sequential Approach

*Han Du*, Craig Enders*
*Assistant Professor*
*University of California, Los Angeles*
*United States*

Despite the broad appeal of missing data handling approaches that assume a missing at random (MAR) mechanism (e.g., multiple imputation and maximum likelihood estimation), some very common analysis models in the behavioral science literature are known to cause bias-inducing

problems for these approaches due to the incompatibility issue. Regression models with incomplete interactive or polynomial effects are a particularly important example because they are among the most common analyses in behavioral science research applications. Several Bayesian multiple imputation methods that yield imputation models compatible with the analysis model have been proposed in the last several years but they are limited in specific scenarios. The purpose of this paper is to outline a general form of a sequential approach for model-based multiple imputation, which can handle a wide range of interactive and non-linear effects up to three-level models. Computer simulation results suggest that this new approach can be quite effective when applied to multilevel models with random coefficients and interaction effects. In most scenarios that we examined, imputation-based parameter estimates were quite accurate and tracked closely with those of the complete data. The new procedure is available in the Blimp software application for macOS, Windows, and Linux.

---

**Evaluating informative hypotheses using the Bayes factor**

*Dr. Xin Gu*
*Department of Educational Psychology*
*East China Normal University*

Applied researchers are well-acquainted with null-hypothesis significance testing (NHST). It is, arguably, their main inferential tool. In the last decade the limitations of NHST have extensively been discussed. To name two: the null-hypothesis only rarely represents the expectations that researchers have; and, evaluation of the p-value versus the benchmark ".05" leads to undesirable phenomena like questionable research practices and publication bias. This presentation will introduce an alternative for NHST, that is, hypothesis evaluation using the Bayes factor. Using this approach, in addition to the traditional null and alternative hypotheses also informative hypotheses that represent the expectations that researchers have can be evaluated. A simple example of an informative hypothesis is Hi: m1 > m2 > m3, where the m's denote the means in each of three groups. Furthermore, the Bayes factor and posterior model probabilities are used for the evaluation of informative hypotheses. The Bayes factor is a measure of support for two hypotheses, if, for example, BFij = 10, then there is 10 times more support in the data for Hi than for Hj. While posterior model probabilities can be used to compare three or more hypotheses. To compute Bayes factor and posterior model probabilities, an R package "bain" is developed, which can be downloaded and installed from R CRAN.

---

**On the Estimation with Factor Analysis in High Dimensional Settings**

*Kentaro Hayashi*
*Associate Professor*
*University of Hawaii at Manoa*
*United States*

When the number of variables is larger than the sample size, the sample covariance matrix is no longer positive definite, and its inverse does not exists. Under the sparsity assumption, the problem can be dealt with by methods such as the lasso or penalized likelihood. However, in high-dimensional settings in behavioral sciences, the sparsity assumption does not necessarily hold. The number of variables is often greater than the sample size while they might still be comparable. Under such circumstances, unweighted least squares (ULS) and ridge approaches may be good options in estimating the parameters in factor analysis. We examine and compare different approaches, and show that some approach gives relatively small mean square errors when the dimensions are larger than the sample size.

---

**Comparing Three MASEM Approaches to Quantifying or Explaining Between-Study Heterogeneity in SEM Parameters**

*Zijun Ke*
*Assistant Professor*
*Sun Yat-Sen University*
*China*

Meta-analytic structural equation modeling (MASEM) is gaining attention from researchers because it is increasingly recognized as a way to build and test theory. However, one of the major challenges of MASEM research- how to model and explain meaningful effect size heterogeneity (heterogeneity in structural equation modeling parameters) - cannot be handled appropriately by the conventional two-stage approaches to MASEM. Recently three novel methods based on different model specifications have been proposed to tackle this issue: full information MASEM, one-stage MASEM, and Bayesian MASEM. Yet, the relative theoretical advantages and disadvantages of these methods are largely unknown. In this study, we will discuss the theoretical differences among the three methods and show via two real MASEM examples that how these theoretical differences will impact real MASEM research.

---

**A Bayesian Estimation of Deaths and Damages in Major Cyclones, Famine and Floods in Bangladesh**

*Hasinur Rahaman Khan*, Mahmuda Jahan*
*Associate Professor of Applied Statistics*
*ISRT*
*University of Dhaka*
*Bangladesh*

In practice, we often deal with the situation where pieces of information gathered from different sources vary from each other profoundly. When none of these sources are completely reliable, it is needed to take each and every one of them into account. In this paper, we demonstrate how to solve this problem and present a systematic and unified way of combining the collected knowledge pieces of deaths and damages by implementing Bayesian approach to major cyclone,

famine and flood data in Bangladesh. Supra-Bayesian models are built by combining the experts' opinion while treating them as observed data. A likelihood function is specified as the distribution of experts' opinion, parametrized on the basis of how we assume the experts tend to overestimate or underestimate the parameter and expressed our prior knowledge on the parameter by assigning a prior distribution to it. Therefore, a posterior distribution for the parameter of interest can be obtained, in which the prior opinion is updated on the ground of the opinions expressed by the experts. Four priors are used to avoid any biases to obtain posteriors. Later, posterior distributions are obtained for each case and pooled together using meta analysis to get a single estimate.

---

**New Social Media Sampling Strategy: An Exploratory Study using Non-API methods for Social Data Extraction**

*Karl Ho*
*Clinical Associate Professor, Social Data Analytics Program, School of Economic, Political and Policy Sciences*
*University of Texas at Dallas*
*United States*

Conventional data extraction methods using Application Program Interface (API) provide convenient channels for getting direct access to data. The fast-growing body of public opinion studies employing social media significantly benefits from the APIs provided by companies such as Facebook, Twitter and Google. The essential weakness of the API methods however is the "blackbox" process that lacks transparency and stability (Morstatter et al. 2013). Social media companies can alter the amount of data available and fashion of acquisitions at own discretion at any point of time. In this study, a non-API method is introduced to tap into high volume of social media data without being subject to the API institutional constraints. Since the new approach does not rely on API tokens and specifications, it points to a new direction of social and web data collection that can provide more information on the population structure of social network groups such as political campaigns, public policy advocates and online propaganda. An illustration will be provided to demonstrate extraction of historical social media data for network analysis and model estimation.

---

**A Structural Equation Modeling approach to Reliability Analysis When Data Are Items Having Different Numbers of Ordered Categories**

*Seohyun Kim, Laura Lu*, Allan Cohen*
*Associate Professor*
*University of Georgia*
*USA*

This study describes a structural equation modeling (SEM) approach to reliability analysis when data are items having different numbers of ordered categories. A simulation study is provided to

compare the performance of this reliability coefficient, coefficient alpha and population reliability for tests having items with different numbers of ordered categories, a one-factor and a bi-factor structures, and different skewness distributions of test scores. Results indicated that the proposed reliability approach provided accurate population reliability in most conditions. An empirical example was used to illustrate the performance of the different coefficients for a test of items with 2 or 3 ordered categories.

---

**An Evaluation of Statistical Differential Analysis Methods in Single-Cell RNA-seq Data**

*Dongmei Li*
*University of Rochester*
*United States*

Background: Single-cell RNA-Seq is gaining popularity in recent years. Compared to bulk RNA-Seq, single-cell RNA-Seq allows the gene expression being measured within individual cells instead of mean gene expression levels across all cells in the sample. Thus, cell-to-cell variation of gene expressions could be examined. Gene differential expression analysis remains the major purpose in most Single-cell RNA-Seq experiments and many tools have been developed in recent years to conduct gene differential expression analysis for Single-cell RNA-Seq data.

Results: Through simulation studies and real data examples, we evaluate the performance of five open-source popular methods used for gene differential expression analysis in single-cell RNA-seq data. The five methods include DEsingle (Zero-inflated negative binomial model), Linnorm (Empirical Bayes method on transformed count data using the limma package), Monocle2 (An approximate Chi-Square likelihood ration test), MAST (A generalized linear hurdle model), and DESeq2 (A generalized linear model with empirical Bayes approach). We assessed the false discovery rate (FDR) control, sensitivity, specificity, accuracy, and area under the receiver operating characteristics (AUROC) curve for all five methods under different sample sizes, distribution assumptions, and proportions of zeros in the data.

Conclusions: We found the MAST and Linnorm performs relatively better than other methods with higher AUROC, when there are some proportion of zeros in the single-cell RNA-seq data after filtering. However, when the proportions of zeros are close to zero, the DESingle, Linnorm, and DESeq2 performs relatively better than others with higher AUROC. When sample size increases to 100 in each group, MAST shows the best performance with the highest AUROC regardless of the proportion of zeros in the data.

---

**Advances of Social Network Analysis in Psychological Sciences**

*Haiyan Liu*
*Assistant Professor*

*University of California, Merced*
*United States*

Social network analysis is an interdisciplinary research topic of mathematics, statistics, computer sciences, and psychology. A social network comprises actors (e.g., students, collaborators) and potential ties among them. Such social relations reflect dependence among people within a social network, which is important to people's subjective well-being and behavior development. Understanding the formation and quality of social relations, such as friendship, has been traditionally conducted in psychology and sociology. In most of those studies, only data on dyads with social relations are available due to the specific data collection process, which limits the techniques used in the analysis and hides many patterns in human behaviors unintentionally. Social network data, however, contain data on both dyads with social relations and social relations. Therefore, they provide platforms for traditional research questions (e.g., What leads to friendship) and potentially lead to better results. In addition, social networks are rich with information on community structures, which closely relates to many psychological attributes. Therefore, social network analysis has also become a battlefield for new types of hypotheses on human behaviors. In this talk, I will introduce the concept of social network analysis and demonstrate its developments through several empirical examples.

---

**Latent growth curve models with VAR residuals for longitudinal mediation analysis**

*Xiao Liu*
*University of Notre Dame*
*United States*

Mediation analysis using longitudinal data has become increasingly popular. To perform longitudinal mediation analysis, different models have been proposed, such as the latent growth curve mediation model (LGCM) and the cross-lagged panel mediation model (CLPM). In the current study, we proposed an alternative longitudinal mediation model (referred to as the LGCM-CLRM), where a system of latent growth curve models is used to describe the deterministic inherent trajectories of each individual and a vector autoregressive model is used to describe the within-individual stochastic deviations from the latent trajectory. Compared with existing longitudinal mediation models, the proposed model allows mediation effects in both level-1 and level-2 models, and thus could disentangle different types of mediation effects. The proposed model can be estimated in the multilevel structural equation modeling framework. Simulation studies were performed to evaluate the estimation quality. We also provided a real data example (with Mplus syntax) for illustration.

---

**Mediation Analysis for Complex Surveys with Balanced Repeated Replications**

*Yujiao Mai*

*St. Jude Children's Research Hospital*
*United States*

Mediation analysis is to investigate the role of a third variable (also called a mediator) as a transmitter in the relationship between the exposure and the outcome. Although mediation analysis and corresponding computer tools have been widely applied in research practice, applications to survey data of complex sampling designs have not been addressed. As complex sampling designs using balanced repeated replications are common in finite-population-based studies such as national surveys, this study introduces a mediation analysis method adjusting for complex surveys with balanced repeated replications and develops the software packages in R, Rshiny, and SAS. The study in the end illustrates the application of the method and packages to Tobacco Use Supplement to the Current Population Survey.

---

**MCMC Bootstrap Based Approach to Power/Sample Size Evaluation**

*Oleksandr Mykolayovich Ocheredko*
*Professor and chairman of social medicine and organization of health services*
*Vinnytsya National Medical University*
*Ukraine*

Power calculation is important and evergreen applied statistical avenue. Here I delivered suggestion on enrichment of statistical tools by combination of bootstrap and MCMC modeling. Novelty suggests application of possible data generation mechanism using MCMC and power estimation in bootstrap procedure. I delineated further generalizations that are not incorporated in statistical software yet and demonstrated basic applications using SAS/STAT POWER Procedure examples. One concerns ANOVA the other deals with survival curves. An illustrious advantage of using MCMC is the possibility to exploit distributions of parameters of interest instead of ubiquitously used point estimates. The other methodological advancement though not demonstrated in the paper is the possibility to combine preliminary or historically observed data with experts' views. But foremost appealing to application environment is flexibility no more confined to basic situations rendered by statistical software. I've chosen BUGS language to demonstrate the program code that can be run on WinBUGS, OpenBUGS, and JAGS engines.

---

**Extension of Biplot Methodology to Multivariate Regression Analysis**

*Opeoluwa F Oyedele*
*University of Namibia*
*Namibia*

At the core of multivariate statistics is the investigation of relationships between different sets of variables. More precisely, the inter-variable relationships and the casual relationships. The latter

is a regression problem, where one set of variables is referred to as the response variables and the other set of variables as the predictor variables. In this situation, the effect of the predictors on the response variables is revealed through the regression coefficients. Results from the resulting regression analysis can be viewed graphically using the biplot. The consequential biplot provides a single graphical representation of the samples together with the predictor variables and response variables. In addition, their effect in terms of the regression coefficients can be visualized, although sub-optimally, in the said biplot.

---

**A More Accurate Estimator of Multiple Correlation Coefficient**

*Lu Peng and Bingjiang Li*
*Nanjing University of Posts and Telecommunications*
*China*

The squared multiple correlation coefficient ($R^2$) is the most widely used measure of the goodness of fit in regression analysis. Unfortunately, the sample $R^2$ is biased for estimating its population counterpart ($\rho^2$), and their difference increases as the number of variables ($p$) increases. Efforts have been made on modifying $R^2$. The most notable result is the adjusted one ($R_{adj}^2$) , which considers the influence of the sample size ($N$) and $p$. However, $R_{adj}^2$ is still biased, and there dose not exist an unbiased estimator of $\rho^2$. Using empirical modeling and statistical learning, we develop a new formula for estimating the population $\rho^2$, denoted as $R_e^2$. The development involves obtaining the empirical bias $F(R)$-$F(\rho)$ via Monte Carlo simulation across many conditions, where $F(.)$ is the Fisher's z-transformation. The empirical bias is then predicted by functions of $N$, $p$ and the value of $R^2$. Lasso regression and best-subsets regression are used to identify the best predictors of empirical bias. The corrected $R_e^2$ is obtained via a bias correction to $R^2$. Result of cross validation shows that the empirically corrected $R_e^2$ can perform a lot better than both $R^2$ and $R_{adj}^2$.

---

**A Review of Aspect-Based Sentiment Analysis with an Application on Teaching Evaluation**

*Wen Qu*
*University of Notre Dame*
*United States*

With the rapid growth of digital age, text mining becomes explosive popular in recent two decades. Various techniques and methods come out to manage and analyze the text to exploit the underlining information. Among them, the aspect-based sentiment analysis (ABSA), which is a research field that studies people's opinion, sentiment toward attributions or aspects of individual entities, attract researchers in both business and academic world. ABSA was first proposed in 2010 (Thet, Na, & Khoo, 2010). It first extracts the relevant aspects of a specific entity and then determines the sentiment for each aspect. To our knowledge, there is no ready-to-use R packages

or functions for ABSA. Since R language dominates in social and behaviors science, in this study, we offer a brief but comprehensive review of ABSA and apply it on a teaching evaluation study to illustrate how to conduct this study using R.

---

**Model Uncertainty in the Comparison of Two Single Dengue Outbreaks**

*Carlos Rafael Sebrango Rodríguez [1], Lizet Sánchez Valdés [2], Ziv Shkedy [3], Vivian Sistachs Vega [4]*

*[1] Universidad de Sancti Spiritus "José Martí Pérez", Cuba*
*[2] Centro de Inmunología Molecular, Cuba*
*[3] Center for Statistics, Hasselt University, Belgium*
*[4] Universidad de La Habana, Cuba*

In recent years there has been increased interest in using statistical models for analysis of single dengue outbreaks based on the reported cumulative cases. The 3 parameter logistic model (3P logistic) and the Richards model have been used to estimate primary epidemiological parameters in single dengue outbreak. A topic that could be of interest to epidemiologists is the comparison of two single dengue outbreaks based on estimates of key epidemiological parameters: The turning point, the final size and the basic reproductive number $R_0$. In order to compare two single dengue outbreaks we create a model that takes into account both outbreaks simultaneously. In this paper, we describe different methodologies based on Frequentist and Bayesian approaches that takes into account the model uncertainty in the comparison of two single dengue outbreaks, based on the parameter estimates of the primary epidemiological parameters. The frequentist approach consists of comparing outbreak doing an extension of 3P logistic and Richards models and the use of model averaging for taking into account model uncertainty. In the Bayesian approach we use a Bayesian hierarchical model and we use Bayesian model averaging applying Gibbs variable selection. The proposed methods are applied to dengue outbreaks that occurred in La Lisa municipality, Havana City, Cuba during 2006 and 2007 outbreaks.

---

**Pivot Analysis in Weighted Linear Regression**

*Yuancheng Si*
*School of Mathematics University of Manchester*
*UK*

According to Carl V. Lutzer (2017), the simple linear regression lines based on repeating single observations from a given datasets would pivots at certain pivot point. In this paper, we would discuss this behavior in more general case and give explanation about the pivot behavior.

---

**WeibullR an R Package for Weibull Analysis for Reliability Engineering**

*David Silkworth*
*Managing Director*
*OpenReliability.org*
*United States*

Life data analysis in the graphical tradition of Waloddi Weibull. The WeibullR package provides a flexible data entry capability with three levels of usage. Quick Fit Functions, wblr object model, and technical back end functions. WeibullR should appeal to the newest practitioners to the R community as well as seasoned researchers willing to examine deeper aspects of analysis. Presentation will follow content available at https://github.com/openrelia/WeibullR.gallery

---

**A Slice Inverse Regression Algorithm Based on k-Medoids Clustering**

*Jiashan Tang*, Michael Ng and Zhichao Xie*
*Professor and Vice Dean*
*College of Science*
*Nanjing University of Posts and Telecommunications*
*China*

Slice inverse regression (SIR) algorithm is a nonparametric dimensionality reduction technique. This algorithm can achieve the goal of reducing dimension without any assumption on parameters and without losing any information about the relationships among variables. Based on SIR, combined with the K-medoids clustering, we propose a K-medoids inverse regression (K-medoids IR) algorithm. Simulation result shows that this new algorithm can effectively avoid the interference caused by noise and can greatly reduce the deviation caused by outliers.

---

**Exploring Spatio-Temporal Patterns of Air Quality Index Data in China**

*Haokun Tang, Binbin Lu, Yulin Xie*
*School of Remote Sensing and Information Engineering*
*Wuhan University*
*China*

Objective: The study aims to explore the spatio-temporal patterns of air pollutions across China with the air quality index (AQI) data.

Background: Air pollution is harmful for human body and natural environment. The previous studies show that there is statistically significant associations between air pollution and mortality[1]. Evidences show that inhalable particulate matter can adversely affect the body's cardiopulmonary function[2]. China has experienced rapid economic development since the reform and opening up. However, this is accompanied by the deterioration of the ecological

environment. Air pollution is an important aspect[3]. At present, studies have been conducted on air pollution in China. Hu et al. studied the correlation between PM2.5 and PM10 in North China and Yangtze River Delta[4]. Lin et al. proposed a method for estimating PM2.5 using remote sensing satellites[5]. All these studies focused that the spatio-temporal distribution of air pollutions in China and assist make decision of pollution control.

Data: In 2012, China issued a new Environmental Air Quality Standard (GB3095-2012), and AQI (Air Quality Index) replaced the API (Air Pollution Index) as a new air pollution assessment standard[6]. AQI is calculated by IAQI (Individual Air Quality Index) which is calculated according to individual pollutants. Air pollutants include PM2.5, PM10, sulfur dioxide, nitrogen dioxide, ozone and carbon monoxide.

In this study, AQI data is collected from China Air Quality Online Monitoring and Analysis Platform(https://www.aqistudy.cn). AQI data for 2016-2018 was used in this study. The data includes six pollutants and corresponding AQI values for each day and month in 363 cities in China. The data is saved in file with CSV(Comma-Separated Values) format.

---

**Correlation Analysis between Tourism Development, Economic Growth and Carbon Emissions: A Comparative Analysis Based on Six Provinces in the Central China**

*Zhibiao Wang*[1,2], Peibo Yao[2]*
*1.Research Centre of Exploiting and Utilizing Featured Resources in Wuling Mountainous Area, Yangtze Normal University*
*2. Research Centre of Innovation and Development of Cultural Industries in the Central Plains, Henan University*

Tourism plays an important role in economic growth in many regions, but the problem of carbon emission arises with the development of tourism. It is of great significance to study the relationship among tourism development, economic growth and carbon emission. Therefore, this paper, making use of the data of tourism revenues, GDP and carbon emissions of six provinces in the central China from 1995 to 2016, conducted unit root test, cointegration test and Granger causality test so as to find their relations. The results showed: there existed a long-term equilibrium relationship among the tourism revenue, regional GDP and carbon emission of the six central provinces. As regards the relationship between tourism revenue and economic development, Granger causalities in Henan, Anhui and Hubei were two-way, and that in Hunan and Jiangxi province were one-way and from tourism revenue to economic development, while that in Shanxi province was one way from economic development to tourism revenue. As regards the relationship between carbon emission and tourism revenue, Granger causalities in Anhui, Shanxi and Hubei province were one-way and from carbon emission to tourism revenue, while that in Henan and Hunan province were one-way and from tourism revenue to carbon emission, while there existed no Granger causality between tourism revenues and carbon emissions in Jiangxi province. With respect to the relationship between carbon emission and economic development, Granger causalities in Henan, Anhui and Hubei province were one-way

and from carbon emission to GDP， and that in Hunan province was one-way and from economic growth to carbon emission; in Shanxi province, there was a two-way Granger causality between these two variables, while in Jiangxi province there was no Granger causality between them. Finding of the present study can, on the one hand help the central provinces to optimize tourism accordingly, and on the other hand enrich the research on sustainable development of tourism.

---

**A Nonparametric Multivariate Statistical Process Control Chart Based on Change Point Model**

*Yafei Xu*
*Beijing AI Lab Vivo Communication Technology Co. Ltd.,*
*China*

This article presents a nonparametric control chart based on the change point model, for multivariate statistical process control (MSPC). The main constituent of the chart is the energy test that focuses on the discrepancy between empirical characteristic functions of two random vectors. This new multivariate control chart highlights in three aspects. Firstly, it is nonparametric, requiring no pre-knowledge of the data generating processes. Secondly, this control chart monitors the whole distribution, and not only specific characteristics like mean or covariance. Thirdly, itis designed for online detection (Phase II), which is central for real time surveillance of stream data. Simulation study discusses in-control and out-of-control measures in context of mean shift and covariance shift. In the real application, three financial data sets (in 5, 29, 90 dimensions) were employed to analyze the performance of the control chart for financial surveillance. The results from both simulation and empirical studies, compared with benchmarks, strongly advocate the proposed control chart. In this paper, an R package `EnergyOnlineCPM' is contributed in CRAN for further study and practice of nonparametric MSPC.

---

**Latent Variable Regression by Partial Least Squares and Four Other Composite Scores: Consistency, Bias and Correction**

*Ke-Hai Yuan, Yong Wen, and Jiashan Tang*
*University of Notre Dame and Nanjing University of Posts and Telecommunications*
*USA and China*

In many fields, interesting attributes are latent that can only be observed via indicators. Structural equation modeling (SEM) is a key methodology for studying the relationship among the latent attributes. Two approaches to SEM have been presented in the literature. One is to fit the mean and covariance structure model to its sample counterpart by minimizing their difference via a discrepancy function, called covariance-based SEM (CB-SEM). The other approach is via repeated least squares regression with a single dependent variable each time, called partial least

squares SEM (PLS-SEM). Because PLS-SEM is essentially latent variable regression using composite scores, the parameter estimates are biased in general. This article analytically compares the size of the bias in regression coefficients of the following methods: PLS-SEM, regression analysis using the Bartlett-factor-scores, regression analysis using the separate and joint regression-factor-scores, and regression analysis with the unweighted composite scores. A correction to parameter estimators following mode A of PLS-SEM is also proposed. Monte Carlo results indicate that regression analysis using other composite scores can be as good as PLS-SEM with respect to bias and efficiency/accuracy. Results also indicate that the corrected estimators following PLS-SEM can be as good as that of CB-SEM.

---

**VBTree Tutorial: Automatic Completion of Group Operation on Structural Dataset**

*Chen Zhang*
*Department of Materials Processing, Graduate School of Engineering*
*Chiba Lab, Institute for Materials Research*
*Tohoku University*
*Japan*

This paper introduces an R package designed for engineering practitioners and researchers, which can make it possible that most of the plotting and data analyzing tasks could be directly achieved from their summary table of dataset. It is even unnecessary for users to master some complicated data reshaping operations in R aforehand. By understanding the built-in managing logic of VBTree as well as essential knowledge of R, even a novice can make their data processing workflow be achieved automatically.

---

**A Non-Parametric Model to Address Overdispersed Count Response in a Longitudinal Data Setting with Missingness**

*Hui Zhang*
*Associate Member, Department of Biostatistics*
*St. Jude Children's Research Hospital*
*United States*

Although widely used for comparing multiple samples in biomedical and psychosocial research, the analysis of variance (ANOVA) model suffers from a series of flaws that not only raise questions about conclusions drawn from its use, but also undercut its many potential applications to modern clinical and observational research. In this paper, we propose a new class of generalized ANOVA models to concurrently address all these fundamental flaws underlying this popular multi-group comparison approach so that it can be applied to many immediate as well as potential applications ranging from addressing an age-old technical issue in applying ANOVA to cutting-edge methodological challenges arising from the emerging effectiveness research

paradigm. By integrating the classic theory of U-statistics with the state-of-the-art concepts such as the inverse probability weighted estimates, we develop distribution-free inference for this new class of models to address missing data for longitudinal clinical trials and cohort studies. We illustrate the proposed class of models with both real and simulated study data, with the latter investigating behaviors of model estimates under small and moderate sample sizes.

---

**Improving Teaching Evaluation using Text Mining**

*Zhiyong Zhang*
*Associate Professor*
*University of Notre Dame*
*United States*

To better evaluate teaching performance is an important topic in education research. Traditionally, quantitative data are collected through Likert scales. However, such data ignore the subtle information regarding teaching. Although open-ended or free-style questions are often used to collect additional information on teaching performance, the resulting data are often in qualitative text format and rarely analyzed. Therefore, it is necessary to build a system to score and extract important information from the qualitative text data. In this study, we use a variety of machine learning algorithms such as LASSO and convolutional neural network to understand the relationship between the quantitative rating data and the qualitative text data. The study leads to a model that can quantify the text data for teaching evaluation.

---

**Modified Maximum Likelihood Estimators for Logistic Distribution Using Ranked Set Samples**

*Zili Zhang*
*The University of Manchester*
*United Kingdom*

In this article, the maximum likelihood estimators (MLE) and modified maximum likelihood estimators (MMLE) of the location and scale parameters for logistic distribution using simple random sampling and different sampling schemes of ranked set sampling (RSS) are obtained. The corresponding MLE and MMLE using RSS when the ranking is imperfect are considered too.

The paper is organised as follow. The MLE of $\beta = 1$ s based on SRS, RSS and RSS consist of first order statistic, denoted by RSSF and MMLE of $\beta$ based on RSS under perfect judgement ranking are discussed in Section 2. The same discussions are applied on location parameter $\mu$ in Section 3. In Section 4, the biases of the estimating equations in last two section and the properties of the estimators of $\beta$ and $\mu$ obtained by these different estimating equations under

imperfect ranking are discussed and compared. In the simulation study in Section 5, the properties in Section 4 are verified and conclusions are given.

---

**On the Number of Components in Mixture Model Based on EM Algorithm**

*Yanglu Zhao*, Dandan Duan, Raomin Hu, Jiashan Tang, Yong Wen, Ke-Hai Yuan*
*College of Science*
1. *Nanjing University of Posts and Telecommunications*
*China*

Mixture modeling has become a popular technique in data analysis. Because of model based, it typically yields more accurate results than conventional methods in cluster analysis. A key element is the number of components in the mixture model, and it determines the final result of data analysis. The expectation maximization (EM) algorithm is commonly used for parameter estimation in mixture model, and in the field of machine learning and data clustering. EM algorithm is an iterative algorithm for computing the maximum likelihood estimates of model parameters with incomplete data, and the number of components in mixture model cannot be observed or is a missing value. Researchers often use AIC and BIC to determine the number of components in mixture model. However, these two criteria are not reliable in applications, and often yield misleading results in real data analysis. Aiming at this problem, this paper proposes a new method to determine the number of components in mixture modeling. The method uses the scree plot of the likelihood function to determine the number of clusters. The simulation results show that the method of scree plot not only obtains the same number of components as AIC and BIC do in most cases, it can also yields more reliable results when conditions are not ideal, which are typical with real data. Subsequently, the new method is applied for parameter estimation of the fountain data from yellow stone national park.

---

**Prospect Theory and Stock Returns: A Test for Trading Behavior of Individual VS. Institutional Investors**

*Xiaoling Zhong[1], Junbo Wang[2]*
*[1]International Institute of Finance, School of Management, University of Science and Technology of China*
*[2]Deparment of Economics and Finance, the City University of Hong Kong*

Since the 1980s, prospect theory has been showed to have predictive power in various financial markets. It has been commonly accepted that prospect theory describes how people evaluate financial assets when making investment decisions. Using a regulatory change in the Chinese B-share market as an exogenous shock, we directly test the different trading behaviors between individual and institutional investors. The empirical evidences show that after the regulatory change, the predictive power of prospect theory becomes significantly stronger in the B-share

market, providing strong evidence that individual investors rely more on prospect theory when making investment decisions. This change of the predictive power of prospect theory for B-share market mainly comes from the probability weighting component, which reflects the “lottery-type” and “insurance-type” demands of investors.

---

**A Regularized Method for Parameters Selection in Structural Equation Model**

*Chengfang Zhu*
*Nanjing University of Posts and Telecommunications*
*China*

In structural equation model, in order to reduce the complexity of exploratory factor analysis and improve the model fit of confirmatory factor analysis, multiple researchers have used lasso regularized method to select parameters for the exploratory part of the model. In this paper, in order to verify the universality of this method, we compare the results of parameter selection by the LM-test method in EQS and the lasso regularized method in the lslx package of R, as well as the reconstructed model fit after adding the selected parameters by two methods through simulation experiment. Finally, through empirical analysis, the performance of two methods in parameter selection and improvement of model fit is further compared. The experimental results show that in terms of parameter selection, the number of correctly selected parameters by the lasso under "reasonable parameter selection" can reach the result obtained by the LM-test. As the sample size increases, the number of selected parameters increases, and the stability of the selected parameters of the two methods is basically consistent. At the same time, the number of incorrectly selected parameters by the lasso is less than that by the LM-test. In terms of the deviation between the constructed model parameter value and the population model parameter value, the average deviation of lasso is larger than that of LM-test from the factor load of non-zero terms, and the average deviation of lasso is smaller than that of LM-test from the factor load of zero terms. In terms of mean variance deviation of error terms or variance of latent variables \ covariance mean deviation, lasso is larger than LM-test. We add the selected parameters, reconstruct the model, and compare the overall fit of the model obtained by the two methods. From the average value of $\chi^2$、RMSEA、CFI, as the sample size increases, the model fit will increase accordingly. At the same sample size, the model fit obtained by lasso is slightly lower than that obtained by LM-test. Finally, empirical analysis results show that LM-test selects more parameters than lasso. Compared with confirmatory factor analysis, both methods improve the model fit to a certain extent, and the degree of improvement is basically the same.