Conference Program

Data Science and Psychology

The 2024 Meeting of the International Society for Data Science and Analytics

July 21-24, 2024

Vienna, Austria

Organizing Committee

- Haiyan Liu, Assistant Professor, University of California, Merced
- Laura Lu, Associate Professor, University of Georgia
- Wen Qu, Assistant Professor, Fudan University
- Jiashan Tang, Professor, Nanjing University of Posts and Telecommunications
- Xin Tong, Associate Professor, University of Virginia
- Ke-Hai Yuan, Professor, University of Notre Dame
- Zhiyong Zhang, Professor, University of Notre Dame

Sponsored by

- Notre Dame Global
- Institute for Educational Initiatives, University of Notre Dame
- School of Science, Nanjing University of Posts and Telecommunications
- International Society for Data Science and Analytics

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

Schedule	
All times are in the local time used in Vienna, Austria.	
Conference Venue: Park Suite 7, Hilton Vienna Park hotel.	
July 21, 2024	
Time	Title & Presenter
Breakfast	On own or provided by the hotel if booked through the conference link
9am – 5pm	Conference registration
2pm – 5pm	**Workshop 1: Conducting Meta-Analytic Structural Equation Modeling with R** Instructor: Mike Cheung -- National University of Singapore Assistant: Lingbo Tong [redacted]
7pm – 10pm	**Workshop 2: An Introduction to Generalized Structured Component Analysis Structural Equation Modelling (GSCA-SEM) and its Applications Using Free Software** Instructor: Heungsun Hwang -- McGill University Assistant: Austin Wyman [redacted]

	[redacted]
July 22, 2024	
Time	Title & Presenter
Breakfast	On own or provided by the hotel if booked through the conference link
Morning	Session Chair: Dr. Johnny Zhang
9:00 – 9:15	**Opening remarks** Johnny Zhang President of ISDSA Professor, University of Notre Dame
	Machine Learning Techniques
9:15 – 10:00	Regular talk **Scale-Invariance, Equivariance and Dependency of Structural Equation Models** Ke-Hai Yuan [kyuan@nd.edu]*; University of Notre Dame, USA; US Ling Ling [lingl@njupt.edu.cn]; Nanjing University of Posts and Telecommunications; CN Zhiyong Zhang [ZhiyongZhang@nd.edu]; University of Notre Dame; US
10:00 – 10:30	Regular talk **Deep learning generalized structured component analysis: An interpretable artificial neural network model with composite indexes** Gyeongcheol Cho [cho.1240@osu.edu]*; Department of Psychology, The Ohio State University; US Heungsun Hwang [heungsun.hwang@mcgill.ca]; McGill University; CA
10:30 – 11:00	Regular talk **Feature Selection for Binary and Multi-Class Classification Problems** Ayman Alzaatreh [aalzaatreh@aus.edu]*; American University of Sharjah; AE
11:00 – 12:00	*Speed talks (4)* **11:00-11:10** **Using machine learning methods in the presence of numerous measured**

	confounders in mediation analysis Milica Miocevic [milica.miocevic@mcgill.ca]*; McGill University; CA **11:10-11:20** **Psychometric AI: Integrating Modern Data Science in Psychological Testing and Assessment** Wei Wang [wwang@gc.cuny.edu]*; The Graduate Center, City University of New York; US Chapman Lindgren [clindgren@gc.cuny.edu]; US Max Lobel [max.lobel.2000@gmail.com]; US Kemar Pickering [kpickering@gradcenter.cuny.edu]; US **11:20-11:30** **Variable Selection via Regularized Psychometric Factor and Network Models in Multidimensional Context** Jiaying Chen [jc168@uark.edu]; University of Arkansas; US Jihong Zhang [jzhang@uark.edu]; University of Arkansas; US Xinya Liang [xl014@uark.edu]*; University of Arkansas; US **11:30-11:40** **Regularization Methods for Factor Models in the Presence of Partial Measurement Invariance** Emma Somer [emma.somer@mail.mcgill.ca]*; McGill University; CA Carl Falk [carl.falk@mcgill.ca]; McGill University; CA Milica Miocevic [milica.miocevic@mcgill.ca]; McGill University; CA
12:00-2:00 pm	Lunch Break, box lunch provided
Afternoon	**Bayesian Methods and Statistics** Session Chair: Dr. Xin Tong
2:00 – 2:30	Regular talk **Revisiting Bayesian Two Sample Inference** Xin Tong [xt8b@virginia.edu]*; University of Virginia; US Sarah Depaoli [sdepaoli@ucmerced.edu]; UC Merced;
2:30 – 3:00	Regular talk **Model Selection for Mixed-effects Models with Confidence Interval for LOO or WAIC Difference**

	Yue Liu [helena701@126.com]*; Institute of Brain and Psychological Sciences, Sichuan Normal University; CN Fan Fang [fangfan_ricca@163.com]; ; CN Hongyun Liu [hyliu@bnu.edu.cn]; Beijing Normal University; CN
3:00 – 3:30	Regular talk **Selection of Best Experience for Digital Platforms** Will Stamey [wstamey@nd.edu]; University of Notre Dame; US **Ken Kelley [kkelley@nd.edu] *; University of Notre Dame; US** Bhargab Chattopadhyay [bhargab@iiitvadodara.ac.in]; Indian Institute of Technology – Vadodara; IN T. Bandyopadhyay [tathagata@iima.ac.in]; IN
3:30 – 4:00	**Determining the Number of Factors in Exploratory Factor Analysis with Model Error** Yilin Li [yli49@nd.edu]*; University of Notre Dame; US Guangjian Zhang [gzhang3@nd.edu]; University of Notre Dame; US
4:00 – 4:30	Regular talk **Power Analysis for Cohort Sequential Designs** Lijuan Wang [lwang4@nd.edu]*; University of Notre Dame; US
4:30 – 4:40	Speed talk **Revisiting Estimation in Hierarchical Modeling and Optimal Design** Laura Lu [zlu@uga.edu]*; University of Georgia; US
5:00PM -	Dinner on own
July 23, 2024	
Morning	**Structural Equation Modeling (SEM) and Differential Equation Models** Session Chair: Dr. Ke-Hai Yuan
9:00 – 9:30	Regular talk **Instructing Language Models to Do Reasoning Wisely** Meng Jiang [mjiang2@nd.edu]*; University of Notre Dame; US
9:30 – 10:00	Regular talk **Bayesian Growth Curve Modeling with Measurement Error in Time**

	Lijin Zhang [lijinzhang@stanford.edu]*; Stanford University; US Wen Qu [wqu@fudan.edu.cn]; Fudan University; CN Zhiyong Zhang [zhiyongzhang@nd.edu]; University of Notre Dame; US
10:00 – 10:30	Regular talk **The Performance of Fit Measures in Detecting Structural Misspecification in SEM Using Both ML and Bayesian SEM** Chunhua Cao, University of Alabama
10:30 – 11:00	Regular talk **Piecewise Strategies for Fitting Differential Equation Models to Time Series Data** Yueqin Hu [yueqinhu@bnu.edu.cn]*; Beijing Normal University; CN
11:00 – 12:00	*Speed talks (3)* **11:00-4:15** **Regularized Integrated Generalized Structured Component Analysis** Heungsun Hwang [heungsun.hwang@mcgill.ca]*; McGill University; CA Gyeongcheol Cho [cho.1240@osu.edu]; Ohio State University; **11:15-11:30** **Simple Procedure To Estimate A Structural Equation Model With Latent Variables** Zouhair El Hadri [z.elhadri@um5r.ac.ma]*; Mohammed V University, Faculty of sciences Rabat; MA **11:30-11:45** **An Estimation Approach for Time-varying Effect Models Using Cubic Splines** Jingwei Li [jingweil@mailbox.sc.edu]; University of South Carolina; US Donna L. Coffman [dcoffman@mailbox.sc.edu]*;University of South Carolina; US
12:00-2:00 pm	Lunch Break, box lunch provided
Afternoon	**Psychological Studies** Session Chair: Dr. Laura Lu
2:00 – 2:30	Regular talk **A Comparison of Minimal-Effect Testing, Equivalence Testing, and the**

	Conventional Null Hypothesis Testing for the Analysis of Bi-factor Jiashan Tang [tangjs@njupt.edu.cn]*; Nanjing university of Posts and Telecommunications; CN Shunji Wang [WongShunchi@163.com]; Nanjing university of Posts and Telecommunications; CN Ke-Hai Yuan [kyuan@nd.edu]; University of Notre Dame; US
2:30 – 3:00	Regular talk **Introduction to an online app for SEM analysis with text data** Zhiyong Zhang [zzhang4@nd.edu]*; University of Notre Dame Notre Dame, IN 46556 USA; US
3:00 – 4:00	*Speed talks (4)* **3:00-3:10** **Unintentional and Intentional Code-switching of Chinese-English Bilinguals** Yanran Chen [ychen44@nd.edu]*; University of Notre Dame; US **3:10-3:20** **How are you really feeling? A dynamic network approach to detecting nomothetic patterns of emotion regulation ability** Austin Wyman [awyman@nd.edu]*; University of Notre Dame; US **3:20-3:30** **Bioinformatic Analysis of Immune-related LncRNA in Head and Neck Squamous Cell Carcinoma** Jiawen Wu [wujiaw96@163.com]; College of Science, Nanjing University of Posts and Telecommunications; CN Jiashan Tang [tangjs@njupt.edu.cn]*; College of Science, Nanjing University of Posts and Telecommunications; CN **3:30-3:40** **Predictive Analytics: Technical and Application on Health Data Imputation Using a Machine Learning Approach** Sandra Ihemeje [nkechisandy@gmail.com]*; ISDSA Nigeria; NG
6:00PM -	Dinner at Zwölf Apostelkeller

July 24, 2024	
Breakfast	On own or provided by the hotel if booked through the conference link
2pm – 5pm	**Workshop 3: Causal Moderated Mediation Analysis** Instructor: Xu Qin -- University of Pittsburgh Assistant: Lingbo Tong [redacted] [redacted]
5pm	End of the meeting

Invited Workshops

Workshop 1: Conducting Meta-Analytic Structural Equation Modeling with R
Instructor: Prof. Mike Cheung -- National University of Singapore

Time: 1pm-4pm, July 21, Vienna Time / 8pm-11pm, July 21, Singapore Time / 8am-11am, July 21, US Eastern Time
Location: Virtual on Zoom

The workshop will cover meta-analytic structural equation modeling (MASEM), which uses the techniques of meta-analysis and structural equation modeling to synthesize correlation matrices and fit hypothesized models on the combined correlation matrix. It can be used to test path models, confirmatory factor analytic models, and structural equation models from a pool of correlation matrices. MASEM offers the benefits of both meta-analysis and SEM.

During the workshop, I will provide an introduction to the basic theory of MASEM and demonstrate how to conduct the analyses with R. While some familiarity with R would be beneficial, the workshop is designed to be accessible to those who are new to the programming language.

Professor Mike Cheung specializes in quantitative methods and conducts research in structural equation modeling, meta-analysis, and multilevel modeling. His work focuses on the integration of meta-analysis and structural equation modeling. He is an Associate Editor of Research Synthesis Methods and Neuropsychology Review and serves on the editorial boards of several journals. For more information, please visit https://mikewlcheung.github.io/.

Workshop 2: An Introduction to Generalized Structured Component Analysis Structural Equation Modelling (GSCA-SEM) and its Applications Using Free Software
Instructor: Prof. Heungsun Hwang -- McGill University

Time: Time: 7pm-10pm, July 21, Vienna Time / 1pm-4pm, July 21, US/Canada Eastern Time
Location： Virtual on Zoom

Researchers in various fields are interested in studying the path-analytic relationships between constructs such as self-esteem, depression, socioeconomic status, etc. As constructs are abstract concepts that are not directly measurable, they are represented by proxies linked to empirical data or observed variables in statistical models. This enables researchers to test hypotheses about the relationships between constructs. There are two traditional ways of statistically representing constructs: (common) factors and components.

Structural equation modelling (SEM) is a general statistical framework for specifying and examining how such statistical representations as factors or components are related to observed variables and how they are related based on prior theory or knowledge. SEM has diverged into two domains, i.e., factor-based vs. component-based, depending on whether all constructs are represented as factors or components.

The two SEM domains include different statistical methods. Covariance structure analysis (CSA) has been a standard method for factor-based SEM, although there are other methods, including model-implied instrumental variable methods, factor score regression, structured factor analysis, and generalized structured component analysis with measurement errors incorporated (GSCA M) . On the other hand, generalized structured component analysis (GSCA) is the most general method for component-based SEM, which can include a long- standing component-based method, partial least squares path modelling, as a special case. It has been shown that when an SEM method developed for one domain is used for the other domain, it results in biased solutions. For example, CSA will provide biased estimates of parameters in models with components (e.g., component loadings and path coefficients relating components), whereas GSCA will yield biased estimates of parameters in models with factors (e.g., factor loadings and path coefficients connecting factors).

Over the decades, all the methods have been used exclusively for either of the two SEM domains, permitting researchers to estimate models with factors or components only. However, researchers may often need to include both factors and components in the model to consider a broad array of constructs from different disciplines. The next generation of SEM methods has recently emerged that permits estimating models with both factors and components in a unified framework. It includes consistent partial least squares and integrated generalized structured component analysis (IGSCA).

GSCA-SEM (generalized structured component analysis structural equation modelling) is an umbrella term that includes three SEM methods—GSCA, GSCA M , and IGSCA—for estimating models with components only, with factors only, or with both factors and components, respectively. GSCA-SEM is highly versatile in accommodating the two statistical representations of constructs. GSCA Pro is a stand-alone software program for GSCA-SEM. The software can be freely downloaded from its website (www.gscapro.com). It provides a graphical user interface that allows users to draw their model as a path diagram easily, fit GSCA-SEM to the model, and obtain results.

This workshop begins by explaining the conceptual foundations of GSCA-SEM, focusing on model specification and evaluation. It then provides step-by-step illustrations of using the free software for various GSCA-SEM applications.

Prof. Hwang’s research program is generally devoted to the development and application of quantitative methods to address diverse issues in psychology and various other fields. His recent interests include the development of data integration tools for high-dimensional data collected from multiple sources; the development of a statistical methodology for investigating associations among genetic, brain, and behavioural/cognitive phenotypes; and the development and application of predictive models or machine learning algorithms for predicting behavioural and cognitive outcomes using genetic, physiological, and psychological data.

Workshop 3: Causal Moderated Mediation Analysis
Instructor: Prof. Xu Qin -- University of Pittsburgh

Time: Time: 1pm-4pm, July 24, Vienna Time / 8am-11am, July 24, US Eastern Time

Location： Virtual on Zoom (Register to get the link)

Research questions regarding how, for whom, and where a treatment achieves its effect on an outcome have become increasingly valued. Such questions can be answered by causal moderated mediation analysis, which assesses the heterogeneity of the mediation mechanism underlying the treatment effect across individual and contextual characteristics. The purpose of this three-hour virtual course is to introduce the general definition, identification, estimation, and sensitivity analysis for causal moderated mediation effects under the potential outcomes framework. Participants will also learn how to use a user-friendly R package to conduct the analysis and visualize analysis results. The method introduction and the package implementation will be illustrated with a re-analysis of the National Evaluation of Welfare-to-Work Strategies (NEWWS) Riverside data.

Dr. Xu Qin is an Assistant Professor of Research Methodology at the School of Education (primary) and an Assistant Professor of Biostatistics at the School of Public Health (secondary). She holds a Ph.D. from the Department of Comparative Human Development at the University of Chicago and a B.S. and an M.S. in Statistics from the Renmin University of China.

Her research focuses on solving cutting-edge methodological problems in causal mediation analysis and multilevel modeling. She is also interested in using rigorous and innovative quantitative methods to evaluate the impacts of interventions and the underlying mechanisms. Methodologically, she has developed statistical methods and software for investigating the heterogeneity in causal mediation mechanisms in both multilevel and single-level settings, as well as sensitivity analysis and power analysis methods for causal mediation analysis. Substantively, she is interested in applying advanced statistical methods in developmental, educational, and health research.

Invited Talks

[Regular]

Feature Selection for Binary and Multi-Class Classification Problems

Ayman Alzaatreh [aalzaatreh@aus.edu]*; American University of Sharjah; AE

Feature selection has become a critical step in most data mining applications to mitigate the curse of dimensionality in high-dimensional datasets. Without direct input from the target variable, filter methods evaluate the importance of features as a pre-processing operation to the learning algorithm and select the best feature subsets through some information metrics. Filters are known to be more computationally efficient than wrapper and embedded methods. In this talk, a Bayesian approach namely, the relative belief ratio will be used as a filter method in binary and multi-class classification problems. The relative belief ratio is used as a filter method to rank features based on their importance in relation to a binary and multi-class target variables. Several benchmark data sets are used to demonstrate the applicability of the proposed method

---

[Speed]

Unintentional and intentional code-switching of Chinese-English bilinguals

Yanran Chen [ychen44@nd.edu]*; University of Notre Dame;

Code-switching (CS) is the alternate use of two or more varieties of language in a single conversation episode. While it can happen unintentionally, driven by the ease of production or environmental influence, bilinguals also leverage code-switching intentionally to achieve specific purposes, such as rhetorical effects, audience design, and the expression of emotions. This study categorizes the motivations of code-switching and summarizes its usages in two Chinese-English conversation datasets (ASCEND and SEAME). A key aspect of this study is to differentiate between unintentional and intentional CS with spontaneous speech corpora and examine the difference in linguistic structures of each type. The study will employ large language models to efficiently scale up annotation of the utterances. The artificial intelligence tools will be instructed to follow a comprehensive taxonomy of the CS motivations and perform hierarchical classification on a large quantity of examples using the taxonomy. The accuracy will be tested by the agreement of AI and human annotators. We anticipate two main findings: 1) CS is utilized both unintentionally, prompted by the ease of production, and intentionally, to serve specific purposes; and 2) CS utterances differing in intentionality exhibit significant structural differences, including the length of the switch, part of speech, and syntactic complexity. Our postulation is that intentional code-switching might be more fluent and occur at specific syntactic boundaries (intersentential or tag-switch). Conversely, unintentional code-switching may be more abrupt and occur mid-sentence (intrasentential switch).

[Regular]

Deep learning generalized structured component analysis: An interpretable artificial neural network model with composite indexes

Gyeongcheol Cho [cho.1240@osu.edu]*; Department of Psychology, The Ohio State University; US
Heungsun Hwang [heungsun.hwang@mcgill.ca]; McGill University; CA

Generalized structured component analysis (GSCA) is a multivariate method for specifying and examining interrelationships between observed variables and components. Despite its data-analytic flexibility honed over the decade, GSCA always defines every component as a linear function of observed variables, which can be less optimal when observed variables for a component are nonlinearly related, often reducing the component's predictive power. To address this issue, we combine deep learning and GSCA into a single framework to allow a component to be a nonlinear function of observed variables without specifying the exact functional form in advance. This new method, termed deep learning generalized structured component analysis (DL-GSCA), aims to maximize the predictive power of components while their directed or undirected network remains interpretable. Our real and simulated data analyses show that DL-GSCA produces components with greater predictive power than those from GSCA in the presence of nonlinear associations between observed variables per component.

[Speed]

Simple Procedure To Estimate A Structural Equation Model With Latent Variables

Zouhair El Hadri [z.elhadri@um5r.ac.ma]*; Mohammed V University, Faculty of sciences Rabat; MA

Structural Equation Modelling (SEM) is a sophisticated approach used to analyse complex models with both observed variables called Manifest Va- riables (MV) or indicators, and unobserved variables called Latent Variables (LV), Factors, Components or Constructs. The aim of the present work is to introduce a simple procedure to estimate the parametetrs associated with SEM models. The proposed procedure is the extension of the procedure introduced by El Hadri & al. in 2023 for Path Analysis models (Model without latent variables).

[Regular]

Piecewise Strategies for Fitting Differential Equation Models to Time Series Data

Yueqin Hu [yueqinhu@bnu.edu.cn]*; Beijing Normal University; CN

Differential equation models are continuous-time dynamic models suitable for analyzing intensive longitudinal data. Numerical optimization methods can be used to estimate parameters in differential equation models. However, as behavioral data often exhibit phase jumps, and longer time series are more prone to phase jumps, this method can lead to inaccurate estimation when applied to behavioral data, especially long sequences. Therefore, this study proposes segmenting the time series data to accurately estimate long sequences with phase problems. The method we propose allows each short segment of the long sequence to use different initial values, thus enabling more accurate capture of local dynamics, and then estimating overall fixed effects and local random effects accordingly using the multilevel version of the numerical optimization method. Data simulation conditions include time series length, signal-to-noise ratio, and phase jump conditions, and the segmentation algorithm considers smoothing methods, sliding window width, and step size. We will present the performance of different segmentation methods compared to the non-segmentation algorithms under various data conditions, and provide empirical demonstrations using intensive longitudinal data on mindfulness and psychological distress from two hundred and seventy university students at thirty-five consecutive time points.

[Speed]
Regularized Integrated Generalized Structured Component Analysis

Heungsun Hwang [heungsun.hwang@mcgill.ca]*; McGill University; CA
Gyeongcheol Cho [cho.1240@osu.edu]; Ohio State University;

Integrated generalized structured component analysis (IGSCA) has been developed for estimating structural equation models with both factors and components. In this study, we propose a regularized extension of IGSCA to address potential multicollinearity or perform automatic variable selection. This extension, termed Regularized IGSCA, aims to minimize a penalized least squares objective function that includes both L1 and L2 penalties. We present an empirical application to illustrate the usefulness of the method.

[Speed]
Predictive Analytics Technical and Application

Sandra Ihemeje [nkechisandy@gmail.com]*; ISDSA Nigeria; NG

This research project explores the technical aspects and applications of predictive analytics. Predictive analytics is a powerful tool that utilizes historical data, statistical techniques, and advanced algorithms to make accurate predictions and optimize operations. The study provides an overview of predictive analytics concepts and techniques, including regression analysis, decision trees, random forests, and machine learning algorithms.

The research investigates the applications of predictive analytics in various industries, such as finance, marketing, healthcare, and manufacturing. Case studies and examples highlight the use

of predictive analytics in financial markets, targeted marketing campaigns, healthcare diagnosis and treatment, and supply chain management.

The study examines the benefits of predictive analytics, including improved decision-making, increased efficiency, enhanced customer experience, risk mitigation, and competitive advantage. It also addresses the challenges and limitations of predictive analytics, as well as the ethical considerations associated with its use.

Furthermore, the research discusses future trends and directions in predictive analytics, such as advancements in machine learning and artificial intelligence, integration of big data, and the ethical and legal implications of using predictive analytics.

The findings of this research project provide valuable insights into the technical aspects and applications of predictive analytics. The study concludes with a summary of the findings and recommendations for future research, contributing to the existing body of knowledge and offering practical suggestions for organizations and researchers interested in harnessing the power of predictive analytics.

[Regular]
Instructing Language Models to Do Reasoning Wisely

Meng Jiang [mjiang2@nd.edu]*; University of Notre Dame; US

Reasoning in natural language is an amazing human ability. The NLP community has been collecting numerous data to train and test language models on various reasoning tasks such as mathematical reasoning, commonsense reasoning, abductive reasoning, and counterfactual thinking. The training was conventionally tuning the model parameters on input-output pairs, like a math word problem and an answer. Large language models, hitting the world with their emergent abilities of doing tasks in a chat mode, have changed the methodology of teaching machines to do reasoning. In this talk, I will start from the key techs behind the large language models, introducing why "instructing" models to do reasoning in a chat model, instead of "training", is quite effective and becomes a fashion. I'll then present a few studies very briefly where psychological knowledge and/or human learning skills have inspired algorithm design to instruct the large language models to do reasoning wisely. The studies vary from solving complex math word problems in English, answering questions in a different language, explaining a verification on statements, to answering questions under counterfactual presuppositions. These works are accepted to top NLP or AI venues in 2023-2024. One received the outstanding paper award in EMNLP 2023.

[Speed]
An estimation approach for time-varying effect models using cubic splines

Jingwei Li [jingweil@mailbox.sc.edu]*; University of South Carolina;

Traditional mediation analysis typically examines the relations among an intervention, a time-invariant mediator, and a time-invariant outcome variable. Although there may be a total effect of the intervention on the outcome, there isa need to understand the process by which the intervention affects the outcome. This indirect effect is frequently assumed to be time-invariant. With improvements in data collection technology, it is possible to obtain repeated assessments overtime resulting in intensive longitudinal data. This calls for an extension of traditional mediation analysis to incorporate time-varying variables as well as time-varying effects. In this paper, we focus on estimation and inference for the time-varying mediation model, which allows mediation effects to vary as a function of time via cubic spline interpolation. Two simulation models and a smoker's health research are studied to compare this method with local smoothing. More accurate results are obtained for the cubic spline interpolation when there are fewer time points.

[Speed]

Determining the Number of Factors in Exploratory Factor Analysis with Model Error

Yilin Li [yli49@nd.edu]*; University of Notre Dame; US
Guangjian Zhang [gzhang3@nd.edu]; University of Notre Dame; US

A key decision in exploratory factor analysis (EFA) is to determine the number of factors. Parallel Analysis (PA) and its variants are often recommended to aid this decision and their efficacy has been largely supported by simulation studies. The goal of the current study is to examine how PA and its variants perform in more realistic situations where EFA models fit approximately rather than perfectly. For comparison, we also consider a factor retention method that involves a model fit measure (Root Mean Square Error of Approximation, RMSEA) specifically designed to deal with model error. Our main findings include (1) PA is satisfactory when the factors are well-represented (high variable-to-factor ratios), but its performance becomes less satisfactory when the factors are not well-represented (low variable-to-factor ratios); (2) The RMSEA-based method is more satisfactory than PA under most conditions unless the sample size is very small; (3) The performance of the RMSEA-based method improves with larger samples, but the performance of PA and its variants do not improve with large samples.

[Regular]

Model Selection for Mixed-effects Models with Confidence Interval for LOO or WAIC Difference

Yue Liu [helena701@126.com]; Institute of Brain and Psychological Sciences, Sichuan Normal University; CN
Fan Fang [fangfan_ricca@163.com]; ; CN
Hongyun Liu [hyliu@bnu.edu.cn]*; Beijing Normal University; CN

LOO and WAIC become widely used for model selection in Bayesian statistics. Most studies select the model with the smallest value of them based on the point estimates, neither considering the difference of the fit indices between candidate models, nor the uncertainty of the estimates. Accordingly, we propose a sequential method comparing models based on confidence intervals for delta LOO or delta WAIC. A simulation study comparing this method and the point method in selecting mixed-effects location–scale models (MELSMs) is conducted.

Our study revealed that the sequential methods had higher accuracy rate of model selection than the point methods when the true model was simple, or had large magnitude of random intercept in the scale-model, or large sample size. For the most complex model, although the sequential methods preferred simper models, they performed as good as the point methods with larger sample size and more serious heterogeneity of residual variances. Moreover, the sequential methods, especially using 90%CI, tended to have higher power, lower Width95, and smaller SE than the point methods. Meanwhile, the differences between LOO and WAIC were obvious only when level-1 sample size was small, where LOO performed better when True model had either homogeneous residual variances, or serious heterogenous residual variance. Finally, as SE of delta LOO and delta WAIC used to construct their CI could be calculated by R package brms (or LOO) conveniently, we suggest apply the proposed methods in evaluation and selection for MELSMs in practice.

[Speed]
Revisiting Estimation in Hierarchical Modeling and Optimal Design

Laura Lu [zlu@uga.edu]*; University of Georgia; US

Multilevel models, also referred to as hierarchical linear models or mixed-effects models, are widely used for analyzing data with a nested or hierarchical structure. In such data, lower-level units, like students or individuals, are nested within higher-level units such as classes, schools, or organizations. Various estimation methods exist for fitting multilevel models to data, including Full Maximum Likelihood Estimation (FML), Restricted Maximum Likelihood Estimation (REML), and Bayesian Estimation methods. Additionally, exploratory estimation methods are also available, such as exploratory Ordinary Least Squares (OLS) estimation, which entails fitting separate regression models at each level of the hierarchy.

In this study, we revisit parameter estimation in Optimal Sample Allocation for multi-site randomized trials. In Optimal Sample Allocation research, the central objective is to determine the sample size or allocation of resources across different treatment groups or experimental conditions to maximize the study's efficiency and precision. The framework of multi-site randomized trials is hierarchical modeling. The dominant estimation method involves fitting an exploratory OLS regression model to each individual to derive lower-level intercepts and slopes, and these resulting intercepts and slopes are either averaged across higher-level units or regressed on higher-level predictors. We first revisit the parameter estimation in this framework and discuss the advantages and disadvantages of this method. We also propose other Optimal Sample Allocation methods by incorporating alternative estimation approaches. Results will be compared, and conclusions will be provided.

[Speed]

Using machine learning methods in the presence of numerous measured confounders in mediation analysis

Milica Miocevic [milica.miocevic@mcgill.ca]*; McGill University; CA

In certain fields, such as epidemiology and health research, there are numerous measured confounding variables (e.g., demographic information and medical history) that need to be included in the statistical model to avoid biasing the effects of interest. In statistical mediation analysis, researchers are usually focused on accurately estimating the indirect effect. Previous work has shown that in mediation analysis, accounting for confounders and pure predictors of the outcome allows for unbiased estimates of the indirect and direct effects, whereas adjusting for pure predictors of the independent variable and mediator can increase the standard errors of the indirect and direct effects, thus lowering the power to detect these effects (Diop et al., 2021). This project aims to examine if machine learning methods can be leveraged to select confounders of the paths constituting the indirect effect in mediation analysis from a large set of measured variables. A simulation study was conducted to examine if ridge regression, lasso, and elastic net successfully select confounders of the relationships between the independent variable and mediator (a-path), the mediator and the outcome (b-path), and the independent variable and the outcome (c'-path) in the following scenarios: (1) 40 measured variables, none of which act as either a covariate or a confounder for variables in the mediation model, (2) small effects of 4 confounders for all three paths in the mediation model and 36 unrelated measured variables, (3) large effects of 4 confounders for all three paths in the mediation model and 36 unrelated measured variables, (4) 4 pure covariates predicting the outcome variable and 36 unrelated measured variables, (5) a mixture of pure predictors and confounders of the a-path in addition to unrelated variables that are also included in the model, and (6) a mixture of pure predictors and confounders of the b-path in addition to unrelated variables that are also included in the model. The indirect effect was large in all conditions and the sample size was 200. The bias of the point estimates and the coverage of the 95% confidence intervals for the indirect effect were evaluated over 1,000 iterations. Preliminary findings indicate that lasso tended to have the highest accuracy out of the three procedures and ridge regression tended to yield the highest standard errors of the indirect effect except in scenario (4) when it was the most efficient method. Ridge regression also yielded intervals for the indirect effect with coverage below the nominal value in scenarios (3), (5), and (6), and all procedures had coverage below the nominal value in scenario (2). The presentation will discuss the pros and cons of each machine learning procedure for confounder selection across different scenarios.

[Speed]

Regularization Methods for Factor Models in the Presence of Partial Measurement Invariance

Emma Somer [emma.somer@mail.mcgill.ca]*; McGill University; CA

Carl Falk [carl.falk@mcgill.ca]; McGill University; CA
Milica Miocevic [milica.miocevic@mcgill.ca]; McGill University; CA

Measurement invariance testing is a prerequisite for meaningful group comparisons. When item parameters differ across groups, one way to correct for bias is using a partial measurement invariance model. Many of the approaches involve an iterative procedure, where items are tested for noninvariance one at a time, and subsequently, latent parameter estimates are interpreted. More recently, several approaches that simultaneously identify noninvariant items and perform latent parameter estimation have been proposed. In particular, regularization methods have been extended to the measurement invariance framework and involve imposing a penalty term that shrinks parameter differences across invariant items to zero in order to reduce model complexity and improve interpretation. Most of the previous literature has focused on detecting noninvariant items rather than obtaining accurate relationships between latent variables, which is often the main goal in psychological research. In the current study, we compared lasso, ridge, and elastic net for obtaining estimates of latent parameters using a simulation study. The aim of this research is to 1. evaluate whether machine learning methods can produce unbiased and efficient estimates of latent parameters and 2. examine whether there are differences across the three methods. We manipulated the number of indicators (4 and 8), sample size (N = 200, 500, 1000), magnitude of noninvariance (small, medium, and large), proportion of noninvariance (25%, 50%, and 75%), and the value of the correlation (0 and 0.3) between two latent variables and tested a range of 50 penalty values. Preliminary results suggest that methods generally perform similarly. However, when the number of indicators is lower and the proportion of noninvariance is higher, the lasso and elastic net outperform ridge in terms of bias and efficiency in some conditions.

[Regular]

Selection of Best Experience for Digital Platforms

Will Stamey [wstamey@nd.edu]*; University of Notre Dame; US
Ken Kelley [kkelley@nd.edu]; University of Notre Dame; US
Bhargab Chattopadhyay [bhargab@iiitvadodara.ac.in]; Indian Institute of Technology – Vadodara; IN
T. Bandyopadhyay [tathagata@iima.ac.in]; ; IN

Sequential methods can estimate the number of samples required for statistical inference while the sampling takes place. Digital experiments are often concerned with comparing website designs, algorithms or features and determining most effective for a business process. Addressing the needs of this task, we propose a sequential approach for comparing multiple treatments against a baseline simultaneously and we demonstrate its ability to control Type I and Type II errors according to user preference using a Monte Carlo simulation study.

[Speed]

A Comparison of Minimal-Effect Testing, Equivalence Testing, and the Conventional Null Hypothesis Testing for the Analysis of Bi-factor

Jiashan Tang [tangjs@njupt.edu.cn]*; Nanjing university of Posts and Telecommunications; CN
Shunji Wang [WongShunchi@163.com]; Nanjing university of Posts and Telecommunications; CN
Ke-Hai Yuan [kyuan@nd.edu]; University of Notre Dame; US

A necessary step in applying bi-factor models is to evaluate the need for domain factors with a general factor in place. The conventional null hypothesis testing (NHT) was commonly used for such a purpose. However, the conventional NHT meets challenges when the domain loadings are weak or the sample size is insufficient. This article proposes using minimal-effect testing (MET) and equivalence testing (ET) to analyze bi-factor models. A key element in conducting ET and MET is the minimal size of factor loadings that can be regarded as noteworthy in practice, termed as minimal noteworthy size. This article presents two approaches to formulating the minimal noteworthy size and compares the pros and cons of MET, ET, and the conventional NHT. Analysis shows that MET, ET, and the conventional NHT are complementary. Combining them to test the noteworthiness of domain loadings can help researchers make a comprehensive judgment. Simulated and real datasets illustratethe applications of the three methods. Statistical power of the methods is also discussed.

[Regular]

Revisiting Bayesian Two Sample Inference

Xin Tong [xt8b@virginia.edu]*; University of Virginia; US
Sarah Depaoli [sdepaoli@ucmerced.edu]; UC Merced;

Two sample inference is a fundamental problem in statistics. Hypothesis testing in the form of two sample tests like Student's t test are among the most popular statistical procedures conducted across all scientific domains and is particularly important for experimental scientists. With a comparison between treatments and the control group, researchers can determine causal relationships between variables. It is well recognized in the literature that traditional two sample comparison methods are often inadequate for analyzing modern datasets where underlying distributions could be multivariate or nonnormal, and the differences across the distributions could be locally concentrated. Tools from the Bayesian perspective are gaining popularity as they can flexibly describe different shapes of distributions. However, our recent study showed that Bayesian two sample inference provided inconsistent results when different test statistics were used even for very basic mean and variance comparisons. Therefore, in this study, Bayesian two sample inference will be revisited. We systematically evaluate the performance and the impact of different test statistic using simulation studies. We will further provide a guideline on the use of test statistics in the Bayesian framework and discuss the implications of this inconsistency problem in more complex models and analyses.

[Speed]
Power analysis for cohort sequential designs

Lijuan Wang [lwang4@nd.edu]*; University of Notre Dame; US

Cohort sequential designs allow researchers to study developmental trajectories over a longer time interval by collecting and analyzing data from different age cohorts during a shorter interval of time. In this study, we examine how study design factors including the number of cohorts, the degree of age overlapping among cohorts, and the number of measurement occasions per cohort influence estimating and testing developmental trajectories over time using latent growth curve modeling. Particularly, we investigate how these design factors influence the statistical power of convergence tests, tests of fixed effects (e.g., average change and group difference in average changes), and tests of relations in changes in cohort sequential designs.

[Speed]
Psychometric AI: Integrating Modern Data Science in Psychological Testing and Assessment

Wei Wang [wwang@gc.cuny.edu]*; The Graduate Center, City University of New York; US
Chapman Lindgren [clindgren@gc.cuny.edu]; US
Max Lobel [max.lobel.2000@gmail.com]; US
Kemar Pickering [kpickering@gradcenter.cuny.edu]; US

The current study explored a burgeoning field—Psychometric AI, which integrates modern data science techniques and psychological testing and assessment to not only improve measurement accuracy, efficiency, and effectiveness, but also effectively reduce human bias and increase objectivity in measurement. By leveraging unobtrusive eye-tracking sensing techniques and performing 1,470 runs with 7 different machine-learning classifiers, We systematically examined the efficacy of various (ML) models in measuring different facets and measures of the emotional intelligence (EI) construct. Our results revealed an average accuracy ranging from 50–90%, largely depending on the percentile to dichotomize the EI scores.

[Regular]
Study on the influence of family on adolescents' mental health

Yong Wen [ywen1108@njupt.edu.cn]; College of Science, Nanjing University of Posts and Telecommunications; CN
Yingying Dai [1221087212@njupt.edu.cn]; College of Science, Nanjing University of Posts and Telecommunications; CN
Jiashan Tang [tangjs@njupt.edu.cn]*; College of Science, Nanjing University of Posts and Telecommunications; CN

The mental health problems of teenagers are becoming increasingly prominent. The family is the environment in which individuals grow up after birth, which is generally composed of children and their parents. Teenagers are directly and profoundly influenced by the family during their growth. From the perspective of family, this paper classifies the integration of factors at the family level into five factors: family socioeconomic status, parental companionship, family conflict, family expectation and family structure. Gender, academic achievement and mental health are selected as relevant factors for adolescents. The data of China Family Panel Studies in 2018 are selected, and then a structural equation model is constructed to analyze the impact of family factors on adolescents' mental health. The gender differences in structural pathways are compared by multi-group analysis, and the moderating effects of gender on family factors influencing mental health pathways are analyzed. Multiple linear regression analysis is used to study the effects of family structure and sibling structure on adolescents' mental health.

[Speed]

Bioinformatic Analysis of Immune-related LncRNA in Head and Neck Squamous Cell Carcinoma

Jiawen Wu [wujiaw96@163.com]; College of Science, Nanjing University of Posts and Telecommunications; CN
Jiashan Tang [tangjs@njupt.edu.cn]*; College of Science, Nanjing University of Posts and Telecommunications;

Recent studies have shown that many immune-related long non-coding RNA (lncRNA) has unique advantages as a novel biomarker in cancer diagnosis, treatment and prognosis. The study of immune-related lncRNAs in head and neck squamous carcinoma (HNSCC) is of great importance. Based on gene expression data and clinical data of HNSCC patients, immune-related lncRNAs with prognostic value were identified using univariate Cox regression analysis, Lasso regression analysis and multivariate Cox analysis to construct a prognostic risk score model. And the performance of the prognostic model was also evaluated using Kaplan-Meier analysis and time-dependent ROC curves. The results were screened to obtain 7 key lncRNAs and constructed survival prognostic models. Kaplan-Meier analysis revealed that the survival rate of the high-risk group was significantly lower than that of the low-risk group, while the ROC curve proved that the model has good predictive ability. The results of the study showed that the survival prognostic model constructed based on seven immune-related lncRNAs could provide effective survival prognosis prediction for HNSCC patients.

[Speed]

How are you really feeling? A dynamic network approach to detecting nomothetic patterns of emotion regulation ability.

Austin Wyman [awyman@nd.edu]*; University of Notre Dame; US

Emotion detection AI is emerging as a promising alternative to self-report measures in intensive

longitudinal data designs, such as ecological momentary assessment (EMA). For example, AI is able to retrieve emotion estimates at much faster intervals than traditional self-report measures and without the cognitive interference from reading and answering, which are present in all psychometric instruments. However, most studies involving AI are only able to interpret large numbers of emotion estimates idiographically, and there lacks a clear framework for how to interpret them nomothetically. Thus, we propose a novel framework for transforming emotion detection AI estimates in order to maximize their utility in EMA research. Our approach uses group iterative multiple model estimation (GIMME) to identify lagged emotion subgroups and estimate the closeness of fit for each individual network. Implications in both confirmatory and exploratory studies are discussed.

[Regular]

Scale-Invariance, Equivariance and Dependency of Structural Equation Models

Ke-Hai Yuan [kyuan@nd.edu]*; University of Notre Dame, USA; US
Ling Ling [lingl@njupt.edu.cn]; Nanjing University of Posts and Telecommunications;
Zhiyong Zhang [ZhiyongZhang@nd.edu]; University of Notre Dame; US

Data in social and behavioral sciences typically contain measurement errors and do not have predefined metrics. Structural equation modeling (SEM) is widely used for the analysis of such data, where the scales of the manifest and latent variables are often subjective. This article studies how the model, parameter estimates, their standard errors (SEs), and the corresponding z-statistics are affected by the scales of the manifest and latent variables. Analytical and empirical results show that (1) the normal-distribution-based likelihood ratio statistic is scale-invariant with respect to scale changes of manifest and latent variables as well as to anchor change of latent variables; (2) the normal-distribution-based maximum likelihood (NML) parameter estimates are scale-equivariant with respect to scale-change of manifest and latent variables as well as to anchor change of latent variables; (3) standard errors (SEs) following the NML method are parallel-scale-equivariant with respect to scale changes of the manifest and latent variables; and (4) the z-statistics are scale-invariant with respect to scale changes of the manifest and latent variables. However, only (1) and (2) hold if latent variables are rescaled by changing anchors. Nevertheless, parameters that are not directly related to latent variables with changing anchors are still scale-equivariant and their z-statistics are still scale-invariant. The results are expected to advance understanding of the output of SEM analysis, and also facilitate result interpretation and comparison across studies as in meta analysis.

[Regular]

Bayesian Growth Curve Modeling with Measurement Error in Time

Lijin Zhang [lijinzhang@stanford.edu]*; Stanford University; US
Wen Qu [wqu@fudan.edu.cn]; Fudan University; CN
Zhiyong Zhang [zhiyongzhang@nd.edu]; University of Notre Dame; US

Growth curving modeling has been widely used in many disciplines to understand the trajectories of growth. Two popular forms utilized in the real-world analyses are the linear and quadratic growth curve models. These models operate on the assumption that measurements are conducted exactly at pre-set time or intervals. In essence, the reliability of these models is deeply tied to the punctuality and consistency of the data collection process. However, in real-world data collection, this assumption is often violated. Deviations from the ideal measurement schedule often emerge, resulting in measurement error in time and consequent biased responses. Our simulation findings indicate that such error can skew estimations, especially in quadratic GCM. To account for the measurement error in time, we introduce a Bayesian growth curve model to accommodate the error in the individual time values. We demonstrate the performance of the proposed approach through simulation studies. Furthermore, to illustrate its application in practice, we provide a real-data example, underscoring the practical benefits of the proposed model.

[Regular]
Introduction to an online app for SEM analysis with text data

Zhiyong Zhang [zzhang4@nd.edu]*; University of Notre Dame Notre Dame, IN 46556 USA; US

Text data are widely collected in research and can come from many different sources. However, text data are largely under-analyzed in social, behavioral and education research. Supported by the Institute of Education Sciences (R305D210023), we have developed a general model that can combine structural equation models with text data in which a two-stage method was used to first extract the information from the text data through different methods such as sentiment analysis and text encoders and then the information was used in SEM. We also developed an online app - BigSEM, that can be used to conduct SEM analysis with text data. In this talk, I will provide a tutorial on how to use the online app through examples.

[Speed] –
Implementation of a Data Aggregation ETL Pipeline and Business Intelligence System

Emmanuel Elom [engrloyal@gmail.com]*; N/A; US

The goal of this project is to create a system that allows data from many sources, such as databases, APIs, files, webpages, and other sources, to be seamlessly combined and translated into a common format before being loaded into a database or files for analytics using the business intelligence (BI) system which will enable the data analysis and information collecting for business decision-making processes.

This implementation will involve several stages which involves; (1) Requirement Gathering, (2) Data Collections, (3) Data Processing, (4) Data Loading, (5) Business Intelligence Analytics.

In conclusion, the establishment of a business intelligence system and data aggregation ETL pipeline will enhance decision-making and provide useful insights into corporate operations. Businesses may improve their operations and make better judgments by combining data from many sources and making it accessible for analysis.

[Speed] –
Unsecured Debt and Psychological Well-Being Among Older Americans: The Role of Financial and Health Literacy

Mary Akinde [maryakinde@uga.edu]*; University of Georgia; US

This study investigated the impact of personal debt on psychological well-being among 9961 adults aged 51 and above, with a specific focus on unsecured debt. The research explored the role of financial literacy and health literacy in this relationship. Data from the 2020 Health and Retirement Study were utilized, involving respondents who completed both a core survey and a psychosocial leave-behind questionnaire. Psychological well-being was assessed using the reversed 8-item Center for Epidemiologic Studies Depression Scale.

The analysis revealed that 32% of respondents have unsecured debt, and the presence of such debt significantly predicted lower psychological well-being. Health literacy was a significant predictor of higher psychological well-being, while financial literacy did not exhibit a similar impact. Among older Americans, unsecured debt was identified as a contributing factor to reduced psychological well-being.

Notably, the deleterious effects of unsecured debt on mental health were found to be mediated by health literacy. Given that medical debt constitutes a sizable proportion of unsecured debt for older individuals, possessing greater knowledge of one's health situation appears to mitigate the psychological distress associated with debt. Interventions aimed at enhancing the ability of older adults to access and comprehend health-related information and services may serve as a protective measure against the negative psychological consequences of grappling with unsecured debt.

[Speed] –
The Use of Psychedelics for Depression: A Systematic Review and Metanalysis on Ketamine and Psilocybin Treatment.

Michel Monal [mmonal@mru.edu]*; Miami Regional University; US
Sergio Torralbas Fitz [torralbasfitz@gmail.com]; Miller School of Medicine, University of Miami; US

José Sigarreta Almira [josemariasigarretaalmira@hotmail.com]; Autonomous University of Guerrero, Faculty of Mathematics; MX
Jose Rodriguez Garcia [jomaro@math.uc3m.es]; Carlos III University of Madrid, Spain; ES

Background and Aims: The re-emergence of psychedelics in the field of psychiatric treatment, particularly for depression, has increasingly gained the attention of the scientific community due to the large number of people suffering from depression and relatively low treatment effectiveness. This study conducted a systematic review and meta-analysis of articles examining the impact of ketamine and psilocybin on depressive symptoms. **Method:** We conducted a comprehensive literature search, following the guidelines for reporting systematic reviews and meta-analyses (PRISMA). Our search targeted academic databases, including EBSCO, PLoS One, JSTOR, PubMed, and Psych Info, between March 2013 and June 2023. We targeted randomized clinical trials that used Ketamine and psilocybin to treat depressive symptoms in oncologic patients. We carefully selected the relevant data and conducted a meta-analysis using STATA software (Version 14.0., Stata Corporation, and College Station, Texas, USA). The statistical software used was STATA version 14.0. We utilized a random effect model to perform the meta-analysis. The Standardized Mean Difference (SMD) was the effect measure, with a 95% confidence interval. **Results:** The selected studies indicated a pooled effect size (ES) of 0.79 (95% confidence interval: 0.70 to 0.88), demonstrating ketamine's and psilocybin's significant and rapid effect in reducing depressive symptoms. This effect persisted over the week following treatment. There was moderate heterogeneity among the studies (I^2 = 59.08%), but the overall evidence supports ketamine's and psilocybin's efficacy in alleviating depressive symptoms. **Conclusions:** Ketamine and psilocybin have been shown to reduce depressive symptoms after administration, with significant effectiveness. These findings highlight ketamine's and psilocybin's great potential as a treatment for depression. Further research into ketamine's and psilocybin's short and long-term impact on depression is recommended, as well as its biochemical interaction and long-term safety and effectiveness.

[Speed] –

Survival analysis of recurrent events: An application to viral rebound among HIV/AIDS patients in Namibia

Dibaba Gemechu [diboobayu@gmail.com]*; Namibia University of Science and Technology (NUST); NA
Pia Abraham [pabraham@nsa.org]; Namibia University of Science and Technology (NUST); NA

Adherence to antiretroviral therapy (ART) for Human Immunodeficiency Virus (HIV) patients is key to maintaining viral suppression and preventing sexual transmission of HIV. However, some patients who obtained viral suppression are unable to maintain an undetectable viral load and experience viral rebound and hence, needs to be investigated. Several studies employed the standard survival analysis methods, focusing solely on the time it takes for the first event to

occur. However, this approach overlooks the fact that individuals may go through multiple events, resulting in valuable information being overlooked or discarded. The main objective of this study is, therefore, to investigate factors associated with recurrent events of viral rebound among HIV patients in Namibia by fitting various extensions of survival models such as Andersen-Gill, Prentice-Williams-Peterson total time (PWP-TT) and Prentice-Williams-Peterson gap time (PWP-GT). The results showed the PWP-TT model outperformed the other recurrent models and furthermore, among the factors included in the model, factors such as age, number of years in ART and the baseline CD4 count was found to be a significant factor associated with viral rebound of HIV patients.

[Regular] –
Loss Aversion Distribution: The Science Behind Loss Aversion Exhibited by Sellers of Perishable Good

Daniel Koh [daniel.koh@danthescientist.com]*; Koh & Associates; JP

This study introduces the loss aversion distribution, a novel approach to analyzing consumer behavior toward perishable goods, diverging from traditional exponential models by incorporating a non-memoryless characteristic. It captures the dynamic nature of consumer loss aversion from manufacture to expiry, highlighting an initial muted response that intensifies mid-lifecycle and diminishes as expiry nears. Utilizing derivative analysis of the probability density function, the research establishes the distribution's monotonicity, boundedness within [0,1], and non-negativity. An empirical finding of note is the b-parameter's approximation to a Gaussian distribution, offering insights into consumer psychology. This framework not only advances understanding of loss aversion in perishable goods but also opens avenues for further psychometric exploration.

[Speed] –
Real-Time Fraud Detection Using Machine Learning

Benjamin Borketey [bbortey9@gmail.com]*; The University of Akron; US

Fraud detection remains a formidable challenge in various fields, particularly in the context of credit card transactions, where the prevalence of illicit activities continues to grow with advancing technology. This study addresses the critical need for effective fraud detection methodologies, focusing on real-time applications to mitigate financial losses and safeguard consumer interests. Leveraging a credit card dataset from Kaggle, I address the class imbalance inherent in fraud datasets using Synthetic Minority Oversampling Techniques (SMOTE) to enhance modeling efficiency. I use several machine learning algorithms to classify transactions as fraud or genuine, including Logistic Regression, Linear Discriminant Analysis, K-nearest Neighbors, Classification and Regression Tree, Naive Bayes, Support Vector Machine, Random

Forest, XGBoost, and Light Gradient-Boosting Machine. Through rigorous evaluation metrics such as AUC, PRAUC, F1, KS, Recall, and Precision, the Random Forest model emerges as the best model with superior performance in detecting fraudulent activities. Furthermore, the Random Forest model predicted well by identifying about 92% of transactions scoring 90 and above as fraudulent, equating to a detection rate of approximately 70.77% for all fraudulent transactions in the test dataset. Overall, the Random Forest model captures more than half of the fraud (resulting in a detection rate of 50%) for each bin in the test dataset. I use SHAP Values for model explainability. The SHAP summary plot highlights the global importance of individual features, with 'V12' being the most influential and 'V14' closely following. Additionally, SHAP force plots provided local interpretability, revealing how specific features impacted predictions for individual instances.

[Regular] –
Power analysis for count data, application to over-dispersed data and limited sample sizes.

Oleksandr Ocheredko [Ocheredko@yahoo.com]*; Vinnytsya National Medical University; UA

Count data are ubiquitous, with data sampling and data analysis quite refine. Still there are points of discussions and openings in power analysis, in particular to accommodate flexible designs and model structures.

In this paper I deliver some aspects of post-hoc power analysis for over-dispersed count data, as well as introduce a new role of power analysis in testing count data based hypotheses under limited sample size.

[Speed] –
Big Data in the Boardroom: A Literature Review of Its Strategic Impact on Business Decision-Making.

Chidimma Ogbonna [ogbonnachidimma@outlook.com]*; University of Texas Rio Grande Valley; US

The era of big data has transformed strategic decision-making within the business sector, prompting a shift from intuition-based to analytics-driven approaches. This literature review synthesizes current research on the influence of big data on business strategy, emphasizing enhanced decision-making accuracy, real-time operational agility, predictive analytics for proactive strategizing, and the tailoring of customer experiences through personalization. Advanced analytic techniques applied to vast data sets provide a competitive edge by uncovering hidden patterns in customer behavior, market dynamics, and operational efficiency. The integration of predictive analytics allows businesses to anticipate market trends and consumer needs, fostering a proactive rather than reactive business model. Additionally, the capability to personalize products and services on an unprecedented scale signifies a departure from

homogeneous marketing strategies to ones that are highly individualized. However, the adoption of big data analytics is not without challenges; it necessitates substantial infrastructural investments, a cultural shift towards data literacy, and the careful navigation of ethical considerations regarding privacy and data stewardship. Ethical considerations are scrutinized in light of regulations such as the GDPR (2016) and CCPA (2018), with insights from Barocas & Selbst (2016) on the potential for data's disparate impact. The review concludes that big data analytics is indispensable for businesses seeking to remain competitive in a rapidly evolving marketplace, necessitating a nuanced approach to its technical, human, and ethical dimensions for firms to maintain a competitive edge, ensure responsible stewardship, enhance decision-making processes, enable agility, and create personalized customer experiences.