Workshop on Statistical Learning for Process Data (log files)

Date: July 25, 2020

Online via Zoom

Registration

The course is free of charge. All are welcome! Registration is required at http://www.scientifichpc.com/processdata/workshop.html. We will email Zoom meeting link and detailed course materials and instructions to the email address provided in the registration form shortly before the workshop.

Abstract:

Process data refers to log files generated by human-computer interactive items. They contain detailed keystrokes and mouse clicks as well as their timestamps. Our research shows that process data contain substantially more information than classic item responses. They bring great opportunities and at the same time great challenges to the psychometrics community. In this course, we summarize our research methods and empirical findings and provide hands-on training for these methods via real and simulated data. We first provide an overview that features extracted from process data do contain more information than classic item responses. In addition, we also address the question how process data helps solving psychometric problems. To do so, we provide several applications of process data analysis to specific psychometric problems: improving test reliability, reducing/removing differential item functioning, and improving career planning.

We developed an R package ProcData, an open source package for exploratory process data analysis. In this course, participants will be provided with hands-on training of process data analysis via ProcData. Intended audience are expected to have basic knowledge of latent variable modeling and familiarity with R/RStudio and are interested in learning or applying data-driven methods to log data analysis. Running ProcData package requires installation of R, Rcpp, and Anaconda (installation instructions/help will be provided). Participants are expected to bring their own laptop with Windows or Mac operating system to fully engage in the activity.

This course contains three main sections: 1. an overview of process data analyses including an introduction of the methodologic development and key empirical results, 2. an introduction to ProcData and hands-on practice, 3. specific applications and hands-on practice of process data to psychometric problems: improving test reliability, reducing/removing differential item functioning, prediction of career development and satisfaction.

Agenda:

11:00 am — 01:00 pm: Overview of process data analysis by Jingchen Liu

01:30 pm — 02:45 pm: Introduction to ProcData: feature extraction methods, recurrent neural network based models by Xueying Tang

03:00 pm — 04:15 pm: Partial score and differential item functioning via process data by Susu Zhang

Syllabus

Overview of process data analysis

Recent advances in informational technology has led to the increasing popularity of computer-based interactive items, which require test-takers to complete specific tasks within a simulated environment. In addition to the final outcomes, the entire log of interactions between the test-taker and the item, i.e., the sequence of actions and their timestamps, are recorded as process data. Process data contain rich information about test-takers' problem-solving processes that are not recoverable from the final responses. In this overview, we summarize our main research developments. It includes feature extraction via multidimensional scaling and neural-network-based autoencoder. An important question is how process data can assist specific psychometric research. To address this problem, we present two applications: improving test reliability by constructing a process-data-based partial score system and removing/reducing differential item functioning by including process data in the scoring rules.

Introduction to ProcData

We introduce ProcData, an R package we design for processing, examining, and analyzing process data. The list of topics includes

1. Installation of ProcData
2. Class proc and its print and summary methods
3. Climate control dataset cc_data
4. Reading and writing response processes via read.seqs and write.seqs
5. Functions for data processing
6. MDS feature extraction via seq2feature_mds
7. Autoencoder feature extraction via seq2feature_seq2seq

We will demonstrate the features of ProcData through a live R session.

Partial score and differential item functioning via process data

We provide two specific applications of process data analysis to psychometric problems. These two examples illustrate how to make use of the additional information in process data and to what extend they add values to the existing literature.

Accurate assessment of students' ability is the key task of testing. Traditional assessments are based on the item final outcomes (correct/incorrect). While problem-solving processes contain additional information about a student's proficiency on the measured traits, we establish a framework to systematically construct a process-data-based scoring system that is substantially more accurate than the traditional IRT-model-based assessment in terms of reliability.

Differential item functioning (DIF) is critical for item validity. Various methods have been developed to identify DIF. However, few results are available to reduce or remove DIF. We develop a framework that identifies and further constructing a scoring rule to reduce/remove DIF. This new scoring rule is based on an individualized score adjustment based on process data.

In this section, we provide a step-by-step instruction of these two methods through simulated data.

Presenter Information:

Jingchen Liu, Columbia University

jcliu@stat.columbia.edu

Dr. Jingchen Liu is Professor of Statistics at Columbia University. He holds a Ph.D. in Statistics from Harvard University. He is the recipient of 2018 Early Career Award given by the Psychometric Society, 2013 Tweedie New Researcher Award given by the Institute of Mathematical Statistics, and a recipient of the 2009 Best Publication in Applied Probability Award given by the INFORMS Applied Probability Society. He has research interests in statistics, psychometrics, applied probability, and Monte Carlo methods. He is currently an associate editor of Psychometrika, British Journal of Mathematical and Statistical Psychology, Journal of Applied Probability/Advances in Applied Probability, Extremes, Operations Research Letters, and STAT.

Xueying Tang, University of Arizona

xytang@math.arizona.edu

Dr. Xueying Tang is an Assistant Professor in Statistics in the Department of Mathematics at the University of Arizona. Prior to joining the University of Arizona, she was a postdoctoral research scientist at Columbia University in the Department of Statistics. Her research interests include high dimensional Bayesian statistics, latent variable models and their applications in education and psychology. She has worked extensively on data-driven methods for the analysis of process data from educational assessments and is one of the developers of the ProcData R package for exploratory analysis of log data.

Susu Zhang, University of Illinois at Urbana-Champaign

szhan105@illinois.edu

Dr. Susu Zhang is a postdoctoral researcher in the Statistics Department at Columbia University. In Fall 2020, she will join the Departments of Psychology and Statistics at the University of Illinois at Urbana-Champaign. Her research interests include latent variable modeling, the analysis of complex data (e.g., log data) in computer-based educational and psychological assessments, and longitudinal models for learning and interventions.