Contribution ID: **161** Type: **Symposium**

Statistical Learning Approaches to Psychometric Modeling Challenges

Symposium title

Statistical Learning Approaches to Psychometric Modeling Challenges

Coordinator

Dylan Molenaar

Affiliation

University of Amsterdam

Abstract

Due to the rapid recent developments in the fields of artificial intelligence, many interesting new statistical modeling tools have become available. In this symposium, we explore the usefulness of these modeling tools for the field of psychometrics. That is, we focus on various psychometric challenges, and propose solutions that combine models and algorithms from the fields of statistical learning, machine learning, and deep learning with models and algorithms from the field of psychometrics. The results are interesting new approaches that can be used in psychometric practice, and that can -in turn- aid in increasing the interpretability of the black-box models used in the field of artificial intelligence.

The outline of this symposium is as follows: In the first talk, Molenaar focusses on modeling challenges with respect to the socalled Laten Space Item Response Theory models and propose an approach based on regularization from the statistical learning literature to address this challenge. Next, Federiakin will address challenges with respect to the factor analysis of process data using recurrent neural networks from the deep learning literature. Then Veldkamp will focus on facilitating the estimation of (complex) latent class models to (big) data structures by using autoencoders from the deep learning literature. Finally, Debelak will focus on identifying problematic model behavior underlying automated essay scoring using approaches from psychometrics and machine learning.

Keywords

Psychometrics, statistical learning

Number of communicatios

4

Communication 1

Regularized Estimation of the Latent Space Item Response Theory Model

Authors

Dylan Molenaar

Affiliation

University of Amsterdam, The Netherlands

Abstract

In latent space item response theory (IRT) modelling, both subjects and items are positioned in R dimensional Euclidian latent space. This framework allows for detailed modelling of local dependences among items and subjects, which are assumed to be absent in conventional IRT models. Latent space IRT has demonstrated its value in diverse fields, including intelligence assessment (Kang & Jeon, 2025; Kim et al., 2014), developmental psychology (Go et al., 2022), mental health (Jeon & Schweinberger, 2024), social influence (Park et al., 2023), national school policy evaluation (Jin et al., 2022), and student monitoring (Jeon et al., 2021). However, its broader application is limited by the computational challenges posed by the Bayesian algorithms commonly used for model estimation.

Therefore, in this presentation, a novel estimation procedure is proposed based on regularized joint maximum likelihood estimation. This approach significantly reduces computational demands making it feasible to conduct more robust model evaluations using K-fold cross-validation. The advantages of this method are illustrated in a simulation study and a real data analysis.

Keywords

item response theory, regularization, cross-validation

Communication 2

"Factor analysis"of Process Data via Psychology-Informed Variational Recurrent Autoencoders for the Analysis of Critical Online Reasoning

Authors

Denis Federiakin (1,2), Olga Zlatkin-Troitschanskaia (1), Lidia Dobria (3)

Affiliation

(1) Department of Business and Economics Education, University of Mainz; (2) Department of Psychology, Frankfurt University, Germany; (3) Department of Mathematics, Wilbur Wright College, Chicago, IL.

Abstract

The cornerstone of psychometrics –factor analytical methods –is designed for the interpretable dimensional reduction of response accuracy vector data. This approach can be likened to Variational AutoEncoders (VAEs) with shallow decoders (Urban & Bauer, 2021). However, it is not suitable for analyzing raw process data due to its inability to account for autoregressive dependencies within sequential data. To address such dependencies, various Recurrent Neural Network (RNN) architectures have been proposed, including Variational Recurrent AutoEncoders (VRAEs; Fabius & Van Amersfoort, 2014). This type of RNN creates vector representations of sequential data and reconstructs sequences while preserving autoregressive dependencies.

In this presentation, we propose two custom, interpretation-based recurrent units –one for encoding and one for decoding sequences –tailored for analyzing behavioral data. Both units utilize gating mechanisms to mitigate the vanishing and exploding gradient problem (Hochreiter et al., 2001) while preserving interpretability. The Recurrent Encoding Behavioral Unit (REBU) is inspired by the Long Short-Term Memory unit (Hochreiter & Schmidhuber, 1997), whereas the Recurrent Decoding Behavioral Unit (RDBU) is developed from contemporary psychological theories on Person-Situation Interactions (Furr & Funder, 2018). The RDBU accounts for situational strength (environmental cues regarding the desirability of potential behaviors) and situational affordances (contextual features enabling the expression of specific traits), resulting in an interpretable decoder structure similar to the VAE-based approach to factor analysis.

The architecture consists of two information channels: Long-Term Memory (LTM) and Short-Term Memory (STM). The LTM channel stores information about the vector representation throughout the entire sequence, while the STM channel is responsible for capturing first-order autoregressive dependencies. In RDBU, LTM

satisfies the principle of "factor scores"by remaining constant throughout the sequence. This ensures that the learned latent representations are independent of the reconstruction process.
For our analysis, we used log data from 315 higher education students who participated in the Critical Online Reasoning assessment (Molerov et al., 2020). In this scenario-based assessment, students were presented with a problem that lacked a clearly definitive correct answer and were instructed to search online for relevant information. Over the course of 20 minutes, they conducted a brief Internet search and compiled a short essay based on the arguments they discovered. During this process, their search history and actions were tracked. In our analysis, websites are treated as situational contexts, and clickstream data as actions. The results suggest that it is possible to both generate and interpret vector representations of students' action sequences with acceptable model quality metrics (ROUGE-L and BLEU scores of approximately 0.7).
We also discuss common challenges associated with VRAEs (and their potential solutions). These challenges include longer training times compared to vector-to-vector VAEs; the rare token problem (Yu et al., 2021) which can be addressed through the introduction of "unknown"tokens and balanced reconstruction loss; and posterior collapse, which can be mitigated using input and output STM dropout (Gal & Ghahramani, 2016) combined with KLD annealing (Bowman et al., 2015). Finally, we outline directions for future research.

Keywords

Recurrent Neural Networks, Process Data

Communication 3

Variational Autoencoders for Models with Latent Classes

Authors

Karel Veldkamp, Raoul Grasman, Dylan Molenaar

Affiliation

University of Amsterdam, The Netherlands

Abstract

Amortized variational inference (AVI) has recently been proposed in the field of Item response theory as a computationally efficient alternative to marginal maximum likelihood estimation (MML). The current study investigates if the computational advantages of AVI for large, high dimensional data carry over to discrete latent variable models. We adapt three techniques from the machine learning literature to the estimation of discrete latent variable models. In separate simulations, we compare the different approaches for latent class analysis, cognitive diagnostic models and mixture IRT models respectively. Results show that AVI is much faster than MML for mixture IRT models. AVI is also slightly faster than MML for LCA models with a large number of classes and items, and is less likely to end up in local minima. Overall we conclude that AVI provides accurate parameter estimates for all three models discussed, but that the computational advantages are most significant for models that have a mixture of discrete and continuous latent variables, such as mixture IRT.

Keywords

Autoencoders, latent classes

Communication 4

Automated Essay Scoring Using Generative Artificial Intelligence: Illustration of a Systematic Evaluation Framework

Authors

Rudolf Debelak (1), Laura Stahlhut (2), Kyle Matoba (1)

Affiliation

(1) University of Zurich, EPFL; (2) Institut für Bildungsevaluation Zürich AG

Abstract

Automated essay scoring systems can support teachers by providing rapid, cost-effective verbal and numerical feedback on student writing. In recent years, these systems have improved significantly with the rise of generative artificial intelligence models based on the transformer architecture. Research consistently shows that these models outperform traditional machine learning approaches across a wide range of natural language processing tasks, including essay scoring.

Despite these advancements, the application of this technology in psychology and education presents several risks, including: a) biased, inconsistent, or inappropriate verbal feedback, b) numerical scores that are highly arbitrary or systematically deviate from human ratings, and c) potential discrimination against specific groups in scoring and feedback.

In this talk, we illustrate these risks using empirical findings from an ongoing pilot study on automated essay scoring in Switzerland, drawing on examples from several widely used generative models. We then introduce a framework for evaluating the reliability, validity, and fairness of automated essay scoring systems. This framework integrates psychometric principles, such as item response theory and probabilistic test theory, with benchmarking standards from computer science to systematically identify problematic model behaviour. We demonstrate the framework's practical application using anonymized data from our pilot study and summarize main takeaways and challenges.

Keywords

Item scoring, item response theory

Primary authors: FEDERIAKIN, Denis (Johannes Gutenberg University Mainz); MOLENAAR, Dylan (University of Amsterdam); VELDKAMP, Karel (Universiteit van Amsterdam); DEBELAK, Rudolf (University of Zurich, EPFL)

Presenters: FEDERIAKIN, Denis (Johannes Gutenberg University Mainz); MOLENAAR, Dylan (University of Amsterdam); VELDKAMP, Karel (Universiteit van Amsterdam); DEBELAK, Rudolf (University of Zurich, EPFL)

Session Classification: Symposium : "Statistical Learning Approaches to Psychometric Modeling Challenges"

Track Classification: Statistical analyses: Statistical analyses