

Masterarbeit: Sampling und Sketching im 2PL-IRT-Modell

(english Version below)

Das zweiparametrische logistische Item-Response-Theorie-Modell (2PL Modell) wird unter anderem im Bereich der empirischen Bildungsforschung (z.B. PISA-Studie) eingesetzt, um Daten aus Leistungstests zu analysieren. Dabei kommen gerade bei computergestützten längsschnittlichen Erhebungen große Datenmengen zusammen. Traditionell werden bei beschränkter Rechenleistung Stichproben aus umfangreichen Datensätzen gezogen. Insbesondere dann, wenn neben den eigentlichen Leistungstestdaten auch Kovariaten von Interesse sind, ist bei Verwendung gängiger Samplingverfahren unklar, ob Analysen basierend auf Teildatensätzen zu ähnlichen Ergebnissen führen, wie auf dem vollständigen Datensatz. Eine zusätzliche Schwierigkeit besteht, wenn die Daten als Strom vorliegen, was insbesondere in computeradaptiven Test- und Trainingssystemen der Fall ist.

Im Kontext von logistischer Regression zeigen Vorarbeiten von Munteanu, Schwiigelshohn, Sohler, Woodruff (2018) sowie Munteanu, Omlor und Woodruff (2021), wie man sampeln oder sketchen kann, um eine hinreichende Approximationsgüte zu sichern. In dieser Masterarbeit sollen die Techniken von Munteanu et al. auf den Fall des 2PL-Modells übertragen werden. Dazu sind Vorkenntnisse im Bereich des 2PL-Modells nicht notwendig. Wichtig sind im Kontext der Arbeit Simulationen, günstig, aber nicht unabdingbar, sind Python-Vorkenntnisse.

Die Arbeit wird von Philipp Doebler gemeinsam mit Alexander Munteanu betreut.

Kontakt: doebler@statistik.tu-dortmund and alexander.munteanu@tu-dortmund

Datum: 2.7.2021

Master thesis: Sampling and sketching in the 2PL IRT model

The two-parameter logistic item response theory model (2PL model) is used, among other things, in the field of empirical educational research (e.g., PISA study) to analyze data from achievement tests. In this context, large amounts of data accumulate, especially in computer-based longitudinal surveys. Traditionally, samples are drawn from large data sets in case computing power is limited, or a fast analysis is desired. When covariates are of interest in addition to the actual performance test data, it is unclear when using common sampling methods whether analyses based on subsamples will yield similar results as those based on the full data set. An additional difficulty arises when the data are available as a stream, which is particularly the case in computer adaptive testing and training systems.

In the context of logistic regression, prior work by Munteanu, Schwiegelshohn, Sohler, Woodruff (2018) and Munteanu, Omlor, and Woodruff (2021) shows how to sample or sketch to ensure sufficient approximation quality. In this master thesis, the techniques of Munteanu et al. will be applied to the case of the 2PL model. For this purpose, prior knowledge of the 2PL model is not necessary. Important in the context of the thesis are simulations, favorable, but not indispensable, is previous Python knowledge.

The work is supervised by Philipp Doeblner together with Alexander Munteanu.

Contact: doebler@statistik.tu-dortmund and alexander.munteanu@tu-dortmund

Date: 2.7.2021