Reviewer's report on doctoral thesis:

A. Hladká: Statistical models for detection of differential item functioning

The main aim of the dissertation is to provide an overview of statistical methodology applicable in the framework of the differential item functioning (DIF) detection, to propose and investigate new methods, and to verify the theoretical properties in a simulation study. An important part of the dissertation is also the software implementation of the proposed methodology in R library `difNLR`. Apart of several already published papers, the author is working on a book (jointly with her supervisor) concerning computational aspects of psychometric methods.

Contents

Chapter 1 starts with a review of generalized logistic regression models. After describing several estimation algorithms and computing the estimates, the author derives corresponding asymptotic distributions and compares the proposed algorithms in a simulation study. The presentation is sufficiently clear but I am not sure whether it is required that $c_i < d_i$ (see, e.g., the parameter bounds described on page 21).

In Chapter 1, the author moves from the conditioning by 'true ability' to the conditioning by 'observed ability' (or matching criterion) without a sufficiently clear explanation: compare, for examples, formulas (1) and (1.1)—although this topic is discussed later on in Chapter 4. Would it be possible to investigate the consequences of replacing the true ability by the matching criterion (e.g. the total score) in a simulation study? Can it happen that the parameters of interest (DIF) are not identifiable?

Ordinal and nominal response variables are discussed in Chapter 2. Compared to the previous chapter, the author uses only one estimation algorithm but it is not clear whether it also encounters similar numerical instabilities as algorithms from Chapter 1.

Nonparametric kernel-based methods are proposed in Chapter 3. The new methodology is based on the nearest neighbour approach of Srihera and Stute (2010) and, compared to parametric models, it certainly provides more flexible estimators. A new estimator of the asymptotic variance is proposed together with a new computational algorithm (based on binning) on page 92. The main contribution is the discussion concerning weights and the estimates of optimal weights combined with wild bootstrap. In my opinion, the proposed methods are correct but the assumptions could be formulated more carefully: both kernels mentioned on page 90 are not twice continuously differentiable (as required by assumption (iii)) and it is unclear what is meant by twice differentiable 'density of the observed ability' on page 91; what would be the density of the typically discretely distributed total score?

Chapter 4 concerns two further topics. The so-called item purification addresses problems caused by estimating the true utility (this is possible by assuming the existence of DIF-free items). The effects of using item purification and multiple testing corrections are investigated in a simulation study.

Further comments

The statistical methodology concerning DIF detection and testing is presented clearly and systematically, including also a discussion of assumptions and theoretical properties of the estimators. The research is easily reproducible and applicable because the methods have been implemented in a standard statistical software.

The results were published in five journal articles, another journal article is under revision and further two journal articles are under preparation. The results should be published also in a book that is currently prepared for publication in CRC Press.

In summary, apart of an extensive review of existing DIF methodology and related simulation studies, the main contributions are:

- computational algorithms in Chapter 1, including also derivation of corresponding asymptotic distributions,
- nonparametric methods in Chapter 3, including estimation of optimal weights and wild bootstrap,
- the R library `difNLR`.

Of course, the research area is still very far from being exhausted and many topics and possible improvements (e.g., item purification and constrained nonparametric estimation) remain open for further research.

Some additional questions:

1. All algorithms in Chapter 1 seem to be numerically unstable but it is unclear whether similar problems were encountered also in Chapter 2.
2. Concerning NLR, it seems that the observations are heteroskedastic (see also last paragraph on page 23). Wouldn't it be better to use weighted least squares?
3. How many items were simulated on page 65? Did you use the simulated true ability or the observed matching criterion (total score) as the explanatory variable in the simulation study?
4. Would it be possible to use Bayesian methods and MCMC to estimate the unobserved quantities such as, e.g., the true ability or the guessing and inattention parameters?

5. Would it be possible to demonstrate possible advantages of the nonparametric estimator in Section 3.2 by simulating some even more complicated item responses than (3.11)?

6. It seems that the nonparametric approach can lead to nonmonotone regression functions. Is it possible to include suitable monotonicity constraints that could also improve the asymptotic properties of the nonparametric regression estimators?

Summary

Although the author investigates theoretical properties of the DIF estimators and tests, the results are interesting mainly from practical point of view. In my opinion, the dissertation clearly demonstrates the author's capability to independent creative work and, therefore, I recommend to award the scientific degree Ph.D. to Adéla Hladká.

Doc. RNDr. Zdeněk Hlávka, Ph.D.