Computational Statistics and Data Analysis 59 (2013) 144–160

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous multifactor DIF analysis and detection in Item Response Theory

F.B. Gonçalves [a,*], D. Gamerman [b], T.M. Soares [c,d]

[a] Departamento de Estatística, Universidade Federal de Minas Gerais, Brazil
[b] Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Brazil
[c] Departamento de Estatística, Universidade Federal de Juiz de Fora, Brazil
[d] CAEd, Universidade Federal de Juiz de Fora, Brazil

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form 7 October 2012
Accepted 10 October 2012
Available online 23 October 2012

Keywords:
Bayesian analysis
Educational assessment
Differential Item Functioning
Item Response Theory
MCMC
Multifactor

ABSTRACT

Item Response Theory (IRT) is a psychometric theory widely used in educational assessment and cognitive psychology to analyse data emerged from answers given to items contained in exams, questionnaires, etc. Standard IRT, however, is based on models which assume that items behave equally to all individuals. This may not be a reasonable assumption, especially when the individuals taking the test have different social and/or cultural backgrounds. Differential Item Functioning (DIF) is an area of IRT which allows an item to be perceived differently by distinct groups, respecting its usual characteristics. DIF hypothesis avoids neglecting items that may behave differently among groups and may also be used to provide important information about differences in the populations involved in the study. In this paper, two integrated Bayesian models for DIF analysis in IRT are proposed and compared. Both models are based on a two component mixture with one component describing DIF and the other accounting for the absence of DIF. Another contribution of this paper is the approach of the simultaneous presence of multiple factors causing DIF. Ideas from ANOVA models are used to characterize different possibilities associated with these factors. The models are also extended to account for explanation and detection in each factor. A simulation study was conducted to assess the model's capabilities and to compare it against existing alternatives. Special attention has been directed to the conditions required to ensure model identification. An analysis of a Mathematics exam applied nationally to Brazilian elementary school students is made considering two DIF factors: geographical region and type of school. The results highlight the relevance of the proposed methodology to address important issues in educational studying and testing.

1. Introduction

Item Response Theory (IRT) is a psychometric theory widely used in educational assessment and cognitive psychology. As its name suggests, it analyses data emerged from answers given by individuals to items contained in exams, questionnaires, etc.

The main educational assessment tests in the world (like PISA, GRE, SAT, TOEFL) make use of this theory to obtain their results. However, the applicability of IRT is not restricted to this kind of test, being applicable to any psychological test to estimate latent characteristics of the test taker that are related to the answers given in the test.

* Correspondence to: Av. Antônio Carlos, 6627 - ICEx - DEST, 31270-901, Belo Horizonte, MG, Brazil. Tel.: +55 31 34095939.
E-mail address: fbgoncalves@est.ufmg.br (F.B. Gonçalves).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.10.011

IRT models generally relate (probabilistically) the answer given to an item by an individual to characteristics of items and individuals. Common characteristics of the item include difficulty and discrimination. For the individual, on the other hand, it can be any latent trait or ability, such as proficiency in mathematics or level of depression.

Standard IRT is based on models that assume that items behave equally to all individuals. Differential Item Functioning (DIF), however, allows an item to function differently among groups in its usual characteristics: discrimination, difficulty and even guessing. Namely, if an item has DIF, students from different groups and with the same proficiency may have a different probability of correctly answering the item. It is useful to have items with no DIF (usually called anchor items) in order to correctly identify the proficiencies and DIF's.

The detection of items with DIF is an important step in DIF analysis. However, a complete analysis also requires satisfactory classification of the DIF found, identification of the factors associated to DIF and, perhaps, a hypotheses-confirmatory analysis. In this context, it is natural to construct regression models that associate covariates, related to the items, to the DIF's magnitude. The covariates would represent the DIF factors in such a way that the results of the regression analysis would confirm or not the formulated hypotheses.

The importance of Bayesian methods in IRT has steadily grown (see, for example, Albert, 1992; Patz and Junker, 1999b,a; Fox and Glas, 2001; Béguin and Glas, 2001). DIF analysis is a particularly appropriate environment for a genuine Bayesian formulation due to the complex structure of the models and the subjective decision features involved, which can be naturally formulated through the Bayesian argument. Janssen et al. (2004) proposed a model with item group predictors using a Bayesian approach. Soares et al. (2009) proposed a methodology for incorporating DIF detection along with parameter estimation from a Bayesian perspective. More recently, Frederickx et al. (2010) proposed a similar DIF model in a multivariate setting where the DIF values were correlated across groups. However, they only considered a single factor, DIF in the difficulty and did not propose any model structure to explain DIF.

This paper presents an integrated Bayesian approach for DIF detection, explanation and analysis. A model based on Gonçalves (2006) is introduced and compared with the model proposed by Soares et al. (2009). Both models simultaneously detect items with DIF in difficulty and discrimination and estimate all the other parameters of the model. They additionally provide explanation of DIF through covariates.

These two models, however, consider the existence of DIF in the item characteristics and/or allow for changes in the population model for the proficiencies according to a single classifying factor. Examples of factors include: gender, ethnicity, geographical region and type of school. These factors may jointly intervene in the analysis. This intervention can take place in different forms, ranging from no interactions among factors to the saturated model with all possible forms of interaction.

Practical considerations of the data set often leads to the consideration of multiple factors. Wang (2000) presents a model for multifactor analysis of DIF considering only the Rasch Model. In his model, at least one item has to be anchored (i.e., believed to have no DIF) for model identification. Besides, the DIF detection is not made simultaneously with the estimation of the other parameter and the DIF parameters are not explained through covariates. All these shortcomings are solved with the models introduced in this paper. These models explicitly consider the existence of multiple factors and discuss all possible forms they can intervene in the analysis. Additionally, the models allow for simultaneous detection and explanation of DIF with regard to all factors jointly considered.

The models introduced in this paper consider the three-parameter logistic model (Birnbaum, 1968), and may not require any previously set anchor item nor setting the mean of the DIF parameters as equal to zero. Moreover, the DIF detection is made simultaneously with the estimation of the other parameters and a regression structure is used to explain the DIF parameters using covariates. Given the large dimension of the parameter space, it becomes important to understand the implications of the model assumptions. Some issues of practical relevance concerning prior distributions, the identifiability of the model and the convergence of the MCMC algorithm will also be discussed. A simulated data set is used to illustrate the capabilities of the model and compare it against existing alternatives.

Subsequently, the models are used to analyse a data set concerning a mathematics exam applied to 4th grade students in Brazil's elementary school. The students are firstly separated by a factor determined by the country's main regions. Important cultural differences between the regions are shown to provide a substantial impact to the study. Then, another grouping factor associated to the type of school is also incorporated to the analysis. Different multifactor DIF models are then fitted and compared.

Section 2 shows the proposed models for DIF Analysis. Section 3 presents aspects of the Bayesian inference procedure. Section 4 presents a simulated study with the proposed models whereas Section 5 shows a real example in educational assessment.

2. Models for DIF analysis

Typically, in educational assessment, a test is formed by I items, but student j only answers a subset $I(j)$ of these items. Let $Y_{ij},\ j = 1, \dots, J$, be the score attributed to the answer given by the student j to the item $i \in I(j) \subset \{1, \dots, I\}$. Only the dichotomous case, where one of the scores in $\{0, 1\}$ is attributed to the item, will be considered. This way, $Y_{ij} = 1$, if the answer is correct and $Y_{ij} = 0$, if the answer is wrong. In general, there can be different types of DIF (see Hanson, 1998 for a wider characterization), but restricted to the characteristics of the three-parameter model—3PL Birnbaum (1968), the types of DIF can be immediately characterized according to the difficulty, discrimination and guessing.