LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz & Gunther Schauberger

A Penalty Approach to Differential Item Functioning in Rasch Models

Technical Report Number 134, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A Penalty Approach to Differential Item Functioning in Rasch Models

Gerhard Tutz & Gunther Schauberger

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

{ gerhard.tutz, gunther.schauberger}@stat.uni-muenchen.de

December 7, 2012

Abstract

A new diagnostic tool for the identification of differential item functioning (DIF) is proposed. Classical approaches to DIF allow to consider only few subpopulations like ethnic groups when investigating if the solution of items depends on the membership to a subpopulation. We propose an explicit model for differential item functioning that includes a set of variables, containing metric as well as categorical components, as potential candidates for inducing DIF. The ability to include a set of covariates entails that the model contains a large number of parameters. Regularized estimators, in particular penalized maximum likelihood estimators, are used to solve the estimation problem and to identify the items that induce DIF. It is shown that the method is able to detect items with DIF. Simulations and two applications demonstrate the applicability of the method.

Keywords: Rasch model, differential item functioning, penalized maximum likelihood, DIF lasso.

1 Introduction

Differential item functioning (DIF) is the well known phenomenon that the probability of a correct response among equally able persons differs in subgroups. For example, the difficulty of an item may depend on the membership to a racial, ethnic or gender subgroup. Then the performance of a group can be lower because these items are related to specific knowledge that is less present in this group. The effect is measurement bias and possibly discrimination, see, for example, Millsap and Everson (1993); Zumbo (1999). Various forms of differential item

functioning have been considered in the literature, see, for example, Holland and Wainer (1993); Osterlind and Everson (2009); Rogers (2005).

We will investigate DIF in item response models, focussing on the Rasch model. In item response models DIF is considered to be uniform, that is the probability of correctly answering is uniformly greater for specific subgroups. A main contribution in this area is Thissen et al. (1993), where tests are used to find items that show DIF. More recently, DIF has been embedded into the framework of mixed models (Van den Noortgate and De Boeck, 2005) and Bayesian approaches have been developed (Soares et al., 2009).

A severe limitation of existing approaches is that they are typically limited to the consideration of few subgroups. Most often, just two subgroups have been considered with one group being fixed as the reference group. The objective of the present paper is to provide tools that allow for several groups but also for continuous variables like age to induce differential item functioning. We propose a model that lets the item difficulties to be modified by a set of variables that can potentially cause DIF. The model necessarily contains a large number of parameters which raises severe estimation problems. But estimation problems can be solved by regularized estimation procedures. Although alternative strategies could be used we focus on regularization by penalization, using penalized maximum likelihood (ML) estimates. The procedure allows to identify the items that suffer from DIF and investigate which variables are responsible.

More recently Strobl et al. (2010) proposed a new approach that is also able to handle several groups and continuous variables but uses quite different estimation procedures. The work stimulated our research and we will compare our method to this alternative approach.

In Section 2 we present the model, in Section 3 we show how the model can be estimated. Then we illustrate the fitting of the model by use of simulation studies and real data examples.

2 Differential Item Functioning Model

We will first consider the binary Rasch model and then introduce a general parametric model for differential item functioning.

2.1 The Binary Rasch Model

The most widespread item response model is the binary Rasch model (Rasch, 1960). It assumes that the probability that a participant in a test scores on an item is determined by the difference between two latent parameters, one representing the person and one representing the item. In assessment tests the person parameter refers to the ability of the person and the item parameter to the difficulty of the item. More generally the person parameter refers to the latent trait

the test is supposed to measure. With $X_{pi} \in \{0, 1\}$ the probability that person p solves item i is given by

$$P(X_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad p = 1, \ldots, P \quad , i = 1, \ldots, I$$

where θ_p is the person parameter (ability) and β_i is the item parameter (difficulty). A more convenient form of the model is

$$\log(\frac{P(X_{pi} = 1)}{P(X_{pi} = 0)}) = \theta_p - \beta_i, \tag{1}$$

where the left hand side represents the so-called logits, $\text{Logit}(P(X_{pi} = 1) = \log(P(X_{pi} = 1)/P(X_{pi} = 0))$. It should be noted that the parameters are not identifiable. Therefore one has to fix one of the parameters. We choose $\theta_P = 0$, which yields a simple representation of the models to be considered later.

Under the usual assumption of conditional independence given the latent traits the maximum likelihood (ML) estimates can be obtained within the framework of generalized linear models (GLMs). GLMs for binary responses assume that the probability $\pi_{pi} = P(X_{pi} = 1)$ is given by $g(\pi_{pi}) = \boldsymbol{x}_{pi}^T \boldsymbol{\delta}$, where $g(.)$ is the link function and $\boldsymbol{x}_{pi}$ is a design vector linked to person p and item i. The link function is directly seen from model representation (1). The design vector, which codes the persons and items and the parameter vector are seen from

$$\log(\frac{P(X_{pi} = 1)}{P(X_{pi} = 0)}) = \theta_p - \beta_i = \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta},$$

where $\mathbf{1}_{P(p)}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$ has length $P - 1$ with 1 at position p, $\mathbf{1}_{I(i)}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$ has length I with 1 at position i, and the parameter vectors are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{P-1})$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)$ yielding the total vector $\boldsymbol{\delta}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$. The design vector linked to person p and item i is given by $\boldsymbol{x}_{pi}^T = (\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T)$.

GLMs are extensively investigated in McCullagh and Nelder (1989), short introductions with the focus on categorical data are found in Agresti (2002) and Tutz (2012). The embedding of the Rasch model into the framework of generalized linear models has the advantage that software that is able to fit GLMs and extensions can be used to fit models very easily.

2.2 A General Differential Item Functioning Model

In a general model that allows the item parameters to depend on covariates that characterize the person we will replace the item parameter by a linear form that includes a vector of explanatory variables. Let $\boldsymbol{x}_p$ be a person-specific parameter that contains, for example, gender, race, but potentially also metric covariates like age. If β_i is replaced by $\beta_i + \boldsymbol{x}_p^T \boldsymbol{\gamma}_i$ with item-specific parameter $\boldsymbol{\gamma}_i$ one obtains the model

$$\log \frac{P(X_{pi}=1)}{P(X_{pi}=0))} = \theta_p - (\beta_i + \boldsymbol{x}_p^T \boldsymbol{\gamma}_i) \tag{2}$$

For illustration let us consider the simple case where the explanatory variable codes a subgroup like gender, which has two possible values. Let $x_p = 1$ for males and $x_p = 0$ for females. Then, if item i functions differently in the subgroups, one has the item parameters

$$\beta_i + \gamma_i \text{ for males and } \quad \beta_i \text{ for females.}$$

Then γ_i represents the difference of item difficulty between males and females. If one prefers a more symmetric representation one can choose $x_p = 1$ for males and $x_p = -1$ for females obtaining

$$\beta_i + \gamma_i \text{ for males and } \quad \beta_i - \gamma_i \text{ for females.}$$

Then γ_i represents the deviation of the sub-populations in item difficulty from the baseline difficulty β_i. Of course in an item that does not suffer from differential item functioning, one has $\gamma_i = 0$ and therefore, items for males and females are equal.

The strength of the general model (2) is that also metric covariates like age can be included. Thinking of items that are related to knowledge on computers or modern communication devices the difficulty may well vary over age. One could try to build more or less artificial age groups, or, as we do, assume linear dependence of the logits. With x_p denoting age in years the item parameter is $\beta_i + \text{age}\gamma_i$. If $\gamma_i = 0$ the item difficulty is the same for all ages.

The multi-group case is easily incorporated by using dummy-variables for the groups. Let R denote the group variable, for example, race with k categories, that is, $R \in \{1, \ldots, k\}$. Then one builds a vector $(x_{R(1)}, \ldots, x_{R(k-1)})$, where components are defined by $x_{R(j)} = 1$ if $R = j$ and $x_{R(j)} = 0$ otherwise. The corresponding parameter vector $\boldsymbol{\gamma}_i$ has $k-1$ components $\boldsymbol{\gamma}_i^T = (\gamma_{i1}, \ldots, \gamma_{i,k-1})$. Then the parameters are

$$\beta_i + \gamma_{i1} \quad \text{in group 1,} \quad \ldots \quad \beta_i + \gamma_{i,k-1}, \quad \text{in group } k-1 \quad \beta_i \text{ in group } k.$$

In this coding the last category, k, serves as reference category, and the parameters $\gamma_{i1}, \ldots, \gamma_{i,k-1}$ represent the deviations of the subgroups with respect to the reference category.

One can also use symmetric coding where one assumes $\sum_{j=1}^{k} \gamma_{ij} = 0$ yielding parameters

$$\beta_i + \gamma_{i1} \quad \text{in group 1,} \quad \ldots \quad \beta_i + \gamma_{i,k-1}, \quad \text{in group } k-1 \quad \beta_i + \gamma_{ik} \text{ in group } k.$$

In effect one is just coding a categorical predictor in $0-1$-coding or effect coding, see, for example, Tutz (2012).

The essential advantage of model (2) is that the person-specific parameter includes all the candidates that are under suspicion to induce differential item functioning. Thus one has a vector that contains age, race, gender and all the other candidates. If one component in the vector $\boldsymbol{\gamma}_i$ is unequal zero the item is group-specific, the parameter shows which of the variables is responsible for the differential item functioning. The model includes not only several grouping variables but also metric explanatory variables.

The challenge of the model is to estimate the large number of parameters and to determine which parameters have to be considered as unequal zero. The basic assumption is that most of the parameters do not depend on the group, but some can. One wants to detect these items and know which one of the explanatory variables is responsible. For the estimation one has to use regularization techniques that are discussed in the next section.

3 Estimation by Regularization

3.1 Maximum Likelihood Estimation

Let the data be given by $(X_{pi}, \boldsymbol{x}_p)$, $p = 1, \ldots, P, i = 1, \ldots, I$. Maximum likelihood estimation of the model is straightforward by embedding the model into the framework of generalized linear models. By using again the coding for persons and parameters in the parameter vectors $\mathbf{1}_{P(p)}$ and $\mathbf{1}_{I(i)}$ the model has the form

$$\begin{aligned}\log \frac{P(X_{pi}=1)}{P(X_{pi}=0))} &= \theta_p - \beta_i - \boldsymbol{x}_p^T \boldsymbol{\gamma}_i \\ &= \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta} - \boldsymbol{x}_p^T \boldsymbol{\gamma}_i.\end{aligned}$$

With the total vector given by $(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T)$ one obtains for observation X_{pi} the design vector $(\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T, 0, 0, , \ldots, -\boldsymbol{x}_p^T \ldots, 0, 0)$, where the component $-\boldsymbol{x}_p^T$ corresponds to the parameter $\boldsymbol{\gamma}_i$.

Although ML estimation is straightforward estimates will exist only in very simple cases, for example, if the explanatory variable codes just wo subgroups. In higher dimensional cases ML estimation will deteriorate and no estimates or selection of parameters are available.

3.2 Penalized Estimation

In the following we will consider regularization methods that are based on penalty terms. The general principle is, not to maximize the log-likelihood function, but a penalized version. Let $\boldsymbol{\alpha}$ denote the total vector of parameters, in our case $\boldsymbol{\alpha} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T)$. Then one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

where $l(.)$ is the common log-likelihood of the model and $J(\boldsymbol{\alpha})$ is a penalty term that penalizes specific structures in the parameter vector. The parameter λ is a tuning parameter that specifies how serious the penalty term has to be taken. A widely used penalty term in regression problems is $J(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{\alpha}$, that is, the squared length of the parameter vector. The resulting estimator is known under the name ridge estimate, see Hoerl and Kennard (1970) for linear models and Nyquist (1991) for the use in GLMs. Of course, if $\lambda = 0$ maximization yields the ML estimate. If $\lambda > 0$ one obtains parameters that are shrunk toward zero. In the extreme case $\lambda \to \infty$ all parameters are set to zero. The ridge estimator with small $\lambda > 0$ stabilizes estimates but does not select parameters, which is the main objective here. Penalty terms that are useful because they enforce selection are L_1-penalty terms.

Let us start with the simple case of a univariate explanatory variable, which, for example, codes gender. Then the proposed *lasso penalty for differential item functioning* (DIF-lasso) is given by

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^{I} |\gamma_i|,$$

which is a version of the L_1-penalty or lasso (for least absolute shrinkage and selection operator), which was propagated by (Tibshirani, 1996) for regression models. It should be noted that the penalty term contains only the parameters that are responsible for differential item functioning, therefore only the parameters that carry the information on DIF are penalized. Again, if $\lambda = 0$ maximization yields the full ML estimate. For very large λ all the γ-parameters are set to zero. Therefore in the extreme case $\lambda \to \infty$ the Rasch model is fitted without allowing for differential item functioning. The interesting case is in between, when λ is finite and $\lambda > 0$. Then the penalty enforces selection. Typically, for fixed λ, some of the parameters are set to zero while others take values unequal zero. With a carefully chosen tuning parameter λ the parameters that yield estimates $\hat{\boldsymbol{\gamma}}_i > 0$ are the ones that show DIF.

For illustration we consider a Rasch model with 10 items and 70 persons. Among the 10 items three suffer from DIF induced by a binary variable with parameters $\gamma_1 = 2,\ \gamma_2 = -1.5,\ \gamma_3 = -2$. Figure 1 shows the coefficient build-ups for the γ-parameters for one data set, that is, how the parameters evolve with decreasing tuning parameter λ. In this data set ML estimates existed. We do not use λ itself on the x-axis but a transformation of λ that has better scaling properties. Instead of giving the λ-values on the x-axis we scale it by $\|\hat{\boldsymbol{\gamma}}\|/\max\|\hat{\boldsymbol{\gamma}}\|$, where $\max\|\hat{\boldsymbol{\gamma}}\|$ corresponds to the L_2-norm of the maximal obtainable estimates, that is, the ML estimates. On the right side of Figure 1 one sees the estimates for $\lambda = 0$ ($\|\hat{\boldsymbol{\gamma}}\|/max\|\hat{\boldsymbol{\gamma}}\| = 1$), which correspond to the ML estimates for the DIF model. At the left end all parameters are shrunk to zero, corresponding to the value of λ, where the simple Rasch model without DIF

is fitted. Thus the figure shows how estimates evolve over diminishing strength of regularization. At the right end no regularization is exerted, at the left side regularization is so strong that all γ-parameters are set to zero. The vertical line shows the tuning parameter selected by BIC (see below), which represents the best estimate for this selection criterion. If one uses this criterion all items with DIF (dashed lines) are selected, obtaining estimates unequal zero. But for all items without DIF the estimates are zero. Therefore in this data set identification was perfect.

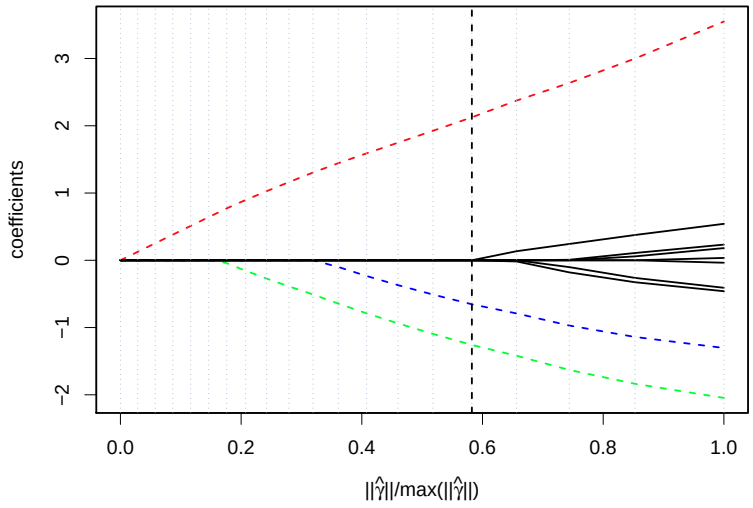

FIGURE 1: *Coefficient build-up for Rasch model with DIF induced by binary variable, dashed lines are the items with DIF, solid lines are the items without DIF.*

In the general case with a vector of covariates that potentially induce DIF a more appropriate penalty is a modification of the grouped lasso (Yuan and Lin, 2006; Meier et al., 2008). Let $\boldsymbol{\gamma}_i^T = (\gamma_{i1}, \ldots, \gamma_{im})$ denote the vector of modifying parameters of item i, where m denotes the length of the person-specific covariates. Then the *group lasso penalty for item differential functioning* (DIF-lasso) is

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^{I} \|\boldsymbol{\gamma}_i\|,$$

where $\|\boldsymbol{\gamma}_i\| = (\gamma_{i1}^2 + \cdots + \gamma_{im}^2)^{1/2}$ is the L_2-norm of the parameters of the ith item with m denoting the length of the covariate vector. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\gamma}}_i = \mathbf{0}$ or $\gamma_{is} \neq 0$ for $s = 1, \ldots, m$. Thus the whole group of parameters collected in $\boldsymbol{\gamma}_i$ is shrunk simultaneously toward zero.

For a geometrical interpretation of the penalty, see Yuan and Lin (2006). The effect is that in a typical application only some of the parameters get estimates $\hat{\boldsymbol{\gamma}}_i \neq \mathbf{0}$. These correspond to items that show DIF.

Choice of Penalty Parameter

An important issue in penalized estimation is the choice of the tuning parameter λ. In our case it determines the numbers of items identified as inducing DIF. Therefore it determines if all items with DIF are correctly identified and also if some are falsely diagnosed as DIF-items. To find the final estimate in the solution path it is necessary to balance the complexity of the model and the data fit. However, one problem is to determine the complexity of the model, which in penalized estimation approaches is not automatically identical to the number of parameters in the model. We worked with several criteria for the selection of the tuning parameter, including cross-validation and AIC criteria with the number of parameters determined by the degrees of freedom for the lasso (Zou et al., 2007). A criterion that yielded a satisfying balancing and which has been used in the simulations and applications is the BIC (Schwarz, 1978) with the degrees of freedom for the group lasso penalty determined by a method proposed by Yuan and Lin (2006). Here, the degrees of freedom (of penalized parameters $\boldsymbol{\gamma}$) are approximated by

$$\tilde{df}_{\boldsymbol{\gamma}}(\lambda) = \sum_{i=1}^{I} I(\|\boldsymbol{\gamma}_i(\lambda)\| > 0) + \sum_{i=1}^{I} \frac{\|\boldsymbol{\gamma}_i(\lambda)\|}{\|\boldsymbol{\gamma}_i^{ML}\|}(m-1).$$

Since the person parameters and the item parameters are unpenalized, the total degrees of freedom are $df(\lambda) = I + P + \tilde{df}_{\boldsymbol{\gamma}}(\lambda) - 1$. The corresponding BIC is determined by

$$BIC(\lambda) = -2 \cdot l(\boldsymbol{\alpha}) + df(\lambda) \cdot \log(P \cdot I),$$

where $l(\boldsymbol{\alpha})$ is the log-likelihood of the current parameter vector $\boldsymbol{\alpha}$.

Some Remarks

We focus on penalized ML estimation. Regularized estimation with penalty terms has the advantage that the penalty term is given explicitly, and therefore it is known how estimates are shrunk. An alternative procedure that could be used is boosting. It selects relevant variables by using weak learners and regularization is obtained by early stopping, see, for example, Bühlmann and Hothorn (2007), and for logistic models Tutz and Binder (2006). Although the form of regularization is not given in an explicit form it typically is as efficient as regularization with corresponding penalty terms. Also mixed model methodology as used by (Soares et al., 2009) to estimate DIF can be combined with penalty terms that enforce selection. However, methodology is in its infancy, see, for example, Ni et al. (2010); Bondell et al. (2010).

4 The Fitting Procedure At Work

In the present section it is investigated if the procedure is able to detect the items that show DIF. This is done in a simulation study where it is known which items are affected by DIF.

Illustration

For illustration we will first consider several examples. In the first example we have 70 persons, 10 items, three with DIF ($\boldsymbol{\gamma}_1^T = (-1, 0.8, 1)$, $\boldsymbol{\gamma}_2^T = (-1.1, 0.5, 0.9)$, $\boldsymbol{\gamma}_3^T = (1, -1, -1)$, $\boldsymbol{\gamma}_4^T = \ldots = \boldsymbol{\gamma}_{10}^T = (0, 0, 0)$). The upper panel in Figure 2 shows the coefficient build-ups for an exemplary data set. Now one item is represented by three lines one for each dimesion of the covariate. Again, items with DIF are given by non-solid lines and items with DIF by solid lines. In this data set the BIC criterion selects all the items with DIF and sets all items without DIF to zero. In the lower panel one sees a data set where identification is not perfect. It is seen that some items without DIF are falsely considered as inducing DIF. But also in this data set the items with DIF are the first ones to obtain estimates unequal zero when penalization is relaxed. The items without DIF obtain estimates unequal zero but estimates are very small.

An example without DIF is seen in Figure 3. The setting is the same as before ($P = 70,\ I = 10$) but all γ-parameters are set to zero. It is seen that the procedure also works well in the case of the Rasch model because all γ-parameters are estimated as zero.

For further illustration we show in the upper panel of Figure 4 the estimates of 100 simulated data sets for the same setting as in Figure 2. The box-plots show the variability of the estimates, the stars denote the underlying true values. The β-parameters in the left block represent the basic item parameter, which are estimated rather well. In the next block the modifying parameters γ_{is} are shown for items with DIF and in the last block the modifying parameters for items without DIF are shown. In this last block the stars that denote true values are omitted since they are all zero. Overall the estimates of the basic β-parameters (first block) and the items without DIF (third block) are quite close to their true values. In particular the estimates of the parameters that correspond to items without DIF are zero or close to zero and are frequently diagnosed as not suffering from DIF. The γ-parameters in the middle block, which correspond to items with DIF, are distinctly unequal zero and therefore the DIF-items are identified. But the latter estimates are downward biased because of the exerted penalization, which shrinks the estimates.

The bias can be removed and estimators possibly improved by an additional refit. The fit of the model in combination with the selection of the tuning parameter yields the set of items that are considered as suffering from DIF. To avoid shrinkage and bias one can compute a final un-penalized ML fit of the re-

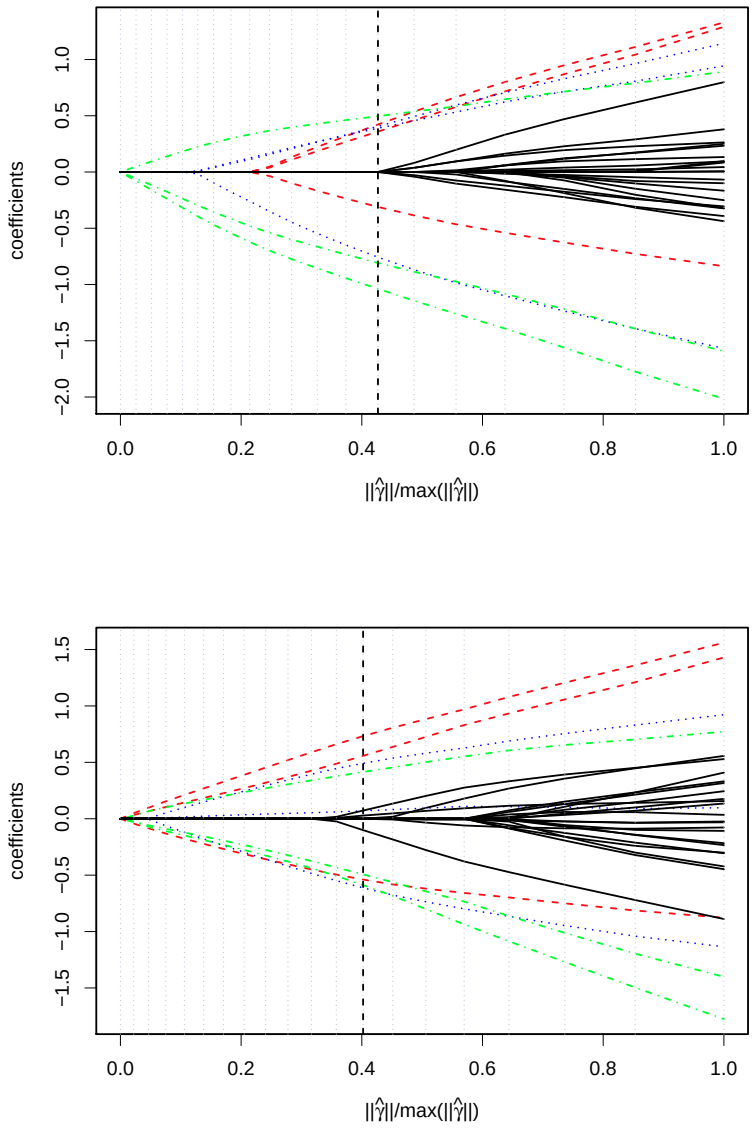

FIGURE 2: *Coefficient build-up for Rasch model with DIF induced by three variables, dashed lines are the items with DIF, solid lines are the items without DIF, upper panel shows perfect identification, in the lower panel identification is not perfect.*

duced model that contains only the parameters that have been selected as being non-zero. In the lower panel of Figure 4 the estimates with a final refit step are given. While the estimation of the basic β-parameters has hardly changed, the downward bias in item parameters for items with DIF is removed. However, the

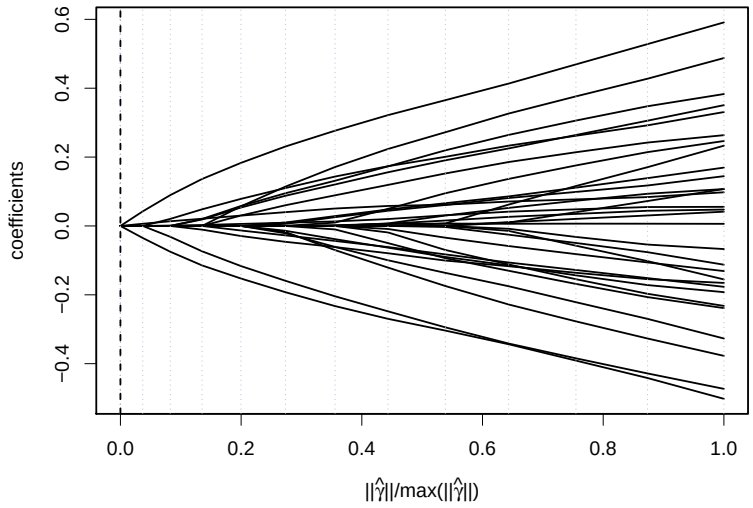

FIGURE 3: *Coefficient build-up for Rasch model without DIF .*

estimates of parameters for items without DIF automatically suffers. If one of these items is diagnosed as DIF-item the final ML-fit yields larger values than the penalized estimate.

Simulation Scenarios

In the following we give results for selected simulation scenarios based on 100 simulations. The person parameters are drawn from a standard normal distribution and we consider scenarios with varying strength of DIF. The item parameters have the form $\beta_i + \boldsymbol{x}_p^T \boldsymbol{\gamma}_i$. We always work with standardized person characteristics $\boldsymbol{x}_p$, that is, the components have variance 1. A measure for the strength of DIF in an item is the variance $V_i = \mathrm{var}(\beta_i + \boldsymbol{x}_p^T \boldsymbol{\gamma}_i)$, which has the value $\sum_j \gamma_{ij}^2$ for independent components. The average of $\sqrt{V_i}$ over the items *with* DIF gives a measure of the strength of DIF in these items. The implicitly used reference value is the standard deviation of the person parameters, which is 1. For the parameter given previously the average is 1.61. We consider this scenario as strong DIF, the value 1 (for parameters $\gamma_{ij}/1.61$) as medium DIF and 0.8 (for parameters $\gamma_{ij}/2$) as weak DIF. An overall measure of DIF in a setting is the average of $\sqrt{V_i}$ over *all* items. For the strong scenario with 10 items one obtains 0.48, for the medium and weak 0.3 and 0.24, respectively.

When calculating mean squared errors we distinguish between person and item parameters. For person parameters it is the average over simulations of

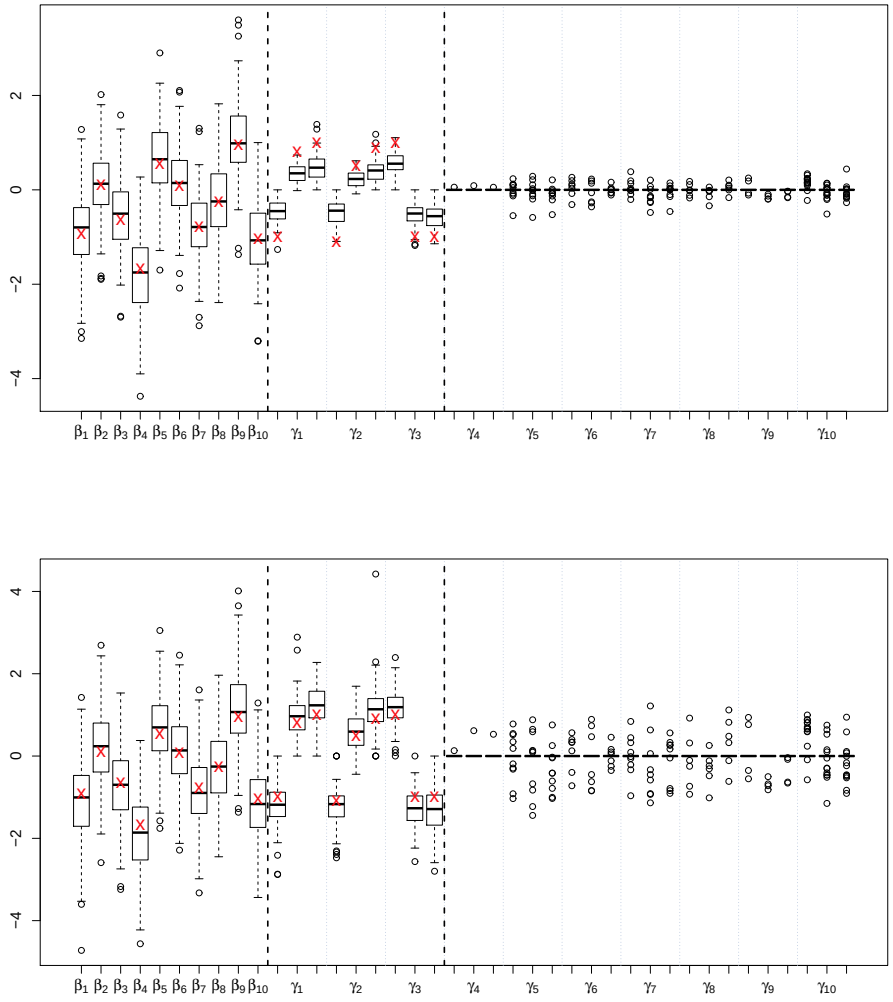

FIGURE 4: *Upper panel: Box plots of estimates for Rasch model with DIF induced by three variables, stars denote true values. Lower panel: the same model with a final ML step on selected items.*

$\sum_p(\hat{\theta}_p - \theta_p)^2/P$. For items it is the squared difference between the estimated item difficulty and the actual difficulty $\sum_p \sum_i[(\beta_i + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i) - (\hat{\beta}_i + \boldsymbol{x}_p^T\hat{\boldsymbol{\gamma}}_i)]^2/(I \cdot P)$.

One of the main objectives of the method is the identification of items with DIF. The criteria by which the performance of the procedure can be judged are the hits or true positives (i.e. the number of correctly identified items with DIF) and the false positives (i.e. the number of items without DIF that are falsely diagnosed as items with DIF).

The settings are the following. The first has been used in the illustrations

given before, but, for completeness it is given again.

- Setting 1: 70 persons, 10 items, 3 with DIF on 3 variables, parameters (strong DIF): $\boldsymbol{\gamma}_1^T = (-1, 0.8, 1)$, $\boldsymbol{\gamma}_2^T = (-1.1, 0.5, 0.9)$, $\boldsymbol{\gamma}_3^T = (1, -1, -1)$, $\boldsymbol{\gamma}_4^T = \ldots = \boldsymbol{\gamma}_{10}^T = (0, 0, 0)$, two variables binary, one standard normally distributed,
- Setting 2: 120 persons, items as in setting 1,
- Setting 3: 300 persons, 40 items, 4 with DIF on 5 variables, parameters (strong DIF): $\boldsymbol{\gamma}_1^T = (-1, 0.8, 0, 0, 1)$, $\boldsymbol{\gamma}_2^T = (0, 1.1, 0.9, 0, 0.9)$, $\boldsymbol{\gamma}_3^T = (0.8, 0, -1, -1, 0)$, $\boldsymbol{\gamma}_4^T = (0, 0, 1, 0.9, 0.7)$, $\boldsymbol{\gamma}_5^T = \ldots = \boldsymbol{\gamma}_{40}^T = (0, 0, 0, 0, 0)$, two variables binary, three standard normally distributed.

In Table 1 the MSEs as well as the hits and false positive rates are given for the fit of the Rasch model (without allowing for DIF), the DIF-lasso and the DIF-lasso with refit. It is seen that the accuracy of the estimation of person parameters does not depend on the strength of DIF. It is quite similar for strong, medium and weak DIF. Also the fitting of the Rasch model or DIF-lasso yields similar estimates of person parameters. The refit, however, yields poorer estimates in terms of MSE for smaller number of persons, but for 300 persons there is hardly a difference. The estimation of item parameters shows a different picture. DIF-lasso distinctly outperforms the Rasch model, in particular if DIF is strong. The refit procedure can again be recommended for large number of persons but not for small numbers. For illustration in Figure 5 the box plots for setting 2 with strong DIF are shown. The picture does not show the four data sets that produced extreme values for all methods.

Since our focus is on the identification of DIF-items the hits and false positive rates are of particular interest. It is seen from the lower panel of Table 1 that the procedure works well. If DIF is strong the hit rate is close to 1, for medium DIF one needs more persons in the setting to obtain an average of 0.80. Of course for weak DIF identification is harder and one will not always find all the items with DIF. One nice result is that the false positive rate is negligible, although not all items with DIF may be found, it hardly occurs that items without DIF are falsely diagnosed.

5 Examples

5.1 Exam Data

Our first data example deals with the solution of problems in an exam following a course on multivariate statistics. There were 18 problems to solve and 57 students. In this relatively small data set two variables that could induce DIF were available, the binary variables level (bachelor student of statistics: 1, master

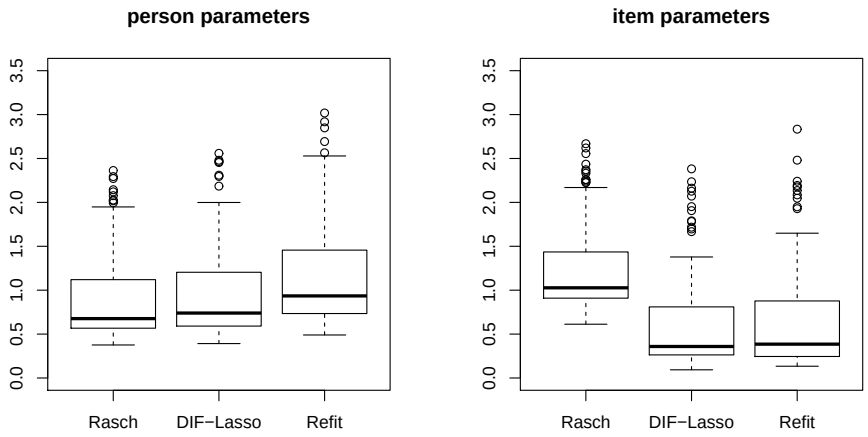

FIGURE 5: *Box plots of MSEs for Setting 2 (strong)*

student with a bachelor in an other area: 0) and gender (male: 0, female: 1). Figure 6 shows the coefficient build-ups. With BIC as selection criterion no item showed DIF. So we were happy that the results did not indicate that the exam was preferring specific subgroups.

In this simple case, in which potential DIF is induced by binary variables, which indicate the sub populations, one can also use test statistics to examine if DIF is present because ML estimates exist. The embedding into the framework of generalized linear models allows to use the likelihood ratio test to test the null hypothesis $\gamma_1 = \ldots, \gamma_I = 0$ (for the theory see, for example Tutz (2012)). We consider the effects of gender and level separately. The p-values are 0.28 for gender and 0.38 for level. The result supports that DIF is not present. Alternatively, we used model checks based on conditional estimates as Andersen's likelihood ratio test (Andersen, 1973), which is implemented in the R-package `eRm`, see Mair et al. (2012) and Mair and Hatzinger (2007). These tests resulted in p-values of 0.315 for gender and 0.417 for level and also support that DIF is not an issue in this data set.

5.2 Knowledge Data

An example that has also been considered by Strobl et al. (2010) uses data from an online quiz for testing one's general knowledge conducted by the weekly German news magazine SPIEGEL. The 45 test questions were from five topics, politics, history, economy, culture, and natural sciences. We use the same sub sample as Strobl et al. (2010) consisting of 1075 university students from Bavaria, who had

		MSE person parameters			MSE item parameters		
Setting		Rasch	DIF-Lasso	Refit	Rasch	DIF-Lasso	Refit
	strong	1.08	1.13	1.46	1.42	0.90	1.17
1	medium	1.11	1.11	1.23	0.96	0.85	0.97
	weak	1.17	1.15	1.20	0.91	0.85	0.92
	strong	0.98	1.04	1.32	1.30	0.70	0.85
2	medium	1.07	1.08	1.24	0.88	0.70	0.81
	weak	1.10	1.06	1.18	0.80	0.68	0.78
	strong	0.25	0.24	0.25	0.38	0.19	0.15
3	medium	0.24	0.22	0.24	0.20	0.14	0.13
	weak	0.23	0.22	0.23	0.16	0.13	0.13

Setting		hits	false positives
	strong	0.97	0.071
1	medium	0.50	0.010
	weak	0.33	0.006
	strong	1.00	0.061
2	medium	0.84	0.026
	weak	0.52	0.014
	strong	0.99	0.017
3	medium	0.80	0.003
	weak	0.42	0.000

TABLE 1: *MSEs for the simulation scenarios (upper panel) and average rates of hits/false positives (lower panel)*

.

all been assigned a particular set of questions. The covariates that we included as potentially inducing DIF are gender, age, semester of university enrollment, an indicator for whether the student's university received elite status by the German excellence initiative (elite), and the frequency of accessing SPIEGEL's online magazine (spon).

Figure 7 shows as an example the coefficient build-ups for the covariate gender. At the path point that was selected by the BIC criterion (dashed vertical line), 17 of the 45 items showed DIF, which is not surprising because it is not a carefully constructed test that really focusses on one latent dimension. In Figure 8 the estimated effects of the items containing DIF are visualized. The upper panel shows the profile plots of the parameters for the included covariates. For each item with DIF one profile is given. The lower panel shows the strengths of the effects in terms of the absolute value of the coefficients. One boxplot refers to the absolute values of the 17 parameters for one covariate. It is seen that the strongest effects are found for the covariate gender, the weakest effects are in

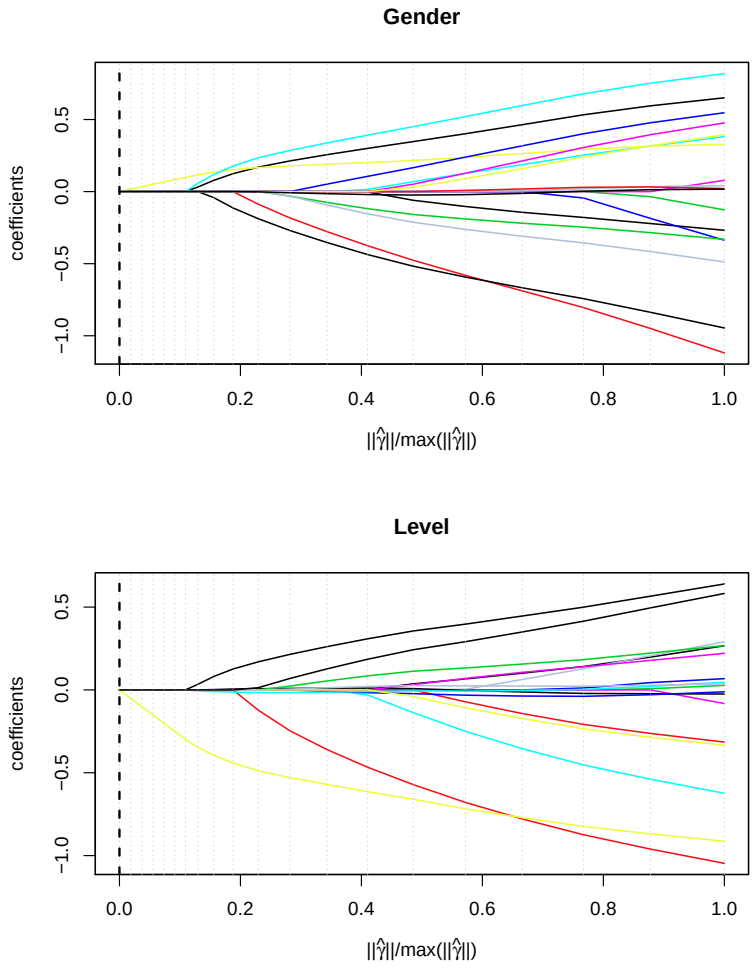

FIGURE 6: *Coefficient build-ups for exam data.*

the variable elite, which measures the status of the university where the student is enrolled. It should be noted that the importance of the single covariates for the DIF can be measured by the absolute values of their coefficients since all covariates were standardized.

In Figure 8 (upper panel) four items are represented by dashed lines. They showed the strongest DIF in terms of the L_2- norm of the estimated parameter vector. All of them refer to economics. For illustration, these four items are considered in more detail. They are

- Zetsche: "Who is this?" (a picture of Dieter Zetsche, the CEO of the Daimler AG, maker of Mercedes cars, is shown).

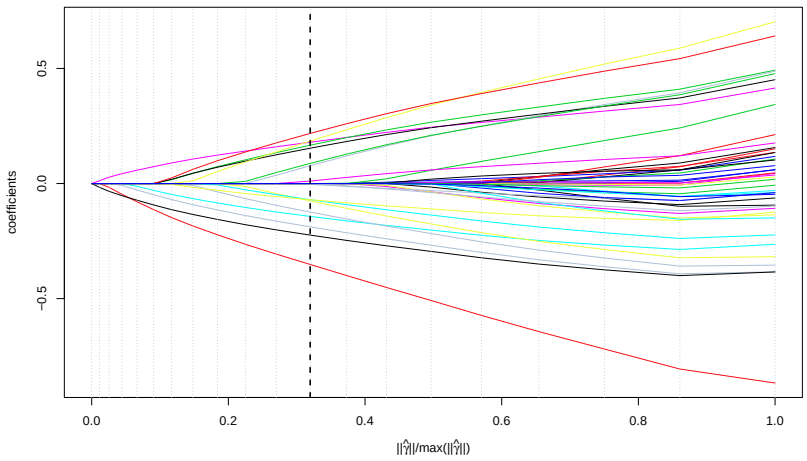

FIGURE 7: *Coefficient build-ups for covariate gender in Quiz Data; dashed vertical line indicates BIC-optimal path point*

- manufacturies: "What is the historic meaning of manufacturies?" (Manufacturies were the precursors of industrial mass production).
- Deutsche Bank: "Which company does this logo represent?" (Deutsche Bank).
- AOL: "Which internet company took over the media group Time Warner?" (AOL).

The profiles for the items Zetsche, Deutsche Bank and AOL are quite similar. They are distinctly easier for male participants and for frequent visitors of SPIEGELonline. The item manufacturies shows a quite different shape being definitely easier for females. It is also easier to solve for students that are not frequent visitors of SPIEGELonline. The item differs from the other three items because it refers more to a broad education than to current issues. In this respect female students and students that do not follow the latest news seem to find the item easier. Therefore the different profile.

6 An Alternative Method

In contrast to most existing methods the proposed procedure allows to include all variables that might lead to DIF and identify the items with DIF. Quite recently Strobl et al. (2010) proposed a new procedure that is also able to investigate the effect of a set of variables. Therefore it seems warranted to discuss the differences between our method and the recursive partitioning approach advocated by Strobl et al. (2010).

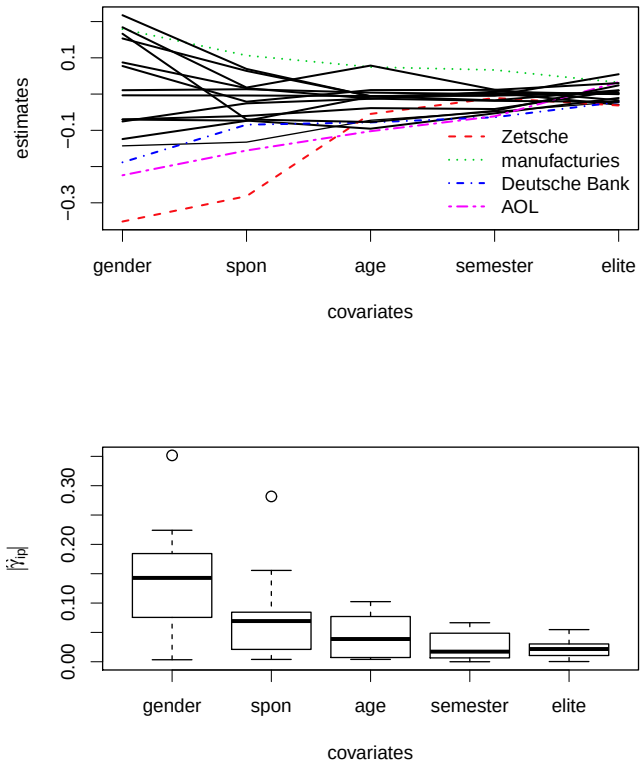

FIGURE 8: *Upper panel: profile plot for coefficient estimates of items with DIF, profiles of the four items with highest DIF are highlighted; lower panel: boxplots of absolute values of coefficient-estimates for items with DIF*

Recursive partitioning is similar to CARTs (Classification and Regression Trees), which were propagated by Breiman et al. (1984). For a more recent introduction see Hastie et al. (2009), or from a psychological viewpoint Strobl et al. (2009). The basic concept of recursive partitioning and tree methods in regression models is to recursively partition the covariate space such that the dependent variable is explained best. In the case of continuous predictors partitioning of the covariate space means that one considers splits in single predictors, that is, a predictor X is split into $X \leq c$ and $X > c$ where c is a fixed value. All values c are evaluated and the best split is retained. If a predictor is categorical splits refer to all possible subsets of categories. Recursive partitioning means that one finds the predictor together with the cut-off value c that explains the dependent variable best. Then given $X \leq c$ (and the corresponding sub sample) one repeats the procedure searching for the best predictor and cut-off value that works best for the sub sample with $X \leq c$. The same is done for the sub sample with $X > c$. The procedure of consecutive splitting can be visualized in a tree.

Of course, there are many details to consider, for example, one has to define what best explanation of the dependent variable means, when to stop the procedure and other issues. For details see Breiman et al. (1984).

In item response models the partitioning refers to the predictors that characterize the persons. That means when using the person-specific variable X, for example, age, it is split into $X \leq c$ and $X > c$. The Rasch model is fit in these sub populations yielding different estimates of item parameters. Then one has to decide if the difference between item estimates before splitting and after splitting is systematic or random. If it is systematic the split is warranted. For the decision Strobl et al. (2010) use structural change tests, which have been used in econometrics (see also Zeileis et al. (2008)). Although the basic concept is the same as in the partitioning in regression models, now a model is fitted and therefore the method is referred to as model based partitioning. For details see Strobl et al. (2010).

For the knowledge data Strobl et al. (2010) identified gender, spon and age as variables that induce DIF. This is in accordance with our results (Figure 8), which also identified these variables as the relevant ones. By construction the partitioning approach yields areas, in which the effect is estimated as constant. The partitioning yielded eight subpopulations, for example, $\{female, spon \leq 1, age \leq 21\}$ and $\{male, spon \leq 2-3, age \leq 22\}$. Within these subspaces all items have estimates that are non-zero. Items that have particularly large values are considered as showing DIF. It is not clear what criterion is used to identify the items that actually show DIF. Strobl et al. (2010) just describe 5 items that seem to have large values. Therefore, one can not compare the two approaches in terms of the number of selected items.

Let us make some remarks on the principles of the recursive partitioning approach to DIF and the penalization method proposed here.

Recursive partitioning can be considered a non-parametric approach as far as the predictors are concerned. No specific form of the influence of predictors on items is assumed. But, in the case of continuous variables implicitly a model is fitted that assumes that the effects are constant over a wide range, that is, over $X \leq c$ and $X > c$ given the previous splitting. In contrast, our penalization approach assumes a parametric model for DIF. Although it can be extended to a model with unspecified functional form, in the present version it is parametric. An advantage of parametric models is that the essential information is contained in a modest number of parameters that show which variables are influential for specific items. A disadvantage of any parametric model is that it can be misspecified. The partitioning approach, considered as a more exploratory tool, is less restrictive, although assuming a constant value over wide ranges is also a restriction.

An advantage of the parametric model, if it is a fair approximation to the underlying structure, is the use of familiar forms of the predictor, namely a linear predictor, which, of course, can include interactions. In contrast, partitioning methods strongly focus on interactions. Typically in each consecutive layer of

the tree a different variable is used in splitting. The result is smaller and smaller subpopulations which are characterized as a combination of predictors. The subpopulations $\{female, spon \leq 1, age \leq 21\}$ and $\{male, spon \leq 2-3, age \leq 22\}$, found for the knowledge data seem rather specific.

A potential disadvantage of tree based methods is their instability. A small change of data might result in quite different splits. That is the reason why tree-based methods have been extended to random trees, which are a combination of several trees on the same data set, see Breiman (2001).

The penalty approach uses an explicit model for DIF, and the model is separated from the estimation procedure. In the partitioning approach the model and the fitting are entwined. For practioners it is often helpful to have an explicit form of the model that shows how parameters determine the modelled structure. Moreover, in the penalty approach an explicit criterion is used to determine how many and which items show DIF. The ability to identify the right items has been evaluated in the previous section.

Of course, none of the models is true. Neither is the effect constant within an interval of age as assumed in the partitioning approach nor is the effect linear as assumed in the suggested model. But, as attributed to Box, although all models are wrong some can be useful. Since the models are not nested a goodness-of-fit tests could yield a decision. But goodness-of-fit as a measure for the adequacy of a model is a tricky business in partitioning models as well as in regularized estimation procedures, in particular in the framework of item response models. Therefore, not much is available in terms of goodness-of-fit, although it might be an interesting topic of future research.

One basic difference seems to be that the penalty approach uses all covariates, with the variables that are of minor relevance obtaining small estimates, but selects items. The partitioning approach selects variables, or, more concisely combinations of covariates, but then estimates all items as having an effect, that is, estimates are unequal zero. Thus penalty approaches focus on the selection of items, partitioning methods on the selection of combinations of covariates.

7 Concluding Remarks

A general model for DIF that is induced by a set of variables is proposed and estimation procedures are given. It is shown that the method is well able to identify items with DIF. The concept is general, with modifications it can be extended to models that include items with more than two categories as, for example, the graded response model (Samejima, 1997) or the partial credit model (Masters, 1982). Also the assumption that items are modified in the linear form $\boldsymbol{x}_p^T \boldsymbol{\gamma}_i$ can be relaxed to allow for additive functions $f_1(x_{p1}) + \cdots + f_m(x_{pm})$ by using, for example, P-spline methodology (Eilers and Marx, 1996).

The estimation used here is penalized unconditional ML estimation. Alterna-

tive regularized estimators could be investigated, for example, estimators based on mixed models methodology. Also the regularization technique can be modified by using boosting techniques instead of penalization.

The results shown here were obtained by an R program that is available from the authors. It uses the the coordinate ascent algorithm proposed in Meier et al. (2008) and the corresponding R package `grplasso` (Meier, 2009). Currently we are developing a faster program that is based on more recently developed optimization techniques, namely the fast iterative shrinkage thresholding algorithm (Beck and Teboulle, 2009).

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Andersen, E. (1973). A goodness of fit test for the rasch model. *Psychometrika 38*, 123–140. 10.1007/BF02291180.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Im. Sci. 2*(1), 183–202.

Bondell, H., A. Krishna, and S. Ghosh (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, 1069–1077.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and J. C. Stone (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science 22*, 477–505.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and Penalties. *Statistical Science 11*, 89–121.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning (Second Edition)*. New York: Springer-Verlag.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Holland, W. and H. Wainer (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates.

Mair, P. and R. Hatzinger (2007, 2). Extended rasch modeling: The erm package for the application of irt models in r. *Journal of Statistical Software 20*(9), 1–20.

Mair, P., R. Hatzinger, and M. J. Maier (2012). *eRm: Extended Rasch Modeling*. R package version 0.15-0.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika 47*(2), 149–174.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models (Second Edition)*. New York: Chapman & Hall.

Meier, L. (2009). *grplasso: Fitting user specified models with Group Lasso penalty*. R package version 0.4-2.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B 70*, 53–71.

Millsap, R. and H. Everson (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement 17*(4), 297–334.

Ni, X., D. Zhang, and H. H. Zhang (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics 66*, 79–88.

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Applied Statistics 40*, 133–141.

Osterlind, S. and H. Everson (2009). *Differential item functioning*, Volume 161. Sage Publications, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rogers, H. (2005). Differential item functioning. *Encyclopedia of Statistics in Behavioral Science*.

Samejima, F. (1997). Graded response model. *Handbook of modern item response theory*, 85–100.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Soares, T., F. Gonçalves, and D. Gamerman (2009). An integrated bayesian model for dif analysis. *Journal of Educational and Behavioral Statistics 34*(3), 348–377.

Strobl, C., J. Kopf, and A. Zeileis (2010). A new method for detecting differential item functioning in the Rasch model. Technical Report 92, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Strobl, C., J. Malley, and G. Tutz (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods 14*, 323–348.

Thissen, D., L. Steinberg, and H. Wainer (1993). Detection of differential item functioning using the parameters of item response models. *Differential item functioning*, 67–113.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Tutz, G. and H. Binder (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics 62*, 961–971.

Van den Noortgate, W. and P. De Boeck (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics 30*(4), 443–464.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.

Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics 35*(5), 2173–2192.

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*.