Moritz Maximilian Berger

On the Detection of Latent Structures in Categorical Data

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 01. Juni 2016

Moritz Maximilian Berger

On the Detection of Latent Structures in Categorical Data

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 01. Juni 2016

Erster Berichterstatter: Prof. Dr. Gerhard Tutz
Zweiter Berichterstatter: Prof. Dr. Martin Spieß

Tag der Disputation: 02. August 2016

Diese Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der LMU München. Ich danke allen, die durch ihre Unterstützung zur Entstehung dieser Arbeit beigetragen haben. Besonders danke ich ...

... meinem Doktorvater Prof. Dr. Gerhard Tutz für die enge Zusammenarbeit, die zahlreichen, produktiven Diskussionen und sein großes Engagement.

... Prof. Dr. Martin Spieß, der sich freundlicherweise bereit erklärt hat die Aufgabe des Zweitgutachters zu übernehmen.

... Prof. Dr. Thomas Augustin für die gute Zusammenarbeit in der Lehre und den Vorsitz in der Promotionskomission.

... meinen aktuellen und ehemaligen Kolleginnen und Kollegen am Seminar für angewandte Stochastik für die gute Zusammenarbeit und die angenehme Arbeitsatmosphäre. Im Speziellen danke ich Gunther und Micha für die konstruktiven Diskussionen über Differential Item Functioning und Response Styles.

... allen Kolleginnen und Kollegen am Institut für Statistik für zahlreiche Mittagspausen und gemeinsame Aktivitäten neben der Arbeit.

... meinen Freunden und meiner Familie für deren uneingeschränkte Unterstützung. In erster Linie danke ich meinen Eltern, die mir das Studium der Statistik ermöglicht haben und mir stets einen starken Rückhalt bieten.

Zusammenfassung

Durch die steigende Verfügbarkeit großer Datenmengen wird es zunehmend wichtiger, die zugrundeliegenden Strukturen in den Daten aufzudecken. Die vorliegende Arbeit beschäftigt sich mit der Erfassung latenter Strukturen in kategorialen Daten. In Regressionsmodellen fungieren kategoriale Variablen entweder als abhängige oder als Teil der unabhängigen Variablen. Je nach Konstellation sind unterschiedliche Strategien notwendig, um die zugrundeliegenden Strukturen zu erfassen. Der erste Teil dieser Arbeit widmet sich Regressionsmodellen mit einer überproportional großen Zahl an Parametern. Genauer werden Modelle mit kategorialen Einflussgrößen und einer großen Anzahl an Kategorien betrachtet. Außerdem werden Modelle für Messwiederholungen mit festen Effekten untersucht. Von Interesse ist hierbei, zu identifizieren, welche „latente Gruppen“ von Kategorien bzw. Beobachtungseinheiten denselben Effekt auf die abhängige Variable aufweisen. Zur Identifizierung dieser Gruppen wird ein neuartiger Ansatz vorgestellt, der auf rekursiver Partitionierung basiert. Im Gegensatz zu konkurrierenden Methoden, die bestimmte Penalisierungsterme verwenden, ist die vorgeschlagene Methodik auch auf sehr hochdimensionale Probleme anwendbar. Der zweite Teil dieser Arbeit beschäftigt sich mit Item-Response Modellen, das heißt mit Regressionsmodellen zur Messung „latenter Fähigkeiten“ von Personen. Die Item-Response-Theorie verwendet Indikatoren, wie die Antworten von Personen auf bestimmte Testitems, um auf deren Fähigkeit zu schließen. Ein Phänomen, dessen man sich in psychologischen Tests bewusst sein sollte, ist das sogenannte Differential Item Functioning (DIF). DIF tritt auf, falls die Schwierigkeit eines Items für Personen mit derselben Fähigkeit von deren Charakteristika, wie Geschlecht oder Herkunft, abhängt. Ein auf rekursiver Partitionierung basierendes Verfahren wird vorgeschlagen, das eine simultane Bestimmung der von DIF betroffenen Items im Bezug auf eine beliebige Anzahl an Kovariablen ermöglicht. Einer der Vorteile gegenüber klassischen Ansätzen ist, dass die vorgeschlagene Methodik diejenigen Regionen im Kovariablenraum indentifiziert, die DIF verursachen, ohne, dass diese vorher definiert werden müssen. Desweiteren wird eine Erweiterung für ungleichmäßiges DIF entwickelt. Der letzte Teil der Arbeit befasst sich mit Regressionsmodellen für Bewertungsskalen, die häufig in der Verhaltensforschung Anwendung finden. Hierbei kann Heterogenität unter den Befragten durch „latente Antwortstile“ zu verzerrten Schätzungen und irreführenden Interpretationen der beobachteten Antworten führen. Die vorliegenden Analysen beschränken sich auf Skalen mit symmetrischen Antwortkategorien und einem spezifischen Antwortstil, nämlich der Tendenz zur mittleren oder extremen Kategorien. Eine stärkere oder schwächere Konzentration in der Mitte kann in ordinalen Regressionmodellen auch als Dispersionsabweichung interpretiert werden. Die Stärke der vorgeschlagenen Modellen ist, dass sie in das Framework der generalisierten linearen Modelle eingebettet werden können und somit Inferenztechnicken und asymptotische Ergebnisse für diese Klasse von Modellen zur Verfügung stehen. Darüber hinaus wird ein Visualisierungstool entwickelt, das die Interpretation der Effekte leicht zugänglich macht.

Summary

With the growing availability of huge amounts of data it is increasingly important to uncover the underlying data generating structures. The present work focusses on the detection of latent structures for categorical data, which have been treated less intensely in the literature. In regression models categorical variables are either the responses or part of the covariates. Alternative strategies have to be used to detect the underlying structures. The first part of this thesis is dedicated to regression models with an excessive number of parameters. More concrete, we consider models with various categorical covariates and a potentially large number of categories. In addition, it is investigated how fixed effects models can be used to model the heterogeneity in longitudinal and cross-sectional data. One interesting aspect is to identify the categories or units that have to be distinguished with respect to their effect on the response. The objective is to detect "latent groups" that share the same effects on the response variable. A novel approach to the clustering of categorical predictors or fixed effects is introduced, which is based on recursive partitioning techniques. In contrast to competing methods that use specific penalties the proposed algorithm also works in high-dimensional settings. The second part of this thesis deals with item response models, which can be considered as regression models that aim at measuring "latent abilities" of persons. In item response theory one uses indicators such as the answers of persons to a collection of items to infer on the underlying abilities. When developing psychometric tests one has to be aware of the phenomenon of Differential Item Functioning (DIF). An item response model is affected by DIF if the difficulty of an item among equally able persons depends on characteristics of the persons, such as the membership to a racial or ethnic subgroup. A general tree-based method is proposed that simultaneously detects the items and subgroups of persons that carry DIF including a set of variables on different scales. Compared to classical approaches a main advantage is that the proposed method automatically identifies regions of the covariate space that are responsible for DIF and do not have to be prespecified. In addition, extensions to the detection of non-uniform DIF are developed. The last part of the thesis addresses regression models for rating scale data that are frequently used in behavioural research. Heterogeneity among respondents caused by "latent response styles" can lead to biased estimates and can affect the conclusion drawn from the observed ratings. The focus is on symmetric response categories and a specific form of response style, namely the tendency to the middle or extreme categories. In ordinal regression models a stronger or weaker concentration in the middle can also be interpreted as varying dispersion. The strength of the proposed models is that they can be embedded into the framework of generalized linear models and therefore inference techniques and asymptotic results for this class of models are available. In addition, a visualization tool is developed that makes the interpretation of effects easy accessible.

Contents

1. Introduction 1

2. Structured Regression Models for Categorical Predictors 13

2.1. Introduction 13

2.2. Structured Predictors 15

2.3. Tree-Structured Clustering 16

2.4. Standard Errors and Stability of Clusters 24

2.5. Related Approaches 27

2.6. Simulations 30

2.7. Further Applications 37

2.8. Possible Extensions 43

2.9. Concluding Remarks 45

3. Modelling Heterogeneity in Fixed Effects Models 47

3.1. Introduction 47

3.2. Accounting for Heterogeneity in Clustered Data 49

3.3. An Illustrative Example 52

3.4. Fitting Procedure 55

3.5. Related Approaches 59

3.6. Simulations 60

3.7. Further Applications 68

3.8. Extension to Group-Specific Slopes 72

3.9. Concluding Remarks 75

4. Identification of Differential Item Functioning in Rasch Models 77

4.1. Introduction 77

4.2. Item Focussed Recursive Partitioning 78

4.3. Fitting Trees 87

4.4. Simulations 91

4.5. Further Application 103

4.6. Extension to Ordinal Item Responses 106

4.7. Concluding Remarks 109

5. Detecting Uniform and Non-Uniform DIF by Logistic Regression 111

5.1. Introduction 111

5.2. Logistic Regression Approaches to DIF 113
5.3. An Illustrative Example 116
5.4. Fitting Procedure 118
5.5. Simulations 120
5.6. Investigation of Non-Uniform DIF 129
5.7. Empirical Applications 136
5.8. Concluding Remarks 142

6. Modelling of Extreme Response Styles in Rating Scales 143

6.1. Introduction 143
6.2. An Extended Rating Scale Model 145
6.3. Effects in the RSRS Model 154
6.4. Estimation of Parameters and Inference 161
6.5. Implementation and Available Programs 162
6.6. Further Applications 163
6.7. Extensions and Comparison with Alternative Approaches 166
6.8. Response Styles for Several Items 169
6.9. Concluding Remarks 170

7. Varying Dispersion in Cumulative Regression Models 173

7.1. Introduction 173
7.2. Separating Location and Dispersion 175
7.3. Inference and Computation of Estimates 183
7.4. Application: Confidence in the Federal Government 184
7.5. Comparison of Models and Consequences of Ignored Dispersion 186
7.6. Non-Symmetric Responses 189
7.7. Partial Proportional Odds Models versus the Modelling of Dispersion 191
7.8. Concluding Remarks 195

8. Conclusion and Outlook 197

Appendices 203

A. Overview on Variables of the Applications in Chapter 2 205

B. Tabular Display of Simulation Results for Chapter 3 207

C. Additional Simulation Results for Chapter 5 213

D. Supplement Exemplary R Code for Chapter 6 219

References 227

1. Introduction

A huge amount of data is collected in many areas of applied science for statistical analyses. The main cause is the development of computer technology and the accompanying possibilities of data processing. In many applications regression models are used to describe the relation between a dependent variable of interest (called response) and several explanatory variables (called predictors). If many variables are available on different scales, for example, a mixture of continuous and categorical variables, one has to carefully select the variables that are incorporated in the analysis. In particular one has to decide in which form they are included in the model. Categorical variables that can be ordered or unordered are typically difficult to handle, as they require specific coding. Moreover, it is often the case that variables are not directly observable or that the true underlying data generating structure is only captured implicitly by observed variables. These variables or structures are called "latent". The present thesis deals with modelling strategies for the detection of the following latent structures in categorical variables:

- Latent groups comprising categories of categorical predictors or measurement units that share the same effect on the response (subject in Chapter 2 and 3).
- Latent traits that are measured in psychometric modelling, where the answers to a set of items are used to infer on the underlying abilities (subject in Chapter 4 and 5).
- Latent response styles that affect the response behavior and therefore the conclusions drawn from the observed ratings in behavioral research (subject in Chapter 6 and 7).

In each case tailored regression models have to be used to detect the underlying latent structures. All proposed approaches in this thesis are based on generalized linear models, see McCullagh and Nelder (1989). In the following some basic concepts are shortly introduced on which the proposed methods are built on.

Categorical Predictors

In regression models categorical variables require special attention. Unlike continuous variables they have to be appropriately recoded into several variables. For a categorical variable $x \in \{1, \ldots, m\}$ the most popular way is to define $m-1$ dummy variables $\tilde{x}_j$, where $\tilde{x}_j = 1$ if

$x = j$, and $\tilde{x}_j = 0$ otherwise. Thereby, each category is compared to a predefined reference and the regression coefficients are interpreted accordingly. An overview on this topic is found in Tutz (2012).

The incorporation of categorical variables inevitably leads to a huge number of parameters in the linear predictor which calls for regularization techniques. Classical approaches are group or fused lasso type penalties, see Tibshirani et al. (2005), Yuan and Lin (2006) and Bondell and Reich (2009). The objective of the use of such penalties for categorical predictors is the exclusion of the predictor from the model or the grouping of categories with the same impact on the response. The main drawback of existing approaches is that they are computationally expensive and become infeasible for a large number of categories. To overcome these problems in Chapter 2 of this thesis a novel approach for regularized modelling of categorical predictors based on recursive partitioning techniques is carefully developed and compared to its competitors.

Categorical Responses

In the simplest case one has a binary response coding two categories $y \in \{0, 1\}$. The most popular regression model for binary responses is the logistic regression model (in short logit model) that links the conditional expectation of the response to the linear predictor by the logistic distribution function, so

$$log\left(\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}\right) = \boldsymbol{x}^\top\boldsymbol{\beta},$$

where $\eta = \boldsymbol{x}^\top\boldsymbol{\beta}$ is the linear predictor composed of explanatory variables $\boldsymbol{x}$ and corresponding coefficients $\boldsymbol{\beta}$. A main advantage of the model is the easy interpretation of effects by odds ratios. Alternative link functions, which will not be considered in this thesis, are the probit or complementary log-log link. An introduction on binary regression models is found in Fahrmeir et al. (2013). In this thesis the logit model is used in various ways. In particular, in Chapter 4 and 5 it is employed to develop extended models for item response data.

The most popular model for the analysis of item response data is the Rasch model (Rasch, 1960). Let the response be given by Y_{pi}, which indicates if respondent p, $p = 1, \dots, P$, solved item i, $i = 1, \dots, I$, correctly or not. Then the Rasch model is given by

$$log\left(\frac{P(Y_{pi}=1|\theta_p,\beta_i)}{1-P(Y_{pi}=1|\theta_p,\beta_i)}\right) = \theta_p - \beta_i,$$

where θ_p denotes the ability of respondent p and β_i denotes the difficulty of item i. The model simply represents a binary logit model. Therefore, by the choice of appropriate assumptions the model can be embedded into the framework of generalized linear models. A basic introduction into the Rasch model is found in Strobl (2012).

More general one has a response $Y \in \{1, \ldots, k\}$ with k categories. A classical approach is the multinomial logit model. The model is locally a binary logit model specifying the odds between category r, $r = 1, \ldots, k-1$, and a predefined reference. If the response is ordinal, i.e., the categories have a natural order, it is advisable to use models that explicitly make use of this information. An useful choice is the cumulative logit model that specifies the cumulative probabilities $P(Y \leq r) = P(Y = 1) + \ldots + P(Y = r)$ by the logistic distribution function. A representation of the model is

$$log\left(\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})}\right) = \eta_r = \theta_r + \boldsymbol{x}^\top \boldsymbol{\beta}_r, \quad r = 1, \ldots, k-1,$$

where θ_r denote category-specific threshold parameters and $\boldsymbol{\beta}_r$ are category-specific regression coefficients. One drawback of the model is that the model requires the ordering of predictors $\eta_1 \leq \ldots \leq \eta_{k-1}$ and therefore constraints on the parameters are needed. An alternative is the adjacent categories logit model that specifies the odds of adjacent categories $r+1$ and r as

$$log\left(\frac{P(Y = r+1|\boldsymbol{x})}{P(Y = r|\boldsymbol{x})}\right) = \eta_r = \theta_r + \boldsymbol{x}^\top \boldsymbol{\beta}_r, \quad r = 1, \ldots, k-1.$$

A common assumption that is made to result in a parsimonious parameterization in both models is $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_{k-1}$. For further details see Agresti (2009) or Tutz (2012). In this thesis both models for ordinal responses, the cumulative and the adjacent categories model, are used in various ways. In particular, in Chapter 6 and 7 they are employed to develop extended models for rating scale data.

Recursive Partitioning

An alternative to linear or additive regression models are recursive partitioning techniques, also known as trees. A main advantage of trees is that interactions in particular of higher order can easily be modelled by successive splitting of the predictor space. The concept goes back to automatic interaction detection (AID) introduced by Morgan and Sonquist (1963). In general tree-based methods may be divided into two groups - methods that use binary splits and methods that yield trees with multiway splits. Examples for the latter are the C4.5 (Quinlan, 1993) algorithm, the successor and refinement of ID3 (Quinlan, 1986), and CHAID (Kass, 1980). Multiway splits offer the advantage that a variable is rarely used for splitting several times and therefore does not appear more than once in the tree. However binary trees are usually preferred because a multiway split, for example in an ordinal variable, can also be achieved by successive binary splitting. The most popular approaches are classification and regression trees (CARTs) proposed by Breiman et al. (1984). For an introduction into the basic concepts see Hastie et al. (2009) and Tutz (2012).

In each step of the tree construction a node A, that is a subset of the predictor space, is split into a left node A_1 and a right node A_2 corresponding to disjoint subsets of A. Each split is determined by one variable and one corresponding split-point that has to be chosen appropriately. After several splits the terminal nodes describe a partition of the predictor space. The tree yields an interpretable structure of the relation between the predictors and the response. In each terminal node the predicted outcome is a constant that depends on the scale of the response. For a continuous response it is simply the mean in the respective region.

The construction of the split also depends on the scale of the variable. For a continuous or ordinal variable x and chosen split-point c, the partition $\{A_1, A_2\}$ has the form

$$A_1 = A \cap \{x \leq c\}, \quad A_2 = A \cap \{x > c\}.$$

For a categorical variable without ordering $x \in \{1, \ldots, K\}$, the partition has the form

$$A_1 = A \cap S_1, \quad A_2 = A \cap S_2,$$

where S_1 and S_2 are disjoint, non-empty subsets $S_1 \subset \{1, \ldots, K\}$ and its complement $S_2 = \{1, \ldots, K\} \setminus S_1$. There are $2^{K-1} - 1$ possible pairs S_1, S_2 that have to be considered when searching for the optimal split.

For the selection of splits several criteria have been proposed. A classical way is to use impurity measures as the Gini index or the entropy and to select the split that maximally decreases the impurity of the tree. An alternative that is used in this thesis are test-based splits. In each iteration one yields a model for the conditional mean $E(y|\boldsymbol{x})$ that is associated with the current tree structure. The model assumes that the response is constant within already built subsets. To select the best split one evaluates the improvement of the model fit by use of a measure for the goodness-of-fit. A common choice is to use the difference in deviances

$$d = D(M_A) - D(M_{A_1,A_2}),$$

where $D(M_A)$, $D(M_{A_1,A_2})$ denote the deviances of the models with and without the split, and to select the split for which d is maximal.

Finally one has to determine the size of the tree. One strategy is to grow a very large tree and to prune it to an adequate size afterwards. A second strategy, which is used in this thesis, is to stop growing the tree if a certain splitting criterion is no longer met. Thus,

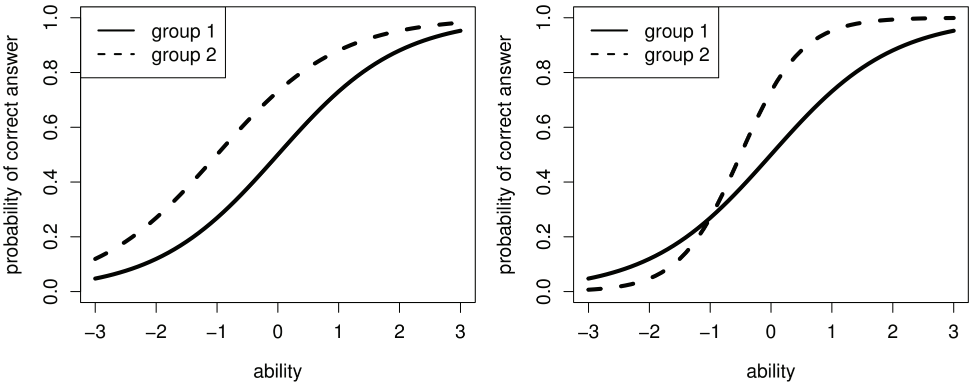


Figure 1.1.: Exemplary item characteristic curves for an item with uniform (left) and non-uniform (right) DIF with regard to two groups.

the tree size is determined beforehand by early stopping. Applied stopping criteria are explained in more detail in the respective chapters.

In Chapter 2 to 5 of this thesis tree-based splits are used to extend generalized linear models. The adaptation to specific problems result in models with higher flexibility. It is important to note that in the proposed approaches there is one main difference to the fitting of common trees. Tree-based splits are embedded into the linear or additive predictor of regression models. Thus, they are only part of the whole model. To ensure valid estimates of all parameters all data are used in each fitting step. In contrast, common trees condition on previous splits and only use the data of already built subsets of the predictor space to determine the next split.

Differential Item Functioning

Intelligence and other achievement tests aim at measuring latent abilities or traits of persons. As they are not directly observable the answers on a collection of items are used to infer on the underlying ability of the person. To draw valid conclusions it is necessary to design the tests very carefully. In particular test items should not be unfair, that is, should not favour specific groups. If the probability to answer an item correctly is different among persons with the same latent ability, it is referred to item bias or differential item functioning (DIF). For a detailed introduction to DIF, see Holland and Wainer (1993) or Osterlind and Everson (2009). DIF is often caused by certain characteristics of the persons, such as the membership to a racial or ethnic subgroup. In the previous literature this topic has been dealt with extensively. An overview on existing methods is given in Millsap and Everson (1993). In current research several approaches have been proposed, for example by Strobl et al. (2015), Tutz and Schauberger (2015) and Magis et al. (2015).

In general DIF may be divided into two main types - uniform and non-uniform DIF. Uniform DIF is present, if the differences of the probability to answer an item correctly between different groups does not depend on the ability of the persons. If non-uniform DIF is present, the differences differ across persons depending on their ability level. A visualization of the probabilities to answer an item correctly as a function of the ability of a person is shown in Figure 1.1. The left panel shows the so-called item characteristic curves for two groups with uniform DIF. In this example the item is more difficult for all persons in group 2. In contrast, in the right panel the two groups show non-uniform DIF. It can be seen, that one obtains crossing item response curves. For persons with low ability the item is easier for group 2 than for group 1 and vice versa for persons with higher ability level. In Chapter 4 and 5 of this thesis regression models that capture uniform as well as non-uniform DIF in a very flexible way are developed and compared to their competitors.

Guideline through the Thesis

This thesis can be divided into three main parts, which are dedicated to the detection of latent structures in different forms. Each part is contained of two chapters.

Chapter 2 and Chapter 3 deal with regression models containing an excessive number of parameters, which calls for structured modelling approaches. In many applications one has a variety of potential explanatory variables, in particular several categorical predictors on an ordinal or nominal scale. In both forms the simple use of dummy variables for each category will cause estimation problems and probably will not reflect the true impact of the predictors on the response. To gain interpretability one wants to exclude non-influential variables and wants to know which categories have to be distinguished. The focus is on the detection of latent groups of categories that share the same effect on the response.

In Chapter 2 a novel approach for the clustering of categories in regression models using tree-based splits is proposed. Previous methods for the fusion of categorical predictors proposed by Gertheiss and Tutz (2009) and Gertheiss and Tutz (2010) are based on penalized maximum likelihood estimation. An overview on this topic was recently given by Tutz and Gertheiss (2016). The main problem of these approaches is that they are not applicable for a large number of categories due to the computational effort. In addition, simulations show that the proposed tree-based approach yields much better results in terms of its clustering performance. In Chapter 2 several applications and further comparisons to competing approaches underline the usefulness and the applicability of the proposed method.

In Chapter 3 the tree-structured modelling approach developed in Chapter 2 is adapted to models for repeated measurements. In longitudinal or cross-sectional studies the heterogeneity of measurement units has to be taken into account. A classical solution that is widely used to model unobserved heterogeneity is the random effects model. Despite its popularity the random effects model often causes problems because strong assumption are made on the generation of the observed data. A more flexible alternative, which is proposed in Chapter 3, is the fixed effects model. In fixed effects models heterogeneity is captured by own parameters for each measurement unit defining appropriate dummy variables. Obviously this can be seen as a special case of categorical predictors. For example unit-specific intercepts can be treated as the parameters of a nominal variable. Again the huge number of parameters raises the questions of computational feasibility and interpretability of the model. Furthermore, the assumption that all measurement units behave different is quite strong. In repeated measurements one wants to know if heterogeneity is present at all, and if it is, to identify latent groups of measurement units that share the same effect on the response. The proposed method is illustrated in several applications and in extensive simulations including the comparison to competing methods.

In **Chapter 4 and Chapter 5** a novel method for the detection of differential item functioning based on recursive partitioning is proposed. Classical testing approaches for the identification of items that carry DIF are restricted to the comparison of two or few subgroups that have to be pre-specified. In particular in the case of continuous covariates it might be challenging to determine the relevant groups that should be investigated. An alternative approach that is also based on recursive partitioning and is able to handle several covariates was recently proposed by Strobl et al. (2015). The main drawback of the method is that it detects the subsets of the predictor space that carry DIF but does not automatically detect the items that are responsible. The methods proposed here combine the two desirable criteria. By recursive partitioning on the item level one achieves a simultaneous detection of DIF items and corresponding subgroups that do not have to be pre-specified.

The most popular model of the item response theory (IRT) is the Rasch model (Rasch, 1960), introduced before. It assumes that the probability to answer an item correctly is determined by exactly two parameters - the ability of the person and the difficulty of the item. Due to the simple form of the model it can only capture uniform DIF. The method of recursive partitioning on the item level, called item focussed trees, is developed for the Rasch model in Chapter 4. The advantages towards competing methods and the good performance are demonstrated in several simulations and two applications.

An alternative non-IRT approach to the detection of DIF was proposed by Swaminathan and Rogers (1990) and extended by Magis et al. (2011) and Magis et al. (2015). The main idea is to use the test score, i.e. the number of solved items, and the group membership of the persons as predictors of a logistic regression model that models the probability of solving an item correctly. The structure of the model allows to investigate uniform as well as non-uniform DIF. In Chapter 5 the logistic regression model is incorporated into the framework of item focussed trees. In particular the investigations on non-uniform DIF show the potential of the method.

Chapter 6 and Chapter 7 are dedicated to ordinal regression models used in behavioural research. In many studies rating scales are employed to investigate attitudes or performance of the participants. When evaluating the observed ratings one should always be aware of specific response styles. Observed ratings caused by a certain response pattern that is independent of the content of the response may lead to wrong conclusions. This thesis focusses on extreme response styles, that is the tendency to the middle or extreme categories.

In Chapter 6 the adjacent categories model is extended by the introduction of an additional parameter that determines the response style. The additional response style parameter can be specified as a function of explanatory variables. The proposed method is quite different from alternative IRT based approaches in which latent traits are used and multiple items are necessary. The strength of the model is that it simultaneously accounts for content-related and response style effects. By embedding the proposed model into the framework of (multivariate) generalized linear models established estimation and inference tools can be used. Simulations illustrate that biased estimates of the content-related effects can be avoided by accounting for the response style. In addition, a visualization tool is developed that makes the interpretation of effects easily accessible. Several applications demonstrate its applicability.

A strong tendency to the middle or extreme categories can also be seen as varying dispersion. In many applications a lack-of-fit is caused by an insufficient modelling of dispersion effects. In Chapter 7 the cumulative regression model is extended by an additional term that determines the dispersion. The design is very similar to the model in Chapter 6. However, the parameters are interpreted as location and dispersion effect. In simulations and applications the proposed model shows a very similar performance to an alternative model that was introduced by McCullagh (1980). Embedding the model into the framework of generalized linear models allows to use asymptotic results that have been developed for this class of models. Moreover, selected examples show that the extended cumulative model with dispersion effects is an parsimonious alternative to cumulative models with a huge number of category-specific parameters.

Apart from some cross-references, each chapter is self-contained including an own introduction to the relevant topics and can therefore be read separately.

Contributing Manuscripts

Parts of this thesis have been published as articles in peer reviewed journals, in proceedings of scientific conferences, as preprints on arXiv hosted by Cornell University or as technical report at the Department of Statistics of the Ludwig-Maximilians-Universiät München. In the following, chapter by chapter all contributing manuscripts are listed and the contributions of the respective authors are described.

- **Chapter 2:**
  Tutz and Berger (2015b): Tree-Structured Modelling of Categorical Predictors in Regression, *Cornell University Library*, arXiv: 1504.04700.

  The project was set up by Gerhard Tutz and developed jointly by Gerhard Tutz und Moritz Berger. Moritz Berger implemented the method and conducted the simulations and real data analyses. The manuscript was written in close collaboration by both authors.
  The Chapter is a modified version of Tutz and Berger (2015b). The manuscript was extended by the simulations in Section 2.6.1 and by Section 2.8 which introduces further-reaching concepts. Some parts were rewritten and the notation was slightly changed. In addition, the application in Section 2.7.1 differs from the original one. Appendix A contains some supplementary material.

- **Chapter 3:**
  Berger and Tutz (2015c): Tree-Structured Clustering in Fixed Effects Models, *Cornell University Library*, arXiv: 1512.05169.

  Gerhard Tutz initiated the use of tree-based methods in fixed effects models. Moritz Berger was responsible for the implementation of the method and the simulation studies as well as the applications on real data. The manuscript was mainly written by Moritz Berger in close collaboration with Gerhard Tutz.
  The original manuscript was extended by further considerations in Section 3.4, by the application in Section 3.7.1 and Section 3.8, which deals with group-specific slopes. Apart from these sections and minor modifications, Chapter 3 together with Appendix B and Berger and Tutz (2015c) match.

- **Chapter 4:**
  Tutz and Berger (2015a): Item focussed Trees for the Identification of Items in Differential Item Functioning, *Psychometrika*, published online, doi: 10.1007/s11336-015-9488-3.

  Chapter 4 was set up by Gerhard Tutz who conceptualized the theoretical framework. Moritz Berger implemented the method and the corresponding `R` package `DIFtree`. He also evaluated the simulation studies and real data examples. The manuscript was written in close collaboration by both authors.
  The original manuscript was complemented by simulations in Section 4.4.3 and Section 4.4.4 and extended by Section 4.6, which deals with ordinal item responses. Apart

from these sections and minor modifications, Chapter 4 and Tutz and Berger (2015a) match.

- **Chapter 5:**
  Berger and Tutz (2015a): Detection of Uniform and Non-Uniform Differential Item Functioning by Item Focussed Trees, *Cornell University Library*, arXiv: 1511.07178.

  The project was jointly developed by the two authors. Moritz Berger implemented the method and conducted the simulations and applications on real data. He mainly wrote the manuscript in close collaboration with Gerhard Tutz.
  The chapter is a revised version of Berger and Tutz (2015a). The manuscript was extended by several simulations in Chapter 5.5 and Chapter 5.6.5. Moreover, some parts were rewritten and further considerations were added. Appendix C contains additional simulation results.

- **Chapter 6:**
  Tutz and Berger (2016a): Response Styles in Rating Scales - Simultaneous Modelling of Content-Related Effects and the Tendency to Middle or Extreme Categories, *Journal of Educational and Behavioral Statistics 41(3)*, 239-268.

  The project was initiated by Gerhard Tutz who developed the theoretical framework and investigated the literature. Moritz Berger was responsible for the implementation of the method and the evaluation of numerical experiments as well as real data examples. The manuscript was mainly written by Gerhard Tutz with contributions of Moritz Berger.
  The original manuscript was complemented by simulations in Section 6.3.1 and extended by Section 6.8, which introduces further-reaching concepts. Apart from these sections and some minor modifications, Chapter 6 together with Appendix D and Tutz and Berger (2016a) match. Preliminary work on the project can be found in the proceedings of the IWSM 2015 (Berger and Tutz, 2015b).

- **Chapter 7:**
  Tutz and Berger (2016b): Seperating Location and Dispersion in Ordinal Regression Models, *Ludwig-Maximilians-Universität München, Department of Statistics*, Technical Report 190.

  Chapter 7 was mainly drafted by Gerhard Tutz with contributions of Moritz Berger. Moritz Berger conducted several simulation studies and applications on real data. He contributed substantially to the presentation of the results.
  Apart from some modifications, particularly regarding the notation, and the arrangement of the sections, Chapter 7 and Tutz and Berger (2016b) match.

Software

All computations were done with the statistical program `R` (R Core Team, 2016) and additional packages. The corresponding packages are indicated in the respective chapters and sections.

The methods and functions implemented for Chapter 2 and Chapter 3 are made available by the self-implemented R add-on package `structree` (Berger, 2016b), which will presumably be made publicly accessible via the Comprehensive R Archive Network (CRAN). An initial version of the package can be downloaded from `http://www.statistik.lmu.de/~mberger/forschung.html`. The package imports the two `R` add-on packages `mgcv` (Wood, 2011) and `penalized` (Goeman et al., 2014).

For the methods proposed in Chapter 4 and Chapter 5 the self-implemented `R` add-on-package `DIFtree` (Berger, 2016a) was developed, which can be downloaded from CRAN. It imports the two `R` add-on packages `penalized` (Goeman et al., 2014) and `plotrix` (Lemon, 2006).

The methods for Chapter 6 and Chapter 7 are implemented by use of the `R` add-on-package `VGAM` (Yee, 2010), which attaches the two base packages `splines` and `stats4`. Embedding the estimation procedure into the framework of `VGAM` ensures quite fast computation. The corresponding functions are available upon request. Moreover, for illustration Appendix D contains parts of the implemented `R` code.

2. Structured Regression Models for Categorical Predictors

2.1. Introduction

In most regression problems one has a mixture of explanatory variables. Some are continuous, some are binary and others are categorical on a nominal scale or ordered categorical. Flexible models with a focus on main effects are generalized additive models (GAMs). In particular they allow to include continuous variables that have a smooth effect of unspecified functional form. However, the focus on main effects turns into the disadvantage that higher order interactions are hard to model. Furthermore, generalized additive models can contain a multitude of parameters.

An alternative tool that is widely used is recursive partitioning also known as trees. The most popular methods are classification and regression trees (CART), outlined in Breiman et al. (1984), and the C4.5 algorithm, which was proposed by Quinlan (Quinlan, 1986; Quinlan, 1993). An introduction into the basic concepts is found in Hastie et al. (2009), an overview on recursive partitioning in the health sciences was given by Zhang and Singer (1999) and an introduction including random forests with applications in psychology by Strobl et al. (2009).

One big advantage of trees is that they automatically find interactions. The concept of interactions is at the core of trees, which have its roots in automatic interaction detection (AID), proposed by Morgan and Sonquist (1963). But the focus on interactions can also turn into a disadvantage because common trees do not allow for a linear or smooth component in the predictor. Below the root node most nodes represent interactions. Thus potentially linear or additive effects of covariates are rarely detected. This is in contrast to generalized additive models, which take main effects much more serious.

This chapter is a modified version of Tutz and Berger (2015b). For more information on the personal contributions of the authors and textual matches, see page 9.

One application we will consider are the Munich rent standard data, which were also analysed in Gertheiss and Tutz (2010). The data set consists of 2053 households with the response variable being monthly rent per square meter in Euro. Available predictors are the urban district (nominal factor), the year of construction, the number of rooms, the quality of residential area (ordinal factors), the floor space (metric) and five additional binary variables. Conventional trees treat all these explanatory variables in a similar way. They split the predictor space by use of one variable into two regions. Within the regions the response is fitted as a constant. If in the first step a continuous explanatory variable is selected, for example floor space, in the next step typically interactions with floor space are fitted, more concise, interactions with the two selected regions of floor space. In the next steps all fits refer to higher order interactions. Therefore, trees have a strong tendency to fit interactions and neglect the main effects. The relevance of explanatory variables is found a posteriori by defining importance measures, which in random forests in some form reflect how often a variable has been selected, see, for example, Ishwaran (2007), Sandri and Zuccolotto (2008) and Strobl et al. (2008). In contrast, if in generalized additive models binary and categorical variables are included by use of a linear predictor one obtains estimates of parameters that reflect the importance of the variables directly.

The tree-structured approach proposed in the present chapter combines the advantages of generalized additive models and trees. The method uses trees in part of the variables but allows to include others as parametric or smooth components in the model. Similar approaches have been considered for longitudinal data, see, for example, Sela and Simonoff (2012) and Bürgin and Ritschard (2015). Our focus is on categorical predictors with many categories as, for example, the urban district in the rent data (25 districts). In particular categorical predictors are difficult to handle because for each category one parameter is needed. Thus simple parametric models tend to become unstable which calls for regularized estimates. Categorical predictors or factors come in two forms, unordered or ordered. In both forms one wants to know if the predictor has an impact, and, if it has, which categories have to be distinguished. The latter problem means that one wants to find clusters of categories (or factor levels) that share the same expected response. In the nominal case all possible partitions of the set of categories are candidates, whereas in the ordered case clusters are formed by fusion of adjacent categories. The proposed method uses trees to find the clusters of factor levels. Thus trees are used for the categorical variables while the other variables are included in the classical form of linear or smooth effects.

Fusion of categories to obtain clusters of categories within a regression model has been mainly investigated by penalization methods, see Bondell and Reich (2009), Gertheiss and Tutz (2009) and Gertheiss and Tutz (2010). However, in contrast to the tree-structured approach, these penalization techniques are restricted to a small number of categories. Penalization methods and tree-type methods that are related or alternatives to the present approach are considered in a separate section (Section 2.5). In Section 2.2 we introduce a tree-structured model for categorical predictors, in Section 2.3 the fitting procedure is presented. Section 2.4 deals with standard errors and the stability of clusters. Results of simulation studies are given in Section 2.6 and in Section 2.7 we consider two further applications. Finally, in Section 2.8 further extensions of the proposed approach are shortly introduced.

2.2. Structured Predictors

As in generalized linear models (GLMs) let the mean response $\mu = \mathrm{E}(y|\boldsymbol{x})$ be linked to the explanatory variables in the form

$$\mu = h(\eta) \quad \text{or} \quad g(\mu) = \eta,$$

where $h(\cdot)$ is the response function and $g(\cdot) = h^{-1}(\cdot)$ is the link function. As in GLMs we also assume that the distribution of $y|\boldsymbol{x}$ follows a simple exponential family (McCullagh and Nelder, 1989). While GLMs always assume that the predictor is linear we assume that the predictor is composed of two components, a tree component and a linear or additive component. For data $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i),\ i = 1, \ldots, n$, the predictor of the model with a linear component has the form

$$\eta_i = tr(\boldsymbol{z}_i) + \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{2.1}$$

where $tr(\boldsymbol{z}_i)$ is the tree component of the predictor and $\boldsymbol{x}_i^T \boldsymbol{\beta}$ is the familiar linear term. Thus, one distinguishes between two groups of explanatory variables, namely $\boldsymbol{z}$, which are determined by a tree, and $\boldsymbol{x}$, which have a linear effect on the response. In extended versions we consider the additive predictor

$$\eta_i = tr(\boldsymbol{z}_i) + \sum_{j=1}^{p} f_{(j)}(x_{ij}), \tag{2.2}$$

where the $f_{(1)}(\cdot), \ldots, f_{(p)}(\cdot)$ are unspecified functions and p is the number of $\boldsymbol{x}$-variables. Then one obtains a tree-structured model with additive components.

We will focus on the case where the $\boldsymbol{z}$-variables are categorical. When a tree is built, successively a node A, that is a subset of the predictor space, is split into subsets with the split determined by only one variable. For a *nominal* categorical variable $z \in \{1, \ldots, K\}$, the partition has the form $A \cap S,\ A \cap \bar{S}$, where S is a non-empty subset $S \subset \{1, \ldots, K\}$ and $\bar{S} = \{1, \ldots, K\} \setminus S$ is the complement. Thus, after several splits the predictor $tr(z_i)$ represents a clustering of the categories $\{1, \ldots, K\}$, and the tree term can be represented by

$$tr(z_i) = \alpha_1 I(z_i \in S_1) + \cdots + \alpha_m I(z_i \in S_m).$$

$S_1, \dots, S_m$ is a partition of $\{1, \dots, K\}$, and $I(\cdot)$ denotes the indicator function with $I(a) = 1$ if a is true, $I(a) = 0$ otherwise.

For an *ordinal* categorical variable $z \in \{1, \dots, K\}$ the partition into two subsets has the form $A \cap \{z \leq c\}, \quad A \cap \{z > c\}$, based on the threshold c on variable z. Thus during the building of a tree clusters of adjacent categories are formed. The tree term has the same form as before but with the subsets that represent the clusters having the form $S_k = \{a_{k-1}, \dots, a_k\}$, $a_{k-1} < a_k$.

In the case of more than one categorical predictor the tree-structured model proposed here forms clusters only for one variable. Then, with q predictors in $\boldsymbol{z}$ the tree component has the form

$$tr(\boldsymbol{z}_i) = tr(z_{i1}) + \cdots + tr(z_{iq}),$$

where $tr(z_r)$ is the tree for the r-th variable, that means it represents clusters of the r-th variable with the cluster form determined by the scale level of the corresponding variable. A traditional tree hardly finds clusters for single components. It typically produces clusters that combine several variables, in particular, mixing nominal and ordinal predictors.

Clustering by trees is a forward selection strategy. But one should be aware that the all subsets strategies fail even in cases of a moderate number of categories. Already in the case of only one predictor one has to consider all subsets $S_1, \dots, S_m$ and fit the corresponding model with predictor $\eta_i = \alpha_1 I(z_i \in S_1) + \cdots + \alpha_m I(z_i \in S_m) + \boldsymbol{x}_i^T \boldsymbol{\beta}$. This is computational feasible only for a very small number of categories. For more than one variable one has to consider all possible combinations, which is bound to fail.

2.3. Tree-Structured Clustering

For simplicity we start with only one categorical predictor. The model for the general case is introduced in a later section.

2.3.1. Trees with Clusters in a Single Predictor

Let us first consider one ordinal (or metric) variable z. Then one split in a tree that includes a linear predictor is found by fitting the model with predictor

$$\eta_i = \alpha_l I(z_i \leq c) + \alpha_r I(z_i > c) + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $I(\cdot)$ again denotes the indicator function. By use of the split-point c the model splits the predictor space into two regions, $z \leq c$ and $z > c$. In the left node, for all $z \leq c$, one specifies the response level α_l, in the right node, for all $z > c$, one specifies the level α_r. It should be emphasized that in $\boldsymbol{x}$ no intercept is included. An equivalent representation of the predictor is

$$\eta_i = \beta_0 + \alpha I(z_i > c) + \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

with the transformation of parameters given by $\beta_0 = \alpha_l$ and $\alpha = \alpha_r - \alpha_l$. The latter form of the predictor is more convenient since it contains an intercept as common regression models do and only one step function has to be specified.

When growing trees one has to specify the possible split-points. Let in the following C denote the set of possible splits c. For a metric predictor, in principle all possible thresholds c can be used, but it suffices to use as candidates all the distinct observations available for the predictor. Therefore, C contains the distinct values of the observed predictor. For ordinal predictors $z \in \{1, \dots, K\}$ the set $C = \{c_1, \dots, c_K\}$ is simply $\{1, \dots, K\}$.

The basic algorithm that we are using for an ordinal variable is the following.

Tree-Structured Clustering - Single Ordered Predictor

Step 1 (Initialization)

(a) Estimation: Fit the candidate GLMs with predictors

$$\eta_i = \beta_0 + \alpha_k I(z_i > c_k) + \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad k = 1, \dots, K$$

(b) Selection

Select the model that has the best fit. Let $c^*_{k_1}$ denote the best split.

Step 2 (Iteration)

For $\ell = 1, 2, \dots,$

(a) Estimation: Fit the candidate models with predictors

$$\eta_i = \beta_0 + \sum_{s=1}^{\ell} \alpha_{k_s} I(z_i > c^*_{k_s}) + \alpha_k I(z_i > c_k) + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

for all values $c_k \in C \setminus \{c^*_{k_1}, \dots, c^*_{k_\ell}\}$

(b) Selection

Select the model that has the best fit yielding the split-point $c^*_{k_{\ell+1}}$.

The algorithm uses two steps, fitting of candidate models and selection of the best model. In GLM-type models it is quite natural to measure the fit by the deviance. Thus, one selects the model that has the smallest deviance. The criterion is equivalent to minimizing the entropy, which has been used as a splitting criterion already in the early days of tree construction (Breiman et al., 1984).

The algorithm yields a sequence of fitted split-points $c^*_{k_1}, c^*_{k_2}, \ldots$ from C and the corresponding parameter estimates $\hat{\alpha}_{k_1}, \hat{\alpha}_{k_2}, \ldots$ from the last fitting step. Typically the selection of split-points is stopped before all possible splits are included (for stopping criteria see below) and one obtains the subset of selected splits $C^* = \{c^*_{k_1} \ldots, c^*_{k_{m-1}}\}$, where m denotes the number of selected clusters. Since the fitted functions are step functions one obtains a partitioning into clusters of adjacent categories. For ordered categories the thresholds are given by $C = \{1, \ldots, K\}$ and one obtains the clustering after ordering the selected thresholds such that $c_{(k_1)} < c_{(k_2)} < \ldots$ by $\{1, \ldots, c_{(k_1)}\}, \{c_{(k_1)} + 1, \ldots, c_{(k_2)}\}, \ldots, \{c_{(k_{m-1})} + 1, \ldots, K\}$. If in the initialization step the maximal value from the set of considered split-points, C, is selected, the algorithm stops immediately because in the iteration steps always the same model would be found. Then, $\hat{\alpha}_1 = 0$ and no split-point is selected. Thus, the variable is not included.

Although the method generates trees the methodology differs from the fitting of common trees if a parametric term is present. In common trees without a parametric term partitioning of the predictor space is equivalent to splitting the set of observed data accordingly. In the next split only the data from the corresponding subspace are used. For example, when a split yields the partition $\{z \leq c\}$, $\{z > c\}$, in the next split only the data from $\{z \leq c\}$ (or $\{z > c\}$) are used to obtain the next split. This is different for the tree-structured model. In all of the fitting steps all data are used. This ensures that one obtains valid estimates of the parametric component together with the splitting rule.

The method explicitly does not use off-sets. When fitting within the iteration steps the previously fitted models serve only to specify the split-points that are included in the current fit. But no estimates from the previous steps are kept. This is in contrast to Yu et al. (2010), where off-sets are used.

Stopping Criterion

When building a tree it is advisable to stop after an appropriately chosen number of steps. There are several strategies to select the number of splits. One strategy that has been used since the introduction of trees is to grow large trees and prune them afterwards, see Breiman et al. (1984) or Ripley (1996), Chapter 7. Alternative strategies based on conditional inference procedures were given by Hothorn et al. (2006).

We use as one strategy *k-fold cross-validation.* That means the data set is split into k subsets. The tree is grown on $k-1$ of these subsets, which is considered the learning sample, and then the tree is evaluated on the left-out sub sample. Since we are working within the GLM framework a natural candidate for the evaluation criterion is the predictive deviance. The number of splits that showed the best performance in terms of the predictive deviance is chosen in the final tree fitted for the whole data set.

An alternative is to use a *stopping criterion based on p-values*, a procedure that is strongly related to the conditional inference procedure proposed by Hothorn et al. (2006). In each step of the fitting procedure one obtains a p-value for the parameter that determines the splitting. In our notation, in the ℓ-th split one tests the null hypotheses $H_0 : \alpha_\ell = 0$ yielding the p-value p_ℓ for the selected split. Typically the sequence of p-values $p_1, p_2, \ldots$ is increasing. A simple criterion is to stop if the p-values are larger than a pre-specified threshold α. However, one should adapt for multiple testing errors because in each split several hypotheses are tested. A simple strategy is to use the Bonferroni procedure and stop if $p_\ell > \alpha/(K-(\ell-1))$ because in the ℓ-th split $K-(\ell-1)$ number of parameters are tested. Then, in each step the overall error rate is under control. As test statistic one can use the Wald statistic or the likelihood ratio statistic. Although the Wald statistic is easier to compute, we prefer the likelihood ratio statistic because it corresponds to the selection criterion, which selects the model with minimal deviance.

Nominal Predictor

For a nominal predictor $z \in \{1, \ldots, K\}$ splitting is much harder because one has to consider all possible partitions that contain two subsets. That means one has $2^{K-1}-1$ candidates for splitting. For large K the number of candidates is excessive. But it has been shown that for regular trees it is not necessary to consider all possible partitions. One simply orders the predictor categories by increasing mean of the outcome and then splits the predictor as if it were an ordered predictor. It has been shown that this gives the optimal split in terms of various split measures, see Breiman et al. (1984) and Ripley (1996) for binary outcomes and Fisher (1958) for quantitative outcomes and the remarks of Hastie et al. (2009).

2.3.2. Trees with Clusters in More than One Predictor

If several predictors are included in the tree component the algorithm also selects among the available variables. Let C_r denote the possible splits in variable z_r and K_r denote the number of values in C_r. The basic form of the algorithm is the following.

Tree-Structured Clustering - Several Ordered Predictors

*S*tep 1 (Initialization)

(a) Estimation: Fit the candidate GLMs with predictors

$$\eta_i = \beta_0 + \alpha_{rk} I(z_{ir} > c_{rk}) + \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad r = 1, \dots, q, \ k = 1, \dots, K_r$$

(b) Selection

Select the model that has the best fit. Let $c^*_{r_1,k_1}$ denote the best split, which is found for variable z_{r_1}. That means that $c^*_{r_1,k_1}$ is from the set of possible splits for z_{r_1}.

*S*tep 2 (Iteration)

For $\ell = 1, 2, \dots,$

(a) Estimation: Fit the candidate models with predictors

$$\eta_i = \beta_0 + \sum_{s=1}^{\ell} \alpha_{r_s,k_s} I(z_{ir_s} > c^*_{r_s,k_s}) + \alpha_{rk} I(z_{ir} > c_{rk}) + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

for all r and all values $c_{rk} \in C_r$ that have not been selected in previous steps.

(b) Selection

Select the model that has the best fit yielding the new split-point $c^*_{r_{\ell+1},k_{\ell+1}}$ that is found for variable $z_{r_{\ell+1}}$.

In the sequence of selected split-points $c^*_{r_1,k_1}, c^*_{r_2,k_2}, \dots$ and corresponding estimates $\hat{\alpha}_{r_1,k_1}, \hat{\alpha}_{r_2,k_2}, \dots$ the first index refers to the variable and the second to the split for this variable. The selected splits for the r-th variable can be collected in C^*_r, which comprises all splits $c^*_{r_\ell,k_\ell}$for which $r_\ell = r$ holds.

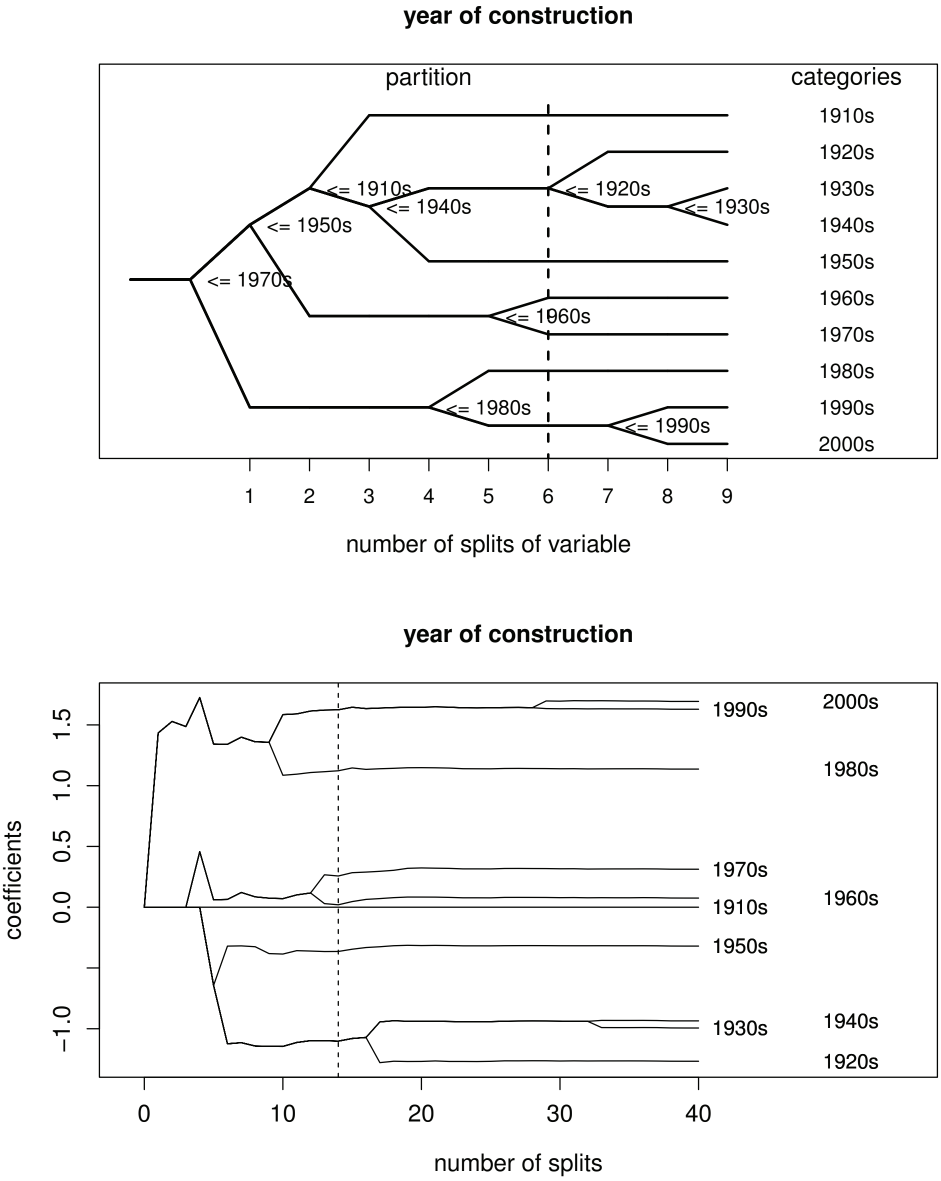


Figure 2.1.: Results for the ordinal predictor year of construction for the analysis of the Munich rent standard data. Upper panel: resulting tree for year of construction, lower panel: paths of coefficients against all splits.

2.3.3. Trees for Rent Data

In the Munich rent data one has one nominal predictor (urban district), three ordinal predictors (year of construction in decades, number of rooms, quality of residential area), one metric variable (floor space) and five binary variables. In the additive part we model the effect of the metric predictor by cubic regression splines and include the binary variables

Table 2.1.: Estimated coefficients, stability measures of the tree component and 95% confidence intervals of the linear term for the analysis of the Munich rent standard data.

Predictor	**Cluster**	**Coefficient**	**Stabilty**
Urban district	7,11,14,16,22,23,24	-1.525	0.431
	6,8,10,15,17,19,20,21,25	-1.005	0.421
	9,13	-0.647	0.506
	2,4,5,12,18	-0.368	0.511
	1,3	0.000	0.552
Year of construction	1910	0.000	1.000
	1920s,1930s,1940s	-1.098	0.730
	1950s	-0.365	1.000
	1960s	0.030	1.000
	1970s	0.267	1.000
	1980s	1.115	1.000
	1990s,2000s	1.622	0.927
Number of rooms	1,2,3	0.000	0.642
	4,5,6	-0.327	0.865
Quality of residential area	fair	0.000	1.000
	good	0.356	1.000
	excellent	1.436	1.000

Predictor	**Coefficient**	**95% confidence interval**
Hot water supply (no)	-1.987	[-2.513,-1.372]
Central heating (no)	-1.355	[-1.820,-0.947]
Tiled bathroom (no)	-0.543	[-0.786,-0.318]
Supplementary equiment in bathroom (yes)	0.511	[0.199,0.807]
Well equipped kitchen (yes)	1.198	[0.839,1.579]

in a linear form. The fusion of categories obtained by the tree is illustrated for the predictor year of construction. Figure 2.1 shows the resulting tree and the coefficient paths over the splits for the predictor year of construction. The upper panel shows the successive splits against the number of splits in this predictor. The lower panel shows the coefficients plotted against the splits in all of the predictors. It is seen, in particular from the first steps, that estimates can change when other variables are included. But after about 14 splits the estimates are very stable. Since the maximal number of splits is 40 the estimates after 40 splits represent the fit of a generalized additive model. When p-values with significance level 0.05 are used as splitting criterion one obtains seven clusters marked by the dashed lines in both panels. The rent per square meter seems to be the same, for example, for houses built between the 1920s and 1940s and for houses built in the 1990s and 2000s. The gap between the high rent cluster and the middle clusters is larger than the gap between the

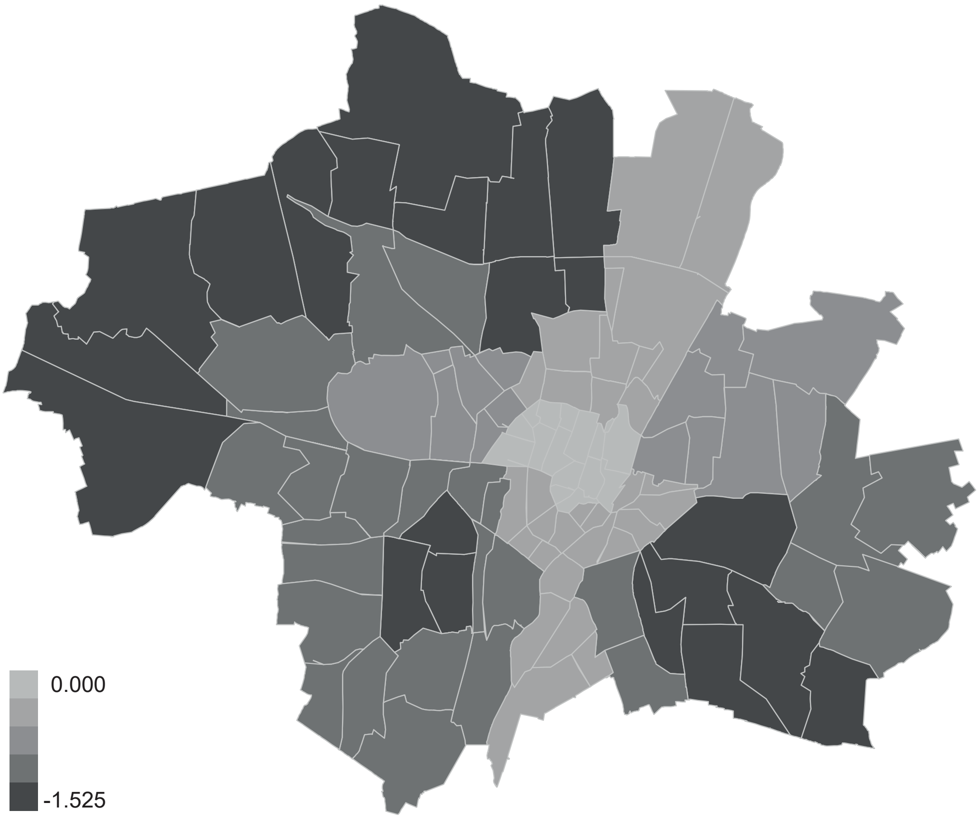


Figure 2.2.: Map of Munich indicating the estimated clusters for variable urban district of the Munich rent standard data. The algorithm detects five groups of districts that share the same effect, respectively. The darker the shade the lower the estimated coefficient.

middle clusters and the low rent clusters. The estimated values are given in Table 2.1. The table also shows the clusters for the other variables in the tree component, the estimates of the linear part as well as stability measures that are explained later. It should be noted that no predictor has been completely excluded from the model.

The size of clusters found by the algorithm vary in a wide range. For variable urban district (reference 1: inner city around Marienplatz) one obtains five clusters, where the smallest clusters $\{1, 3\}$ and $\{9, 13\}$ consist of only two categories, but the biggest cluster contains nine categories. A graphical illustration of the resulting partition is given in Figure 2.2. The map was created by `R` package `R2BayesX` (Umlauf et al., 2015; Belitz et al., 2015). A darker shade corresponds to a lower estimated coefficient. It can be seen that rents are most expensive around the city center and therefore estimated coefficients for the other clusters are all negative (darker shades). There are several outskirts that build the cluster with the lowest rents. A detailed overview of all districts is given in Appendix A on page 205, where the numbers correspond to the labels in Table 2.1.

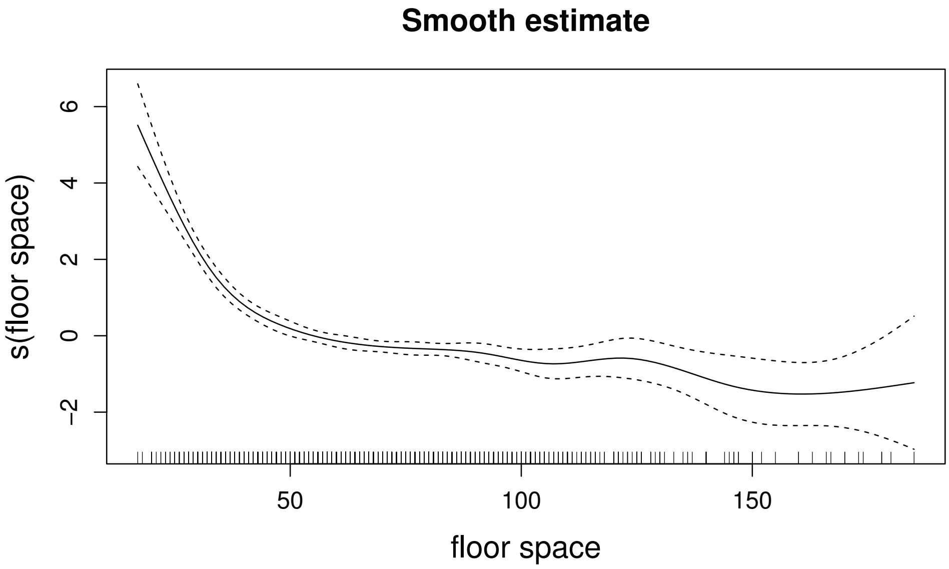


Figure 2.3.: Resulting function of the smooth estimation of predictor floor space of the Munich rent standard data in the additive part of the tree.

Since it is not to be expected that the rent per square meter depends linearly on the floor space it is fitted as a smooth function. For the estimation we use penalized cubic regression splines, penalized by the integrated squared second derivative penalty (Eilers and Marx, 1996). We chose a modest number of ten basis functions. For computation we used the R package mgcv (Wood, 2011). When fitting a smooth function one has to specify a smoothing parameter, which in our procedure is selected new in each iteration step. The resulting function, pictured in Figure 2.3, is monotonically decreasing, which means that the net rent per square meter decreases with growing floor space. The function decreases strongly until a floor space of about 50 and is rather flat for a greater floor space, but it is definitely not linear.

2.4. Standard Errors and Stability of Clusters

The tree-structured model is an extension of GLMs and GAMs. While in standard GLMs approximate standard errors for the parameters are obtained from asymptotic theory, for semiparametric models as considered here an alternative way to obtain standard errors has to be used. One way is to use bootstrap procedures as described in Efron and Tibshirani (1994). By repeated fitting on sub samples that have been obtained by drawing with replacement one can compute approximate standard errors. But when computing standard errors one has to distinguish between the two parts of the model, the parametric and the

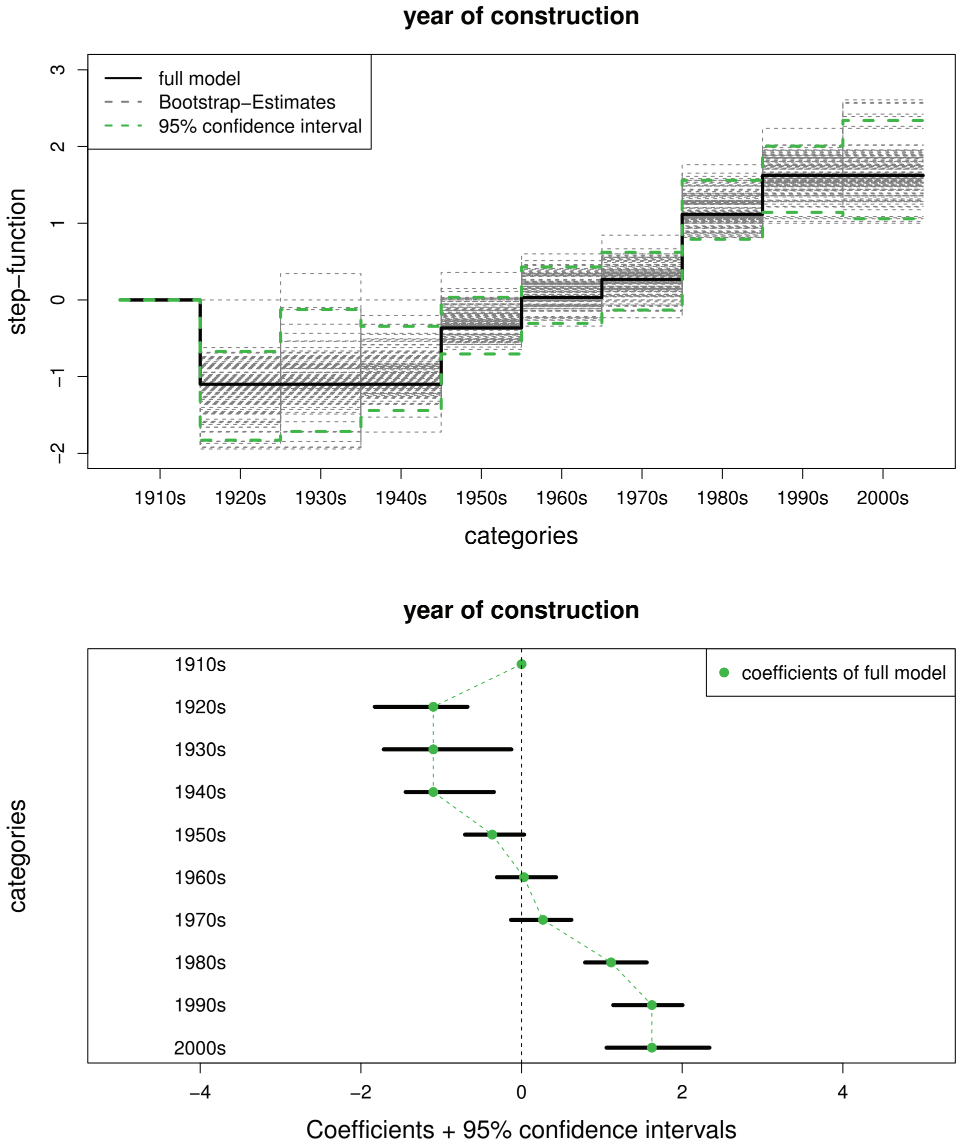


Figure 2.4.: Estimated step functions and resulting 95% confidence intervals for the ordinal predictor year of construction for the analysis of the Munich rent standard data based on 1000 bootstrap samples.

tree part. For the parametric part, which means for the parameter $\boldsymbol{\beta}$, standard procedures to compute the standard deviations and confidence intervals over the bootstrap samples can be used. For the rent data the resulting confidence intervals are given in Table 2.1. For categorical predictors we consider the estimated step functions, which are determined by sums of the parameter estimates $\hat{\alpha}_{rk}$. Bootstrap intervals can be given for all estimated sums $\tilde{\alpha}_{rs} = \sum_{k=1}^{s} \hat{\alpha}_{rk}$. Typically some of the parameter estimates $\hat{\alpha}_{rk}$ are zero, but this will not to be the case in the bootstrap samples. Consequently one obtains confidence intervals

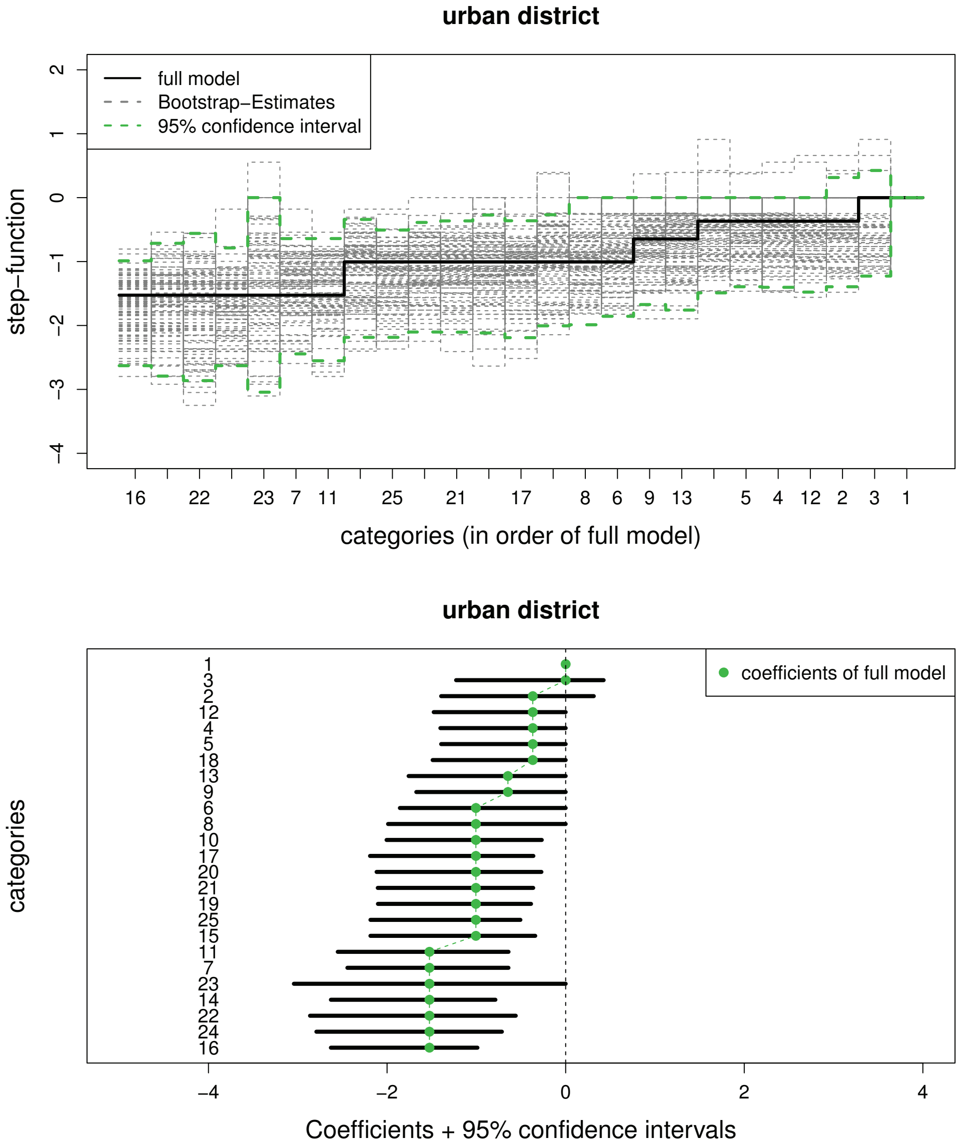


Figure 2.5.: Estimated step functions and resulting 95% confidence intervals for the nominal predictor urban district for the analysis of the Munich rent standard data based on 1000 bootstrap samples.

that do not necessarily have equal length within clusters. The somewhat harder problem is the case of nominal predictors. Since in bootstrap samples the ordering of the predictor categories will differ one has to carefully rearrange the parameter estimates to obtain the confidence intervals for the estimates $\tilde{\alpha}_{rs}$ in the original sample.

For illustration we show the bootstrap results for the variables year of construction (Figure 2.4) and urban district (Figure 2.5). The upper panels of the two figures show only the first 100 bootstrap based function estimates. The lower panels show the 95% confidence intervals for the single effects for 1000 bootstrap samples. It is seen that for year of construction the first big cluster, which contains the decades 1920-1940, has varying lengths of confidence intervals, but all of them do not contain zero. Thus they should be distinguished from the reference category, which is the first decade, and has fixed value zero. For the nominal predictor urban district confidence intervals are larger than for the ordinal variable year of construction. This was to be expected for a nominal variable with many categories. However, as already suspected from Figure 2.2, it is seen that several big clusters are definitely less expensive than the district inner city.

Bootstrapping yields confidence intervals for the step functions but does not contain information about the reliability of cluster identification. Therefore it seems warranted to supplement the confidence intervals by diagnostic tools that reflect the stability of clusters. One is a distance matrix obtained from the bootstrap samples. Let B denote the number of bootstrap samples and n_{lk} denote the number of samples for which category l and k were in the same cluster. Then a simple similarity measure for categories is $s_{lk} = n_{lk}/B$. If $s_{lk} = 1$ category l and k were in the same cluster in all of the bootstrap samples. The stability of a cluster is obtained by averaging over all the distances of pairs of categories within the cluster. Of course, if a cluster contains only one category the similarity measure has the value 1. It is seen from Table 2.1 that the stability can strongly vary across clusters. For the nominal variable urban district the clusters show similarity in the range $(0.43, 0.55)$ whereas for the ordinal variable year of construction one obtains also very large values as 0.73 and 0.93. The latter value refers to the cluster of decades 1990 and 2000 and means that it was in the same cluster in 93% of the bootstrap samples.

2.5. Related Approaches

In the following the relation of the proposed method to related and alternative approaches is shortly sketched. Our method aims at the identification of clusters in categorical predictors in the presence of other, in particular, also continuous variables. Therefore discussion refers to this objective.

The strongest relation is to approaches that are able to detect clusters in categories by the definition of appropriate penalty terms and maximization of the corresponding penalized log-likelihood. Let us for simplicity consider the case of one categorical predictor and several continuous predictors. Then the corresponding linear predictor of a GLM is given by

$$\eta_i = \alpha_0 + \alpha_1 \tilde{z}_{i1} + \cdots + \alpha_{K-1} \tilde{z}_{i,K-1} + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $\tilde{z}_j$ are the dummy variables for the categorical predictor $z \in \{1, \dots, K\}$. Let the penalized log-likelihood be given by $l_p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = l(\boldsymbol{\alpha}, \boldsymbol{\beta}) - J(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denotes the log-likelihood of the GLM and $J(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a penalty term. For a categorical predictor a penalty that enforces clustering of categories of z is given by

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \lambda \sum_{l<k} |\alpha_l - \alpha_k|.$$

For $\lambda = 0$ one obtains the ML estimate, if $\lambda \to \infty$ all categories of z are fused to one cluster. The method has been proposed by Bondell and Reich (2009) for ANOVA-type models and was adapted to variable selection by Gertheiss and Tutz (2010), Tutz and Gertheiss (2014). The main problem with this approach is that it becomes computationally infeasible if the number of categories gets large. This is due to the definition of the penalty term, which includes all pairwise differences. If the number of categories is 40, the penalty already contains 780 differences. As the approach by penalized maximum-likelihood estimation is a competitor to the method proposed here we include it in our simulations in Section 2.6.1. For the comparison we use the `R` add-on package `gvcm.cat` (Oelker, 2015). The implementation is based on a uniform framework proposed by Oelker and Tutz (2015) that uses an approximation introduced by Fan and Li (2001).

An obvious relation is to classical recursive partitioning as CARTS. The main differences have already been outlined. The method proposed here allows to include a parametric or smooth component that accounts for the main effect in a model. Thus, the method allows to identify clusters of categories within one predictor that have the same effect on the response. If one fits a classical tree that includes all the variables no clustering is obtained because the tree fits interactions between all the variables.

As a forward strategy one might suspect a strong relation to boosting concepts. Boosting methods were originally developed in the machine learning community as a means to improve classification (e.g., Shapire, 1990). Later it was shown that it can be seen as the fitting of an additive structure by minimizing specific loss functions (see Friedman, 2001; Friedman et al., 2000; Bühlmann and Yu, 2003). Minimization is obtained iteratively by utilizing a steepest gradient descent approach. In a forward searching procedure components that are potentially relevant are included in the predictor. The potentially relevant components are fitted by so-called base learners. A simple example is the fitting of a linear model where a base learner refers to the fitting of one component of the linear predictor, $x_i\beta_j$. By including one of the components at a time and selection of the component that maximally improves the fit one obtains the final model. The method proposed here seems to be very similar. The base learner that is used is one split in a variable, which has the form $\alpha_{rk}I(z_{ir} > c_{rk})$. Selection of the most relevant term is also based on goodness-of-fit. However, there is one crucial difference between the tree-structured model and boosting, namely that boosting uses weak learners. A weak learner is somehow vaguely defined as a refit that only slightly improves the overall fit, but properties of the procedure definitely depend on the weakness of the learner (Bühlmann and Yu, 2003). In our procedure a weak learner would be the inclusion of the best split $\alpha_{rk} I(z_{ir} > c_{rk})$, but with a new parameter value α_{rk} that is only slightly larger than the parameter used in the previous step. Of course, one could fit categorical predictors by weak learners, or equivalently by boosting, but the effect would be a smooth fit over categories, because in each boosting step the parameters are updated only weakly but most of them are selected during the iterations. Therefore, the procedure fails to obtain the intended clustering of categories.

A further approach that is related to the tree-structured model is model-based partitioning proposed by Zeileis et al. (2008). The basic concept is to fit a parametric model in every leaf of a tree, for example, a linear regression model. By fitting a model to subsets that are defined in the usual way by splitting variables one obtains a partitioned or segmented parametric model. Within this framework it is possible to detect areas where model fits differ because the linear models fitted to leafs differ in their parameters. It is a flexible modelling tool in which all kinds of parametric models can be used. However, as in common trees the focus is not on main effects but on interaction although in the wider sense that models differ in different leafs. In particular for categorical predictors, which are considered here, one obtains different structures when using model-based partitioning in the sense of Zeileis et al. (2008) or structured regression as proposed here. In model-based partitioning splits in a categorical predictor are enforced if the parameters of the fitted model differ in the resulting clusters of categories. After several splits one obtains quite different models that hold within clusters of categories. In our structured regression clusters of categories are built by assuming that the effect on the response is the same within clusters and that the main effects are constant. Thus the focus is on similarity of categories not on dissimilarity of categories with respect to the models that hold within clusters of categories.

Finally, several modelling strategies were proposed that also use a combination of a parametric term and a tree component. One is the partially linear tree-based regression model developed by Chen et al. (2007). The focus of the paper is on genetic risk factors. The main difference to the procedure proposed here is the restriction to a linear term and an alternative algorithm that uses off-sets in the iterative algorithm instead of updating the linear component. The approach has been extended to account for multivariate outcomes by Yu et al. (2010). An alternative model is the regression trunk model proposed in Dusseldorp and Meulman (2004) and Dusseldorp et al. (2010). The model is designed for metric response only. In contrast to our approach it uses the same variables in the tree component and the linear term, which yields hard to interpret effects. Moreover, they use the more conventional fitting strategy that first grows a large tree and then prunes it. Therefore, the relevance of predictors in terms of significance should be hard to obtain. A combination of linear fits and tree-structured component with the focus on diagnostic for linear models was considered by Su et al. (2009).

2.6. Simulations

The proposed tree-structured model allows to detect clusters of categories that share the same effect on the response while letting other variables, in particular metric variables, have a linear or smooth effect on the response. In order to investigate the performance of the model we now give the results of several simulations. We first consider data with one categorical covariate. The main objective here is to compare the proposed model, abbreviated by *TSC*, for tree-structured clustering, to the model based on penalized maximum-likelihood estimation, abbreviated by *PENL*. For the computations we used the function `gvcm.cat()` of `R` package `gvcm.cat` (Oelker, 2015) and included adaptive weights in the penalty term. Subsequently we give detailed results for more complex data with several predictors comparing several stopping criteria. All the results are based on 100 repetitions.

Evaluation Criteria

The estimated coefficients are compared to the true parameters by calculating mean squared errors (MSEs). Therefore we distinguish between the tree-based parameters $\boldsymbol{\alpha}$ and the parameters $\boldsymbol{\beta}$ of the linear term. For the r-th categorical predictor the MSE of the α-parameters is $\sum_{k=1}^{K_r}(\hat{\alpha}_{rk} - \alpha_{rk})^2/K_r$ and for the β-parameters it is $\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2/p$, where p denotes the number of covariates in the linear term.

To judge the clustering of the categorical $\boldsymbol{z}$-variables in the tree component, False Positive Rates (FPR) and False Negative Rates (FNR) are computed.

- False Positiv: A difference between two estimated parameters α_{rk} which is truly zero is set to nonzero
- False Negativ: A difference between two estimated parameters α_{rk} which is truly nonzero is set to zero

In addition the number of clusters respectively the number of splits determined by the different approaches are of interest.

2.6.1. Comparison to Penalized Estimation

Here we consider data with one categorical variable z. The true number of clusters in each case is $m = 5$, so categories $1, \dots, K$ are split into five partitions $S_1, \dots, S_5$. The true coefficients of the clusters are $\boldsymbol{\alpha} = (0, 1, 2, 3, 4)^\top$. The number of categories K varies from 20 to 100. In particular, we focus on the case where the number of categories K is much higher than the true number of clusters m. The model has an additional linear term $\boldsymbol{x}^\top\boldsymbol{\beta}$,

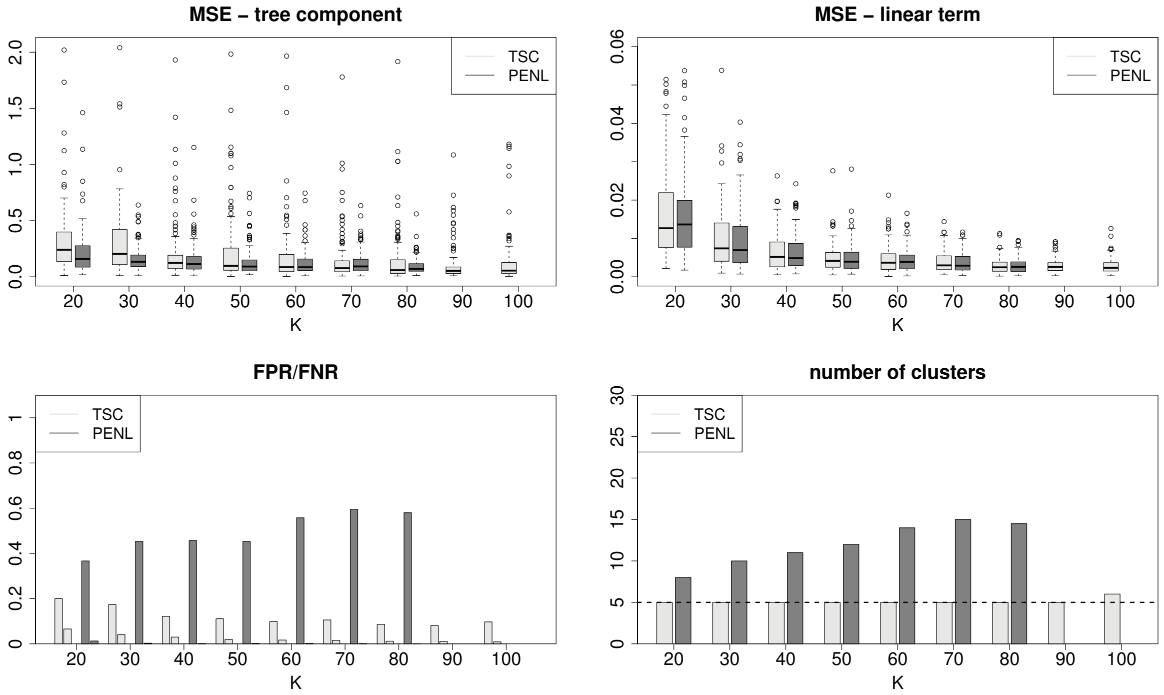


Figure 2.6.: Results of the simulation with normal response and one ordinal predictor for the tree-structured model (light grey) and the penalty approach (dark grey).

where $\boldsymbol{x}$ is $N(\mathbf{0}_4, \boldsymbol{\Sigma}_4)$-distributed with variances 1 and covariances 0.3. The true regression coefficients of the linear term are $\boldsymbol{\beta} = (-0.6, 0.4, -0.8, 1.2)^\top$. In general z has a nominal structure, but it is also possible to assume that the class labels have an ordinal structure. In the following investigations we distinguish between these two cases.

In order to gain comparability of the tree-structured model and the penalty approach in both cases we use 5-fold cross-validation to select the best model.

Normal Response

We start with simulation scenarios where the responses y_i, $i = 1, \ldots, n$ are normally distributed with $\varepsilon_i \sim N(0, 1)$. We consider a balanced design with five observations in each category, thus the total number of observations is $n = K \cdot 5$.

Figure 2.6 shows the results for the settings where z is treated as ordinal predictor. Each panel shows the results for the nine settings with varying K (along the x-axis). For the tree-structured model (TSC) all the results are given in light grey, for the penalty approach (PENL) they are given in dark grey. As the penalty approach is computational infeasible for a very large number of categories, no results are displayed for the settings with $K = 90$ and $K = 100$. The mean squared errors of the tree component given in the top left are very

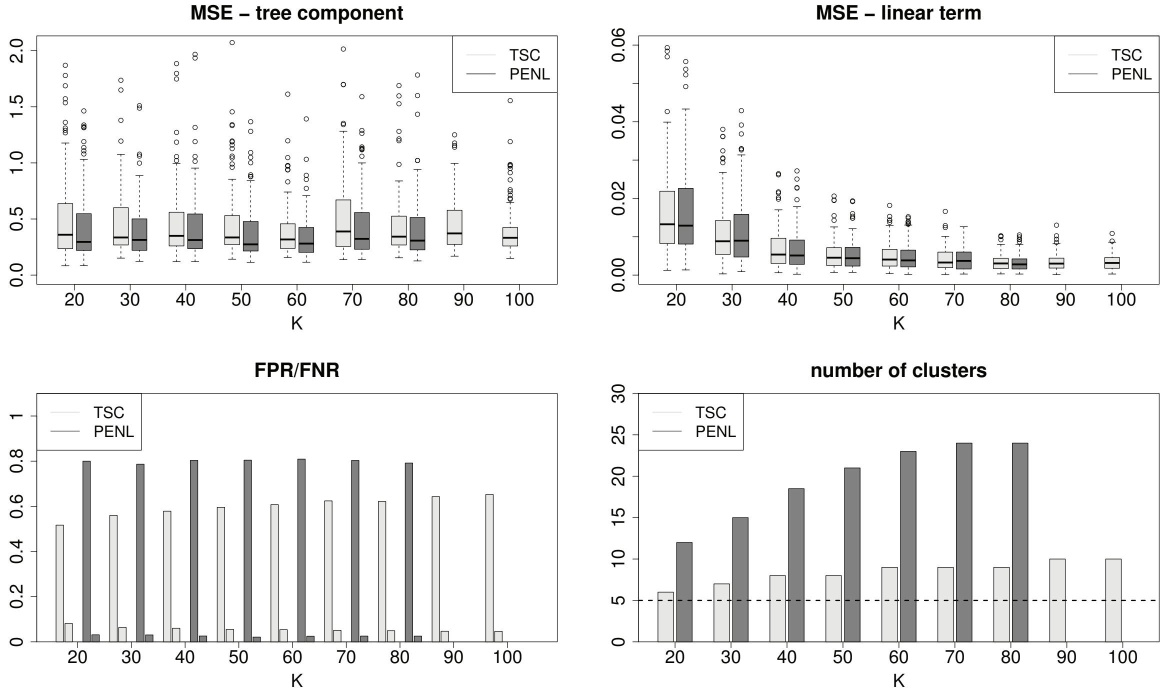


Figure 2.7.: Results of the simulation with normal response and one nominal predictor for the tree-structured model (light grey) and the penalty approach (dark grey).

stable across all settings. For a small number of categories ($K = 20$ and $K = 30$) the tree-structured model performs worse, but the results are rather the same for large K. The mean squared errors of the linear term (top right) decrease with increasing K. The observed values are very small and nearly the same for the two approaches. However, distinct differences are seen for the FPR and FNR as well as for the number of clusters. They are pictured in the lower panel in Figure 2.6, where the bars correspond to the average over all repetitions. It can be seen that the penalty approach performs very poorly in particular for large K. One observes false positive rates up to 0.6. For the tree-structured approach they are below 0.2 across all settings. In addition, the tree-structured model on average is able to detect the true number of clusters even for very large K. Whereas the penalty approach distinctly overestimates the number of clusters.

The picture changes for the settings where z is treated as nominal predictor, that is without the pre-assumption that categories are ordered (Figure 2.7). Mean squared errors of the tree component are larger than in the ordinal case for all settings. This is caused by the poor clustering performance (lower panel). False positive rates of the tree-structured model exceed the value 0.5, for the penalty approach one observes even values about 0.8. Again the penalty approach seriously overestimates the true number of clusters, but also the tree-structured model now tends to detect a higher number of clusters. It is worth noting that the mean squared errors of the linear term given in the top right of Figure 2.7 largely remain the same as in the ordinal case. Hence the linear part of the model is not affected by the different assumptions for the scale of the z-variable.

Binary Response

In as second simulation we consider discrete response variables $y_i \sim B(1, \pi_i)$, where $\pi_i = \exp(\eta_i)/(1 + \exp(\eta_i))$. The structure of the simulated data sets remains the same, but in contrast to the simulations with normal response we use a balanced design with 20 observations in each category, giving the total number of observations $n = K \cdot 20$. The corresponding results are given in Figure 2.8 and Figure 2.9. It can be seen, that the previous findings for the simulations with normal response can be confirmed and therefore the conclusions remain largely the same.

In summary, the two approaches are competitive in terms of their estimation accuracy with a tendency of stronger variation for the tree-structured model. Concerning the clustering of categories the tree-structured approach performs much better, especially in the ordinal case. Obviously the nominal case is much more challenging for both approaches. Moreover, it is again noteworthy that no estimates are available for the penalty approach if the number of categories exceeds a certain size.

2.6.2. Evaluation of Stopping Criteria

One of the most important questions when building a tree is the choice of a optimal stopping criterion. In the previous section we used 5-fold cross-validation and in our applications we use a stopping criterion based on p-values to determine the best model. Here we consider a simulation with several covariates to compare different stopping criteria, including those already used.

We consider the case of 4 ordinal and 4 nominal predictors in the tree component of the model. For both types of variables we use two predictors with 10 and two predictors with 5 categories. The true coefficients of the ordinal predictors are $\boldsymbol{\alpha}_1 = (0, 1, 1, 2, 2, 3, 3, 4, 4)^\top$, $\boldsymbol{\alpha}_2 = (0, 0, 0, 0, 2, 2, 2, 2, 2)^\top$, $\boldsymbol{\alpha}_3 = (1, 1, 2, 2)^\top$ and $\boldsymbol{\alpha}_4 = (0, 0, 0, 0)^\top$. For the nominal predictor they are $\boldsymbol{\alpha}_5 = (0, 0.5, 0.5, -0.5, -0.5, 1.5, 1.5, -1.5, -1.5)^\top$, $\boldsymbol{\alpha}_6 = (0, 0, 0, 0, -2, -2, -2, -2, -2)^\top$, $\boldsymbol{\alpha}_7 = (1, 1, -1, -1)^\top$ and $\boldsymbol{\alpha}_8 = (0, 0, 0, 0)^\top$. In both cases the true numbers of clusters are 5, 2 and 3. The fourth predictor is not influential. Note that the effect of the first category in each case is set to zero. Altogether there are 52 possible splits in the tree component. The true model contains 14 splits, 7 within the ordinal and 7 within the nominal predictors. We generate data sets with $n = 2000$ observations and a normal distributed response with $\varepsilon \sim N(0, 1)$. Our model also has an additional

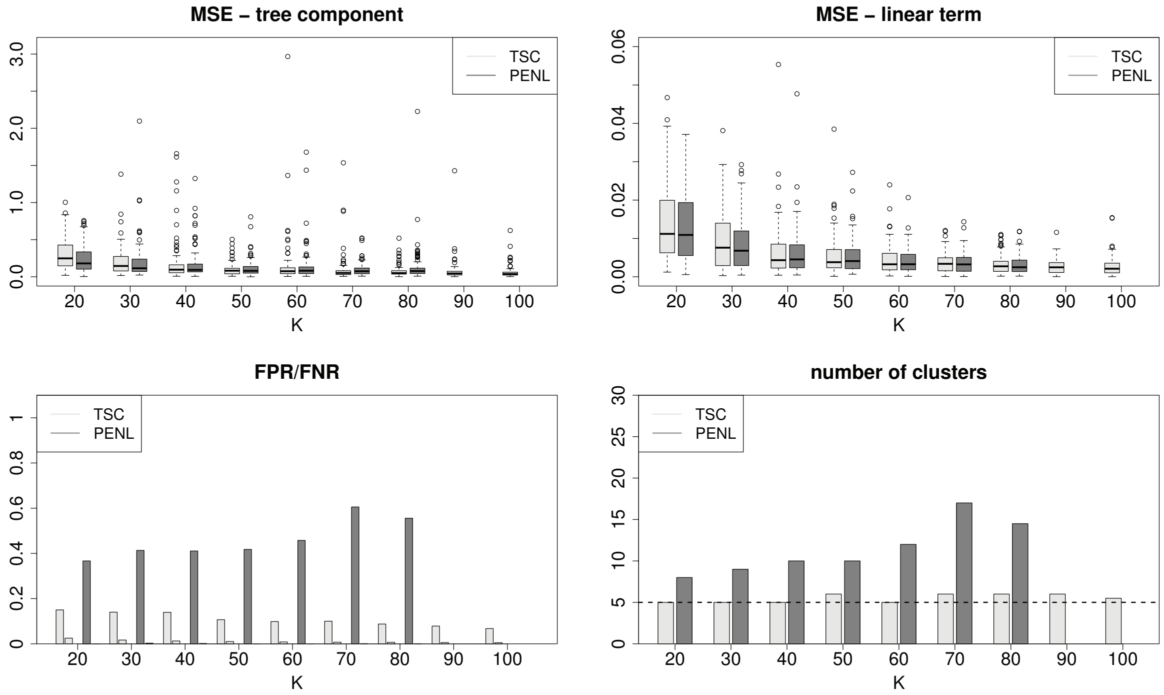


Figure 2.8.: Results of the simulation with binomial response and one ordinal predictor for the tree-structured model (light grey) and the penalty approach (dark grey).

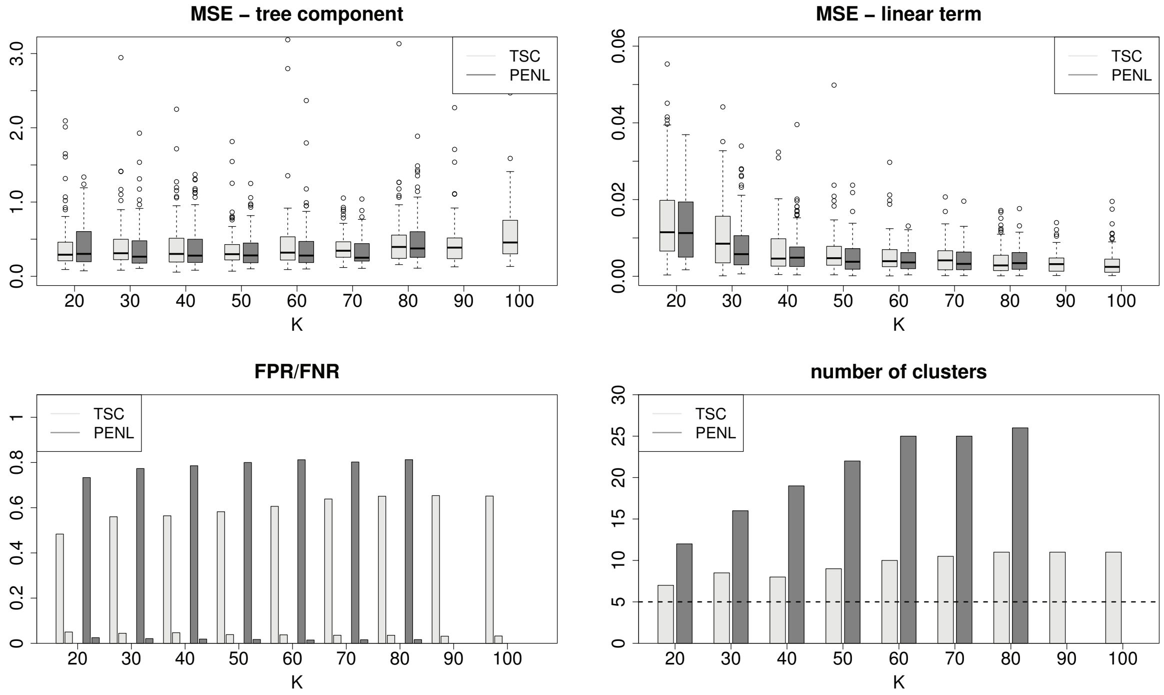


Figure 2.9.: Results of the simulation with binomial response and one nominal predictor for the tree-structured model (light grey) and the penalty approach (dark grey).

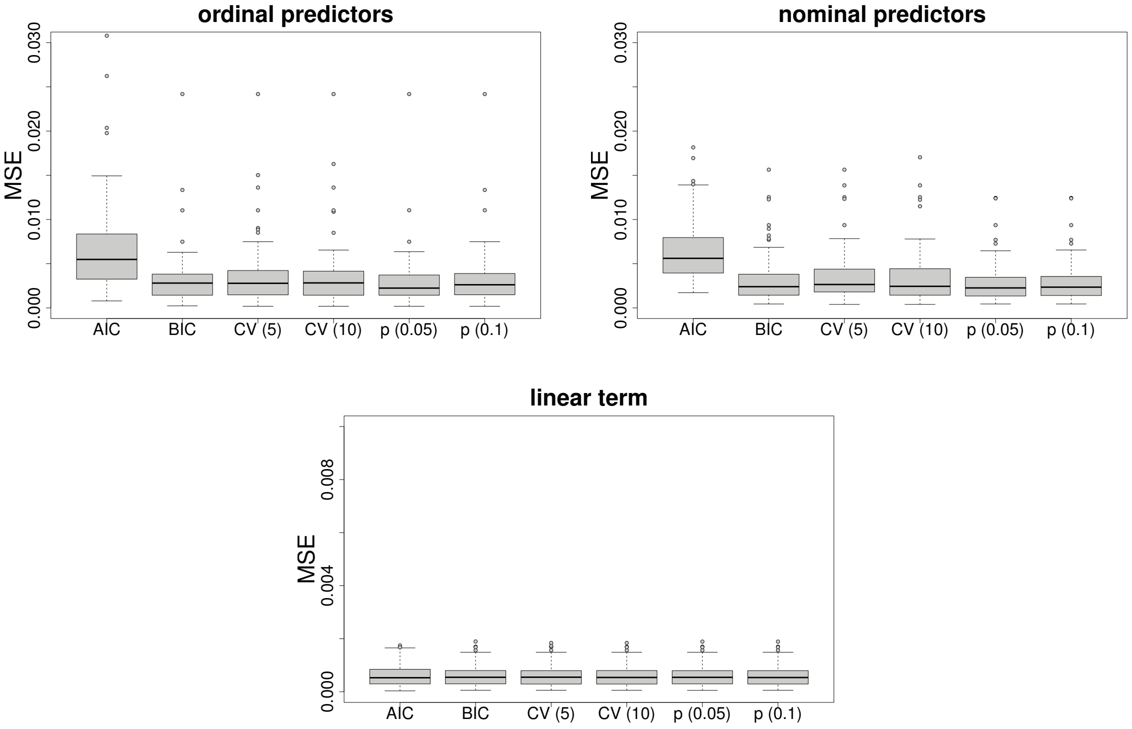


Figure 2.10.: Mean squared errors (MSEs) of parameter estimates of ordinal, nominal predictors and the linear term for the simulation study with several predictors.

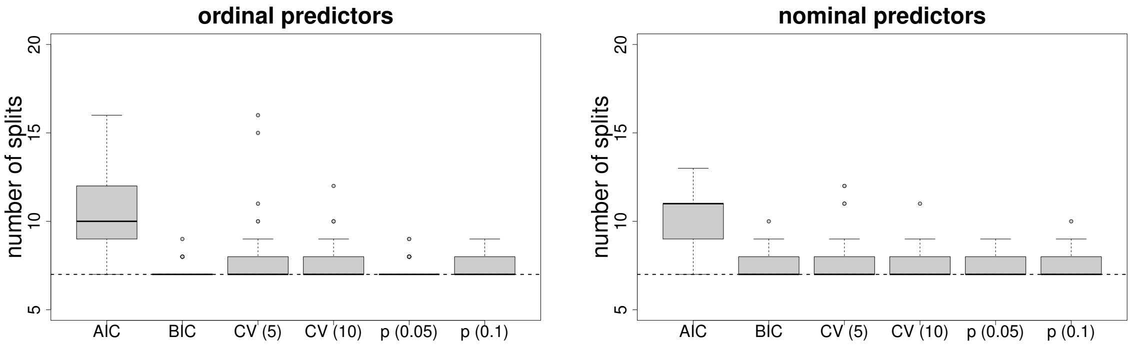


Figure 2.11.: Number of splits of ordinal and nominal predictors in the tree component for the simulation study with several predictors.

linear term $\boldsymbol{x}^T\boldsymbol{\beta}$, where $\boldsymbol{x}$ is $N(\mathbf{0}_5, \boldsymbol{\Sigma}_5)$-distributed with variances 1 and covariances 0.3. The true regression coefficients of the linear term are $\boldsymbol{\beta} = (-2, 1, -1, 3, 2)^\top$.

In our analysis we distinguish between the MSEs for the nominal and the MSEs for the ordinal predictors respectively as the average over the four predictors. Boxplots of the MSEs based on 100 simulations are shown in Figure 2.10. We compare six different stopping criteria: AIC, BIC, 5-fold cross validation, 10-fold cross validation, p-values with significance

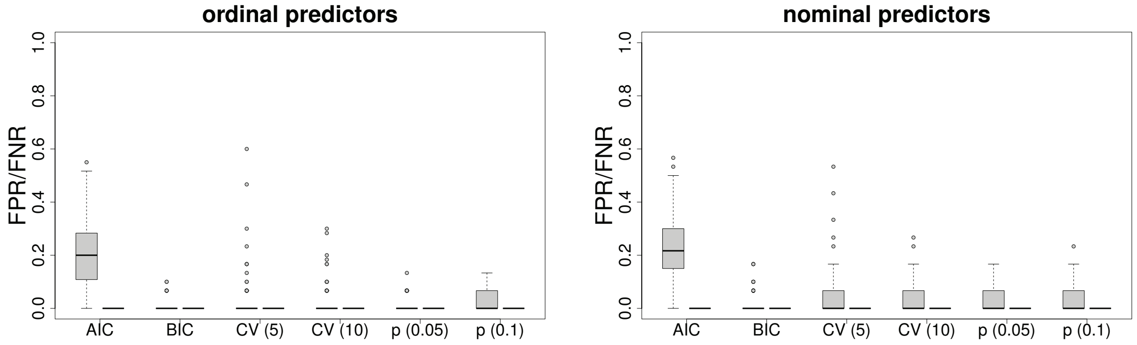


Figure 2.12.: FPR (left boxplots) and FNR (right boxplots) of ordinal and nominal predictors in the tree component for the simulation study with several predictors.

level $\alpha = 0.05$ and p-values with $\alpha = 0.1$. In Figure 2.10 the latter are denoted by $p(0.05)$ and $p(0.1)$. The smallest median of MSEs for the ordinal predictors as well as for the nominal predictors were found for the strategy with p-values and common significance level $\alpha = 0.05$ (fifth boxplots). MSEs for the linear term are very small and almost identical over stopping criteria. As already seen in Section 2.6.1 estimation of the linear term of the tree-structured model shows very good performance and seems to be not strongly linked to the clustering in the tree component.

Figure 2.11 shows the number of splits in the tree component of the model separately for the ordinal and the nominal predictors. The horizontal line shows the optimal number of splits of the underlying data generating model. It is seen that for the ordinal predictors one obtains nearly perfect results with BIC and p-value $\alpha = 0.05$. The true number of 7 splits is found in almost all simulations. For the nominal predictors the performance is very similar over stopping criteria with the exception of AIC, which performs worse than the other procedures. Since for p-values with $\alpha = 0.05$ there is no outlier it shows again the best performance. In summary, the number of splits is very close to the optimal number for all the procedures showing again that the model is able to find the right number of splits.

Figure 2.12 shows boxplots of TPR and FPR seperatly for the ordinal and nominal predictors. As for the MSEs we computed the average over the four predictors. Since the tree-structured model has a weak tendency to overestimate the number of splits (see Figure 2.11) FNRs are found to be zero in all simulations. With exception of AIC also the median of the FPRs is zero over stopping criteria. This again illustrates the overall good performance of the proposed tree-structured approach.

Table 2.2.: Estimated coefficients, stability measures of the tree component and 95% confidence intervals of the linear term for the analysis of the household data.

Predictor	Cluster	Coefficient	Stability
Country	BE,HB,HH	-1.647	0.658
	BB,HE,MV,NW,SL,SN,ST,SH,TH	-0.425	0.521
	BY,BW,NI,RP	0.000	0.637
Number of persons	1	0.000	1.000
	2,3,4,5,6,7,8,9,10,11,12	1.424	0.810
Kind of household	3, 8	-1.438	0.990
	1, 2, 4, 5, 6, 7	0.000	0.443

Predictor	Coefficient	95% confidence interval
net income of all persons	0.580	[0.520,0.650]
PC in household	1.008	[0.899,1.132]
life policy during the year before	0.754	[0.629,0.898]

2.7. Further Applications

In the following the proposed tree-structured model is illustrated in two further applications and its performance is compared to alternative models.

2.7.1. Car in Household

As second application we consider data from the German socio-economic panel from 2012 carried out by the German institute DIW, which comprises 12322 households. They are available from `http://www.diw.de/de/soep`. The response variable we consider is the binary variable if a car is in the household or not. Independent variables that we include in our model are the net income of all persons in the household in thousands of Euro (metric), the country (16 categories), kind of household (nominal factor), number of persons in the household (ordinal factor), PC in the household (yes/no), life insurance during the year before (available/not available).

A particularly interesting variable is the country with 16 categories. In a parametric model it generates 15 parameters. With the approach suggested here the number should reduce because it aims at identifying clusters of countries that share the same effect.

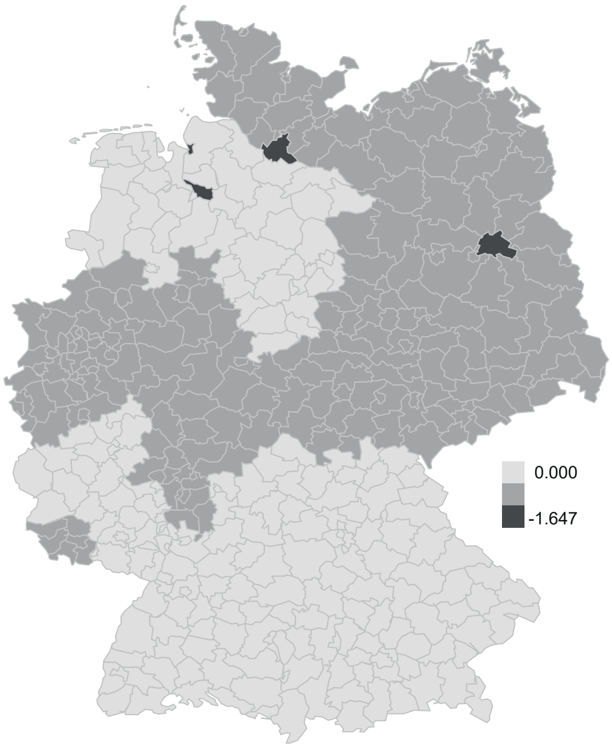


Figure 2.13.: Map of Germany indicating the estimated clusters for variable country of the household data. The algorithm detects three groups of countries that share the same effect, respectively. The darker the shade the lower the estimated coefficient.

We fit a logistic regression model for the probability of holding a car and use p-values as the stopping criterion. The tree component of the model includes the nominal factors country, type of household and the ordinal factor number of persons. The metric variable net income and the two binary variables are included in the linear term of the model. The maximum number of splits in this case is 33. The algorithm stops very early and we obtain the model with four splits as the best model.

The results of the fitted tree-structured model are given in Table 2.2, where the countries are abbreviated by the official country codes by ISO 3166. A detailed overview of all countries and the categories of variable kind of household is given in Appendix A on page 206. Table 2.2 shows in particular estimated coefficients, stability measures for clusters in the tree component and 95% confidence intervals for the linear term based on 1000 bootstrap samples. It is seen that the three variables in the linear term all have a significant influence on the probability of holding a car. The higher the net income, the higher the probability of holding a car. Also a PC in the household and a life insurance increase the probability.

For the nominal predictor country (reference 1: Bavaria) in the tree component one obtains only three clusters that show an interesting structure. A graphical visualization of the resulting partition is given in Figure 2.13. The map of Germany was created by `R` package

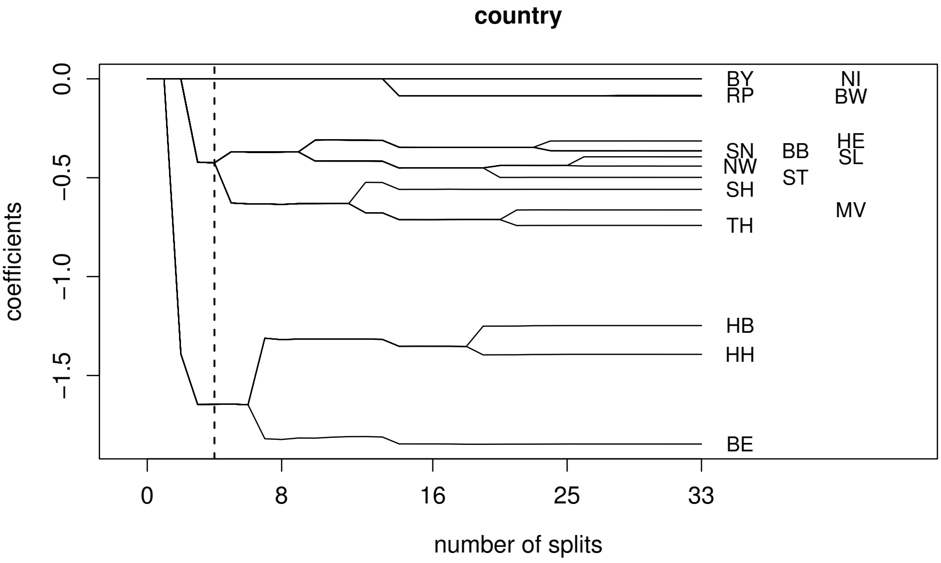


Figure 2.14.: Coefficient paths for the nominal predictor country for the analysis of the household data.

R2BayesX (Umlauf et al., 2015; Belitz et al., 2015). A darker shade corresponds to a lower probability of holding a car. The first cluster, which has the strongest decrease in probability, is formed by the cities Berlin (BE), Bremen (HB) and Hamburg (HH), which are not only cities but also countries. Since in German cities public transportation is easily available and distances are small the necessity of owning a car given fixed income is reduced. The coefficient -0.647 means that the probability of owning a car decreases by a factor of 0.2 when compared to the reference cluster with effect zero. Next to Bavaria the reference cluster also contains Baden-Wuerttemberg and Rhineland-Palatinate in the south of Germany as well as Lower Saxony. The biggest cluster with nine countries has also a reduced probability, but the reduction is not as strong as for the countries that are also cities. As seen from the coefficient paths in Figure 2.14 the big cluster could also divided into two sub-clusters, but were merged by the chosen stopping criterion. For the variables number of persons in the household and kind of household one obtains only two clusters, respectively. It is only distinguished between one person households that show a strongly increased probability of owning a car and the rest of the households. Compared to other kinds of households single parents (category 3) are very unlikely to hold a car. Stability measures in Table 2.2 are very large. For the nominal predictor country the values are greater than 0.5 and do not vary a lot, so the algorithm forms stable clusters.

Figure 2.15 shows the fitted functions for 100 bootstrap samples and 95% confidence intervals based on 1000 bootstrap samples for the predictor country . It is seen that the chosen reference Bavaria is the first country in the order of countries and therefore has the highest

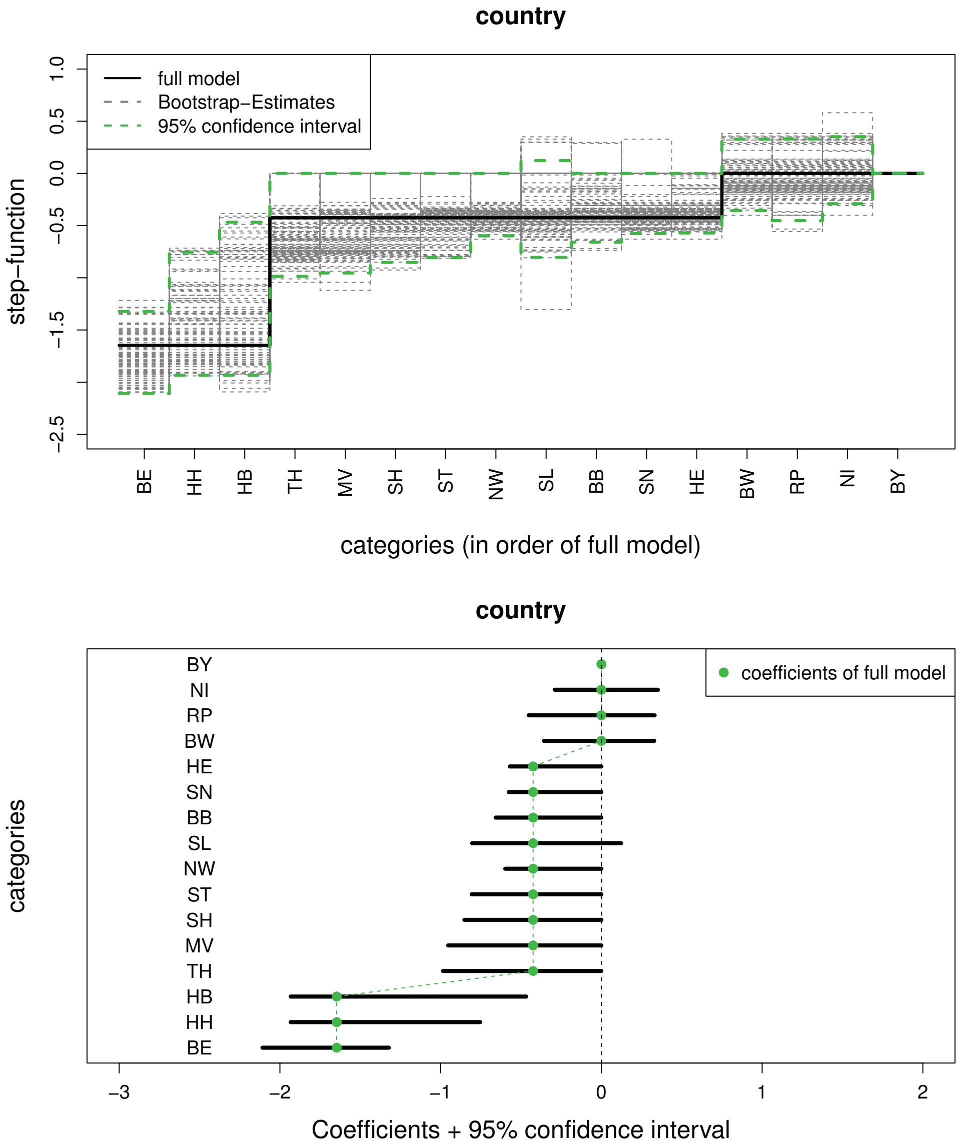


Figure 2.15.: Estimated step functions and resulting 95% confidence intervals for the nominal predictor country for the analysis of the household data based on 1000 bootstrap samples.

probability of outcome in the data. Only the confidence intervals of the big states Lower Saxony and Baden-Wuerttemberg als well as of Rhineland-Palatine and Saarland contain values greater than zero. The effects of the three cities Berlin, Bremen and Hamburg are significantly different from zero. The bootstrap interval of Bremen is very large due to a small number of observations.

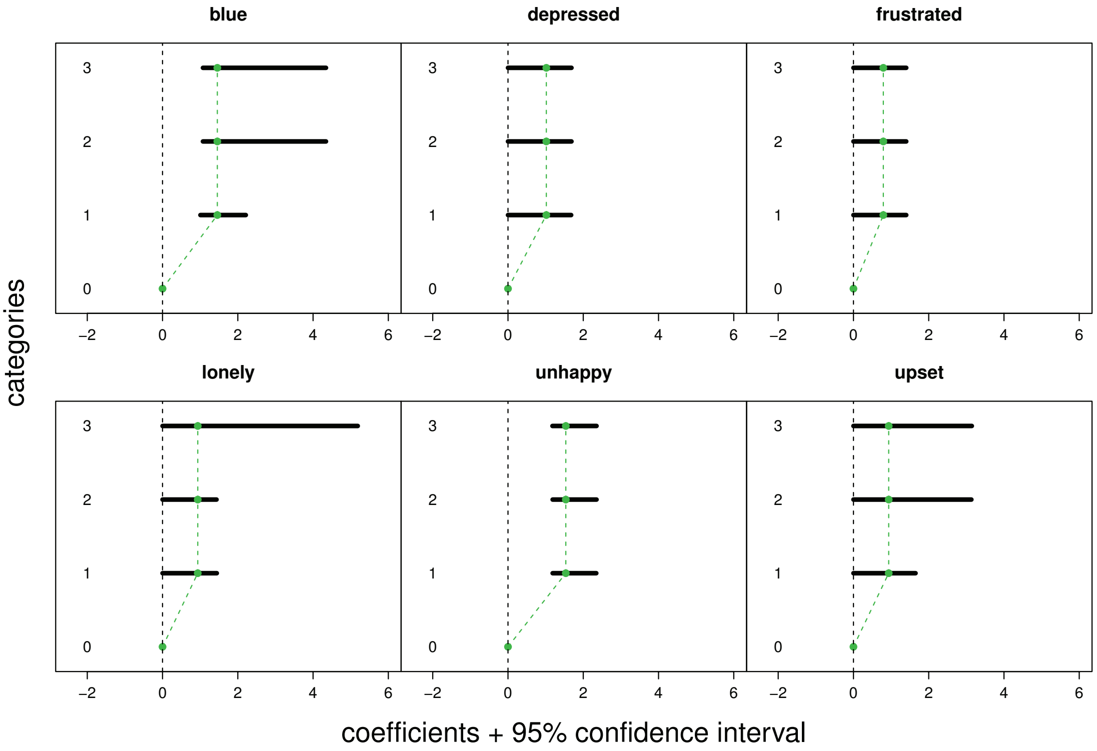


Figure 2.16.: Fitted coefficients of the full model (green dashed lines) and estimated 95% confidence intervals based on 1000 bootstrap samples for the six items of the MSQ data that are included in the model.

2.7.2. Motivational States Questionnaire

The third application concerns a comprehensive mood questionnaire, the so-called Motivational States Questionnaire (MSQ). It was developed to study emotions in laboratory and field settings. The data was collected between 1989 and 1998 at the Personality, Motivation, and Cognition Laboratory, Northwestern University (see Rafaeli and Revelle, 2006). The data is part of the R package `psych` (Revelle, 2013). The original version of the MSQ included 70 items. Due to a huge number of missing values we use a revised version of 68 items of 1292 participants for our analysis. The response format was a four-point scale that asks the respondents to indicate their current standing with the following scale: 0 (not at all), 1 (a little), 2 (moderatly), 3 (very much).

As response variable y we consider the indicator if the participant feels sad or not, generated from the answers given for the item that asks for being "sad". The probability of feeling sad is modeled by a logistic regression model as in the household data. The linear predictor consists of 67 ordinal predictors. Each predictor has four categories and corresponds to one item that was asked for in the questionnaire. There are no additional covariates, but the example illustrates that the model is able to handle a large number of ordinal predictors in the tree component.

The fitted coefficients and estimated 95% confidence intervals based on 1000 bootstrap samples for the predictors that are included in the model are shown in Figure 2.16. It is seen that only six variables among the 67 available variables were selected. Only the items that ask for being "blue", "depressed", "frustrated", "lonely", "unhappy" and "upset" are considered as being influential. Moreover, there is substantial clustering of the categories of the predictors. The coefficients of each predictor is a constant for level 1 to 3 reducing the ordinal predictors to binary predictors that distinguish between category 0 and the rest only. Bootstrap based confidence intervals are not the same for levels 1 to 3 in each case. Hence, there are bootstrap samples where the clusters consisting of level 1 to 3 are split a second time. Only for emotions "blue" and "unhappy" the confidence intervals do not contain zero. Thus it can be concluded that there are only 2 out of 67 emotions that have a significant effect on the probability of being sad.

2.7.3. Comparison with Alternative Models

In the previous sections the tree-based model was used to identify clusters in categorical predictors. Although prediction is not the main objective of the modelling strategy one expects any appropriate model to also perform well in terms of prediction accuracy. Therefore, we briefly compare the tree-based model with its main competitors with regard to prediction accuracy. Since in simulations typically one model, namely the data generating model, is preferred we consider the performance for the real data sets. The predictive deviance in both cases was measured by 5-fold cross-validation using 100 repetitions. As competing models we used the generalized additive model, a plain tree and model based partitioning. The generalized additive model was estimated by function `gam()` from package `mgcv` (Wood, 2011). The plain tree was estimated by use of the function `rpart()` from package `rpart` (Therneau et al., 2014). The complexity parameter 'cp' determines the minimal reduction of lack of fit. The optimal parameter was found to be 0.01 in both examples. Model based partitioning was estimated by the function `mob()` of package `party` (Zeileis et al., 2008). Predictors in the tree component of our model were used for partitioning. Predictors in the parametric part of our model were passed to models in each leaf. Complexity parameter 'trim', specifies the trimming in the parameter instability test. The optimal parameter was found to be 0.05 (rent) and 0.03 (car). Figure 2.17 shows the results for the rent data and the household data, respectively. It is seen that the tree-based model and GAM have comparable performance, which was to be expected since the tree-based model is essentially a GAM but with built-in clustering. The plain tree, with its focus on interaction shows much worse performance whereas model based partitioning performs poorly in one case and rather well in the other case.

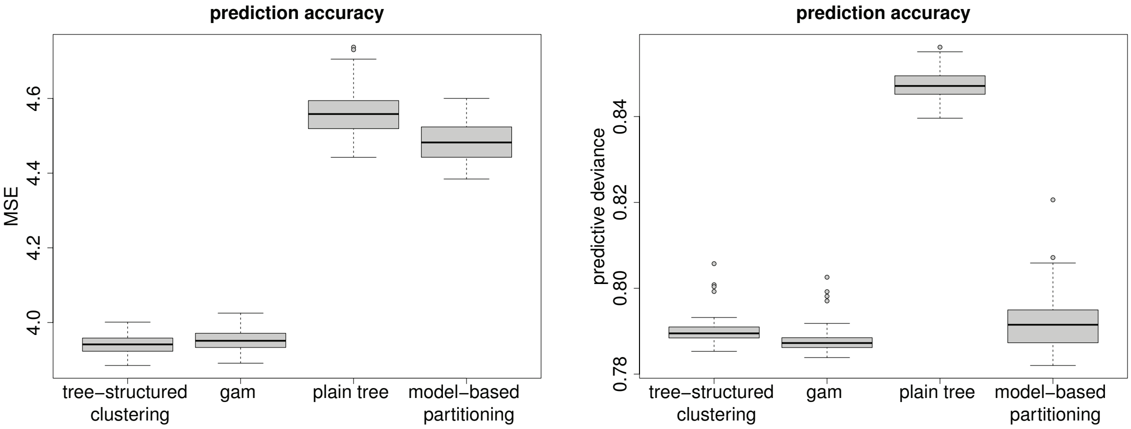


Figure 2.17.: Comparison of prediction accuracy of tree-structured clustering with other methods for the Munich rent standard data (left) and the household data (right).

2.8. Possible Extensions

In this section we will briefly sketch two extensions of the proposed tree-structured model that aim at improving clustering performance and model fit.

2.8.1. Stability Selection

The results of the simulations in Section 2.6.1 showed a satisfactory performance of the proposed model, but in particular for nominal predictors it is worth thinking about modifications to improve the fit in terms of clustering. One strategy we consider here is closely related to the concept of stability selection introduced by Meinshausen and Bühlmann (2010). Stability selection is a very general approach that can be applied to a broad range of existing methods. The main objective is to improve structure estimation by aggregation of estimates obtained by many subsamples. For simplicity of notation we consider the case of one categorical predictor only. The suggested algorithm that is build on the algorithm given in Section 2.3 is the following:

1. Fit the model for the whole sample.

2. Determine the number of clusters m from model in step 1.

3. Draw a bootstrap sample or subsample of predefined size, e.g. $\lfloor n/2 \rfloor$.

4. Fit the model for the sample drawn in step 3.

5. Keep the determined split-points C_s^* from the model in step 4.

6. Repeat step 3 to 5 for a predefined number of repetitions $s = 1, \ldots, S$.

7. Compute the selection probability of each split-point.

8. Choose $m-1$ split-points with the highest selection probability. Reduce the number of clusters m, if there are less than $m-1$ unequivocal maxima.

9. Refit the model for the whole sample using the split-points determined in step 8.

An initial analysis of the algorithm based on simulated data described in Section 2.6.1 showed slightly improved results, especially for the scenarios with $K = 20$ and $n = 100$. However, further investigations are needed to evaluate the usefulness of this extension.

2.8.2. Incorporation of Interactions

The focus of the proposed tree-structured model is on modelling of main effects of categorical predictors with many categories. This is in contrast to conventional trees where the terminal nodes usually correspond to higher order interactions. In an extension of model (2.1) or (2.2) however it is also possible to take interactions between the categorical predictors in the tree component into account. To be in line with the hierarchical principle in one step this means to simultaneously select two splits with regard to two variables and the corresponding interaction.

In order to preserve clarity we now change some notation. Concretely, for the pair of variables (j, r) and corresponding split-points (ℓ, k) the first split including an interaction means to fit the model with predictor:

$$\eta_i = \beta_0 + \alpha_{j\ell} I(z_{ij} > c_{j\ell}) + \alpha_{rk} I(z_{ir} > c_{rk}) + \gamma_{j\ell,rk} I(z_{ij} > c_{j\ell}) I(z_{ir} > c_{rk}) + \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where $\alpha_{j\ell}$, $\ell = 1, \ldots, m_\ell$ and α_{rk}, $k = 1, \ldots, m_r$ denote the main effects of regions $\{z_{ij} > c_{j\ell}\}$ and $\{z_{ir} > c_{rk}\}$ and $\gamma_{ik,j\ell}$ denotes the interaction between these two regions with regard to variables $j, r \in \{1, \ldots, q\}$.

The fitting procedure given in Section 2.3 can easily be adapted to this more general model. During iteration in each step there are four kinds of models that have to be investigated:

1. Selection of one split in one variable (as before).

2. Selection of two splits in two variables and the corresponding interaction.

3. Selection of an interaction between two splits in two variables that were already selected in previous steps.

4. Selection of one split in one variable and the interaction between the selected split and a split in another variable that was already selected in a previous step.

The degrees of freedom of the likelihood-ratio test in each case depends on the number of parameters that are involved in the splitting. In each step from all the candidate models the model is chosen that yields the best fit.

An initial analysis showed that the fitting procedure of the extended model with interactions works quite well. Nevertheless a huge number of interaction effects in the model can lead to a loss of interpretability. In particular the relation between two variables with several interaction effects with regard to different split-points is hard to overlook. Therefore, further research is needed to evaluate the performance of the model in simulations and the usefulness in applications.

2.9. Concluding Remarks

The proposed tree-structured approach is a modelling tool that allows to identify clusters in categorical predictors for nominal and ordinal predictors. In particular when several predictors with potentially many categories are available it is an efficient tool to reduce the superfluous complexity of classical parametric models. Simulation results show that the algorithm works well, in particular compared to the approach by penalized maximum likelihood estimation.

It should be noted that the tree-structured approach does not yield a tree in the sense of traditional recursive partitioning, where models are fitted recursively to sub samples defined by nodes. In the tree-structured model one obtains for each of the categorical predictors that are used in the tree component a separate tree. The obtained trees show which categories have to be distinguished given the other predictors are included in the model.

The results shown in this chapter were obtained by the `R` package `structree` (Berger, 2016b) version 1.0.1 that is available upon request and will presumably be made publicly accessible via CRAN.

3. Modelling Heterogeneity in Fixed Effects Models

3.1. Introduction

The analysis of longitudinal data and cross-sectional data that come in clusters requires to take the dependence of observations and the heterogeneity of measurement units into account. Typically, measurements within units tend to be more similar than measurements between units. If the heterogeneity is ignored poor performance of estimators and misleading standard errors are to be expected.

The most popular, widely used model to account for unobserved heterogeneity is the random effects model, see, for example, Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005) and McCulloch and Searle (2001). Typically in the random effects model it is assumed that the random effects follow a normal distribution. This strong assumption results in an economical model but inference may be sensitive to the specification of the distribution of random effects, see Heagerty and Kurland (2001), Agresti et al. (2004) and Litière et al. (2007). Several approaches to weaken the assumption of normally distributed random effects have been proposed. More flexible distributions are obtained, for example, by using mixtures of normals as proposed by Chen and Davidian (2002) and Magder and Zeger (1996). Huang (2009) proposed diagnostic methods for random-effect misspecification and Claeskens and Hart (2009) proposed tests for the assumption of the normal distribution. More recently, Lombardía and Sperlich (2012) proposed the class of semi-mixed effects models, a continuum of models that combine random and fixed effects.

An alternative approach to model heterogeneity uses finite mixtures. In finite mixtures of generalized linear models it is assumed that the density or mass function of the responses given the explanatory variables is determined by a finite mixture of components. Each of

This chapter is a modified version of Berger and Tutz (2015c). For more information on the personal contributions of the authors and textual matches, see page 9.

the components has its own response distribution and own parameters that determine the influence of explanatory variables. If only part of the parameters, for example the intercepts, are allowed to vary over components one obtains a discrete distribution of the heterogeneity part of the model. Models of that type were considered by Follmann and Lambert (1989) and Aitkin (1999). Follmann and Lambert (1989) investigated the identifiability of finite mixtures of binomial regression models and gave sufficient identifiability conditions for mixing of binary and binomial distributions. Grün and Leisch (2008b) considered identifiability for mixtures of multinomial logit models.

Finite mixture models replace the assumption of a fixed continuous distribution of random effects by the assumption of a discrete distribution. One may see this as an alternative and flexible specification of the heterogeneity component only. However, by assuming a discrete distribution of the intercepts instead of a continuous distribution as in random effects models one also implicitly assumes that there are clusters of units that share the same effect. In some applications it is definitely of interest to identify these units. We will consider an example in which the units are schools and one wants to know which schools are similar in their performance with regard to the education of students.

Here we consider an alternative to finite mixture models with the same objectives, that are use of a flexible discrete distribution and identification of units that share the same effect. However, the starting point is different. We use a fixed effects model in which each unit has its own parameter. An advantage is that no structural assumptions on the unit-specific effects have to be made. Clusters of parameters and therefore units with the same effect are found by tree methodology. The method proposed in the present chapter is related to the tree-based approach developed in Chapter 2. In the following it is adapted to a model including fixed effects for repeated measurements.

Classical recursive partitioning techniques or trees were first introduced by Morgan and Sonquist (1963). Very popular methods are classification and regression trees (CART) by Breiman et al. (1984) and C4.5 by Quinlan (1986) and Quinlan (1993). A newer version of recursive partitioning based on conditional inference was proposed by Hothorn et al. (2006). An overview on recursive partitioning in health science was given by Zhang and Singer (1999) and with a focus on psychometrics by Strobl et al. (2009). An easily accessible introduction into the basic concepts is found in Hastie et al. (2009).

The tree methodology used here differs from these approaches. In CART and other classical approaches the whole covariate space is recursively partitioned into subspaces. In order to obtain a partitioning in the intercepts (or slopes) only, one has to apply a different form of trees. It has to be designed in a way that the subspaces are built for specific effects only, for example the intercepts, while other parameters that represent common effects of explanatory variables are not partitioned into subspaces. Our main focus is on the clustering of intercepts, however, we will also refer to the case of unit-specific slopes.

One big advantage using recursive partitioning techniques is the computational efficiency. The proposed tree-structured model especially enables the evaluation of high-dimensional data. Alternative approaches to identify clusters within a fixed effects model framework as proposed by Tutz and Oelker (2016) fail in high dimensional settings.

This chapter is organized as follows: In Section 3.2 we introduce the tree-structured model for unit-specific intercepts and in section 3.3 we present an illustrative example. Details about the fitting procedure are given in Section 3.4. After a short introduction of related approaches in Section 3.5 we give the results of wider simulation studies (Section 3.6). Section 3.7 contains further applications. Finally, in Section 3.8 we consider the extension to models with unit-specific slopes and give a small example.

3.2. Accounting for Heterogeneity in Clustered Data

Consider clustered data given by $(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, $i = 1, \dots, n$, $j = 1, \dots, n_i$, where y_{ij} denotes the response of measurement j for unit i and two sets of predictive variables $\boldsymbol{x}_{ij}^\top = (1, x_{ij1}, \dots, x_{ijp})$ and $\boldsymbol{z}_{ij}^\top = (1, z_{ij1}, \dots, z_{ijq})$. In longitudinal data the units can, for example, represent persons that are measured repeatedly. In the following, we consider alternative methods to account for the potential heterogeneity of units. We start with methods that use random effects, then consider fixed effects model and finite mixtures.

3.2.1. Random Effects Models

In a generalized linear mixed model (GLMM) the mean response $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ is linked to the explanatory variables by

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i, \tag{3.1}$$

where $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$ is a linear term which contains the fixed effect $\boldsymbol{\beta}$. The second term $\boldsymbol{z}_{ij}^\top \boldsymbol{b}_i$ contains the random effects for covariates $\boldsymbol{z}_{ij}$ that are varying across units and $g(\cdot)$ is a known link function. In a GLMM it is assumed that the distribution of $y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}$ follows a simple exponential family and that the observations y_{ij} are conditionally independent. For the random effects $\boldsymbol{b}_i$, which model the heterogeneity of the units, one typically assumes a normal distribution $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{rand})$.

In a GLMM the distribution of the random effects is used to account for the heterogeneity of the units and the focus is mainly on the parametric term $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Although the distributional assumption for the random effects makes the estimation of the model very efficient there are also some disadvantages. If the assumed distribution is very different from the real data generating distribution, inference can be biased. The assumption of a continuous distribution also does not allow for the same effects of different units. Hence, clustering of units is not possible. Another crucial point of the GLMM is the assumption that the random effects $\boldsymbol{b}_i$ and the covariates $\boldsymbol{x}_{ij}$ are uncorrelated. This assumption can lead to poor estimation accuracy, see, for example, Grilli and Rampichini (2011). Functions for the estimation of generalized linear mixed models are provided in the `R` package `lme4` (Bates et al., 2015), which we will use for the computations in the applications and simulations.

3.2.2. Fixed Effects Models

In contrast to mixed models, fixed effects models model heterogeneity among units by using one parameter $\boldsymbol{\beta}_i$ for each unit. The mean response $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ is linked to the explanatory variables in the form

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{\beta}_i, \tag{3.2}$$

where $\boldsymbol{x}_{ij}$ again is a vector of covariates that have the same effect across all units and $\boldsymbol{z}_{ij}$ contains covariates that have different effects over units. Each measurement unit has his own parameter vector $\boldsymbol{\beta}_i^\top = (\beta_{i0}, \dots, \beta_{iq})$. The specification of one parameter vector per unit results in a very large number of parameters which can affect estimation accuracy. Moreover, typically there is not enough information to distinguish between all units. To cope with these problems one can assume that there are groups of units that share the same effect on the response. Forming clusters of units leads to a reduced number of parameters and stable estimates. There are several strategies to identify these clusters, the fixed effects model with regularization or the finite mixture model (see next sections).

3.2.3. Tree-Structured Clustering

In the approach considered here one assumes that the fixed effects model holds, but not all the unit-specific parameters are assumed to be different. Clusters (or groups) of measurement units are identified by recursive partitioning methods. We first consider unit-specific intercepts only. Let us start with the simplest case in which all intercepts are equal, that is, the linear predictor has the form $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_0$. If there are two clusters the corresponding linear predictor is given by

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_{i0}^{(k)}, \quad k = 1, 2, \tag{3.3}$$

where k denotes if the unit is in the first or the second group. A simple test, for example a likelihood ratio test, for the hypothesis $H_0 : \beta_{i0}^{(1)} = \beta_{i0}^{(2)}$ can be used to determine if the model with two groups is more adequate for the data than the model in which all the intercepts are equal. By iterative splitting into subsets guided by test statistics one obtains a clustering of units that have to be distinguished with regard to their intercept.

In general, regression trees can be seen as a representation of a partition of the predictor space. A tree is built by successively splitting one node A, that is already a subset of the predictor space, into two subsets A_1 and A_2 with the split being determined by only one variable. In a fixed effects model, when specifying specific intercepts for each unit, the unit number itself can be seen as a nominal categorical variable with n categories. The partition has the form $A \cap S_1, \quad A \cap S_2$, where S_1 and S_2 are disjoint, non-empty subsets $S_1 \subset \{1, \dots, n\}$ and its complement $S_2 = \{1, \dots, n\} \setminus S_1$. Using this notation another representation of model (3.3) is given by

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_{i0}^{(1)} I(i \in S_{10}) + \beta_{i0}^{(2)} I(i \in S_{20}),$$

where $I(\cdot)$ denotes the indicator function with $I(a) = 1$, if a is true and $I(a) = 0$ otherwise. After several splits one obtains a clustering of the units $\{1, \dots, n\}$ and the predictor of the resulting model can be represented by

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \sum_{k=1}^{m_0} \beta_{i0}^{(k)} I(i \in S_{k0}), \tag{3.4}$$

where $S_{10}, \dots, S_{m_0 0}$ is a partition of $\{1, \dots, n\}$ consisting of m_0 clusters that have to be distinguished in terms of their individual intercepts. Model 3.4 can be seen as a special case of the model proposed in Chapter 2 including only one nominal predictor. In the following we will use the model abbreviation TSC for tree-structured clustering.

3.2.4. Finite Mixture Models

An alternative approach that also allows to identify clusters of units are finite mixture models. These were, for example, considered by Follmann and Lambert (1989) and Aitkin (1999). The general assumption in finite mixtures of generalized regression models is that the mixture consists of m components where each component follows a parametric distribution of the exponential family of distributions. The density of the mixture can be given by

$$f(y|\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{k=1}^{m} \pi_k f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k),$$

where $f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k)$ denotes the k-th component of the mixture with parameter vector $\boldsymbol{\beta}_k$ and dispersion parameter ϕ_k. For the unknown component weights π_k $\sum_{k=1}^{m} \pi_k = 1$ and $\pi_k > 0$, $k = 1, \dots, m$ has to hold.

Table 3.1.: Summary statistics of the test score of the 56 multiple-choice items and covariate gender of the illustrative example (CTB data).

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	x_{max}
Test score	21	32	34	34.14	37	46
Gender		male: 761			female: 739	

Here we consider models with components that differ in their intercepts. Within the framework of finite mixtures one specifies for the $k-$th component of the mixture a model with predictor $\eta_{ij}^{(k)} = \beta_{i0}^{(k)} + x_{ij}^{\top}\boldsymbol{\beta}$. For models with normal response the mixture components are given by $N(y_{ij}|\eta_{ij}^{(k)}, \sigma^2)$, where the variance σ^2 is fixed for all components. For models with a binary response the mixture components are $B(y_{ij}|n, \pi_{ij}^{(k)})$, where $\pi_{ij}^{(k)} \in (0, 1)$ and $\text{logit}(\pi_{ij}^{(k)}) = \eta_{ij}^{(k)}$. For further details, see Grün and Leisch (2007).

Estimation of the mixture model is usually obtained by the EM-algorithm with the number of components m being specified beforehand. The optimal number of components is chosen afterwards, for example by information criteria like AIC or BIC. Grün and Leisch (2008a) provide the R-package `flexmix`, which is used for the computations in our applications and simulations. Regularization and variable selection for mixture models have been considered by Khalili and Chen (2007) and Städler et al. (2010) but not with the objective of clustering units with regard to their effects.

3.3. An Illustrative Example

Before giving details how to grow trees and estimate the proposed model (3.4) we want to illustrate the procedure by use of an application. We consider a data set from CTB/McGraw-Hill, a division of the Data Recognition Corporation (DRC). For a description of the original data, see De Boeck and Wilson (2004). The data includes results of an achievement test that measures different objectives and subskills of subjects in mathematics and science. For our investigation we use the results of 1500 grade 8 students from 35 schools. They had to respond to 56 multiple-choice items (31 mathematics, 25 science). The response y_{ij} is the overall test score of student j in school i, defined as the number of correctly solved items. The main objective is to adequately describe the heterogeneity of the 35 schools. As additional covariate we include the gender of the students (male: 0, female: 1). The

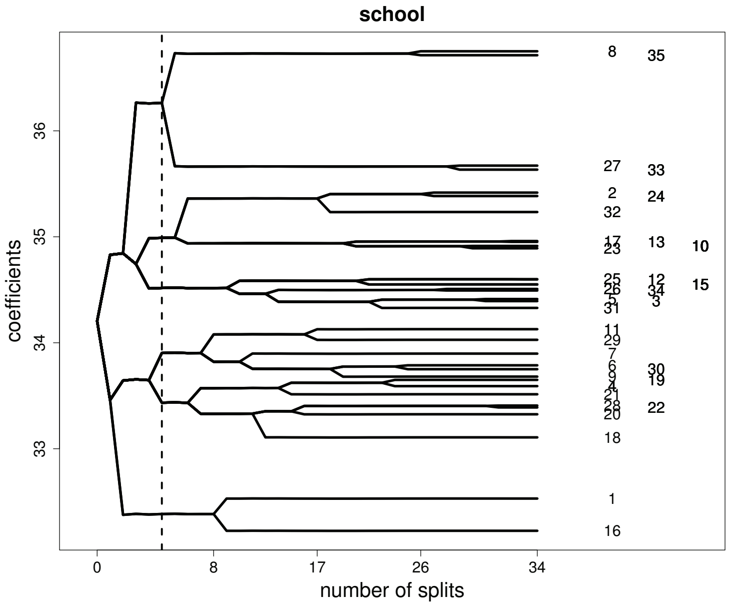


Figure 3.1.: Paths of coefficients of school-specific intercepts against all splits of the illustrative example (CTB data). The optimal number of splits is marked by a dashed line.

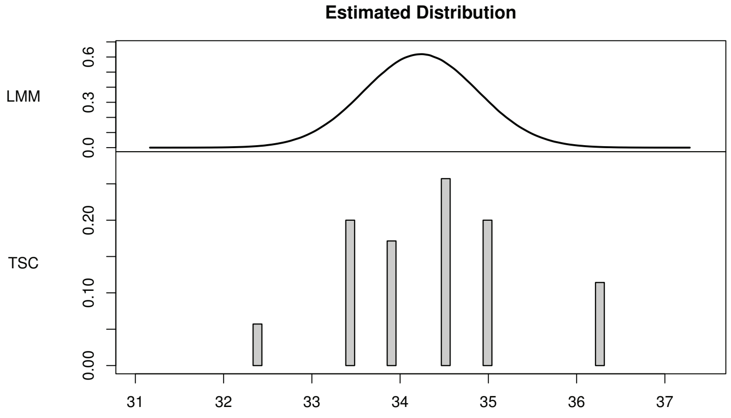


Figure 3.2.: Comparison of the estimated distribution of the mixed model and the school-specific intercepts of tree-structured clustering (CTB data).

summary statistics of the test scores and the covariate gender is given in Table 3.1. By using the proposed tree-structured approach the model that was obtained has the form

$$\mu_{ij} = \beta_G \cdot G_{ij} + \sum_{k=1}^{m_0} \beta_{i0}^{(k)} I(i \in S_{k0}), \quad i = 1, \ldots, 35,$$

where $G_{ij} \in \{0, 1\}$ denotes the gender of student j in school i, $S_{10}, \ldots, S_{m_0 0}$ is a partition of the 35 schools and $\beta_{i0}^{(k)}, k = 1, \ldots, m_0$, denote the effects of the corresponding clusters.

Table 3.2.: Estimation results of the illustrative example (CTB data) using the classical mixed model, tree-structured clustering and the finite mixture model.

Predictor	**LMM**		**TSC**		**FIN**	
	Coefficient	95%-CI	Coefficient	95%-CI	Coefficient	95%-CI
gender	-0.106	[-0.475, 0.298]	-0.088	[-0.478, 0.313]	-0.084	[-0.473, 0.309]
β_0	34.235	[33.964,34.542]	—	—	—	—
σ^2_{rand}	0.416	[0.394, 1.353]	—	—	—	—

School-specific intercept	**TSC**		**FIN**	
	Cluster	Coefficient	Cluster	Coefficient
β_{i0}	1,16	32.384	1,4,6,7,9,16,	33.508
	4,18,19,20,21,22,28	33.434	18,19,20,21,	
	6,7,9,11,29,30	33.904	22,28,30	
	3,5,12,14,15,25,26,31,34	34.517	2,3,5,8,10,11,12,13,14,	34.689
	2,10,13,17,23,24,32	34.999	15,17,23,24,25,26,27,	
	8,27,33,35	36.264	29,31,32,33,34,35	

The coefficient paths of the school-specific intercepts obtained by tree-structured clustering are shown in Figure 3.1. The coefficient paths build a tree that successively partitions the schools in terms of the performance of students. The left end refers to the global intercept estimated as an average over the 35 schools. On the right end of the coefficient paths all possible splits have been performed and the estimated coefficients correspond to those of a simple fixed effects model without clustering. The optimal number of splits that is selected by the algorithm, is marked by the dashed line. It is seen that estimates change strongly in the first steps, but after about ten splits the estimates are very stable.

A graphical comparison of the estimated normal distribution of the random effects using a classical linear mixed model (LMM) and the distribution of the school-specific intercepts of the tree-structured model (TSC) is shown in Figure 3.2. It illustrates the main advantage of the tree-structured model. There is no distributional assumption on the school-specific intercepts, especially no assumption of symmetry. The number of schools in each cluster are quite different and not symmetric. Clustering of similar schools strongly reduces the complexity of the fixed effects model and makes interpretation of school-specific differences very easy. There are two small clusters of schools where the performance in the test considerably deviates upwards or downwards, the differences between the clusters with medium performance are smaller.

Table 3.2 shows an overview of the estimation results obtained by using the classical linear mixed model (LMM), the proposed tree-structured model (TSC) and a finite mixture model (FIN), where only the intercepts are allowed to vary over the components. Confidence intervals are obtained by using bootstrap procedures, where the model is fitted repeatedly on sub samples of size n that are obtained by drawing with replacement. The results here are obtained by 2000 sub samples. It is seen that all of the methods did not find a significant effect for covariate gender. The performance of males and females seems not to differ systematically. The variance obtained by the mixed model is significantly different from zero, which suggests that heterogeneity of schools is definitely present. The lower panel in Table 3.2 shows the estimated partition of schools obtained by the tree-structured model and the finite mixture model. In the latter case, model selection by AIC and BIC both yield the same result. Tree-structured clustering identifies six clusters of schools until further splits are no longer significant (for details of the algorithm see Section 3.4). The finite mixture approach identifies only two clusters of schools. This illustrates the tendency of the finite mixture approach to find a small number of clusters, which will be investigated later. For comparison in Table 3.2 the schools that belong to the two clusters found by the finite mixture model are coloured in black and grey.

3.4. Fitting Procedure

In this section we give details of the algorithm that yields the tree-structured model. Let us again consider the model with unit-specific intercepts after the first split, which has the form

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_{i0}^{(1)} I(i \in S_{10}) + \beta_{i0}^{(2)} I(i \in S_{20}). \tag{3.5}$$

When determining the first split for the nominal predictor $i \in \{1, \dots, n\}$ one has to consider all possible partitions of the two subsets S_{10} and S_{20}. Altogether there are $2^{n-1} - 1$ possible splits, which can be a very large number. It has been shown in earlier research that it is not necessary to consider all possible partitions, see Breiman et al. (1984) and Ripley (1996) for binary outcomes and Fisher (1958) for quantitative outcomes. It is sufficient to order the predictor categories, here the measurement units, with respect to the means of the response and to treat the predictor as if the categories were ordered. In a first step, units are ordered according to their maximum-likelihood estimates, so that $\hat{\beta}_{(10)} \leq \hat{\beta}_{(20)} \leq \ldots \leq \hat{\beta}_{(n0)}$. Then one considers splits of adjacent measurement units to obtain the optimal split. To use this simplification one starts with an equivalent representation of model (3.5) given by

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_0 + \alpha_{i0} I(i > c),$$

with $\beta_{i0}^{(1)} = \beta_0$ and $\beta_{i0}^{(2)} = \beta_0 + \alpha_{i0}$. The set C of possible thresholds c is from $\{1, \dots, n-1\}$. The fitting procedure considered in the following uses this model as building block. By iterative splitting of adjacent measurement units the searched-for clustering is obtained.

Basic Algorithm

The basic algorithm for the model with unit-specific intercept is the following.

Tree-Structured Clustering – Unit-specific intercept

$\mathcal{S}$tep 1 (Initialization)

(a) Estimation: Fit the candidate GLMs with predictors

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_0 + \alpha_{i0} I(i > c_{i0}), \quad c_{i0} = 1, \dots, n-1$$

(b) Selection

Select the model that has the best fit. Let $c^*_{i_1 0}$ denote the best split.

$\mathcal{S}$tep 2 (Iteration)

For $\ell = 1, 2, \dots,$

(a) Estimation: Fit the candidate models with predictors

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \beta_0 + \sum_{s=1}^{\ell} \alpha_{i_s 0} I(i > c^*_{i_s 0}) + \alpha_{i0} I(i > c_{i0}),$$

for all values $c_{i0} \in C \setminus \{c^*_{i_1 0}, \dots, c^*_{i_\ell 0}\}$

(b) Selection

Select the model that has the best fit yielding the split-point $c^*_{i_{\ell+1} 0}$.

In each selection step of the algorithm one has to identify the best split and during iterations one has to decide when to stop. Common splitting criteria for tree-based methods are impurity measures that have already been introduced by Breiman et al. (1984). An alternative is to use a test statistic to evaluate which split most improves the explanatory power of the predictors. We will draw on the latter concept and use a procedure that is strongly related to the conditional inference framework proposed by Hothorn et al. (2006). In each iteration one examines the null hypotheses $H_0 : \alpha_{i0} = 0$ for all remaining possible split-points. This can, for example, be tested by a likelihood-ratio test. To determine the best split we simultaneously consider all test statistics T_{i0} from the set of possible splits c_{i0} and choose the split-point for which T_{i0} had the largest value. For illustration Figure

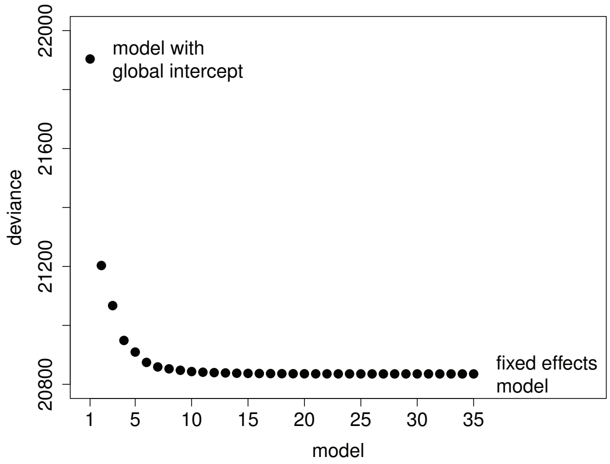


Figure 3.3.: Deviances of all the models selected during the fitting procedure of the illustrative example (CTB data).

3.3 shows the deviances obtained for all models of the illustrative example (Section 3.3). The value on the left corresponds to the deviance of the model with a global intercept, the deviance on the right end corresponds to the fixed effects model with an individual intercept for each school. The deviances strongly decrease in the first steps but after about 10 splits the model fit does not improve considerable any more. In the first step the model with the best fit is found for split-point 15 ($c^*_{15_1 0}$). The corresponding test statistic is obtained by building the difference between the first two values given in Figure 3.3, namely $T_{15,0} = 21903.77 - 21203.07 = 700.7$. The test statistics in the following steps can be computed accordingly.

Stopping Criterion

Since each likelihood ratio test statistic asymptotically follows a chi-squared distribution, in each step one additionally obtains a p-value associated with the test statistic T_{i0} of the selected split. To determine the optimal number of splits *one strategy* is to stop if the p-value exceeds a certain pre-specified threshold. This strategy was already proposed in Chapter 2. In each step one should take into account the number of possible splits and adapt for multiple testing errors. Given overall significance level α one simply uses the Bonferroni procedure and stops if $p_\ell > \alpha/(n-(\ell-1))$ because in the $\ell - th$ iteration there are $n-(\ell-1)$ possible splits. Thus, the overall error rate is under control.

A *second strategy* is to check if the heterogeneity of measurement units is already modelled sufficiently in each step. Before executing one further split one tests the global null hypothesis that the current model completely captures the heterogeneity of the data against the alternative that the data is more heterogeneous. To decide for the first split one has to

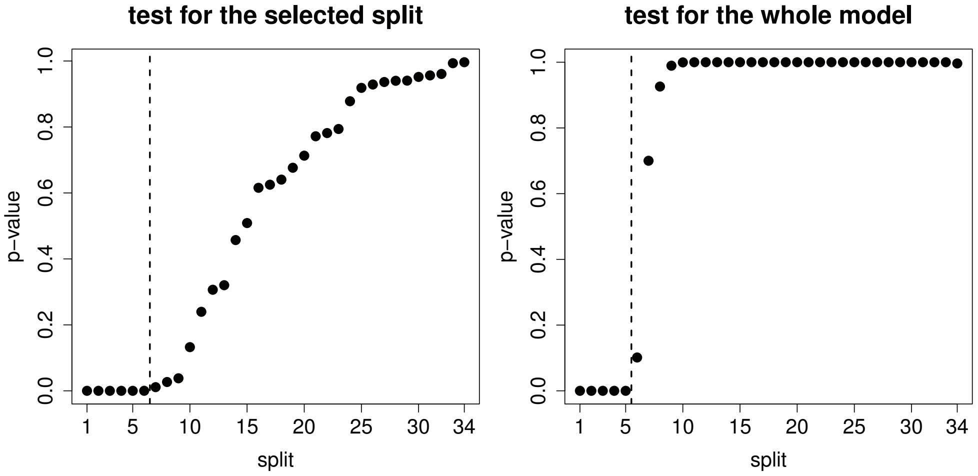


Figure 3.4.: P-values obtained for the illustrative example (CTB data) using different stopping criteria. The left panel shows the p-values associated with the selected split, the right panel shows the p-values when using a global test incorporating all unit-specific parameters of the current model. The selected number of splits is marked by a dashed line, respectively.

examine the null hypothesis $H_0 : \beta_{10} = \beta_{20} = \ldots = \beta_{n0}$, which corresponds to the case of no heterogeneity. The hypothesis is tested by a likelihood-ratio test with significance level α and $n-1$ degrees of freedom, because $n-1$ differences of parameters are tested. Depending on the significance of this global test the selected split or no splitting is performed. In the illustrative example the test statistic in the first step is obtained by building the difference of deviance of the model with global intercept and the deviance of the fixed effects model, that is $T_1 = 21903.77 - 20835.30 = 1068.47$ (see Figure 3.3) on 34 degrees of freedom. After several splits only differences of units within already built clusters are tested. In the $\ell - th$ step $n - \ell$ differences have to be tested because $\ell - 1$ splits are already performed. If a significant effect is found the selected split is performed, otherwise splitting is stopped. We prefer to use the second strategy in our simulations and applications because this stopping criterion leads to a clear separation of the selection of splits and the splitting decision. In particular the splitting decision is only minor influenced by the previously identified ordering of measurement units.

In detail the p-values for the illustrative example obtained by the two stopping criteria are given in Figure 3.4. In addition the selected number of splits is marked by dashed lines, respectively. The left panel shows the p-values that correspond to the test statistics T_{i0} of the selected splits, the right panel shows the p-values that correspond to the test statistics using the global hypotheses. Based on the first strategy the algorithm detects seven clusters, whereas according to the second strategy there are only six clusters (as given in Table 3.2). As was to be expected, in both cases the sequence of p-values is increasing, but with a considerable flatter slope in the left panel.

The result of the fitting procedure is a sequence of m_0-1 selected split-points $c^*_{i_10}, \dots, c^*_{i_{m_0-1}0}$ and corresponding parameter estimates $\hat{\alpha}_{i_10}, \dots, \hat{\alpha}_{i_{m_0-1}0}$. Ordering of the selected split-points yields the desired clustering of ordered units $\{1, \dots, c^*_{(i_10)}\}$, $\{c^*_{(i_10)}+1, \dots, c^*_{(i_20)}\}$, … , $\{c^*_{(i_{m_0-1}0)}+1, \dots, n\}$. The corresponding intercepts $\beta^{(k)}_{i0}$ for each cluster are then given by

$$\hat{\beta}^{(k)}_{i0} = \hat{\beta}_0 + \sum_{s=1}^{k-1} \hat{\alpha}_{(i_s0)}, \quad k = 1, \dots, m_0.$$

During the iterations only the selected split-points but no estimates from previous steps are kept. All coefficients of the models, including the parameters $\boldsymbol{\beta}$ of the linear term, are refitted in each step and the final estimates are those from the last iteration.

3.5. Related Approaches

In the following we will briefly consider alternative methods that account for unobserved heterogeneity and are related to our tree-structured model. One of the approaches is a competitor to the method proposed here and will also be included in the simulations.

Clustering of units can also be obtained by penalized maximum likelihood estimation as proposed more recently by Tutz and Oelker (2016). Let $\boldsymbol{\beta}_0^T = (\beta_{10}, \dots, \beta_{n0})$ denote the intercepts of the fixed effects model. An estimation procedure that identifies clusters is obtained by maximizing the penalized log-likelihood $l_p(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = l(\boldsymbol{\beta}, \boldsymbol{\beta}_0) - \lambda J(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$, where $l(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ denotes the unpenalized log-likelihood, $J(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ is a specific penalty term and λ is a tuning parameter. The penalty term that enforces clustering of unit-specific intercepts is given by

$$J(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \sum_{r>s} |\beta_{r0} - \beta_{s0}|,$$

where only pairwise differences of the unit-specific intercepts are included. If $\lambda = 0$, one obtains the unpenalized maximum-likelihood estimates and each unit has his own intercept. If $\lambda \to \infty$, all units are fused to one cluster with the same intercept. For a comparison we use the corresponding R-package `gvcm.cat` proposed by Oelker (2015) in our simulations. The use of such penalties in ANOVA was already proposed by Bondell and Reich (2008) and for variable selection by Gertheiss and Tutz (2010) and Tutz and Gertheiss (2014). A problem with the method is that the penalty contains $n(n-1)/2$ differences and therefore the algorithm becomes extremely demanding for large values of n. It typically fails if the number of groups is larger than 50 or 60.

The method proposed here should be distinguished from the mixed effects regression trees (MERT) proposed by Hajjem et al. (2011) and the RE-EM trees, which were independently proposed by Sela and Simonoff (2012). The basic concept is to combine a linear mixed effects model for clustered data and a standard regression tree. The substantial difference is that the tree is not applied to the random or unit-specific effects of the model but to the fixed effects term. The predictor of the estimated model has the form $\eta_{ij} = f(\boldsymbol{x}_{ij}) + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i$, where $\boldsymbol{b}_i \sim N(0, \boldsymbol{\Sigma}_{rand})$. It is the function $f(\boldsymbol{x}_{ij})$ that is estimated by a standard regression tree. The model yields random effects that are node-invariant and therefore does not focus on the similarity of units but rather on the dissimilarity of observations within units.

An alternative Bayesian approach to model clustered random effects is based on Dirichlet processes. Dirichlet processes were proposed by Ferguson (1973) and studied, for example, by Sethuraman (1994) and Hjort et al. (2010). The main advantage of Dirichlet processes is their cluster property, which allows to flexibly model discrete distributions. Assuming a Dirichlet process for the distribution of random effects creates ties among the random effects. The resulting Dirichlet process mixture yields clusters of units. Dirichlet process priors have been used within the linear mixed model framework by Bush and MacEachern (1996) and Müller and Rosner (1997). A frequentist approach to linear mixed models with Dirichlet process mixtures was given by Heinzl and Tutz (2013), a combination of Dirichlet processes and fusion penalties was considered in Heinzl and Tutz (2014), Heinzl and Tutz (2016). The approach works for linear models, but extensions to generalized mixed models seem not available.

3.6. Simulations

In the following we investigate the performance of the proposed tree-structured model and compare it to competing methods. The focus is on data settings with clusters of units that share the same effect on the response and where the strict assumptions of the mixed model do not hold. We are in particular interested in the estimation accuracy and the clustering performance. We will compare the generalized fixed effects model (GFM), the generalized mixed model (GMM), the tree-structured model (TSC), the model based on penalized maximum-likelihood estimation (PENL), the finite mixture model with model selection by AIC (FINA) and the finite mixture model with model selection by BIC (FINB).

We consider several simulation scenarios where the overall number of observations is 800, made up of the components $n = 200/n_i = 4$, $n = 100/n_i = 8$, $n = 40/n_i = 20$ or $n = 20/n_i = 40$. In addition to the unit-specific intercepts we include one continuous covariate x_1 with $x_{ij1} \sim N(0,1)$ and one binary covariate x_2 with $x_{ij2} \sim B(1, 0.5)$. Unit-specific intercepts β_{i0} are drawn symmetrically from a normal distribution or are drawn from a chi-square distribution that is skewed. In order to obtain clusters of units, the intercepts are sorted according to size and divided into balanced groups. The average over the intercepts of each group is defined as the new unit-specific intercept $\beta_{i0}^{(k)}$, $k = 1, \dots, m_0$. We consider scenarios with $m_0 \in \{5, 10\}$. Therefore, the true simulated size of clusters varies between 2 for the scenarios with $n = 20$, $m_0 = 10$, and 40 for the settings with $n = 200$, $m_0 = 5$.

Correlation between Intercepts and Covariates

An important assumption of the mixed model is that the unit-specific intercepts are independent from the predictors $\boldsymbol{x}$. In order to break this assumption we simulate data with correlations $\rho = \text{corr}(\beta_{i0}, x_{ij1}) \neq 0$. For the simulation we use a sequential procedure adopted from Tutz and Oelker (2016). Consider the case of normal distributed intercepts β_{i0}. Here, values are first generated by $\beta_{i0} \sim N(\mu_b, \sigma_b^2)$ and $x_{ij1} \sim N(0, 1)$. Afterwards x_{ij1} is transformed according to the bivariate normal distribution of (β_{i0}, x_{ij1}) with the corresponding correlation. We consider scenarios with $\rho \in \{0, 0.8\}$. In the case of chi-squared distributed intercepts the joint distribution of (β_{i0}, x_{ij1}) is not bivariate normal, but we can use the same transformation for x_{ij1} yielding the same empirical correlations.

Evaluation Criteria

We compare the estimated coefficients to the true parameters by calculating mean squared errors (MSEs). We distinguish between the MSE of the unit-specific intercepts $\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{i0} - \beta_{i0})^2$, referred to as *intercepts*, and the MSE of the effects of the two covariates $\frac{1}{2}\sum_{d=1}^{2}(\hat{\beta}_d - \beta_d)^2$, referred to as *linear term*. Concerning the mixed model, coefficients $\hat{\beta}_{i0}$ are computed as the sum of the estimated posteriori modes and the fixed intercept $\hat{\beta}_0$. In addition the number of clusters determined by the different approaches are of interest. All the presented evaluations are based on 100 replications.

3.6.1. Normal Response

We start with simulation scenarios where the responses y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, n_i$ are normally distributed with $\varepsilon_{ij} \sim N(\mu_\varepsilon = 0, \sigma_\varepsilon^2 = 3^2)$. Here we set $\beta_1 = \beta_2 = 2$ as the true parameters of the two covariates. In the first case we consider cluster-specific intercepts that were generated from the fusion of parameters that follow a standard normal distribution.

It is important to mention that in the above setting the effective number of parameters for the mixed model heavily depends on the variance σ_ε^2 of the response and the variance σ_b^2 of the random intercepts. Following Ruppert et al. (2003), the effective degrees of freedom for the random intercepts for a linear random intercept model are

$$df_b = \frac{(n-1)n_i}{n_i + \frac{\sigma_\varepsilon^2}{\sigma_b^2}}.$$

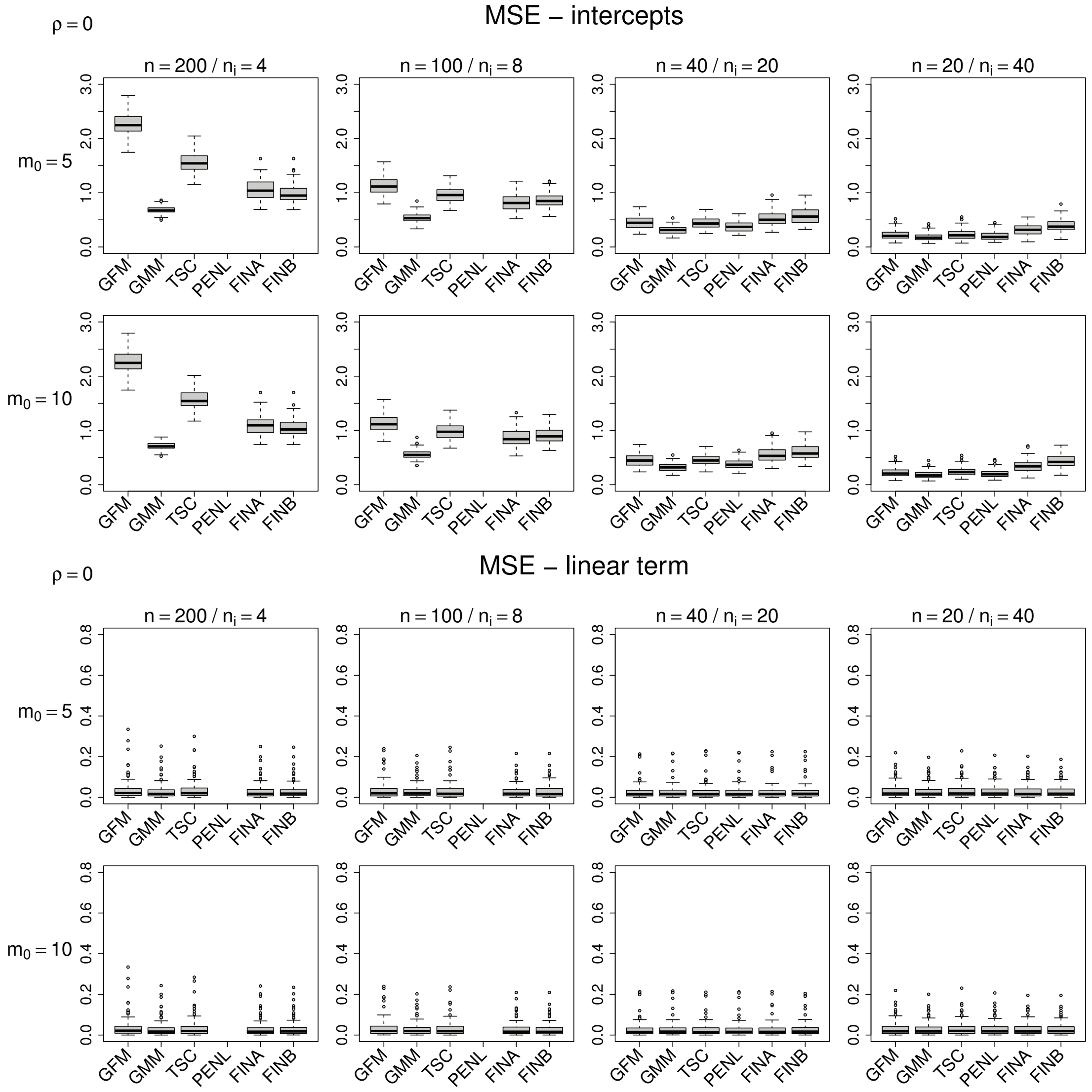


Figure 3.5.: MSEs of intercepts (upper panel) and the linear term (lower panel) for the settings with normal response, normal intercepts and $\rho = 0$.

If $\sigma_b^2 \to 0$ or $\sigma_\varepsilon^2 \to \infty$ the result is a model with only one intercept and if $\sigma_b^2 \to \infty$ or $\sigma_\varepsilon^2 \to 0$ the result is a model with n intercepts, corresponding to the fixed effects model. With $\sigma_\varepsilon^2 = 9$ and $\sigma_b^2 = 1$ one obtains the effective degrees of freedom 61.2, 46.5, 26.9 and 15.5 depending on the combination of parameters n and n_i. Therefore, one is not too close to the fixed effects model, which allows a fair comparison of the mixed model and the tree-structured model. In the second case with a skewed distribution for the unit-specific intercepts we use $\beta_{i0} \sim \chi^2(0.5)$ with $\sigma_b^2/2 = 0.5$ degrees of freedom. After centering of the coefficients one obtains the same empirical values $\mu_b = 0$ and $\sigma_b^2 = 1$ as in the standard normal case.

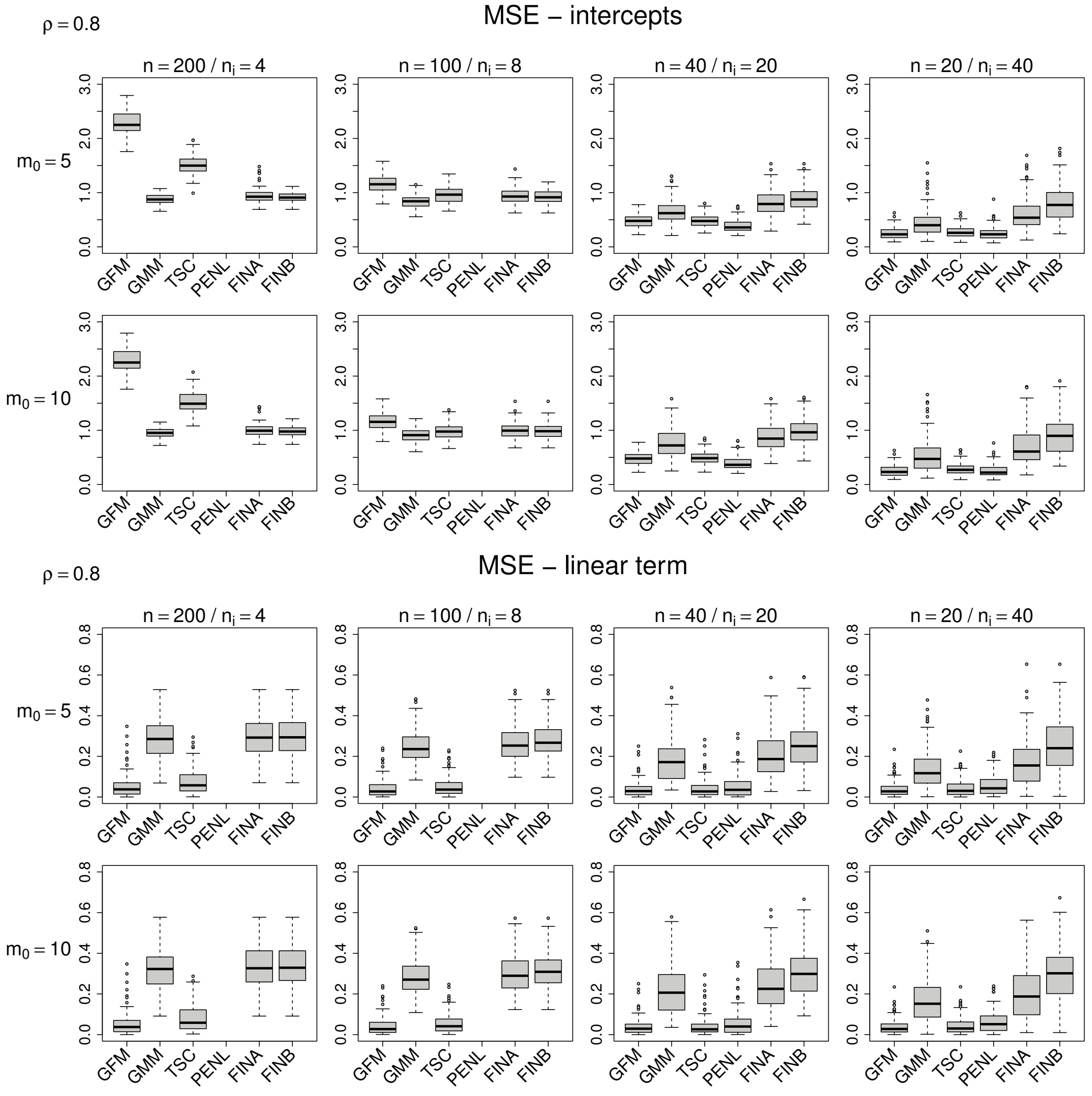


Figure 3.6.: MSEs of intercepts (upper panel) and the linear term (lower panel) for the settings with normal response, normal intercepts and $\rho = 0.8$.

Figure 3.5 shows the boxplots of the MSEs for the eight different settings generated by normally distributed intercepts and without correlation ($\rho = 0$). As the approach by penalized likelihood estimation is computational infeasible for a large number of units n, no results are displayed for the settings with $n = 200$ and $n = 100$. It is seen from the lower panel that all the approaches nearly show the same performance for the linear term. However, distinct differences are seen for the intercepts (upper panel). Although there are clusters of units the mixed model shows good performance for all settings. The fixed effects model performs poorly, especially for the settings with $n_i = 4$, the finite mixture model performs poorly for the settings with $n = 40$ and $n = 20$. The estimates of the tree-structured model show better performance than the fixed effects model for smaller values

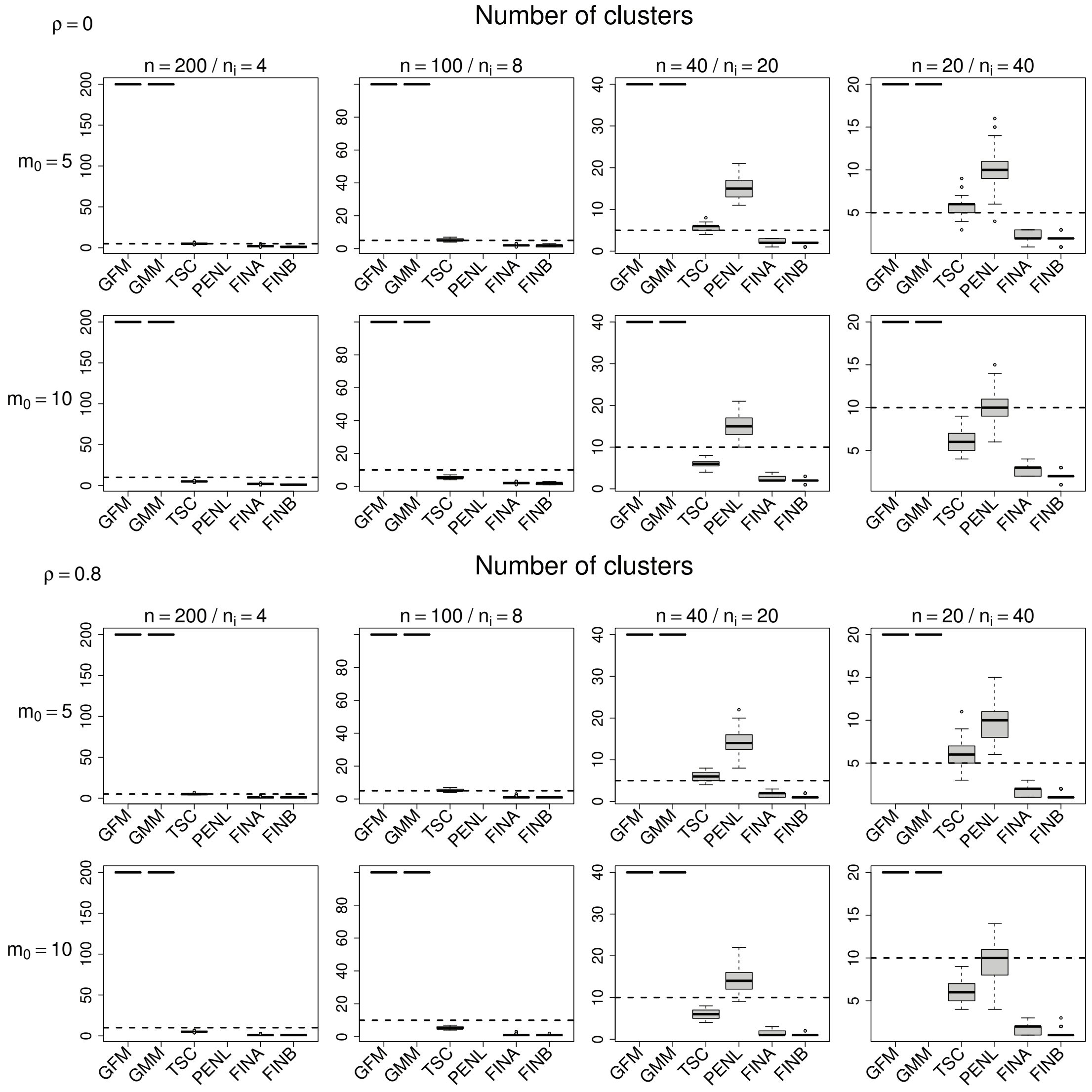


Figure 3.7.: Selected number of clusters for the settings with normal response, normal intercepts, $\rho = 0$ (upper panel) and $\rho = 0.8$ (lower panel). The true number of clusters m_0 is marked by dashed lines.

of n_i and comparable performance for larger values. The performance is the same as for the penalty approach if estimates exist. The picture changes in the settings with correlation $\rho = 0.8$ between covariate x_1 and the unit-specific intercepts (Figure 3.6). For the linear term (lower panel) the performance of the mixed model and the finite mixture model suffers strongly. In contrast, the estimation accuracy of the fixed effects model, the tree-structured model and the penalized likelihood approach is not affected by the correlation. In particular, the tree-structured model outperforms the penalty approach in all the settings in which the penalty approach works. The results for the intercepts (upper panel) do not change that

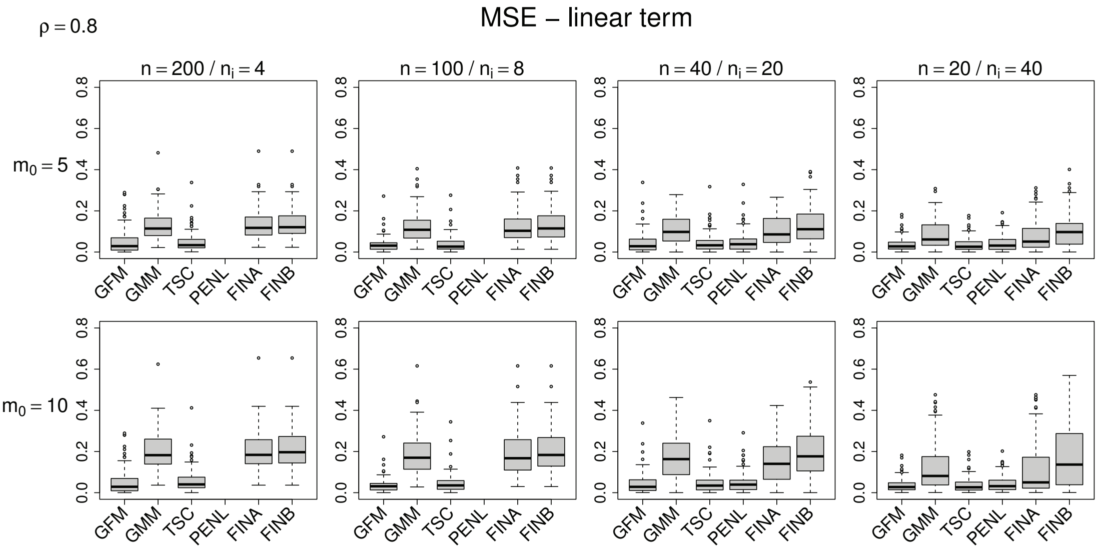


Figure 3.8.: MSEs of the linear term for the settings with normal response, chi-squared intercepts and $\rho = 0.8$.

much but the mixed model and the finite mixture model is now competitive only for small values of n_i.

Boxplots of the selected number of clusters are given in Figure 3.7 for $\rho = 0$ (upper panel) and $\rho = 0.8$ (lower panel). Since the fixed effects model and the mixed model do not build clusters of units, the given number of clusters for the two approaches is equal to the number of units. There are only minor differences between the settings with and without correlation. The number of clusters identified by the tree-structured model is very close to the true number for the settings with five clusters ($m_0 = 5$) but the true number of clusters is slightly underestimated in the settings with ten clusters. In contrast, the penalty approach selects a distinctly higher number of clusters with a strong variation. The finite mixture model consistently selects only too small number of clusters. On average only about two clusters are selected by AIC as well as by BIC.

The evaluations of the same settings with cluster-specific intercepts that were generated by a chi-squared distribution yield very similar results. In particular the performance of the mixed model seems not to be affected too strongly by the skewed distribution of the random intercepts. For illustration Figure 3.8 shows the MSEs of the linear term for the settings with $\rho = 0.8$. See Appendix B, page 209 for an overview of all results.

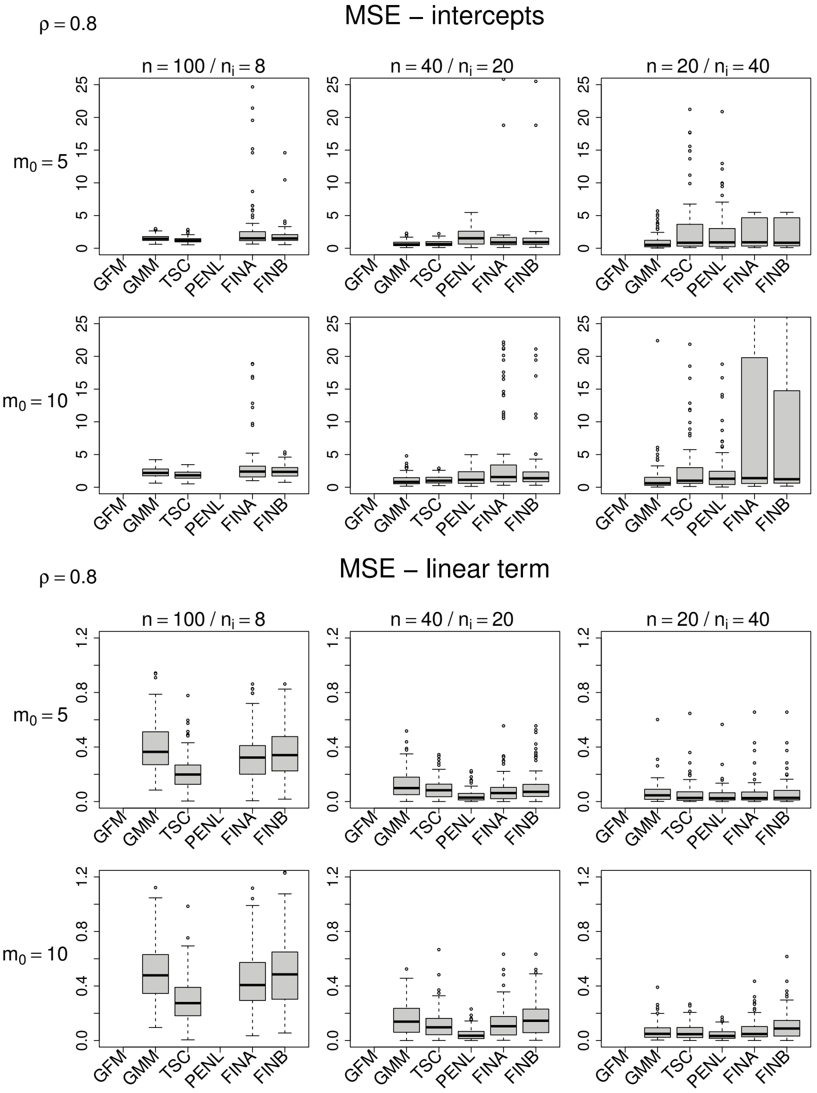


Figure 3.9.: MSEs of intercepts (upper panel) and the linear term (lower panel) for the settings with binary response, chi-squared intercepts and $\rho = 0.8$.

3.6.2. Binary Response

In the following we briefly consider discrete response variables $y_{ij} \sim B(1, \pi_{ij})$, where $\pi_{ij} = \exp(\eta_{ij})/(1 + \exp(\eta_{ij}))$. The structure of the simulated data sets remains the same but some modifications to the specifications in Section 3.6.1 are necessary. The parameters of the linear term are set to $\beta_1 = \beta_2 = 0.1$. For the cluster-specific intercepts we chose $\beta_{i0} \sim N(-0.8, 2^2)$ or as skew counterpart $\beta_{i0} \sim \chi^2(2)$, centered such that $\mu_b = -0.8$. Since $n_i = 4$ is a relatively small size when modelling binary responses, we do not consider the corresponding settings. Furthermore, we omit the estimates of the fixed effects model because they are very unstable and often do not exist in this case. Accordingly, the order

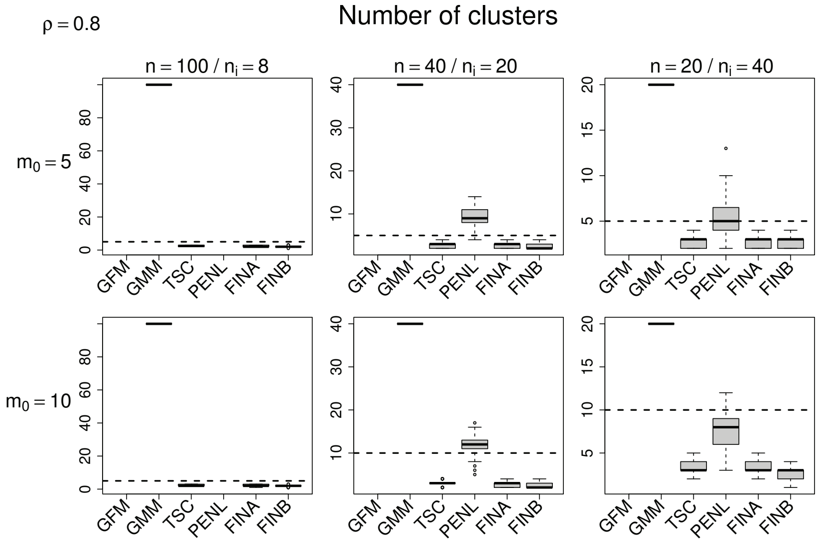


Figure 3.10.: Selected number of clusters for the settings with binary response, chi-squared intercepts and $\rho = 0.8$. The true number of clusters m_0 is marked by dashed lines.

of measurement units used in the algorithm of the tree-structured model is not based on the estimates of the unrestricted model but by adding a small ridge penalty.

In contrast to the settings with normal response, the results for the binary response as a whole seem to be more affected by a skewed distribution of the intercepts. In the following we will focus on the settings with chi-squared distributed intercepts and $\rho = 0.8$, and refer to Appendix B, pages 210 and 211 for further results. Figure 3.9 shows the MSEs of the unit-specific intercepts (upper panel) and the linear term (lower panel). Again the mixed model and the finite mixture model perform poorly with regard to the linear term, but there are only minor differences for $n = 20$. Regarding the intercepts the average results are comparable for all the approaches. It is noticeable that one observes huge outliers for the finite mixture models, especially with model selection by AIC. It is most conspicuous for the settings with $n = 20$, where the boxplots have been truncated.

The corresponding boxplots of the selected number of clusters are given in Figure 3.10. Here the tree-structured model only detects very few clusters (for $m_0 = 5$ and $m_0 = 10$) and is almost as restrictive as the finite mixture model. As before the penalty approach selects a higher number of clusters and has a stronger variation but the selected number of clusters is closer to the true number of clusters.

Table 3.3.: Estimation results of the beta blocker data using the classical mixed model, tree-structured clustering and the finite mixture model.

Predictor	**GMM**		**TSC**		**FIN**	
	Coefficient	95%-CI	Coefficient	95%-CI	Coefficient	95%-CI
treatment (yes/no)	-0.130	[-0.183,-0.084]	-0.131	[-0.184,-0.085]	-0.129	[-0.183,-0.084]
β_0	-2.326	[-2.413,-2.270]	—	—	—	—
σ^2_{rand}	0.236	[0.192, 0.357]	—	—	—	—

Center-specific intercept	**TSC**		**FIN**	
	Cluster	Coefficient	Cluster	Coefficient
β_{0i}	13,14,18,19,22	-2.969	13,14,18,19,22	-2.963
	1,2,3,4,5,6,8,10,11,21	-2.401	1,2,3,4,5,6,8,	-2.379
	7,9,17	-1.946	9,10,11,17,21	
	12,15,16,20	-1.567	7,12,15,16,20	-1.739

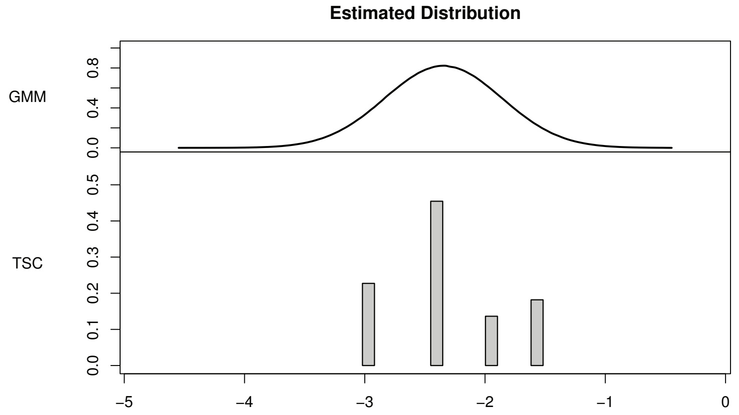


Figure 3.11.: Comparison of the estimated distribution of the mixed model and the center-specific effects tree-structured clustering (beta blocker data).

3.7. Further Applications

In the following we give the results of two further real data examples with binary response and compare them to the alternative approaches.

3.7.1. Beta Blocker

As second application we use a dataset that has already been considered by Aitkin (1999), Grün and Leisch (2008a) and Tutz and Oelker (2016). The data was collected in a 22-center clinical trial to investigate the effect of beta blockers on the mortality after myocardial infarction. In each center patients were divided into a test group (treatment = 1) and a control group (treatment = −1). The total number of patients is 20290, whereby the number of patients per center varies strongly over centers. The binary response of interest is if the patient deceased ($y_{ij} = 1$) or not ($y_{ij} = 0$). It is modelled by a logistic regression model $\text{logit}(P(y_{ij} = 1)) = \eta_{ij}$. The heterogeneity among the center, more precise the basic risk for a decease, is captured in the center-specific intercepts and shall be modelled adequately.

The results by the alternative approaches considered here are given in Table 3.3. The table contains estimated coefficients and 95% confidence intervals obtained by 2000 bootstrap samples. There is a significant treatment effect. The estimated parameters of all methods indicate that the probability of a decease decreases for the test group by the factor 0.88. The variance component of the mixed is small but significantly different from zero. This allows the conclusion to be drawn that centers do not differ very much but their heterogeneity can not be neglected. The partitions and corresponding effects of center-specific intercepts found by the tree-structured model and the finite mixture model are given in the lower panel of Table 3.3. Regarding the finite mixture model we prefer to use model selection by BIC as it showed more stable estimates in the simulations with binary response. It can be seen, that the estimated coefficients for all clusters are negative, as the probability of staying alive in principle is much higher than the probability of a decease. The finite mixture model detects three clusters, whereas according to the tree-structured model there are four clusters of centers that have to be distinguished in terms of their basic risk. It is noticeable that the cluster with the lowest probability containing five centers is exactly the same for both methods with very similar estimates.

A comparison of the estimated normal distribution of the mixed model and the center-specific effects of the tree-structured model is visualized in Figure 3.11. The main advantage of the tree-structured model compared to a mixed model is again pointed out in the figure. There is no distributional assumption on the center-specific intercepts, which allows that the number of centers in each cluster is quite different and not symmetric.

3.7.2. National Survey in Guatemala

In a third application we consider data derived from the National Survey of Maternal and Child Health in Guatemala in 1987. The data is available from the R-package `mlmRev` (Bates et al., 2014) and was also analysed by Rodriguez and Goldman (2001). The data contains observations of children that were born in the 5-year period before the survey. In our analysis we include 1211 children living in 45 communities. One observes a minimal number of 20, a maximal number of 50 and an average number of 26.9 pregnancies per community. The response y_{ij} is a binary outcome with $y_{ij} = 0$ for traditional prenatal care and $y_{ij} = 1$ for modern prenatal care, for example by doctors or nurses. As in the previous

Table 3.4.: Description and distribution of the covariates used for the analysis of the Guatemala survey.

Variable	Description	Categories	Frequency
ethn	Mother's ethnicity	non-indigenous (Ladino)	612
		indigenous, not speaking Spanish	286
		indigenous, speaking Spanish	313
momEd	Mother's level of education	not finished primary	571
		finished primary	607
		finished secondary	33
husEd	Husband's level of education	not finished primary	430
		finished primary	598
		finished secondary	67
		unknown	116
husEmpl	Husband's employment status	unskilled	45
		professional	120
		agricultural, self-employed	420
		agricultural, employee	407
		skilled service	219
telev	Frequency of TV usage	never	1034
		not daily	52
		daily	125
momAge	Mother 25 years or older	no	583
		yes	628
toilet	Modern toilet in house	no	112
		yes	1099

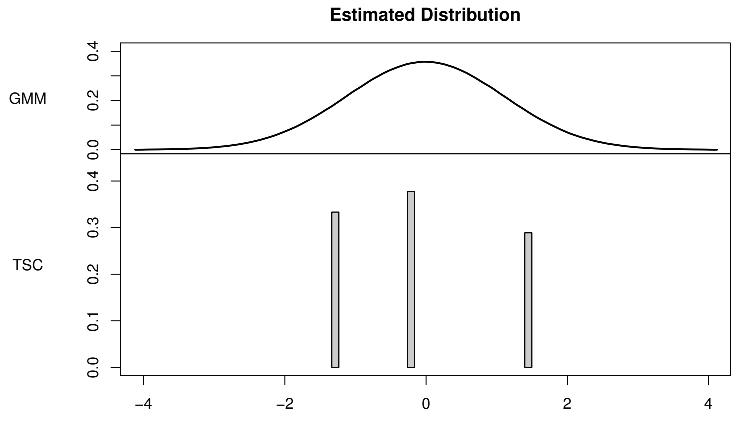


Figure 3.12.: Comparison of the estimated distribution of the mixed model and the community-specific intercepts of tree-structured clustering (Guatemala survey).

example the response is modelled by a logistic regression model. The heterogeneity of communities is modelled by the alternative approaches considered here. In total there are 733 pregnancies with traditional and 478 observed pregnancies with modern prenatal care.

Table 3.5.: Estimation results of the Guatemala survey using the generalized mixed model, tree-structured clustering and the finite mixture model.

Predictor	**GMM**		**TSC**		**FIN**	
	Coefficient	95%-CI	Coefficient	95%-CI	Coefficient	95%-CI
ethn						
not spanish	-1.370	[-2.101,-0.774]	-1.090	[-2.469,-0.387]	-0.995	[-2.280,-0.556]
spanish	-0.720	[-1.235,-0.244]	-0.434	[-1.425, 0.005]	-0.335	[-1.338, 0.011]
momEd						
primary	0.645	[0.331, 1.048]	0.673	[0.298, 1.122]	0.646	[0.317, 1.078]
secondary	1.385	[0.303, 2.955]	1.405	[0.268, 3.046]	1.735	[0.364, 2.944]
husEd						
primary	0.785	[0.445, 1.236]	0.817	[0.437, 1.303]	0.843	[0.444, 1.301]
secondary	0.194	[-0.809, 1.186]	0.049	[-0.922, 1.286]	0.291	[-0.846, 1.311]
unknown	0.398	[-0.113, 0.951]	0.520	[-0.101, 1.006]	0.428	[-0.106, 0.962]
husEmpl						
professional	-0.210	[-1.150, 0.670]	-0.095	[-1.301, 0.820]	-0.408	[-1.336, 0.667]
agricult, self	-0.119	[-0.975, 0.721]	-0.065	[-1.044, 0.798]	-0.266	[-1.065, 0.716]
agricult, empl	-0.158	[-1.024, 0.656]	-0.100	[-1.092, 0.750]	-0.238	[-1.103, 0.723]
skilled	-0.199	[-1.079, 0.606]	-0.125	[-1.123, 0.661]	-0.300	[-1.134, 0.607]
telev						
not daily	0.355	[-0.497, 1.292]	0.226	[-0.601, 1.286]	0.241	[-0.548, 1.283]
daily	0.867	[0.312, 1.560]	0.928	[0.290, 1.570]	0.735	[0.307, 1.524]
momAge	0.099	[-0.208, 0.403]	0.061	[-0.241, 0.411]	0.061	[-0.219, 0.401]
toilet	-0.869	[-1.833,-0.055]	-1.008	[-1.875, 0.092]	-0.839	[-1.808,-0.154]
β_0	-0.011	[-1.223, 1.166]	—	—	—	—
σ^2_{rand}	1.250	[1.233, 2.416]	—	—	—	—

Community-specific intercept	**TSC**			**FIN**		
	Cluster	Size	Coefficient	Cluster	Size	Coefficient
β_{i0}	1	15	-1.286	1	33	-0.696
	2	17	-0.214	2	12	1.465
	3	13	1.448			

The two binary and five categorical explanatory variables that characterize the children's mothers and their families are given in Table 3.4.

An overview of the estimated coefficients when using a generalized mixed model (GMM), tree-structured clustering (TSC) and a finite mixture model (FIN) is given in Table 3.5. The 95% confidence intervals were obtained by 2000 bootstrap samples. It can be seen from the results that the age of the mother at the time of the survey as well as the employment status of the husband do not have a significant effect on the form of prenatal care. The educational level of the mother as well as of the husband, however, have a strong impact. For births where the mother at least finished primary or the husband finished primary modern prenatal care was provided more likely compared to births of parents without any graduation. Indigenous mothers (speaking and not speaking Spanish) are also more likely to use traditional prenatal care than non-indigenous mothers. The existence of a modern toilet in the household does not favour the use of modern prenatal care, whereas it is preferred by families using the television regularly.

Table 3.6.: Summary statistics of the mathematics score and the weekly time in hours spent on math homework (NELS).

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	x_{max}
Mathematics score	34.99	42.03	48.64	51.51	60.66	77.20
Homework (HW)	0	1	1	2.02	3	7

A comparison of the estimates obtained by the three methods does not show strong distinctions and no clear tendency. Differences occur for variable ethnicity (first rows in Table 3.5), for which the two estimates of the mixed model are larger than for TSC and FIN and for mothers that finished secondary (fourth row) for which the estimate of the finite mixture model is larger than for TSC and GMM.

The estimated community-specific intercepts obtained by tree-structured clustering and the finite mixture model are given in the lower panel of Table 3.5. Using the tree-structured model results in three clusters of communities that differ in terms of their probability to use modern prenatal care, whereas the finite mixture (selected by BIC) identifies only two clusters. The detected partitions and the high variance obtained by the mixed model indicate that heterogeneity of communities is definitely present. Nevertheless, only a few clusters of communities have to be distinguished. There is a strong similarity between the third cluster of the tree-structured model ($\beta_{i0}^{(3)} = 1.448$) and the second cluster of the finite mixture model ($\beta_{i0}^{(2)} = 1.465$) but as a whole the partition of tree-structured clustering seems to be more adequate. In Figure 3.12 the estimated distribution of the community-specific intercepts of the tree-structured model and the estimated normal distribution of the mixed model are graphically illustrated.

3.8. Extension to Group-Specific Slopes

So far we limited our considerations to the case of a group-specific intercept, where $z_{ij} = 1$. However, the general fixed effects model (3.2) allows for more than one parameter to be unit-specific. It is straightforward to extend the tree-structured model to include a covariate vector $\boldsymbol{z}_{ij} = (1, z_{ij1}, \ldots, z_{ijq})$. Then one obtains a model with predictor

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \sum_{r=0}^{q} \sum_{k=1}^{m_r} z_{ijr} \beta_{ir}^{(k)} I(i \in S_{kr}), \tag{3.6}$$

where $S_{1r}, \ldots, S_{m_r r}$ is a partition of the units $\{1, \ldots, n\}$ with respect to the r-th component of $\boldsymbol{z}_{ij}$ and $\beta_{ir}^{(1)}, \ldots, \beta_{ir}^{(m_r)}$ are the corresponding parameters of each cluster. Due to individual splits, the number and form of clusters do not have to be the same for the different components of $\boldsymbol{z}_{ij}$. The fitting procedure given in Section 3.4 can easily be adapted to this general model. In each iteration one simply has to determine the best split among all covariates and all corresponding splits simultaneously. In a first step the order of the units $\{1, \ldots, n\}$ with respect to single covariates has to be defined. It is not assumed that the order is the same for each of the covariates. The result is one tree for each covariate that represents a partition of units. This extended model is simply a special case of the model proposed in Chapter 2 with several nominal predictors.

Application: National Education Longitudinal Study

As an example we consider data of the National Election Study (NELS) of 1988. For a detailed description, see Curtin et al. (2002). For our analysis we use a subsample of 260 grade 8 students from 10 schools, with an average number of 26 students per school. The response y_{ij} is the standardized mathematics score of student j in school i, that is measured between 0 and 100 and thus assumed to be Gaussian. Next to the school itself the weekly time in hours spent on maths homework (HW) will serve as explanatory variable. The summary statistics of the response and the covariate are given in Table 3.6. To explain the mathematics score it is reasonable to assume that the effect of covariate HW differs across schools. Therefore in our model besides school-specific intercepts we include school-specific slopes with respect to covariate HW. By using the extended tree-structured approach the model that was obtained has the form

$$y_{ij} = \sum_{k=1}^{m_0} \beta_{i0}^{(k)} \; I(i \in S_{k0}) + \sum_{k=1}^{m_{HW}} HW_{ij} \; \beta_{i,HW}^{(k)} \; I(i \in S_{k,HW}), \quad i = 1, \ldots, 10,$$

where $S_{10}, \ldots, S_{m_0 0}$ denotes the partition of schools regarding their intercepts with effects $\beta_{i0}^{(k)}$ and $S_{1,HW}, \ldots, S_{m_{HW},HW}$ denotes the partition of schools regarding their effect of the time spent on maths homework with effects $\beta_{i,HW}^{(k)}$. Figure 3.13 shows the coefficient paths obtained for the school-specific intercepts of the two components in the model. In analogy to Figure 3.1 the paths of the school-specific intercepts are given in the left panel. Furthermore one observes paths of school-specific slopes that are given in the right panel. As there are 10 schools the maximal number of splits in each tree component is nine, giving an overall number of 18 splits in the model (displayed on the x-axis). In total the algorithm performs 11 splits, marked by dashed lines in Figure 3.13, until further splits are no longer significant. It can be seen that estimates change strongly until stopping. The final model defines 7 clusters of schools sharing the same intercept, that is the same average mathematical competence. For the effect of the time spent on maths homework

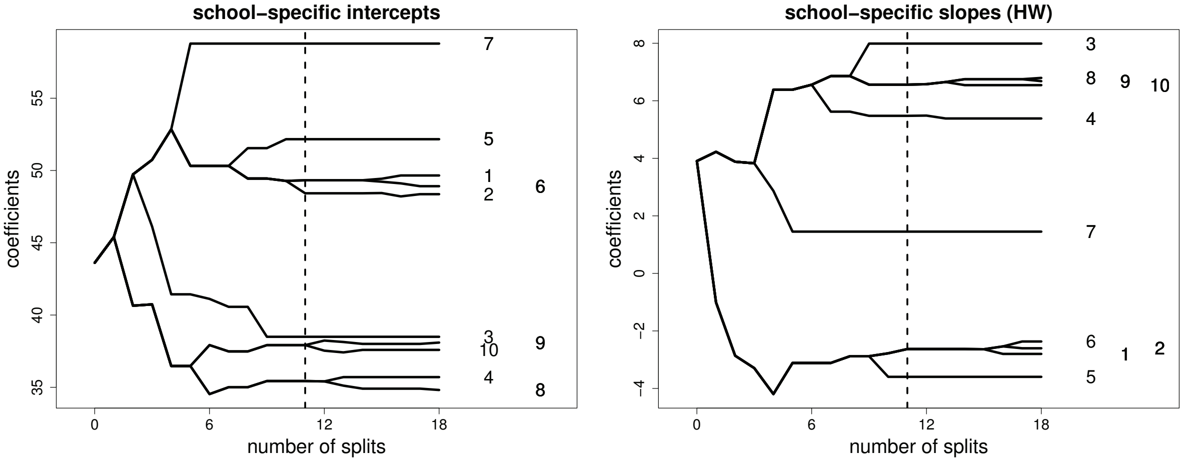


Figure 3.13.: Paths of coefficients of school-specific intercepts (left panel) and school specific slopes of variable HW (right panel) against all splits (NELS). The optimal number of splits is marked by dashed lines.

Table 3.7.: Estimation results for school-specific intercepts given in the left columns and for school-specific slopes of variable HW given in the right columns (NELS).

	Cluster	Coefficient		Cluster	Coefficient
β_{i0}	4,8	35.433	$\beta_{i,HW}$	5	-3.596
	9,10	37.917		1,2,6	-2.630
	3	38.949		7	1.452
	2	48.423		4	5.477
	1,6	49.324		8,9,10	6.560
	5	52.165		3	7.988
	7	58.780			

one obtains 6 clusters. The partitions of the two components, in detail given in Table 3.7, are quite different. Regarding the intercepts (left columns) there are three clusters composed of two schools while the other schools have their individual effect. Regarding the slopes (right columns) there are two clusters composed of three schools while the others have their individual effects. It is conspicuous that for cluster $\{1, 2, 6\}$ and school 5 the average mathematical competence is comparably high but the estimated effects of HW is actually negative. Obviously in this schools the weekly time spent on maths homework is an indicator for students with week performance. The opposite effect is seen for school 3 with an low average mathematical competence but a very large positive effects of HW. Here the time spent on maths homework has a favorable influence.

3.9. Concluding Remarks

The proposed tree structured model competes well with the competitors. In particular, it performs better than the finite mixture approach and has the advantage that the number of units is not restricted as in the penalty approach. The applications were chosen to illustrate the potential of the method to find clusters that share the same effect on the response. The potential of the method to yield better estimates when the heterogeneity and explanatory variables are correlated is demonstrated in the simulations.

The results shown in this chapter were obtained by the R-package `structree` (Berger, 2016b) version 1.0.1 that is available upon request and will presumably be made publicly accessible via CRAN.

4. Identification of Differential Item Functioning in Rasch Models

4.1. Introduction

Differential item functioning (DIF) is a well known problem in item response theory. It occurs if the probability of a correct response among equally able persons differs in subgroups, for example, if the difficulty of an item depends 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). In particular Magis et al. (2010) gave an excellent overview of the existing DIF detection methods.

The traditional approach to identify items that carry DIF is based on test statistics. For each item a test is performed that shows if the item has different difficulties in subgroups that have to be defined by the experimenter. Test statistics have been proposed by Thissen et al. (1993), Lord (1980), Holland and Thayer (1988), Kim et al. (1995) and Raju (1988). Mixed model approaches were proposed by Van den Noortgate and De Boeck (2005) and Bayesian approaches have been developed by Soares et al. (2009).

The classical testing approach with a focus on sub groups is not without problems. First, when testing it is assumed that all other items are free of DIF, which is an assumption that typically does not hold, see also Magis et al. (2010). Second, the proposed tests are limited to the consideration of few subgroups. Typically one considers just two subgroups with one group being fixed as the reference group. That means if one suspects item difficulties to

This chapter is a modified version of Tutz and Berger (2015a). For more information on the personal contributions of the authors and textual matches, see page 9.

depend on age one has to know the age groups before testing. Thus age has to be split into two or more intervals without knowing which ones are relevant. Moreover, the approaches are restricted to subgroups. Therefore, it is hard to investigate the dependence on more than one possibly DIF inducing variable.

More recently, several methods have been proposed to cope with these problems. Tutz and Schauberger (2015) proposed 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 abundance of parameters in the model is handled by using penalization techniques. An alternative regularization method that uses the logistic regression approach to DIF detection was proposed by Magis et al. (2015). A further approach that is also able to handle several groups and continuous variables was proposed by Strobl et al. (2015). It avoids the comparison of pre-specified focal and reference group by using recursive partitioning techniques, also known as trees. The proposed recursive partitioning scheme automatically identifies the subgroups of subjects exhibiting DIF.

The method proposed in this chapter also uses recursive partitioning techniques, but in a different form than Strobl et al. (2015). Strobl et al. (2015) recursively partition the covariate space to identify regions of the covariate space in which DIF occurs. In the investigated regions a parametric latent trait model that includes covariates is fitted. Regions are suspected to be relevant if the parameter estimates in the regions differ strongly. Therefore, regions in the covariate space are identified that show different difficulties. A disadvantage of the method is that it detects regions of the covariate space that are linked to DIF but does not automatically detect the items that are responsible. In contrast, the recursive partitioning method proposed here focusses on the detection of the items that are responsible for DIF. Recursive partitioning is used on the item level not on the global level, which treats all items simultaneously, as in the method proposed by Strobl et al. (2015). The item focussed approach allows to detect the items that carry DIF but keeps the advantage that no pre-specified subgroups are needed.

In Section 4.2 we introduce the new method and present an illustrative example, in Section 4.3 we give a detailed description of the fitting procedure. Results of wider simulation studies with comparisons to competing methods are given in Section 4.4. In Section 4.5 we consider another application. Finally, in Section 4.6 we consider an extension to ordinal item responses.

4.2. Item Focussed Recursive Partitioning

We will consider differential item functioning for the Rasch model. Therefore we start with the introduction of some notation.

4.2.1. Differential Item Functioning for the Rasch Model

In the binary Rasch model the probability for a person to score on an item is determined by a parameter for the latent ability of the person and a parameter for the item difficulty. In the case of P persons and I items, the Rasch model is given by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}, \quad p = 1, \ldots, P, \quad i = 1, \ldots, I, \tag{4.1}$$

where Y_{pi} represents the response of person p on item i. It is coded by $Y_{pi} = 1$ if person p solves item i and $Y_{pi} = 0$ otherwise. Both, the person parameters, $\theta_p,\ p = 1, \ldots, P$, and the item parameters, $\beta_i,\ i = 1, \ldots, I$, are unknown and have to be estimated.

An alternative form of the model is

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \eta_{pi} = \theta_p - \beta_i, \tag{4.2}$$

where the predictor $\eta_{pi} = \theta_p - \beta_i$ represents the difference between the ability of the person and the difficulty of the item. As model (4.2) is not identifiable in this general form, a restriction on the parameters is needed. A common choice that is also used in the following is $\theta_P = 0$.

In Rasch models, DIF appears if an item has different difficulties depending on characteristics of the person that tries to solve the item. The simplest form of DIF is found if item difficulties differ in a focal and a reference group. If item i is a DIF item the predictor of the model is given by

$$\eta_{pi} = \theta_p - \gamma_i^{(j)}, \quad j = 1, 2, \tag{4.3}$$

where $j = 1$ denotes the focal group and $j = 2$ the reference group. DIF occurs, if $\gamma_i^{(1)} \neq \gamma_i^{(2)}$, which can be tested, for example, by likelihood ratio tests. The recursive partitioning scheme considered in the following uses this simple model, which considers two subgroups, as building block. By iterative application of the splitting into two subgroups one obtains a tree for each item. It should be mentioned that we consider uniform DIF in Rasch models. For more general models as 2PL or 3PL models DIF can be generated in different ways, for example, by difference in item discrimination.

4.2.2. Recursive Partitioning

Recursive partitioning also known as tree-based modeling has its roots in automatic interaction detection (AID), proposed by Morgan and Sonquist (1963). The most popular modern version is due to Breiman et al. (1984) and is known by the name *classification and* *regression trees*, often abbreviated as CART. Alternative approaches are the C4.5 algorithm (Quinlan, 1986, 1993), or the recursive partioning framework based on conditional inference proposed by Hothorn et al. (2006). The method is conceptually very simple. By binary recursive partitioning the feature space is partitioned into a set of rectangles, and on each rectangle a simple model (for example, a constant) is fitted. An overview with a focus on psychometrics was given by Strobl et al. (2009).

Regression trees may be seen as a hierarchical way to describe a partition of the predictor space. The tree represents the partition in a unique way. Each node of the tree corresponds to a subset of the predictor space. The *root* is the top node consisting of the whole predictor space, and the *terminal nodes* or *leaves* of the tree correspond to the subregions.

To grow a tree one typically uses the "standard splits", which means that each partition of node A into subsets A_1, A_2 is determined by only one variable. The splits to be considered depend on the scale of the variable:

> For *metrically scaled* and *ordinal* variables, the partition into two subsets has the form
> $$A \cap \{x_j \leq c\}, \quad A \cap \{x_j > c\},$$
> based on the threshold c on variable x_j.
>
> For *categorical* variables without ordering $x_j \in \{1, \ldots, K_j\}$, the partition has the form
> $$A \cap S, \quad A \cap \bar{S},$$
> where S is a non-empty subset $S \subset \{1, \ldots, K_j\}$ and $\bar{S} = \{1, \ldots, K_j\} \setminus S$ is the complement.

In the following we will mostly use the split for metrically scaled or ordinal variables to illustrate how trees are obtained. Let $\boldsymbol{x}_p^T = (x_{p1}, \ldots, x_{pm})$ denote a person-specific covariate vector of length m. For the detection of DIF the first split means one examines for all the items, all the variables and possible splits of the corresponding variable the Rasch model with predictor

$$\eta_{pi} = \theta_p - [\gamma_{il}^{[1]} I(x_{pj} \leq c_j) + \gamma_{ir}^{[1]} I(x_{pj} > c_j)],$$

where $I(\cdot)$ denotes the indicator function with $I(a) = 1$ if a is true and $I(a) = 0$ otherwise. The model is just an alternative representation of (4.3), with the focal and reference group constructed by a split of the j-th variable at split-point c_j. The parameter $\gamma_{il}^{[1]}$ denotes the item difficulty in the left node ($x_{pj} \leq c_j$) and $\gamma_{ir}^{[1]}$ the item difficulty in the right node ($x_{pj} > c_j$). One chooses that combination of item, variable and split that has the smallest p-value when tested for DIF, that is, in the examination of the null hypothesis $H_0 : \gamma_{il}^{[1]} - \gamma_{ir}^{[1]} = 0$. This selection yields the first split into left and right daughter nodes corresponding to the regions $I(x_{pj} \leq c_j)$ and $I(x_{pj} > c_j)$.

Further splitting means that one of the nodes, say $I(x_{pj} > c_j)$, is further split, for example, in variable s at cut point c_s, yielding the daughters

$$I(x_{pj} > c_j)I(x_{ps} \leq c_s) \quad \text{and} \quad I(x_{pj} > c_j)I(x_{ps} > c_s),$$

and the linear predictor

$$\eta_{pi} = \theta_p - [\gamma_{il}^{[1]} I(x_{pj} \leq c_j) + \gamma_{il}^{[2]} I(x_{pj} > c_j)I(x_{ps} \leq c_s) + \gamma_{ir}^{[2]} I(x_{pj} > c_j)I(x_{ps} > c_s)],$$

where $\gamma_{il}^{[2]}, \gamma_{ir}^{[2]}$ are the weights on the new split. Then the item difficulty in the region $\{x_{pj} \leq c_j\}$ is $\gamma_{il}^{[1]}$ but for the region $\{x_{pj} > c_j\}$ one has to distinguish between $\{x_{pj} > c_j, x_{ps} \leq c_s\}$ with item difficulty $\gamma_{il}^{[2]}$ and $\{x_{pj} > c_j, x_{ps} > c_s\}$ with item difficulty $\gamma_{ir}^{[2]}$.

The corresponding trees are trees for specific items, namely the items that were selected to carry DIF. If an item is never selected it is considered as compatible with the Rasch model.

In the following we use the model abbreviation IFT for item focussed trees.

4.2.3. An Illustrative Example

Before giving the details how to grow trees we want to illustrate the procedure by use of a data set that has been used previously in the DIF literature (Strobl et al., 2015). We consider the data of an online quiz testing one's general knowledge. The test was conducted by the German news magazin Spiegel in 2009. The whole test consisted of 45 questions from five different topics, that are politics, history, economy, culture and natural sciences. A detailed analysis and discussion of the original data set is found in Trepte and Verbeet (2010).

We use a subset of the data including 1075 university students from Bavaria. To test for DIF we incorporate the five covariates gender (0: female, 1: male), age, number of matriculated semester, elite status of the university (0: no, 1: yes) and the frequency of accessing Spiegel's online magazine (spon) from 1 (never) to 7 (daily). The distributions of the five covariates and the test results are displayed in Figure 4.1.

Item Focussed Recursive Partitioning

When using item focussed recursive partitioning 21 of the 45 items show DIF. The result is not surprising because the questions of the quiz were not chosen very carefully to avoid DIF. Altogether the algorithm performs 33 splits until further splits are not significant at significance level $\alpha = 0.05$ (for details of the test see Section 4.3). The first ten splits all

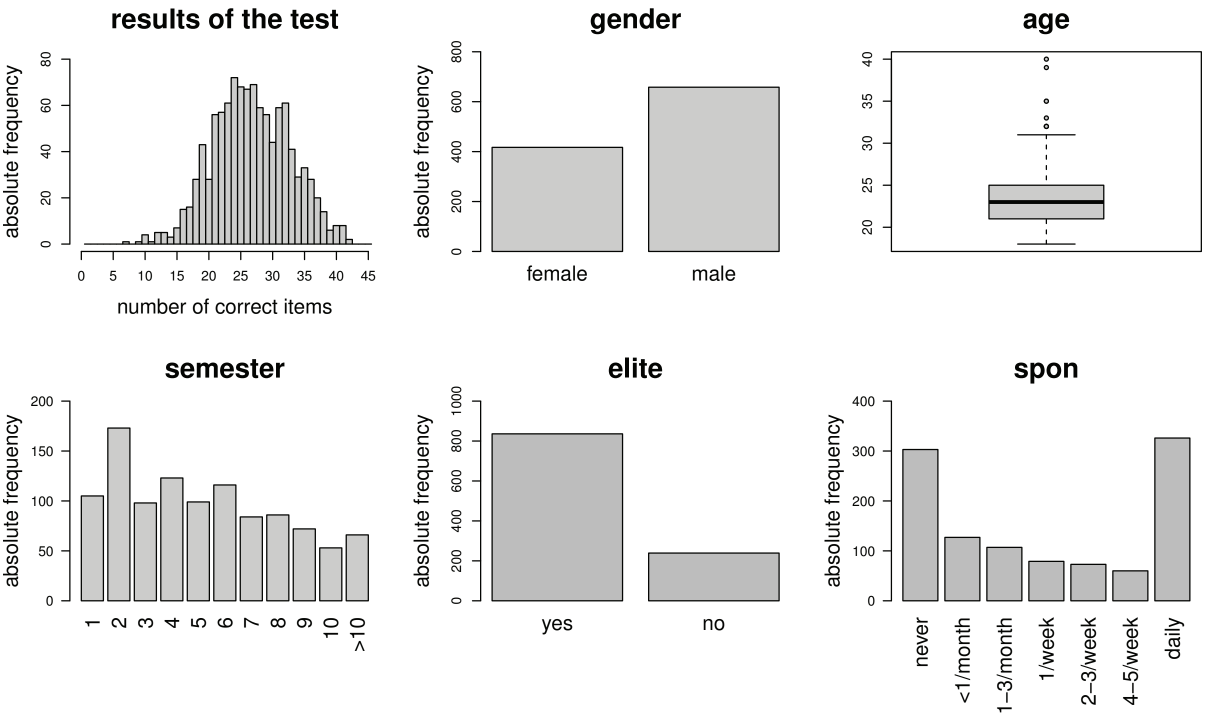


Figure 4.1.: Graphical representation of the results of the general knowledge test (upper left) and the distribution of the five covariates in the analyzed data.

refer to the covariate gender, so the strongest effects were found for the difference between males and females. No significant splits were found for the variable elite. The difficulties of the items seem not to depend on the elite status of the university. The three items with the strongest effects, which were found in the first iterations of the algorithm, were the following:

19: Who is this? - Picture of Dieter Zetsche, CEO of Mercedes-Benz

43: Which kind of bird is this? - Blackbird

40: What is also termed Trisomy 21? - Down syndrome

The resulting trees for these items 19, 43 and 40 are shown in Figure 4.2. For each item one can see how the difficulty of the item depends on the characteristic of certain variables. The estimated item difficulties are given in each leaf of the trees, which represent the identified subgroups. For example, in item 19 (recognition of Dieter Zetsche as CEO of Mercedes Benz) the difficulty for females (gender=0) is 2.665 while for males (gender=1) it is distinguished between students who frequently read Spiegel online (spon=7) with an item difficulty of 0.126 and a much larger item difficulty of 1.155 for students who read it less regularly (spon ≤ 6). The other two items show DIF only for gender. Both items

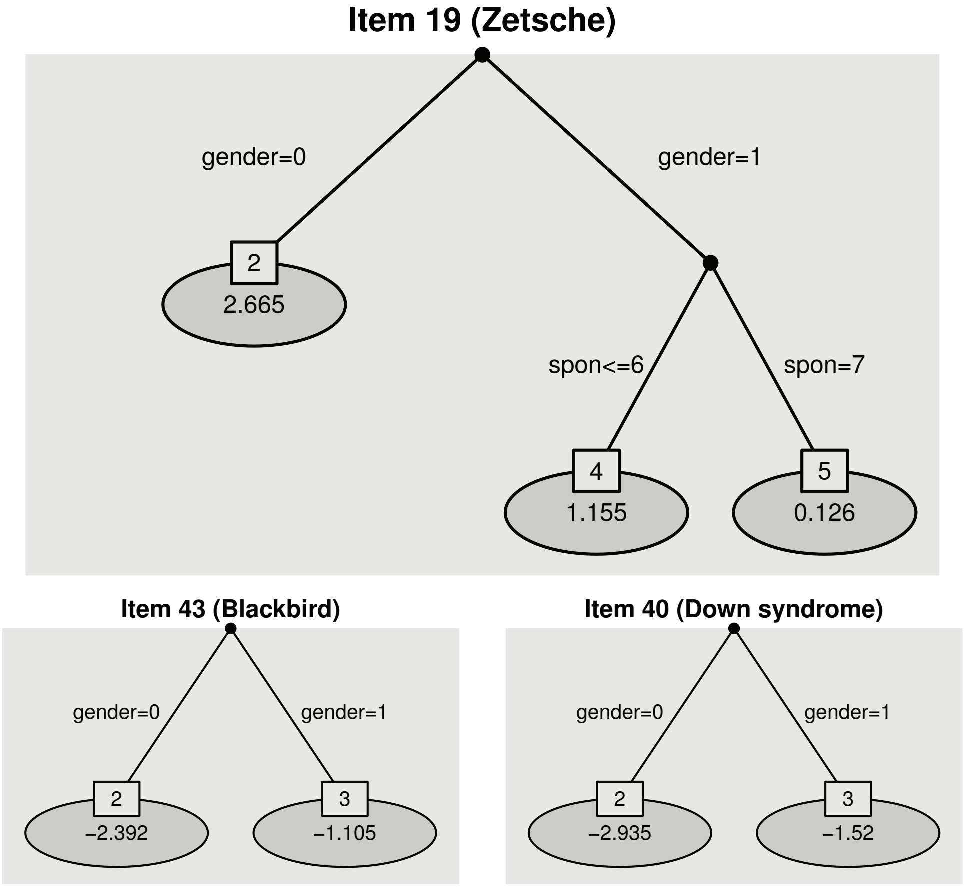


Figure 4.2.: Trees for Items 19, 43 and 40 of the general knowledge test. The item difficulties are given for each subgroup represented by the leaves of the trees.

concerning the recognition of birds and knowledge of genetic diseases are easier to solve for females. It is also seen that item 19 is much harder to solve than the other two items.

Another quite interesting tree structure is received for item 6 of the test (see Figure 4.3). The corresponding question asks to identify the Prime Minister of Bavaria, Horst Seehofer. For all students who read the online magazine very regular (spon>5) the question is very easy. By contrast the question is more difficult for students who do not read Spiegel online very often (spon $\leq$ 5), in particular if they are female (gender=0) and comparably young (age $\leq$ 21).

The strength of the approach is that one sees for each item which variables generate DIF. The tree structure also yields an ordering of the relevance of the variables with the first split being the most relevant. By recursive splitting of regions trees are always devices to detect interactions. For example, in item 19 a relevant interaction effect is that of gender and frequency of reading Spiegel online. Moreover, trees automatically detect the groups

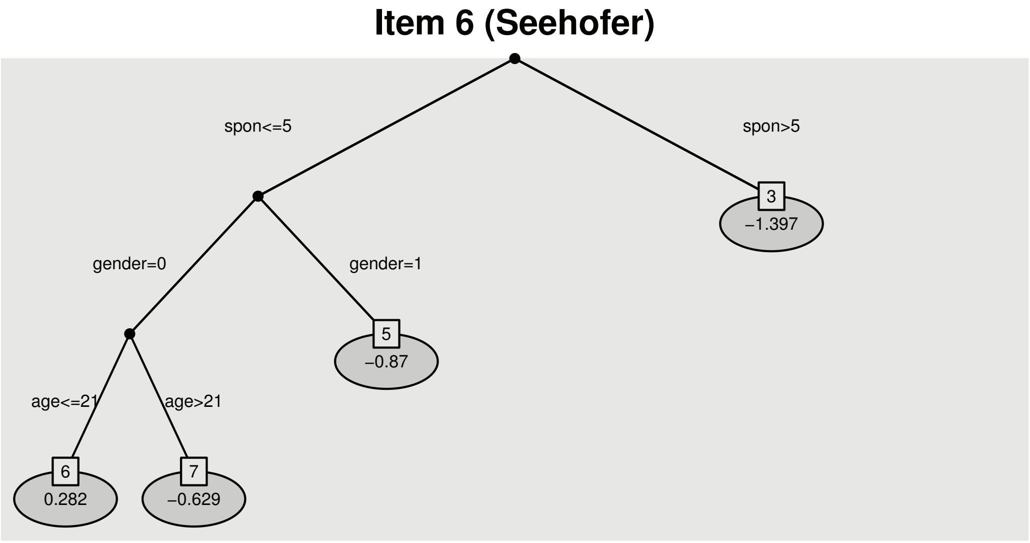


Figure 4.3.: Tree for Item 6 of the general knowledge test. The item difficulties are given in each leaf of the tree.

that have to be distinguished. It is not necessary to define the focus and the reference group beforehand.

Rasch Trees

To illustrate the difference between the alternative approach to use trees we analyse in the following the same data set by using the Rasch tree concept of Strobl et al. (2015). The corresponding tree is given in Figure 4.4. The significance level used for the tests for parameter instability was the same as for our item focussed trees, $\alpha = 0.05$.

The basic concept of conventional Rasch trees is to search for the split in the explanatory variables that shows the strongest differences in *all of the item difficulties*. In this application one obtains a tree with splits in two variables, gender and spon. These variables are found to induce DIF and one finds four groups that differ in terms of item difficulty. In each leaf of the corresponding tree the estimated difficulties are shown. The crucial point is that the resulting tree is one tree for all of the items. It does not identify the items that are responsible for the split and therefore for DIF. Consequently, from Figure 4.4 it is hard to identify those items that are affected by DIF and those that are not. Moreover, there is no criterion provided to identify the responsible items. In contrast, the item focussed approach shows which items are responsible. It is seen from Figure 4.2 that both variables, gender and spon, are also found for items 19, 34 and 40 but in a more differentiated way. In addition, Figure 4.3 shows that also item 6 is a DIF item that is also specific for age.

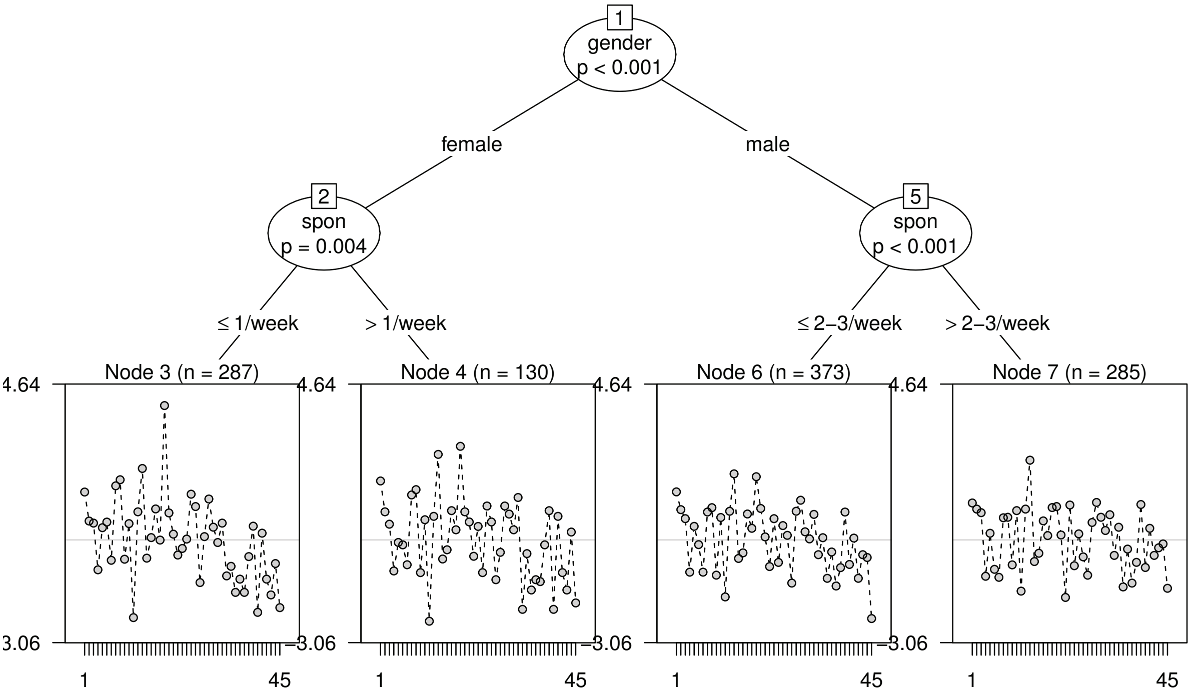


Figure 4.4.: Result of the analysis of the general knowledge test by a Rasch tree.

The example illustrates one specific difference between the two approaches namely the obtained results. The conventional tree in Figure 4.4 shows that gender and spon induce DIF but it is not yet clear which items are concerned. The item focussed tree approach identifies the items and yields a specific tree for each item. For item 19 the splits use also gender and spon (see Figure 4.2), but for females it seems not necessary to split further. For items 43 and 40 (see Figure 4.2) only gender seems relevant. For item 6 also age is found to induce DIF and the strongest variable, which is split first, is spon, not gender. Therefore, instead of assuming splits to be the same for the whole set of items one obtains specific splits for each item. The resulting small trees show how covariates determine DIF items and the visualization as trees makes it easily accessible.

Logistic Regression

For comparison we also consider the logistic regression method that was proposed by Swaminathan and Rogers (1990) and, more recently, extended by Magis et al. (2015). The basic concept is to fit a logistic model for answering an item correctly given the test score and the group membership. The model has the form

$$\log\left(\frac{P(Y_{pi}=1|S_p,g)}{P(Y_{pi}=0|S_p,g)}\right)=\beta_{0i}+S_p\beta_i+\gamma_{ig},$$

where g denotes the group, S_p is the test score of person p and γ_{ig} are the group-specific parameters. Of course, if one considers G groups, one of the parameters $\gamma_{i1},\ldots,\gamma_{iG}$ has to be set to zero. For example, by setting $\gamma_{i1}=0$ the first group is implicitly chosen as the reference group. If one has only two groups there is only one parameter for each item. Following Magis et al. (2015), the parameters β_{0i} can be seen as the counterparts of the item difficulties and the parameters β_i as the counterparts of item discrimination parameters. The parameters of interest, however, are the parameters $\gamma_{i1},\ldots,\gamma_{iG}$. If one of them is unequal zero the item is supposed to show DIF. Therefore, DIF can be diagnosed by testing, for example, by using a likelihood ratio test, whether the null hypothesis H_0: $\gamma_{i1}=\cdots=\gamma_{iG}=0$ holds.

The basic concept can also be used for continuous or a mix of categorical and continuous variables. Then one considers the logistic model

$$\log\left(\frac{P(Y_{pi}=1|S_p,\boldsymbol{x}_p)}{P(Y_{pi}=0|S_p,\boldsymbol{x}_p)}\right)=\beta_{0i}+S_p\beta_i+\boldsymbol{x}_p^\top\boldsymbol{\gamma}_i,$$

where $\boldsymbol{x}_p$ is a vector of explanatory variables that might induce DIF. It should be noted that group membership is just a special case; with reference group 1 one uses the vector of explanatory variables $\boldsymbol{x}_p^T=(x_{p2},\ldots,x_{pG})$, where $x_{pg}=1$ if person p is from group g and 0 otherwise. The corresponding vector of parameters is $\boldsymbol{\gamma}_i^T=(\gamma_{i2},\ldots,\gamma_{iG})$. In the general case, DIF diagnosis uses a test for the pair of hypotheses

$$H_0:\boldsymbol{\gamma}_i=\mathbf{0}\qquad H_1:\boldsymbol{\gamma}_i\neq\mathbf{0},$$

where $\mathbf{0}$ is the vector in which all components are zero. Hypotheses are tested separately for each item with significance level α.

The logistic model approach to DIF detection is not without problems. The test scores are used as a proxy for the ability of a person. However, test scores as the number of solved items are sufficient statistics for ability parameters only if the Rasch model holds, that is, if no DIF is present (see also Magis et al., 2015). Nevertheless, it provides a general method to investigate DIF. Therefore, we will use the method in simulations and in the present illustration.

Table 4.1 compares the logistic model (Logistic) and item focussed trees (IFT). It shows only items that were found to be DIF items by one of the methods. The order of the items

Table 4.1.: Comparison of detected DIF items of the general knowledge test using logistic regression and item focussed recursive partitioning.

item	lr statistic	p-value	Logistic	IFT
19	114.5921	0.0000	×	×
28	82.4253	0.0000	×	×
26	81.3029	0.0000	×	×
34	74.2029	0.0000	×	×
40	72.8286	0.0000	×	×
25	61.0688	0.0000	×	×
43	55.1655	0.0000	×	×
36	54.1240	0.0000	×	×
24	49.4813	0.0002	×	×
33	49.2615	0.0002	×	×
45	48.0907	0.0002	×	×
13	46.8115	0.0004	×	
8	43.8113	0.0010	×	×
12	43.6708	0.0010	×	×
5	40.0507	0.0032	×	×
27	40.0141	0.0033	×	
35	39.1269	0.0043	×	
41	35.5084	0.0121	×	
9	34.6475	0.0154	×	×
1	33.6944	0.0200	×	
37	30.6730	0.0438	×	
22	30.1479	0.0499	×	×
44	28.9102	0.0674		×
6	28.2181	0.0793		×
42	24.4983	0.1777		×
23	23.9883	0.1966		×
39	13.4830	0.8130		×

reflects the p-values of the likelihood ratio test when investigating DIF by use of the logistic model approach. There is a strong overlap; 16 items were found to be DIF items in both methods, 6 items showed DIF when using the logistic approach but not when using item focussed trees, 5 items showed DIF when using item focussed trees but not when using the logistic approach.

4.3. Fitting Trees

In this section we give the details of the algorithm that yields item focussed trees. In particular we show how trees are grown and when to stop.

4.3.1. The Basic Algorithm

In all tree-based methods one has to decide in particular how to split and how to determine the size of the trees. Split criteria that are in common use are splitting by impurity measures like the Gini-based impurity or the entropy and test-based splits. The latter use a test statistic to evaluate which split is the strongest to explain the impact of predictors. Already Breiman et al. (1984) considered very general families of impurity measures including the entropy, which is strongly related to test-based split when the deviance is used as test statistic, see, for example, Ciampi et al. (1987) and Clark and Pregibon (1992). As far as tree size is concerned, in early recursive partitioning approaches the final tree is typically obtained by growing large trees and then prune them to an adequate size, for details see Breiman et al. (1984) or Ripley (1996), Chapter 7. Alternative methods are based on maximally selected statistics. The basic idea is to consider the distribution of the selection process. When a split-point is selected based on a test statistic T_i for possible split-point i, one investigates the distribution of $T_{max} = max_{i=1,\ldots,m} T_i$. The p-value of the distribution of T_{max} provides a measure for the relevance of a predictor that does not depend on the number of split-points since the number has been taken into account, see Hothorn and Lausen (2003), Shih (2004), Shih and Tsai (2004), Strobl et al. (2007). A unified framework for recursive partitioning that embeds tree-structured regression models into a well-defined theory of conditional inference procedures was proposed by Hothorn et al. (2006). The splitting is stopped when the global null hypothesis of independence between the response and any of the predictors cannot be rejected at a pre-specified nominal significance level α. The method explicitly accounts for the involved multiple test problem. By separating variable selection and the splitting procedure one arrives at an unbiased recursive partitioning scheme that also avoids the selection bias toward predictors with many possible splits or missing values. We will draw on the concept of conditional inference procedures in our approach to select splits.

Let us consider again the construction of the first split. One examines for all the items, all the variables and possible splits of the corresponding variable the Rasch model with predictor

$$\eta_{pi} = \theta_p - [\gamma_{il} I(x_{pj} \leq c_j) + \gamma_{ir} I(x_{pj} > c_j)].$$

The test for DIF at split-point c_j corresponds to the null hypothesis $H_0 : \gamma_{il} - \gamma_{ir} = 0$. If H_0 holds for all split-points the item shows no DIF since $\gamma_{il} = \gamma_{ir}$ holds for all split-points. Let T_{jc_j} denote the corresponding test statistic, for example, the log-likelihood test statistic. To obtain a test for variable j one has to consider simultaneously all the test statistics T_{jc_j} with c_j from the set of possible splits. We will use the maximal value statistic $T_j = \max_{c_j} T_{jc_j}$, which is composed from the strongly correlated test statistics. To obtain a decision on the null hypothesis controlling for a given significance level a permutation test is used. That means the distribution of T_j is determined by using random permutations of variable j

that break the relation of the covariate and the response. More concrete, one permutes the values of variable j in the data matrix and computes the corresponding value of the test statistic. By computing the values of the test statistic for a large number of permutations one obtains an approximation of the distribution under the null hypothesis that variable j has no effect and an corresponding p-value. In our applications and simulations we used 1000 permutations.

Given overall significance level α the significance level for the permutation test that tests splits in one variable is chosen by α/m, where m denotes the number of covariates that are available. For the item and variable with the largest value of T_j the permutation test is carried out. If no significant effect is found no splitting is performed. Otherwise for this combination the split-point is chosen for which T_{jc_j} had the smallest p-value. Since variable selection is separated from the splitting decision one could also use alternative criteria for the selection of splits. If variable, item and split-point are selected the model is fitted for this selection yielding estimates $\hat{\theta}_p, \hat{\gamma}_{il}, \hat{\gamma}_{ir}$.

For illustration we use again the example from Section 4.2.3. The largest test statistic over all items and variables occurred for item 19 and gender. The corresponding value of the test was $T_{gender} = 85.5$. For comparison, the values of the other variables for item 19 were $T_{age} = 9.2$, $T_{semester} = 3.8$, $T_{elite} = 0.9$ and $T_{spon} = 48.86$. The permutation test for the combination item 19 and gender was highly significant with an p-value close to zero and distinctly smaller than 0.05/5=0.01. Therefore, one has a significant split and the first split is for gender in item 19. Since for gender there is only one possible split, one has not to investigate which split is the best.

In later steps of the growing of a tree the basic procedure is the same, one searches for the statistic with the maximal value trying all combinations of items and variables. For the items that have not yet been split the search is the same as before, but for items that already have been split one starts from already selected splits. Let the already built node for item i be characterized by $S_i = \{(c_{ij_1}, a_{i1}), \ldots, (c_{ij_B}, a_{iB})\}$, where c_{ij_b} is the threshold in variable j_b and $a_{ib} \in \{0, 1\}$ encodes if one is below or above the threshold. The corresponding node is

$$\text{node}_i(\boldsymbol{x}_p) = \prod_{b=1}^{B} I(x_{pj_b} > c_{ij_b})^{a_{ib}} (1 - I(x_{pj_b} > c_{ij_b}))^{1-a_{ib}},$$

where B denotes the total number of branches. When considering splits of this node one examines for all variables j and all possible splits the Rasch model with item difficulties

$$\gamma_{il}\text{node}_i(\boldsymbol{x}_p) I(x_{pj} \leq c_j) + \gamma_{ir}\text{node}_i(\boldsymbol{x}_p) I(x_{pj} > c_j),$$

where c_j is a split-point for variable j. The corresponding null hypothesis is $H_0 : \gamma_{il} - \gamma_{ir} = 0$, which is tested by test statistic T_{jc_j}. Again one first investigates if variable j has an effect by using a permutation test for $T_j = \max_{c_j} T_{jc_j}$ with significance level α/m, for the node and variable with the largest value of T_j. If a significant effect is found one determines the best split and fits the corresponding model for this split-point. It should be noted that in the fitting step all other parameters of the model, including the person parameters θ_p, are refitted.

In the illustrative example several other items were split in the next steps, in the eleventh step again item 19 was selected in the second node ($gender = 1$), which was already built in the first split. The maximal test statistic was $T_{spon} = 29.5$, the others for this node were $T_{age} = 4.4$, $T_{semester} = 2.3$ and $T_{elite} = 0.9$. Covariate gender can not be considered anymore and therefore the local significance level in the already built node has to be adapted to 0.05/4. The corresponding p-value was 0.001. The selected split, which had the smallest p-value for the likelihood ratio statistic for spon in item 19 (given $gender = 1$), was obtained for the sixth split (spon ≤ 6; spon $= 7$).

The procedure stops if no test for the combination of item and variable (given the root or and already identified node) is significant any more. In the illustrative example the algorithm terminates after 33 splits in 21 items. The largest maximal value statistic in the 34-th step was 10.62, but not significant on level 0.01. Item 19 was selected for the last time in the eleventh step.

If no combination of item and variable is significant any more the tree for an item i that has been split is defined by terminal nodes $S_{i1}, \ldots, S_{iL_i}$ and the predictor of the model can be represented by

$$\eta_{pi} = \theta_p - \mathrm{tr}_i(\boldsymbol{x}_p) = \theta_p - \sum_{\ell=1}^{L_i} \gamma_{i\ell} \, \mathrm{node}_{i\ell}(\boldsymbol{x}_p), \tag{4.4}$$

where $\gamma_{i1}, \ldots, \gamma_{iL_i}$ denote the item difficulties in the terminal nodes. The algorithm terminates if no significant permutation test is obtained anymore. For those items where no splitting is performed the constant $\mathrm{tr}_i(\boldsymbol{x}_p) = \beta_i$, corresponding to the item parameter of the simple Rasch model, is fitted.

In the illustrative example the resulting tree for item 19, given in Figure 4.2, is composed of three terminal nodes including the two covariates gender and spon.

4.3.2. Comments on Type I Error Rates

It seems warranted to briefly discuss the concept of the type I error rate that is behind the considered procedures. It should especially clarify when to adapt the given significance level and when not. In DIF detection type I error typically is seen as equivalent to false alarm rates, see, for example, Magis and De Boeck (2014). If one wants to control for this form of type I error it suffices to use tests with an significance level α for each item. Then, for each non-DIF item the probability of being falsely classified as DIF item is controlled by α. If one has N non-DIF items one can expect $N\alpha$ items to be falsely classified as DIF items yielding a false alarm rate $N\alpha/N = \alpha$. This procedure is used by most of the test based approaches including the item focussed trees proposed here. It is in line with the concept of controlling the false discovery rate proposed by Benjamini and Hochberg (1995).

A quite different concept of type I error is the familywise error rate, which stimulated research in multiple testing. The familywise error rate is the probability of falsely rejecting at least one among all the considered hypotheses when performing multiple hypotheses tests. In DIF detection it corresponds to the probability that at least one item is falsely classified as DIF item. The concept is much stronger. If one wants to control the familywise error rate by a global significance level α one has to use much smaller significance levels in the single tests. One has to adapt the significance level, for example, by using the Bonferroni or the Holm procedure (Holm, 1979).

These procedures are used in the proposed item focussed trees when several variables are available. In order to obtain a significance level α for each item, for fixed item the significance level of the tests for each variable is chosen as α/m, where m denotes the number of covariates that might induce DIF. Thus, given that an item has no DIF, in the first step the significance level for testing one variable is α/m and the probability that the item shows DIF in the first step in any of the variables (and therefore DIF is diagnosed at all) is restricted by α. So the probability of a false DIF result is restricted by α because further tests are performed only if a significant result was found in the first step. The consequence is that on the item level the familywise error rate is under control, with the family of the null hypotheses being composed of all the null hypotheses that there is no DIF in single variables (for fixed item).

4.4. Simulations

In this section we investigate the performance of the fitting procedure in terms of the ability to detect items that show DIF and to estimate the item difficulty parameters in each node. We consider several simulation scenarios where data $Y_{pi},\ p = 1, \dots, P,\ i = 1, \dots, I$ were generated according to the binary Rasch model with DIF in some of the items. All the presented results are based on 100 replications.

The following components of the model are the same in each simulation scenario:

- $P = 500$ (number of persons); $I = 20$ (number of items)
- $\theta_p \sim N(0, 1)$ (person abilities)

- $\beta_i \sim N(0,1)$ (item difficulties for items without DIF)

If item i is assumed to show DIF the corresponding normal distributed item difficulty is transformed by step functions. The resulting item difficulties refer to groups of persons represented by the nodes $S_{i1}, \dots, S_{iL_i}$.

Strength of DIF

In each simulation scenario we generate data with three different strengths of DIF, strong, medium and weak. The strength of DIF in one item i can be measured by the variance of the item parameters $V_i = \text{var}\left(\sum_\ell \gamma_{i\ell} \, \text{node}_{i\ell}\right)$, which for fixed nodes is determined by the parameters $\gamma_{i\ell}$. The average of V_i over the items with DIF is used as a measure of the overall strength of DIF in these items. In all of the simulation scenarios parameters are specified in such a way that for strong DIF the DIF strength is 0.41, for medium DIF the strength is 0.23 and for weak DIF it is 0.10.

Mean Squared Errors

We compare the estimated coefficients to the true parameters by calculating mean squared errors (MSEs). For the person abilities it is $\frac{1}{P}\sum_{p=1}^{P}(\hat{\theta}_p - \theta_p)^2$ and for the item difficulties it is $\frac{1}{P \cdot I}\sum_{p=1}^{P}\sum_{i=1}^{I}(\hat{tr}_i(\boldsymbol{x}_p) - tr_i(\boldsymbol{x}_p))^2$, respectively, averaged over all simulations.

Hit Rates

Let each item be characterized by a vector $\boldsymbol{\delta}_i^T = (\delta_{i1}, \dots, \delta_{im})$, with $\delta_{ij} = 1$ if item i has DIF in component j and $\delta_{ij} = 0$ otherwise. An item is a non-DIF item if $\boldsymbol{\delta}_i^T = (0, \dots, 0)$, if one of the components is 1 it is a DIF item. With indicator function $I(\cdot)$, criteria to judge the identification of items with DIF are:

- True positive rate on the item level (items correctly identified as DIF items):

  $TPR_I = \frac{1}{\#\{i:\boldsymbol{\delta}_i \neq \boldsymbol{0}\}}\sum_{i:\boldsymbol{\delta}_i \neq \boldsymbol{0}} I(\hat{\boldsymbol{\delta}}_i \neq \boldsymbol{0})$

- False positive rate on the item level (non-DIF items incorrectly identified as DIF items):

  $FPR_I = \frac{1}{\#\{i:\boldsymbol{\delta}_i = \boldsymbol{0}\}}\sum_{i:\boldsymbol{\delta}_i = \boldsymbol{0}} I(\hat{\boldsymbol{\delta}}_i \neq \boldsymbol{0})$

- True positive rate for the combination of item and variable:

  $TPR_{IV} = \frac{1}{\#\{i,j:\delta_{ij} \neq 0\}}\sum_{i,j:\delta_{ij} \neq 0} I(\hat{\delta}_{ij} \neq 0)$

- False positive rate for the combination of item and variable:

  $FPR_{IV} = \frac{1}{\#\{i,j:\delta_{ij}=0\}} \sum_{i,j:\delta_{ij}=0} I(\hat{\delta}_{ij} \neq 0)$

4.4.1. One Single Predictor

In the first simulation scenarios we consider only one predictor x that induces DIF in several items. In this case also traditional methods to detect DIF can be used.

Comparison with Alternative Methods

We will start with a comparison of the proposed method with other established methods for the detection of DIF. Most methods are restricted to the comparison of two or more groups. We consider the Mantel-Haenszel method (MH), the method of logistic regression (Logistic) and Lord's χ^2-test (Lord). An overview of these methods is given in Magis et al. (2010) and Magis et al. (2011). For the comparison we use the implementation in the `R` add-on package `difR` (Magis et al., 2013).

For the comparison of two groups we simulate four items with DIF induced by one binary predictor $x \in \{0, 1\}$. For the comparison of multiple groups we simulate DIF with respect to an ordered factor $x \in \{1, \ldots, 5\}$. The definition of differences of item difficulties in these groups are given in Table 4.2 for both scenarios. The overall strength of DIF in the four items can be determined by the value of c. Choosing c=1 in the strong setting, c=0.75 in the medium case and c=0.5 in the weak case leads to the DIF strengths as given above. In addition, we consider the case without DIF. It corresponds to the value c=0.

The selection performance for both scenarios is given in Table 4.3 for all of the methods. In the case of item focussed trees (IFT) each permutation test is based on 1000 permutations. Table 4.3 shows true positive and false positive rates on the item level as the average over 100 simulations, respectively. In the case without DIF only false positive rates are available. It is seen that the proposed method competes well with the established methods. In the case of two groups the true positive and false positive rates are very similar for all methods with the exception of Lord's method. The latter shows distinctly smaller false positive rates with a tendency to slightly smaller true positive rates than the other methods. In the case of five groups the pattern is similar. It can be seen that for weak DIF the true positive rates are poor for all of the methods, in particular Lord's method performs very poorly. However, in this case item focussed trees still shows the best result yielding the true positive rate 0.61. In the case of no DIF and two groups trees show the same false positive rates as MH and logistic. For five groups the false positive rates are slightly smaller than in these methods. As in the other settings Lord's method yields different values.

Table 4.2.: True simulated differences of item difficulties for the comparison of two or five groups.

	Difference of Difficulty			
Scenario	Item 1	Item 2	Item 3	Item 4
two groups	$1c \cdot I(x = 1)$	$-1c \cdot I(x = 1)$	$1.5c \cdot I(x = 0)$	$-1.5c \cdot I(x = 0)$
five groups	$1c \cdot I(x > 2)$	$-1c \cdot I(x > 3)$	$1.5c \cdot I(x > 4)$	$-1.5c \cdot I(x > 1)$

Table 4.3.: True positive and false positive rates on the item level for the comparison of two or five groups as average over all 100 replications.

		Two groups		Five groups	
Method		TPR_I	FPR_I	TPR_I	FPR_I
MH	strong	0.9975	0.0463	0.9950	0.0581
	medium	0.9800	0.0444	0.9125	0.0588
	weak	0.8400	0.0450	0.5300	0.0575
	no DIF	—	0.0470	—	0.0535
Logistic	strong	0.9975	0.0513	0.9975	0.0656
	medium	0.9750	0.0506	0.9225	0.0594
	weak	0.8375	0.0488	0.5700	0.0600
	no DIF	—	0.0475	—	0.0585
Lord	strong	0.9975	0.0325	0.9850	0.0286
	medium	0.9650	0.0325	0.8225	0.0268
	weak	0.7900	0.0319	0.3925	0.0300
	no DIF	—	0.0305	—	0.0245
IFT	strong	0.9950	0.0444	0.9900	0.0500
	medium	0.9625	0.0438	0.9250	0.0581
	weak	0.8100	0.0481	0.6100	0.0538
	no DIF	—	0.0475	—	0.0485

Continuous Predictor

The previous simulations showed that item focussed trees work quite well in pure detection of DIF items when compared to established methods. One of the advantages of item focussed trees is that the method is not limited to the case of a simple comparison of multiple groups but can also handle a much more complex structure of predictors.

In the following we consider one standard normal distributed predictor x and two items with DIF. We assume a sigmoidal relation between the value of x and the item difficulty of item 1 and 2. The linear predictors are given by

$$\eta_{p1} = \theta_p - \beta_1 + c \cdot \arctan(x_p) \quad \text{and} \quad \eta_{p2} = \theta_p - \beta_2 - c \cdot \arctan(x_p).$$

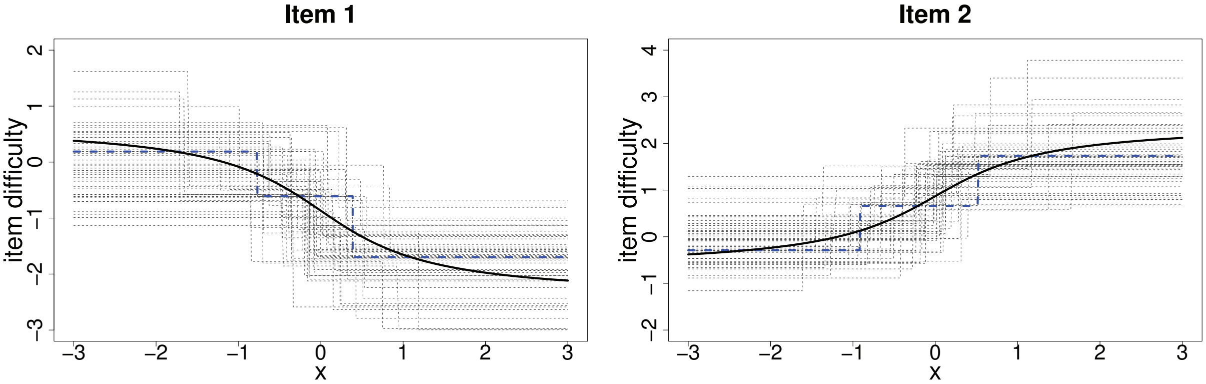


Figure 4.5.: True item difficulties for item 1 and 2 (bold lines) and estimated item difficulties for 50 replications (dashed lines) of the simulation scenario with one standard normal distributed predictor and strong DIF.

For item 1 item difficulties are monotonically decreasing, thus for persons with small x item 1 is harder to solve than for persons with a higher value of x. For item 2 item difficulties are monotonically increasing, thus for persons with a small value of x it is easier to solve than for persons with a higher value. The data generating process in this scenario is not determined by step functions but on smooth functions. Therefore the problem is a difficult one for trees, which rely on step functions. The overall strength of DIF in items 1 and 2 is again determined by a factor c. In order to achieve comparable results we use the same values of c as in the previous simulations leading to the same DIF strengths of 0.41 (strong), 0.23 (medium), 0.10 (weak) and 0 (no DIF).

Figure 4.5 shows the function of the true underlying item difficulties for item 1 and 2 with strong DIF and the estimated step functions for 50 randomly chosen replications of the simulation drawn with dashed lines for $x \in [-3, 3]$. It is seen that the estimated step-functions in Figure 4.5 capture the underlying structure quite well.

Estimated MSEs of person-parameters θ_p and item-parameters $tr_i(\boldsymbol{x}_p)$ as well as true positive and false positive rates on the item level averaged over all simulations are given in Table 4.4. Again all permutation tests are based on 1000 permutations. In the case of one single predictor x vector $\boldsymbol{\delta}_i$ only has one element, so true positive and false positive rates for the combination of item and variable correspond to those on the item level. As the method of logistic regression can also handle continuous covariates we additionally compute the rates for this approach so that it can be compared to the item focussed trees in terms of DIF detection.

Similar to the results in Table 4.3 true positive rates in Table 4.4 are very high even in the case of weak DIF. For item focussed trees false positive rates are all smaller than 0.05 so the global significance level holds. Logistic regression yields slightly larger true positive rates

Table 4.4.: Estimated MSEs, true positive rates and false positive rates for the simulation scenario with one standard normal distributed predictor as average over 100 simulations.

		Continuous Predictor			
Method		MSE persons	MSE items	TPR_I	FPR_I
IFT	strong	0.4511	0.1585	1.0000	0.0411
	medium	0.4378	0.1533	0.9750	0.0439
	weak	0.4257	0.1439	0.7550	0.0439
	no DIF	0.4346	0.1278	—	0.0440
Logistic	strong	—	—	1.0000	0.0556
	medium	—	—	0.9900	0.0572
	weak	—	—	0.8800	0.0567
	no DIF	—	—	—	0.0545

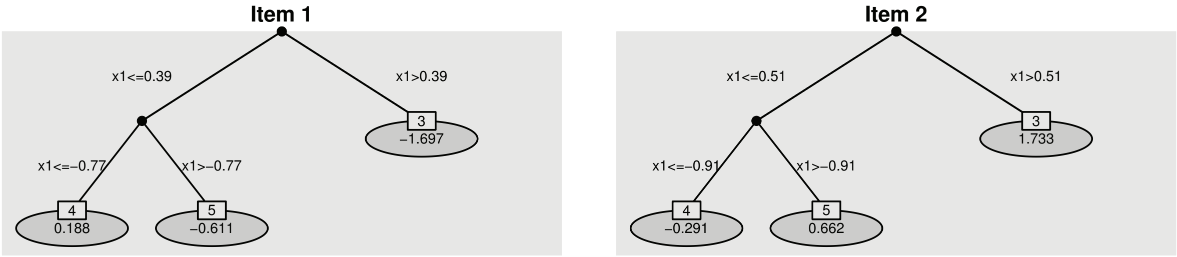


Figure 4.6.: Trees for item 1 and 2 for one estimation of the simulation with one standard normal distributed predictor and strong DIF. Estimated item difficulties are given in each leaf of the trees.

but also larger false positive rates. As was to be expected MSEs of person parameters and item parameters slightly grow with increasing strength of DIF (for item focussed trees).

For item focussed trees single estimation results can also be visualized as tree. Figure 4.6 shows the resulting trees for item 1 and 2 for one exemplary replication of the simulation with strong DIF. The estimated item difficulties are given in each leaf of the trees. In this example two splits are performed for both items. Because of small differences of item difficulties at the borders the algorithm does not perform more splits. A tree with 2 splits or 3 leafs corresponds to a estimated function with 2 steps. The corresponding step functions are marked by dashed lines in Figure 4.5.

The simulation scenario shows that the proposed method is not only able to find relevant DIF items but also to detect complex, especially not linear, structures of DIF. Also in terms of estimation accuracy the algorithm performs quite well.

Table 4.5.: True simulated differences of item difficulties for the three simulation scenarios with four predictors.

Item	Difference of Difficulty		
	Scenario 4	Scenario 5	Scenario 6
1	$1c \cdot I(x1 = 1)$	$1c \cdot I(x2 > 0.1)$	$0.75c \cdot I(x1 = 1) + 0.75c \cdot I(x2 > 0.1)$
2	$-1c \cdot I(x1 = 1)$	$-1c \cdot I(x2 > 0.1)$	$-0.75c \cdot I(x1 = 1) - 0.75c \cdot I(x2 > 0.1)$
3	$1.5c \cdot I(x3 = 1)$	$1.5c \cdot I(x4 > -0.1)$	$0.8c \cdot I(x3 = 1) + 0.8c \cdot I(x4 > -0.1)$
4	$-1.5c \cdot I(x3 = 1)$	$-1.5c \cdot I(x4 > -0.1)$	$-0.8c \cdot I(x3 = 1) - 0.8c \cdot I(x4 > -0.1)$

4.4.2. Several Predictors

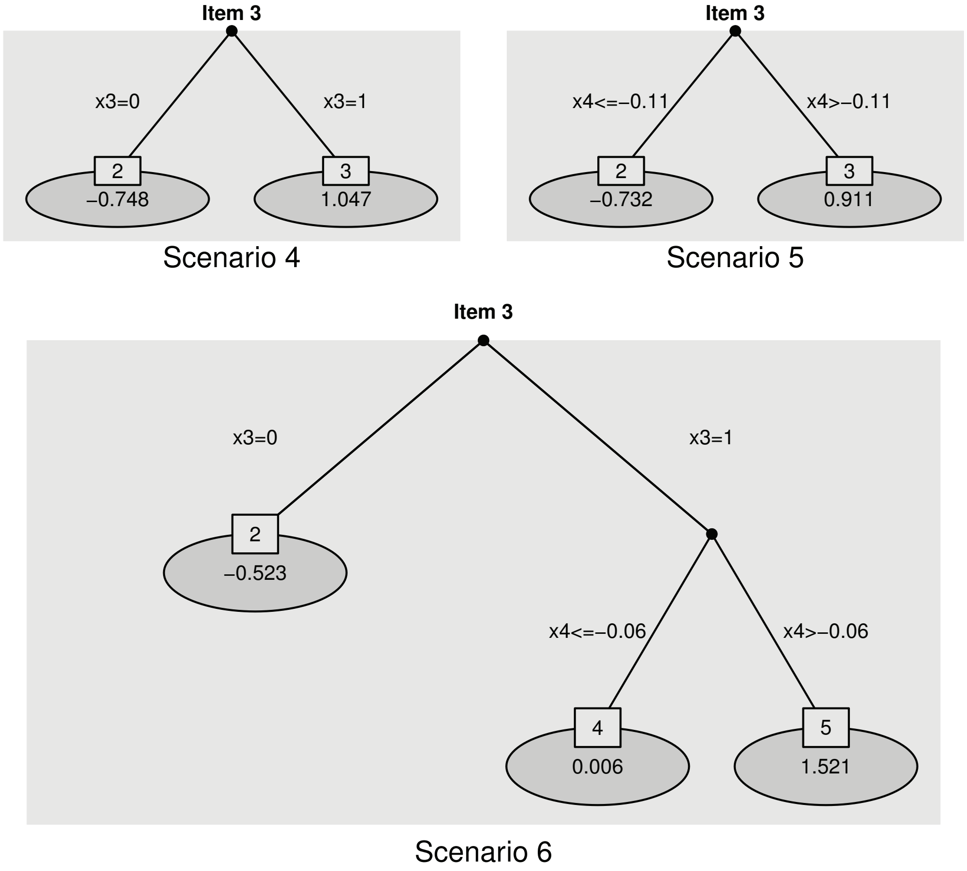
In the following simulations we consider data with four predictors $x1, \ldots, x4$ that potentially induce DIF in 4 out of 20 items. The distributions of the four predictors are

$$x1,\, x3 \sim B(1, 0.5) \quad \text{and} \quad x2,\, x4 \sim N(0, 1).$$

We consider three simulation scenarios with different structures of DIF with respect to items 1, 2, 3 and 4. Differences of item difficulties are defined as given in Table 4.5. In scenario 4 DIF occurs in the binary components $x1$ and $x3$, in scenario 5 DIF occurs in the continous components $x2$ and $x4$ and in scenario 6 it is a more complex structure with DIF in a combination of binary and normal distributed variables. The overall strength of DIF in the four items again depends on the value of c. To obtain strong, medium and weak DIF, c is chosen in the same way as in the previous scenarios.

Figure 4.7 shows one exemplary estimation result of item 3 for each scenario with strong DIF where the true underlying tree structure is detected. The estimated item difficuties are given in each leaf of the trees. The true item parameters for item 3 of the two groups in scenario 4 and 5 are -0.68 and 0.82. In scenario 6 they are -0.68, 0.12 and 0.92. As for all other simulations estimated values are close to the true ones. True and estimated split-points of scenario 5 and 6 regarding to the standard normal variable $x4$ do not differ very much for the exemplary trees in Figure 4.7. Due to the data generating process they are necessarily not exactly the same. For the binary variable $x3$ there is only one possible split.

An overview of the simulation results based on 100 replications is given in Table 4.6. MSEs of person-parameters θ_p and item-parameters $tr_i(\boldsymbol{x}_p)$, true positive and false positive rates on the item level as well as for the combination of items and variables are summarized for the three scenarios and each strength of DIF. All permutation tests are again based on 1000 permutations. To account for the four covariates in the model the local significance level for one test is 0.05/4. If one item is first split in one of the binary components $x1$ or $x3$,



Scenario 4

Scenario 5

Scenario 6

Figure 4.7.: Exemplary estimation results of simulation scenarios 4, 5 and 6 with four predictors and strong DIF. Estimated item difficulties are given in each leaf of the trees.

the local significance level for splits in further nodes has to be adapted as there is no more possible split with regard to $x1$ or $x3$. Consequently it is 0.05/3 in both built nodes. Again we compare with the logistic regression approach as far as DIF detection is concerned.

It is seen that for item focussed trees MSEs of person parameters tend to grow with increasing strength of DIF but are quite stable over all simulations. Hence estimation accuracy is not affected too much by variable and DIF structure. MSEs of item parameters are about the same as in Table 4.4 but do not differ systematically. True positive rates on the item level are very high for medium and strong DIF for each of the three scenarios. Detection of relevant DIF inducing items works well in these settings. In the weak settings only about half of the items with DIF are identified. In scenario 6 DIF is affected by two variables. Here true positive rates for the combination of item and variables are smaller than for scenario 4 and 5. Even for strong DIF the hit rate for item and variable is only about 0.70. However, it is worth noting that in settings 4 and 5 the hit rates for the combination of item and variable are well comparable to the hit rates for items indicating that the identifi-

Table 4.6.: Simulation results for simulation scenarios 4, 5 und 6 with four predictors as the average over 100 simulations.

			MSE		**true positive**		**false positive**	
Scenario	**Method**		persons	items	TPR_I	TPR_{IV}	FPR_I	FPR_{IV}
4	IFT	strong	0.4253	0.1336	0.9825	0.9825	0.0269	0.0096
		medium	0.4069	0.1260	0.8450	0.8425	0.0270	0.0089
		weak	0.4056	0.1272	0.4975	0.4900	0.0263	0.0077
	Logistic	strong	—	—	0.9975	—	0.0619	—
		medium	—	—	0.9350	—	0.0581	—
		weak	—	—	0.6975	—	0.0569	—
5	IFT	strong	0.4176	0.1583	0.9625	0.9625	0.0275	0.0087
		medium	0.4111	0.1474	0.8375	0.8350	0.0313	0.0084
		weak	0.4174	0.1649	0.5300	0.5275	0.0263	0.0064
	Logistic	strong	—	—	0.9450	—	0.0575	—
		medium	—	—	0.8075	—	0.0575	—
		weak	—	—	0.4675	—	0.0563	—
6	IFT	strong	0.4207	0.1516	0.9975	0.7025	0.0269	0.0088
		medium	0.4153	0.1392	0.8750	0.5425	0.0275	0.0083
		weak	0.4086	0.1422	0.4375	0.2363	0.0312	0.0080
	Logistic	strong	—	—	1.0000	—	0.0581	—
		medium	—	—	0.9775	—	0.0563	—
		weak	—	—	0.6450	—	0.0569	—

cation of the variable that induces DIF works. False positive rates are very small across all simulations, in particular the global significance level holds. At most one item without DIF is misleadingly identified as DIF item or one split with regard to a variable that was not inducing DIF is executed during estimation. The logistic regression method yields larger true positive rates than the tree in scenario 4 and 6 but smaller values in scenario 5. In all settings false positive rates are distinctly larger for the logistic regression method.

4.4.3. Comparison with Penalization and Boosting

In current research alternative approaches to detect DIF based on the binary Rasch model (4.2) have been proposed that also allow to include a set of variables as potential candidates for DIF. The general (linear) DIF model proposed by Tutz and Schauberger (2015) has the form

$$\log\left(\frac{P(Y_{pi}=1|\boldsymbol{x}_p)}{P(Y_{pi}=0|\boldsymbol{x}_p)}\right)=\eta_{pi}=\theta_p-(\beta_i+\boldsymbol{x}_p^\top\boldsymbol{\gamma}_i), \tag{4.5}$$

where $\boldsymbol{x}_p$ is the vector of explanatory variables of person p. The item parameters β_i are replaced by $\beta_i+\boldsymbol{x}_p^\top\boldsymbol{\gamma}_i$. DIF is present in item i if the item-specific parameter vector $\boldsymbol{\gamma}_i \neq \mathbf{0}$. Because of the huge number of parameters maximum likelihood estimates will be rather

Table 4.7.: Estimated MSEs, true positive rates and false positive rates for the simulation scenario 1 and 3 as average over 100 simulations.

Scenario	**Method**		MSE persons	MSE items	TPR_I	FPR_I
1	IFT	strong	0.4040	0.1055	0.9950	0.0444
		medium	0.3954	0.1040	0.9625	0.0438
		weak	0.4068	0.1147	0.8100	0.0481
	PenL	strong	0.4046	0.0862	0.9925	0.0369
		medium	0.3899	0.0881	0.9425	0.0244
		weak	0.3790	0.0828	0.4450	0.0100
	Boost	strong	0.4050	0.1036	0.9950	0.0288
		medium	0.3957	0.1021	0.9675	0.0262
		weak	0.4029	0.1153	0.7875	0.0275
3	IFT	strong	0.4511	0.1585	1.0000	0.0411
		medium	0.4378	0.1533	0.9750	0.0439
		weak	0.4257	0.1439	0.7550	0.0439
	PenL	strong	0.4141	0.1079	0.9950	0.0150
		medium	0.4101	0.0933	0.8700	0.0156
		weak	0.3979	0.1175	0.3000	0.0006
	Boost	strong	0.4518	0.1434	1.0000	0.0367
		medium	0.4306	0.1211	0.9900	0.0394
		weak	0.4243	0.1206	0.8450	0.0367

unstable or will even not exist. To solve this problem Tutz and Schauberger (2015) propose a penalization method using a group lasso type penalty, that was introduced by Yuan and Lin (2006). A quite different alternative proposed by Schauberger and Tutz (2015) is to use boosting techniques, that have been developed in statistics by Friedman et al. (2000). As the two approaches are competing methods of our proposed item focussed trees we include them in our simulations. For the computations we used the corresponding R add-on packages `DIFlasso` (Schauberger, 2014) and `DIFboost` (Schauberger, 2015). Although all the methods can handle more complex settings we prefer to compare their performance in the rather simple scenarios with one binary predictor (Scenario 1) and one continuous predictor (Scenario 3), considered in Section 4.4.1 before.

Table 4.7 shows the mean squared errors of person parameters and item parameters as well as the true positive und false positive rates on the item level for IFT, the penalty approach (PenL) and the boosting approach (Boost). Apart from the TPR and FPR of setting 1 the results for IFT are the same as in Table 4.3 and 4.4 but for the sake of completeness they are given again. It can be seen that the penalty approach yields the lowest MSEs across all settings. In particular the MSEs of item-parameters are considerably smaller compared to item focussed trees and the boosting approach. Furthermore, in scenario 3 the penalty and the boosting approach both outperform item focussed trees in terms of MSEs of the item parameters. This is not surprising as the data in scenario 3 is simulated in accordance

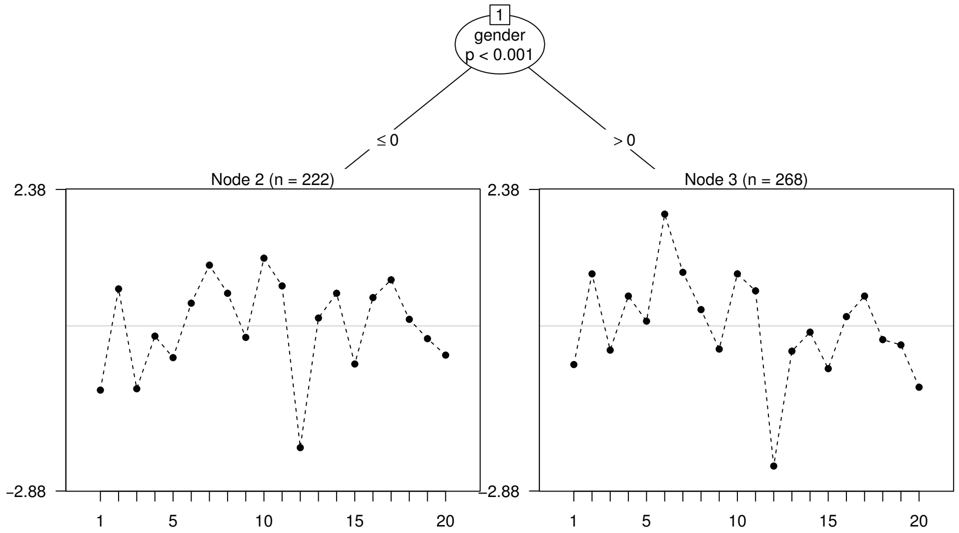


Figure 4.8.: Exemplary Rasch tree for the scenario with six DIF items determined by gender (first 6 items) and age (item 6).

with continuous DIF effects. In terms of DIF detection all methods show very good overall performance for medium and strong DIF. Even in the settings with weak DIF item focussed trees and the boosting approach yield TPR larger than 0.75. Whereas the penalty approach performs very poor in the settings with weak DIF, where the average true positive rates are only 0.445 and 0.300. The false positive rates for the penalty and the boosting approach are distinctly smaller than for item focussed trees across all settings. Both alternatives are even more conservative than the tree-based approach.

4.4.4. Specific Scenarios

In the final simulations we will consider two specific scenarios to point out some important features of the proposed item focussed trees compared to alternative approaches that were already discussed previously in this chapter.

Comparison to Rasch Trees

In the first scenario we aim at demonstrating the difference between the Rasch tree of Strobl et al. (2015) and item focussed trees. As in all previous simulations, we consider 500 persons and 20 items with $\theta_p \sim N(0,1)$ and $\beta_i \sim N(0,1)$ for non-DIF items. DIF items are generated by two variables, one binary with $x_1 \in \{0,1\}$ called gender and one continuous

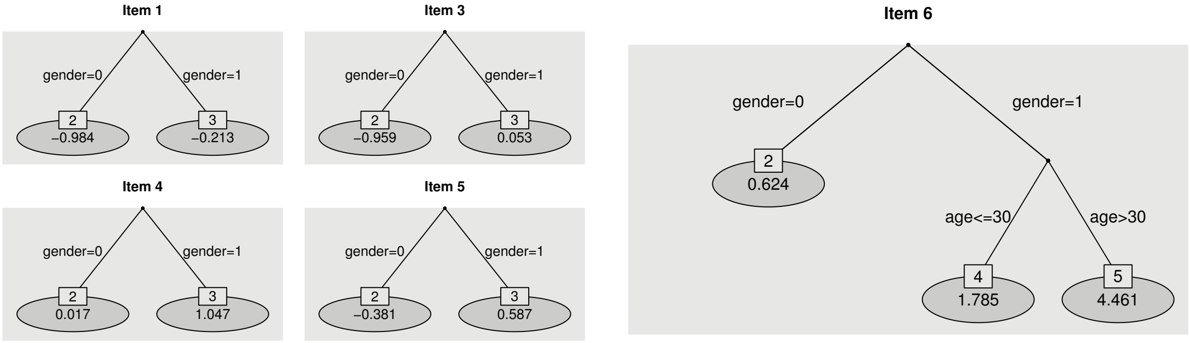


Figure 4.9.: Exemplary trees obtained by the item focussed tree approach for the scenario with six DIF items determined by gender (first 6 items) and age (item 6).

with $x_2 \in [20, 40]$ called age. The first six items are DIF items, in items 1, 2, 3, 4, 5 DIF is induced by gender, in item 6 DIF is induced by gender and age. For the first five items DIF is determined by the step functions $1 \cdot I(x_1 = 1)$, for item 6 we use the step functions, $1 \cdot I(x_1 = 1)$ and $2 \cdot I(x_1 = 1 \,\&\, x_2 > 30)$. The scenario is designed such that all DIF items share one DIF inducing variable but for one item DIF depends also on a second variable.

It is interesting to see if the DIF items are correctly identified together with the variables that induce DIF. Therefore we fitted both methods on 100 replications. By construction in the classical Rasch trees items are not identified but variables are. In all of the replications gender was used in the splitting procedure yielding a hit rate of 1 for gender, however, in only 16 of the 100 replications age was used in the splitting procedure yielding a hit rate of 0.16 for age. A typical tree is shown in Figure 4.8. It shows a split in variable gender but not in age. Moreover, it is hard to see which items carry DIF.

In contrast item focussed trees show which items induce DIF. Overall the true positive rate on the item level was 0.963, the false positive rate 0.030. The hit rate for DIF in item 6 was 1, the hit rate for DIF in gender was 1 and the hit rate for DIF in age 0.97. In 92 of the 100 replications item 6 was split in gender and age, in 5 replications a split in age occurred for another item. Therefore, the item focussed tree approach was well able to detect that in one of the items DIF was induced by both variables gender and age whereas the classical Rasch tree in 84 of the 100 replications used only a split in gender because this variable is stronger in the sense that it induces DIF in several items.

Figure 4.9 shows an example of the trees obtained by item focussed trees. It shows a split in variable gender for items 1, 3, 4, 5 and a split in both variables for item 6. The chosen example is not perfect since there is no split in item 2.

Comparison to Mantel Haenszel

An important advantage of the proposed item focussed trees is that it is able to handle a set of covariates and to identify the relevant regions in which DIF occurs by recursive splitting. Especially in the case of continuous covariates one is not restricted to the comparison of two or several subgroups defined by pre-specified split-points. The algorithm automatically determines the best model by searching over the grid of possible split-points. Nevertheless, in some situations the search for DIF might be costly compared to alternative approaches with prior hypotheses.

We simulate data with 200 persons, 10 items one of them with DIF and one continuous covariate $x \in [18, 70]$, for example like age. Let us consider the hypothesis that DIF is in groups $[18, 30), [30, 70)$ and distinguish between two cases. In the first case (referred to as Scenario 1) DIF is indeed in the intervals $[18, 30), [30, 70)$. Hence DIF is simulated by the step functions $c \cdot I(x \geq 30)$, where $c \in [0, 2]$ determines the strength of DIF. In the second case (referred to as Scenario 2) the hypothesis is wrong and DIF is actually in intervals $[18, 40), [40, 70)$, generated accordingly.

Figure 4.10 shows the true positive rates for item focussed trees (solid lines) and the Mantel-Haenszel (MH) procedure (dashed lines) as a function of the DIF strength c for scenario 1 in the left panel and scenario 2 in the right panel. When using MH the prior hypothesis, which is wrong in scenario 2, is tested in both cases. It can be seen from the left panel that if one has actually DIF in the groups that are tested the power of MH is larger than for IFT. The difference between the two approaches can be seen as the cost of searching for DIF. However, if the assumption is wrong (right panel) the power of MH is worse and IFT shows superior power. This result also holds for other fixed hypotheses methods.

Thus, if one cannot trust the prior hypothesis, that is, the grouping into intervals, IFT seem preferable. If one is only interested in predefined groups or knows that it can only occur in these groups the knowledge might also be used within the IFT framework. Then one tests two groups, and, as has been shown in Section 4.4.1, the performance of the competing procedures is comparable.

4.5. Further Application

As second application we consider data from the Intelligence-Structure-Test 2000 R (I-S-T 2000 R; source of supply is Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0049-551) 999-50-999, www.testzentrale.de), developed by Amthauer et al. (2001); Beauducel et al. (2010). The test is a fundamentally revised version of its predecessors I-S-T 70 (Amthauer et al., 1973) and I-S-T 2000 (Amthauer et al., 1999). The

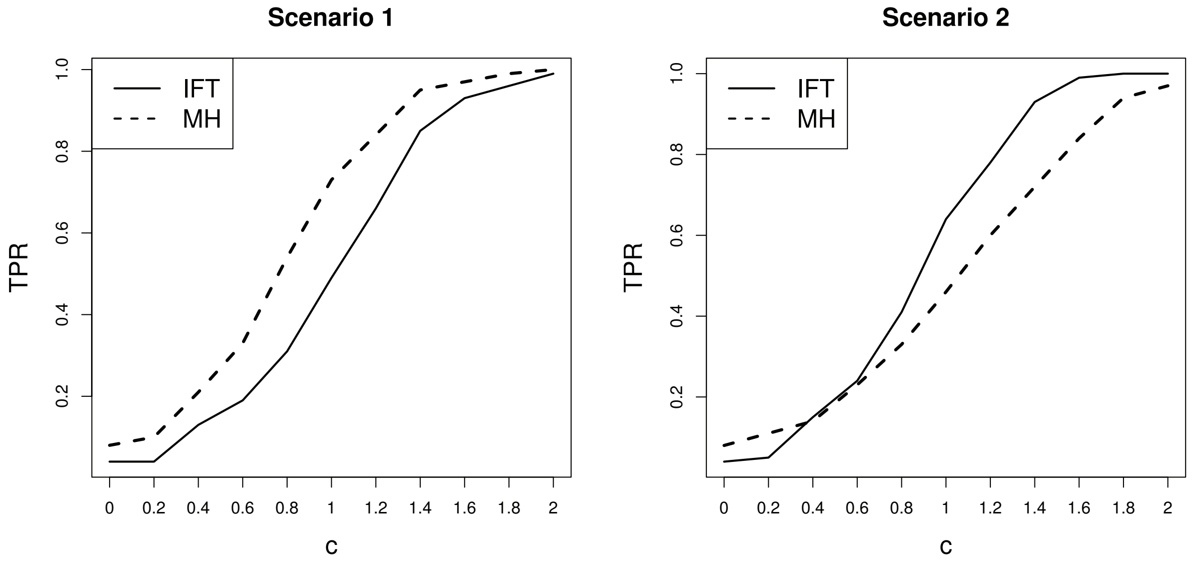


Figure 4.10.: True positive rates as a function of DIF strength c for item focussed trees (solid lines) and the Mantel-Haenszel procedure (dashed lines). In scenario 1 DIF is in intervals $[18, 30), [30, 70)$, in scenario 2 DIF is in intervals $[18, 40), [40, 70)$.

present study was carried out by the Department of Education of the Ludwig-Maximilians University in Munich and has been anaylysed before by (Bühner et al., 2006). The test was conducted at the Phillips University in Marburg. For our analysis we use data from 273 students from 40 different subject areas. The I-S-T 2000 R consists of 9 modules with 20 items each. The first module (items 1 to 20) is about the completion of sentences and asks for sentences where one word is missing. There are five possible solutions for each sentence. The respondent is asked to choose the word that completes the sentences correctly. Further details on the I-S-T 2000 R and its predecessors can be found, for example, in Schmidt-Atzert et al. (1995) and Schmidt-Atzert (2000).

To test for DIF in these items we incorporate the covariates gender (male: 0, female: 1) and age. The distribution of the two covariates and the test result for items 1 to 20 are displayed in Figure 4.11. There are 97 male and 176 female students with age ranging from 18 to 39. The student with the worst result had only 5 correct answers, whereas six students answer all 20 tasks of module 1 correctly.

Using item focussed trees results in only 3 of 20 items showing DIF. The algorithm executed only four splits before stopping ($\alpha = 0.05$). All permutation tests were based on 1000 permutations. Both covariates gender and age are at least once used for splitting and therefore both covariates are included in the model.

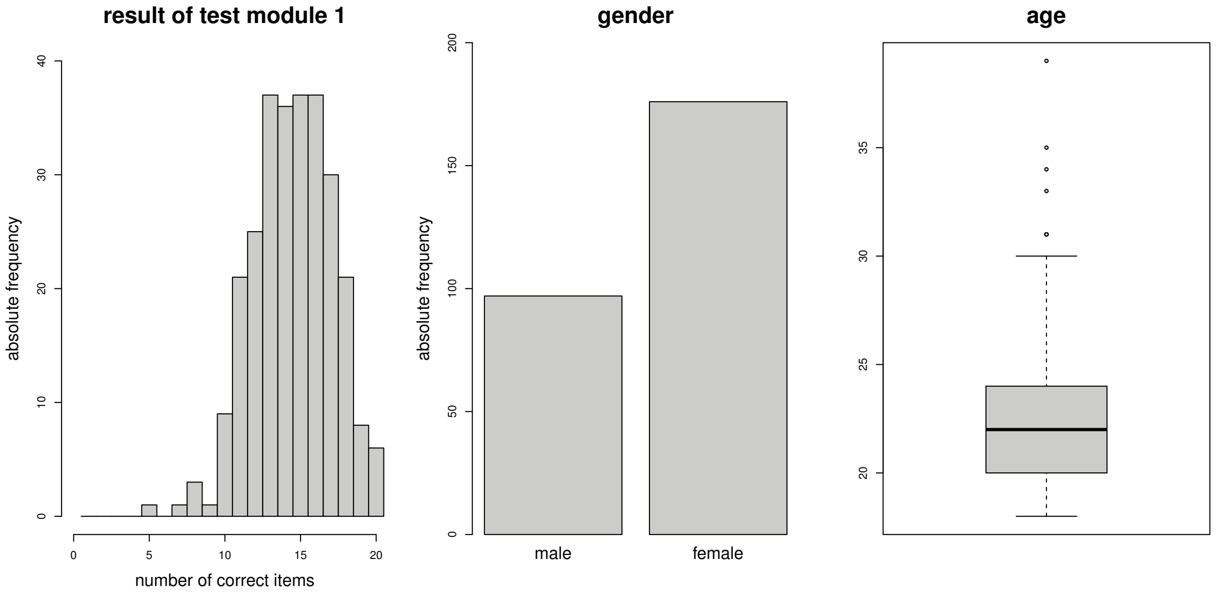


Figure 4.11.: Graphical representation of the results of the first module (items 1 to 20) of the I-S-T 2000 R (left) and the distribution of the two covariates in the analyzed data.

The three items that were identified as DIF items are the following (correct answers are marked in bold):

9: Fathers are ...? (more) experienced than their sons.

a) always **b**) usually c) much d) less e) fundamentally

11: Every river has ...?

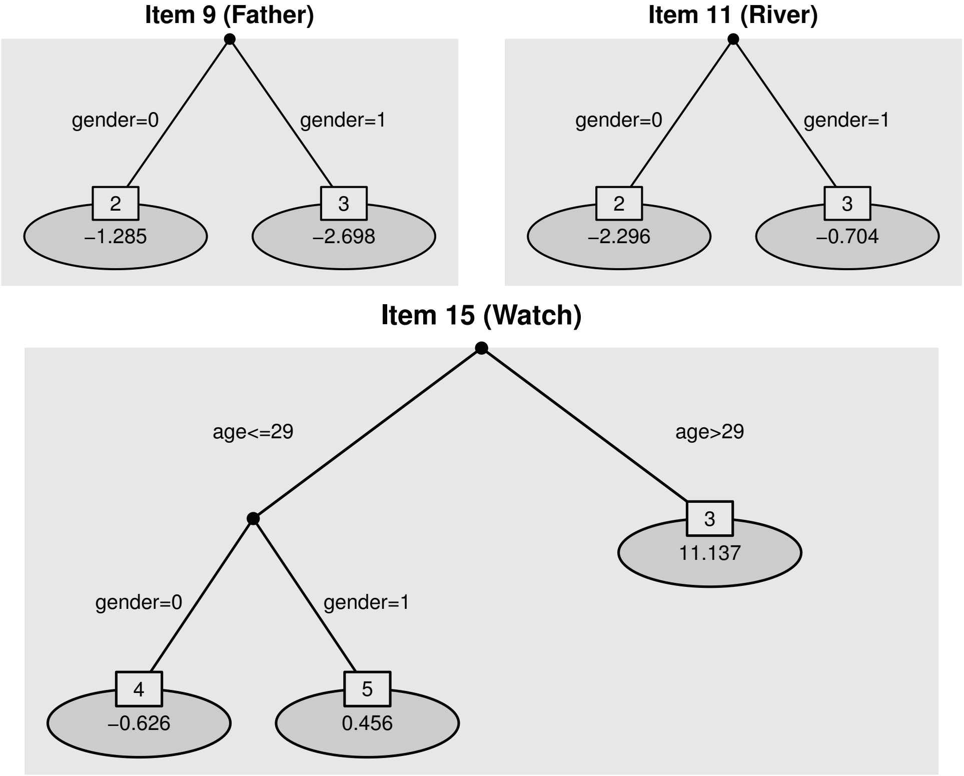
a) fishes b) bridges c) ships **d**) gradients e) rapids

15: A watch always needs (a) ...?

a) battery b) case c) numbers **d**) energy e) hands

The resulting trees for items 9, 11 and 15 are shown in Figure 4.12. Items 9 and 11 show DIF only for gender. The estimated item difficulties show that item 9, which relates to social relations, is easier for females (gender=1) and item 11, which relates to natural sciences, is easier for males (gender=0). Item 15, which relates to technics, is very difficult for all students who are comparably old (age $>$ 29) while for younger students (age $\leq$ 29) it is distinguished between males with an item difficulty of -0.626 and females with a larger item difficulty of 0.456.

The item difficulty of item 15 for students older than 29 given in Figure 4.12 is 11.137. This corresponds to probability 1 for solving the item. In fact no student in the sample, who



Item 9 (Father)

Item 11 (River)

Item 15 (Watch)

Figure 4.12.: Trees for Items 9, 11 and 15 in the subtest sentence completion (IST 2000 R, Amthauer et al., 2001). Estimated item difficulties are given in each leaf of the trees.

was older than 29, answered item 15 correctly. Thus, when searching for the optimal split, the split regarding age and threshold 29 is obviously the best choice. Splitting in this case leads to a pure node with all responses having value 0. A maximum likelihood estimate for the item difficulty in this node does not exist as it tends to infinity. In order to guarantee the existence of all estimates we added a small ridge penalty on the item parameters that ensures that an estimate exists.

4.6. Extension to Ordinal Item Responses

So far we considered an extension of the binary Rasch model, that is, we focussed on dichotomous responses. In psychological tests that contain dichotomous items it is only distinguished if the respondent solved the item correctly or not. However in behavioural research polytomous items are often used to measure performance, personality or attitudes. An example are symmetric response categories on a rating scale from *strongly disagree*, *moderatly disagree*, . . .,*moderatly agree*, *strongly agree*. While several methods are available for dichotomous responses (considered in the previous sections) the ordinal case has not been given very much attention so far. An extension for the method of Strobl et al. (2015) to polytomous items was more recently proposed by El-Komboz et al. (2014). A prominent model that is often used to model ordinal item responses is the partial credit model (PCM) proposed by Masters (1982). We will now introduce an extension of the proposed item focussed trees to detect DIF for ordinal item response by use of the PCM. In order to preserve clarity we now change some notation.

4.6.1. The Partial Credit Model

In the following we consider I items with ordered categories and P persons. For simplicity we assume that the number of categories is equal across items. Let $Y_{pi} \in \{0, 1, \dots, k\}$, $p = 1, \dots, P$, $i = 1, \dots, I$ denote the ordinal response of person p on item i, than the PCM assumes for the probabilities

$$P(Y_{pi} = r) = \frac{\exp(\sum_{l=1}^{r} \theta_p - \delta_{il})}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} \theta_p - \delta_{il})}, \quad r = 1, \dots, k,$$

where θ_p is the person parameter and $(\delta_{i1}, \dots, \delta_{ik})$ are the item parameters of item i. For notational convenience the definition of the model uses implicitly $\sum_{k=1}^{0} \theta_p - \delta_{ik} = 0$. With this convention an alternative form of the model is

$$P(Y_{pi} = r) = \frac{\exp(r\theta_p - \sum_{k=1}^{r} \delta_{ik})}{\sum_{s=0}^{k} \exp(\sum_{k=1}^{s} \theta_p - \delta_{ik})}.$$

The link to the binary Rasch model (4.1) becomes obvious if one considers responses in adjacent categories. Given response categories r and $r-1$, the representation

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r-1)}\right) = \theta_p - \delta_{ir}, r = 1, \dots, k, \tag{4.6}$$

shows that the model is locally a binary Rasch model with person parameter θ_p and item difficulty δ_{ir}.

4.6.2. Item-Focussed Trees for the PCM

In representation (4.6) the linear predictor for person p and the r-th threshold of item i is given by $\eta_{pir} = \theta_p - \delta_{ir}$. As before in item focussed trees the predictor is successively modified by allowing different predictors, or more precisely differences in parts of the predictor, in different regions of the covariate space.

For a more concise description, let $\boldsymbol{x}_p^T = (x_{p1}, \ldots, x_{pm})$ again denote a vector of measurements on person p. Starting from the root, the first split means to examine for all the items, all the variables and possible splits the PCM with predictors

$$\eta_{pir} = \theta_p - [\gamma_{ir,1}^{[1]} I(x_{pj} \leq c_j) + \gamma_{ir,2}^{[1]} I(x_{pj} > c_j)], \quad r = 1, \ldots, k, \tag{4.7}$$

where r now is the threshold and the left and right node are denoted by indices 1 and 2. That means that item i shows DIF generated by the j-th variable x_{pj} at split-point c_j. The item has parameters $\gamma_{i1,1}^{[1]}, \ldots, \gamma_{ik,1}^{[1]}$ for the region $\{x_{pj} \leq c_j\}$ and parameters $\gamma_{i1,2}^{[1]}, \ldots, \gamma_{ik,2}^{[1]}$ for the region $\{x_{pj} > c_j\}$. With predictor (4.7) one explicitly models the *non-homogeneous* case, which means that all the parameters in both nodes can vary freely without any restrictions. An interesting alternative might be the *homogeneous* case, where it is assumed that all thresholds for item i are shifted by an item-specific constant γ_i. Thus for region $\{x_{pj} \leq c_j\}$ the estimated item parameters after the first split are $\delta_{i1} + \gamma_i, \ldots, \delta_{ik} + \gamma_i$.

Further splitting of model (4.7) means that one of the nodes, for example the left node $I(x_{pj} \leq c_j)$, is further split in variable s, yielding a new partition into left and right node and the PCM with predictor

$$\eta_{pir} = \theta_p - [\gamma_{ir,1}^{[2]} I(x_{pj} \leq c_j) I(x_{ps} \leq c_s) + \gamma_{ir,2}^{[2]} I(x_{pj} \leq c_j) I(x_{ps} > c_s) + \gamma_{ir,2}^{[1]} I(x_{pj} > c_j)],$$

where c_s is a new split-point for variable x_{ps} and $\gamma_{i1,1}^{[2]}, \ldots, \gamma_{ik,1}^{[2]}, \gamma_{i1,2}^{[2]}, \ldots, \gamma_{ik,2}^{[2]}$ are the weights on the new split.

4.6.3. Fitting Procedure for the PCM

Estimates of the partial credit model can easily be obtained by embedding the model into the framework of multivariate generalized linear models (GLM). Let the data be given by $(Y_{pi}, \boldsymbol{x}_p)$, $p = 1, \ldots, P$, $i = 1, \ldots, I$. For the item responses one assumes a multinomial distribution $Y_{pi}|\boldsymbol{x}_p \sim M(1, \boldsymbol{\pi}_{pi})$, where $\boldsymbol{\pi}_{pi}^\top = (\pi_{pi1}, \ldots, \pi_{pik})$ with components $\pi_{pir} = P(Y_{pi} = r|\boldsymbol{x}_p)$. The link function of the GLM can be derived from representation (4.6) and has the form

$$g(\pi_{pir}) = \log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)}\right) = (\mathbf{1}_p^{(P-1)})^\top \boldsymbol{\theta} - (\mathbf{1}_r^{(k)})^\top \boldsymbol{\delta}_i,$$

where $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_{P-1})$, $\boldsymbol{\delta}_i^\top = (\delta_{i1}, \ldots, \delta_{ik})$ and $\mathbf{1}_r^{(k)}$ denotes the unit vector of length k with a 1 in component r. As in the dichotomous case, it is required to set $\theta_P = 0$ to ensure the identifiability of the model. The whole parameter vector of the model is than given by $(\boldsymbol{\theta}^\top, \boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_I^\top)$.

Using this representation of the PCM the proposed item focussed trees can straightforward be adapted to the ordinal case. The algorithm described in Section 4.3.1 remains largely the same. The only difference is the basic model (GLM) that is estimated during iteration. For the implementation one can make use of the R-package `VGAM` (Yee, 2010; Yee, 2014), which allows to estimate so-called vector generalized additive models.

4.7. Concluding Remarks

Item focussed recursive partitioning is a modelling tool that allows for simultaneaous detection of items and variables that are responsible for DIF. In particular when several covariates on different scales are available as potentially DIF inducing variables it is an efficient and flexible tool for DIF investigations. Simulation results show that the proposed fitting procedure works quite well in terms of selection performance as well as in terms of estimation accuracy.

Since the proposed item focussed trees, in short IFT, are also based on recursive partitioning as the Rasch trees of Strobl et al. (2015), abbreviated by RT, it seems worth summarizing similarities and differences of the two approaches. Both methods are test based. They use test statistics to identify split-points in the variable space that are linked to DIF. The main differences are

- RT splits the variable space so that the fitted models in subspaces are maximally different. IFT searches for the best splits in single items.
- Consequently, in each subspace RT yields a set of estimated item parameters that characterize the item difficulties in the subspace. Typically the estimated difficulties for all items differ over subspaces. Therefore, all items show (estimated) DIF. Which ones are really to be considered as DIF items has to be decided by a separate decision rule (which still has to be found).
- In contrast IFT identifies the items for which a split is warranted (based on test statistics). Thus the recursive partitioning method itself identifies the DIF items.
- The differences in the partitioning scheme has consequences for the algorithm. While RT simply fits separately within subspaces by using the observations within the corresponding subspaces, IFT fitting uses always all the observations. In IFT the model itself accounts for the different functioning in subspaces. Therefore the algorithms differ distinctly and use different test statistics.

Let us finally discuss potential extensions to the 2PL and 3PL models. In the more general 3PL model the predictor has the form $\eta_{pi} = \delta_i + \alpha_i(\theta_p - \beta_i)$ with the additional chance parameter δ_i and the item discrimination parameter α_i. In this model DIF could be induced by differences in item difficulty parameter or the item discrimination parameter (ignoring the possibility that also the chance parameter could be modified by explanatory variables). An extended version of item focussed trees should be able to detect both forms of DIF. The first can be modelled, as before, by replacing the item difficulty β_i by $\gamma_{il} I(x_{pj} \leq c_j) + \gamma_{ir} I(x_{pj} > c_j)$, the latter by replacing the item discrimination α_i by $\tilde{\gamma}_{il} I(x_{pj} \leq c_j) + \tilde{\gamma}_{ir} I(x_{pj} > c_j)$. However, fitting of the corresponding model cannot be embedded into the framework of generalized linear models. One has to design specific software that is able to fit such models, for example, by integrating out the person parameters to obtain marginal estimates. In addition, one needs test statistics that compare the model without splits and the model with splits to obtain appropriate summary tests for the necessity to split on the ability or discrimination level and a criterion to select the appropriate split. In summary, the concept may be extended to more general models, but since estimation and testing is much more difficult serious research is necessary to accomplish the task. This is certainly an interesting topic for future research.

The results shown in this chapter were obtained by the R-package `DIFtree` (Berger, 2016a) version 1.1.0 that is available on CRAN.

5. Detecting Uniform and Non-Uniform DIF by Logistic Regression

5.1. Introduction

In recent years differential item functioning (DIF) and DIF identification methods have been areas of intensive current research. Differential item functioning occurs if the probability of a correct response among persons with the same value of their underlying trait differs in subgroups, for example, if the difficulty of an item depends on the membership to a racial, ethnic or gender subgroup. If a test contains DIF items it may be unfair, that is, favor specific groups. When developing and using tests that measure latent abilities one should be aware of the phenomenon of DIF. Ideally tests should not contain suspicious items. If this cannot be obtained one should at least know which items are DIF items and by which covariates DIF is generated. For more details on DIF, measurement bias and possibly discrimination, see, for example, Holland and Wainer (1993), Osterlind and Everson (2009), Rogers (2005), Millsap and Everson (1993) and Zumbo (1999).

A variety of methods to detect DIF has been proposed, for a more recent overview see Magis et al. (2010). One can in particular distinguish between *item response theory (IRT) modelling approaches* and *test score methods* (Magis et al., 2015). The former assume that an IRT model holds in each group. Tests as Lord‘s test or likelihood ratio tests are used to detect differences of item parameters between groups. IRT approaches have been used, among others, by Lord (1980), Raju (1988) and Holland and Wainer (1993). Test score methods use a matching variable as, for example, Mantel-Haenzel test procedures (Holland and Thayer, 1988) or logistic regression modelling (Swaminathan and Rogers, 1990). We will use the logistic regression framework since it also allows to investigate non-uniform DIF. Uniform DIF is present if the (scaled) differences in the probabilities of solving an item of subjects from different groups but with the same ability level do not depend on the

This chapter is a modified version of Berger and Tutz (2015a). For more information on the personal contributions of the authors and textual matches, see page 10.

common ability level. In non-uniform DIF scenarios the differences are not constant across ability levels and crossing item response curves may occur.

More recently IRT based DIF modelling has been extended to allow for continuous variables that induce DIF (compare also Chapter 4). The corresponding latent trait models contain many parameters since each item comes with an own vector of parameters. Therefore maximum likelihood estimates are bound to fail. Tutz and Schauberger (2015) used a penalty approach to regularize parameter estimation whereas Schauberger and Tutz (2015) used boosting techniques. A non-IRT modelling approach with regularization by penalties has been proposed by Magis et al. (2015).

This chapter focusses on score based methods. A recursive partitioning (tree based) method is proposed that allows to identify the items that carry DIF together with the variables that induce DIF. The variables can represent groups as in classical DIF detection techniques but can also include continuous variables like age. A strength of the method is that for continuous variables it is not necessary to define a priori the intervals that are relevant, the method itself generates the intervals that are linked to DIF. The resulting tree visualizes in a simple way the structure of DIF in an item showing which variables and interactions of variables generate DIF. The method is related to the recursive partitioning method proposed in Chapter 4. The basic concepts remain the same but are adapted to the logistic regression approach for DIF detection.

The method should be distinguished from the Rasch trees proposed by Strobl et al. (2015). One difference between the methods is that Rasch trees are IRT based methods designed for uniform DIF only. However, also for the detection of uniform DIF there are strong differences between the methods. By using tree methodology the Rasch tree method also does not need pre-specified subgroups and can handle continuous variables. Rasch trees recursively partition the covariate space to identify regions of the covariate space in which DIF occurs by fitting separate item response models in these regions. Regions are suspected to be relevant if the parameter estimates in the regions differ strongly. Therefore, regions in the covariate space are identified that show different difficulties but the method does not flag items that are responsible. In contrast, the recursive partitioning method proposed here focusses on the detection of the items that are responsible for DIF. Recursive partitioning is used on the item level not on the global level, which treats all items simultaneously and therefore does not show which item is responsible for the occurrence of DIF. Chapter 4 already provides a more detailed discussion of the different ways of using tree methodology and illustrate the difference in applications and simulations.

In Section 5.2 we introduce the item focussed tree approach based on the logistic regression model for uniform DIF and in Section 5.3 we present an illustrative example. A detailed description of the fitting procedure is given in Section 5.4. In Section 5.5 we consider the results of various simulations. Models for the extension to non-uniform DIF are considered in Section 5.6. Finally, Section 5.7 contains two applications on real data.

5.2. Logistic Regression Approaches to DIF

In this section basic logistic regression approaches to the detection of uniform DIF are described and the alternative tree based method is introduced.

5.2.1. Linear Logistic Regression Approaches to DIF

The basic test score based method to detect uniform DIF was proposed by Swaminathan and Rogers (1990). The method was already shortly sketched in Section 4.2.3. It can be seen as a starting point of the method proposed here.

Let $Y_{pi} \in \{0,1\}$, $p = 1,\dots,P$, $i = 1,\dots,I$ denote the response when person p tries to solve item i. Swaminathan and Rogers (1990) proposed to model the probability of solving an item as a function of the group membership and the test score by fitting the logistic regression model

$$\log\left(\frac{P(Y_{pi}=1|S_p,g)}{P(Y_{pi}=0|S_p,g)}\right) = \eta_{pi} = \beta_{0i} + S_p\beta_i + \gamma_{ig}, \tag{5.1}$$

where g denotes the group, S_p is the test score of person p, β_{0i} is the intercept, β_i is the slope of item i and γ_{ig} are the group-specific parameters. In this model the parameters $\beta_{01},\dots,\beta_{0I}$ represent the item difficulties and the parameters $\beta_1,\dots,\beta_I$ correspond to discrimination parameters. Within this framework the test scores are considered as proxies for the abilities of persons. For the detection of DIF the most interesting parameters are the group-specific parameters $\gamma_{i1},\dots,\gamma_{iG}$, where G denotes the number of groups. They represent the differential item functioning. In the simplest case of two groups, a reference group and a focal group, one chooses $\gamma_{i1} = 0$ for the reference group. Thus, for example, with groups defined by gender with female as the reference group one has

$$\beta_{0i} + \gamma_{i,male} \text{ for males} \quad \text{and} \quad \beta_{0i} \text{ for females.} \tag{5.2}$$

If $\gamma_{i,male} \neq 0$ one has DIF in item i generated by gender. The original framework for two groups was proposed by Swaminathan and Rogers (1990), the extension to multiple groups was considered by Magis et al. (2011). In the multiple group case one of the G groups, for example the first group, has to be chosen as reference group by setting $\gamma_{i1} = 0$.

DIF detection within the logistic regression framework typically uses likelihood ratio statistics that test the null hypothesis $H_0 : \gamma_{i1} = \dots = \gamma_{iG} = 0$. If the hypothesis is rejected item

i is considered as DIF item. Each item is tested separately at significance level α with the degrees of freedom equal to $G-1$, depending on the number of groups.

The basic concept can be simply extended to include continuous (and categorical) variables that might induce DIF. Let $\boldsymbol{x}_p^\top = (x_{p1}, \ldots, x_{pm})$ be a vector of person-specific explanatory variables of length m. An extension of model (5.1) for uniform DIF has the form

$$\log\left(\frac{P(Y_{pi}=1|S_p,\boldsymbol{x}_p)}{P(Y_{pi}=0|S_p,\boldsymbol{x}_p)}\right) = \eta_{pi} = \beta_{0i} + S_p\beta_i + \boldsymbol{x}_p^\top\boldsymbol{\gamma}_i. \tag{5.3}$$

The new intercept parameters in model (5.3) are $\beta_{0i} + \boldsymbol{x}_p^\top\boldsymbol{\gamma}_i$ and they differ according to the characteristics of the person $\boldsymbol{x}_p$. The comparison of multiple groups is just a special case. Setting the first group as reference one defines the vector of explanatory variables $\boldsymbol{x}_p^\top = (x_{p2}, \ldots, x_{pG})$, where $x_{pg} = 1$ if person p is from group g and 0 otherwise. The corresponding vector of parameters for one item i is $\boldsymbol{\gamma}_i^\top = (\gamma_{i2}, \ldots, \gamma_{iG})$. Uniform DIF is present in this item if $\boldsymbol{\gamma}_i \neq \mathbf{0}$. To investigate DIF one uses a global test for the whole parameter vector, $H_0 : \boldsymbol{\gamma}_i = 0$. The alternative hypothesis is that at least one of the parameters are unequal to zero. The hypotheses are tested separately for each item at significance level α. Due to the design of the tests the approach identifies the items that carry DIF but does not contain any information about the components of $\boldsymbol{x}_p$ that are responsible for DIF. Although being a straightforward extension of the fixed groups DIF model (5.1) the extension (5.3) seems not to have been investigated so far.

We will refer to the multiple groups model (5.1) as the *classical* logistic regression modelling approach and to model (5.3) as the *extended* approach. It should be mentioned that the extended approach (including continuous or categorical covariates) is already implicitly contained in the approach proposed by Magis et al. (2011). The approach of Magis et al. (2015) provides an extra layer of complexity with penalization on the DIF parameters. The main contribution in this chapter, which is outlined in the following sections, is that the linear part of the basic model is replaced by tree structured fitting.

5.2.2. A Tree Representation of DIF

DIF detection based on the logistic regression model as described in the previous section has some limitations and drawbacks. If one uses the traditional version with G groups DIF can be induced only by group membership. A continuous variable like age has to be divided into intervals to obtain groups without knowing which intervals are important. The extended version with a linear predictor is restricted by the assumption that the DIF effect is linear. Moreover, the tests that are used to identify items that carry DIF do not show which variables are responsible for DIF, at least not in a simple way. The proposed recursive partitioning method avoids the problem that reference and focal groups have to be specified a priori. By recursive splitting the method itself identifies the groups that induce DIF if they are present.

The general concept of recursive partitioning has its roots in automatic interaction detection. The most popular modern version is due to Breiman et al. (1984) and is known by the name *classification and regression trees*, or CART. An alternative approach is the recursive partitioning framework based on conditional inference proposed by Hothorn et al. (2006). The basic method is conceptually very simple. By binary recursive partitioning the feature space is partitioned into a set of rectangles, and on each rectangle a simple model (for example, a constant) is fitted. An easily accessible introduction into basic concepts is found in Hastie et al. (2009), an overview with a focus on psychometrics was given by Strobl et al. (2009). It should be noted that the method proposed here is based on the same idea but there is one crucial difference. When fitting a model we do not fit two separate models within the rectangles obtained by partitioning. We fit one closed model and only the intercept is partitioned into rectangles. This yields item focussed trees in contrast to global trees as used by conventional Rasch trees.

Building a tree means to successively find a partition of the predictor space, where each node represents a subset of the predictor space. The terminal nodes of the tree build a disjoint partition of the predictor space and correspond to the relevant subregions of interest. When growing a tree one typically splits one node A into two subsets A_1 and A_2. The split is determined by exactly one variable and the construction of the split depends on the scale of the variable. In the following considerations we will focus on metrically scaled and ordinal variables. In this case the partition into two subsets has the form

$$A_1 = A \cap \{x_j \leq c\} \quad \text{and} \quad A_2 = A \cap \{x_j > c\},$$

with regard to threshold c on variable x_j. Given the covariates $\boldsymbol{x}_p$ one can account for uniform DIF by building a partition of the respondents with differing intercepts. The first split with regard to the j-th variable and corresponding split-point c_j means to fit the model with predictor

$$\eta_{pi} = S_p\beta_i + [\gamma_{il}^{[1]} I(x_{pj} \leq c_j) + \gamma_{ir}^{[1]} I(x_{pj} > c_j)], \tag{5.4}$$

where $I(\cdot)$ denotes the indicator function with $I(a) = 1$ if a is true and $I(a) = 0$ otherwise. The parameter $\gamma_{il}^{[1]}$ denotes the intercept in the left node ($x_{pj} \leq c_j$) and $\gamma_{ir}^{[1]}$ the intercept in the right node ($x_{pj} > c_j$). For example one split with regard to the binary covariate gender yields the intercepts

$$\gamma_{il}^{[1]} = \gamma_{i,male} \text{ for males} \quad \text{and} \quad \gamma_{ir}^{[1]} = \gamma_{i,female} \text{ for females.}$$

This parametrization is an equivalent representation of (5.2). The main difference is that the two subgroups of interest are not predefined but determined by a split in variable j at split-point c_j. To determine the first split one examines all the null hypotheses $H_0 : \gamma_{il}^{[1]} = \gamma_{ir}^{[1]}$. If H_0 cannot be rejected for any combination of variable and split-point the item is considered to be free of DIF. In the proposed algorithm likelihood ratio tests are used to examine the null hypotheses. In the very first step one chooses the combination of item, variable and split-point with the smallest p-value of the corresponding test. If a significant effect is found the first split into left and right node is carried out for the selected item. In Section 5.4 the splitting criterion is described in more detail.

One further split, for example in the right node ($x_{pj} > c_j$), with regard to the s-th variable at split-point c_s yields the two daughter nodes $I(x_{pj} > c_j)I(x_{ps} \leq c_s)$ and $I(x_{pj} > c_j)I(x_{ps} > c_s)$. The new nodes are both defined by the product of two indicator functions. In general each node can be represented by a product of several indicator functions, namely

$$node(\boldsymbol{x}_p) = \prod_{b=1}^{B} I(x_{pj_b} > c_{j_b})^{a_b} I(x_{pj_b} \leq c_{j_b})^{1-a_b},$$

where B is the total number of indicator functions or branches, c_{j_b} is the selected split-point in variable j_b and $a_b \in \{0, 1\}$ indicates which of the indicator functions, below or above the threshold, is involved. The resulting predictor of the model for item i after several splits with terminal nodes $\ell = 1, \ldots, L_i$ is than given by

$$\eta_{pi} = S_p\beta_i + \sum_{\ell=1}^{L_i} \gamma_{i\ell}\, node_{i\ell}(\boldsymbol{x}_p) = S_p\beta_i + tr_i(\boldsymbol{x}_p), \tag{5.5}$$

where $tr_i(\boldsymbol{x}_p)$ is the tree component containing subgroup-specific intercepts represented by the terminal nodes $node_{i\ell}(\boldsymbol{x}_p)$. The proposed algorithm yields an individual tree for each item that was selected to carry DIF. If an item is never chosen for splitting it is assumed to be free of DIF, and the fitted ”tree” is a constant $tr_i(\boldsymbol{x}_p) = \beta_{0i}$.

We use the model abbreviation *IFT* for item focussed trees based on the logistic regression framework.

5.3. An Illustrative Example

The procedure is now first illustrated by the use of artificial data. We consider data $Y_{pi},\ p = 1, \ldots, 800,\ i = 1, \ldots, 20$, that are generated by a two-parameter model (2PL) with DIF. The basic 2PL model has the form

$$P(Y_{pi} = 1|\theta_p, b_i, a_i) = \frac{\exp\left(a_i(\theta_p - b_i)\right)}{1 + \exp\left(a_i(\theta_p - b_i)\right)},$$

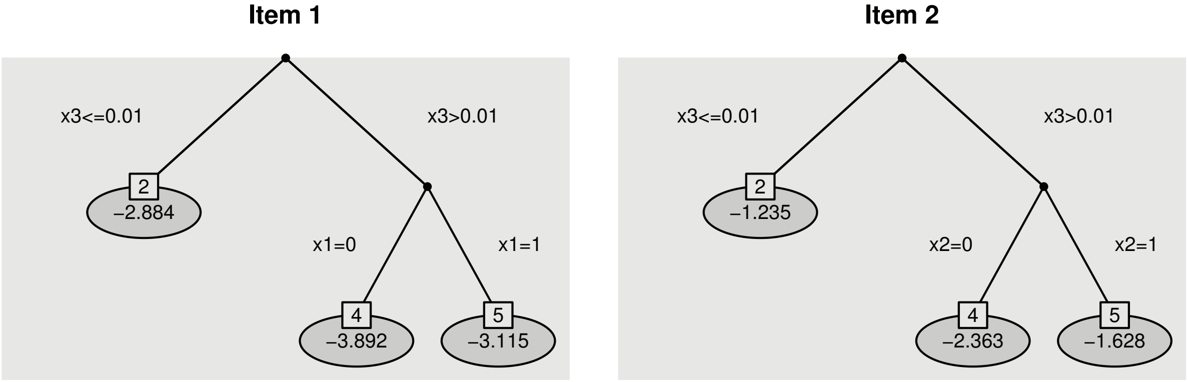


Figure 5.1.: Estimated trees of item 1 and 2 for the illustrative example. Estimated coefficients $\gamma_{i\ell}$ are given in each leaf of the trees.

where θ_p denotes the person ability, b_i the item difficulty and a_i the item discrimination. We first generate person parameters θ_p and item difficulties b_i from a standard normal distribution and item discriminations a_i from a uniform distribution. However, instead of generating data from the 2PL model we assume that the difficulties of two of the 20 items depend on covariates in a complex pattern.

In detail, we consider three covariates, two binary variables $x1$, $x2 \sim B(1, 0.5)$ and one standard normal distributed variable $x3 \sim N(0, 1)$. In item 1 DIF is induced by $x1$ and $x3$ and the modified value of the difficulty is determined by the step functions $b_{1,\text{mod}} = b_1 + 0.8 \cdot I(x_3 > 0) + 0.8 \cdot I\left(\{x_3 > 0\} \,\cap\, \{x_1 = 0\}\right)$, in item 2 DIF is induced by $x2$ and $x3$ and we use the step functions $b_{2,\text{mod}} = b_2 + 0.8 \cdot I(x_3 > 0) + 0.8 \cdot I\left(\{x_3 > 0\} \,\cap\, \{x_2 = 0\}\right)$, which represents an interaction between variables $x2$ and $x3$. In order to evaluate the fitting procedure 100 data sets were generated.

Figure 5.1 shows one exemplary estimation result of the two items with DIF (item 1 and 2) when fitting *IFT*. The estimation in this example is quite perfect because the true underlying tree structure is detected for both items and no further item is falsely identified as DIF item. It can be seen from the trees that there are three groups represented by three terminal nodes, respectively. For item 1 it is distinguished between $\{x3 \leq 0.01\}$ and $\{x3 > 0.01\}$, and within this group between $\{x1 = 0\}$ and $\{x1 = 1\}$. The corresponding intercepts $\hat{\gamma}_{1\ell}$ and $\hat{\gamma}_{2\ell}$, $\ell = 1, \ldots, 3$, of the estimated model (5.5) are given in each leaf of the trees. According to model (5.5), the probability to solve the item correctly increases with increasing intercepts. From the estimates in Figure 5.1 one can derive that item 1 is most difficult for region $\{x_3 > 0.01\} \,\cap\, \{x_1 = 0\}$ and item 2 is most difficult for $\{x_3 > 0.01\} \,\cap\, \{x_2 = 0\}$. These results are exactly in line with the true simulated effects. In the simulations in Section 5.5 this artificial data is, inter alia, again considered in more detail.

5.4. Fitting Procedure

In this section we give details about the fitting procedure for our proposed item focussed trees to investigate uniform DIF. The basic concepts are the same as for item focussed trees in the Rasch model proposed in Chapter 4.

5.4.1. Concepts

When building trees for single items in each step one has to identify the best split due to an optimality criterion and decide if there is a relevance to perform the split or not. The second determines when to stop and therefore at the same time determines the size of the trees.

Since the approach is based on logistic regression models it is quite natural to use test based splits. In each step of the fitting procedure one obtains p-values for the two parameters that are involved in the splitting. In our previous notation one examines all the null hypotheses $H_0 : \gamma_{il} = \gamma_{ir}$ for each combination of item, variable and split-point. One simply selects the combination as the optimal one that has the smallest p-value. As test statistic we use the likelihood ratio (LR) test statistic. Computing the LR test statistic requires to estimate both models, the full model and the restricted model under H_0. We nevertheless prefer the LR statistic because it corresponds to select the model with minimal deviance. This criterion on the other hand is equivalent to minimizing the entropy, which belongs to the family of impurity measures that were already introduced as splitting criteria by Breiman et al. (1984).

In order to decide if the split should be performed or not we use a concept based on maximally selected statistics. The idea is to perform a test that investigates the null hypotheses of independence of the response and one of the covariates at the global variable level. For one fixed item i and variable j one simultaneously considers all LR test statistics T_{jc_j}, where c_j are from the set of possible split-points, and computes the maximal value statistic $T_j = max_{c_j} T_{jc_j}$. The p-value that can be obtained by the distribution of T_j provides a measure for the relevance of variable j. The result is not influenced by the number of split-points, since it has already taken into account, see Hothorn and Lausen (2003), Shih (2004), Shih and Tsai (2004), Strobl et al. (2007). As the distribution of T_j in general is unknown we use a permutation test to obtain a decision on the null hypotheses. The distribution of T_j is determined by computing the maximal value statistics based on random permutations of variable j. A random permutation of variable j breaks the relation of the covariate and the response in the original data. By computing the maximal value statistics for a large number of permutations one obtains an approximation of the distribution under the null hypotheses and an corresponding p-value. All computations in the present chapter are based on 1000 permutations. Given overall significance level α the local significance level of one permutation test for fixed item and variable is chosen as α/m. Using this adaption the probability for each item without DIF of being falsely classified as DIF item is controlled by α. As usual in DIF detection one controls for the type I error that is also known as false alarm rate. However, on the item level one should adapt for multiple testing. Choosing α/m ensures that the probability of falsely identifying at least one variable as responsible for DIF is controlled by α.

It should be noted that in general the number of permutations should depend on the number of covariates m. In our simulations and applications the maximal number of covariates is 3. Therefore, with a sample of 1000 permutations the p-values are determined with sufficient accuracy. From our experience it is recommended to use at least 200 permutations for settings with one covariate and to increase the number of permutations by 200 per covariate. Thus, a lower bound for settings with 3 covariates are 600 permutations.

5.4.2. The Basic Algorithm

The basic algorithm for uniform DIF is the following.

Basic Algorithm - Uniform DIF

Step 1 (Initialization)

Set counter $\nu = 1$

(a) Estimation

For all items $i = 1, \dots, I$, fit all the candidate logistic models with predictor

$$\eta_{pi} = S_p\beta_i + \gamma_{i1} I(x_{pj} \le c_{ijk}) + \gamma_{i2} I(x_{pj} > c_{ijk}), \\ j = 1, \dots, m, \quad k = 1, \dots, K_j.$$

(b) Selection

Select the model that has the best fit. Let c_{i_1,j_1,k_1} denote the best split, which is found for item i_1 and variable x_{j_1}.

(c) Splitting decision

Select the item and variable with the largest value of T_j. Carry out permutation test for this combination with significance level α/m. If significant, fit the selected model yielding estimates $\hat{\beta}_i$, $\hat{\gamma}_{i_1,1}$, $\hat{\gamma}_{i_1,2}$ and nodes $node_{i_1,1}, node_{i_1,2}$, set $\nu = 2$. If not, stop, no DIF detected.

Step 2 (Iteration)

(a) Estimation:

For all items $i = 1, \dots, I$ and already built nodes $\ell = 1, \dots, L_{i\nu}$, fit all the candidate logistic models with new intercepts

$$\gamma_{i,L_{i\nu}+1} node_{i\ell} I(x_{pj} \leq c_{ijk}) + \gamma_{i,L_{i\nu}+2} node_{i\ell} I(x_{pj} > c_{ijk})$$

for all j and remaining, possible split-points c_{ijk}.

(b) Selection

Select the model that has the best fit yielding the split-point c_{i_ν,j_ν,k_ν}, which is found for item i_ν in node $node_{i_\nu,\ell_\nu}$ and variable x_{j_ν}.

(c) Splitting decision

Select the node and variable with the largest value of T_j. Carry out permutation test for this combination with significance level α/m. If significant, fit the selected model yielding the additional estimates $\hat{\gamma}_{i_\nu,L_{i_\nu,\nu}+1}, \hat{\gamma}_{i_\nu,L_{i_\nu,\nu}+2}$, set $\nu = \nu+1$. If not, stop.

5.5. Simulations

In the following we consider data $Y_{pi},\ p = 1, \dots, P,\ i = 1, \dots, I$ that are generated according to the two-parameter model (2PL), which is a dichotomous IRT model of the form

$$P(Y_{pi} = 1|\theta_p, a_i, b_i) = \frac{\exp\left(a_i(\theta_p - b_i)\right)}{1 + \exp\left(a_i(\theta_p - b_i)\right)}, \tag{5.6}$$

where θ_p are the person abilities, b_i are the item difficulties and a_i are the item discrimination parameters.

We consider several simulation scenarios where in a first step the person parameters θ_p and the item difficulties b_i are independently drawn from a standard normal distribution and the item discrimination parameters a_i are uniformly distributed, $a_i \sim U(0, 1)$. If an item i is assumed to show uniform DIF the corresponding parameter b_i is subsequently transformed by specific step functions in each scenario. A detailed description is given in the respective section.

In each simulation scenario we vary the number of persons, $P \in \{400, 800\}$, the number of items, $I \in \{20, 40\}$, and the percentage of DIF items, which is 0%, 10% or 20%. In the cases with DIF we additionally consider three different strengths of DIF, given for each scenario in the respective section. In total this results in 28 different settings (4 without DIF and 24 with DIF), respectively. In each setting 100 data sets were generated. During estimation each permutation test is based on 1000 permutations.

In order to evaluate the performance of the proposed tree based model (5.5) we compute true positive rates (TPR), also named hit rates, and false positive rates (FPR), which correspond to the Type I error rates if no DIF is present. We distinguish between TPR and FPR on the item level and for the combination of item and variable. Let each item be characterized by a vector $\boldsymbol{\delta}_i^T = (\delta_{i1}, \ldots, \delta_{im})$, where m denotes the number of covariates, with $\delta_{ij} = 1$ if item i has DIF in variable j and $\delta_{ij} = 0$ otherwise. An item is a non-DIF item if $\boldsymbol{\delta}_i^T = (0, \ldots, 0)$, if one of the components is 1 it is a DIF item. With indicator function $I(\cdot)$, the criteria to judge the identification of items with DIF are:

- True positive rate on the item level:

  $TPR_I = \frac{1}{\#\{i:\boldsymbol{\delta}_i \neq \mathbf{0}\}} \sum_{i:\boldsymbol{\delta}_i \neq \mathbf{0}} I(\hat{\boldsymbol{\delta}}_i \neq \mathbf{0})$

- False positive rate on the item level:

  $FPR_I = \frac{1}{\#\{i:\boldsymbol{\delta}_i = \mathbf{0}\}} \sum_{i:\boldsymbol{\delta}_i = \mathbf{0}} I(\hat{\boldsymbol{\delta}}_i \neq \mathbf{0})$

- True positive rate for the combination of item and variable:

  $TPR_{IV} = \frac{1}{\#\{i,j:\delta_{ij} \neq 0\}} \sum_{i,j:\delta_{ij} \neq 0} I(\hat{\delta}_{ij} \neq 0)$

- False positive rate for the combination of item and variable:

  $FPR_{IV} = \frac{1}{\#\{i,j:\delta_{ij} = 0\}} \sum_{i,j:\delta_{ij} = 0} I(\hat{\delta}_{ij} \neq 0).$

The methods that are considered in the simulations are

- *Logistic*, which denotes the *classical* regression method proposed by Swaminathan and Rogers (1990) and Magis et al. (2011). If the predictor is a vector with possibly continuous variables it denotes the *extended* logistic model.

- *IFT* for item focussed trees based on the logistic model, which describes the recursive partitioning method proposed here.

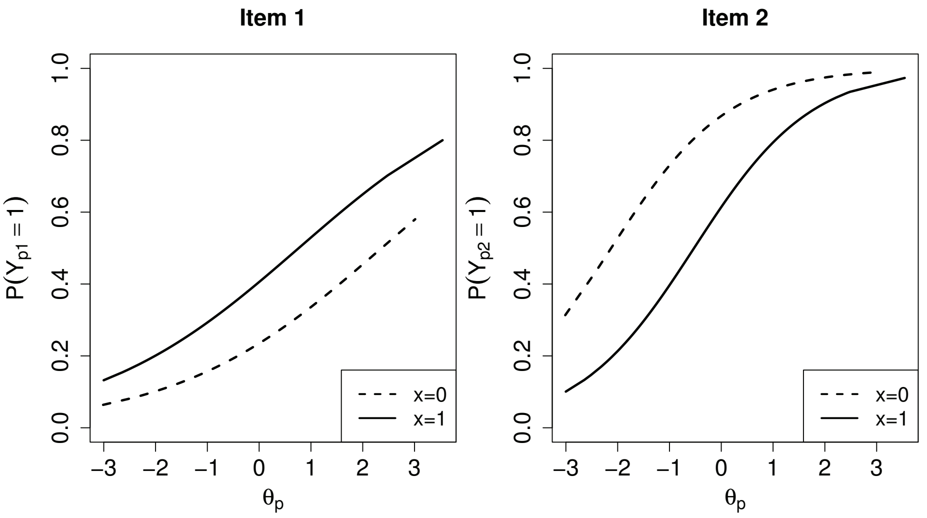


Figure 5.2.: Item Characteristic Curves of item 1 and item 2 for one setting in the simulation with one binary predictor.

5.5.1. One Binary Predictor

First we consider data with two or more groups defined by one covariate. The main objective here is to compare the proposed *IFT* approach to the classical *Logistic* approach, which is well established for the comparison of multiple groups. Later we give detailed results of the proposed *IFT* considering more complex data constellations with several predictors.

We start with one binary covariate $x \in \{0, 1\}$. In this simple case the investigations reduce to the comparison of two groups. Uniform DIF is present if the item difficulties b_i differ between the two groups. The difference is simulated by $b_{i,\text{mod}} = b_i + c \cdot I(x = 0)$ for one half of the DIF items and $b_{i,\text{mod}} = b_i + c \cdot I(x = 1)$ for the other half of the DIF items. The strength of DIF is determined by the constant $c \in \{0.4, 0.8, 1.6\}$. A difference in difficulties of 0.4 is very small, whereas a difference of 1.6 between the two groups is quite large. DIF is generated symmetrically because one half of DIF items favour the first group ($x = 1$) and the other DIF items favour the second group ($x = 0$). For illustration Figure 5.2 shows the Item Characteristic Curves (ICC) of the two items with DIF for the setting with $P = 800$, $I = 20$, 10% DIF items and $c = 1.6$. From the probabilities it can be seen that item 1 is more difficult for $x = 0$ and item 2 is more difficult for $x = 1$. The item locations (value of θ_p with probability 0.5) differ between the two groups but the item discriminations (steepness at the item location) are the same for both groups.

For the comparison of the results we use Receiver Operating Characteristic (ROC) curves, which have also been used by Magis et al. (2015) and Schauberger and Tutz (2015), to evaluate the performance of DIF detection methods. True positive rates and false positive

Table 5.1.: Average FPR on the item level at significance level $\alpha = 0.05$ for the four settings without DIF in the simulation with one binary predictor.

FPR_I	I=20		I=40	
	P=400	P=800	P=400	P=800
IFT	0.050	0.051	0.049	0.050
Logistic	0.052	0.048	0.051	0.050

rates on the item level were computed for increasing significance level $\alpha \in \,]0,1[$. The corresponding ROC curve is then obtained by plotting (FPR_I, TPR_I) as a function of α. Figure 5.3 shows the ROC curves for six out of 24 settings with DIF as the average over 100 repetitions, respectively. The upper panels show the settings with $P = 400$, $I = 40$, 20% DIF and varying DIF strength $c = 1.6$ (solid line), $c = 0.8$ (dashed line) and $c = 0.4$ (dotted line). The lower panels show settings with the same DIF strength $c = 0.8$ and $P = 800$, $I = 20$, 20% DIF (solid line), $P = 800$, $I = 20$, 10% DIF (dashed line) and $P = 400$, $I = 40$, 10% DIF (dotted line). The resulting curves for *IFT* are given in the left panel and the resulting curves for the classical *Logistic* method are given in the right panel. From Figure 5.3 it can be seen that the DIF strength (value of c) and the sample size P have a strong effect on the detection performance, whereas the percentage of DIF items does not have a strong impact.

Although the global performance varies over the different settings, there are only minor differences between the two methods as far as their performance is concerned. All settings we considered, not only the one presented in Figure 5.3, showed nearly no differences between the two methods. A tabular display of the average TPR and FPR at significance level $\alpha = 0.05$ for all settings with DIF are given in Appendix C on page 214. This result is not really surprising. After one split in to the binary predictor x the obtained model (5.5) for one item is exactly the same as model (5.3), which is used for testing when using the classical *Logistic* approach. In this case the only remaining difference is the use of different test statistics to obtain a decision. Nevertheless, the classical and the new approach obviously show the same performance. This is important because the tree based approach, which can also be used in more complex settings with many variables, can also be used in the case of two groups without loss of efficiency.

The construction of ROC curves is an efficient tool but is informative only if DIF is present. Therefore, we separately consider the case without DIF. The average false positive rates with significance level $\alpha = 0.05$ for the four settings without DIF are given in Table 5.1. The absence of DIF is a baseline situation to check a possible inflation of false positive rates. According to the obtained results this is not the case. The *IFT* approach (approxi-

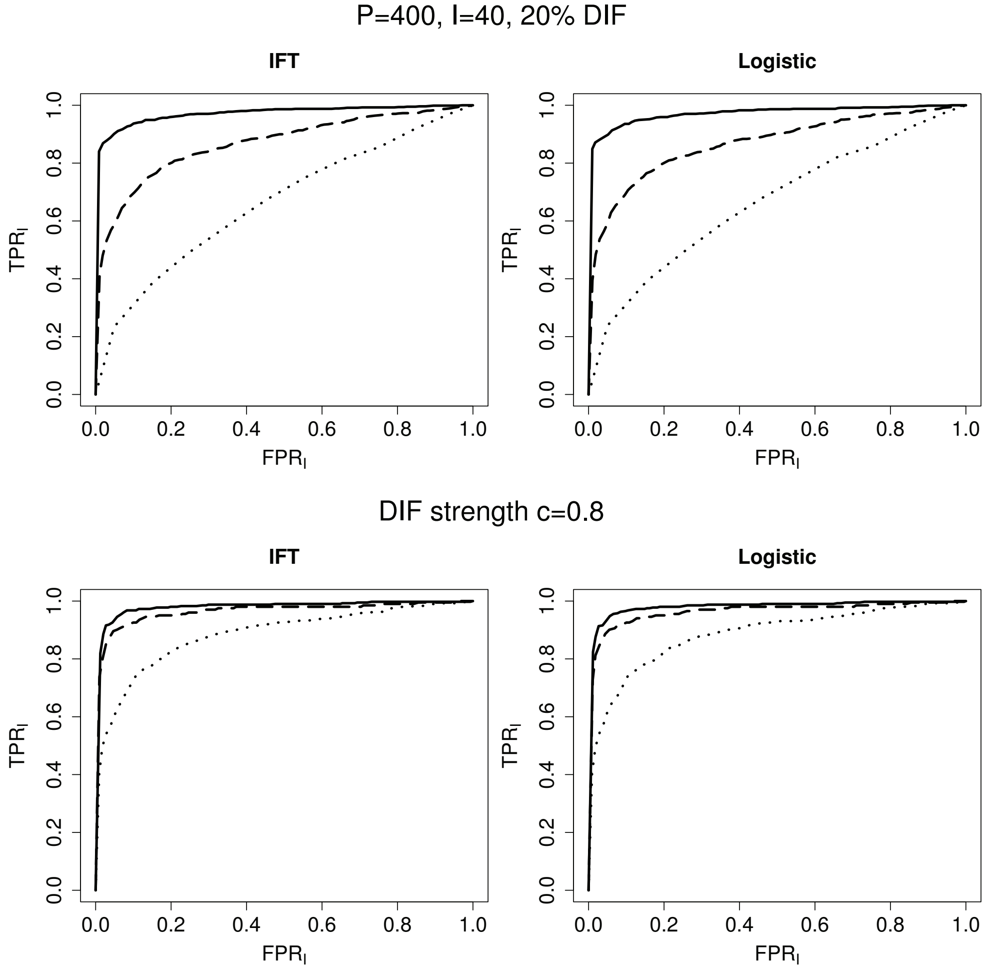


Figure 5.3.: Average ROC curves for six settings in the simulation with one binary predictor. The upper panel shows the curves for three settings with fixed components and varying DIF strength (different line types), the lower panel shows the curves for three settings with the same DIF strength.

mately) holds the significance level as does the classical *Logistic* approach. Again, the two approaches nearly yield the same results.

5.5.2. One Ordered Predictor

Here we consider an ordered factor $x \in \{1, \ldots, 6\}$. The difference in item difficulties is simulated by $b_{i,\text{mod}} = b_i + c \cdot I(x > 3)$ for one half of DIF items and $b_{i,\text{mod}} = b_i + c \cdot I(x \leq 3)$ for the other half of DIF items. Hence there are only two groups that show a true difference, respectively. All the other specifications remain the same as in the previous section 5.5.1. The ROC curves of the six selected examples are given in Figure 5.4. The chosen settings are the same as in Figure 5.3. The left panel now refers to the settings with varying DIF

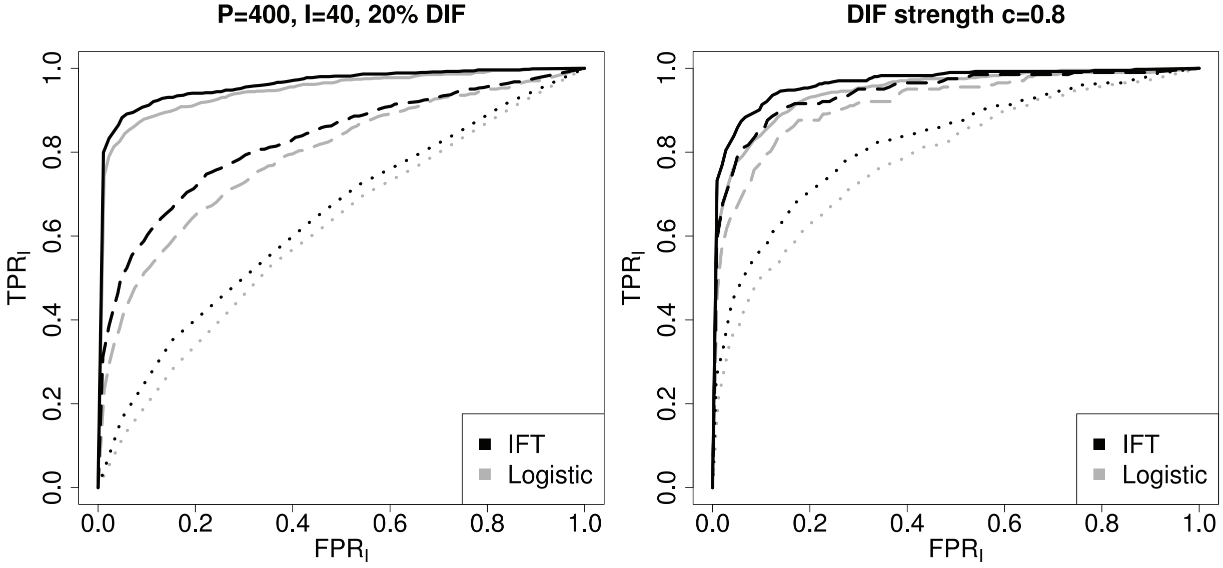


Figure 5.4.: Average ROC curves for six settings in the simulation with one ordered predictor. The left panel shows the curves for three settings with fixed components and varying DIF strength (different line types), the right panel shows the curves for three settings with the same DIF strength.

strength and fixed I, P and percentage of DIF items. The right panel refers to the three settings with constant DIF strength.

In contrast to the comparison of two groups, now there are visible differences between the performances of the two methods. The ROC curves show that *IFT* (black lines) outperforms the classical *Logistic* (grey lines) across the whole range of α. The ROC curves of the new approach are everywhere above the ROC curves of the classical approach. These findings are consistent throughout all settings. The differences are strongest for the settings with medium DIF ($c = 0.8$). An overview of the average TPR and FPR at significance level $\alpha = 0.05$ for all settings with DIF are given in Appendix C on page 215.

The reason for the better performance of *IFT* is that it is able to use the ordering of the categories. Since DIF is linked to the ordinal scale of the factor a method that is able to exploit the ordering should perform better than the classical method that just distinguishes between the groups. It is noteworthy that in Figure 5.4 the performance of the settings with a large number of persons and medium DIF strength (solid and dashed line in the right panel) is fairly similar to the performance with a small number of persons and strong DIF (solid line in the left panel). This underlines that an increase of sample size strongly contributes to improve the detection performance.

Table 5.2.: Average FPR at significance level $\alpha = 0.05$ for the four settings without DIF in the simulation with three covariates.

		I=20		I=40	
		P=400	P=800	P=400	P=800
IFT	FPR_I	0.027	0.021	0.024	0.022
	FPR_{IV}	0.009	0.007	0.008	0.007

5.5.3. Several Predictors

In the following simulations we consider three covariates, two binary variables $x1$, $x2 \sim B(1, 0.5)$ and one standard normal distributed variable $x3 \sim N(0, 1)$. Since IFT allows to determine the variables that are responsible for DIF, true positive and false positive rates for the combination of item and variable can be computed. In the following all the presented results are based on computations with significance level $\alpha = 0.05$. To account for the three covariates in the model the local significance level for one permutation test is 0.05/3.

Before simulating items with DIF we first investigate the baseline situation without DIF. The average false positive rates for the four settings (varying number of persons and items) without DIF are given in Table 5.2. It is seen that IFT yields small false positive rates. The procedure is somewhat conservative and does not fully use the specified significance level. On average only one item is misleadingly identified as DIF item. False positive rates for the combination of item and variable are much smaller. With 40 items the value 0.008 means that only one split with regard to a variable that was not inducing DIF was falsely executed during estimation.

DIF in the First Variable

In the settings with DIF, first DIF is simulated as in the simulation with one binary predictor only (Section 5.5.1). If DIF is present, the item difficulties b_i differ between the two groups defined by the binary covariate $x1$. Hence the underlying true model is defined by one split in $x1$. Boxplots of true positive and false positive rates of the 24 settings with DIF are given in Figure 5.5. The results on the item level are in light grey and are given on the left of each panel, the results for the combination of item and variable are in dark grey and are given on the right of each panel. In addition, the significance level $\alpha = 0.05$ is marked as a reference by dashed lines. It is seen from Figure 5.5 that IFT shows good overall performance for medium and strong DIF if the number of persons is large. For small DIF effects the number of persons definitely has to be large. True positive rates are high in the settings with

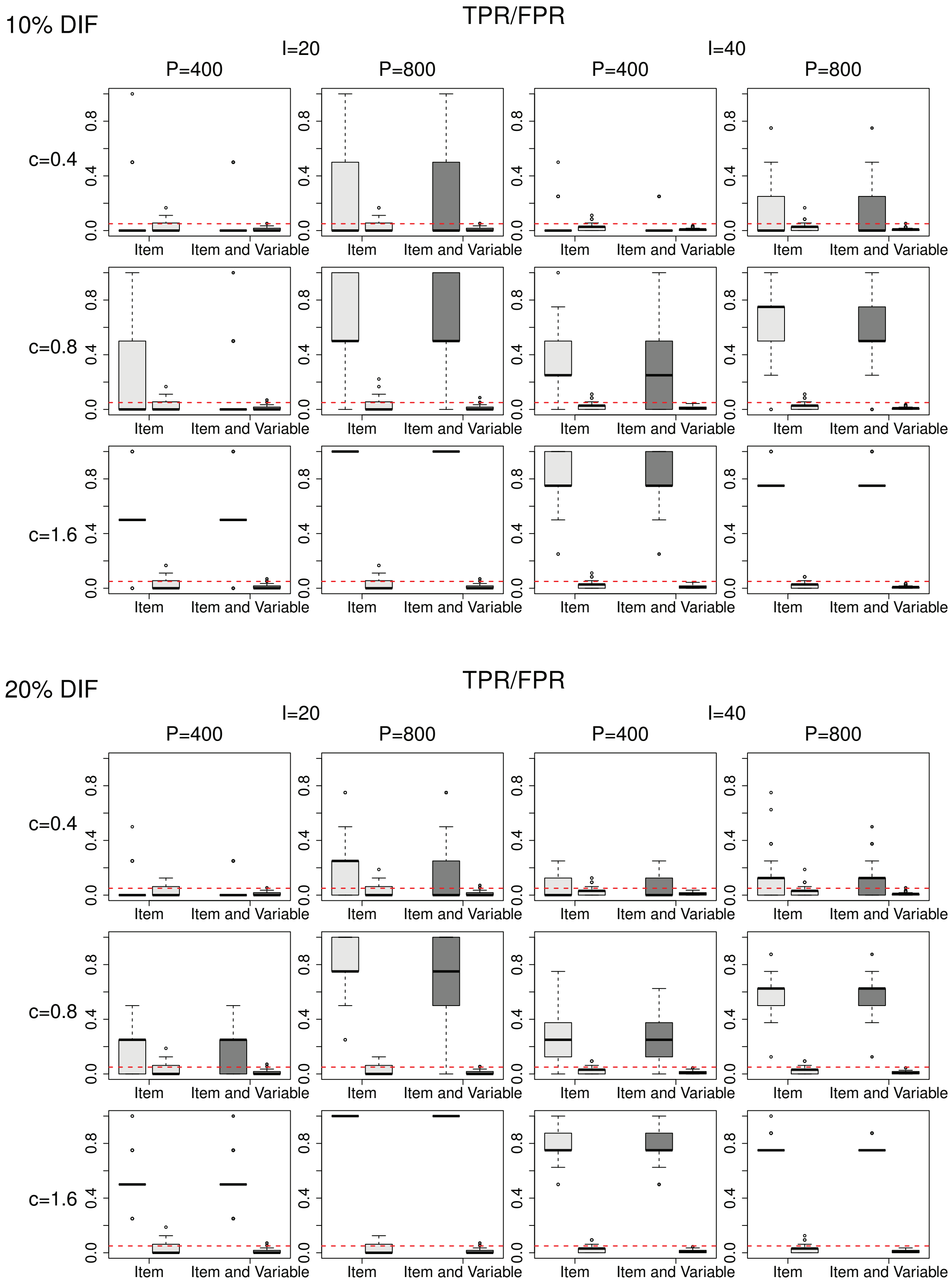


Figure 5.5.: Boxplots of TPR and FPR at significance level $\alpha = 0.05$ (marked by dashed lines) in the simulation with three covariates and DIF in $x1$. Results on item level are given in light grey, results for the combination of item and variable are given in dark grey.

$P = 800$ and $c = 1.6$. Here a clear separation between DIF and non-DIF items is seen. For the setting in the lower left of Figure 5.5 with $P = 400$, $I = 40$, 20% DIF items and $c = 1.6$ one observes a TPR of 0.5 in 68 of the 100 data sets and therefore the box reduces to one value. In the settings with small DIF ($c = 0.4$) and a small number of persons ($P = 400$) the method is hardly able to detect the corresponding items, however, as is seen from Figure 5.4 also alternative methods show poor performance if DIF is weak. False positive rates are very small throughout all settings, in particular the global significance level holds (with a tendency of the method to be conservative). It is noteworthy that the true positive rates for the combination of item and variable in all settings are very similar to the true positive rates for items. Therefore, *IFT* is able to simultaneously identify the items and variables that are responsible for DIF. Similar pictures resulted if the covariates x_1, x_2 and x_3 were correlated with medium sized correlation $\rho = 0.6$. In analogy to Figure 5.5 the results for the simulations with correlation are shown in Appendix C in Figure C.1 on page 216. It should be noted that in classical approaches for fixed groups the simultaneous detection of DIF item and responsible variable is not investigated. If one considers more than one categorical variable, for example, gender and race, typically DIF induced by gender and race are investigated separately with significance levels fixed to the same value separately for the two investigations. However it should be mentioned that in the *extended Logistic* model one could also investigate the effect of both variables by including both variables, and possibly an interaction term, in the linear predictor.

DIF in Two Covariates

In the following we consider again the complex DIF structure considered in the illustrative example and use two DIF items. In item 1 DIF is induced by $x1$ and $x3$ and determined by the step functions $b_{1,\text{mod}} = b_1 + c \cdot I(x_3 > 0) + c \cdot I(\{x_3 > 0\} \cap \{x_1 = 0\})$, in item 2 DIF is induced by $x2$ and $x3$ and we use the step functions $b_{2,\text{mod}} = b_2 + c \cdot I(x_3 > 0) + c \cdot I(\{x_3 > 0\} \cap \{x_2 = 0\})$. The strength of DIF again is determined by the additional parameter $c \in \{0.4, 0.8, 1.6\}$. By choosing these values for c the differences between the individual groups remain the same as in the previous simulations.

In the same way as in Figure 5.5, the true positive rates and false positive rates of the twelve settings (with varying I, P and c) based on 100 replications are given in Figure 5.6. The true positive rates on the item level (given in light grey) are very high for all settings with $c = 0.8$ and $c = 1.6$. Especially for the settings with $P = 800$ the selection of items is quite perfect. However, for small DIF ($c = 0.4$, first row) the detection of responsible items remains quite challenging. It is also seen that the hit rates for the combination of item and variable (given in dark grey) are not so much smaller than the hit rates for items. Since here DIF is generated by two variables *IFT* cannot detect both variables in all the cases. However, the small false positive rates show that the procedure does not tend to perform

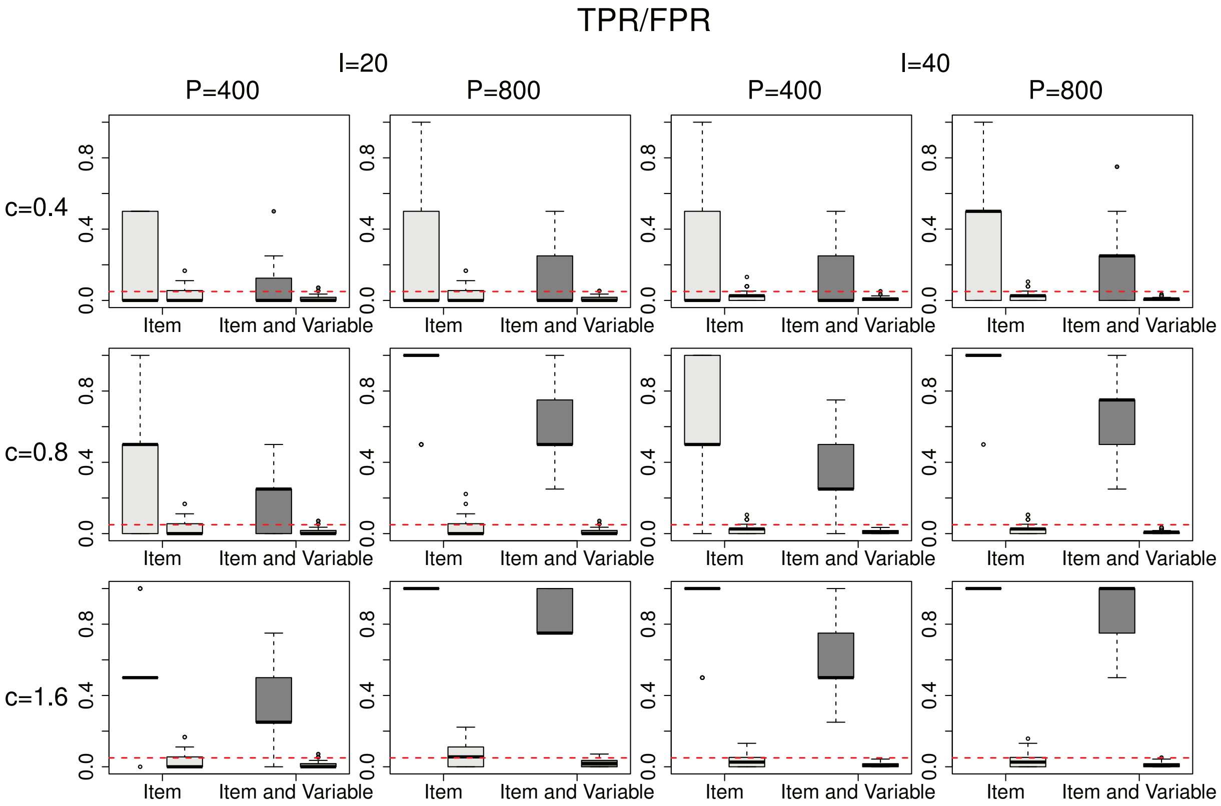


Figure 5.6.: Boxplots of TPR and FPR at significance level $\alpha = 0.05$ (marked by dashed lines) in the simulation with three covariates and DIF in two items and two covariates. Results on item level are given in light grey, results for the combination of item and variable are given in dark grey.

splits with regard to variables that are not responsible for DIF. If a significant effect is found the corresponding split is always in the right variable.

5.6. Investigation of Non-Uniform DIF

A strength of the logistic framework for DIF detection proposed by Swaminathan and Rogers (1990) is that it can be extended to detect non-uniform DIF. We first consider the classical and extended approach and then item focussed trees.

5.6.1. Logistic Regression for Non-Uniform DIF

Let us again first consider the comparison of multiple groups. To account for non-uniform DIF model (5.1) has to be extended by group-specific slopes and has the form

$$\eta_{pi} = \beta_{0i} + S_p\beta_i + \gamma_{ig} + S_p\alpha_{ig}, \tag{5.7}$$

where α_{ig} are the additional group-specific slopes. The first group is chosen as reference group by setting $\gamma_{i1} = \alpha_{i1} = 0$, see, for example, Magis et al. (2011). The model can be extended to account for non-uniform DIF that is generated by a vector of covariates in a similar way as for uniform DIF. Then one uses the model

$$\eta_{pi} = \beta_{0i} + S_p\beta_i + \boldsymbol{x}_p^\top\boldsymbol{\gamma}_i + S_p\boldsymbol{x}_p^\top\boldsymbol{\alpha}_i, \tag{5.8}$$

which contains an interaction between the person characteristics $\boldsymbol{x}_p$ and the test score S_p. The new slope parameters in model (5.8) are contained in $S_p(\beta_i + \boldsymbol{x}_p^\top\boldsymbol{\alpha}_i)$. Model (5.8) reduces to the logistic model used in Section 5.2.2 if $\boldsymbol{\alpha}_i = \boldsymbol{0}$. Thus uniform DIF is present if $\boldsymbol{\gamma}_i \neq \boldsymbol{0}$ given $\boldsymbol{\alpha}_i = \boldsymbol{0}$. However, the item shows non-uniform DIF if $\boldsymbol{\alpha}_i \neq \boldsymbol{0}$ whether $\boldsymbol{\gamma}_i = \boldsymbol{0}$ or not.

5.6.2. Logistic Regression Trees for Non-Uniform DIF

When using item focussed trees, non-uniform DIF means that splits are not only admissible in the variables $x_{p1}, \ldots, x_{pm}$, but also in the interaction terms $S_px_{p1}, \ldots, S_px_{pm}$. A (first) split with regard to the interaction between the test score and the j-th variable yields the model with predictor

$$\eta_{pi} = \beta_{0i} + S_p\,[\alpha_{il}^{[1]}I(x_{pj} \le c_j) + \alpha_{ir}^{[1]}I(x_{pj} > c_j)],$$

where the parameter $\alpha_{il}^{[1]}$ denotes the slope in the left node ($x_{pj} \le c_j$) and $\alpha_{ir}^{[1]}$ denotes the slope in the right node ($x_{pj} > c_j$).

5.6.3. Test Strategies

In the literature different strategies were proposed how to test for the significance of DIF by means of model (5.7), see, for example, Zumbo (1999) and Magis et al. (2011). We will use similar strategies when testing for DIF in the extended logistic regression model (5.8) and the tree-based approach.

Testing for DIF

The first strategy is to test for both types of DIF effects simultaneously. The corresponding null hypothesis given model (5.7) is $H_0 : \gamma_{i2} = \ldots = \gamma_{iG} = \alpha_{i2} = \ldots = \alpha_{iG} = 0$. For model (5.8) the corresponding null hypothesis is given by $H_0 : \boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i = 0$. That means DIF is investigated by using a global test for the whole parameter vector $(\boldsymbol{\gamma}_i, \boldsymbol{\alpha}_i)$. DIF is considered as being present (in any form) if the test rejects the null hypothesis, meaning that at least one of the parameters $\gamma_{ij},\ \alpha_{ij},\ j = 1, \ldots, m$, differs from zero.

For item focussed trees the equivalent is that at least one split is performed in one of the components. When selecting the optimal split in each step of the algorithm, one has to consider all combinations of item, variable, split-point and component with regard to intercept and slope. The final model consists of one or two separate trees, one referring to the intercept and one referring to the slope. In general the trees will be different but can also have the same structure. The resulting tree is given by

$$\eta_{pi} = tr_i(\boldsymbol{x}_p) + tr_i(S_p, \boldsymbol{x}_p), \tag{5.9}$$

where $tr_i(\boldsymbol{x}_p)$ is the tree component containing subgroup-specific intercepts and $tr_i(S_p, \boldsymbol{x}_p)$ is the tree component containing subgroup-specific slopes. In contrast to the tree in model (5.5) for uniform DIF now one has two possible trees. If there is only a significant effect in one of the two components a constant $tr_i(\boldsymbol{x}_p) = \beta_{0i}$ or $tr_i(S_p, \boldsymbol{x}_p) = S_p\beta_i$ is fitted in the other component.

In comparison to the classical and extended *Logistic* method, the tree-based model has two advantages:

- The obtained tree(s) distinguish between items with uniform and non-uniform DIF. The trees themselves show which form of DIF is present. Thus both types of DIF can be detected simultaneously within one fitting procedure.
- The obtained tree(s) identify the variables that induce uniform and/or non-uniform DIF. In particular, both types of DIF can be caused by different variables.

Testing for Non-Uniform DIF

A second strategy is to explicitly test for non-uniform DIF. Using the extended Logistic model (5.8) one investigates the null hypothesis $H_0 : \boldsymbol{\alpha}_i = \mathbf{0}$ for each item. Non-uniform DIF is considered as being present if the hypothesis is rejected, meaning that at least one parameter α_{ij} differs from zero.

Table 5.3.: Modified values of item discrimination and item difficulty parameters in the illustrative example with non-uniform DIF.

Item	Non-Uniform DIF	Item	Uniform DIF
1	$a_{1,\mathsf{mod}} = a_1 + 0.6 \cdot I(x_1 = 1)$	3	$b_{3,\mathsf{mod}} = b_3 + 0.8 \cdot I(x_1 = 1)$
2	$a_{2,\mathsf{mod}} = a_2 + 0.6 \cdot I(x_2 = 0)$	4	$b_{4,\mathsf{mod}} = b_4 + 0.8 \cdot I(x_2 = 0)$

For item focussed trees the detection of non-uniform DIF means that a significant split in the *slope* component is found. Consequently, during estimation only the models with *simultaneous splits* in the intercepts and the slopes are considered as potential candidates. Therefore, one split in item i with regard to variable j corresponds to the model with predictor

$$\eta_{pi} = [\gamma_{il}^{[1]} I(x_{pj} \le c_j) + \gamma_{ir}^{[1]} I(x_{pj} > c_j)] + S_p\, [\alpha_{il}^{[1]} I(x_{pj} \le c_j) + \alpha_{ir}^{[1]} I(x_{pj} > c_j)], \qquad (5.10)$$

which contains two intercepts $(\gamma_{il}^{[1]}, \gamma_{ir}^{[1]})$ and two slopes $(\alpha_{il}^{[1]}, \alpha_{ir}^{[1]})$ with respect to the same subgroups. To select the optimal split and to determine the splitting decision one compares the likelihoods of model (5.4) and (5.10). The procedure is continued in each step of the algorithm, considering all combinations of item, variable and split-point.

If non-uniform DIF is present, the final model consists of two trees containing subgroup-specific intercepts and subgroup specific slopes that are determined by the same splits.

For the different strategies we will use the same terminology as Magis et al. (2011) in his investigation of the case in which DIF is induced by multiple groups:

- *UDIF* means testing for uniform DIF, $H_0 : \boldsymbol{\gamma}_i = 0$, given model (5.3) within the logistic regression approach. For trees it refers to testing the corresponding splits.
- *DIF* means simultaneous tests for uniform and non-uniform DIF, $H_0 : \boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i = 0$, given model (5.8) for logistic regression. For trees it refers to testing the corresponding splits for both types of DIF.
- *NUDIF* means tests for non-uniform DIF, $H_0 : \boldsymbol{\alpha}_i = \mathbf{0}$, given model (5.8) for logistic regression. For trees it refers to testing the corresponding splits.

5.6.4. Illustrative Example

As in section 5.3 we consider data $Y_{pi},\ p = 1, \dots, 800,\ i = 1, \dots, 20$, that are generated by a 2PL-model with DIF. As before the item discrimination parameters a_i are first drawn

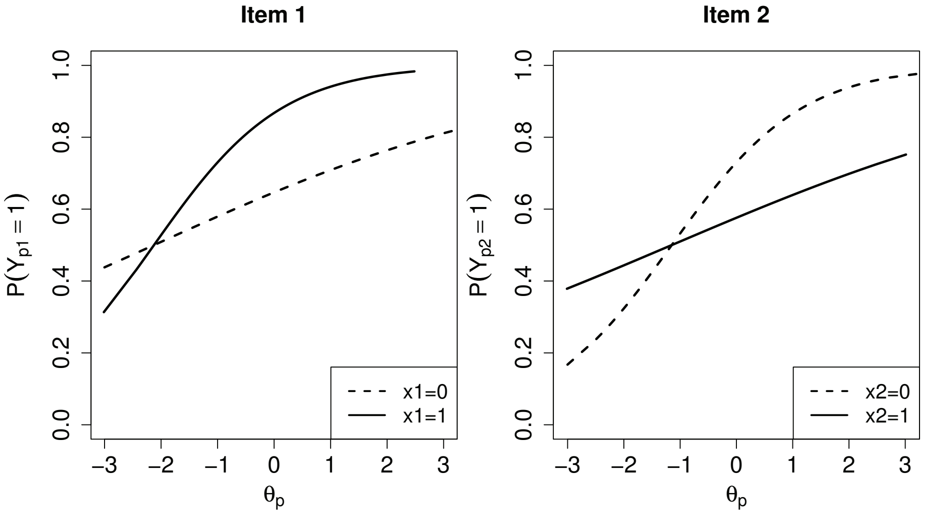


Figure 5.7.: Item Characteristic Curves of item 1 and item 2 for the the illustrative example with non-uniform DIF.

from a uniform distribution. However, in order to simulate non-uniform DIF we do not generate data from the 2PL-model but assume that the item discrimination parameters depend on covariates. The same strategy for generating non-uniform DIF was also used by Rogers and Swaminathan (1993), Narayanan and Swaminathan (1996) or Jodoin and Gierl (2001).

Again, we consider 100 data sets with three covariates, two binary variables $x1$, $x2 \sim B(1, 0.5)$ and one standard normal distributed variable $x3 \sim N(0, 1)$. We simulate data where two of the 20 items show non-uniform DIF and two of the 20 items only show uniform DIF. The modified values of the discrimination and difficulty parameters are determined by step function given in Table 5.3. In item 1 and 3 DIF is induced by $x1$ and in item 2 and 4 DIF is induced by $x2$. Hence, in all four cases two groups have to be distinguished. The resulting ICC of the two items with non-uniform DIF (item 1 and 2) are given in Figure 5.7 separately for the two groups. It can be seen from the curves that the item locations are equal for both groups but the item discriminations (as it was simulated) differ between the groups. When fitting *IFT* the non-uniform DIF structure is detected correctly if there is one split in the slope component of the model of item 1 in $x1$ and item 2 in $x2$.

DIF

Figure 5.8 shows one exemplary estimation result obtained by *IFT* when testing for both types of DIF simultaneously. In this example items 1, 2, 3, and 4 are correctly identified as DIF items. All items are split once yielding trees with two terminal nodes, respectively.

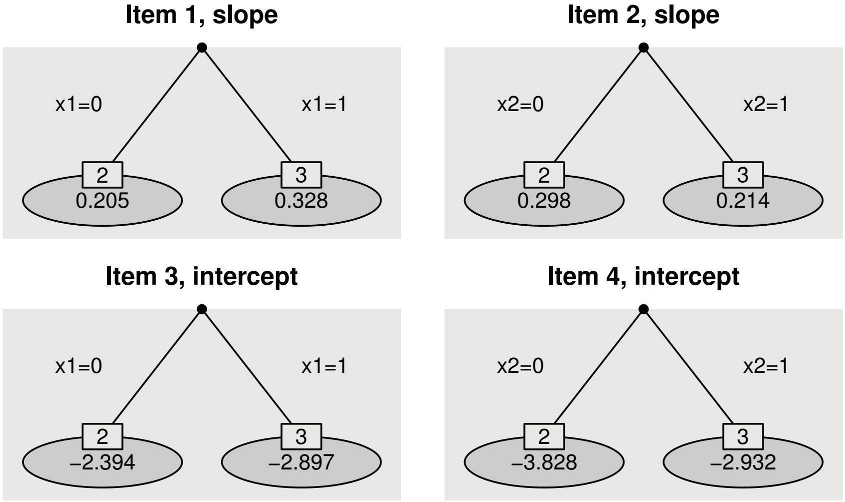


Figure 5.8.: Estimated trees for the illustrative example with non-uniform DIF, testing for both types of DIF. Estimated coefficients $\alpha_{i\ell}$ (upper) and $\gamma_{i\ell}$ (lower) are given in each leaf of the trees.

Items 1 and 2 (upper panel) are split with regard to the slopes indicating non-uniform DIF. In item 1 the (simulated) item discrimination is higher for $\{x1 = 1\}$, yielding a higher slope for the corresponding subgroup ($\hat{\alpha}_{1,x1=1} = 0.328$). Whereas, in item 2 the item discrimination is larger for $\{x2 = 0\}$, which results in a larger slope for this subgroup ($\hat{\alpha}_{2,x2=0} = 0.298$). In items 3 and 4 (lower panel) one split is performed with regard to the intercepts indicating uniform DIF. The results are also in line with the true simulated effects. The model provides an identification of DIF items together with the responsible covariates and a classification by type of DIF.

Non-Uniform DIF

When using *IFT*, which explicitly tests for non-uniform DIF, only items 1 and 2, that were simulated as non-uniform DIF items, are detected. The corresponding trees are given in Figure 5.9. The subgroup-specific slopes (left panel) are defined by the same splits as in the *DIF* framework considered previously. Due to the construction of the model the estimated coefficients $\alpha_{i1}, \alpha_{i2}, i = 1, 2$, however, differ slightly. If splits are significant the same splits are performed in the intercepts yielding trees with subgroup-specific intercepts. Since they are not of main interest they are displayed a little smaller (right panel of Figure 5.9).

5.6.5. Simulations

In the following we briefly illustrate the properties of the models for the *DIF* and *NUDIF* framework by means of a small simulation. The structure of the simulated datasets we

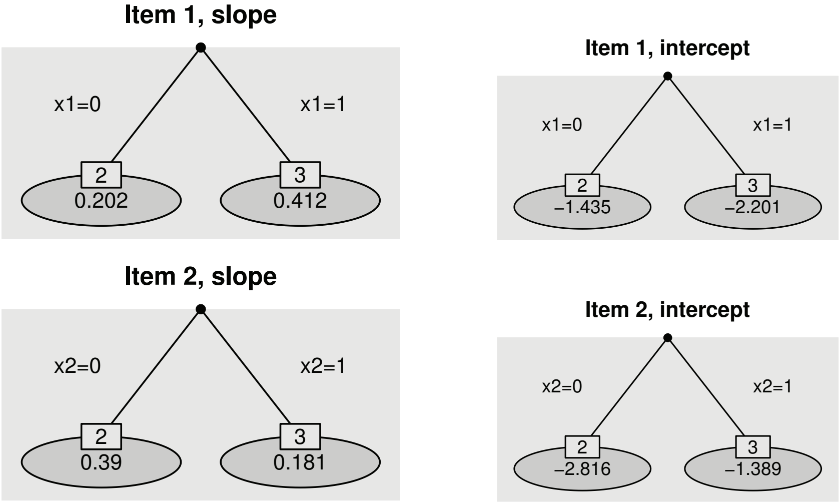


Figure 5.9.: Estimated trees for the illustrative example with non-uniform DIF, testing for non-uniform DIF. Estimated coefficients $\alpha_{i\ell}$ (left) and $\gamma_{i\ell}$ (right) are given in each leaf of the trees.

consider here is the same as in section 5.5. We limit the discussion to the comparison of two groups defined by one binary covariate $x \in \{0, 1\}$. According to model (5.6) non-uniform DIF is present if the item discriminations a_i differ between the two groups. The difference in item discriminations is simulated by the equation $a_{i,\text{mod}} = a_i + c \cdot I(x = 0)$ for one half of DIF items and by the equation $a_{i,\text{mod}} = a_i + c \cdot I(x = 1)$ for the other half of DIF items, with constant $c \in \{0.3, 0.6\}$. From our experience the values 0.3 and 0.6 represent medium DIF effect sizes. Boxplots of true positive and false positive rates on the item level for the setting with $P = 800$, $I = 20$ and 20% DIF obtained by *IFT* (left of each panel) and the classical *Logistic* model (right of each panel) are given in Figure 5.10. The results when testing for both types of DIF are shown in the left panel and the results when testing for non-uniform DIF are shown in the right panel. Within the *DIF* framework the classical *Logistic* model outperforms the proposed tree-based approach. The average hit rate in the setting with $c = 0.6$ (lower left) is 0.66 for *Logistic* but only 0.43 for *IFT*. This was to be expected because the test on the whole parameter vector (γ_i, α_i) obviously has a stronger power than the tests on single splits. However, in the *NUDIF* framework the two methods almost yield the same results. The average hit rate in the settings with $c = 0.6$ (lower right) for both models is 0.44. Due to the construction of the models the main difference in the case of two groups is the use of different test statistics to obtain a decision. As we already illustrated for uniform DIF, our proposed item focussed trees approach can also be used to detect non-uniform DIF without loss of efficiency. The findings presented here can be confirmed by the results of all other settings considered in our simulation. For details a tabular display of the average TPR and FPR for all settings with non-uniform DIF are given in Appendix C on page 217.

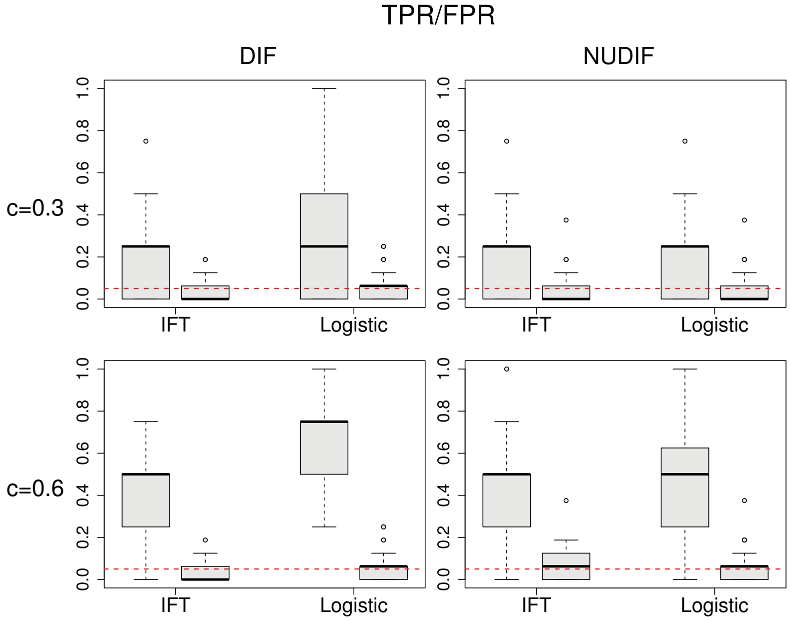


Figure 5.10.: Boxplots of TPR and FPR for the simulation with non-uniform DIF and one binary predictor ($P = 800$, $I = 20$, 20% DIF), testing for both types of DIF (left) and testing for non-uniform DIF (right).

Table 5.4.: Summary statistics of the test score of the second module (items 21 to 40) of the I-S-T 2000 R and the two considered covariates.

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	x_{max}
Test score	6	12	14	13.87	16	19
Age	18	20	22	22.88	24	39
Gender	male: 97			female: 176		

5.7. Empirical Applications

Finally we will illustrate and compare the proposed approaches on real data examples.

5.7.1. I-S-T 2000 R

We use data from the Intelligence-Structure-Test 2000 R (I-S-T 2000 R; source of supply is Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0049-551) 999-50-999, www.testzentrale.de). The test was developed by Amthauer et al. (2001); Beauducel

Table 5.5.: Comparison of detected DIF items of the second module of the I-S-T 2000 R using IFT and the extended Logistic approach for uniform and non-uniform DIF.

	Item focussed Trees			Extended Logistic		
Item	UDIF	DIF	NUDIF	UDIF	DIF	NUDIF
First	×	× (u)		×	×	
Second	×	× (u)		×	×	
Third	×	× (non)		×	×	
Fourth				×		
Fifth				×		

et al. (2010) and is a revised version of its predecessors I-S-T 70 (Amthauer et al., 1973) and I-S-T 2000 (Amthauer et al., 1999). The available study was conducted at the Phillips University in Marburg (Bühner et al., 2006). There were 273 participants from 40 different subject areas. The first module of the test was already analyzed in an application in Chapter 4. The second module contains 20 items (items 21 to 40) in which analogies play the major role. There are three predefined terms with a certain relation between the first two. This relationship needs to be recognized to find the fourth term. From five possible answers the respondent is asked to choose the term that relates to the third term as the second term relates to the first term. One example is

dark:bright = wet:?

a) rain b) day c) moist d) wind e) dry.

Therefore, one has to select that alternative that relates to wet as bright relates to dark.

For the investigation of DIF in these items we incorporate the covariates gender (male: 0, female: 1) and age. The summary statistics of the resulting test scores of items 21 to 40 and the two covariates are given in Table 5.4.

When using *IFT* for uniform DIF 3 out of 20 items show DIF. The algorithm performs only three splits before stopping and, therefore, each item is split only once. All permutation tests were based on 1000 permutations at local significance level 0.05/2.

The estimated trees for three items detected as DIF items are given in Figure 5.11. It is seen that both covariates gender and age seem to induce DIF because both are used for splitting at least once. The second and third item show DIF induced by gender, whereas the first item shows DIF induced by age. According to the estimated coefficients the second item is easier for females (gender=1), the third item is easier for males (gender=0) and the first item is easier for all students who are rather young (age$\leq$23).

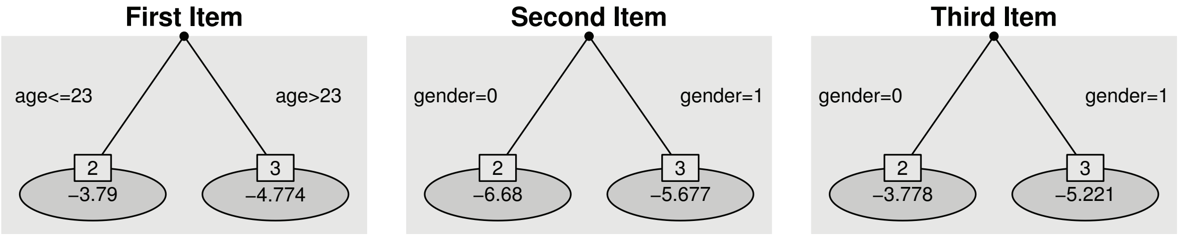


Figure 5.11.: Trees of the three detected DIF items of the second module of the I-S-T 2000 R using the model for uniform DIF. Estimated intercepts γ_{il} are given in each leaf of the trees.

An overview of the detected DIF items obtained by the six strategies discussed in this chapter is given in Table 5.5. When using *IFT* which tests for both types of DIF, one obtains very similar results. As in the *UDIF* framework the first, second and third item are also identified as DIF items with the same variables that induce DIF. The estimated models for the first and second item are even identical. A difference occurs for the third item, where the split in gender is not performed in the intercept but in the slope component. The model gives the estimated intercept $\beta_{0,Third} = -4.993$. The resulting tree of slopes α_{il} is given in Figure 5.12. The estimated coefficients again mean that the item favours males (gender=0) but the difference slightly increases for participants with a higher test score. Interestingly, the splits in the intercept (*UDIF*, Figure 5.11) and in the slope (*DIF*, Figure 5.12) result in very similar estimated probabilities. As a consequence it is not surprising that the third item is not detected by the model within the *NUDIF* framework.

The evaluation of the data set by the extended *Logistic* model (5.3) for uniform DIF yields five DIF items (fourth column in Table 5.5). Based on the results in the simulations, it seems that the fourth and fifth item might be falsely identified as items with uniform DIF. Concerning the identification of items, the results within the *DIF* and *NUDIF* framework are equal to those of *IFT*. However, when testing non-uniform DIF for the third item one obtains the p-value 0.052 indicating an almost significant effect. Table 5.6 shows an detailed overview of the estimated DIF effect sizes when using the two approaches for uniform DIF. For IFT (left columns) the given values correspond to the (norm of the) differences of the estimated values in the nodes of the trees in Figure 5.11. For the third item one observes the difference 1.443 which is quite large. The extended *Logistic* approach does not explicitly provide information about the variables that are responsible for DIF but the estimates and corresponding standard errors given in Table 5.6 indicate which ones might be relevant.

It is noteworthy that in summary the test seems not to be strongly affected by DIF. From the 20 items that use analogies only three are suspect of DIF and the effects are not overly strong. This was to be expected of a carefully designed test.

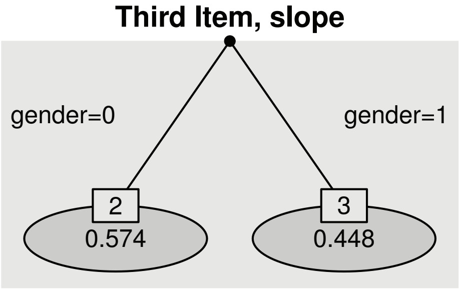


Figure 5.12.: Tree of the third detected DIF item of the second module of the I-S-T 2000 R using the model for both types of DIF. Estimated slopes α_{il} are given in each leaf of the trees.

Table 5.6.: Overview on estimated effect sizes of the second module of the I-S-T 2000 R using IFT and the extended Logistic approach for uniform DIF. For IFT the differences of the effects in the nodes are given, for the Logistic approach the estimates and standard errors are given.

	Item focussed Trees		**Extended Logistic**	
Item	Age	Gender	Age	Gender
First	0.984	×	-0.943 (0.152)	-0.026 (0.154)
Second	×	1.002	0.091 (0.165)	0.507 (0.174)
Third	×	1.443	0.485 (0.212)	-0.583 (0.225)
Fourth	×	×	0.175 (0.200)	-0.455 (0.237)
Fifth	×	×	0.088 (0.133)	0.367 (0.138)

5.7.2. CTB Science Data

In a second application we consider a data set from CTB-McGraw Hill, which was already analyzed in the illustrative example in Chapter 3. For a description of the original data, see also De Boeck and Wilson (2004). The data includes the results of 1500 grade 8 students from 35 schools. The students had to respond to 76 items, measuring different objectives and subskills related to mathematics and science. For the present investigation we restrict to the 25 multiple-choice items from subject area science.

To test for DIF in these items we incorporate the three covariates gender (male: 0, female: 1), type of the school (1: catholic, 2: private, 3: public) and size of the school (number of students in hundreds). The summary statistics of the test scores for the 25 items and the three covariates are given in Table 5.7.

When fitting *IFT* for uniform DIF 14 of 25 items are identified as DIF items. Altogether the algorithm performs 27 splits until further splits are no longer significant. With three covariates, each permutation test is performed at local significance level 0.05/3. The p-value in the 28-th iteration was 0.02 and thus not significant on level $0.01\overline{6}$. All splits refer to

Table 5.7.: Summary statistics of the test score of the 25 multiple-choice items from subject area science of the CTB data and the three considered covariates.

Variable	Summary statistics					
	x_{min}	$x_{0.25}$	x_{med}	$\bar{x}$	$x_{0.75}$	x_{max}
Test score	7	14	16	16.01	18	23
Size	100	500	900	868.3	1300	1600
Type	catholic: 105		private: 84		public: 1311	
Gender	male: 761			female: 739		

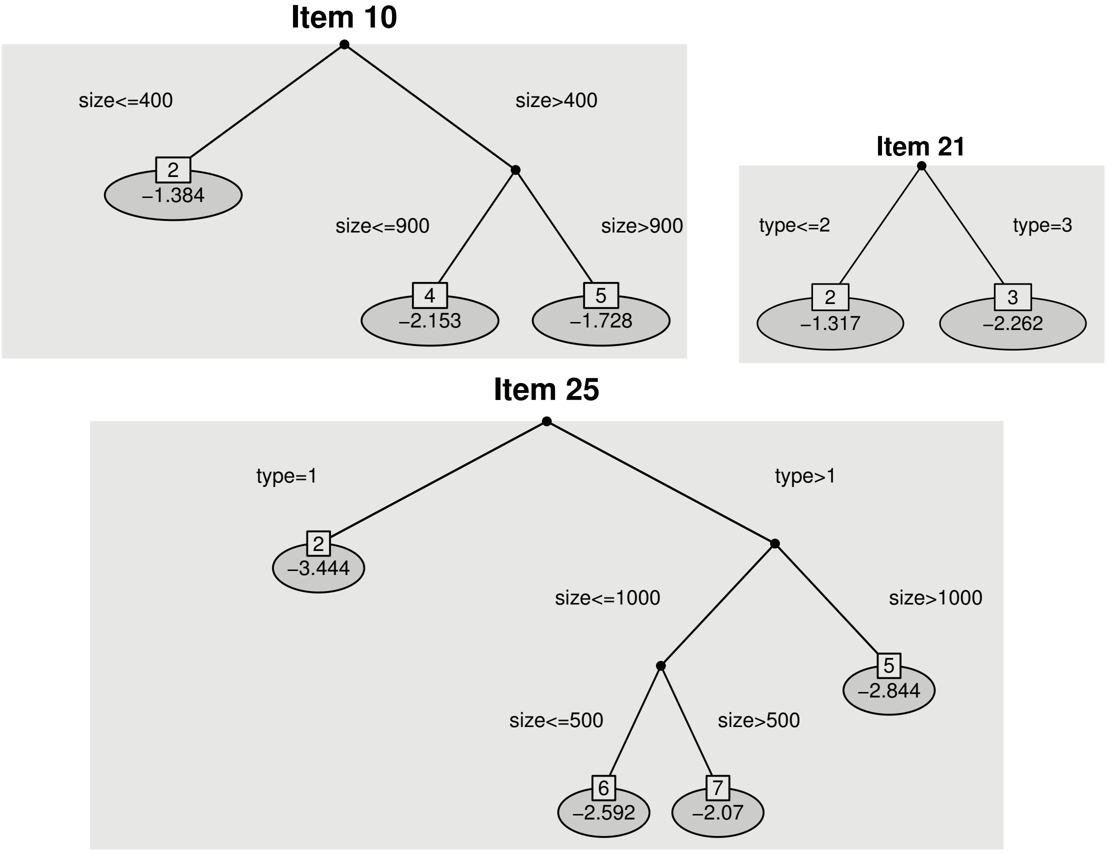


Figure 5.13.: Trees of items 10, 21 and 25 of the CTB data using the model for uniform DIF. Estimated intercepts γ_{il} are given in each leaf of the trees.

covariates type and size, whereas no significant splits were found for variable gender. There does not seem to be any difference between males and females.

The trees for three selected items are given in Figure 5.13. In item 10 DIF is induced by size and one has to distinguish between three subgroups. The item is easiest for students in small schools (size$\leq$400) but most difficult for students in medium-sized schools (400<size$\leq$900).

Table 5.8.: Comparison of detected DIF items of the CTB data using IFT and the extended Logistic approach for uniform and non-uniform DIF.

	Item focussed Trees				**Extended Logistic**		
Item	UDIF	DIF		NUDIF	UDIF	DIF	NUDIF
21	×	×	(non)	×	×	×	×
3	×	×	(u)		×	×	
4	×	×	(u)		×	×	
8	×	×	(u)		×	×	
9	×	×	(u)		×	×	
14	×	×	(non)		×	×	
16	×	×	(non)		×	×	
25	×	×	(u)		×	×	
11					×	×	×
13				×		×	×
19	×	×	(u)		×		
5	×	×	(u)				
10	×	×	(u)				
24					×	×	
1							×
6	×						
15	×						
17	×						

Item 21 is easier for students in a catholic or private school (type≤2) compared to students in public schools (type=3). An interesting partition is received for item 25. For all students in a catholic school (type=1) the question is very difficult. By contrast the question is easier for all students in a private or public school (type>1), in particular for those in medium-sized schools (500<size≤1000).

To obtain DIF effect sizes we computed the maximal difference of estimated effects between any two nodes for each tree. The obtained values vary over a wide range from 0.458 to 2.985. This also confirms that large DIF effects such as 1.6 might occur in real data sets.

An overview of the detected DIF items by the six evaluated models is given in Table 5.8. It shows only items that were found to be DIF items by at least one of the models. Within the *DIF* framework (second column) eleven DIF items are identified. These are the same items as with the restricted model for uniform DIF discussed above, but without item 6, 15 and 17. Unlike above, there are three items that are classified as non-uniform DIF items by the more general model. Here, for example in item 21 the split regarding the type of school is not performed in the intercept but in the slope component. According to the model testing for non-uniform DIF (third column) the two items 13 and 21 carry non-uniform DIF. In contrast to item 13, item 21 is also detected within the *UDIF* and *DIF* framework.

The comparison to the extended *Logistic* approach shows a strong overlap. Within the *UDIF* framework (first and fourth column) there is a agreement in nine items. In the *DIF* framework this is the case for eight items. However it should again be mentioned, that the extended *Logistic* approach within the *DIF* framework does not distinguish between uniform and non-uniform DIF. When testing for non-uniform DIF (sixth column) one obtains four significant results. In contrast to items 1 and 11, items 13 and 21 are also found by *IFT*. In total item 21 is the only item that shows DIF according to all six models and four items are only identified as DIF items by one of the six models.

5.8. Concluding Remarks

The proposed recursive partitioning approach, in short IFT, is an extension of the basic logistic regression model for the detection of uniform and non-uniform DIF. In contrast to the classical approach, IFT allows to incorporate several covariates on different scales, including ordinal and continuous covariates, that potentially induce DIF. The method leads to simultaneous selection of items and (interactions of) variables that cause DIF. The result typically is a small tree for each DIF item and therefore the DIF structure is easy accessible.

The results of the simulations including uniform as well as non-uniform DIF show that IFT has the same performance than the classical approach in the simple case of two groups but also works quite well in more complex settings with various covariates. Neverthless, it should be noted that in the latter case the method is conservative and does not exploit the significance level fully. The applications demonstrate the flexibility and interpretability of IFT, also compared to the extended Logistic model that tests DIF by a vector of covariates. In particular, within the framework that tests for both types of DIF the obtained trees show which type of DIF is present.

The results shown in this chapter were obtained by the R-package `DIFtree` (Berger, 2016a) version 2.0.1 that is available on CRAN.

6. Modelling of Extreme Response Styles in Rating Scales

6.1. Introduction

In behavioral research rating scales have been used for a long time to investigate attitudes and behaviors. However, observed ratings may not represent the true opinion, in particular response styles may affect the response behavior, see, for example Messick (1991), Baumgartner and Steenkamp (2001). An extensive overview on response styles in survey research was given more recently by Van Vaerenbergh and Thomas (2013). A response style can be considered as a consistent pattern of responses that is independent of the content of a response (Johnson, 2003).

In this chapter we consider symmetric response categories of the form *strongly disagree, moderately disagree,..., moderately agree, strongly agree* and focus on response styles that are characterized by a disproportionate tendency to middle categories or to extreme categories, that is, the highest and lowest response categories. The preference to extreme categories is often called extreme response style and has been a topic of research for some time. Its counterpart, the tendency to choose middle categories has been investigated, for example, by Baumgartner and Steenkamp (2001).

In many studies the presence of response styles has been found. Response styles can differ, for example, across nations (Clarke, 2000; Van Herk et al., 2004), ethnicity (Marin et al., 1992) or educational level (Meisenberg and Williams, 2008). In particular, in the psychometric literature extreme response styles have been discussed within the framework of item response models. Bolt and Johnson (2009) and Bolt and Newton (2011) considered a multi-trait model, which is a version of the nominal response model proposed by Bock

This chapter is a modified version of Tutz and Berger (2016a). Inital considerations can be found in Berger and Tutz (2015b). For more information on the personal contributions of the authors and textual matches, see page 10.

(1972). Johnson (2003) considered a cumulative type model for extreme response styles. Eid and Rauber (2000) considered a mixture of partial credit models that is able to detect response styles. More recently tree type approaches have been proposed. They typically assume a nested structure where first a decision about the direction of the response and then about the strength is obtained. Models of this type have been proposed by Suh and Bolt (2010), De Boeck and Partchev (2012), Thissen-Roe and Thissen (2013), Jeon and De Boeck (2015), Böckenholt (2012), Khorramdel and von Davier (2014) and Plieninger and Meiser (2014).

In contrast to research in item response theory, where the focus is on the modelling of individual differences in terms of latent traits based on answers to several items without accounting for explanatory variables, in this chapter we aim at investigating the influence of explanatory variables on the content related choice and the response style for one item. The strength of the model is that it simultaneously accounts for both effects. It allows

- to investigate content related effects that are undisturbed by the response style for a single item,
- to investigate the response style undisturbed by content related effects,
- to use covariates to disentangle content and style,
- to avoid biased estimates of the content related effects, which are the parameters of interest in most studies.

Approaches to simultaneous modelling of content related effects and response styles seem to be scarce. Most approaches rely on the calculation of specific indices that can be corrected by regression techniques, see, for example Baumgartner and Steenkamp (2001). An exception are the latent class approaches considered, for example, by Moors (2004), Kankaraš and Moors (2009), Moors (2010) and Van Rosmalen et al. (2010). Latent class models are a strong tool but specific software is necessary and the existence of latent classes is always a strong assumption and interpretation has to rely on their existence. The crucial difference between these latent variable approaches and the proposed adjacent categories model is that the response style is not perceived as an individual trait, but exists solely in relation to the covariates. The model does not need the additional assumptions that accompany latent variable modelling.

The proposed modelling of response styles generated by covariates for one item uses a concept of the response style that differs from the usual concept. In the psychometric literature a response style typically is considered as a tendency in how a rating scale is used across items yielding a consistent pattern of responses that is independent of the content of a response (Johnson, 2003). When using this concept multiple items are a necessity. In our approach the tendency to extreme or middle categories is separated from the content related effects by using the symmetry of the response categories and letting covariates determine the tendency to specific categories. Nevertheless, since the model provides an explicit modelling of a tendency to extreme or middle categories the term response style seems also appropriate within our modelling framework.

In Section 6.2 the basic model is introduced. An illustrative example is given and a visualization tool is developed. In Section 6.3 the effects of parameters are discussed and the potential bias of estimates is investigated. Section 6.4 is devoted to inference, tools for the estimation of parameters are provided in Section 6.5. In Section 6.6 further applications that illustrate the method are given. In Section 6.7 we consider possible extensions and compare the approach to alternatives proposed, in particular, in item response theory. Finally, Section 6.8 introduces further extensions of the approach to the partial credit model.

6.2. An Extended Rating Scale Model

Let $Y_i \in \{1,\dots,k\}, i = 1,\dots,n$ denote the observed responses on a rating scale. Categories $1,\dots,k$ represent graded agree-disagree attitudes with a natural symmetry like *strongly disagree, moderately disagree,..., moderately agree, strongly agree.* If the number of response categories is odd there is a neutral middle category, if k is even there is none and the respondent is forced to exhibit at least a weak form of agreement or disagreement. Let $\boldsymbol{x}_i$ denote a vector of explanatory variables that is observed together with the response Y_i. Several models that link the explanatory variables to the ordinal response are available. Common model classes are the cumulative models, the sequential and adjacent categories models, see, for example, Agresti (2009) and Tutz (2012). We will focus on the adjacent categories model, which has the advantage that no constraints on the parameters are needed. Moreover, a specific version of the model is widely used in item response modelling. The partial credit model (Masters, 1982), which was already introduced in Chapter 4, uses the adjacent logit link to model item difficulties but does not include explanatory variables. In the following we first consider the basic model and then the extensions that account for response styles.

6.2.1. Adjacent Categories Model

The model proposed here is an extension of the adjacent categories model. The basic form of the model with logit link is given by

$$\log\left(\frac{\pi_{i,r+1}}{\pi_{ir}}\right) = \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad r = 1,\dots,k-1,$$

where $\pi_{ir} = P(Y_i = r|\boldsymbol{x}_i)$ denotes the conditional probability of response category r. The model assumes that the adjacent categories logits $\log(\pi_{i,r+1}/\pi_{ir})$ are determined by an intercept θ_r, which is specific for the adjacent categories, and a linear effect of the explanatory variables, $\boldsymbol{x}_i^T\boldsymbol{\beta}$. The ordering of the categories is modelled implicitly by assuming that the weight parameter does not depend on r. If one lets the parameter depend on the category one obtains the classical multinomial logit model, which does not exploit the ordering of the categories (Agresti, 2009).

The interpretation of the parameters of the model is seen best when the parameters are given as functions of probabilities. For covariate vector $\boldsymbol{x}^T = (x_1, \ldots, x_p)$ and corresponding parameter vector $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ it may be derived that the parameter of the j-th covariate is determined by

$$e^{\beta_j} = \frac{\pi_{r+1}(x_j+1)/\pi_r(x_j+1)}{\pi_{r+1}(x_j)/\pi_r(x_j)}, \tag{6.1}$$

where $\pi_r(x_j)$ denotes the probability of response category r for the vector of explanatory variables with the j-th covariate having value x_j and $\pi_r(x_j+1)$ is the probability of response category r if the j-th covariate is increased by one unit to $x_j + 1$. All other variables are fixed. Thus, e^{β_j} is the odds ratio that compares the odds for categories $r + 1$ and r when the j-th covariate is increased by one unit.

6.2.2. Accounting for Response Styles

For simplicity let us first consider the case of three response categories, $k = 3$. Then the model is given by the two equations that specify $\log(\pi_{i2}/\pi_{i1})$ and $\log(\pi_{i3}/\pi_{i2})$. The extended model proposed here contains the additional parameter δ_i and has the form

$$\log\left(\frac{\pi_{i2}}{\pi_{i1}}\right) = \theta_1 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta_i, \quad \log\left(\frac{\pi_{i3}}{\pi_{i2}}\right) = \theta_2 + \boldsymbol{x}_i^T\boldsymbol{\beta} - \delta_i.$$

The parameter δ_i specifies the response style. If $\delta_i \to \infty$ one obtains $\pi_{i2} \to 1$, which means a strong tendency to the middle category. If $\delta_i \to -\infty$ one obtains $\pi_{i2} \to 0$, which means a strong tendency to the response categories 1 and 3 corresponding to the extreme response style. It is important that the response style is separated from the preference represented by the linear term $\boldsymbol{x}_i^T\boldsymbol{\beta}$. While $\boldsymbol{x}_i^T\boldsymbol{\beta}$ represents the content-related effect, δ_i represents the response style towards the middle category or away from it.

The effect of the additional parameter is illustrated in Figure 6.1 for a univariate explanatory variable with $\beta = 1$. It is seen that a person with $\delta_i = 2$ has a stronger tendency to choose the middle category than a person with $\delta_i = 0$ whereas a person with $\delta_i = -2$ hardly

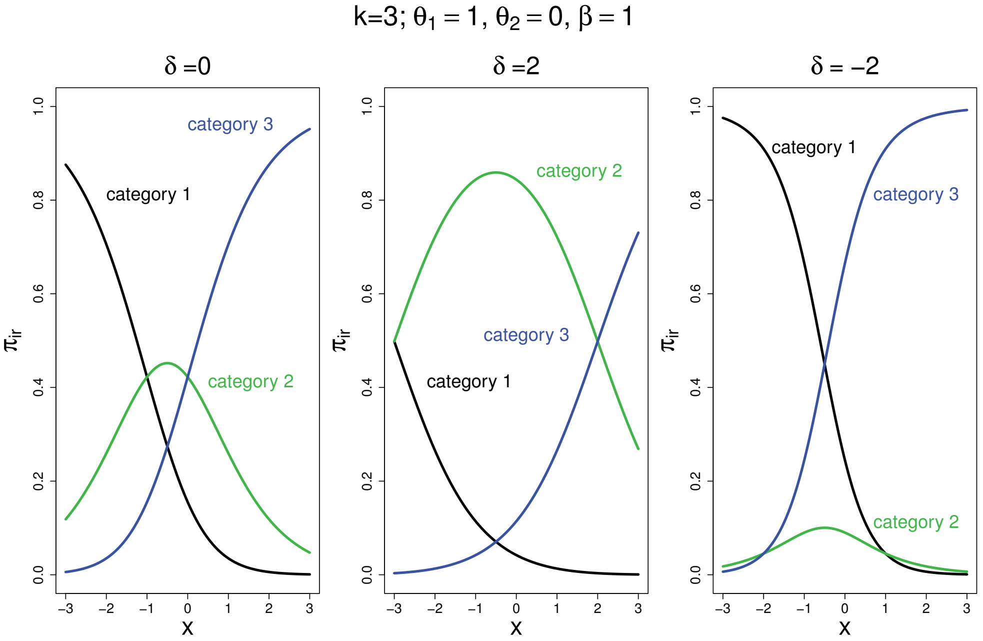


Figure 6.1.: Response functions for several values of δ_i.

uses the middle category. Although the numeric values change the shapes of the response functions for categories 1 and 3 are very similar for all values of δ_i.

The strength of the model is that the parameter δ_i can be specified as a function of explanatory variables. Let $\boldsymbol{z}_i$ be an additional vector of variables, which are assumed to determine the response style. The $\boldsymbol{z}_i$ can be different from $\boldsymbol{x}_i$ but can also be the same. With $\delta_i = \boldsymbol{z}_i^T\boldsymbol{\gamma}$ one obtains the model

$$\log\left(\frac{\pi_{i2}}{\pi_{i1}}\right) = \theta_1 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\gamma}, \quad \log\left(\frac{\pi_{i3}}{\pi_{i2}}\right) = \theta_2 + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{\gamma}.$$

The model has some interesting properties. From

$$\log\left(\frac{\pi_{i3}}{\pi_{i1}}\right) = \theta_1 + \theta_2 + 2\boldsymbol{x}_i^T\boldsymbol{\beta}$$

one sees that the log odds for the categories that actually represent agreement and disagreement are not affected by the term that determines the response style. On the other hand

$$\log\left(\frac{\pi_{i2}/\pi_{i1}}{\pi_{i3}/\pi_{i2}}\right) = \theta_1 - \theta_2 + 2\boldsymbol{z}_i^T\boldsymbol{\gamma}$$

shows that specific odds ratios do not depend on the content-related term.

It is noteworthy that the parameters of the content-related term are the same as in the simple adjacent categories model. This may be seen from simple derivation of the parameters for the simple adjacent categories model. For three response categories an even more intuitive form than (6.1) is given by

$$e^{2\beta_j} = \frac{\pi_3(x_j+1)/\pi_1(x_j+1)}{\pi_3(x_j)/\pi_1(x_j)},$$

which shows the explicit dependence on the categories that refer to agreement or disagreement. For the parameters of the response style effects one obtains

$$e^{2\gamma_j} = \frac{\pi_2(z_j+1)/\pi_1(z_j+1)}{\pi_3(z_j+1)/\pi_2(z_j+1)} / \frac{\pi_2(z_j)/\pi_1(z_j)}{\pi_3(z_j)/\pi_2(z_j)}.$$

The explicit form of the parameters also ensures that the model is identifiable.

The General Model for k Response Categories

In the general case one has to distinguish between an odd and even number of response categories. For k *odd* let $m = [k/2] + 1$ denote the middle category. Then the rating scale model that accounts for the tendency to the middle or extreme categories has the form

$$\begin{aligned} \log\left(\frac{\pi_{i,r+1}}{\pi_{ir}}\right) &= \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\gamma}, \quad r = 1,\dots,m-1, \\ \log\left(\frac{\pi_{i,r+1}}{\pi_{ir}}\right) &= \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{\gamma}, \quad r = m,\dots,k-1. \end{aligned} \tag{6.2}$$

The term $\theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}$ represents the usual effects of covariates $\boldsymbol{x}_i$ in an adjacent categories model. If $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is large higher categories are preferred, if it is small low categories are chosen.

Positive values of the term $\delta_i = \boldsymbol{z}_i^T\boldsymbol{\gamma}$ increase the probabilities of higher categories for $r = 1,\dots,m-1$ but decrease them for $r = m,\dots,k-1$. Thus δ_i determines if middle categories or extreme categories are preferred. The effect is also seen when considering extreme values of δ_i. For $\delta_i = \boldsymbol{z}_i^T\boldsymbol{\gamma} \to \infty$ one obtains $\pi_{im} \to 1$ and therefore a tendency to the middle category while $\delta_i \to -\infty$ entails $\pi_{i2},\dots,\pi_{i,k-1} \to 0$ and therefore a preference of the extreme categories.

It should be noted that the modeling approach differs from alternative perspectives on response styles. In the literature response styles are often defined as preferring the outer or the midpoint categories across many unrelated or weakly related items. In our model a negative value of the response style parameter indicating extreme response style captures not only a preference for the extremes ”strongly agree” compared to the adjacent category ”agree” but also a preference for ”agree” compared to ”somewhat agree”. The response style $\boldsymbol{\gamma}$-parameter thus picks up not only the tendency to select the extremes, but a general tendency to prefer more extreme categories given the substantive stand of the respondent.

For k *even* the model has a slightly different form. Let in this case $m = k/2$ denote the split between agreement and disagreement categories. Then the proposed model has the form

$$\begin{aligned}
\log\left(\frac{\pi_{i,r+1}}{\pi_{ir}}\right) &= \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\gamma}, \quad r = 1,\dots,m-1,\\
\log\left(\frac{\pi_{i,m+1}}{\pi_{im}}\right) &= \theta_m + \boldsymbol{x}_i^T\boldsymbol{\beta}, \qquad (6.3)\\
\log\left(\frac{\pi_{i,r+1}}{\pi_{ir}}\right) &= \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{\gamma}, \quad r = m+1,\dots,k-1.
\end{aligned}$$

The effect of the term $\delta_i = \boldsymbol{z}_i^T\boldsymbol{\gamma}$ is the same as in the case where k is odd. Large values indicate a tendency to the extreme response style, small values a tendency to the middle.

For simplicity we will use the abbreviation RSRS for the model (k odd or even) for *Rating Scale model accounting for Response Styles.* Before discussing the effects in detail we first consider an application.

An Illustrative Example

Although estimation methods have not yet been given we consider an application to illustrate the effects obtained by using the extended model. We consider data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy that have been used before by Gambacorta and Iannario (2013). They are available from http://www.bancaditalia.it/statistiche/indcamp. The response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). As explanatory variables we consider: gender (0: male, 1: female), the marital status (single, married, separated, widowed), the place of living (north, south, center), the general degree of confidence in other people from 1 (low) to 10 (high), the atmosphere the interview took place in (1 to 10), the citizenship and the age in decades. The respondents were also asked about their assessment if the household income is sufficient to see the family through to the end of the month rated from 1 (with great difficulty) to 5 (very easily). The analysis is based on a subset with 3816 respondents of the SHIW of 2010. Variable age was centered around 60 and variable confidence around 5. We fitted a simple adjacent categories model with all of the covariates and the extended version that accounts for response styles where all the variables are allowed to have content-related and response style effects. For the variables age and confidence we also included quadratic and cubic terms because

Table 6.1.: Parameter estimates and standard errors for the illustrative example (SHIW study).

	Covariates	**Extended Adjacent**		**Adjacent**	
		estimate	se	estimate	se
Content-related effects	Gender	-0.0302	0.0155	-0.0292	0.0154
(x-variables)	Married	0.0256	0.0240	0.0475	0.0223
	Separated	0.0291	0.0373	0.0200	0.0325
	Widow	0.0116	0.0338	0.0170	0.0292
	Center	0.1666	0.0192	0.1887	0.0195
	South	0.0169	0.0172	0.0170	0.0166
	Incomesufficient	0.0100	0.0060	0.0153	0.0059
	Atmosphere	0.0162	0.0054	0.0173	0.0047
	Citizenforeign	-0.0413	0.0414	-0.0545	0.0373
	Confidence	0.0035	0.0072	0.0029	0.0070
	Confidence^2	-0.0084	0.0011	-0.0082	0.0011
	Confidence^3	0.0008	0.0004	0.0011	0.0004
	Age	-0.0123	0.0086	-0.0160	0.0088
	Age^2	-0.0041	0.0031	-0.0029	0.0028
	Age^3	0.0010	0.0013	0.0015	0.0013
Response style effects	Gender	0.0034	0.0317		
(z-variables)	Married	-0.4208	0.0477		
	Separated	0.0067	0.0701		
	Widow	0.1063	0.0642		
	Center	-0.0385	0.0387		
	South	0.1336	0.0350		
	Incomesufficient	-0.0908	0.0124		
	Atmosphere	-0.1079	0.0106		
	Citizenforeign	0.3206	0.0806		
	Confidence	0.0073	0.0146		
	Confidence^2	-0.0228	0.0024		
	Confidence^3	0.0006	0.0010		
	Age	0.0003	0.0182		
	Age^2	-0.0259	0.0062		
	Age^3	0.0078	0.0028		

tests showed that the effects are different from zero. First of all, it is interesting if the style related effects are needed in the model. The likelihood ratio test for the null hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ has the χ^2-value 1,101.11 on 15 degrees of freedom. Therefore, style effects are definitely present. The estimated effects and standard errors for both models are given in Table 6.1. It is seen that the estimates as well as the standard errors of the content-related effects differ for the adjacent categories model and its extended version. In some cases the estimates are larger in other cases smaller if one ignores the response style (see also Section 6.3). As far as the effects on the response style are concerned it is seen that gender had no effect on the response style but, for example, sufficiency of income, age and confidence had effects on the response style that can not be ignored. The weight -0.09 on sufficiency of income with very small standard error indicates that confidence in the sufficiency of income increases the tendency to choose extreme categories. Instead of discussing the various effects in detail in the next sections visualization tools are developed.

Visualization of Effects

The extended model contains more parameters than a simple rating scale model. In particular, when various explanatory variables are included it is hard to keep track of all the relevant effects. Therefore we provide some visualization tools to investigate the effect strength. We explicitly consider the case of an odd number of response categories,model (6.2), and start with the visualization of linear effects. It is immediately seen that the odds of adjacent categories have the form

$$\begin{aligned}\frac{\pi_{i,r+1}}{\pi_{ir}} &= e^{\theta_r}(e^{\beta_1})^{x_{i1}}\dots(e^{\beta_p})^{x_{ip}}(e^{\gamma_1})^{z_{i1}}\dots(e^{\gamma_q})^{z_{iq}}, \quad r=1,\dots,m-1,\\ \frac{\pi_{i,r+1}}{\pi_{ir}} &= e^{\theta_r}(e^{\beta_1})^{x_{i1}}\dots(e^{\beta_p})^{x_{ip}}(e^{-\gamma_1})^{z_{i1}}\dots(e^{-\gamma_q})^{z_{iq}}, \quad r=m,\dots,k-1,\end{aligned}$$

where the explanatory variables for content-related effects have length p and the response style effects length q. Thus, if the j-th $x-$variable increases by one unit the multiplicative effect on the odds between adjacent categories is given by e^{β_j}.

If the j-th $z-$variable increases by one unit the multiplicative effect on the odds between adjacent categories depends on the category. It is e^{γ_j} for categories smaller than m and $e^{-\gamma_j}$ for the higher categories. If the x and z-variables are the same the effects are seen by plotting the tuple $(e^{\gamma_j}, e^{\beta_j})$. If a covariate is present only as an x- or z-variable one of the components in the tuple is 1.

For the SHIW study we show the effects of the marital status, gender and the area of living in Figure 6.2. In the figure pointwise confidence intervals are included. We use stars with the horizontal and vertical lengths corresponding to the 95% confidence intervals of e^{γ_j} and e^{β_j}, respectively. It is seen from the left panel that there is no difference between men and women in the response style (e^{γ_j} close to one), but women tend to choose lower scales of happiness (e^{β_j} around 0.97). For the variable marital status we chose "single" as the reference category obtaining the value $(e^{\gamma_j}, e^{\beta_j}) = (1, 1)$. It is seen that all others have higher happiness scores, although especially the effect of the category "widow" is not significantly different from the category "single". As far as the response styles are concerned, separated and widowed persons showed a tendency to the middle whereas married people give a more distinct response when compared to the reference category "single". From the right panel it is seen that people living in the center of Italy have significantly higher happiness scores

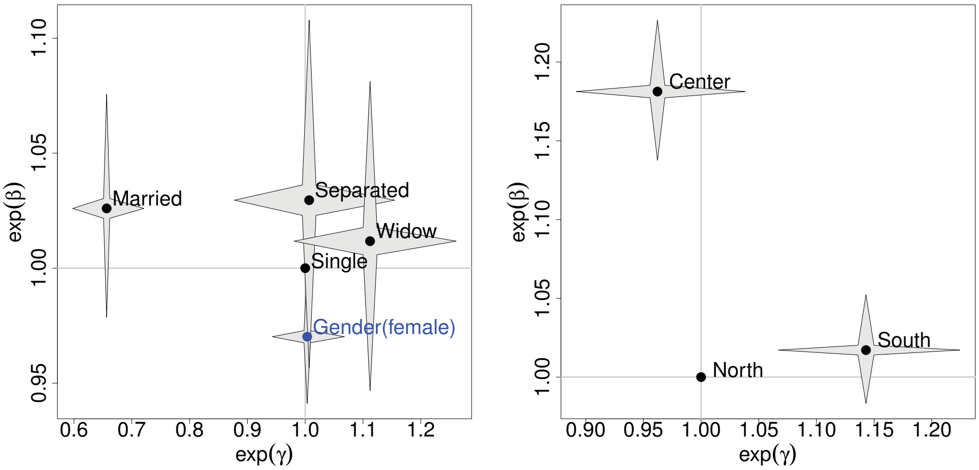


Figure 6.2.: Visualization of estimated effects for the illustrative example (SHIW study) including pointwise confidence intervals.

than people living in the south or the reference category "north". The difference in the preference of the response styles between categories "center" and "north" can be neglected but there is a significant difference between categories "south" and "north". People living in the south tend to choose less extreme response categories. It should be noted that the confidence intervals we show do not include the correlation between estimates to obtain a more easily accessible visualization. Moreover, correlations tend to be small (see Section 6.3).

Visualization of Non-Linear Effects

In the illustrative example the explanatory variables confidence and age contain in addition to linear terms quadratic and cubic terms. Then it is not sensible to plot the effects of parameters separately. One can understand the effects as functions of the corresponding explanatory variables. For example, the content-related effect of confidence is a polynomial containing cubic terms given by term $f_c^C(\text{conf}) = \text{conf}\beta_{c,1}^C + \text{conf}^2\beta_{c,2}^C + \text{conf}^3\beta_{c,3}^C$ (C indicating content) and the response style effect is given by $f_c^R(\text{conf}) = \text{conf}\beta_{c,1}^R + \text{conf}^2\beta_{c,2}^R + \text{conf}^3\beta_{c,3}^R$ (R indicating response style). Omitting for simplicity the linear effects of the other covariates one has the model

$$\begin{aligned}\frac{\pi_{i,r+1}}{\pi_{ir}} &= e^{\theta_r}(e^{f_c^C(\text{conf})})(e^{f_a^C(\text{age})})(e^{f_c^R(\text{conf})})(e^{f_a^R(\text{age})}), \quad r = 1, \dots, m-1,\\ \frac{\pi_{i,r+1}}{\pi_{ir}} &= e^{\theta_r}(e^{f_c^C(\text{conf})})(e^{f_a^C(\text{age})})(e^{-f_c^R(\text{conf})})(e^{-f_a^R(\text{age})}), \quad r = m, \dots, k-1,\end{aligned}$$

where $f_a^C(\text{age})$, $f_a^R(\text{age})$ represent the content and response style related effects of the variable age.

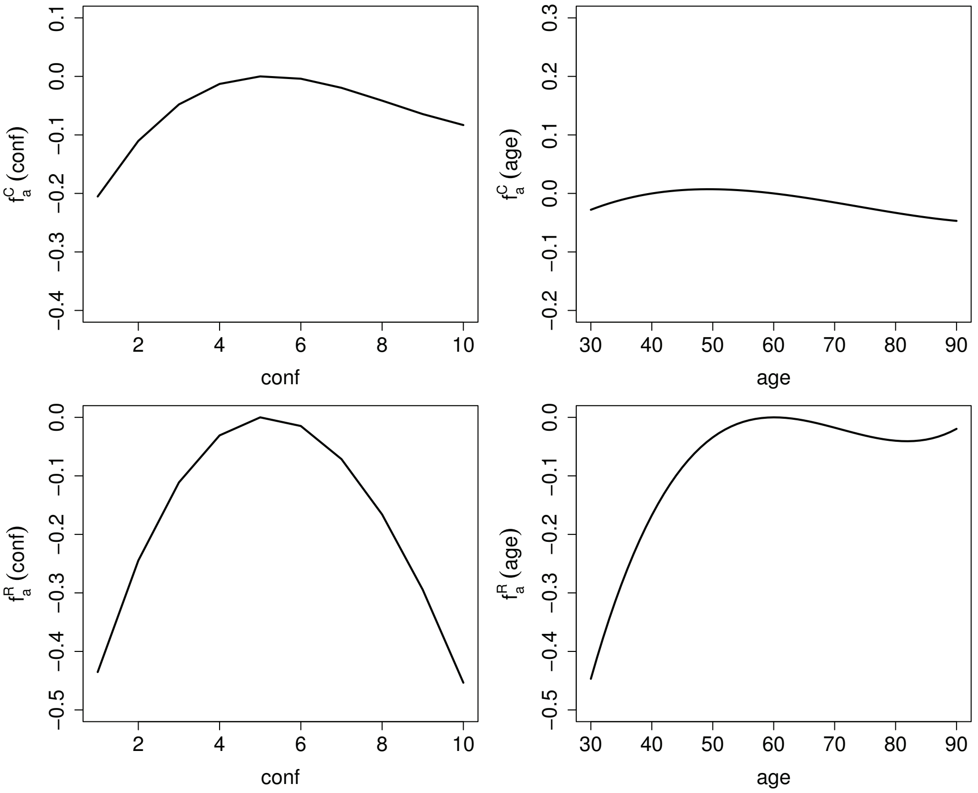


Figure 6.3.: Non-linear effects of content and response style of confidence and age for the illustrative example (SHIW study). The upper panels show the content, the lower panels the response style effects.

Parameters in polynomial terms are hard to interpret but one can plot the corresponding non-linear effects. Figure 6.3 shows the effects of content (first row) and response style (second row). In the figures we used the same scale in order to reveal the strength of the impact of the covariates. It is seen that with increasing confidence up to about value 5 the happiness increases and above 5 slightly decreases. For the response style one gets a distinctly quadratic effect. The tendency to extreme categories (negative values of $f_a^R(\text{age})$) is very strong for high and low values of confidence, and zero for middle categories of confidence. The content effect of age is not significant. Instead of omitting it we show the estimated curve, which is an almost horizontal line close to zero. Concerning the response style, it is seen that younger people have a tendency to extreme response styles, the effect vanished at age 50. It is close to zero for all values greater than 50.

As an alternative to these conventional plots for non-linear effects we propose to visualize them in a similar way as for linear effects by using axes that correspond to effects of response style and effects of content. The corresponding plot is obtained for the covariate confidence by plotting $(e^{f_c^R(\text{conf})}, e^{f_c^C(\text{conf})})$ as a function of confidence (10 points). However, instead of one point as in the visualization of linear effects one obtains a curve in two dimensions

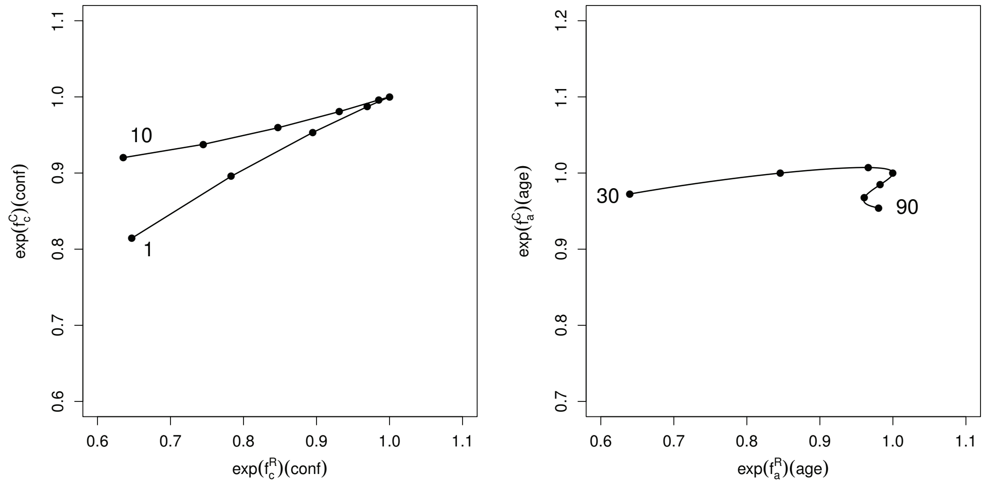


Figure 6.4.: Curves of non-linear effects for confidence (left) and age (right) for the illustrative example (SHIW study).

representing the multiplicative effects on the proportion of the probabilities of adjacent categories concerning content and response style related effects. Figure 6.4 shows the plots for the variables confidence and age. They show how both effects evolve with increasing value of the corresponding covariate. Again we use the same scale for both effects. The curves for confidence show the initial increase and subsequent weak decrease of happiness with the turning point at about 5. In particular for values of confidence between 5 and 10 the variation on the y-axis represents that the variation of the happiness score is weak. Much stronger variation is found for the response styles (x-axis). The tendency to extreme categories weakens with increasing confidence and then gets stronger with the same turning point at 5. The curve for age shows that the effect on happiness is weak with hardly any variation on the y-axis. However, the effect on the response style is rather strong. The tendency to use extreme categories found for 30 years of age diminishes strongly up to about 50 years of age and then hardly changes. The visualization by curves is useful for polynomial terms but can also be used for alternative smooth functions as considered briefly in Section 6.7.

6.3. Effects in the RSRS Model

One of the strengths of the extended RSRS model is that the content-related effects are separated from the tendency to middle or extreme categories. We will investigate the separation for the case k odd, for k even the separation works in a similar way.

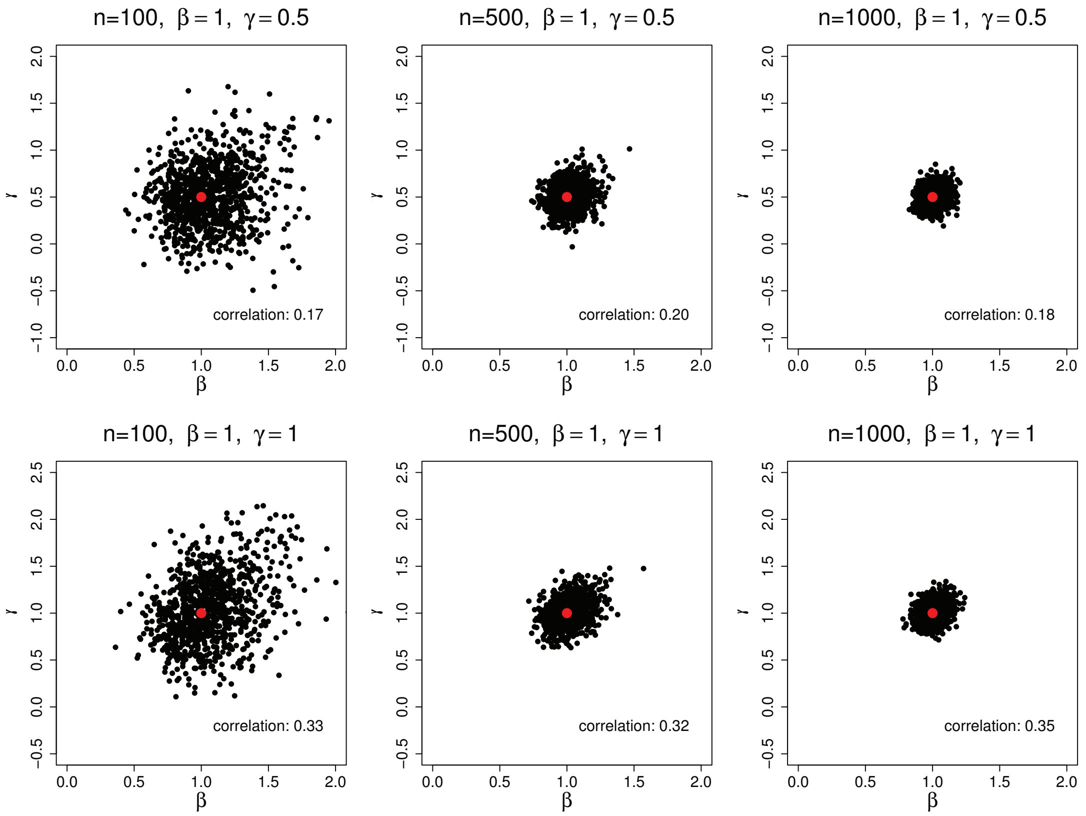


Figure 6.5.: Estimates for several values of β, γ and samples sizes. The explanatory variable follows a standard normal distribution, the true values are given in grey.

Let the model be given by (6.2) and again $m = [k/2] + 1$ denote the middle category. Then one may derive that the parameters of the $\boldsymbol{x}$-variables are determined by

$$e^{2r\beta_j} = \frac{\pi_{m+r}(x_j+1)/\pi_{m-r}(x_j+1)}{\pi_{m+r}(x_j)/\pi_{m-r}(x_j)}, \quad r = 1, \dots, m-1, \tag{6.4}$$

where $\pi_r(x_j)$ again denotes the probability of response category r for the vector of explanatory variables with the j-th covariate having value x_j and $\pi_r(x_j + 1)$ is the probability of response category r if the j-th covariate is increased by one unit to $x_j + 1$. All other covariate are fixed. The representation (6.4) compares the probabilities for the categories $m + r$ and $m - r$, that means categories with equal distance to the middle category. For $k = 7$ and therefore $m = 4$ it compares the probabilities of categories 5 and 3, 6 and 2 as well as 7 and 1. Thus it shows the effect of the explanatory variable in a symmetric way, namely how strong is the preference of, for example, category 5 compared to 3 if the explanatory variable increases by one unit.

It is essential that the parameter β_j does not depend on the term $\boldsymbol{z}_i^T\boldsymbol{\gamma}$, even if $\boldsymbol{x}_i = \boldsymbol{z}_i$. That means also in the simple adjacent categories model, where $\boldsymbol{z}_i^T\boldsymbol{\gamma} = 0$, the parameters β_j are given by (6.4). Therefore the content-related effects in the model are distinctly separated from the tendency to middle or extreme categories.

For the parameters that determine the response style one obtains

$$\gamma_j = 1/(2r)\left(\log\frac{\pi_m(z_j+1)/\pi_{m-r}(z_j+1)}{\pi_{m+r}(z_j+1)/\pi_m(z_j+1)} - \log\frac{\pi_m(z_j)/\pi_{m-r}(z_j)}{\pi_{m+r}(z_j)/\pi_m(z_j)}\right), \quad r = 1,\dots,m-1,$$

where in a similar way as before $\pi_r(z_j)$ denotes the probability of response category r for the vector of explanatory variables with j-th covariate z_j and $\pi_r(z_j+1)$ is the probability of response category r if the j-th covariate is increased by one unit to z_j+1. All other covariate are fixed. The parameter γ_j depends only on response probabilities of categories m, $m+r$ and $m-r$ for different values of z_j. It represents how the concentration of the probability mass is increased in the middle if z_j is increased by one unit. In the same way as β_j is separated from $\boldsymbol{z}_i^T\boldsymbol{\gamma}$ the parameter γ_j is separated from the term $\boldsymbol{x}_i^T\boldsymbol{\beta}$, signaling the separation of the weights on $\boldsymbol{x}$-variables and $\boldsymbol{z}$-variables. One effect of the separation of the effects is that estimates of γ_j, β_j if $x_j = z_j$ typically show weak correlation. For an illustration see Figure 6.5 where the estimates (1000 replications) of one normally distributed explanatory variable with $x = z$ are shown for various parameters β, γ and increasing sample size n. However, the separation of effects does not mean that the response style can be ignored when estimating the content-related effects of variables (see next section).

6.3.1. Accuracy of Estimates if the Response Style is Ignored

If one is not aware of response styles one fits a regression model that contains only the effect of explanatory variables on the response. In the following it is demonstrated that this procedure can result in strongly biased estimates and poor accuracy of the estimates of $\boldsymbol{\beta}$, which are the parameters of interest in most studies.

One Continuous Predictor

For simplicity we first consider the case of only one explanatory variable, which follows a standard normal distribution. Figure 6.6 shows the mean squared errors (MSEs), the variances and the bias of the ML estimate of β if one fits a simple adjacent categories model, which ignores the presence of differing response styles, and if one fits the extended model that accounts for the response style. The data generating model is the extended model with 7 categories for varying values of γ and $\theta_r = 0$, $\beta = 1$. The upper panels show the case where $x = z$, therefore one is estimating the content related effect of an explanatory variable that also has an effect on the response style. It is seen that the MSEs for both models is about the same for very small values of γ. For large absolute values of

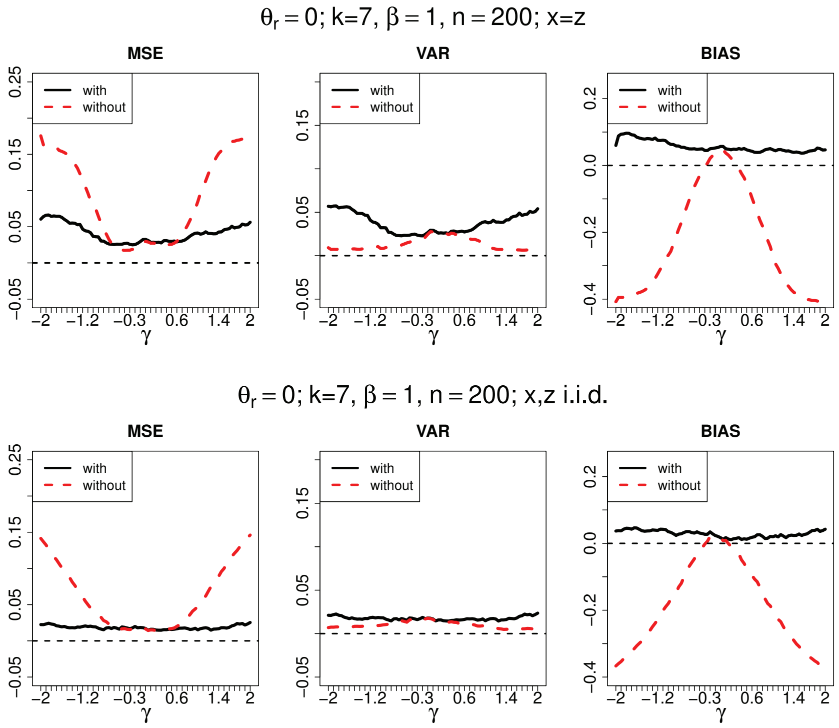


Figure 6.6.: MSEs, variances and bias as a function of γ for the simulation with one predictor; in the upper panel one has $x = z$, in the lower panel x and z differ and are independent. Dashed (red) lines indicate the model without accounting for the response style, solid (black) lines indicate the model with response style effects.

γ the MSE is much larger if the response style is ignored. The poor performance is mainly caused by the bias. One obtains strongly biased estimates even for moderate values of γ that underestimate the size of the effect. The effect is shown for the true value $\beta = 1$. The same strength of the bias is found if $\beta = -1$, but then the parameter β is overestimated instead of underestimated. The tendency is the same, one sees attenuation of the effects, in extreme cases if $\gamma = 2$ the absolute value of the estimate, $|\hat{\beta}|$, is almost the half of the true value $|\beta|$.

One might suspect that the bias is so strong because the variable has two effects, one on the preference and one on the response style. Therefore, we also investigated the case with a predictor $\eta_r = \theta_r + x\beta + z\gamma$, where x, z are independently normally distributed variables. The lower panel of Figure 6.6 shows the resulting curves. It is seen that one obtains biased estimates also if a variable that is independent of x generates varying response styles but is ignored. Therefore one ignores heterogeneity of response styles in the population.

In Figure 6.6 the effect is always attenuation of effects, a familiar phenomenon which also occurs in random effects models if heterogeneity is ignored, see, for example Tutz (2012), Chapter 14. However, in the case of ignored response styles in some cases one can also see

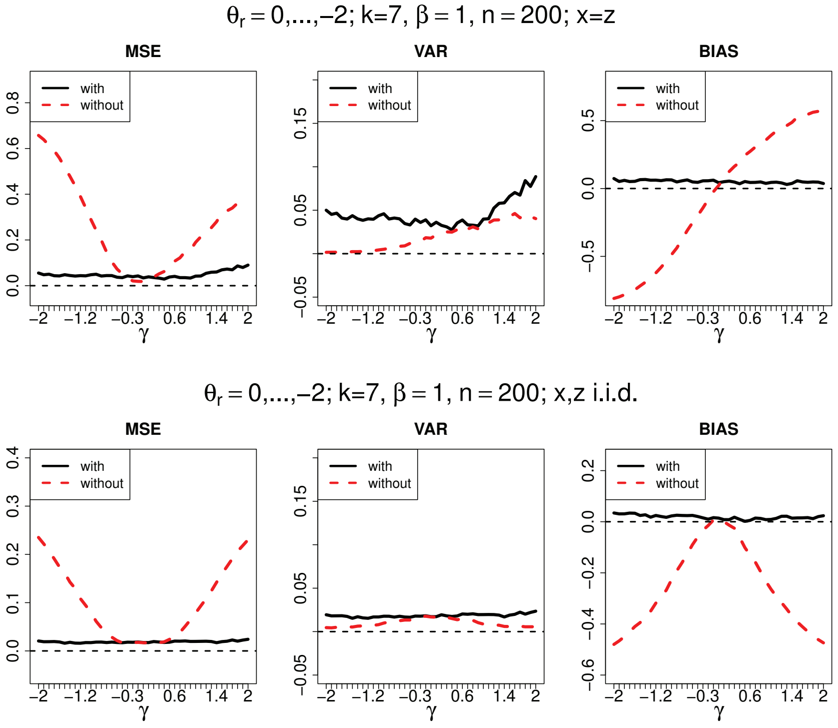


Figure 6.7.: MSEs, variances and bias as a function of γ for the simulation with one predictor and desccending thresholds; in the upper panel one has $x = z$, in the lower panel x and z differ and are independent. Dashed (red) lines indicate the model without accounting for the response style, solid (black) lines indicate the model with response style effects.

stronger effects. In Figure 6.7 MSEs, variances and bias are shown for the same models as in Figure 6.6, but now the thresholds have been changed to $\theta_1 = 0, \theta_2 = -0.4, \theta_3 = -0.8, \ldots$. For these descending thresholds higher categories are preferred for all of the values of the explanatory variables. It is seen that the bias is again negative for all values of γ if x and z are uncorrelated (lower panel) but one obtains overestimation of the true value of $\beta = 1$ in the case where $x = z$ if γ is positive (upper panel). Therefore, if there is a tendency to higher categories and the effect β is positive, and one ignores the tendency to select middle categories (γ positive), this is interpreted by the model without response style effect as a stronger β. The consequence is that larger values of β are obtained, the estimated effect tends to be larger than the true effect. For illustration of the effects we considered values of γ from a wide range. Although large values of γ might occur, in the real data sets we considered $|\gamma|$ was not beyond 1. An indicator of potential non-negligible bias might be strong differences in estimates for the model with response style and the model without response style.

Several Predictors

Further investigations show that the same effects are also found if more than just two variables are included in the model. Therefore, we consider data with 7 categories, four $\boldsymbol{x}$-variables, $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\top$, that are standard normal distributed without correlation and an one-dimensional z-variable. The true coefficients are $\boldsymbol{\beta} = (1, 0.5, -0.4, 0.3)^\top$. In the first case we set $z = x_1$, which means that the first x-variable has a content related effect as well as a response style effect. In the second case z is independently drawn from a standard normal distribution. As before, thresholds θ_r are either all set to zero or descending from zero. The corresponding results of MSEs, variances and bias for varying values of γ are shown in Figure 6.8. It is seen that the previous findings for the simulations with normal response can be confirmed and therefore the conclusions remain largely the same.

6.3.2. Effect of Sample Sizes

It has been demonstrated that biased estimates can be avoided by accounting for the response style when estimating the content-related effects. A quite different question is which observations contribute to the estimation accuracy when differing response styles are present and accounted for in the model. Intuitively accuracy of estimates will be weaker if many respondents prefer the middle category because then there is a tendency that less information about $\boldsymbol{\beta}$ is available. The effect can be illustrated by looking at the effect of β in the simple case of three response categories and a simple binary predictor x representing, for example, gender. As already shown in Section 6.2 the true effect is given by

$$e^{2\beta} = \frac{\pi_3(f)/\pi_1(f)}{\pi_3(m)/\pi_1(m)},$$

where $\pi_r(f), \pi_r(m)$ denote the probability of an response in category r for females and males, respectively. If in one of the two populations there is a strong tendency to the middle category the relative frequencies corresponding to $\pi_3(\cdot)/\pi_1(\cdot)$ will be estimated very unstable because only few observations will be observed in categories 1 and 3. Consequently, the accuracy of $\hat{\beta}$ will suffer.

To demonstrate the effect we show simulation results. We consider a binary predictor $x \in \{0, 1\}$, effect strengths $\beta = 1$ and $\gamma = 1$. Figure 6.9 shows the MSEs for a range of sample sizes, where n_0 denotes the sample size of population $x = 0$ and n_1 the sample size of population $x = 1$. In the left panel the thresholds were $\theta_1 = \theta_2 = 0$ yielding probability vectors $(0.33, 0.33, 0.33)$ for $x = 0$ and $(0.06, 0.468, 0.468)$ for $x = 1$. Therefore, in the population $x = 1$ the proportion $\pi_3(x = 1)/\pi_1(x = 1)$ is rather extreme and unstable to estimate. It is seen from Figure 6.9 that increasing the number of observations in the

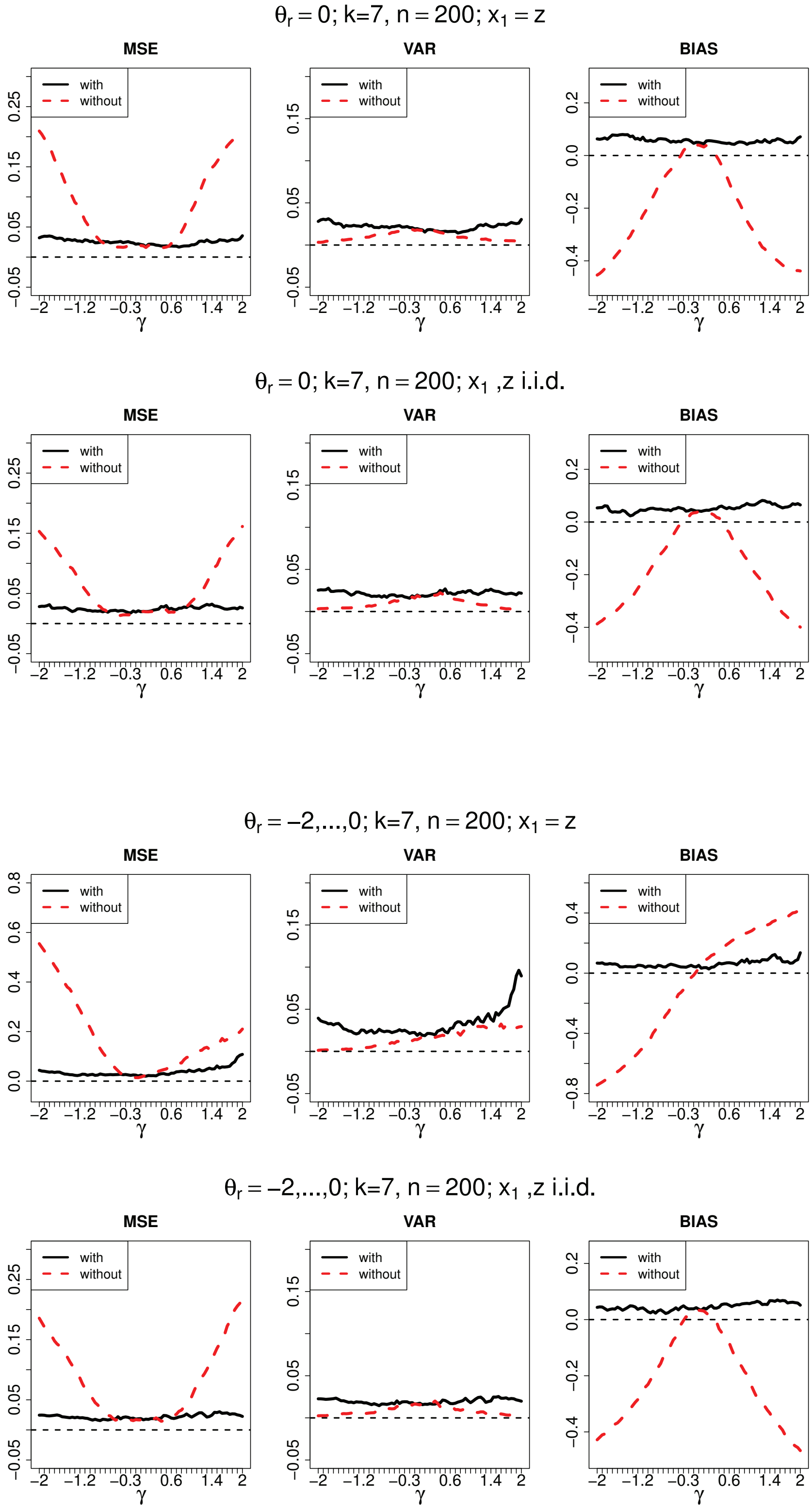


Figure 6.8.: MSEs, variances and bias as a function of γ for the simulations with several predictors; the upper panel corresponds to the setting with $\theta_r = 0$, the lower panel to the setting with descending thresholds. In the upper rows one has $x_1 = z$, in the lower rows x_1 and z differ and are independent, resp.

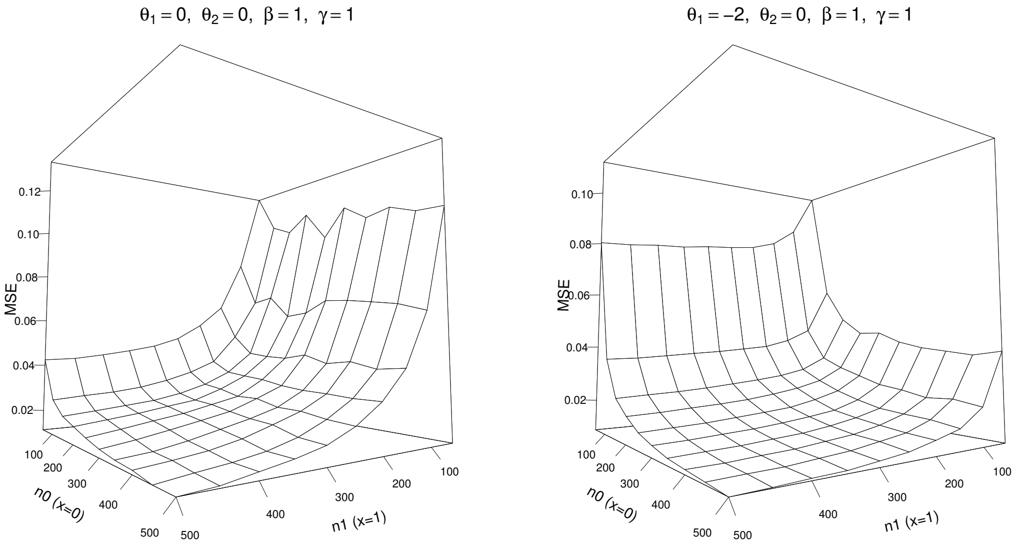


Figure 6.9.: MSE as a function of the sample sizes n_0, n_1 for sub populations $x = 0$, $x = 1$, resp.

population $x = 0$ does improve estimation accuracy only very little while increasing the number of observations in the population $x = 1$ improves the estimation accuracy very strongly. In the right panel of Figure 6.9 the thresholds are $\theta_1 = -2, \theta_2 = 0$ yielding probability vectors $(0.787, 0.106, 0.106)$ for $x = 0$ and $(0.33, 0.33, 0.33)$ for $x = 1$. Now the proportion $\pi_3(x = 0)/\pi_1(x = 0)$ is rather extreme and unstable to estimate. As is seen from the right panel increasing the number of observations in the population $x = 0$ strongly improves the estimates while increasing the number of observations in the population $x = 1$ hardly matters.

Thus, if extreme proportions occur in one population, which can be induced by response styles, estimation accuracy profits from the increase in these populations. The effect can not be exploited in a first investigation, but if one has a pilot study, which gives first results on the probabilities to expect, it can be used to stratify the sample in future studies to improve the accuracy of estimates.

6.4. Estimation of Parameters and Inference

Estimation and testing of the model is simplified by embedding the model into the framework of (multivariate) generalized linear models (GLMs). Let the data be given by $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \dots, n$. Given $\boldsymbol{x}_i, \boldsymbol{z}_i$, one assumes a multinomial distribution, $\boldsymbol{y}_i \sim \mathrm{M}(1, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i^T = (\pi_{i1}, \dots, \pi_{ik})$ with components $\pi_{ik} = P(Y_i = r|\boldsymbol{x}_i, \boldsymbol{z}_i)$. It is straightforward to show that the extended model can be given in the form

$$g(\boldsymbol{\pi}_i) = \boldsymbol{X}_i \boldsymbol{\delta}, \tag{6.5}$$

where $\boldsymbol{X}_i$ is a design matrix composed of the values $\boldsymbol{x}_i, \boldsymbol{z}_i$. $\boldsymbol{\delta}$ is the total vector of parameters containing the parameters $\theta_1, \ldots, \theta_{k-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and $g(\cdot)$ is a vector-valued *link function* $g = (g_1, \ldots, g_{k-1}) : \mathbb{R}^{k-1} \to \mathbb{R}^{k-1}$ given by

$$g_r(\pi_1, \ldots, \pi_{k-1}) = \log(\frac{\pi_{r+1}}{\pi_r}), \quad r = 1, \ldots, k-1.$$

An equivalent form of the link between explanatory variables and response is

$$\boldsymbol{\pi}_i = h(\boldsymbol{X}_i \boldsymbol{\delta}), \tag{6.6}$$

where $h = (h_1, \ldots, h_{k-1}) = g^{-1}$ is the so-called response function. Equations (6.5) and (6.6) represent the structural assumption of a multivariate GLM. Maximum likelihood estimates and inference for multivariate GLMs is extensively discussed in Fahrmeir and Tutz (2001) and Tutz (2012). For example, one can use likelihood ratio tests, score tests or Wald tests to test linear hypotheses of the form $H_0 : \boldsymbol{C\delta} = \boldsymbol{\xi}$ against $H_1 : \boldsymbol{C\delta} \neq \boldsymbol{\xi}$, where $\boldsymbol{C}$ is a fixed matrix of full rank and $\boldsymbol{\xi}$ is a fixed vector.

An interesting aspect is the covariance of estimates which is asymptotically given by the expected information or Fisher matrix, $\boldsymbol{F}(\boldsymbol{\delta}) = \mathrm{E}\left(-\partial l / \partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T\right)$, which has the form

$$\boldsymbol{F}(\boldsymbol{\delta}) = \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i(\boldsymbol{\delta}) \boldsymbol{X}_i.$$

The blocks $\boldsymbol{W}_i(\boldsymbol{\delta})$ of the weight matrix are given by $\boldsymbol{W}_i(\boldsymbol{\delta}) = (\frac{\partial g(\boldsymbol{\pi}_i)}{\partial \boldsymbol{\pi}^T} \boldsymbol{\Sigma}_i(\boldsymbol{\delta}) \frac{\partial g(\boldsymbol{\pi}_i)}{\partial \boldsymbol{\pi}})^{-1}$. If the the two sets of explanatory variables are the same, that is $\boldsymbol{x}_i = \boldsymbol{z}_i$ one can see from the model equations (6.2) and (6.3) that the column that codes the variable x_j and the column that codes the corresponding z-variable are orthogonal. Therefore, the estimates of the effects β_j and γ_j are asymptotically uncorrelated. The effects become orthogonal, really separating the content-related effect and the response style effect.

6.5. Implementation and Available Programs

The model can be estimated and evaluated by use of the the flexible `R` package `VGAM` (Yee, 2010; Yee, 2014), which has also be used in estimation and testing of our applications. Function `vglm()` allows to estimate so-called vector generalized linear models (Yee and Wild, 1996). The extended RSRS model can be seen as a special case of this general family of models. One has to use the family function `acat(reverse=FALSE)`, which specifies the link function that corresponds to the adjacent categories model in the ordering considered here. The argument `parallel=FALSE∼1` ensures that only intercepts are category-specific.

When using the function one has to distinguish between $\boldsymbol{x}$- and $\boldsymbol{z}$-variables. The $\boldsymbol{x}$-variables are not category-specific whereas the $\boldsymbol{z}$-variables represent a special case of category-specific covariates for which only the sign differs for categories below and above the middle category. For category-specific covariates one takes advantage of the argument `xij`. One just has to specify the design matrices by including the $\boldsymbol{z}$-variables in the specific form of models (6.2) and (6.3) and estimation of the extended model by `vglm()` is obtained. An `R` function that automatically generates the design matrix and estimates the model is available upon request. Embedding the estimation procedure into the framework of `VGAM` also has the advantage of quite fast computation. For more details see Appendix D.

6.6. Further Applications

Finally, we give the results of two further real data examples, which illustrate the applicability of the RSRS model.

6.6.1. Healthcare

As a second application we use data from the ALLBUS, the general social survey of social science carried out by the German institute GESIS. They are available from http://www.gesis.org/allbus. For our analysis we consider data from 2012 consisting of 2899 persons. The response is the confidence in the health care system measured on a scale from 1 (no confidence at all) to 7 (excessive confidence). Explanatory variables that we include in our model are: gender (0: male, 1: female), income in thousands of Euro, age in decades and the medical condition of the person on a scale from 1 (very good) to 5 (bad). Again we estimated a simple adjacent categories model and the extended model where all covariates were allowed to have content-related and response style effect. In a second step we refitted the model including only the covariates with a significant effect in each part. The estimated coefficients and the corresponding standard errors are given in Table 6.2. Concerning variable selection covariate gender and income are excluded from the $\boldsymbol{x}$-variables and covariate age is excluded from the $\boldsymbol{z}$-variables. The likelihood ratio test statistic for the global hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ is 44.6 on 8 degrees of freedom. Thus, response style effects should not be neglected. The ordinal predictor medical condition with reference "very good" has significant content-related effects as well as significant response-style effects. Figure 6.10 shows the tuple $(e^{\hat{\gamma}_j}, e^{\hat{\beta}_j})$ of the extended model including pointwise confidence intervals represented by stars. The estimated coefficients show that the confidence in the health care system decreases with deteriorating medical condition. In addition there is a significant tendency to choose extreme categories for persons with a bad medical

Table 6.2.: Parameter estimates and standard errors for the healthcare data.

	Covariates	Extended Adjacent		Adjacent	
		estimate	se	estimate	se
Content-related effects	Age	0.0694	0.0168	0.0702	0.0168
(x-variables)	Age^2	0.0206	0.0043	0.0225	0.0044
	Age^3	-0.0052	0.0024	-0.0055	0.0022
	Good	-0.0073	0.0472	-0.0416	0.0414
	Mostly Good	-0.1621	0.0479	-0.1499	0.0446
	Partly Good	-0.2663	0.0548	-0.2491	0.0543
	Bad	-0.3011	0.0718	-0.2834	0.0788
Response style effects	Gender	0.1380	0.0434		
(z-variables)	Income	0.0733	0.0238		
	Income^2	-0.0071	0.0030		
	Income^3	0.0001	0.0001		
	Good	0.1263	0.0676		
	Mostly Good	-0.0356	0.0685		
	Partly Good	-0.1602	0.0822		
	Bad	-0.3140	0.1172		

condition. For females compared to males there is a significant tendency to middle categories. The explanatory variables income and age contain also quadratic and cubic terms. Figure 6.11 shows the estimated non-linear effects of content (first row) and response style (second row). The covariate income has no significant effect on the confidence. However, with increasing income there is an increasing tendency to middle categories. The effect is not far from being linear but the quadratic and cubic term are significant. Concerning age, the confidence in the health care system decreases up to age 40 and increases between 40 and 80. The decrease after 80 should not be over-interpreted since it is based on few observations. There seems to be no effect of age on the response style (given the other covariates). We do not show the two-dimensional curves for this example because they are not informative.

6.6.2. Motivation of Students

As a third example we consider data from a student questionnaire. It has been evaluated what effect the expectation of students for getting an appropriate job has on their motivation. The response is the effect on motivation on a scale from 1 (often negative) to 5

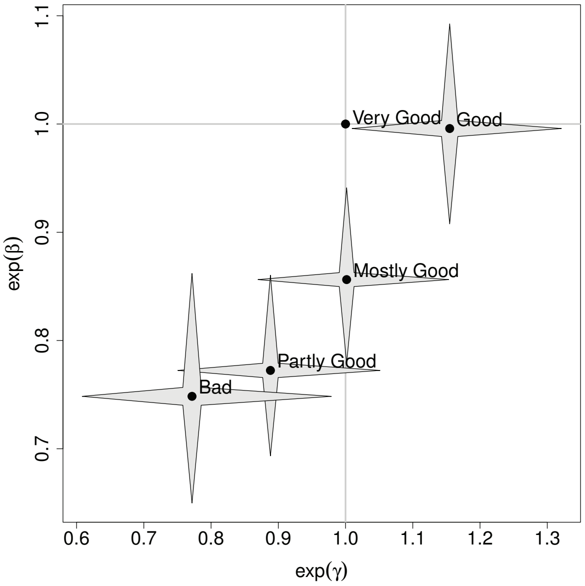


Figure 6.10.: Visualization of estimated effects of covariate medical condition for the healthcare data.

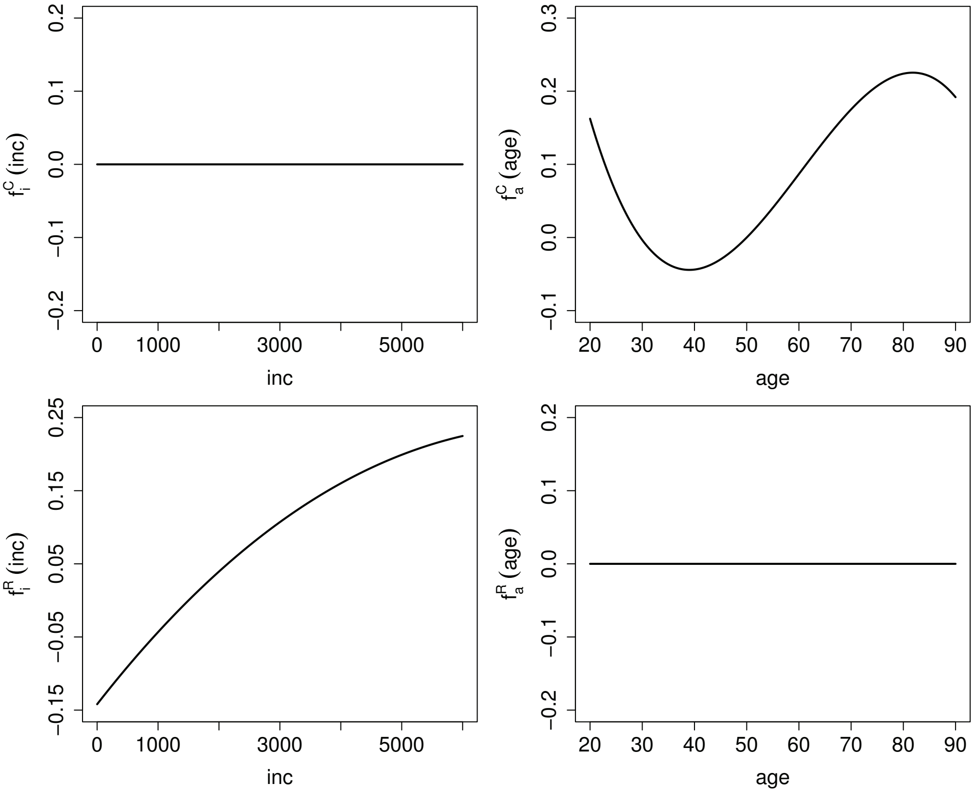


Figure 6.11.: Non-linear effects of content and response style of income and age for the healthcare data. The upper panels show the content, the lower panels the response style effects.

Table 6.3.: Data from a student questionaire.

	Effect on motivation				
Subject Area	Often negative	Sometimes negative	None or mixed	Sometimes positive	Often positive
Psychology	9	26	53	8	6
Physics	8	22	100	20	6
Teaching	26	20	35	0	4

(often positive) with intermediate values ”sometimes negative/positive” and no effect. For our analysis we use data from 343 students from the subject areas psychology, physics and teaching serving as explanatory variable. The data is given in Table 6.3. Overall there is a strong preference for the middle categories, which is characteristic for this sort of question. The comparison of the simple adjacent categories model and the extended model yields the likelihood ratio test statistic 6.14 on 2 degrees of freedom. Thus, response style effects again should not be neglected. The estimated coefficients for both models are given in Table 6.4, a visualization of the effects of the extended model including pointwise confidence intervals is shown in Figure 6.12, where subject “teaching“ was chosen as reference category.

The estimates in the content-related part of the model show that students of psychology and physics see more positive effects on their motivation than students of the teaching profession. In fact job prospects for students of the teaching profession are poor nowadays. The estimated response style effects show a significant tendency to middle categories for students of physics as compared to students of the teaching profession.

A comparison of the content-related effects in Table 6.4 for the simple and the extended model shows that the estimates of the simple model are considerably larger. Thus one observes a positive bias in the estimated β-coefficients of the $\boldsymbol{x}$-variables when ignoring response-style effects. One reason for the positive bias is the peculiar distribution of the data. Table 6.3 shows that most observations are in the middle category (none or mixed) and at the same time there is a general shift to the left or to low categories. Therefore, ignoring the tendency to the middle category leads to an overestimation of the $\boldsymbol{\beta}$-coefficients.

6.7. Extensions and Comparison with Alternative Approaches

In the following we shortly sketch possible extensions of the proposed modelling approach. The first concerns the handling of non-linear effects. If one has continuous covariates

Table 6.4.: Parameter estimates and standard errors for the student questionaire.

	Covariates	Extended Adjacent		Adjacent	
		estimate	se	estimate	se
Content-related effects	Psychology	0.4462	0.1867	0.6338	0.1688
(x-variables)	Physics	0.6616	0.1821	0.8798	0.1633
Response style effects	Psychology	0.2147	0.2308		
(z-variables)	Physics	0.5259	0.2226		

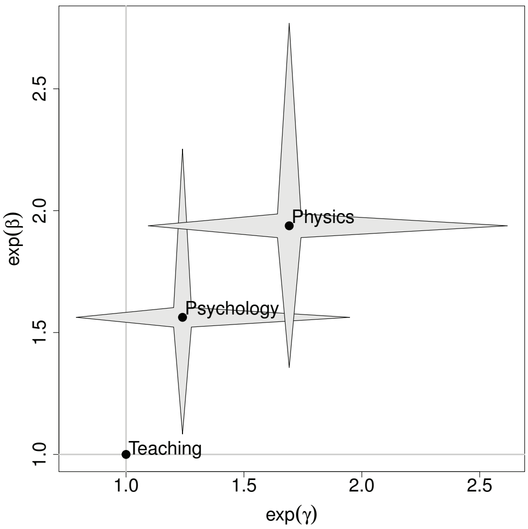


Figure 6.12.: Visualization of estimated effects of covariate subject area for the student questionaire.

one can replace the linear term $\boldsymbol{x}^T\boldsymbol{\beta}$ by an additive term $f_1^C(x_1) + \cdots + f_p^C(x_p)$ and the linear term $\boldsymbol{z}^T\boldsymbol{\gamma}$ by $f_1^R(z_1) + \cdots + f_q^R(z_q)$, where $f_j^C(\cdot), f_j^R(\cdot)$ are unspecified functions. In the illustrative example we already considered the effects as functions but they were restricted to be polynomials. Within the more general framework of additive modelling the functions can be considered as unknown without being specified as polynomials. Typically the unknown functions are approximated by an expansion in basis functions. For example, one assumes $f_j^C(x) = \sum_{r=1}^{M} \beta_{jr}\phi_{jr}(x)$, where ϕ_{jr} are fixed basis functions, for example, Gaussian kernels or B-splines. The latter has been propagated, in particular, by Eilers and Marx (1996). Then one estimates the parameters β_{jr}, which can be estimated in the usual way because the influential term is linear in the parameters. One option is to use few basis functions, for example, four to six and estimation will still be stable. A more flexible approach is to use many basis functions, say 40, but use penalization techniques that still allow to estimate the larger number of parameters. When the basis functions are chosen as B-splines one obtains the so-called penalized splines (P-splines), for details see Eilers and Marx (1996). By adapting these smoothing methods to the current problem the modelling of response styles can be extended to include additive terms in the tradition of generalized additive models (Hastie and Tibshirani, 1986). We do not consider the approach in detail because it involves more advanced penalization techniques, which might detract from the main contents in this chapter.

There are several modelling approaches to response styles that have been proposed, in particular in item response theory. A traditional way to account for differences in the use of rating scales are mixture models. For example, Eid and Rauber (2000) investigated measurement invariance in organizational surveys by using the polytomous mixed Rasch model. The basic assumption is that the whole population can be subdivided into disjunctive latent classes yielding parameters that are linked to the classes. Typically one fits models with two or three classes obtaining class-specific parameters that have to be interpreted. As Eid and Rauber (2000) demonstrated when fitting a model with two latent classes the classes might represent different response styles. The main difference to the approach propagated here is that response styles are not explicitly modelled. The resulting classes can represent extreme response styles or a tendency to the middle categories but do not have to. It might occur that no specific pattern referring to response styles is found for the latent classes. Although finite mixture models are an interesting approach to model heterogeneity, in particular the number of latent classes is not so easy to determine, and if one fits a model with more classes one might obtain quite different estimates and therefore different interpretations. Similar problems are found for the class of multidimensional extensions of response models that account for response styles as considered, for example, by Bolt and Johnson (2009). By including further latent traits in the predictor one obtains multidimensional models. The additional traits can represent response styles. Again the difference is that response styles are not explicitly searched for. Of course one might see this as an advantage. However, there is again some arbitrariness concerning the number of latent traits and the interpretation. The arbitrariness is augmented if the estimates have to be rotated (see for example, Bolt and Johnson, 2009), to obtain a simple interpretation. If one suspects different response styles we find it more attractive to model them explicitly. If one accounts for them by construction one can see if they are present or not. In the next section we introduce possible extensions of the proposed model to item response data.

More explicit modelling of response styles is found in tree type models as considered, for example, by Thissen-Roe and Thissen (2013) and more recently by Jeon and De Boeck (2015). The models assume a sequential decision model. In a first stage it is distinguished between a positive and a negative response, in subsequent steps the strength of the response is determined. Models of this type can be seen more general as nested models (Suh and Bolt, 2010). For ordinal responses with covariates they have been used earlier by Tutz (1989). The models are similar in spirit to the approach proposed here, they model response styles by parameters and have to distinguish between odd and even number of categories. The main differences are in the sequential decision procedure and the parameterization. In step models one assumes 1PL or 2PL models for the separate steps. In the approach considered here there is no sequential mechanism assumed and the parameters are embedded into an adjacent categories model.

Finally, we want to mention approaches to validate the interpretation of response style. In the case of several items this may be done by either selecting two item subsets that are weakly or unrelated (Moors, 2003; Moors, 2004) or use many items (Johnson, 2003; Van Herk et al., 2004) that are unrelated (Baumgartner and Steenkamp, 2001; Clarke, 2001; Weijters et al., 2010). This allows researchers to be certain that a persistent tendency across unrelated items can be ascribed to style (unrelated to item content). In our approach only one item is used to detect response styles but the model is constructed in a way to pick up the response style linked to the particular question that is asked.

6.8. Response Styles for Several Items

The model considered here by construction disentangles the effects of response style and content for one item. However, the basic concept to include a subject-specific term (added for response categories $r = 1, \ldots, m-1$ and subtracted for categories $r = m, \ldots, k-1$ if k is odd) can also be used when one wants to model the response style for more than one item. A common choice to model ordinal item response data is the partial credit model (PCM) proposed by Masters (1982). We will now introduce possible extensions of the proposed RSRS model for several items by use of the PCM. For simplicity we assume that the number of categories is equal across items.

Let $Y_{pi} \in \{1, \ldots, k\}$, $p = 1, \ldots, P$, $i = 1, \ldots, I$ denote the ordinal response of person p on item i, than the PCM assumes for the probabilities of adjacent categories

$$\log\left(\frac{P(Y_{pi} = r+1)}{P(Y_{pi} = r)}\right) = \eta_{pir} = \theta_p - \delta_{ir}, \quad r = 1, \ldots, k-1, \tag{6.7}$$

where θ_p is the person parameter and $(\delta_{i1}, \ldots, \delta_{ik})$ are the item parameters of item i. Representation (6.7) shows that the model is locally (given response categories $r-1$ and r) a binary Rasch model with person parameters θ_p and item difficulty δ_{ir}.

For an odd number of categories k with middle category $m = [k/2] + 1$ an extended partial credit model that accounts for response styles has the form

$$\begin{aligned} \eta_{pir} &= \theta_p + \gamma_p - \delta_{ir} = \theta_p - (\delta_{ir} - \gamma_p), \quad r = 1, \dots, m-1, \\ \eta_{pir} &= \theta_p - \gamma_p - \delta_{ir} = \theta_p - (\delta_{ir} + \gamma_p), \quad r = m, \dots, k-1, \end{aligned} \tag{6.8}$$

where the additional person-specific parameter γ_p determines the response style. The extension is also straightforward for an even number of categories. The parameter γ_p can be seen as a shifting of thresholds. If γ_p is positive one has a shifting of the thresholds δ_{ir} to the left for the disagreement categories yielding the new threshold $\delta_{ir} - \gamma_p$ and a shifting to the right for the agreement categories yielding the new thresholds $\delta_{ir} + \gamma_p$. The effect is that categories in the middle have higher probabilities of being chosen, which is the same as in model (6.2). If γ_p is negative one has the reverse effect. For $\gamma_p \to -\infty$ the whole probability mass is in the categories 1 and k. In the same way as in model (6.2) and (6.3) the additional parameter γ_p can be specified as a function of explanatory variables. If one uses the linear term $\gamma_p = \boldsymbol{z}_p^\top \boldsymbol{\alpha}$ the proposed estimation procedure in Section 6.4 can directly be used. Than one obtains estimates for the item difficulties, the person abilities and the additional response style parameters.

An alternative strategy that is certainly more attractive is to model the heterogeneity of persons by including an own subject-specific response style parameter. In order to reduce the number of parameters one can use random effects, that is one assumes that the response style parameters are drawn from a normal distribution $\gamma_p \sim N(0, \sigma_\gamma^2)$. If the focus is on valid estimates of the item parameters δ_{ir} one can also use a distribution for the ability parameters θ_p. Then one assumes a two-dimensional distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, with variances σ_θ^2, σ_γ^2 and a covariance $\sigma_{\theta\gamma}$. However, for the maximization of the corresponding marginal likelihood specific estimation procedures are needed and have to be developed.

Another quite interesting generalisation is to let the response style depend on the item. In many applications the assumption that it is the same for all items might be rather strong. However, if the response style depends on items one gets an inflation of parameters that call for regularization techniques or other novel estimation techniques. In summary, extended partial credit models are certainly worth investigating but the investigation of the possible models and the development of appropriate estimation tools need further research that is beyond the scope of the present work.

6.9. Concluding Remarks

A model is proposed that simultaneously accounts for content-related effects and response styles that have a tendency to middle or extreme categories. Thus content related effects can be studied without being influenced by the presence of specific response styles and vice versa. In traditional ways to investigate extreme response styles, for example, by computing an index for extreme response styles as the relative number of scores given on the extreme categories as used among others by Bachman and O'Malley (1984) and Van Herk et al. (2004) it is not known how the content-related effects are linked to the index. This is avoided by simultaneous modelling.

A particular strength of the approach is that it provides an easy to use tool and may avoid biased estimates. Of course it can not solve all the problems connected to rating scales. For example, it does not address problems linked to the number of response categories and response category labels (Weijters et al., 2010) or the tendency to show greater acquiescence (Baumgartner and Steenkamp, 2001) but can ameliorate some of the effects that come with specific response styles. Since researchers should "do whatever they can to control for response styles" (Van Vaerenbergh and Thomas, 2013) an easy to use tool should also be used.

7. Varying Dispersion in Cumulative Regression Models

7.1. Introduction

Since the seminal paper of McCullagh (1980) ordinal regression models have been widely applied in various fields of research, see, for example, Liu and Agresti (2005) and Agresti (2009). An important class of ordinal regression models is the class of cumulative models. The most prominent example is the proportional odds model, which will be considered exemplarily in the following before considering general cumulative models.

Let $Y_i \in \{1, \ldots, k\}$, $i = 1, \ldots, n$ denote the response and $\boldsymbol{x}_i$ a vector of explanatory variables. Then the basic form of the proportional odds model is given by

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad r = 1, \ldots, k-1, \tag{7.1}$$

where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$. An attractive feature of the model is the simple interpretation of parameters, which results from the proportional odds property. This property is seen from considering two sets of explanatory variables $\boldsymbol{x}, \tilde{\boldsymbol{x}}$ and the corresponding cumulative odds $\gamma(r|\boldsymbol{x}) = P(Y \leq r|\boldsymbol{x})/P(Y > r|\boldsymbol{x})$ and $\gamma(r|\tilde{\boldsymbol{x}}) = P(Y \leq r|\tilde{\boldsymbol{x}})/P(Y > r|\tilde{\boldsymbol{x}})$. Simple derivation shows that the proportion of the cumulative odds for the two sets of variables is given by

$$\frac{\gamma(r|\boldsymbol{x})}{\gamma(r|\tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^T\boldsymbol{\beta}),$$

and therefore does not depend on the category r. Consequently, the interpretation of parameters does not depend on the category. More concise, $\exp(\beta_j)$ represents the factor

This chapter is a modified version of Tutz and Berger (2016b). For more information on the personal contributions of the authors and textual matches, see page 10.

Table 7.1.: Quality of right eye vision in men and women.

	Highest (1)	Vision 2	Quality 3	Lowest (4)
Men	1053	782	893	514
Women	1976	2256	2456	789

by which all the cumulative odds $P(Y \leq r|\boldsymbol{x})/P(Y > r|\boldsymbol{x})$ change if variable x_j increases by one unit.

The simple interpretation gets lost in an extended version of the model in which parameters are category-specific. That means the predictor $\eta_{ir} = \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}$ in model (7.1) is replaced by $\eta_{ir} = \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}_r$. The corresponding *partial proportional odds model* frequently shows a better fit to the data but interpretation of parameters is more difficult. Moreover, severe restrictions are postulated. While the simple proportional odds model only postulates the ordering of the intercepts $\theta_1 \leq \cdots \leq \theta_{k-1}$ the extended version postulates $\theta_1 + \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \leq \cdots \leq \theta_{k-1} + \boldsymbol{x}_i^T\boldsymbol{\beta}_{k-1}$ for all values $\boldsymbol{x}_i$, which can severely restrict the possible values of explanatory variables. Therefore, often simple Fisher scoring does not work and estimation of parameters fails. For special link functions the cumulative model is equivalent to the sequential model, which allows to avoid the ordering of thresholds, see Tutz (1991) and, more recently Peyhardi et al. (2015). The class of partial proportional odds models has been investigated in particular by Brant (1990), Peterson and Harrell (1990) and Bender and Grouven (1998), graphical checks were proposed by Kim (2003) and Liu et al. (2009).

Despite its disadvantages the partial proportional odds model is often used if the fit of the proportional odds model is unsatisfactory. However, the lack-of-fit can also be caused by an insufficient modelling of dispersion effects. This chapter focussed on the modelling of varying dispersion in ordinal regression. The proposed model is related to the extended adjacent categories model developed in Chapter 6 to account for response styles.

For illustration let us consider a simple example that has already been used by McCullagh (1980). Table 7.1 shows Stuart's (1953) quality of right eye vision data for men and women. From the data it is obvious that women are more concentrated in the middle categories while men have relatively high proportions in the extreme categories. By construction the proportional odds model and other cumulative models without dispersion effects are not able to capture the different variability of subpopulations.

Ignoring dispersion effects is less severe in linear models. Varying dispersion, which for linear models is called heteroscedasticity, affects the precision of least squares estimates but they are still unbiased. However, ordinal regression models are non-linear models. For this class of models biased estimates are to be expected if dispersion is not modelled. In general, the modelling of variability is much harder than the modelling of the mean of the response. For ordinal responses an additional difficulty is that one can not use the variance of a univariate response because the response is multinomial and therefore multivariate. Although categories are ordered treating it as an univariate response would mean to ignore the scale level.

Here a model is proposed that models dispersion by including special effects in the linear predictor, which yields a model that can be estimated within the generalized linear model framework. The estimation procedure is strongly related to the one applied in Chapter 6. In Section 7.2 the model is introduced and an illustrative application is given. Tools for the estimation parameters and inference are provided in Section 7.3. Section 7.4 contains a detailed application. In Section 7.5 the model is compared to the location-scale model and consequences of ignored dispersion effects are briefly considered. After the consideration of non-symmetric responses in Section 7.6, in Section 7.7 alternative strategies to model ordinal response data by including category-specific effects are discussed and compared in further applications.

7.2. Separating Location and Dispersion

In this section we briefly show how cumulative ordinal models, which include the proportional odds model, and the extended location-scale model can be motivated from an underlying metric response model. Then we consider the model with shifted thresholds, which handles dispersion in a quite different way.

7.2.1. Cumulative Type Models for Ordinal Responses

Cumulative type models like the proportional odds model can be motivated by latent variables. The basic assumption is that the observed categories represent a coarser (categorical) version of an underlying (continuous) regression model. Let $\tilde{Y}_i$ be an underlying latent variable that follows a regression model:

$$\tilde{Y}_i = -\boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where ε_i is a noise variable with continuous distribution function F. Furthermore, let the link between the observable categories and the latent variable be given by

$$Y_i = r \quad \Leftrightarrow \quad \theta_{r-1} < \tilde{Y}_i \leq \theta_r,$$

where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = \infty$ are thresholds on the latent scale.

One obtains immediately

$$P(Y_i \leq r|\boldsymbol{x}_i) = P(-\boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i \leq \theta_r) = P(\varepsilon_i \leq \theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}) = F(\theta_r + \boldsymbol{x}_i^T\boldsymbol{\beta}).$$

The model is essentially a univariate response model since it is assumed that a univariate response $\tilde{Y}_i$ is in the background. The response Y_i is just a coarser version of $\tilde{Y}_i$ where the thresholds θ_r determine the preference for categories and the covariates produce a shifting on the latent scale. If $F(\cdot)$ is chosen as the logistic distribution function one obtains the proportional odds model (7.1).

A model that accounts for additional dispersion is obtained by assuming for the latent variable $\tilde{Y}_i = -\boldsymbol{x}_i^T\boldsymbol{\beta} + \tau_{\boldsymbol{x}_i}\varepsilon_i$, where $\tau_{\boldsymbol{x}_i}$ is the variance of the underlying regression model, which may depend on $\boldsymbol{x}_i$. The corresponding cumulative model with dispersion, also called *location-scale model*, is given by

$$P(Y_i \leq r|\boldsymbol{x}_i) = F\left(\frac{\gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta}}{\tau_{\boldsymbol{x}_i}}\right), \quad r = 1, \dots, k-1, \tag{7.2}$$

see McCullagh (1980). In cases where the concentration in response categories varies across populations, the model is more appropriate than the simple cumulative model. The simple cumulative model is based on the underlying continuous regression model $\tilde{Y}_i = -\boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, where the distribution of ε_i does not depend on $\boldsymbol{x}_i$. Thus the model assumes that with varying $\boldsymbol{x}_i$ the probability mass is merely shifted on the latent scale, therefore $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is often called the location effect. If the probability mass is more concentrated in one population and spread out in other populations, the simple cumulative model is unable to model the varying dispersion. The inclusion of a variance that varies over populations can capture this effect. Since the model includes a shifting and a dispersion or scaling effect it is often called a location-scale model.

One has to find appropriate ways to link the dispersion parameter to covariates. For example, one can use $\tau_{\boldsymbol{x}_i} = \exp(\boldsymbol{x}_i{}^T\boldsymbol{\gamma})$, which makes $\tau_{\boldsymbol{x}_i}$ positive. However, the model is highly non-linear and one is no longer within the framework of (multivariate) generalized linear models. Special software is needed to fit the model. For example, Cox (1995) used non-linear regression programs available in SAS. For further investigation of the model see also Nair (1987) and Hamada and Wu (1990).

7.2.2. Modeling Dispersion by Shifted Thresholds

In the following an alternative way to account for varying dispersion is proposed. Let us consider first the case with an even number of response categories k. Then $m = [k/2]$ splits the response categories into equally sized sets $\{1, \dots, m\}$ and $\{m+1, \dots, k\}$. Moreover, it is assumed that the ordered categories refer to a symmetric response, for example by categories of agreement as *strongly disagree, moderately disagree,..., moderately agree, strongly agree.* Let $\boldsymbol{z}_i$ be an additional vector of explanatory variables, which can be identical to $\boldsymbol{x}_i$ but does not have to.

Let now the thresholds in the proportional odds model be determined by

$$\begin{aligned}
\theta_r &= \beta_{0r} - \boldsymbol{z}_i^T\boldsymbol{\alpha}, \quad r = 1, \ldots, m-1,\\
\theta_m &= \beta_{0m},\\
\theta_r &= \beta_{0r} + \boldsymbol{z}_i^T\boldsymbol{\alpha}, \quad r = m+1, \ldots, k-1.
\end{aligned}$$

That means the center threshold θ_m remains fixed, but lower and upper thresholds are shifted by $\delta_i = \boldsymbol{z}_i^T\boldsymbol{\alpha}$. If δ_i is positive the intervals defined by thresholds are widened, indicating weaker dispersion, if δ_i is negative the intervals are shrunk, indicating stronger dispersion. With $\pi_i(r) = P(Y_i \le r|\boldsymbol{x}_i, \boldsymbol{z}_i)$ the model has the form

$$\begin{aligned}
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{\alpha}), \quad r = 1, \ldots, m-1,\\
\pi_i(m) &= F(\beta_{0m} + \boldsymbol{x}_i^T\boldsymbol{\beta}),\\
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\alpha}), \quad r = m+1, \ldots, k-1.
\end{aligned}$$

Since it is composed of a location component and a shifting of thresholds it is called the *location-shift model.* It is easily derived that $P(Y_i = m|\boldsymbol{x}_i, \boldsymbol{z}_i) + P(Y_i = m+1|\boldsymbol{x}_i, \boldsymbol{z}_i) = F(\beta_{0,m+1} + \boldsymbol{x}_i^T\boldsymbol{\beta} + \delta_i) - F(\beta_{0,m-1} + \boldsymbol{x}_i^T\boldsymbol{\beta} - \delta_i)$. Therefore if $\delta_i \to \infty$ one obtains $P(Y_i = m|\boldsymbol{x}_i, \boldsymbol{z}_i) + P(Y_i = m+1|\boldsymbol{x}_i, \boldsymbol{z}_i) \to 1$, which means a tendency toward the middle categories and therefore weak dispersion. In contrast, strong dispersion ($\delta_i \to -\infty$) means a tendency towards the extreme categories, which can also be interpreted as extreme response style (compare Chapter 6).

The effect of the additional term $\delta = \boldsymbol{z}^T\boldsymbol{\alpha}$ is illustrated in Figure 7.1 for a response with $k = 8$ categories and a binary covariate $x \in \{-1, 1\}$ with $\beta = 1$. We set $x = z$ and chose $\theta_1 = -3, \theta_2 = -2, \ldots, \theta_6 = 2, \theta_7 = 3$. Figure 7.1 shows the distribution of probabilities without dispersion ($\alpha = 0$) and with dispersion effects $\alpha = 0.4$ and $\alpha = -0.4$. It is seen that for $\alpha = 0.4$ the distribution is more concentrated in the middle if $x = 1$ and stronger dispersed if $x = -1$ when compared to the baseline distribution (first row). For $\alpha = -0.4$ one sees the reverse effect, stronger dispersion if $x = 1$ and more concentration in the middle if $x = -1$.

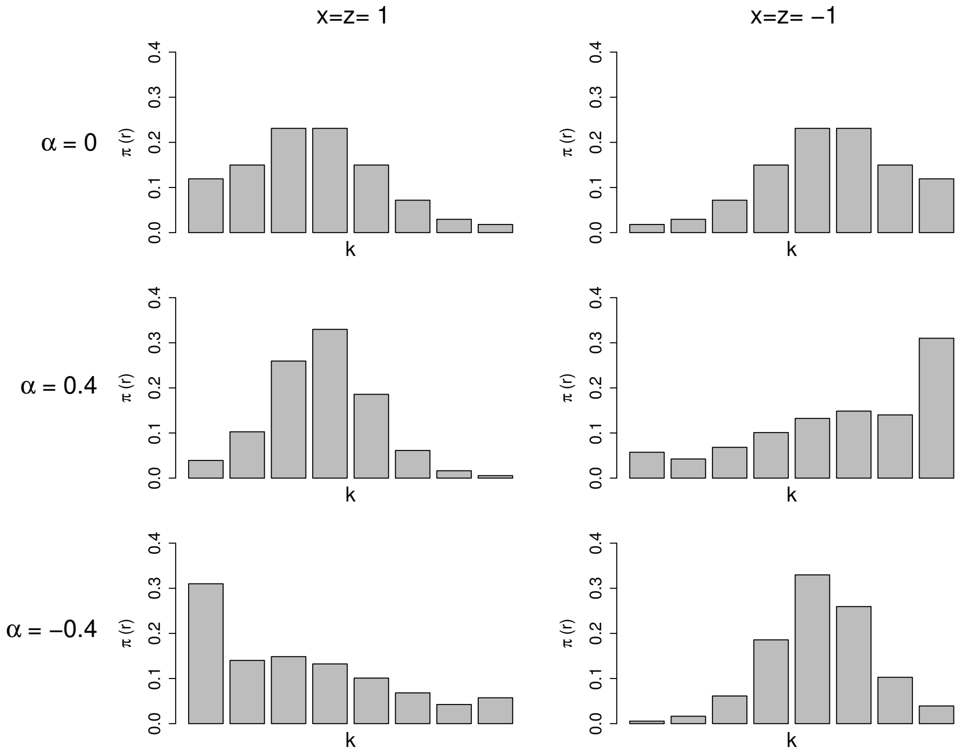


Figure 7.1.: Probability distribution of a response with eight categories for several values of α.

Effects and Interpretation of Parameters

Let first $\boldsymbol{x}$ and $\boldsymbol{z}$ be distinct. It is easily derived that then the proportional odds assumption still holds for the $\boldsymbol{x}$-variables With $\gamma(r|\boldsymbol{x},\boldsymbol{z}) = P(Y \leq r|\boldsymbol{x},\boldsymbol{z})/P(Y > r|\boldsymbol{x},\boldsymbol{z})$ denoting the cumulative odds for category r one obtains for two sets of explanatory variables $\boldsymbol{x}, \tilde{\boldsymbol{x}}$

$$\log\left(\frac{\gamma(r|\boldsymbol{x},\boldsymbol{z})}{\gamma(r|\tilde{\boldsymbol{x}},\boldsymbol{z})}\right) = (\boldsymbol{x}-\tilde{\boldsymbol{x}})^T\boldsymbol{\beta}.$$

Therefore the proportion of cumulative odds $\gamma(r|\boldsymbol{x},\boldsymbol{z})$ and $\gamma(r|\tilde{\boldsymbol{x}},\boldsymbol{z})$ are the same for all categories r. A consequence is that the parameter β_j from the vector $\boldsymbol{\beta}^T = (\beta_1,\ldots,\beta_p)$ is given by

$$e^{\beta_j} = \frac{\gamma(r|(x_1,\ldots,x_j+1,\ldots,x_p),\boldsymbol{z})}{\gamma(r|(x_1,\ldots,x_j,\ldots,x_p),\boldsymbol{z})}. \tag{7.3}$$

That means, if x_j increases by one unit the cumulative odds for each category change by the factor e^{β_j}. For $e^{\beta_j} > 1$ the increase of variable x_j favors low response categories. Thus the main advantage of the proportional odds model, namely, the simple interpretation of parameters, is kept.

For the $\boldsymbol{z}$-variables the interpretation is different. One obtains for two sets of explanatory variables $\boldsymbol{z}, \tilde{\boldsymbol{z}}$

$$\log\left(\frac{\gamma(r|\boldsymbol{x},\boldsymbol{z})}{\gamma(\boldsymbol{x},\tilde{\boldsymbol{z}})}\right) = \begin{cases} -(\boldsymbol{z}-\tilde{\boldsymbol{z}})^T\boldsymbol{\alpha}, & r \in \{1,\dots,m-1\} \\ \ (\boldsymbol{z}-\tilde{\boldsymbol{z}})^T\boldsymbol{\alpha}, & r \in \{m+1,\dots,k-1\}. \end{cases}$$

Thus for α_j from the vector $\boldsymbol{\alpha}^T = (\alpha_1,\dots,\alpha_q)$ one obtains

$$\begin{aligned} e^{-\alpha_j} &= \frac{\gamma(r|\boldsymbol{x},(z_1,\dots,z_j+1,\dots,z_q))}{\gamma(r|\boldsymbol{x},(z_1,\dots,z_j,\dots,z_q))}, \quad r \in \{1,\dots,m-1\}, \\ e^{\alpha_j} &= \frac{\gamma(r|\boldsymbol{x},(z_1,\dots,z_j+1,\dots,z_q)}{\gamma(r|\boldsymbol{x},(z_1,\dots,z_j,\dots,z_q))}, \quad r \in \{m+1,\dots,k-1\}. \end{aligned}$$

That means, if z_j increases by one unit the cumulative odds for categories $r < m$ change by the factor $e^{-\alpha_j}$ and for categories $r > m$ by the factor e^{α_j}. For $\alpha_j > 0$ the increase of variable z_j decreases the cumulative odds for categories $r < m$ and increases the cumulative odds for categories $r > m$, which means that the response probabilities for extreme categories get smaller. The effect is not a shifting of the probability mass of the response but a stronger concentration in the middle.

If $\boldsymbol{x} = \boldsymbol{z}$ the interpretation of parameters is similar. For simplicity we consider an one dimensional x. It is immediately seen that

$$e^{\beta} = \frac{\gamma(m|x+1)}{\gamma(m|x)}.$$

Thus e^{β} represents the odds ratio for categories smaller or equal m if x increases by one unit. It corresponds to the parameter in a binary logit model that distinguishes between categories $\{1,\dots,m\}$ and $\{m+1,\dots,k\}$. For the other cumulative odds one obtains

$$\frac{\gamma(r|x+1)}{\gamma(r|x)} = \begin{cases} e^{\beta}e^{-\alpha}, & r \in \{1,\dots,m-1\} \\ e^{\beta}e^{\alpha}, & r \in \{m+1,\dots,k-1\}. \end{cases} \tag{7.4}$$

Thus $e^{-\alpha}$ and e^{α} modify as factors the basic preference for categories from $\{1,\dots,m\}$ or $\{m+1,\dots,k\}$. For symmetric categories as considered here one obtains a more intuitive form by using for large categories the complementary odds defined by $\tilde{\gamma}(r|x) = P(Y \geq r|x)/P(Y < r|x)$, which give the odds for categories larger or equal r. They are linked to the usual cumulative odds by $\tilde{\gamma}(r|x)^{-1} = \gamma(r-1|x)$. One obtains for categories $r \in \{m+1,\dots,k\}$

$$\frac{\tilde{\gamma}(r|x+1)}{\tilde{\gamma}(r|x)} = e^{-\beta}e^{-\alpha}, \quad r \in \{m+1,\dots,k-1\}.$$

Thus the scaling factor that modifies the basic preference is again $e^{-\alpha}$. If one considers, for example, only the extreme categories one has

$$\frac{\gamma(1|x+1)}{\gamma(1|x)} = e^{\beta}e^{-\alpha} \quad \text{and} \ \frac{\tilde{\gamma}(k|x+1)}{\tilde{\gamma}(k|x)} = e^{-\beta}e^{-\alpha}.$$

Thus the modification of the odds for category 1 as compared to all other categories and the odds for category k as compared to all other categories (the complementary cumulative odds) are both modified by the factor $e^{-\alpha}$, which means for $\alpha > 0$ that both are shrunk by the factor $e^{-\alpha}$.

The parameter α itself is given by

$$e^{-2\alpha} = \frac{\gamma(s|x+1)/\gamma(s|x)}{\gamma(r|x+1)/\gamma(r|x)} = \frac{\gamma(s|x+1)\tilde{\gamma}(r|x+1)}{\gamma(s|x)\tilde{\gamma}(r|x)}$$

for any $s < m, r > m$. The product $\gamma(s|x)\tilde{\gamma}(r|x)$ is a measure for the concentration of the probabilities in extreme categories. It is large if the probabilities of extreme categories are large. Therefore, $e^{-2\alpha}$ represents the change of the concentration in extreme categories if x increases by one unit.

Eye Vision Example

Let us consider the simple quality of eye vision example from Table 7.1. The fitted values of the simple proportional odds model and for the location-shift model with dispersion effect are shown in Table 7.2. It is seen that in both models the location effect ($\hat{\beta} = -0.038$ and $\hat{\beta} = 0.042$) is rather weak and not significant at the 0.05 level. In contrast the dispersion parameter in the model with dispersion $\hat{\alpha} = 0.353$ can definitely not be neglected. The deviance of the proportional odds model is 128.39 on 2 df but reduces to 5.896 on 1 df for the model with location and dispersion effect. The estimated shrinkage factor is $e^{-\hat{\alpha}} = 0.70$, which means that for females the odds for the extreme categories 1 and 4 are shrunk by the factor 0.70 when compared to males.

Model for an Odd Number of Response Categories

Let now categories refer to a symmetric response with categories of agreement as *strongly disagree, moderately disagree,..., moderately agree, strongly agree* but with a neutral category in the middle. Then the number of categories k is an odd number. The model

Table 7.2.: Parameter estimates, standard errors and z-values for the eye vision data.

Covariate	**Proportional Odds Model**			**Location-Shift Model**		
	estimate	se	z value	estimate	se	z value
Intercept1	-0.905	0.034	-26.613	-0.721	0.037	-19.397
Intercept2	0.293	0.033	8.911	0.236	0.033	7.104
Intercept3	2.005	0.039	50.398	1.710	0.045	37.563
gender-location	-0.038	0.038	-1.003	0.042	0.038	1.109
gender-dispersion				0.353	0.031	11.348

with a dispersion component has the same basic structure but now one parameterizes for $m = [k/2] + 1$, which denotes the middle category,

$$\begin{aligned} \theta_r &= \beta_{0r} - \boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = 1, \dots, m-1, \\ \theta_r &= \beta_{0r} + \boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = m, \dots, k-1. \end{aligned}$$

The interpretation is similar as in the case with an even number of response categories. For e^{β_j} one obtains the same interpretation, that is, (7.3) is still the same. Also for the scaling parameters one obtains the same values, but they hold for different response categories. One obtains

$$\begin{aligned} e^{-\alpha_j} &= \frac{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q))}{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}, \quad r \in \{1, \dots, m-1\}, \\ e^{\alpha_j} &= \frac{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q)}{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}, \quad r \in \{m, \dots, k-1\}. \end{aligned}$$

The same holds for (7.4), which is still valid but for accordingly modified categories.

7.2.3. Shifting of Thresholds with Scaling

In the models considered in the previous sections the thresholds have been shifted away from the middle by the value $\delta_i = \boldsymbol{z}_i^T \boldsymbol{\alpha}$. The effect is a widening of the middle category if k is odd and of the two categories in the middle if k is even. However, the other categories have not been widened. Alternatively one can understand dispersion as a widening of all the categories by using scale values for the widening of the intervals between two thresholds. Let us consider again the case k even and $m = [k/2]$.

Let the thresholds be determined more generally by

$$\begin{aligned}
\theta_r &= \beta_{0r} - s_r \boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = 1, \dots, m-1, \\
\theta_m &= \beta_{0m}, \\
\theta_r &= \beta_{0r} + s_r \boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = m+1, \dots, k-1.
\end{aligned}$$

where s_r are scale values that reflect the distance between categories r and m. A simple choice is $s_1 = \ldots = s_{k-1} = 1$, which yields the model used in the previous section.

A particularly attractive choice of scales is obtained by shifting of the thresholds proportional to the distance from the middle threshold. Then one uses $s_r = m - r$ for $r = 1, \dots, m$ and $s_r = r - m$ for $r = m+1, \dots, k-1$ to obtain the model

$$\begin{aligned}
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^T \boldsymbol{\beta} - (m-r) \boldsymbol{z}_i^T \boldsymbol{\alpha}), \quad r = 1, \dots, m, \\
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^T \boldsymbol{\beta} + (r-m) \boldsymbol{z}_i^T \boldsymbol{\alpha}), \quad r = m+1, \dots, k-1.
\end{aligned} \tag{7.5}$$

The effect is that the intervals between all thresholds are widened by the value $\delta_i = \boldsymbol{z}_i^T \boldsymbol{\alpha}$. In the case of four response categories the model with scaling is equivalent to the basic model without scaling. However, for more than four categories the models differ. We will refer to the model (7.5) as the model with scaling.

The interpretation of parameters is similar to the interpretation of parameters in the basic model. If $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct (7.3) still holds, which means, if x_j increases by one unit the cumulative odds for each category change by the factor e^{β_j}. For the $\boldsymbol{\alpha}$-parameters one obtains

$$\begin{aligned}
e^{-(m-r)\alpha_j} &= \frac{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q))}{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}, \quad r \in \{1, \dots, m\}, \\
e^{(r-m)\alpha_j} &= \frac{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q)}{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}, \quad r \in \{m+1, \dots, k-1\}.
\end{aligned}$$

For adjacent categories holds

$$e^{\alpha_j} = \frac{\gamma(r|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q)) / \gamma(r|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}{\gamma(r-1|\boldsymbol{x}, (z_1, \dots, z_j + 1, \dots, z_q)) / \gamma(r-1|\boldsymbol{x}, (z_1, \dots, z_j, \dots, z_q))}.$$

In the case $x = z$, one obtains now

$$\frac{\gamma(r|x+1)}{\gamma(r|x)} = \begin{cases} e^{\beta} e^{-(m-r)\alpha}, & r \in \{1, \dots, m\} \\ e^{\beta} e^{(r-m)\alpha}, & r \in \{m+1, \dots, k-1\}. \end{cases}$$

In particular for middle category m one obtains again

$$e^{\beta} = \frac{\gamma(m|x+1)}{\gamma(m|x)}.$$

For positive α the value $e^{-(m-r)\alpha}$ is smaller than 1, which means it is a shrinkage factor for categories $r < m$. The value $e^{(r-m)\alpha}$ is greater than 1 and therefore increases the odds ratios for large r.

For the case k odd widening of the intervals between thresholds by a fixed value is more difficult. Let again $m = [k/2] + 1$ denote the the middle category. The widening of the intervals by the value $\delta_i = \boldsymbol{z}_i^T \boldsymbol{\alpha}$ is obtained by

$$\begin{aligned} \theta_r &= \beta_{0r} - [(m-r-1) + 1/2]\boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = 1, \dots, m-1, \\ \theta_r &= \beta_{0r} + [(r-m) + 1/2]\boldsymbol{z}_i^T \boldsymbol{\alpha}, \quad r = m, \dots, k-1. \end{aligned}$$

Again, for $\boldsymbol{x}$ and $\boldsymbol{z}$ distinct (7.3) holds and the interpretation of the β parameters are the same.

7.3. Inference and Computation of Estimates

The strength of the proposed modelling of dispersion effects is that the resulting models can be embedded within the framework of multivariate generalized linear models (GLMs). That means they have the form

$$g(\boldsymbol{\pi}_i) = \boldsymbol{X}_i \boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi}_i = h(\boldsymbol{X}_i \boldsymbol{\beta}),$$

where $\boldsymbol{\pi}_i^T = (\pi_{i1}, \dots, \pi_{ik})$ is the vector the of response probabilities with components $\pi_{ir} = P(Y_i = r|\boldsymbol{x}_i)$, $\boldsymbol{X}_i$ is a design matrix constructed from the predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, $\boldsymbol{\beta}$ is the total parameter vector, $g = (g_1, \dots, g_{k-1}) : \mathbb{R}^{k-1} \to \mathbb{R}^{k-1}$ is a vector-valued *link function* and $h(\cdot) = g(\cdot)^{-1}$ is the response function. The components of the vector $\boldsymbol{X}_i \boldsymbol{\beta}$ are the linear predictors $(\eta_{i1}, \dots, \eta_{i,k-1})$. For details of the representation as a multivariate GLMs see Fahrmeir and Tutz (2001), Tutz (2012). Thus the whole machinery of multivariate GLMs, including algorithms, can be used to obtain estimates and standard errors. Also testing of effects, analysis of residuals and goodness-of-fit tests developed for GLMs can be used.

In a very similar way as for the model in Chapter 6, estimates can be obtained by using the `R` package `VGAM` (Yee, 2010; Yee, 2014). The function `vglm()` allows to estimate various multivariate GLMs (Yee and Wild, 1996). By appropriate specification of the design matrix the proposed location-shift model with dispersion effects can be fitted by using `vglm()`. A proportional odds model as considered here can by specified by the family function `cumulative(reverse=FALSE, parallel=FALSE∼1)`, where the second argument ensures that only the thresholds are category-specific. In the location-shift model the $\boldsymbol{z}$-variables can be seen as a special case of category-specific covariates that differ according to a constant factor (depending on the number of response categories and the type of shifting). For the specification of category-specific covariates argument `xij` can be used when calling `vglm()`. Estimates can easily be obtained after building the design matrix that includes the $\boldsymbol{z}$-variables in the specific form. An `R` function that automatically generates the design matrix and estimates the model is available upon request.

7.4. Application: Confidence in the Federal Government

We consider data from the general social survey of social science, in short ALLBUS, a study by the German institute GESIS. The data is available from `http://www.gesis.org/allbus`. Our analysis is based on a subset containing 2935 respondents of the ALLBUS in 2012. On the basis of this data set the confidence in the healthcare system was already analyzed in an application in Chapter 6. For the present investigation the response is the confidence in the federal government measured on a symmetric scale from 1 (no confidence at all/excessive distrust) to 7 (excessive confidence). As explanatory variables we consider the gender (0: male, 1: female), the income in thousands of Euros, the age in decades (centered at 50) with a linear and a quadratic term and the self reported interest in politics from 1 (very strong interest) to 5 (no interest at all). For modelling we chose category "no" as reference. The deviance of the location-shift model (without scaling) is $10,179.51$. For the model with scaled shifting of thresholds one obtains a remarkably smaller value of $10,140.91$. Hence we will present results for the model with scaling. The likelihood ratio test statistic for the null hypotheses $H_0 : \boldsymbol{\alpha} = \mathbf{0}$ is 54.5 on 8 degrees of freedom and therefore dispersion should definitely be taken into account.

The estimated coefficients and corresponding standard errors of the simple proportional odds model and the location-shift model with scaling are given in Table 7.3. It is seen that for both models the location effects of all four covariates should be included in the model. The location-shift model typically yields estimates that are closer to zero. Among the dispersion effects only the variables gender and political interest obtain large z-values and seem to be needed in the model.

To simplify the interpretation of effects, Figure 7.2 shows the tupel $(e^{\hat{\alpha}}, e^{\hat{\beta}})$ for the linear effects of the model with dispersion. The first value, $e^{\hat{\alpha}}$, represents the multiplicative dispersion effect on the odds. For values larger than one one has larger dispersion, for values smaller than one one has smaller dispersion than in the simple proportional odds model. The second value, $e^{\hat{\beta}}$, represents the multiplicative location effect on the odds obtained by the shifting of the underlying continuous response model. For values larger than one

Table 7.3.: Parameter estimates, standard errors and z-values for the government data.

	Covariate	Proportional Odds Model			Location-Shift Model		
		estimate	se	z value	estimate	se	z value
location effects	Gender	-0.157	0.068	-2.303	-0.138	0.069	-1.991
	Income	-0.076	0.021	-3.510	-0.076	0.025	-3.040
	Age	0.076	0.019	4.009	0.079	0.019	4.127
	Age^2	-0.079	0.010	-7.680	-0.079	0.010	-7.703
	Little	-0.874	0.124	-7.071	-0.693	0.130	-5.311
	Medium	-1.129	0.114	-9.889	-0.960	0.121	-7.939
	Strong	-1.267	0.129	-9.843	-1.098	0.135	-8.143
	Very Strong	-0.892	0.148	-6.030	-0.745	0.154	-4.842
dispersion effects	Gender				0.189	0.040	4.699
	Income				0.013	0.013	1.024
	Age				-0.020	0.011	-1.776
	Age^2				0.005	0.006	0.872
	Little				0.265	0.067	3.948
	Medium				0.255	0.060	4.226
	Strong				0.321	0.071	4.488
	Very Strong				0.075	0.074	1.007

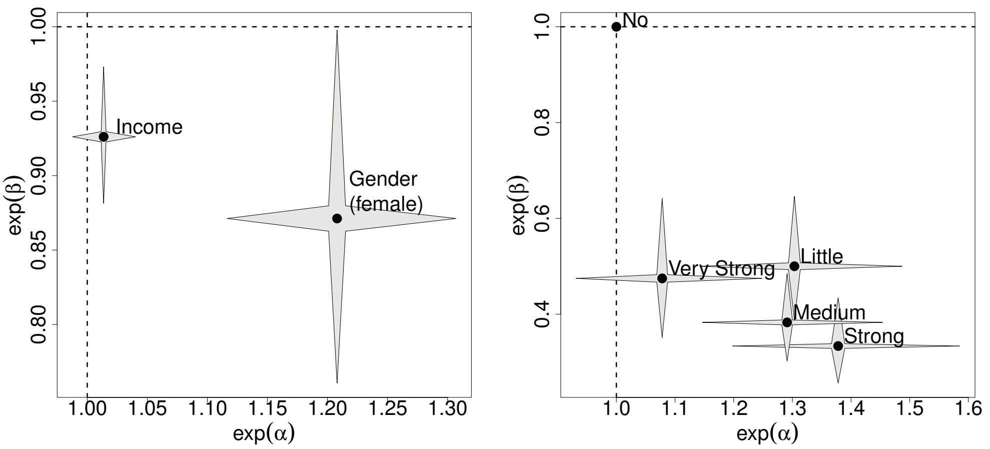


Figure 7.2.: Visualization of estimated effects for the government data including pointwise confidence intervals.

small response categories are favored, for values smaller than one large response categories are favored. The coefficients for gender and income are shown in the left panel, the coefficients for political interest are visualized in the right panel. In Figure 7.2 we also included pointwise 95% confidence intervals that are represented by stars where the horizontal and

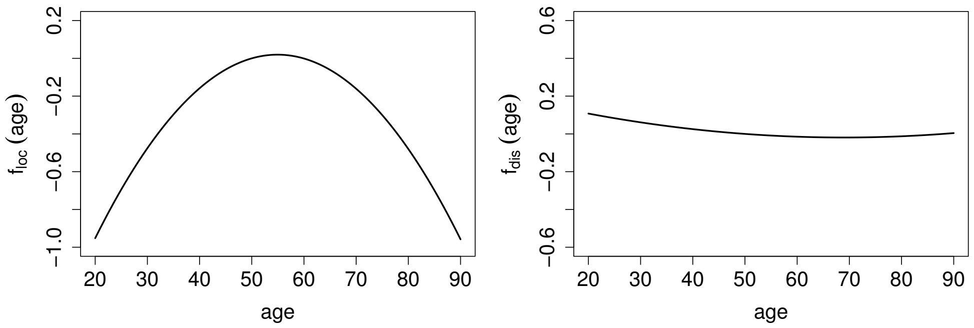


Figure 7.3.: Non-linear location (left) and dispersion (right) effects for covariate age for the government data.

vertical lengths corresponds to the confidence intervals of $e^{\hat{\alpha}}$ and $e^{\hat{\beta}}$, respectively. From the left panel it can be seen that females tend to choose higher response categories and therefore show a higher confidence in the government than males. At the same time they show smaller dispersion than males, responses are more concentrated in the middle. The confidence also increases with increasing income. However, the dispersion effect is very close to one and can be neglected. The right panel shows that the confidence is higher among all respondents that had at least some political interest. Furthermore, respondents that did not choose one of the extremes (“no“ or “very strong interest“) show reduced dispersion. This could be interpreted as a response style, as considered in Chapter 6. People who tend to choose middle categories have the same tendency in all questions.

For non-linear effects as the effect of age, star plots as in Figure 7.2 are not useful. Therefore, in Figure 7.3 the fitted non-linear location (left) and dispersion (right) effects of the variable age, denoted by $f_{loc}(age)$ and $f_{dis}(age)$, are given as a function of age. The dispersion effect is not significant, nevertheless, for illustration we show the corresponding curve. The location curve in the left panel shows that confidence is weakest at about 55 years of age but is definitely stronger for younger and older persons.

7.5. Comparison of Models and Consequences of Ignored Dispersion

Varying dispersion can be modelled by the proposed location-shift model but also by the location-scale model (7.2). In the location-shift model the dispersion is modelled by an explicit shifting of the thresholds which is determined by the parameter $\boldsymbol{\alpha}$. In the location-scale model the dispersion is generated by the variance $\tau_{\boldsymbol{x}} = \exp(\boldsymbol{x}^T\boldsymbol{\gamma})$ of the underlying continuous regression model. The effect, determined by the parameter $\boldsymbol{\gamma}$, is now multiplicative on the thresholds since the predictor has the form $\eta_r = \gamma_{0r}\exp(-\boldsymbol{x}^T\boldsymbol{\gamma}) + \boldsymbol{x}^T\boldsymbol{\beta}\exp(-\boldsymbol{x}^T\boldsymbol{\gamma})$.

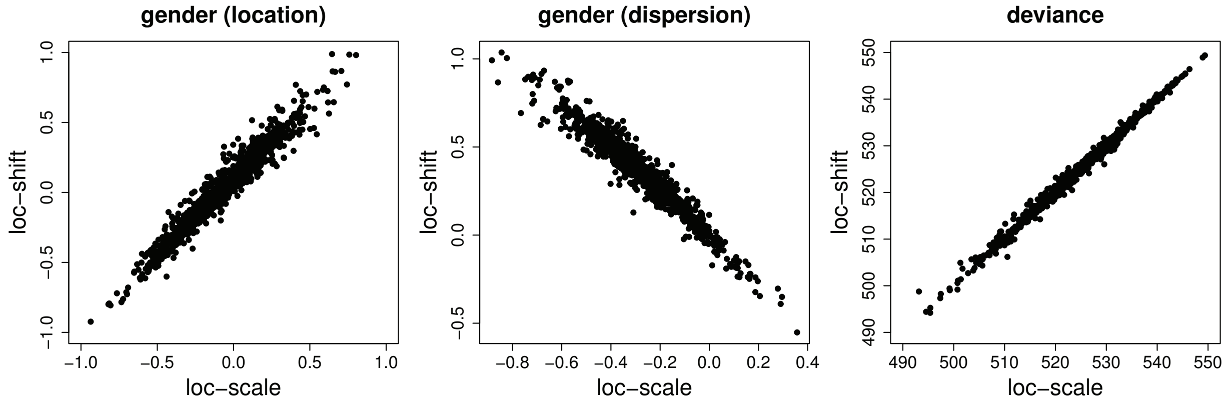


Figure 7.4.: Parameter estimates and deviances of model fits for sub samples of size $n = 200$ from the eye vision data.

Moreover, the dispersion also modifies the location term. Although the models are not equivalent in applications we found the differences in terms of goodness-of-fit can be rather small.

For illustration we first consider the eye vision data example. We draw sub samples of size $n = 200$ from the data set and computed the location effect, the dispersion effect and the deviances of the location-scale (abbreviated by *loc-scale*) and the location-shift (abbreviated by *loc-shift*) model. As it is seen from Figure 7.4 the estimates and deviances of the two models show strong correlation. In particular the deviances of the two models are very close. Therefore, in cases with almost no location effect the models yield similar estimates and goodness-of-fit measures.

Since in the eye vision data example the data generating model is not known, we illustrate the fitting in a small simulation study in which the data generating models are known. We consider two binary covariates with $\boldsymbol{\beta}^T = (0.5, 0.5)$, $k = 5$ response categories and thresholds $\theta_r \in \{-2, \dots, 2\}$. First data are generated by the location-scale model with varying strength of dispersion in the first variable. Then the location-scale and the location-shift model are fitted. The first row of Figure 7.5 shows the resulting deviances. In order to match the strength of dispersion we computed the parameter α of the location-shift model that shows approximately the same dispersion as the corresponding parameter γ of the location-scale model. The relation between these two parameters is non-linear, large values of α correspond to small values of γ. Then data were generated by the location-shift model and again both models are fitted. The resulting deviances are shown in the second row of Figure 7.5. It is seen that the deviances of the two models are quite close with just slightly better fits of the data generating model. If, however, the dispersion is ignored and a simple proportional odds model (abbreviated by *no disp*) is fitted, the fit suffers strongly.

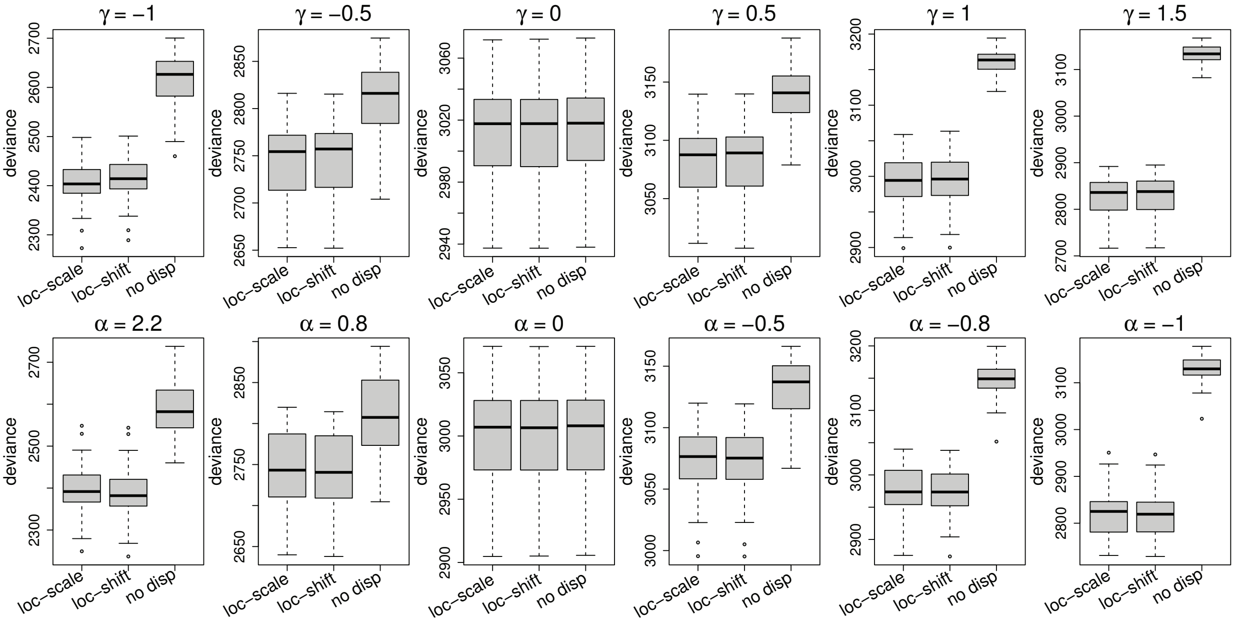


Figure 7.5.: Deviances of model fits for data generated by the location-scale model (first row) and data generated by the corresponding location-shift model (second row).

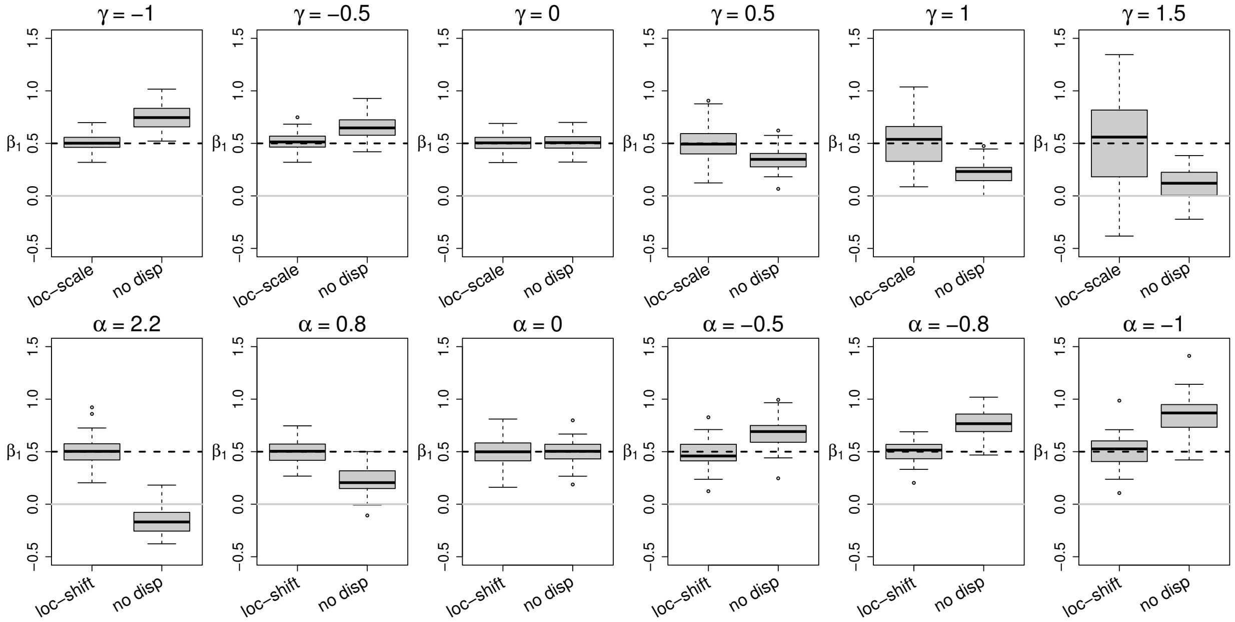


Figure 7.6.: Estimates of the location parameter β_1 for data generated by the location-scale model (first row) and data generated by the corresponding location-shift model (second row).

If strong variation is present the omission of corresponding effects might not only yield large deviances but also reduce the accuracy of the estimates of the location effect. This effect is illustrated by using the same data generating model as before but now with a focus on the estimation of the first parameter. Figure 7.6 shows the estimates of the location effect β_1.

In the first row the location-scale model was the data generating model, in the second row the location-shift model. It is seen that there is no bias if no dispersion effect is present. However, with increasing dispersion the estimates are biased.

In both models we used $\beta_1 = 0.5$. However, one should be aware that the parameters cannot compared directly since they represent different effects in the two models. In the location-scale model the predictor has the form $\eta_r = \gamma_{0r}/\exp(\boldsymbol{x}^T\boldsymbol{\gamma}) + \boldsymbol{x}^T\boldsymbol{\beta}/\exp(\boldsymbol{x}^T\boldsymbol{\gamma})$. In particular, the dispersion is also included in the location term. For a simple binary predictor $x \in \{0, 1\}$, the location term is $\boldsymbol{x}^T\boldsymbol{\beta}/\exp(\boldsymbol{x}^T\boldsymbol{\gamma}) = x\beta/\exp(x\gamma)$, which for $x = 1$ takes the value $\beta/\exp(\gamma)$. Thus, if one ignores the possible variation and fits a model that does not account for it one estimates the parameter $\beta/\exp(\gamma)$ instead of β. Therefore, if γ is positive one can expect a bias towards zero, if γ is negative, one will overestimate the strength of the location effect. This effect is seen from the first row of Figure 7.6. The bias can be severe if γ is large, for example, if $\gamma = 1.5$, estimates are very close to zero, which is not surprising since $\beta/\exp(\gamma) = 0.5/4.48 = 0.11$. In the location-shift model the tendency of the bias is different. As is seen from the second row small values of α (stronger dispersion) yield stronger location effects. For positive values of α the estimated effects are weaker ($\alpha = 0.8$), for large values of α ($\alpha > 2$), however, even the sign of the effect changes. The effects are similar if one considers negative values of β_1 (not shown). Overall, it is seen that ignoring dispersion effects may yield strongly biased estimates.

7.6. Non-Symmetric Responses

In the previous section we considered symmetric responses, which often occur in survey data if the extent of the agreement to a statement is evaluated. However, also non-symmetric responses may show dispersion that varies over sub populations.

7.6.1. Modelling Varying Dispersion in Non-Symmetric Responses

The dispersion modelled so far means varying variability centered at a middle category, which is quite natural for a symmetric response. For non-symmetric responses one may pick a category m and model the variability with a centering between m and $m + 1$ as in model (7.5).

For distinct variables $\boldsymbol{x}$ and $\boldsymbol{z}$ the interpretation of the parameters is the same as in model (7.5) because the derivation of the parameters does not depend on the chosen m. Thus one has several models depending on the chosen category m. The goodness-of-fit of the model measured by the deviance can be used to select a model. It turned out that the estimated location effect $\boldsymbol{\beta}$ depends very weakly on the choice of m whereas the values of the dispersion effects $\boldsymbol{\alpha}$ do depend on m.

However, the case $\boldsymbol{x} = \boldsymbol{z}$ is different. Then it does not matter which category m is chosen, all models (7.5) with any fixed m are equivalent. The only difference is in the interpretation of parameters. The equivalence is seen by transforming the parameters. Let $\beta_{r0}^{(m)}, \boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\alpha}^{(m)}$ denote the parameters of the model (7.5) for fixed category m. It can be shown that for two values m and l

$$\beta_{r0}^{(m)} = \beta_{r0}^{(l)}, r = 1, \dots, k-1, \quad \boldsymbol{\alpha}^{(m)} = \boldsymbol{\alpha}^{(l)}, \quad \boldsymbol{\beta}^{(m)} = \boldsymbol{\beta} + (m-l)\boldsymbol{\alpha}^{(l)}.$$

That means, the intercepts and the dispersion parameters $\boldsymbol{\alpha}$ do not depend on the choice of m. The only parameters that depend on the choice of m are the $\boldsymbol{\beta}$ parameters, and the transformation uses the $\boldsymbol{\alpha}$ parameters. To obtain the interpretation as dispersion parameters again a middle category is a good choice because one obtains

$$\frac{\gamma(r|x+1)}{\gamma(r|x)} = \begin{cases} e^{\beta} e^{-(m-r)\alpha}, & r \in \{1, \dots, m\} \\ e^{\beta} e^{(r-m)\alpha}, & r \in \{m+1, \dots, k-1\}, \end{cases}$$

and in particular

$$e^{\beta} = \frac{\gamma(m|x+1)}{\gamma(m|x)}.$$

That means e^{β} refers to the increase of x by one unit for the fixed category m and α is determined by

$$e^{-\alpha} = \frac{\gamma(r|x+1)/\gamma(r|x)}{\gamma(r+1|x+1)/\gamma(r+1|x)}.$$

7.6.2. Application: Knee Injuries

As an application we consider data from a clinical trial ($n = 127$) that investigates the effect of a therapy on the recovery of knee injuries. The response is the pain during movement measured on a scale from 1 (no pain) to 5 (severe pain), for more details see Tutz (2012). We model the treatment effect (1: therapy, 0: placebo) and the effect of the covariate age in years with a linear and a quadratic effect.

The estimated coefficients and corresponding standard errors for the simple proportional odds model and the location-shift model with scaled shifting of thresholds and $m = 3$ are given in Table 7.4. For the simple proportional odds model the deviance is 362.9 on 501 degrees of freedom and for the models with location and dispersion effects the deviance is 356.3 on 498 degrees of freedom. There are significant location effects for treatment and the linear and the quadratic effect of age. Concerning the dispersion part only the treatment

Table 7.4.: Parameter estimates and standard errors and z-values for knee injury data.

	Covariate	Proportional Odds Model			Loc-Shift Model (m=3)		
		estimate	se	z value	estimate	se	z value
	Intercept1	2.541	1.940	1.309	3.980	2.250	1.769
	Intercept2	3.803	1.957	1.943	3.564	2.011	1.773
	Intercept3	4.809	1.971	2.440	3.059	2.598	1.177
	Intercept4	6.823	2.016	3.385	3.729	3.669	1.017
location effects	Treatment	0.938	0.331	2.834	1.309	0.372	3.513
	Age	-0.372	0.129	-2.871	-0.345	0.149	-2.312
	Age^2	0.006	0.002	3.006	0.006	0.002	2.437
dispersion effects	Treatment				0.636	0.254	2.508
	Age				0.032	0.094	0.343
	Age^2				-0.001	0.002	-0.194

effect with estimate $\hat{\alpha}_{treat} = 0.636$ seems to be relevant. The inclusion of dispersion effects yields a stronger location effect of the variable treatment.

7.7. Partial Proportional Odds Models versus the Modelling of Dispersion

If the proportional odds model does not fit the data well, one strategy is to introduce category-specific parameters, which corresponds to use the partial proportional odds model. The other option, which is proposed here, is to include dispersion effects. Both modelling strategies will yield a better fit. In the following we briefly consider these two options.

An interesting case is the modelling of three response categories ($k = 3$) and $\boldsymbol{x} = \boldsymbol{z}$. Then the two predictors of the location shift model are $\eta_1 = \beta_{01} + \boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\alpha}$ and $\eta_2 = \beta_{02} + \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{x}^T\boldsymbol{\alpha}$, which is the same as the reparameterized predictors $\eta_r = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}_r$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta} - \boldsymbol{\alpha}$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta} + \boldsymbol{\alpha}$. Therefore, the location-shift model is equivalent to the partial proportional odds model. Nevertheless, there are some benefits when using the location-shift parameterization. If the hypothesis $H_0 : \alpha_j = 0$ holds the j-th variable has global and not category-specific effects. The test result is immediately seen from the $z-$ or p-value of the corresponding parameter. Within the partial proportional odds model, one has to test the hypothesis $H_0 : \beta_{j1} = \beta_{j2}$ to investigate if the j-th variable has global effects, which typically makes refitting of the model under constraints necessary. This is illustrated in a small example.

Table 7.5.: Parameter estimates, standard errors and z-values for the retinopathy data.

	Covariate	Proportional Odds Model			Location-Shift Model		
		estimate	se	z value	estimate	se	z value
location effects	SM	-0.254	0.191	-1.328	-0.159	0.198	-0.802
	DIAB	-0.139	0.013	-10.368	-0.148	0.014	-10.524
	GH	-0.459	0.074	-6.175	-0.485	0.076	-6.324
	BP	-0.072	0.013	-5.357	-0.071	0.014	-5.204
dispersion effects	SM				0.491	0.235	2.087
	DIAB				-0.037	0.016	-2.254
	GH				-0.101	0.092	-1.099
	BP				-0.007	0.015	-0.465

	Covariate	Partial Proportional Odds Model		
		estimate	se	z value
	SM1	-0.405	0.205	-1.972
	SM2	0.086	0.254	0.340
	DIAB1	-0.129	0.014	-8.889
	DIAB2	-0.166	0.018	-9.264
	GH1	-0.435	0.080	-5.426
	GH2	-0.535	0.097	-5.470
	BP1	-0.068	0.014	-4.627
	BP2	-0.075	0.017	-4.432

7.7.1. Application: Retinopathy

In a 6-year followup study on diabetes and retinopathy status reported by Bender and Grouven (1998) the interesting question is how the retinopathy status is associated with risk factors. The considered risk factor is smoking (SM = 1: smoker, SM = 0: non-smoker) adjusted for the known risk factors diabetes duration (DIAB) measured in years, glycosylated hemoglobin (GH), which is measured in percent, and diastolic blood pressure (BP) measured in mmHg. The response variable retinopathy status has three categories (1: no retinopathy; 2: nonproliferative retinopathy; 3: advanced retinopathy or blind). The simple proportional odds model yields deviance 904.14, the model with category-specific intercepts yields 892.45, the same as the the location-shift model. The difference, 11.69, on 4 df shows that at least some of the parameters should be category-specific. From the fitted parameters of the location-shift model (Table 7.5) one sees immediately that smoking and DIAB are susceptible of having category-specific effects but not GH and BP. This is not seen from the estimates of the category-specific model.

The location-shift model also provides a different interpretation of the effects of smoking and DIAB. In the location-shift model DIAB shows a strong shifting effect and also varying dispersion. Smoking shows no significant shifting effect, also in the simple proportional

Table 7.6.: Parameter estimates, standard errors and z-values for the election data.

	Covariate	**Proportional Odds Model**			**Location-Shift Model**		
		estimate	se	z value	estimate	se	z value
location effects	Gender	0.628	0.088	7.137	0.583	0.091	6.391
	Age	-0.012	0.002	-4.385	-0.013	0.002	-4.425
	Age^2	0.001	0.001	5.041	0.001	0.001	4.966
	College	-1.419	0.095	-14.864	-1.466	0.105	-13.954
	Home	-0.410	0.096	-4.234	-0.432	0.097	-4.415
	Length	-1.134	0.149	-7.596	-1.212	0.155	-7.796
dispersion effects	Gender				0.141	0.072	1.966
	Age				-0.001	0.002	-0.445
	Age^2				0.001	0.001	1.788
	College				0.108	0.085	1.279
	Home				0.176	0.077	2.266
	Length				0.217	0.122	1.772

odds model the effect is not significant. In the category-specific model smoking for the first split into categories 1 and $\{2, 3\}$ seems to be substantial (z-value -1.972) but not for the other split into categories {1,2} and 3. Within the location-shift model this is explained by a different dispersion over response categories for smokers and non-smokers.

7.7.2. Application: Information about Politics

Finally, we consider an application in which the extension to category-specific effects seems not necessary, however, dispersion effects are present. We use data from the American National Election Study `http://www.electionstudies.org/` containing 1790 respondents from the study in 2000, see Jackman (2009). The response is on an ordinal rating scale that represents the general level of information about politics and public affairs from 1 (very low) to 5 (very high). The obtained level was assessed by the interviewer assigned to each respondent. Explanatory variables are gender (0: male, 1: female), age (centered at 47), college degree (College; yes/no), if the respondent or his family owns their home (Home) and the length of the interview (on a log scale).

When fitting a simple proportional odds model one obtains 4891.198 on 7150 df. To evaluate if effects are really global we fitted a model with category-specific effects. The difference in deviances between the two models is 24.42 on 18 df. Therefore, one can assume that no category-specific effects are needed and the simple proportional odds model seems appropriate. However, one might also investigate if there are dispersion effects. We fitted a location-shift model with only six additional parameters (dispersion effects) to obtain the

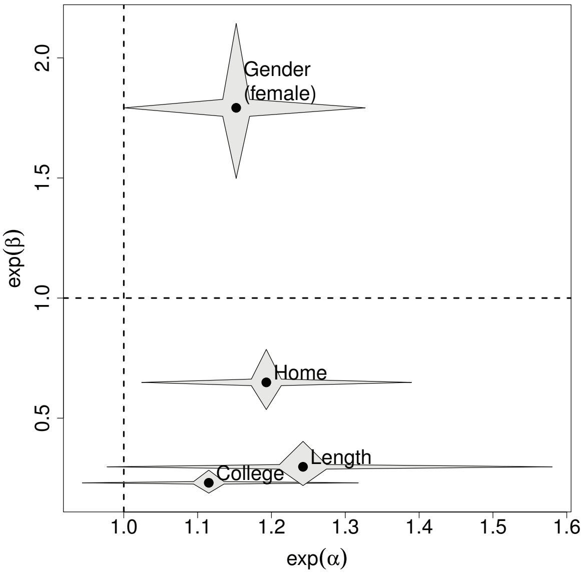


Figure 7.7.: Visualization of estimated effects of the location-shift model for the election data including pointwise confidence intervals.

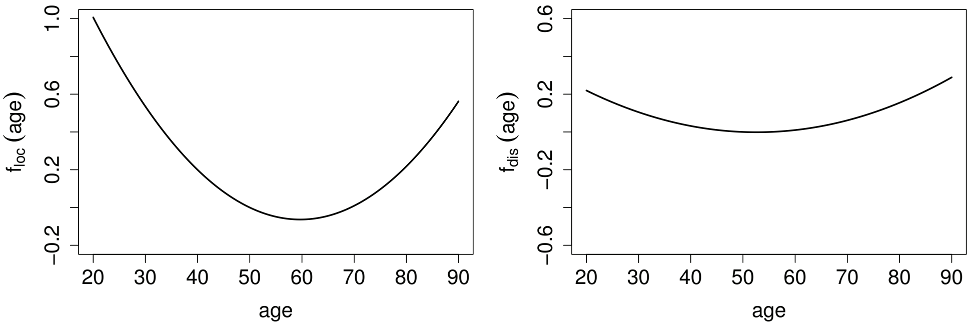


Figure 7.8.: Non-linear location (left) and dispersion (right) effects of the location-shift model for the election data for covariate age.

deviance 4873.526 on 7144 df. When comparing to the simple proportional odds model now one obtains a difference in deviances of 17.636 on 6 df, which indicates that dispersion effects are present. The fitted parameters and corresponding z-values of the location-shift model, given in Table 7.6, show that the location effects of all variables should be included in the model. Among the dispersion effects the two variables gender and home seem to be relevant. From Figure 7.7, which visualizes 95% confidence intervals, it is seen that females seem to be less informed about politics and show weaker dispersion. Respondents who own their home also show weaker dispersion but are better informed about politics. Figure 7.8 shows the non-linear effects of the variable age. The location curve (left panel) shows that the level of information is highest at about 60 and much lower for younger and older respondents.

7.8. Concluding Remarks

An alternative model for the explicit inclusion of dispersion effects is proposed. In terms of goodness-of-fit the model is frequently quite similar to the location-scale model. Nevertheless the model has some advantages. It can be embedded into the framework of generalized linear models and therefore all the inference techniques and asymptotic results that have been shown to hold for this class of models can be used. The interpretation of parameters differs from that of the parameters of the location-scale model. When interpreting parameters of the location-scale model one typically refers to the underlying latent regression model. While the proportional odds model without dispersion can also be fitted and interpreted without referring to the latent model, with dispersion, however, it seems unavoidable to refer to the latent model. In contrast, parameters of the location-shift model can be interpreted straightforward in terms of log-odds.

We also investigated alternative modelling strategies. One may extend simple models with global effect to more flexible models like the partial proportional odds model or examine if dispersion effects as in the location-scale or location-shift model are present. The former strategy may yield models that are much harder to interpret. Some authors argue that simpler models as the proportional odds model are often to be preferred even if the fit is not too good because the obtained first-order effects are often informative for overall summaries that explain the most important dimension of an effect (Agresti, 2009). The second strategy, investigating if dispersion effects are needed, has the advantage that the first-order effects concerning the location are kept and summary measures concerning the location are still available. In addition, if dispersion effects are present estimates of the location effects will be less biased.

8. Conclusion and Outlook

This thesis is dedicated to regression models for categorical variables that either serve as the responses or part of the covariates. In each chapter generalized linear models are adapted to specific problems, which results in tailored solutions with high flexibility. In this concluding chapter the most important results are summarized and possible further research is briefly discussed.

Detection of Latent Groups

In the first part of the thesis an approach for the detection of latent groups in regression models with an excessive number of parameters is discussed. The proposed model is composed of two parts, a tree component and a linear or additive component. In accordance to Chapter 2 the model containing a linear term and categorical predictors $\boldsymbol{z}$ has the form

$$\eta_i = tr(\boldsymbol{z}_i) + \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

For several categorical predictors the tree component $tr(\boldsymbol{z}_i)$ is composed of single trees for each variable. Therefore, the model is designed to find clusters of categories in single components. Following the notation from Chapter 3 the corresponding model accounting for heterogeneity in longitudinal or cross-sectional studies can be written as

$$\eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + tr(\boldsymbol{z}_{ij}).$$

Again the model finds clusters of measurement units that share the same effect on the response by treating each group-specific component in $\boldsymbol{z}$ separately. For models with group-specific intercepts the tree component $tr(z_{ij})$, $z_{ij} = 1$ consists of one single tree only. Extensive simulations and various applications demonstrate the potential of the methods. Main advantages over competing methods are the improved clustering performance and the computational efficiency.

A more general approach is to define a model with a tree component in the sense of traditional recursive partitioning as applied in Chapter 4 and Chapter 5. Let now $\boldsymbol{x}$ and $\boldsymbol{z}$ define two sets of covariates that can be from different scales. Then the general model with predictor $\eta = \boldsymbol{x}^\top \boldsymbol{\beta} + tr(\boldsymbol{z})$ can be composed of a familiar tree that is fitted for the $\boldsymbol{z}$-variables and a familiar linear term that is fitted for the $\boldsymbol{x}$-variables. Although the predictor of the model is different from those in Chapter 2 and Chapter 3 the same strategies for the selection of splits and the splitting decision can be applied. By construction the model includes relevant interactions between the the $\boldsymbol{z}$-variables and focusses on the main effects of the $\boldsymbol{x}$-variables. Therefore, if many covariates are available it might be quite challenging to decide which ones to include in which part. Consequently, selection strategies to separate covariates with a linear and smooth effect have to be developed. Similar modelling strategies with a focus on specific applications have been proposed by Chen et al. (2007) and Yu et al. (2010).

Another quite interesting generalization is to exploit the flexible structure of the predictor to model varying-coefficient models, see Hastie and Tibshirani (1993). Consider three continuous covariates x_1, x_2 and x_3. The simplest way to determine the response is to use a regression model that includes only the main effects of the three variables. However, the impact of x_3 on the response might depend on x_1. For example, the effect of x_3 might be different for two groups defined by x_1 and a corresponding split-point c_1. Then one yields the model with predictors

$$\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \beta_{31}x_{i3}I(x_{i1} \leq c_1) + \beta_{32}x_{i3}I(x_{i1} > c_1),$$

where β_{31} and β_{32} are the effects of variable x_3 for the two groups defined by the so-called effect modifier $x - 1$. Furthermore, if the effect of variable x_3 in region $\{x_1 \leq c\}$ additionally depends on x_2, a further split with regard to split-point c_2 yields the two daughters $x_{i3}I(x_{i1} \leq c_1)I(x_{i2} \leq c_2)$ and $x_{i3}I(x_{i1} \leq c_1)I(x_{i2} > c_2)$. The resulting model is composed of a linear component containing the main effects of x_1 and x_2, and a tree component containing different effects of x_3. The tree component represents subgroups defined by x_1 and x_2 that differ with regard to their linear effect of x_3. After several splits, the predictor of the model can be written as

$$\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + tr_{x_{i1},x_{i2}}(x_{i3}).$$

At the same time it is also possible that the main effects of x_1 or x_2 depend on the respective other variables. Consequently, for each variable that is modified one obtains a single tree. The algorithm proposed in Chapter 2 can easily be adapted to this kind of models using the same strategies for the selection of splits and the splitting decision. The algorithm simultaneously detects the variables that have to be modified and the effect modifiers as well as corresponding subgroups that are responsible. Moreover, the approach allows to combine continuous and categorical effect modifiers. For continuous effect modifiers one typically assumes smooth functions that can be modelled by splines, see, for example Hoover et al. (1998) and Lu et al. (2008). Regularization methods for the selection of effect modifiers in varying-coefficients models seem to be scarce, yet. A tree-based solution with the focus on quality of life research in breast cancer studies was proposed by Su et al. (2011).

Modelling of Latent Traits

In the second part of this thesis item focussed trees for the detection of uniform and non-uniform DIF are considered. The proposed methods simultaneously detect the items and the corresponding subgroups of persons that are responsible for DIF. Chapter 4 focusses on the detection of uniform DIF based on the Rasch model. The proposed model that accounts for DIF has the closed form

$$\eta_{pi} = \theta_p - \mathrm{tr}_i(\boldsymbol{x}_p),$$

when person p and item i are considered. The tree component $tr_i(\boldsymbol{x}_p)$ defines regions of the covariate space that have to be distinguished with respect to their item difficulty. If an item is free of DIF, it is compatible with the Rasch model and the constant $\mathrm{tr}_i(\boldsymbol{x}_p) = \beta_i$ is fitted. Various simulations and comparisons to competing methods illustrate the good performance and the advantages of the proposed method.

As already outlined in Section 4.6, the proposed item focussed trees can straightforward be extended to polytomous items by use of the PCM. By appropriate specification of the design matrix the PCM can be embedded into the framework of multivariate generalized linear models. Estimates can easily be obtained by use of the R-package `VGAM` (Yee, 2010; Yee, 2014). Therefore, the basic algorithm described in Section 4.3.1 can be applied in the same way. As before, the model yields a single tree for each DIF item. In the non-homogeneous case one obtains a different set of item parameters for each node without any restrictions on the parameters. This assumption is particularly interesting if one suspects different response patterns among different groups. In the homogeneous case the difference in item parameters is determined by a constant shifting. Further research is needed to set up appropriate software and to investigate the performance, in particular, compared to the method proposed by El-Komboz et al. (2014).

In Chapter 5 the logistic regression approach proposed by Swaminathan and Rogers (1990) is extended to detect uniform and non-uniform DIF by item focussed trees. In particular in the non-uniform DIF case well reasoned estimation strategies are required. The benefits over traditional approaches are shown in simulations and chosen applications.

The algorithm that yields item focussed trees is mainly characterized by:

- Selection of the best splits by likelihood ratio tests.

- Use of maximal value statistics to determine splitting decisions.

- Use of permutation tests to obtain splitting decisions.

These well chosen components of the algorithm ensure that the selection of splits is separated from the splitting decision and result in an unbiased recursive partitioning scheme as similarly proposed by Hothorn et al. (2006). However, a disadvantage of the approach is the time requirement for the computation of the likelihood ratio test statistics and the permutation tests. Alternatively one could use score test statistics, which has the advantage that only the model under the null hypothesis has to be evaluated. Moreover, permutation tests could be saved if the distribution of the selection process were known, that is the asymptotic distributions of the maximal value statistics. These adjustments are certainly worth investigating in future research.

Finally, the estimation strategy for the Rasch model used in Chapter 4 has to be addressed. By appropriate definition of the design matrix the Rasch model can be embedded into the framework of generalized linear models and joint maximum likelihood (JML) estimates can easily be obtained. This strategy is applied in Chapter 4. As with JML the number of parameters simultaneously increases with the number of persons, two major problems arise. First, the estimation of the model is computationally expensive and unstable in high dimensional settings. Second, for a fixed number of items the estimates for the item difficulties that are the parameters of interest in most applications are inconsistent for $P \to \infty$, see, for example, Anderson (1973). Alternative strategies that do not face these problems are conditional maximum likelihood (CML) estimation and marginal maximum likelihood (MML) estimation. CML makes use of the property that the test score, i.e. the number of solved items, is sufficient for the ability of a person. By conditioning on the test score the person parameters do not occur in the conditional likelihood. MML, on the other hand, assumes that the person parameters are drawn from a normal distribution $N(0, \sigma^2)$. The resulting marginal log likelihood can, for example, be solved by numerical integration, see Hatzinger (1989) for details about the estimation procedures. The use of alternative estimation strategies certainly improves the existing approach but further research is needed to incorporate appropriate tools into the framework of item focussed trees.

Detection of Latent Response Styles

In the third part of this thesis ordinal regression models are extended to account for responses that are characterized by a disproportionate tendency to the middle or the highest and lowest response categories. A strong tendency to the middle or extreme categories can be seen as a specific response style or interpreted as varying dispersion. The linear predictor of two models, the adjacent categories and the cumulative model, are extended by additive terms $\boldsymbol{z}_i^\top \boldsymbol{\gamma}$ or $\boldsymbol{z}_i^\top \boldsymbol{\alpha}$ that determine the response style or the dispersion . These effects are caused by the set of variables $\boldsymbol{z}$. The effects are clearly separated from the content-related effects that are simultaneously determined by the same or a different set of covariates $\boldsymbol{x}$.

The additional effects can be seen as a special case of category-specific covariates. Therefore, the estimates of the models can be obtained by the use of existing software. The visualization of effects makes the results of the models easy accessible. This is illustrated in several applications. Moreover, the benefits of the extended models are demonstrated in simulations, where strongly biased estimates of the content-related effects are observed if a present response style is ignored.

The extended adjacent categories model proposed in Chapter 6 can also be used to model the response style for more than one item (see also Section 6.8). A popular choice, which was also used for the detection of DIF, is the PCM. Next to the item parameters δ_{ir} the extended PCM contains two person-specific parameters, namely, the ability parameter θ_p and the response style parameter γ_p. Again, the response style parameter γ_p can optionally be modelled as a function of explanatory variables $\boldsymbol{x}_p$ with a linear or non-linear effect on the response. By appropriate assumptions the extended PCM can be estimated by the maximization of the joint or the marginal likelihood. The main advantage over previous approaches for item response data is that the response style is explicitly modelled. In latent class approaches, for example, it might be quite challenging to determine the number of classes and to interpret the resulting effects. Further research is needed to develop specific software and to investigate the performance of the method.

A quite different approach for the modelling of ordinal variables generated by rating scales are mixture type models introduced by Piccolo (2003). The basic concept of these models is that the choice of a response category is determined by a mixture of the preference of a person and the persons indecision. The two components are usually referred to as feeling and uncertainty. They are both modelled by different distributions that have to be defined appropriately. Hence the mixture provides high flexibility. For example, in so-called CUB models the first component is modelled by a binomial distribution and the latter by a uniform distribution. For an overview on CUB models, see Iannario and Piccolo (2012). More recently, an extended class of mixtures was proposed by Tutz et al. (2016), where the preference component is determined by a cumulative or adjacent categories model. The tendency to the middle or extreme categories can be interpreted as a special form of uncertainty. Thus, by an appropriate choice for the distribution of the uncertainty component mixture models should also be able to capture extreme response styles. The evaluations of these class of models and the comparisons to the proposed methods might be very interesting and is worth considering in future research.

In summary, this thesis provides a variety of modelling strategies for the the detection of latent structures with a focus on categorical variables. However, there are still several limitations that require further research and development of the approaches. As the methods can be used in many areas of application, also further interactions to related subjects should be investigated in future research.

Appendices

A. Overview on Variables of the Applications in Chapter 2

Table A.1.: Districts in the city of Munich. The numbers correspond to the labels in Table 2.1.

Number	District
1	Altstadt-Lehel (inner city)
2	Ludwigsvorstadt-Isarvorstadt
3	Maxvorstadt
4	Schwabing-West
5	Au-Haidhausen
6	Sendling
7	Sendling-Westpark
8	Schwanthalerhöhe
9	Neuhausen-Nymphenburg
10	Moosach
11	Milbertshofen-Am Hart
12	Schwabing-Freimann
13	Bogenhausen
14	Berg am Laim
15	Trudering-Riem
16	Ramersdorf-Perlach
17	Obergiesing-Fasangarten
18	Untergiesing-Harlaching
19	Thalkirchen-Obersendling-Forstenried-Fürstenried-Solln
20	Hadern
21	Pasing-Obermenzing
22	Aubing-Lochhausen-Langwied
23	Allach-Untermenzing
24	Feldmoching-Hasenbergl
25	Laim

Table A.2.: German country code listed as in the ISO 3166-2.

Abbreviation	Country
BB	Brandenburg
BE	Berlin
BW	Baden-Wuerttemberg
BY	Bavaria
HB	Bremen
HH	Hamburg
HE	Hesse
NI	Lower Saxony
MV	Mecklenburg-Vorpommern
NW	North Rhine-Westphalia
RP	Rhineland-Palatinate
SL	Saarland
SN	Saxony
ST	Saxony-Anhalt
SH	Schleswig-Holstein
TH	Thuringia

Table A.3.: Categories of the nominal variable kind of household.

Number	Kind of Household
1	Single-Person Household
2	Couple Without Children
3	Single Parent
4	Couple with Children aged ≤ 16
5	Couple With Children aged > 16
6	Couple With Children aged ≤ 16 and > 16
7	Multiple Generation Household
8	Other Combination

B. Tabular Display of Simulation Results for Chapter 3

In the following we give the results of all settings of the simulations described in Section 3.6. Each table contains the MSEs of the unit-specific intercepts, the MSEs of the linear term and the selected number of clusters as the average of 100 replications, respectively.

Table B.1.: Average results for the settings with normal response, normal distributed intercepts and $\rho = 0$.

		MSE - intercepts		MSE - linear term		Number of Clusters	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 200$	GFM	2.26	2.26	0.04	0.04	200.00	200.00
$n_i = 4$	GMM	0.68	0.71	0.03	0.03	200.00	200.00
	TSC	1.56	1.57	0.04	0.04	4.96	5.02
	PEL						
	FINA	1.05	1.10	0.03	0.03	1.89	1.91
	FINB	0.99	1.06	0.03	0.03	1.31	1.36
$n = 100$	GFM	1.14	1.14	0.03	0.03	100.00	100.00
$n_i = 8$	GMM	0.54	0.56	0.03	0.03	100.00	100.00
	TSC	0.97	0.99	0.03	0.03	5.28	5.38
	PEL						
	FINA	0.82	0.87	0.03	0.03	2.04	2.10
	FINB	0.86	0.91	0.03	0.03	1.67	1.72
$n = 40$	GFM	0.45	0.45	0.03	0.03	40.00	40.00
$n_i = 20$	GMM	0.31	0.32	0.03	0.03	40.00	40.00
	TSC	0.44	0.46	0.03	0.03	5.82	6.00
	PEL	0.37	0.38	0.03	0.03	15.00	15.06
	FINA	0.53	0.55	0.03	0.03	2.27	2.44
	FINB	0.57	0.61	0.03	0.03	1.86	1.98
$n = 20$	GFM	0.22	0.22	0.03	0.03	20.00	20.00
$n_i = 40$	GMM	0.19	0.19	0.03	0.03	20.00	20.00
	TSC	0.23	0.24	0.03	0.03	5.76	6.00
	PEL	0.21	0.21	0.03	0.03	9.95	9.99
	FINA	0.32	0.34	0.03	0.03	2.45	2.66
	FINB	0.39	0.43	0.03	0.03	1.96	2.06

Table B.2.: Average results for the settings with normal response, normal distributed intercepts and $\rho = 0.8$.

		MSE - intercepts		MSE - linear term		Number of Clusters	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 200$	GFM	2.28	2.28	0.05	0.05	200.00	200.00
$n_i = 4$	GMM	0.88	0.95	0.29	0.32	200.00	200.00
	TSC	1.51	1.53	0.08	0.08	4.86	4.95
	PEL						
	FINA	0.95	1.01	0.30	0.34	1.14	1.10
	FINB	0.92	0.98	0.30	0.34	1.00	1.00
$n = 100$	GFM	1.16	1.16	0.04	0.04	100.00	100.00
$n_i = 8$	GMM	0.84	0.91	0.25	0.29	100.00	100.00
	TSC	0.96	0.98	0.05	0.06	5.18	5.20
	PEL						
	FINA	0.94	1.00	0.26	0.30	1.25	1.25
	FINB	0.92	0.99	0.28	0.31	1.00	1.02
$n = 40$	GFM	0.48	0.48	0.04	0.04	40.00	40.00
$n_i = 20$	GMM	0.67	0.76	0.19	0.23	40.00	40.00
	TSC	0.48	0.50	0.04	0.04	5.82	5.93
	PEL	0.39	0.40	0.05	0.06	14.17	14.14
	FINA	0.82	0.89	0.21	0.25	1.53	1.51
	FINB	0.90	0.99	0.26	0.31	1.11	1.02
$n = 20$	GFM	0.25	0.25	0.04	0.04	20.00	20.00
$n_i = 40$	GMM	0.46	0.54	0.14	0.17	20.00	20.00
	TSC	0.27	0.29	0.05	0.05	5.74	5.97
	PEL	0.25	0.26	0.06	0.06	9.59	9.62
	FINA	0.62	0.71	0.17	0.21	1.80	1.73
	FINB	0.81	0.91	0.25	0.29	1.22	1.16

Table B.3.: Average results for the settings with normal response, chi-squared distributed intercepts and $\rho = 0$.

		MSE - intercepts		**MSE - linear term**		**Number of Clusters**	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 200$	GFM	2.27	2.27	0.04	0.04	200.00	200.00
$n_i = 4$	GMM	0.50	0.59	0.03	0.03	200.00	200.00
	TSC	1.52	1.59	0.04	0.04	4.60	4.88
	PEL						
	FINA	0.69	0.77	0.03	0.03	1.49	1.80
	FINB	0.63	0.76	0.03	0.03	1.14	1.32
$n = 100$	GFM	1.10	1.10	0.03	0.03	100.00	100.00
$n_i = 8$	GMM	0.41	0.47	0.02	0.02	100.00	100.00
	TSC	0.91	0.95	0.02	0.03	4.77	5.14
	PEL						
	FINA	0.54	0.50	0.02	0.02	1.72	1.90
	FINB	0.55	0.55	0.02	0.02	1.28	1.53
$n = 40$	GFM	0.45	0.45	0.03	0.03	40.00	40.00
$n_i = 20$	GMM	0.26	0.28	0.03	0.03	40.00	40.00
	TSC	0.42	0.42	0.03	0.03	4.95	5.15
	PEL	0.30	0.29	0.03	0.03	13.17	13.27
	FINA	0.26	0.28	0.03	0.03	1.85	2.00
	FINB	0.28	0.29	0.03	0.03	1.60	1.68
$n = 20$	GFM	0.23	0.23	0.03	0.03	20.00	20.00
$n_i = 40$	GMM	0.16	0.16	0.03	0.03	20.00	20.00
	TSC	0.22	0.23	0.03	0.03	4.69	4.92
	PEL	0.15	0.15	0.03	0.03	7.87	8.23
	FINA	0.14	0.18	0.03	0.03	1.88	2.10
	FINB	0.15	0.20	0.03	0.03	1.67	1.81

Table B.4.: Average results for the settings with normal response, chi-squared distributed intercepts and $\rho = 0.8$.

		MSE - intercepts		**MSE - linear term**		**Number of Clusters**	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 200$	GFM	2.30	2.30	0.05	0.05	200.00	200.00
$n_i = 4$	GMM	0.56	0.73	0.13	0.20	200.00	200.00
	TSC	1.51	1.55	0.05	0.06	4.62	4.85
	PEL						
	FINA	0.64	0.82	0.13	0.20	1.18	1.24
	FINB	0.60	0.77	0.14	0.21	1.01	1.01
$n = 100$	GFM	1.12	1.12	0.04	0.04	100.00	100.00
$n_i = 8$	GMM	0.53	0.70	0.12	0.18	100.00	100.00
	TSC	0.92	0.95	0.04	0.05	4.72	4.99
	PEL						
	FINA	0.61	0.74	0.12	0.19	1.32	1.33
	FINB	0.60	0.77	0.13	0.20	1.01	1.03
$n = 40$	GFM	0.48	0.48	0.04	0.04	40.00	40.00
$n_i = 20$	GMM	0.44	0.62	0.11	0.17	40.00	40.00
	TSC	0.45	0.46	0.05	0.05	4.82	5.12
	PEL	0.33	0.32	0.05	0.05	12.85	13.07
	FINA	0.45	0.56	0.11	0.15	1.62	1.56
	FINB	0.51	0.70	0.13	0.20	1.26	1.20
$n = 20$	GFM	0.26	0.26	0.04	0.04	20.00	20.00
$n_i = 40$	GMM	0.30	0.44	0.08	0.13	20.00	20.00
	TSC	0.26	0.26	0.04	0.04	4.69	4.92
	PEL	0.20	0.19	0.04	0.04	8.04	8.21
	FINA	0.31	0.44	0.08	0.11	1.74	1.77
	FINB	0.38	0.62	0.11	0.17	1.39	1.34

Table B.5.: Average results for the settings with binary response, normal distributed intercepts and $\rho = 0$.

		MSE - intercepts		MSE - linear term		Number of Clusters	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 100$	GFM						
$n_i = 8$	GMM	0.74	0.88	0.03	0.03	100.00	100.00
	TSC	1.06	1.29	0.02	0.02	2.96	2.98
	PEL						
	FINA	2.88	2.39	0.03	0.03	2.98	3.03
	FINB	2.11	1.66	0.03	0.02	2.64	2.63
$n = 40$	GFM						
$n_i = 20$	GMM	0.48	0.56	0.02	0.02	40.00	40.00
	TSC	0.70	0.87	0.02	0.02	3.32	3.50
	PEL	1.23	1.20	0.02	0.02	10.78	14.28
	FINA	10.70	5.26	0.02	0.02	3.49	3.52
	FINB	9.10	3.93	0.02	0.02	3.00	2.97
$n = 20$	GFM						
$n_i = 40$	GMM	0.71	0.62	0.03	0.03	20.00	20.00
	TSC	2.40	2.18	0.03	0.03	3.44	3.84
	PEL	1.44	1.15	0.03	0.03	5.70	9.15
	FINA	19.94	12.58	0.03	0.03	3.57	3.84
	FINB	15.58	8.71	0.03	0.03	3.12	3.21

Table B.6.: Average results for the settings with binary response, normal distributed intercepts and $\rho = 0.8$.

		MSE - intercepts		MSE - linear term		Number of Clusters	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 100$	GFM						
$n_i = 8$	GMM	2.13	2.55	0.48	0.54	100.00	100.00
	TSC	1.59	1.93	0.25	0.29	2.46	2.38
	PEL						
	FINA	3.43	3.89	0.46	0.51	2.35	2.26
	FINB	2.60	2.95	0.50	0.56	1.93	1.85
$n = 40$	GFM						
$n_i = 20$	GMM	0.92	1.12	0.14	0.15	40.00	40.00
	TSC	0.98	1.16	0.11	0.12	3.04	3.13
	PEL	1.32	1.26	0.05	0.05	10.42	13.19
	FINA	12.51	8.08	0.11	0.14	2.96	2.91
	FINB	8.06	5.39	0.16	0.22	2.45	2.29
$n = 20$	GFM						
$n_i = 40$	GMM	0.87	0.84	0.07	0.08	20.00	20.00
	TSC	2.67	1.87	0.06	0.07	3.21	3.53
	PEL	1.74	1.26	0.05	0.05	5.61	8.91
	FINA	22.57	13.19	0.06	0.09	3.34	3.41
	FINB	15.15	7.81	0.09	0.14	2.81	2.64

Table B.7.: Average results for the settings with binary response, chi-squared distributed intercepts and $\rho = 0$.

		MSE - intercepts		**MSE - linear term**		**Number of Clusters**	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 100$	GFM						
$n_i = 8$	GMM	0.68	0.92	0.02	0.02	100.00	100.00
	TSC	0.91	1.39	0.02	0.02	2.79	2.85
	PEL						
	FINA	1.72	2.30	0.02	0.02	2.74	2.90
	FINB	1.42	1.70	0.02	0.02	2.40	2.51
$n = 40$	GFM						
$n_i = 20$	GMM	0.48	0.61	0.02	0.02	40.00	40.00
	TSC	0.59	0.82	0.02	0.02	3.01	3.37
	PEL	1.60	1.43	0.02	0.02	9.83	12.35
	FINA	5.61	6.94	0.02	0.02	3.04	3.34
	FINB	4.66	4.47	0.02	0.02	2.74	2.91
$n = 20$	GFM						
$n_i = 40$	GMM	1.61	2.00	0.03	0.03	20.00	20.00
	TSC	2.81	2.96	0.03	0.03	2.94	3.56
	PEL	1.93	2.04	0.03	0.02	5.75	8.04
	FINA	21.18	19.61	0.04	0.03	3.04	3.61
	FINB	19.95	16.30	0.03	0.03	2.77	3.06

Table B.8.: Average results for the settings with binary response, chi-squared distributed intercepts and $\rho = 0.8$.

		MSE - intercepts		**MSE - linear term**		**Number of Clusters**	
		$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$	$m_0 = 5$	$m_0 = 10$
$n = 100$	GFM						
$n_i = 8$	GMM	1.55	2.30	0.41	0.50	100.00	100.00
	TSC	1.28	1.91	0.22	0.30	2.50	2.30
	PEL						
	FINA	4.85	4.37	0.33	0.46	2.48	2.24
	FINB	2.68	2.45	0.37	0.50	2.05	1.86
$n = 40$	GFM						
$n_i = 20$	GMM	0.72	1.15	0.13	0.16	40.00	40.00
	TSC	0.75	1.15	0.09	0.12	2.80	3.01
	PEL	1.72	1.53	0.04	0.05	9.38	11.89
	FINA	9.57	6.76	0.09	0.14	2.85	2.84
	FINB	7.06	4.68	0.11	0.17	2.50	2.47
$n = 20$	GFM						
$n_i = 40$	GMM	1.66	2.26	0.07	0.07	20.00	20.00
	TSC	3.08	2.92	0.06	0.07	2.81	3.33
	PEL	2.26	2.34	0.05	0.05	5.59	7.72
	FINA	21.87	21.18	0.06	0.08	2.90	3.25
	FINB	21.79	16.13	0.07	0.11	2.68	2.68

C. Additional Simulation Results for Chapter 5

In the following we give additional results of the simulations in Section 5.5 and 5.6.5. For a detailed description we refer to the respective sections.

Table C.1.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the twelve settings with 10% DIF in the simulation with one binary predictor.

10% DIF, $\alpha = 0.05$			**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic
c=0.4	I=20	P=400	0.135	0.145	0.049	0.051
		P=800	0.415	0.410	0.046	0.047
	I=40	P=400	0.240	0.245	0.050	0.050
		P=800	0.395	0.400	0.051	0.051
c=0.8	I=20	P=400	0.310	0.310	0.051	0.051
		P=800	0.905	0.895	0.047	0.044
	I=40	P=400	0.598	0.613	0.049	0.050
		P=800	0.745	0.750	0.051	0.051
c=1.6	I=20	P=400	0.595	0.595	0.055	0.053
		P=800	1.000	1.000	0.047	0.051
	I=40	P=400	0.963	0.965	0.051	0.050
		P=800	0.828	0.828	0.051	0.052

Table C.2.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the twelve settings with 20% DIF in the simulation with one binary predictor.

10% DIF, $\alpha = 0.05$			**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic
c=0.4	I=20	P=400	0.177	0.172	0.050	0.049
		P=800	0.440	0.448	0.046	0.046
	I=40	P=400	0.236	0.240	0.050	0.050
		P=800	0.401	0.406	0.049	0.050
c=0.8	I=20	P=400	0.378	0.385	0.048	0.049
		P=800	0.930	0.932	0.045	0.045
	I=40	P=400	0.588	0.589	0.051	0.050
		P=800	0.731	0.731	0.049	0.051
c=1.6	I=20	P=400	0.700	0.698	0.052	0.049
		P=800	1.000	1.000	0.042	0.046
	I=40	P=400	0.900	0.897	0.049	0.052
		P=800	0.792	0.791	0.050	0.050

Table C.3.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the twelve settings with 10% DIF in the simulation with one ordered predictor.

10% DIF, $\alpha = 0.05$			**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic
c=0.4	I=20	P=400	0.105	0.065	0.052	0.049
		P=800	0.185	0.175	0.052	0.047
	I=40	P=400	0.147	0.107	0.048	0.052
		P=800	0.287	0.200	0.044	0.050
c=0.8	I=20	P=400	0.275	0.165	0.051	0.048
		P=800	0.805	0.675	0.053	0.047
	I=40	P=400	0.472	0.383	0.050	0.052
		P=800	0.720	0.672	0.045	0.051
c=1.6	I=20	P=400	0.560	0.515	0.051	0.048
		P=800	1.000	1.000	0.057	0.048
	I=40	P=400	0.915	0.877	0.048	0.053
		P=800	0.812	0.795	0.044	0.052

Table C.4.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the twelve settings with 20% DIF in the simulation with one ordered predictor.

10% DIF, $\alpha = 0.05$			**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic
c=0.4	I=20	P=400	0.110	0.085	0.052	0.046
		P=800	0.352	0.250	0.052	0.051
	I=40	P=400	0.166	0.128	0.048	0.054
		P=800	0.278	0.184	0.048	0.051
c=0.8	I=20	P=400	0.292	0.240	0.053	0.047
		P=800	0.863	0.777	0.051	0.048
	I=40	P=400	0.500	0.417	0.047	0.053
		P=800	0.704	0.641	0.045	0.052
c=1.6	I=20	P=400	0.618	0.568	0.053	0.046
		P=800	1.000	1.000	0.049	0.048
	I=40	P=400	0.881	0.843	0.049	0.055
		P=800	0.780	0.775	0.047	0.052

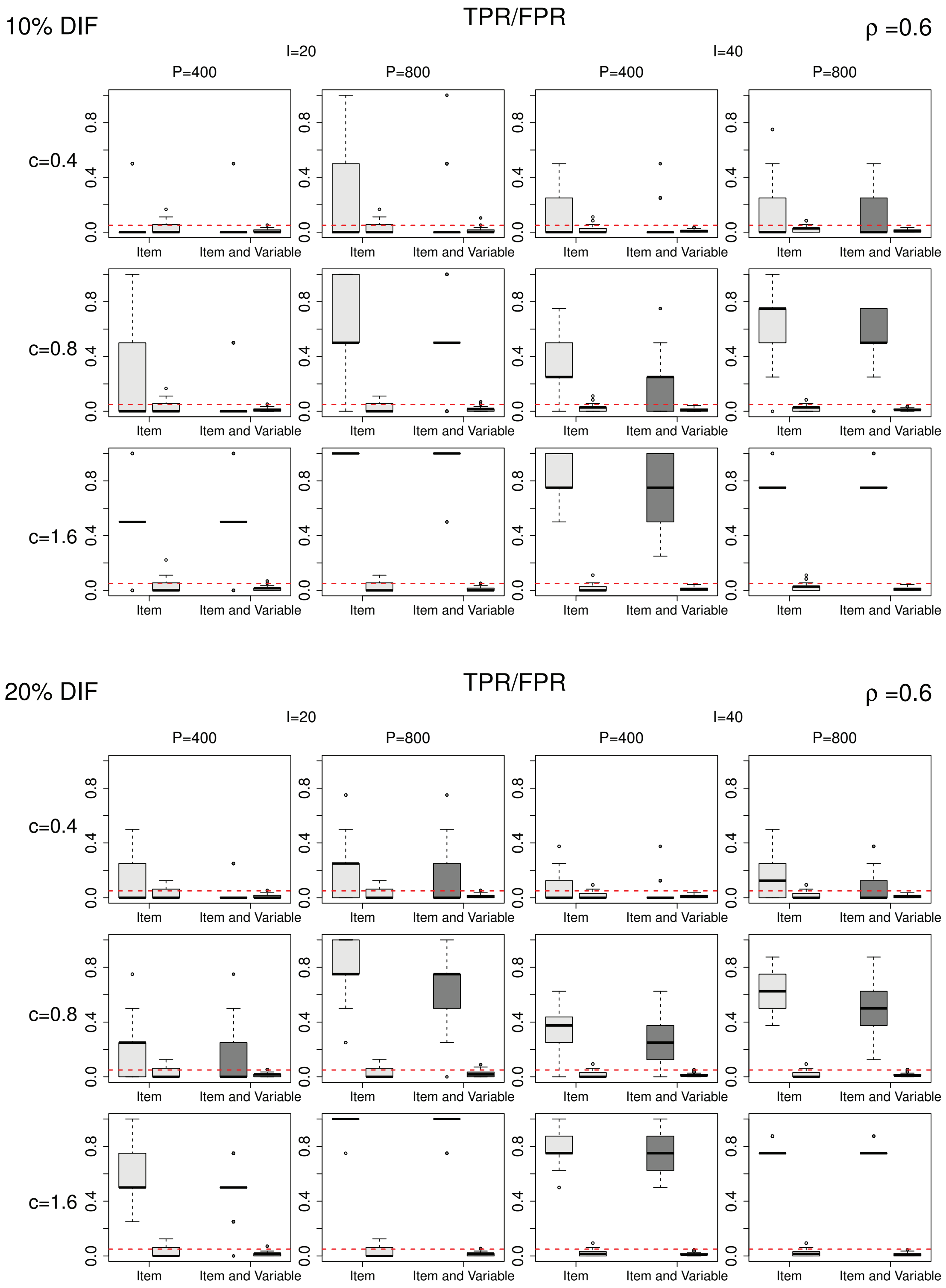


Figure C.1.: Boxplots of TPR and FPR at significance level $\alpha = 0.05$ (marked by dashed lines) in the simulation with three covariates, DIF in $x1$ and correlation between the predictors ($\rho = 0.6$). Results on item level are given in light grey, results for the combination of item and variable are given in dark grey.

Table C.5.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the eight settings with 10% DIF in the simulation with non-uniform DIF and one binary predictor.

10% DIF, $\alpha = 0.05$			**DIF**				**NUDIF**			
			TPR		**FPR**		**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic	IFT	Logistic	IFT	Logistic
c=0.3	I=20	P=400	0.318	0.385	0.031	0.046	0.200	0.200	0.051	0.048
		P=800	0.207	0.292	0.031	0.054	0.125	0.130	0.053	0.047
	I=40	P=400	0.151	0.241	0.033	0.052	0.142	0.145	0.049	0.047
		P=800	0.196	0.346	0.032	0.053	0.320	0.333	0.044	0.046
c=0.6	I=20	P=400	0.575	0.688	0.032	0.048	0.440	0.440	0.047	0.051
		P=800	0.438	0.662	0.032	0.054	0.380	0.380	0.052	0.049
	I=40	P=400	0.414	0.615	0.036	0.056	0.307	0.318	0.048	0.048
		P=800	0.474	0.845	0.034	0.052	0.647	0.650	0.046	0.049

Table C.6.: Average TPR and FPR on the item level at significant level $\alpha = 0.05$ for the eight settings with 20% DIF in the simulation with non-uniform DIF and one binary predictor.

20% DIF, $\alpha = 0.05$			**DIF**				**NUDIF**			
			TPR		**FPR**		**TPR**		**FPR**	
			IFT	Logistic	IFT	Logistic	IFT	Logistic	IFT	Logistic
c=0.3	I=20	P=400	0.318	0.385	0.031	0.046	0.145	0.145	0.047	0.046
		P=800	0.207	0.292	0.031	0.054	0.182	0.177	0.048	0.045
	I=40	P=400	0.151	0.241	0.033	0.052	0.146	0.156	0.045	0.046
		P=800	0.196	0.346	0.032	0.053	0.284	0.284	0.043	0.044
c=0.3	I=20	P=400	0.575	0.688	0.032	0.048	0.340	0.340	0.051	0.049
		P=800	0.438	0.662	0.032	0.054	0.440	0.442	0.056	0.051
	I=40	P=400	0.414	0.615	0.036	0.056	0.354	0.362	0.044	0.045
		P=800	0.474	0.845	0.034	0.052	0.694	0.701	0.046	0.046

D. Supplement Exemplary R Code for Chapter 6

In the following we show how R code can be used to obtain estimates. For illustration we use part of the data of the SHIW study (illustrative example in Section 6.2). The model is estimated by use of the function `vglm()` of the `R` package `VGAM`. Before using `vglm` the data $(\mathbf{Y}_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \ldots, n$ have to brought in a specific form.

Response Matrix

The responses $\mathbf{Y}_i$ have to be given in a data matrix $\mathbf{Y}$ in wide format, such that each observation represents one row and the columns correspond to the response categories. In the SHIW study the response is the happiness index measured on a Likert scale with ten categories from 1 (very unhappy) to 10 (very happy). The first 6 observations of the data matrix $\mathbf{Y}$ are given by:

```
Y[1:6,]
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    0    0    0    0    1    0    0    0     0
## [2,]    0    0    0    0    0    1    0    0    0     0
## [3,]    0    0    0    0    0    1    0    0    0     0
## [4,]    0    0    0    0    0    0    0    0    1     0
## [5,]    0    0    0    0    0    0    1    0    0     0
## [6,]    0    0    1    0    0    0    0    0    0     0
```

Design-Matrices

The explanatory variables $\mathbf{x}_i, \mathbf{z}_i$ have also be given as data matrices $\mathbf{X}$ and $\mathbf{Z}$, again each observation represents one row and the columns correspond to the covariates. For the two covariates gender (0: male, 1: female) and age in decades (centered by 60), which are allowed to have content-related and response style effects, the first observations of the data matrices $\mathbf{X}$ and $\mathbf{Z}$ are:

```
X[1:6,]
```

```
##   Gen Age
## 1   1 2.5
## 2   0 1.1
## 3   0 1.3
## 4   0 2.3
## 5   0 0.3
## 6   1 1.7
```

```
Z[1:6,]
```

```
##   Gen Age
## 1   1 2.5
## 2   0 1.1
## 3   0 1.3
## 4   0 2.3
## 5   0 0.3
## 6   1 1.7
```

From the data matrices $\mathbf{Y}, \mathbf{X}$ and $\mathbf{Z}$ several important values can be extracted. The number of observations corresponds to the number of rows of $\mathbf{X}$ ($n = 3816$), the number of categories corresponds to the number of columns of $\mathbf{Y}$ ($k = 10$), the number of content-related covariates correspond to the number of columns of $\mathbf{X}$ ($px = 2$) and the the number of covariates with response style effect correspond to the the number of columns of $\mathbf{Z}$ ($pz = 2$).

In the proposed models (2) and (3) the explanatory variables $\mathbf{z}_i$ represent a special case of category-specific covariates for which only the sign differs depending on the response category. In the case of an odd number of categories with middle categorie $m = [k/2]+1$ the sign is positive for categories $r = 1, \ldots, m-1$ and negative for categories $r = 1, \ldots, k-1$. In the even case with middle category $m = k/2$ the sign is positive for categories $r = 1, \ldots, m-1$, negative for categories $r = m+1, \ldots, k-1$ and the variables $\mathbf{z}_i$ are set to zero for the middle category m.

The data matrix $\mathbf{Z}$ has to be extended to a data matrix, named **Zext**, where each observation represents one row and the columns contain the values of $\mathbf{z}_i$ for each linear predictor η_{ir}. The corresponding code is:

```
Zext <- Z[,rep(1:pz,each=k-1)]

if(k%%2!=0){               # odd number of categories
  m <- floor(k/2)+1
  for(i in 0:(pz-1)){
    Zext[,(m:(k-1))+i*(k-1)] <- -Zext[,(m:(k-1))+i*(k-1)]
  }
}

if(k%%2==0){               # even number of categories
  m <- k/2
  for(i in 0:(pz-1)){
    Zext[,((m+1):(k-1))+i*(k-1)] <- -Zext[,((m+1):(k-1))+i*(k-1)]
    Zext[,m+i*(k-1)] <- 0
  }
}
```

To improve readability of the model output it is useful to choose informative labels for the columns of data matrices $\mathbf{X}, \mathbf{Z}$ and **Zext**. One might use:

```
l1              <- paste0(rep(names(Z),each=k-1),"z")
l2              <- rep(1:(k-1),times=pz)
colnames(Zext) <- namesZext <- paste0(l1,l2)
colnames(Z)    <- namesZ    <- paste0(names(Z),"z")
colnames(X)    <- namesX    <- paste0(names(X),"x")
```

In the SHIW study there are ten response categories yielding nine linear predictors $\eta_{ir}, r = 1, \ldots, 9$. With the two explanatory variables gender and age the extended data matrix **Zext** in total consists of 18 columns and nine columns per covariate. Columns 1 to 4 contain positive values, column 5 contains zeros and columns 6 to 9 contain negative values, respectively. The data matrix **Zext** (partially) is:

```
Zext[1:6,]


##   Genz1 Genz2 Genz3 Genz4 Genz5 Genz6 Genz7 Genz8 Genz9
## 1     1     1     1     1     0    -1    -1    -1    -1
## 2     0     0     0     0     0     0     0     0     0
## 3     0     0     0     0     0     0     0     0     0
## 4     0     0     0     0     0     0     0     0     0
## 5     0     0     0     0     0     0     0     0     0
## 6     1     1     1     1     0    -1    -1    -1    -1
##   Agez1 Agez2 Agez3 Agez4 Agez5 Agez6 Agez7 Agez8 Agez9
## 1   2.5   2.5   2.5   2.5     0  -2.5  -2.5  -2.5  -2.5
## 2   1.1   1.1   1.1   1.1     0  -1.1  -1.1  -1.1  -1.1
## 3   1.3   1.3   1.3   1.3     0  -1.3  -1.3  -1.3  -1.3
## 4   2.3   2.3   2.3   2.3     0  -2.3  -2.3  -2.3  -2.3
## 5   0.3   0.3   0.3   0.3     0  -0.3  -0.3  -0.3  -0.3
## 6   1.7   1.7   1.7   1.7     0  -1.7  -1.7  -1.7  -1.7
```

Model Specification

The use of `vglm` requires the specification of S formulas or lists of S formulas that are assigned to arguments when the function is called. The construction of the formulas is based on the labels of the columns of the data matrices that were explicitly set before.

The first formula, called ”formula1”, is the general symbolic description of the model, which will later be assigned to the argument `formula` when calling vglm.

```
f11 <- paste(namesX,collapse="+")
f12 <- paste(namesZ,collapse="+")
formula1 <- formula(paste("Y~",f11,"+",f12))
formula1


## Y ~ Genx + Agex + Genz + Agez
```

The second formula, called ”formula2”, is a one sided formula containing every term used by the model (except the response y), which will later be assigned to the argument `form2`.

```
f21 <- paste(namesZext,collapse="+")
formula2 <- formula(paste("~",f11,"+",f12,"+",f21))
formula2
```

```
## ~Genx + Agex + Genz + Agez + Genz1 + Genz2 + Genz3 + Genz4 +
##     Genz5 + Genz6 + Genz7 + Genz8 + Genz9 + Agez1 + Agez2 + Agez3 +
##     Agez4 + Agez5 + Agez6 + Agez7 + Agez8 + Agez9
```

The third formula, called ”formula3”, is a list of formulas, which will later be assigned to the argument `xij`. `VGAM` handles category-specific covariates by the `xij` argument. Each formula corresponds to one covariate, where the right-hand side consists of $k-1$ terms making up a covariate-dependent term. The $k-1$ terms must be unique and should be enumerated in sequential order.

```
formula3 <- c()
for(i in 0:(pz-1)){
  f31 <- paste(namesZext[(1:(k-1))+(k-1)*i],collapse="+")
  f32 <- formula(paste(namesZ[i+1],"~",f31))
  formula3 <- c(formula3,f32)
}
formula3

## [[1]]
## Genz ~ Genz1 + Genz2 + Genz3 + Genz4 + Genz5 + Genz6 + Genz7 +
##     Genz8 + Genz9
##
## [[2]]
## Agez ~ Agez1 + Agez2 + Agez3 + Agez4 + Agez5 + Agez6 + Agez7 +
##     Agez8 + Agez9
```

Estimation with vglm

Package `VGAM` and additional dependent packages have to be loaded.

```
require(VGAM)

## Loading required package:  VGAM
## Loading required package:  stats4
## Loading required package:  splines
```

The data matrix **DM** is assigned to argument `data` when calling `vglm`. The data matrix has to contain the columns of every term used by the model or in the formulas (with the exception of the response matrix **Y**).

```
DM <- data.frame(X,Zext,Z)
```

Now the extended adjacent categories model can be estimated by `vglm`. In `family` one has to choose the adjacent categories family `acat(reverse=FALSE)`, where `reverse=FALSE` means that the ratios $\pi_{i,r+1}/\pi_{ir}$ are modelled. With the argument `parallel` one defines if the estimated effects are category-specific or not. The specification `parallel=FALSE∼1` ensures that only the intercepts $\theta_r, r = 1, \ldots, k-1$ are category-specific. For a more detailed description see also Yee (2010). Finally the call of function `vglm` is:

```
mod <- vglm(formula=formula1,
            family=acat(parallel=FALSE~1,reverse=FALSE),
            xij=formula3,
            form2=formula2,
            data=DM)
```

The summary of the estimated model is the following. The coefficients `(Intercept):1` to `(Intercept):9` are the category-specific intercepts, the coefficients `Genx` and `Agex` are the content-related effects and `Genz` and `Agez` are the response style effects of the two covariates.

```
summary(mod)

##
## Call:
## vglm(formula = formula1, family = acat(parallel = FALSE ~ 1,
##     reverse = FALSE), data = DM, form2 = formula2, xij = formula3)
##
## Pearson residuals:
##                           Min       1Q   Median       3Q   Max
## loge(P[Y=2]/P[Y=1])  -6.345  0.02261  0.03057  0.07315  1.960
## loge(P[Y=3]/P[Y=2])  -3.917  0.05674  0.07425  0.15630  1.861
## loge(P[Y=4]/P[Y=3])  -2.652  0.10692  0.13512  0.24334  2.554
## loge(P[Y=5]/P[Y=4])  -2.727 -0.63422  0.18816  0.29138  2.588
## loge(P[Y=6]/P[Y=5])  -2.916 -0.56699  0.28447  0.47542  2.095
## loge(P[Y=7]/P[Y=6])  -2.540 -0.49697  0.32373  0.52538  1.558
## loge(P[Y=8]/P[Y=7])  -1.850 -0.48453 -0.23893  0.75931  1.879
## loge(P[Y=9]/P[Y=8])  -1.123 -0.30230 -0.12834 -0.07773  4.010
## loge(P[Y=10]/P[Y=9]) -2.066 -0.22875 -0.07256 -0.03995  5.229
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.446155   0.109053   4.091 4.29e-05 ***
## (Intercept):2  0.404232   0.087237   4.634 3.59e-06 ***
## (Intercept):3 -0.185793   0.080483  -2.308   0.0210 *
## (Intercept):4  0.166246   0.079348   2.095   0.0362 *
## (Intercept):5  0.270726   0.069320   3.905 9.41e-05 ***
## (Intercept):6  0.363964   0.062457   5.827 5.63e-09 ***
## (Intercept):7  0.289437   0.054977   5.265 1.40e-07 ***
## (Intercept):8 -0.969050   0.070319 -13.781  < 2e-16 ***
## (Intercept):9 -0.051585   0.086803  -0.594   0.5523
## Genx          -0.036342   0.013789  -2.636   0.0084 **
## Agex          -0.001861   0.004385  -0.424   0.6712
## Genz           0.143557   0.026715   5.374 7.72e-08 ***
## Agez           0.062670   0.008546   7.333 2.25e-13 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  9
##
## Dispersion Parameter for acat family:   1
##
## Residual deviance: 16609.1 on 34331 degrees of freedom
##
## Log-likelihood: -8304.549 on 34331 degrees of freedom
##
## Number of iterations: 4
```

References

Agresti, A. (2009). *Analysis of Ordinal Categorical Data, 2nd Edition.* New York: Wiley.

Agresti, A., B. Caffo, and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis 47*, 639–653.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (1973). *Intelligenz-Struktur-Test (IST 70).* Göttingen: Hogrefe.

Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (1999). *Intelligenz-Struktur-Test 2000 (IST 2000).* Göttingen: Hogrefe.

Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (2001). *Intelligenz-Struktur-Test 2000 R (IST 2000 R) - Handanweisung.* Göttingen: Hogrefe.

Anderson, E. B. (1973). *Conditional Inference and Models for Measuring.* Copenhagen: Metalhygiejnish Forlag.

Bachman, J. G. and P. O'Malley (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly 48*(2), 491–509.

Bates, D., M. Mächler, and B. Bolker (2014). *mlmRev: Examples from Multilevel Modelling Software Review.* R package version 1.0-6.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Baumgartner, H. and J.-B. E. Steenkamp (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research 38*(2), 143–156.

Beauducel, A., D. Liepmann, S. Horn, and B. Brocke (2010). *Intelligence Structure Test - English Version of the Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R).* Göttingen: Hogrefe.

Belitz, C., A. Brezger, T. Kneib, S. Lang, and N. Umlauf (2015). *BayesX: Software for Bayesian Inference in Structured Additive Regression Models.* R package version 1.0-0.

Bender, R. and U. Grouven (1998). Using binary logistic regression models for ordinal data with non–proportional odds. *Journal of Clinical Epidemiology 51*(10), 809–816.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Berger, M. (2016a). *DIFtree: Item Focused Trees for the Identification of Items in Differential Item Functioning.* R package version 1.1.0/2.0.1/2.0.2.

Berger, M. (2016b). *structree: Tree-Structured Clustering for the Identification of Latent Groups.* R package version 1.0.1.

Berger, M. and G. Tutz (2015a). Detection of uniform and non-uniform differential item functioning by item focussed trees. *Cornell University Library.* arXiv:1511.07178.

Berger, M. and G. Tutz (2015b). An extended adjacent categories model accounting for response styles. *Proceedings of the 30th International Workshop on Statistical Modelling, Volume 2.* Statistical Modelling Society.

Berger, M. and G. Tutz (2015c). Tree-structured clustering in fixed effects models. *Cornell University Library.* arXiv:1512.05169.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika 37*(1), 29–51.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods 17*(4), 665–678.

Bolt, D. M. and T. R. Johnson (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement 33*(5), 335–352.

Bolt, D. M. and J. R. Newton (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement 71*(5), 814–833.

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics 64*(1), 115–123.

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics 65*(1), 169–177.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics 46*(4), 1171–1178.

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

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association 98*(462), 324–339.

Bühner, M., M. Ziegler, S. Krumm, and L. Schmidt-Atzert (2006). Ist der I-S-T 2000 R Rasch-skalierbar? *Diagnostica 52*(3), 119–130.

Bürgin, R. and G. Ritschard (2015). Tree-based varying coefficient regression for longitudinal ordinal responses. *Computational Statistics & Data Analysis 86*, 65–80.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric bayesian model for randomised block designs. *Biometrika 83*(2), 275–285.

Chen, J. and M. Davidian (2002). A monte carlo EM algorithm for generalized linear models with flexible random effects distribution. *Biostatistics 3*(3), 347–360.

Chen, J., K. Yu, A. Hsing, and T. M. Therneau (2007). A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects. *Genetic epidemiology 31*(3), 238–251.

Ciampi, A., C.-H. Chang, S. Hogg, and S. McKinney (1987). Recursive partitioning: A versatile method for exploratory data analysis in biostatistics. In I. McNeil and G. Umphrey (Eds.), *Biostatistics*. New York: D. Reidel Publishing.

Claeskens, G. and J. D. Hart (2009). Goodness-of-fit tests in mixed models. *TEST 18*(2), 213–239.

Clark, L. and D. Pregibon (1992). Tree-based models. In J. Chambers and T. Hastie (Eds.), *Statistical Models in S*, pp. 377–420. Pacific Grove, California: Wadsworth & Brooks.

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality 15*(1), 137–152.

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review 18*(3), 301–324.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine 14*(11), 1191–1203.

Curtin, T. R., S. J. Ingels, S. Wu, and R. Heuer (2002). *National Education Longitudinal Study of 1988: Base-Year to Fourth Follow-up Data File User's Manual (NCES 2002-323).* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

De Boeck, P. and I. Partchev (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software 48*(1), 1–28.

De Boeck, P. and M. Wilson (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer Verlag.

Dusseldorp, E., C. Conversano, and B. J. Van Os (2010). Combining an additive and tree-based regression model simultaneously: Stima. *Journal of Computational and Graphical Statistics 19*(3), 514–530.

Dusseldorp, E. and J. J. Meulman (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika 69*(3), 355–374.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*, Volume 57. Chapman & Hall/CRC.

Eid, M. and M. Rauber (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment 16*(1), 20.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

El-Komboz, B. A., A. Zeileis, and C. Strobl (2014). Detecting differential item and step functioning with rating scale and partial credit trees. *Ludwig-Maximilians-Universität München, Department of Statistics*. Technical Report 152.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications.* New York: Springer.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models.* New York: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association 53*(284), 789–798.

Follmann, D. and D. Lambert (1989). Generalizing logistic regression by non-parametric mixing. *Journal of the American Statistical Association 84*(405), 295–300.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics 28*(2), 337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232.

Gambacorta, R. and M. Iannario (2013). Measuring job satisfaction with cub models. *Labour 27*(2), 198–224.

Gertheiss, J. and G. Tutz (2009). Penalized Regression with Ordinal Predictors. *International Statistical Review 77*(3), 345–365.

Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorial explanatory variables. *Annals of Applied Statistics 4*(4), 2150–2180.

Goeman, J., R. Meijer, and N. Chaturvedi (2014). *penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model.* R package version 0.9-45.

Grilli, L. and C. Rampichini (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioural and Social Sciences 7*(4), 121–133.

Grün, B. and F. Leisch (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis 51*(11), 5247–5252.

Grün, B. and F. Leisch (2008a). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software 28*(4), 1–35.

Grün, B. and F. Leisch (2008b). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification 25*(2), 225–247.

Hajjem, A., F. Bellavance, and D. Larocque (2011). Mixed effects regression trees for clustered data. *Statistics and Probability Letters 81*(4), 451–459.

Hamada, M. and C. F. J. Wu (1990). A critical look at accumulation analysis and related methods. *Technometrics 32*(2), 119–130.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B 55*(4), 757–796.

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

Hatzinger, R. (1989). The rasch model, some extensions and their relation to the class of generalized linear models. In *Statistical Modelling*. New York: Springer.

Heagerty, P. and B. F. Kurland (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika 88*(4), 973–985.

Heinzl, F. and G. Tutz (2013). Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. *Statistical Modelling 13*(1), 41–67.

Heinzl, F. and G. Tutz (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal 56*(1), 44–68.

Heinzl, F. and G. Tutz (2016). Additive mixed models with approximate dirichlet process mixtures: the em approach. *Statistics and Computing 26*(1), 73–92.

Hjort, N. L., C. Holmes, P. Müller, and S. G. Walker (2010). *Bayesian nonparametrics*, Volume 28. Cambridge University Press.

Holland, P. W. and D. T. Thayer (1988). Differential item performance and the Mantel-Haenszel procedure. In *Test validity*, pp. 129–145. Routledge.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics 6*(2), 65–70.

Hoover, D. R., J. A. Rice, C. O. Wu, and L. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Hothorn, T. and B. Lausen (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics and Data Analysis 43*(2), 121–137.

Huang, X. (2009). Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics 65*(2), 361–368.

Iannario, M. and D. Piccolo (2012). CUB models: Statistical methods and empirical evidence. In R. Kenett and S. Salini (Eds.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. New York: Wiley.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics 1*, 519–537.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences.* Wiley: Hoboken, New Jersey.

Jeon, M. and P. De Boeck (2015). A generalized item response tree model for psychological assessments. *Behavior Research Methods, published online*. doi: 10.3758/s13428-015-0631-y.

Jodoin, M. and M. Gierl (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for dif detection. *Applied Measurement in Education 14*(4), 329–349.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika 68*(4), 563–583.

Kankaraš, M. and G. Moors (2009). Measurement equivalence in solidarity attitudes in europe insights from a multiple-group latent-class factor approach. *International Sociology 24*(4), 557–579.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1038.

Khorramdel, L. and M. von Davier (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research 49*(2), 161–177.

Kim, J.-H. (2003). Assessing practical significance of the proportional odds assumption. *Statistics & probability letters 65*(3), 233–239.

Kim, S.-H., A. S. Cohen, and T.-H. Park (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement 32*(3), 261–276.

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News 6*(4), 8–12. R package version 3.6.

Litière, S., A. Alonso, and G. Molenberghs (2007). Type I and Type II Error Under Random Effects Misspecification in Generalized Linear Mixed Models. *Biometrics 63*(4), 1038–1044.

Liu, I., B. Mukherjee, T. Suesse, D. Sparrow, and S. K. Park (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in medicine 28*(3), 412–429.

Liu, Q. and A. Agresti (2005). The analysis of ordinal categorical data: An overview and a survey of recent developments. *Test 14*(1), 1–73.

Lombardía, M. J. and S. Sperlich (2012). A new class of semi-mixed effects models and its application in small area estimation. *Computational Statistics & Data Analysis 56*(10), 2903–2917.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lu, Y., R. Zhang, and L. Zhu (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics - Theory and Methods 37*(14), 2249–2261.

Magder, L. and S. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association 91*(435), 1141–1151.

Magis, D., S. Beland, and G. Raiche (2013). *difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics.* R package version 4.5.

Magis, D., S. Bèland, F. Tuerlinckx, and P. Boeck (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods 42*(3), 847–862.

Magis, D. and P. De Boeck (2014). Type I error inflation in DIF identification with Mantel–Haenszel an explanation and a solution. *Educational and Psychological Measurement 74*(4), 713–728.

Magis, D., G. Raîche, S. Béland, and P. Gérard (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing 11*(4), 365–386.

Magis, D., F. Tuerlinckx, and P. De Boeck (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics 40*(2), 111–135.

Marin, G., R. J. Gamba, and B. V. Marin (1992). Extreme response style and acquiescence among hispanics the role of acculturation and education. *Journal of Cross-Cultural Psychology 23*(4), 498–509.

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

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*(2), 109–127.

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

McCulloch, C. and S. Searle (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 72*(4), 417–473.

Meisenberg, G. and A. Williams (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences 44*(7), 1539–1550.

Messick, S. (1991). Psychology and methodology of response styles. In *Improving inquiry in social science: A volume in honor of Lee J. Cronbach*, pp. 161–200. Educational Testing Service.

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

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitdinal Data.* New York: Springer–Verlag.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity 37*(3), 277–302.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review 20*(4), 303–320.

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research 22*(2), 93–119.

Morgan, J. N. and J. A. Sonquist (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*(302), 415–435.

Müller, P. and G. L. Rosner (1997). A bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association 92*(440), 1279–1292.

Nair, V. N. (1987). Chi-squared-type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association 82*(397), 283–291.

Narayanan, P. and H. Swaminathan (1996). Identification of items that show nonuniform dif. *Applied Psychological Measurement 20*(3), 257–274.

Oelker, M.-R. (2015). *gvcm.cat: Regularized Categorical Effects/Categorical Effect Modifiers/Continuous/Smooth Effects in GLMs.* R package version 1.9.

Oelker, M.-R. and G. Tutz (2015). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification, published online*. doi: 10.1007/s11634-015-0205-y.

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

Peterson, B. and F. E. Harrell (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics 39*(2), 205–217.

Peyhardi, J., C. Trottier, and Y. Guédon (2015). A new specification of generalized linear models for categorical data. *Biometrika, published online*. doi: 10.1093/biomet/asv042.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica 5*, 85–104.

Plieninger, H. and T. Meiser (2014). Validity of multiprocess irt models for separating content and response styles. *Educational and Psychological Measurement 74*(5), 875–899.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*(1), 81–106.

Quinlan, J. R. (1993). *Programs for Machine Learning.* San Francisco: Morgan Kaufmann PublisherInc.

R Core Team (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rafaeli, E. and W. Revelle (2006). A premature consensus: are happiness and sadness truly opposite affects? *Motivation and Emotion 30*(1), 1–12.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika 53*(4), 495–502.

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

Revelle, W. (2013). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University. R package version 1.3.10.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge: University Press.

Rodriguez, G. and N. Goldman (2001). Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 164*(2), 339–355.

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

Rogers, H. and H. Swaminathan (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement 17*(2), 105–116.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge: University Press.

Sandri, M. and P. Zuccolotto (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics 17*(3), 611–628.

Schauberger, G. (2014). *DIFlasso: A penalty approach to Differential Item Functioning in Rasch Models.* R package version 1.0-1.

Schauberger, G. (2015). *DIFboost: Detection of Differential Item Functioning (DIF) in Rasch Models by Boosting Techniques.* R package version 0.1.

Schauberger, G. and G. Tutz (2015). Detection of differential item functioning in rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology, published online.* doi: 10.1111/bmsp.12060.

Schmidt-Atzert, L. (2000). Intelligenz-Strukturtest 2000 R. *Zeitschrift für Personalpsychologie 1*, 52–55.

Schmidt-Atzert, L., W. Hommers, and M. Hess (1995). Der I-S-T 70: Eine Analyse und Neubewertung. *Diagnostica 41*(2), 108–130.

Sela, R. J. and J. S. Simonoff (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning 86*(2), 169–207.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Shapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227.

Shih, Y.-S. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis 45*(3), 457–466.

Shih, Y.-S. and H. Tsai (2004). Variable selection bias in regression trees with constant fits. *Computational Statistics and Data Analysis 45*(3), 595–607.

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.

Städler, N., P. Bühlmann, and S. van de Geer (2010). L1-penalization for mixture regression models. *Test 19*(2), 209–256.

Strobl, C. (2012). *Das Rasch-Modell – Eine verständliche Einführung für Studium und Praxis* (2. erweiterte Auflage ed.). München, Mering: Rainer Hampp Verlag.

Strobl, C., A.-L. Boulesteix, and T. Augustin (2007). Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis 52*(1), 483–501.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9:307.*

Strobl, C., J. Kopf, and A. Zeileis (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika 80*(2), 289–316.

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*(4), 323–348.

Su, X., K. Meneses, P. McNees, and W. O. Johnson (2011). Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society C 60*(3), 457–474.

Su, X., C.-L. Tsai, and M. C.Wang (2009). Tree-structured model diagnostics for linear regression. *Machine Learning 74*(2), 111–131.

Suh, Y. and D. M. Bolt (2010). Nested logit models for multiple-choice item response data. *Psychometrika 75*(3), 454–473.

Swaminathan, H. and H. J. Rogers (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement 27*(4), 361–370.

Therneau, T., B. Atkinson, and B. Ripley (2014). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-8.

Thissen, D., L. Steinberg, and H. Wainer (1993). Detection of differential item functioning using the parameters of item response models. In *Differential Item Functioning*, pp. 67–113. Lawrence Erlbaum Associates.

Thissen-Roe, A. and D. Thissen (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics 38*(5), 522–547.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B 67*(1), 91–108.

Trepte, S. and M. Verbeet (2010). *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL Studentenpisa-Test*. Wiesbaden: VS Verlag.

Tutz, G. (1989). Compound regression models for categorical ordinal data. *Biometrical Journal 31*(3), 259–272.

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis 11*(3), 275–295.

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

Tutz, G. and M. Berger (2015a). Item focussed trees for the identification of items in differential item functioning. *Psychometrika, published online*. doi: 10.1007/s11336-015-9488-3.

Tutz, G. and M. Berger (2015b). Tree-structured modelling of categorical predictors in regression. *Cornell University Library*. arXiv:1504:04700.

Tutz, G. and M. Berger (2016a). Response styles in rating scales - simultaneous modelling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics 41*(3), 239–268.

Tutz, G. and M. Berger (2016b). Seperating location and dispersion in ordinal regression models. *Ludwig-Maximilians-Universität München, Department of Statistics*. Technical Report 190.

Tutz, G. and J. Gertheiss (2014). Rating scales as predictors – the old question of scale level and some answers. *Psychometrika 79*(3), 357–376.

Tutz, G. and J. Gertheiss (2016). Regularized regression for categorical data (with discussion). *Statistical Modelling, to appear*.

Tutz, G. and M. Oelker (2016). Modeling clustered heterogeneity: Fixed effects, random effects and mixtures. *International Statistical Review, published online*. doi: 10.1111/insr.12161.

Tutz, G. and G. Schauberger (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika 80*(1), 21–43.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2016). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification, published online*. doi: 10.1007/s11634-016-0247-9.

Umlauf, N., D. Adler, T. Kneib, S. Lang, and A. Zeileis (2015). Structured additive regression models: An r interface to bayesx. *Journal of Statistical Sofware 63*(21), 1–46.

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.

Van Herk, H., Y. H. Poortinga, and T. M. Verhallen (2004). Response styles in rating scales evidence of method bias in data from six eu countries. *Journal of Cross-Cultural Psychology 35*(3), 346–360.

Van Rosmalen, J., H. Van Herk, and P. Groenen (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research 47*(1), 157–172.

Van Vaerenbergh, Y. and T. D. Thomas (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research 25*(2), 195–217.

Verbeke and G. Molenberghs (2000). *Linear Mixed Models for longitudinal data.* New York: Springer–Verlag.

Weijters, B., E. Cabooter, and N. Schillewaert (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing 27*(3), 236–247.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73*(1), 3–36. R package version 1.8-7.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software 32*(10), 1–34. R package version 0.9-8.

Yee, T. W. (2014). *VGAM: Vector Generalized Linear and Additive Models.* R package version 0.9-4.

Yee, T. W. and C. J. Wild (1996). Vector generalized additive models. *Journal of the Royal Statistical Society B 58*(3), 481–493.

Yu, K., W. Wheeler, Q. Li, A. W. Bergen, N. Caporaso, N. Chatterjee, and J. Chen (2010). A partially linear tree-based regression model for multivariate outcomes. *Biometrics 66*(1), 89–96.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*(1), 49–67.

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

Zhang, H. and B. Singer (1999). *Recursive Partitioning in the Health Sciences.* New York: Springer–Verlag.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binaryand Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human ResourcesResearch and Evaluation, Department of National Defense.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 01. Juni 2016

Moritz Maximilian Berger