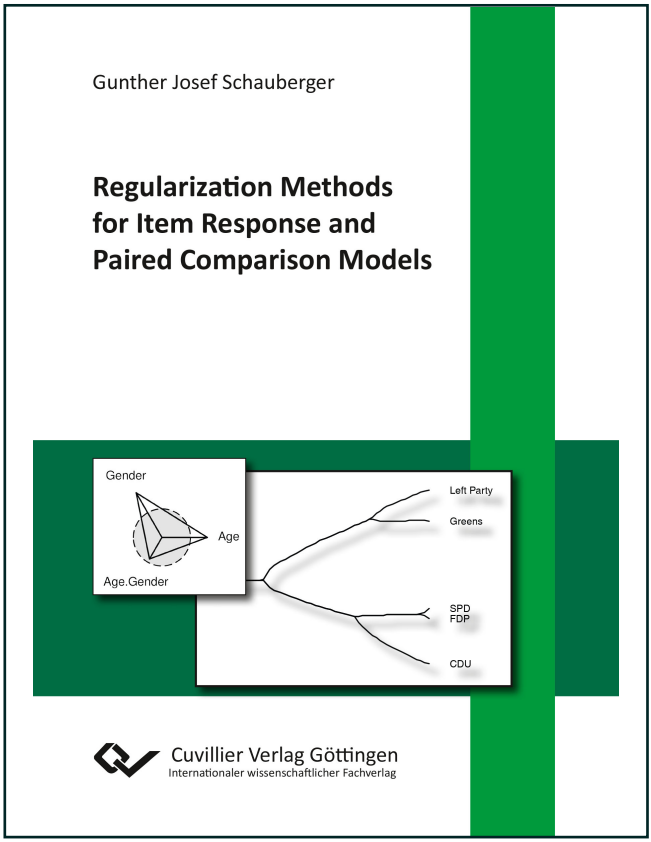
Gunther Schauberger (Autor)

Regularization Methods for Item Response and Paired Comparison Models



https://cuvillier.de/de/shop/publications/7152

1. Introduction

Commonly, item response models and paired comparison models are treated as different model classes, suited for different data situations. However, there is a great similarity between item response data and paired comparisons and, accordingly, between the respective modeling approaches. Item response data appear when test persons face a certain number of items which are designed to measure a specific latent trait of the test persons. Such latent traits can, for example, be certain skills (e.g. intelligence) of the test persons or attitudes towards a specific issue (e.g. xenophobia). In the simplest case only two outcomes are possible, for example right or wrong answers or approving or disapproving of a statement.

Paired comparison data occur if two objects or items compete in a certain way. The most frequent occurrence of paired comparisons is when two objects are presented and raters have to declare a preference for one or the other object. But also in other situations paired comparisons appear, as, for example, in sport competitions between two players or teams. Again, in the simple case only two outcomes are possible, namely the win/preference of one object over the other.

Both in item response data and in paired comparisons, the outcome refers to the result of a specific competition between two actors. Therefore, item response data can be seen as a special type of paired comparison data. Tutz (1989) distinguishes between homogeneous and heterogeneous paired comparisons. In this sense, item response data are heterogeneous paired comparisons as the matched pairs are pairs of one item and one respondent. In contrast, homogeneous paired comparisons treat matched pairs of two objects or items.

The basic and most popular models for these data are the Rasch model (RM) for item response data and the Bradley-Terry or Bradley-Terry-Luce model (BTL) for paired comparison data. The Rasch model (Rasch, 1960) assumes that the probability that a person solves an item is determined by the difference between one latent parameter representing the person and one latent parameter representing the item. Let the random variable Y_{pi} represent the response where $Y_{pi} = 1$ if person p solves item i and $Y_{pi} = 0$ otherwise. With the Rasch model the probability that person p solves item i is modeled by

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

where θ_p is the person parameter and β_i is the item parameter. In contrast, the Bradley-Terry model (Bradley and Terry, 1952) for a competition between two objects a_r and a_s models the probability that a_r beats a_s by

$$P(Y_{(rs)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}.$$

The parameters $\gamma_r, r = 1, \ldots, m$, are the trait parameters of the objects $\{a_1, \ldots, a_m\}$. The random variable $Y_{(rs)}$ denotes the response where $Y_{(rs)} = 1$ if object a_r is preferred over a_s and $Y_{(rs)} = 0$ otherwise.

Comparing these two basic models, "the direct relationship between the RM and the BTL is obvious" (Fischer and Molenaar, 1995). Both models are logit models, their linear predictors represent the difference between the latent traits of both actors. The main difference is, that the two actors are one item and one person for the Rasch model but two items for the Bradley-Terry model. In this thesis, both models for homogeneous and heterogeneous paired comparisons, in particular the Rasch model and the Bradley-Terry model, will be extended in various ways. The proposed extensions are supposed to allow for more flexibility in the modelling of item response and paired comparison data and for the inclusion of more information than in classical modelling approaches. A main focus will be on the inclusion of covariates.

All proposed methods will use regularization techniques for estimation. The main goal of regularization is to prevent overfitting and to allow for unique solutions in ill-posed problems, see Hastie et al. (2009) for an introduction into a broad variety of regularization methods. In this thesis, two different regularization techniques will be used, namely penalization and boosting. In penalization methods for regression models, the regular log-likelihood is maximized with respect to a certain side constraint. The resulting penalized likelihood

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

for a model with a general parameter vector $\boldsymbol{\beta}$ consists of the regular log-likelihood $l(\boldsymbol{\beta})$ and a penalty term $J(\boldsymbol{\beta})$ in combination with a tuning parameter λ. Famous examples for penalization methods are the ridge regression (Hoerl and Kennard, 1970) or lasso regression (Tibshirani, 1996). While ridge restricts the L_2 norm of the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ using the penalty term

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{p} \beta_i^2,$$

lasso restricts the L_1 norm of the parameter vector with the penalty term

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{p} |\beta_i|.$$

A main feature of penalization methods is shrinkage. The estimated coefficients are shrunk toward zero leading to a decreased variance of the estimates. In total, although the shrinkage effect goes along with biased estimates the decreased variance can lead to a decreased mean square error. Some penalization methods as, for example, the lasso also provide a dimension reduction in the covariate space. In the case of lasso, this means that lasso is able to provide parameter estimates equal to zero. Therefore, lasso allows for automatic parameter selection. In recent years, several penalty terms suited for different regression models and different data structures were developed.

Boosting evolved within the machine learning community rather than in the statistical modelling community. First approaches were proposed by Freund et al. (1996) and Tukey (1977). In the context of regression models, boosting was developed by Friedman et al. (2000) and extended, for example, by Bühlmann and Yu (2003) and Bühlmann and Hothorn (2007a). The main feature of boosting is the principle that many weak learners are combined into one joint and (hopefully) strong learner. In regression models, boosting combines many weak learners into a joint model. The main goal is to gradually improve a certain loss function, for example the L_2 loss or specific likelihood functions. In this context, a learner is considered to be a weak learner if it improves the respective loss function only by a little amount. This concept helps to avoid overfitting as the procedures is not supposed to be performed until convergence. Many boosting procedures, including the one proposed in this thesis, also allow for variable selection.

Guideline through the Thesis

This thesis consists of 10 chapters and three appendices. Chapters 2 and 3 contain general introductions into the most important topics treated in Chapters 4 and 5. Chapter 2 provides an introduction into the Rasch model, together with its most important assumptions and properties and the typical estimation methods. Chapter 3 gives a short introduction into the topic of differential item functioning. As Chapters 4 and 5 propose new methods for the detection of differential item functioning, Chapter 3 also presents some of the most popular methods for the detection of differential item functioning.

Chapter 4 proposes a new diagnostic tool for the identification of differential item functioning (DIF). In particular, an explicit model for differential item functioning is proposed that includes a set of variables. In contrast to most classical approaches to detect DIF, which

only allow to consider few (mostly two) subpopulations, the proposed model can handle both continuous and categorical covariates. The ability to include a set of covariates entails that the model contains a large number of parameters. Penalized maximum likelihood estimators are used to solve the estimation problem and to identify the items that induce DIF. It is shown that the method is able to detect items with DIF. Simulations and two applications demonstrate the applicability of the method.

Chapter 5 continues the idea from Chapter 4 to identify differential item functioning using several covariates at the same time and proposes a boosting algorithm instead of the penalized likelihood approach. The covariates can be both continuous and (multi-)categorical, and also interactions between covariates can be considered. The method works for the general parametric model for DIF in Rasch models proposed in Chapter 4. Since the boosting algorithm selects variables automatically, it is able to detect the items which induce DIF. It is demonstrated that boosting competes well with traditional methods in the case of subgroups. Furthermore, it outperforms the method proposed in Chapter 4 in the case of metric covariates. The method is illustrated by an extensive simulation study and an application to real data.

While Chapters 2-5 treat some basics and some new proposals in the context of item response data and the inclusion of covariates, the following chapters consider methods suited for paired comparison data. First, Chapter 6 introduces the basic Bradley-Terry model together with the most important existing extensions of the model.

In traditional paired comparison models heterogeneity in the population is simply ignored and it is assumed that all persons have the same preference structure. In Chapter 7, a new method to model heterogeneity in paired comparison data is proposed. The preference of an item over another item is explicitly modelled as depending on attributes of the subjects. Therefore, the model allows for heterogeneity between subjects as the preference for an item can vary across subjects depending on subject-specific covariates. Since by construction the model contains a large number of parameters we propose to use penalized estimation procedures to obtain estimates of the parameters. The used regularized estimation approach penalizes the differences between the parameters corresponding to single covariates. It enforces variable selection and allows to find clusters of items with respect to covariates. We consider simple binary but also ordinal paired comparisons models. The method is applied to data from a pre-election study from Germany.

In Chapter 8, a general paired comparison model for the evaluation of sport competitions is proposed. It efficiently uses the available information by allowing for ordered response categories and team-specific home advantage effects. Penalized estimation techniques are used to identify clusters of teams that share the same ability. The model is extended to include team-specific explanatory variables. Therefore, in contrast to Chapter 7, object-specific covariates are considered instead of subject-specific covariates. It is shown that regularization

techniques allow to identify the contribution of team-specific covariates to the success of teams. The usefulness of the method is demonstrated by investigating the performance and its dependence on the budget for football teams of the German Bundesliga.

In Chapter 9 an approach for the analysis and prediction of international soccer match results is proposed. In contrast to Chapter 8, the result of one match is not modeled as an ordered response. Instead, the number of scored goals is modeled directly using a Poisson distribution. To account for the paired comparison structure of the data, the linear predictor consists of differences between the covariates of both competing teams. Therefore, similar as in Chapter 8 object-specific covariates are included in the model. Lasso approaches are used to achieve variable selection and shrinkage. Based on preceding FIFA World Cups, two models for the prediction of the FIFA World Cup 2014 are fitted and investigated. Based on the model estimates, the FIFA World Cup 2014 is simulated repeatedly and winning probabilities are obtained for all teams. Both models favor the actual FIFA World Champion Germany.

In Chapters 4 and 5 the concept of effect stars is used to visualize parameter estimates for DIF items, in chapter 7 effect stars are used to visualize estimates from the proposed method BTLLasso. Originally, effect stars were proposed to visualize parameter estimates in categorical response models, in particular for multinomial and ordinal logit models. Therefore, in Appendix A the original concept of effect stars in the context of multinomial logit models is introduced. The multinomial logit model is the most widely used model for nominal multi-category responses. One problem with the model is that many parameters are involved, another that interpretation of parameters is much harder than for linear models because the model is non-linear. Both problems can profit from graphical representations. Effect stars visualize the effect strengths by star plots, where one star collects all the parameters connected to one term in the linear predictor. In contrast to conventional star plots, which are used to represent data, the plots represent parameters and are considered as parameter glyphs. The method is extended to ordinal models and illustrated by several data sets.

In order to keep the single chapters self-contained, every chapter contains separate introductions to the relevant topics and a separate conclusion. Therefore, every chapter can also be read separately but some topics will repeat themselves.

Contributing Manuscripts

Parts of this thesis were published as articles in peer reviewed journals, other parts were published in proceedings of scientific conferences or as technical reports at the Department of Statistics of the Ludwig-Maximilians-Universität München. In the following, chapter by

chapter all contributing manuscripts are listed together with a declaration of the personal contributions of the respective authors:

Chapter 4: Tutz and Schauberger (2015b). A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika 80*(1), 21 – 43

The project was initiated by Gerhard Tutz and further developed jointly by Gerhard Tutz and Gunther Schauberger. Gunther Schauberger implemented the method and performed the simulations and the real data analyses. Gunther Schauberger developed the corresponding R-package `DIFlasso`. The manuscript was written close collaboration of both authors. The original manuscript is extended by Section 4.6, which discusses concepts of variable selection within the proposed method. Apart from this section and some minor modifications Chapter 4 and Tutz and Schauberger (2015b) match. The technical report 134 (Tutz and Schauberger, 2012a) and the conference paper from the IWSM 2013 (Schauberger and Tutz, 2013) contain preliminary work on the project.

Chapter 5: Schauberger and Tutz (2015b). Detection of Differential Item Functioning in Rasch Models by Boosting Techniques. *British Journal of Mathematical and Statistical Psychology*, published online

The project was initiated jointly by Gerhard Tutz and Gunther Schauberger. Main author of the manuscript was Gunther Schauberger in close collaboration with Gerhard Tutz. Gunther Schauberger was responsible for the implementation of the method, of the simulation studies and the application to real data. Furthermore, Gunther Schauberger developed the corresponding R-package `DIFboost`. Apart from minor modifications Chapter 5 and Schauberger and Tutz (2015b) match. The conference paper from the IWSM 2014 (Schauberger and Tutz, 2014) contains preliminary work on the project.

Chapter 7: Schauberger and Tutz (2015c). Modelling Heterogeneity in Paired Comparison Data – an L_1 Penalty Approach with an Application to Party Preference Data. *Department of Statistics, LMU Munich*, Technical Report 183

The project was initiated and realized in close collaboration. Gunther Schauberger as the first author mainly wrote most of the manuscript and performed the presented analyses. He was also responsible for the implementation of the method and the corresponding R-package `BTLLasso`. The original manuscript is extended by Subsection 7.4.3 which discusses the inclusion of twofold interactions in the application and by a paragraph applying the concept of effect stars to the estimates of the proposed method. Apart from these extensions and minor modifications Chapter 7 and Schauberger and Tutz (2015c) match. The conference

paper from the IWSM 2015 (Schauberger and Tutz, 2015a) contains preliminary work on the project.

Chapter 8: Tutz and Schauberger (2015a). Extended Ordered Paired Comparison Models with Application to Football Data from German Bundesliga. *Advances in Statistical Analysis, 99*(2), 209 – 227

The manuscript was a joint project of Gerhard Tutz and Gunther Schauberger. Both authors contributed to the manuscript. The data collection and all implementations were done by Gunther Schauberger. The original manuscript is extended by Section 8.6 where the analyses from the previous sections are applied to the data from another Bundesliga season. Apart from this section and minor modifications Chapter 8 and Tutz and Schauberger (2015a) match. The technical report 151 (Tutz and Schauberger, 2014) contains preliminary work on the project.

Chapter 9: Groll, Schauberger, and Tutz (2015). Prediction of Major International Soccer Tournaments Based on Team-Specific Regularized Poisson Regression: An Application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports 11*(2), 97 – 115

Andreas Groll and Gunther Schauberger initiated and conducted the project in close collaboration. In particular, they were equally responsible for the data collection, the implementation of the methods and the manuscript. Gerhard Tutz supervised the methodological part of the manuscript and helped to improve the manuscript by extensive discussions. Apart from minor modifications Chapter 9 and Groll et al. (2015) match. The technical report 166 (Groll et al., 2014) contains preliminary work on the project.

Appendix A: Tutz and Schauberger (2013): Visualization of Categorical Response Models: From Data Glyphs to Parameter Glyphs. *Journal of Computational and Graphical Statistics, 22*(1), 156 – 177

The manuscript was mainly drafted by Gerhard Tutz with contributions of Gunther Schauberger. Gunther Schauberger was responsible for the implementation including the corresponding `R` package `EffectStars` (Schauberger, 2014b) and for all visualizations in the manuscript. He was strongly involved in all parts of the final manuscript. Apart from minor modifications Appendix A and Tutz and Schauberger (2013) match. The technical report 117 (Tutz and Schauberger, 2012a) and the conference paper from the IWSM 2012 (Schauberger and Tutz, 2012) contain preliminary work on the project.

Software

Most computations in this thesis were done with the statistical program `R` (R Core Team, 2015), parts were implemented in `C++` but are integrated in `R`. For most of the methods proposed in this thesis add-on packages for `R` were developed which can be downloaded from the Comprehensive R Archive Network (CRAN). In particular, the following `R`-packages were developed:

`DIFlasso` provides the method DIFlasso proposed in Chapter 4 (Schauberger, 2014a).

`DIFboost` provides the method DIFboost proposed in Chapter 5 (Schauberger, 2015b).

`BTLLasso` provides the method BTLLasso proposed in Chapter 7 (Schauberger, 2015a). The fitting algorithm of `BTLLasso` is implemented in `C++` code which is integrated into `R` using the packages `Rcpp` (Eddelbuettel, 2013) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014).

`EffectStars` provides the concept of effect stars proposed in Appendix A (Schauberger, 2014b).

2. The Rasch Model

In the following, the basic Rasch model (Rasch, 1960) will be explained in more detail. The Rasch model is considered to be a starting point of the item response theory (IRT) which over the last decades replaced the classical test theory (CTT) as the most popular method in the analysis of tests or questionnaires in general. The main difference between the CTT and the IRT is that the IRT models a probabilistic distribution of the correct response probability. The most general IRT model is the so called 3PL model (Birnbaum, 1968). It models the probability of a specified response depending on item parameters and a person parameter. Typically, such a specified response will simply be the (either correct or wrong) answer on a test question. If person $p,\ p = 1, \ldots, P$, tries to solve item $i,\ i = 1, \ldots, I$, the response is denoted as

$$Y_{pi} = \begin{cases} 1 & \text{person } p \text{ solves item } i \\ 0 & \text{otherwise} \end{cases}$$

Accordingly, the 3PL model is denoted by

$$P(Y_{pi} = 1) = c_i + (1 - c_i)\frac{\exp\left(a_i(\theta_p - \beta_i)\right)}{1 + \exp\left(a_i(\theta_p - \beta_i)\right)}.$$

Here, θ_p represents the person ability and β_i represents the item difficulty. The parameters c_i and a_i represent the guessing parameter and the discrimination parameter of item i. The model is called 3PL model as one item i is characterized by three item parameters, a_i, β_i, c_i. From the 3PL model, the 2PL model and the 1PL model can be obtained as special cases. In the 2PL model, it is assumed that no guessing is possible and the restriction $c_i = 0,\ i = 1, \ldots, I$ is applied. In the 1PL model (in the following referred to as the Rasch model), additionally equal discrimination parameters are assumed by restricting $a_i = 1,\ i = 1, \ldots, I$.

In the analysis of item response data, the Rasch model is the most popular choice. If person $p,\ p = 1, \ldots, P$, tries to solve item $i,\ i = 1, \ldots, I$, this is specified by the Rasch model by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

where θ_p represents the latent person ability and β_i represents the latent item difficulty. For identifiability, a restriction on the parameters is needed. Frequently, either a person parameter or an item parameter is set zero. Basically, the Rasch model simply represents a binomial logit model and can, therefore, easily be embedded into the framework of generalized linear models (GLMs) (McCullagh and Nelder, 1989). The Rasch model makes the person abilities and the item difficulties comparable. For example, if the ability of person p equals the difficulty of item i (i.e. $\theta_p = \beta_i$), the Rasch model will predict a probability of 0.5 that person p will solve item i.

2.1. Assumptions and Properties of the Rasch Model

The Rasch model is accompanied by four main assumptions, namely monotonicity, unidimensionality, conditional independence and sufficiency, compare Hatzinger (1989) and Kelderman (1984).

Monotonicity The solving probability $P(Y_{pi} = 1|\theta_p, \beta_i)$ is strictly monotone increasing for $\theta_p \in \mathbb{R}$. Furthermore, $P(Y_{pi} = 1|\theta_p, \beta_i) \to 0$ for $\theta_p \to -\infty$ and $P(Y_{pi} = 1|\theta_p, \beta_i) \to 1$ for $\theta_p \to \infty$ holds. Therefore, with increasing ability, the probability to solve an item increases.

Unidimensionality Given the item difficulty, the probability to solve an item solely depends on the true value of the respective person on the latent trait. That means that $P(Y_{pi} = 1|\theta_p, \beta_i, \phi) = P(Y_{pi} = 1|\theta_p, \beta_i)$ holds for any additional variable ϕ. Given the ability parameter and the item difficulty, the solving probability does not depend on any other variables ϕ.

Conditional independence Given the latent trait, the items have to be stochastically independent. Therefore, for equally able persons the solving probabilities for different items are independent. Solving one item does not increase or decrease the probability to solve another item. Conditional independence is also widely known as local independence.

Sufficiency The total score of a person $S_p = \sum_i Y_{pi}$ contains the entire information for the ability of the person. The score is a sufficient statistic for the person parameter θ_p, persons with the same score have the same ability. Accordingly, also the number of persons that solved an item i, namely $R_i = \sum_p Y_{pi}$, is a sufficient statistic for the item difficulty.

In the Rasch model (as in all IRT models), items can be visualized using so-called item characteristic curves (ICCs). An ICC shows the probability of a correct response on the respective item depending on the person parameter θ_p. Figure 2.1 exemplarily shows the

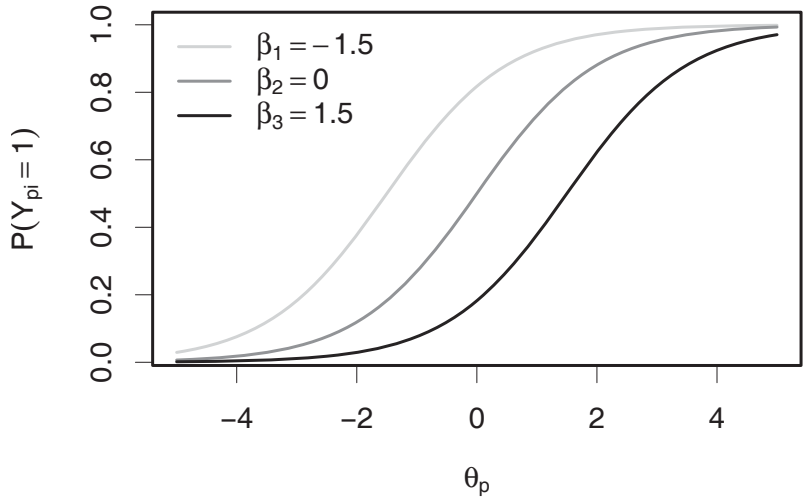


Figure 2.1.: Exemplary item characteristic curves for three items in a Rasch model

ICCs for three items with different item difficulties. The main feature of ICCs in Rasch models is that they all share the same form (they have the same slope) and are only shifted vertically depending on the respective item difficulty.

2.2. Estimation Approaches for the Rasch Model

To estimate the Rasch model, three different maximum likelihood approaches exist: Joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML). JML simultaneously provides estimates both for the person parameters and the item parameters. CML and MML only provide item parameters, person parameters have to be estimated separately.

Joint Maximum Likelihood Estimation

The joint maximum likelihood estimation of Rasch models is the easiest and most intuitive estimation method. If an appropriate design matrix is built, it can easily be performed using standard software for GLMs. Using the restriction $\theta_P = 0$, the design matrix can be seen from

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - \beta_i = \mathbf{1}_{P(p)}^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{1}_{I(i)}^{\mathrm{T}}\boldsymbol{\beta} = \boldsymbol{x}_{pi}^{\mathrm{T}}\boldsymbol{\delta},$$

where $\mathbf{1}_{P(p)}^{\mathrm{T}} = (0,\ldots,0,1,0,\ldots,0)$ has length $P-1$ with 1 at position p, $\mathbf{1}_{I(i)}^{\mathrm{T}} = (0,\ldots,0,1,0,\ldots,0)$ has length I with 1 at position i, and the parameter vectors are $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_{P-1})$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_I)$ yielding the total vector $\boldsymbol{\delta}^{\mathrm{T}} = (\boldsymbol{\theta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})$. The design vector linked to person p and item i is given by $\boldsymbol{x}_{pi}^{\mathrm{T}} = (\mathbf{1}_{P(p)}^{\mathrm{T}},-\mathbf{1}_{I(i)}^{\mathrm{T}})$. Finally, the Rasch model can be estimated by combining all single design vectors x_{pi} into a design matrix and by stacking all responses Y_{pi} appropriately into a response vector.

Estimation based on JML faces two main problems. First, if a person solves all or no items, its ability estimate will diverge to $\theta_p = \infty$ or $\theta_p = -\infty$, respectively. Equivalently, items that were solved by all or no persons will not have finite estimates although this case is much more unlikely as in general the number of persons clearly exceeds the number of items. After all, in both cases the respective person or item has to be removed from the design matrix. Second, the estimates for the item parameters from JML are inconsistent and biased for $P \to \infty$ and I fixed, see e.g. Andersen (1973b, 1980). Therefore, in recent years JML is decreasingly used in practice.

Conditional Maximum Likelihood Estimation

Nowadays, the conditional maximum likelihood method is the most popular choice. It is based on the property, that the sum score $S_p = \sum_i Y_{pi}$ of a person p is sufficient for the ability θ_p of person p. When conditioning on the sum scores the solving probabilities only depend on the item difficulties. Therefore, CML initially only provides estimates for the item parameters. Based on the item parameters, estimates for the person parameters can be obtained in a second step.

Let $\boldsymbol{y}_p = (y_{p1},\ldots,y_{pI})$ represent the response pattern of person p with the corresponding sum score $s_p = \sum_i y_{pi}$. Following Hatzinger (1989), the probability to observe the pattern $\boldsymbol{y}_p$, conditional on the respective sum score S_p, is denoted by

$$\begin{aligned} P(\mathbf{Y}_p = \boldsymbol{y}_p | S_p = s_p) &= \frac{P(\mathbf{Y}_p = \boldsymbol{y}_p)}{P(S_p = s_p)} \\ &= \frac{\exp(\theta_p s_p)\exp(-\sum_i \beta_i y_{pi})/\prod_p(1-\exp(\theta_p-\beta_i))}{\exp(\theta_p s_p)\gamma(s_p;\beta_1,\ldots,\beta_I)/\prod_p(1-\exp(\theta_p-\beta_i))}. \end{aligned} \tag{2.1}$$

Here, $\gamma(s_p;\beta_1,\ldots,\beta_I) = \sum_{\boldsymbol{y}|s_p}\exp(-\sum_i \beta_i y_{pi})$ represents the elementary symmetric function and $\boldsymbol{y}|s_p$ represents all possible response patterns with a sum score s_p. It can be seen

that all terms depending on θ_p can be eliminated from (2.1). Combining all possible sum scores $t = 0, \ldots, I$, the conditional likelihood can finally be denoted by

$$L_c = \frac{\exp(-\sum_i \beta_i r_i)}{\prod_t \gamma(t; \beta_1, \ldots, \beta_I)^{n_t}},$$

where $r_i = \sum_p y_{pi}$ denotes the number of persons that solved item i and n_t is the number of subjects with $s_p = t$. Maximizing the conditional likelihood provides consistent estimates for the item parameters when $P \to \infty$. Afterwards, the person parameters can be estimated assuming the item parameters to be fixed. The conditional maximum likelihood shares the problem of the joint maximum likelihood that for items solved by all or no persons and for persons that solved all or no items, no finite estimates can be found.

Marginal Maximum Likelihood Estimation

Similar to the conditional maximum likelihood approach, the marginal likelihood approach uses the trick to estimate the item parameters separately by eliminating the person parameters from the likelihood. In the case of the marginal likelihood, this is done by assuming a certain distribution for the person parameters. Typically, the person parameters are assumed to be normally distributed. With a given distribution, the person parameters can be integrated out from the likelihood.

The person parameters are assumed to be a random sample of the distribution $G(\theta)$. Then the probability to observe the pattern $\boldsymbol{y}_p$ can be denoted by

$$P(\boldsymbol{Y}_p = \boldsymbol{y}_p) = \int_{-\infty}^{\infty} P(\boldsymbol{y}_p|\theta_p) dG(\theta_p).$$

Using the parameters of the Rasch model, this can be denoted by

$$P(\boldsymbol{Y}_p = \boldsymbol{y}_p) = \exp(\beta_i r_i) \int_{-\infty}^{\infty} \frac{\exp(\theta_p s_p)}{\prod_{i=1}^{I}(1 - \exp(\theta_p - \beta_i))} dG(\theta_p).$$

Finally, the marginal likelihood is defined as product over all persons of the probability above. Then, the likelihood is a function depending on the item parameters and the distribution $G(\theta)$ and can be maximized with regard to the respective parameters. Due to the distributional assumption, using the marginal likelihood the estimates for the persons with perfect scores or scores of zero are finite.

3. Differential Item Functioning

Psychological or educational tests are typically used to investigate a latent trait of a person like the intelligence or other specific skills. For this purpose, appropriate items are needed to provide a valid measurement of the respective trait. Items are considered to be unfair if, for a specific item, two persons with the same underlying latent trait have different probabilities to answer the item correctly. Then, the item functions differently for two persons with the same value of the latent trait. Therefore, this phenomenon is called differential item functioning (DIF). In former publications, DIF was also denominated by the terms measurement bias or item bias, see, e.g., Lord (1980), Swaminathan and Rogers (1990) or Millsap and Everson (1993). Nowadays, the more neutral term of differential item functioning has widely prevailed.

Over the past decades, a vast amount of methods has been proposed to detect DIF. For an overview of the most popular methods see, e.g., Holland and Wainer (2012), Millsap and Everson (1993) or, more up to date, Magis et al. (2010). Typically, DIF is investigated by testing if special (known) characteristics of the participants like gender or ethnicity alter the probability to score on an item. Alternatively, also (unknown) latent classes could be assumed to describe DIF as proposed by Rost (1990). Here one assumes, that a model holds for all persons within a latent class but models for different classes differ. Since it is unclear what the latent classes represent, interpretation is rather hard and much less intuitive than for DIF between known groups. Therefore, latent class models have not become an established tool in DIF detection.

DIF can be divided into uniform and nonuniform DIF. Uniform DIF means that the difference between the solving probabilities for an item is constant along the person abilities for two equally able persons. In nonuniform DIF, the magnitude of the DIF effect depends on the respective person ability. Figure 3.1 exemplarily shows the item characteristic curves for items with uniform (left) and nonuniform (right) DIF between two subgroups of the population. It can be seen, that for nonuniform DIF the item characteristic curves can also be crossing. While the item is easier for group 1 than for group 2 on a low ability level it is harder on a high ability level. Within the context of IRT models, nonuniform DIF can be found in 2PL or 3PL models because only here the item characteristic curves can have different slopes. In case of the Rasch model introduced in Chapter 2, only uniform DIF is possible as all discrimination parameters are assumed to be fixed $a_i = 1,\ i = 1, \ldots, I$.