LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gunther Schauberger & Gerhard Tutz

Modelling Heterogeneity in Paired Comparison Data - an L_1 Penalty Approach with an Application to Party Preference Data

Technical Report Number 183, 2015
Department of Statistics
University of Munich

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

Modelling Heterogeneity in Paired Comparison Data - an L_1 Penalty Approach with an Application to Party Preference Data

Gunther Schauberger & Gerhard Tutz

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

gunther.schauberger@stat.uni-muenchen.de

August 18, 2015

Abstract

In traditional paired comparison models heterogeneity in the population is simply ignored and it is assumed that all persons have the same preference structure. Here, 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 measurements on 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.

Keywords: BTL-Lasso; paired comparison; Bradley-Terry-Luce model; Lasso; heterogeneity.

1 Introduction

Paired comparisons are a well established method to measure the relative preference or dominance of objects or items. The aim is to find the underlying

preference scale by presenting the items in pairs. The method has been used in various areas, for example, in psychology, to measure the intensity or attractiveness of stimuli, in marketing, to evaluate the attractiveness of brands, in social sciences, to investigate the value orientation (e.g. Francis et al. (2002)). In all these applications the items or stimuli are presented in an experiment. But paired comparison are also found in sports whenever two players or teams compete in a tournament. Then the non-observable scale to be found refers to the strengths of the competitors. Paired comparisons can also be obtained from ranked data (Francis et al., 2010) or from scale data (Dittrich et al., 2007). In this kind of data, respondents rank a predefined number of items or assign values from a Likert scale to the items, always referring to a certain attitude of the respondents towards the items. Building differences between the ranks or scales yields (binary or ordered) paired comparison data. We consider an application that shows how to analyse scales for the preference of parties by paired comparisons. In a German pre-election study the respondents were asked to scale the most renowned German parties. The focus of the analysis is on the inclusion of subject-specific covariates to account for the heterogeneity in the population and to investigate which variables determine the preference. More precisely, we investigate which clusters of parties are distinguished by specific covariates allowing that some covariates have no effect on the preference at all.

The most widely used model for paired comparison data is the Bradley-Terry-Luce model. It has been proposed by Bradley and Terry (1952) and is strongly linked to Luce's choice axiom (Luce, 1959). The basic model has been extended in various ways allowing for dependencies among responses, time dependence or simultaneous ranking with respect to more than one attribute. Overviews are found in the review of Bradley (1976), the monograph of David (1988) and more recently in the review of Cattelan (2012). The method proposed in this work can be applied both to binary and ordered response. Former approaches for ordered responses in paired comparisons include Tutz (1986) and Agresti (1992). Dittrich et al. (2004) also combine ordered responses and the inclusion of covariates, yet in a quite different modelling approach using log-linear models and without variable selection.

When persons choose between a pair of items most models assume that the strengths of the items are fixed and equal for all persons. Heterogeneity over persons has rarely been modeled explicitly. Exceptions are Turner and Firth (2012) or Francis et al. (2010), where categorical covariates are considered, but the application is very low dimensional with just two covariates, one with two and one with four categories. Also in Francis et al. (2002) covariates are included. Their model allows even for smooth effects of subject-specific covariates, but the fitting procedure that is proposed is also restricted to few variables. More recently, Casalicchio et al. (2015) presented a boosting approach that is able to include explanatory variables. An alternative approach has been proposed by Strobl et al. (2011). It is based on recursive partitioning techniques (also known

as trees) and automatically selects the relevant variables among a potentially large set of variables. The method proposed here is an alternative to handle the inherently high dimensional estimation problem that comes with the inclusion of explanatory variables. Maximum likelihood estimation is replaced by penalized estimation methods. By using a specific L_1-type penalty, the method is able to fit in high dimensional settings and to form clusters of items regarding the variables that generate heterogeneity.

In Section 2 the basic Bradley-Terry-Luce model for binary and ordered response is introduced. Then the model is extended to include subject-specific covariates. Section 3 contains the integration of the proposed model into the framework of generalized linear models and the penalty term is introduced. Section 3 also describes the implementation of the algorithm, the search for the optimal tuning parameter and the calculation of bootstrap confidence intervals. In Section 4 the application is given in detail.

2 Bradley-Terry Models with Covariates

2.1 The Basic Model

Let $\{a_1, \ldots, a_m\}$ denote the set of objects or items to be compared in a paired comparison experiment. The basic Bradley-Terry model (Bradley and Terry, 1952) specifies the probability that item a_r is preferred over a_s as

$$P(a_r \succ a_s) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)},$$

where, for reasons of identifiability, we use the restriction $\sum_{r=1}^{m} \gamma_r = 0$. The parameters $\gamma_r,\ r = 1, \ldots, m$, represent the attractiveness of the items $\{a_1, \ldots, a_m\}$. The interpretation as strength parameters is straightforward. For $\gamma_r = \gamma_s$, the probability that a_r is preferred over a_s is 0.5, for growing distance $\gamma_r - \gamma_s$ the probability increases.

With the random variable $Y_{(r,s)} = 1$ if $r \succ s$ and $Y_{(r,s)} = 0$ otherwise one obtains the logit model

$$\log \frac{P(Y_{(r,s)} = 1)}{P(Y_{(r,s)} = 0)} = \gamma_r - \gamma_s.$$

2.2 Bradley-Terry Models with Ordered Response

In some applications, paired comparison data can or should not be reduced to binary decisions. For example in sport events like football matches where also draws are possible, simple binary paired comparisons are not appropriate. A model that allows for ordinal responses is the cumulative Bradley-Terry-Luce model (Tutz, 1986) which has the form

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1 + \exp(\theta_k + \gamma_r - \gamma_s)} \tag{1}$$

with the same restriction $\sum_{r=1}^{m} \gamma_r = 0$.

The parameters $\theta_1, \ldots, \theta_K$ represent threshold parameters for the different levels of the response $Y_{(r,s)} \in \{1, \ldots, K\}$. The response $Y_{(r,s)} = 1$ corresponds to a strong preference of a_r over a_s and $Y_{(r,s)} = K$ corresponds to a strong preference of a_s over a_r. The basic Bradley-Terry model can be seen as a special case of model (1) for binary response with $K = 2$.

The strength parameters $\gamma_1, \ldots, \gamma_m$ have the same interpretation as in the binary model. With increasing γ_r the probability for low response categories, and therefore the strong preference of a_r over a_s is increasing while the probability for large response categories denoting dominance of a_s decreases. The threshold parameters determine the preference for specific categories. The threshold for the last category K is restricted to $\theta_K = \infty$ so that $P(Y_{(r,s)} \leq K) = 1$ holds. It is sensible to put further restrictions on the threshold parameters to ensure equal probabilities for corresponding categories if the order of the paired comparison is reversed. Therefore, we use the restrictions $\theta_k = -\theta_{K-k}$ and, if K is even, additionally $\theta_{K/2} = 0$. These restrictions ensure, for example, that $Y_{(r,s)} = 1$ (maximal preference of a_r over a_s) has the same probability as $Y_{(s,r)} = K$. Due to these restrictions, $\lfloor \frac{K-1}{2} \rfloor$ (free) threshold parameters have to be estimated. In the special case of binary response ($K = 2$) all threshold parameters are omitted and the model reduces to the ordinary Bradley-Terry model. If an order effect is required, for example to model the home advantage in sport competitions, an additional parameter can be included. For the application considered here no order effect is needed and therefore is omitted.

Formally, model (1) is a cumulative logit model, also called a proportional odds model. For a response variable consisting of K ordered categories, one models $K - 1$ cumulative probabilities $P(Y_{(r,s)} \leq 1), \ldots, P(Y_{(r,s)} \leq K - 1)$. The probability for a single response category is represented by the difference $P(Y_{(r,s)} = k) = P(Y_{(r,s)} \leq k) - P(Y_{(r,s)} \leq k - 1)$. Therefore, $P(Y_{(r,s)} \leq k)$ has to be greater or equal $P(Y_{(r,s)} \leq k - 1)$ for $k = 1, \ldots, K$ to have non-negative probabilities for all single categories. As the probabilities only differ with respect to the threshold parameters, this is ensured if $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$.

2.3 Heterogeneity in the Bradley-Terry Model

The models considered so far assume that all persons have the same preference structure. Heterogeneity in the population is simply ignored. A more sensible assumption is that preferences depend on covariates that characterize the person that chooses.

Let $Y_{i(r,s)}$ denote the response of person i for given pair of items (r, s) and $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, \ldots x_{ip})$ be a person-specific covariate vector. It is assumed that the

strength of the preference of item a_r for person i is determined by $\gamma_{ir} = \beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_r$. That means there is a global strength parameter β_{r0} but the effective strength is modified by the covariates. The parameter $\boldsymbol{\beta}_r^T = (\beta_{r1}, \ldots, \beta_{rp})$ contains the effect of the covariates on item a_r. The corresponding model has the form

$$\begin{aligned} P(Y_{i(r,s)} \leq k \mid \boldsymbol{x}_i) &= \frac{\exp(\theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\theta_k + (\beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_r) - (\beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + (\beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_r) - (\beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_s))} \\ &= \frac{\exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}} (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}} (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))} \end{aligned} \tag{2}$$

As in model (1), the sum-to-zero constraints $\sum_{r=1}^{m} \beta_{rj} = 0$ with $j = 0, 1, \ldots, p$ are used for identifiability.

The model allows for different preference structures in sub populations. For illustration let us consider the simple case where the person-specific variable codes a subgroup like gender, which has two possible values. Let $x_i = 1$ for males and $x_i = 0$ for females. Then the strengths parameters for item r are

$$\beta_{r0} + \beta_r \text{ for males and } \quad \beta_{r0} \text{ for females.}$$

The β_r represents the difference in attractiveness of item a_r between males and females. When items a_r and a_s are compared the dominance in the male population is determined by $(\beta_{r0} - \beta_{s0}) + (\beta_r - \beta_s)$, in the female population by $(\beta_{r0} - \beta_{s0})$. Thus the female population is like a reference population with dominance determined by the difference in the basic parameters $(\beta_{r0} - \beta_{s0})$. The preference in the male population is modified by the term $\beta_r - \beta_s$, and can be quite different. If one prefers a more symmetric representation one can choose $x_i = 1$ for males and $x_i = -1$ for females obtaining for the strengths parameters for item r

$$\beta_{r0} + \beta_r \text{ for males and } \quad \beta_{r0} - \beta_r \text{ for females.}$$

Then β_r represents the deviation of the attractiveness of item r from the baseline attractiveness β_{r0}. When items a_r and a_s are compared the dominance in the male population is determined by $(\beta_{r0} - \beta_{s0}) + (\beta_r - \beta_s)$, in the female population by $(\beta_{r0} - \beta_{s0}) - (\beta_r - \beta_s)$. Thus the difference of the basic parameters $\beta_{r0} - \beta_{s0}$ is augmented by $\beta_r - \beta_s$ in the male population and reduced by the same value in the female population.

The model accounts for the heterogeneity in the population by explicitly linking the attractiveness of alternatives to explanatory variables. The weight parameters $\boldsymbol{\beta}_r$ reflect how the attractiveness of a specific alternative depends on the covariates.

3 Penalized Estimation

The main problem with the general model (2) is the number of parameters that are involved. One has (with the given restrictions) $\lfloor \frac{K-1}{2} \rfloor$ threshold parameters and for each item the $(p+1)$-dimensional parameter vector $(\beta_{r0}, \boldsymbol{\beta}_r)$. In general, not all covariates might have a (different) influence on all m items. Therefore, we propose to use a penalized likelihood approach instead of ordinary maximum likelihood estimation to reduce the number of involved parameters and to select the relevant variables. In a first step we embed the estimation into the framework of generalized linear models (GLMs) and then introduce penalty terms.

3.1 Embedding into Generalized Linear Models

First, the ordinal Bradley-Terry model is embedded into the framework of Generalized Linear Models (GLMs). In the ordinal Bradley-Terry model without covariates the linear predictor $\eta_{(r,s)k} = \theta_k + \gamma_r - \gamma_s$ can be given as

$$\eta_{(r,s)k} = \theta_k + x_1^{(r,s)}\gamma_1 + \cdots + x_m^{(r,s)}\gamma_m = \theta_k + (\boldsymbol{x}^{(r,s)})^{\mathrm{T}}\boldsymbol{\gamma},$$

where $x_l^{(r,s)} = 1$ if $l = r$, $x_l^{(r,s)} = -1$ if $l = s$, and $x_l^{(r,s)} = 0$ otherwise, encodes the considered pair. The whole vector $\boldsymbol{x}^{(r,s)}$ has the simple form $\boldsymbol{x}^{(r,s)} = \mathbf{1}_r - \mathbf{1}_s$, where $\mathbf{1}_r = (0, \ldots, 0, 1, 0, \ldots, 0)$ has length m with 1 at position r. In this model the strength of an item is the same for all persons, which is a strong assumption ignoring potential heterogeneity.

In the general model with covariates, and therefore explicit modelling of heterogeneity, the linear predictor has the form

$$\begin{aligned}\eta_{i(r,s)k} &= \theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) \\ &= \theta_k + \sum_{j=0}^{p} x_{ij}(\beta_{rj} - \beta_{sj}) = \theta_k + \sum_{j=0}^{p}\sum_{l=1}^{m} x_{ij}x_l^{(r,s)}\beta_{lj}\end{aligned}$$

where $x_{i0} = 1$ is a fixed intercept. Here, $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ represents a covariate vector associated to person i and, therefore, the linear predictors for the same pair are different for persons. For $j > 0$ the predictor is determined by interactions between x_{ij} and the items, which reflects the underlying structure that the item strength is modified by the covariates.

The link between the linear predictor and the probability $P(Y_{i(r,s)} \leq k \mid \boldsymbol{x}_i)$ is determined by the logistic distribution function. It should be noted that the ordered response is transformed into a multivariate response $\boldsymbol{y}_{i(r,s)}^{\mathrm{T}} = (y_{i(r,s)1}, \ldots, y_{i(r,s)q})$ with $q = K-1$ binary variables where $y_{i(r,s)k} = 1$ if $Y_{i(r,s)} \leq k$ and $y_{i(r,s),k} = 0$ if $Y_{i(r,s)} > k$. With $\pi_{i(r,s)k} = \exp(\eta_{i(r,s)k})/(1 + \exp(\eta_{i(r,s)k}))$, the covariance structure for such a multivariate response is given by

$$Cov(\boldsymbol{y}_{i(r,s)}) = \begin{pmatrix} \pi_{i(r,s)1}(1-\pi_{i(r,s)1}) & \pi_{i(r,s)1}(1-\pi_{i(r,s)2}) & \cdots & \pi_{i(r,s)1}(1-\pi_{i(r,s)q}) \\ \pi_{i(r,s)1}(1-\pi_{i(r,s)2}) & \pi_{i(r,s)2}(1-\pi_{i(r,s)2}) & & \vdots \\ \vdots & & \ddots & \vdots \\ \pi_{i(r,s)1}(1-\pi_{i(r,s)q}) & \cdots & \cdots & \pi_{i(r,s)q}(1-\pi_{i(r,s)q}) \end{pmatrix}$$

Because of the restrictions $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$, the design matrix for the threshold parameters has a special form. As stated above, for a response with K categories, $\lfloor \frac{K-1}{2} \rfloor$ different threshold parameters have to be estimated. Therefore, the part of the design matrix corresponding to the paired comparison (r, s) of one person is a $(K-1) \times \lfloor \frac{K-1}{2} \rfloor$ matrix. This matrix is given by

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & -1 \\ \vdots & & \iddots & 0 \\ 0 & -1 & & \vdots \\ -1 & 0 & \cdots & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & -1 \\ \vdots & & \iddots & 0 \\ 0 & -1 & & \vdots \\ -1 & 0 & \cdots & 0 \end{pmatrix}$$

for K uneven or even, respectively. As stated above, for $K = 2$ the model reduces to a GLM with binomial distributed response and all threshold parameters are eliminated from the model.

3.2 Selection by Penalization

In regression models with β as the parameter vector penalization approaches maximize the penalized likelihood

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

where $l(\beta)$ is the usual log-likelihood and $J(\boldsymbol{\beta})$ is a penalty term that penalizes specific structures in the parameter vector. The parameter λ is a tuning parameter that specifies how seriously the penalty term has to be taken. A simple penalty term that could be used is the squared length of the parameter vector $J(\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = \sum \beta_i^2$, known as ridge penalty, see, for example Hoerl and Kennard (1970), Nyquist (1991), Segerstedt (1992), LeCessie (1992). Then, for $\lambda = 0$ maximization yields the ML estimate. If $\lambda > 0$ one obtains parameters that are shrunk toward zero. For appropriately chosen λ the ridge estimator stabilizes

estimates. A disadvantage of the ridge estimator is that it does not select variables. Thus no reduction of the model is obtained. An alternative penalty is the L_1-penalty, also known as lasso (Tibshirani, 1996), which is able to select variables. Instead of the squared parameters one penalizes the absolute values of the parameters with the penalty term $J(\boldsymbol{\beta}) = \sum |\beta_i|$. For penalized likelihood estimation, it is essential that all covariates are on comparable scales. Therefore, in the following it is assumed that all covariates are standardized.

However, the simple lasso cannot be used directly since penalty terms for paired comparison models have to account for the specific structure of the model. In particular, in model (2) one has the parameters of the regular (ordinal) BTL model, namely the threshold parameters and, for each item r, a parameter β_{r0} for its basic attractiveness. They form the basic model and, therefore, will not be penalized. In the general model one has additional parameters for the interaction between the items and the covariates. These parameters will be penalized to obtain the interactions that are actually needed. The proposed penalty term has the form

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{p} \sum_{r<s} w_{rsj} |\beta_{rj} - \beta_{sj}|,$$

where $r, s \in \{1, \ldots, m\}$, w_{rsj} is a weight parameter and the parameters are collected in $\boldsymbol{\alpha}^{\mathrm{T}} = (\theta_1, \ldots, \theta_{K-1}, \beta_{10}, \ldots, \beta_{mp})$. The penalty has the effect that the parameters referring to the same covariate are shrunk towards each other. For large values of λ, the differences are shrunk to exactly zero so that the effect of a covariate is the same for two (or more items). Therefore, the penalty yields clusters of items which share the same effect of a certain covariate. With growing tuning parameter, these clusters become bigger until all items form one single cluster. In that case, due to the sum-to-zero constraints all parameters are zero and the covariate is irrelevant for the attractiveness of the items. The penalty is a L_1-type fusion penalty rather than a simple lasso. Similar penalties have been used for the modelling of factors in GLMs by Bondell and Reich (2009), Gertheiss and Tutz (2010) and Oelker et al. (2014). More recently, penalties of this form have also been used in the modelling of paired comparison models, however, not for the modelling of heterogeneity by inclusion of covariates (Masarotto and Varin, 2012; Tutz and Schauberger, 2014).

For illustration, Figure 1 shows the coefficient paths corresponding to a covariate j for a toy example with $m = 5$ items. The paths are drawn along the (normed) penalty term $\sum_{r<s} |\beta_{rj} - \beta_{sj}|$ for covariate j. It can be seen that the penalty enforces a clustering of the items when the penalty is increased. In the unpenalized model, all items form clusters of their own. With increasing penalty, items 1 and 4 form a cluster, later item 3 is integrated into that cluster. Next, also items 2 and 5 form a cluster and finally all items form one single cluster. If all items share the same parameter (all parameters are zero) that means that

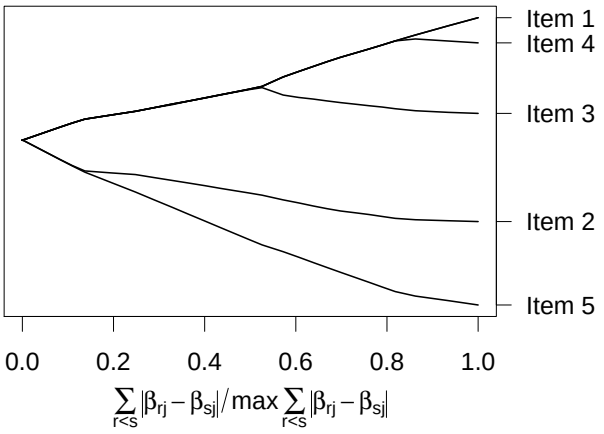

Figure 1: Exemplary coefficient paths for a covariate j in a setting with $m = 5$ different items

the respective covariate is eliminated from the model. Therefore, the proposed penalty term enforces both clustering of items and variable selection at the same time.

Zou (2006) proposed the so-called adaptive lasso as an extension of the regular lasso. In contrast to regular lasso, it provides consistency in terms of variable selection. In the adaptive lasso, the single penalty terms are weighted with the inverses of the unpenalized ML-estimates. In a similar way the weight parameters w_{rsj} are defined by $w_{rsj} = |\beta_{rj}^{\text{ML}} - \beta_{sj}^{\text{ML}}|^{-1}$. The effect is that small differences in the ML-estimates are penalized stronger than bigger differences which has the effect that the clustering of the parameters is enforced.

3.3 Implementation

L_1 penalized cumulative logit models have, e.g., been used in Archer and Williams (2012) and are implemented for `R` (R Core Team, 2015) in Archer (2014a) and Archer (2014b). However, these implementations are limited to lasso type penalties for coefficients. They cannot be used to penalize differences between parameters as required in the paired comparison case. Moreover, in order to obtain consistent estimates we want to include the weights w_{rsj}. For that purpose, a new fitting algorithm was implemented that is able to fulfill these requirements. It is based on the idea of approximating penalties proposed by Oelker and Tutz (2015), which is implemented in the `R`-package `gvcm.cat` (Oelker, 2015), yet not for cumulative logit models. For shorter computation time, the fitting algorithm itself is implemented in `C++` and integrated into `R` using the packages `Rcpp` (Eddelbuettel et al., 2011; Eddelbuettel, 2013) and `RcppArmadillo` (Eddelbuettel and

Sanderson, 2014). The code is available by request from the authors and should be available on CRAN soon.

3.4 Choice of Penalty Parameter

The performance of penalized estimation methods is essentially determined by the choice of the tuning parameter λ. It determines which covariates modify the attractiveness and forms the clusters within the chosen covariates. Mostly, two different approaches are used to determine tunings parameters, namely model selection criteria and cross-validation. Model selection criteria like the AIC (Akaike, 1973) or the BIC (Schwarz, 1978) try to find a compromise between the complexity of the model and the model fit. The complexity of a model is determined by its degrees of freedom. While for ML estimation, the degrees of freedom simply correspond to the number of parameters, the degrees of freedom for penalized likelihood approaches, in particular with a penalty applied on differences, are not straightforward. Therefore, we use cross-validation. In cross-validation, the data set is divided into a predefined number of subsets. Each subset is once used as a test data set while the remaining subsets serve as training data. The model is fitted (for a predefined grid of values for the tuning parameter λ) on the training data while the test data are used for prediction. Then, the predictive performance in the test data can be measured, for example by using the deviance. Moreover, this procedure provides a measure of the predictive performance of the model for every value from the predefined grid of tuning parameters. The tuning parameter with the best performance is chosen. We adapted this general principle to our specific case. The persons or subjects are treated as the observation units so that all paired comparisons corresponding to one person are in the same subset.

3.5 Confidence intervals

In contrast to maximum likelihood estimators, for estimators from penalized likelihood approaches one cannot use the information matrix to obtain standard errors or confidence intervals. Therefore, alternative techniques have to be used. We propose to use the bootstrap method for that purpose. The main idea of bootstrap is to replace an unkown distribution by the respective empirical distribution function. Then, for a predefined number of bootstrap iterations B, a subsample from the empirical distribution function is drawn. In our case, for a single bootstrap iteration, n persons are drawn from the original sample with replacement. The proposed procedure is applied to the sampled data set, including the model selection using cross–validation. Therefore, the additional variance originating from the process of model selection is incorportated in the resulting confidence intervals. Finally, for every parameter bootstrap confidence intervals can be calculated using the empirical $\alpha/2$ and $1 - \alpha/2$ quantiles from the B bootstrap estimates for the respective parameter.

4 Application to Pre-Election Data from Germany

The proposed method is applied to data from the German Longitudinal Election Study (GLES), see Rattinger et al. (2014). The GLES is a long-term study of the German electoral process. It collects pre- and post-election data for the several federal elections.

4.1 Data

The data we are using here originate from the pre-election survey for the German federal election in 2013. In this specific part of the study, the participants ($n = 1155$ after eliminating all incomplete cases) were asked to rank the most important parties (CDU/CSU, SPD, Greens, Left Party, FDP, we eliminated the smaller parties AfD and the Pirate Party) for the upcoming federal election on a scale from -5 to $+5$. Plass et al. (2015) used the data in the context of modelling approaches for undecidedness. The ranks Z_r reflect the general opinions of the participants of party r where $+5$ represents a very positive and -5 represents a very negative opinion. The main goal of this application is to analyse which characteristics of the participants are connected to the opinions of the single parties. For that purpose, we generated paired comparisons out of the rankings. A similar approach for the analysis of rank data using paired comparisons was proposed by Francis et al. (2010). They also discuss the advantages of a paired comparison approach to model this form of data. For each participant, the differences between the ranks of all parties were calculated, ending up with ordered paired comparisons with values between -10 and 10. The response was narrowed down to an ordered response with five categories. The data now represent paired comparisons between all parties measured on an ordered five-point scale:

$$\begin{aligned}
Z_r - Z_s \in \{6, 10\} &\mapsto Y_{(r,s)} = 1: \text{ "I strongly prefer party r over party s"} \\
Z_r - Z_s \in \{1, 5\} &\mapsto Y_{(r,s)} = 2: \text{ "I slightly prefer party r over party s"} \\
Z_r - Z_s = 0 &\mapsto Y_{(r,s)} = 3: \text{ "I have equal opinions of parties r and s"} \\
Z_r - Z_s \in \{-5, -1\} &\mapsto Y_{(r,s)} = 4: \text{ "I slightly prefer party s over party r"} \\
Z_r - Z_s \in \{-10, -6\} &\mapsto Y_{(r,s)} = 5: \text{ "I strongly prefer party s over party r"}
\end{aligned}$$

Within the GLES study, several characteristics of the participants are observed that possibly could affect the preference for the single parties. For our application, the following covariates are considered:

- Age: age of participant in years
- Gender: female (1); male (0)

- East Germany: East Germany/former GDR (1); West Germany/former FRG (0)
- Personal economic situation: good or very good (1); neither/nor, bad or very bad (0)
- School leaving certificate: Abitur/A levels (1); else (0)
- Unemployment: currently unemployed (1); else (0)
- Attendance in Church/Mosque/Synagogue/...: at least once a month (1); else (0)
- Have you been a German citizen since birth: yes (1); no (0)

4.2 Results

In the following, the results for the proposed method are presented for a model where all covariates described above are considered as possibly influential variables. The optimal model is determined by 10-fold cross-validation. Figure 2

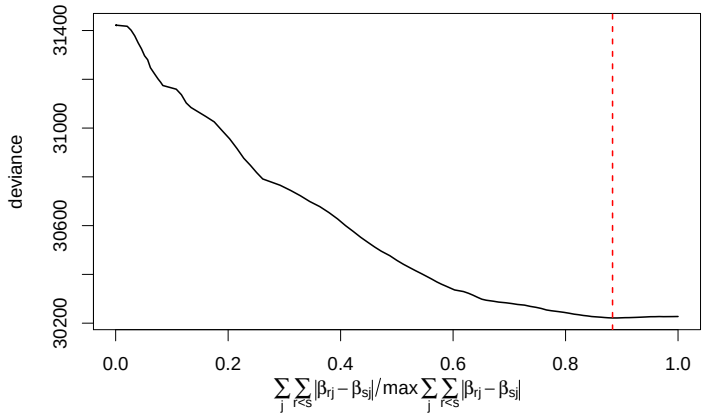

Figure 2: Deviance path for 10-fold cross-validation, dashed vertical line represents model with lowest deviance.

shows the deviances obtained by cross-validation plotted against the (normed) size of the penalized differences. Strong penalization corresponds to values close to 0, weak penalization to values close to 1. The dashed vertical line represents the model with the lowest deviance. Figure 3 shows the corresponding coefficient paths for the threshold parameters θ_1 and θ_2 and the party-specific intercepts $\beta_{10}, \ldots, \beta_{m0}$. These parameters are not penalized. In principle, they might be

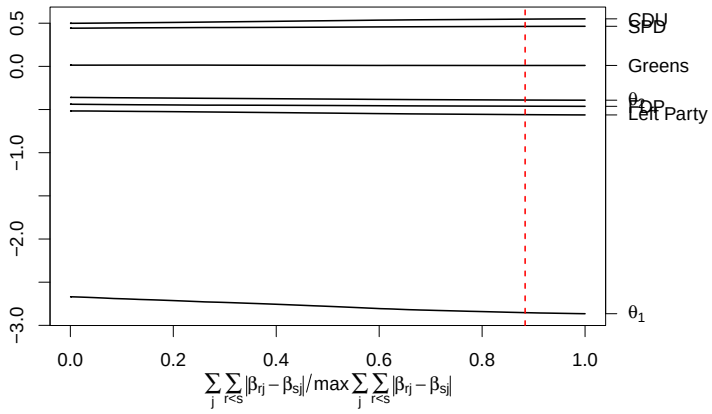

Figure 3: Coefficient paths for all unpenalized parameters (threshold parameters θ_1 and θ_2 and party-specific intercepts). Dashed vertical line represents optimal model according to 10-fold cross-validation.

different for different tuning parameters λ. In the current application, it is seen that both the threshold parameters and the intercepts hardly change along their paths.

Figure 4 shows the corresponding coefficient paths for the eight covariates. The coefficient paths are drawn separately for each covariate. It is seen how the penalty term enforces clustering of the different parties. The dashed vertical lines represent the optimal model according to the 10-fold cross-validation.

The coefficient paths allow for interesting insights into how the preference of the voters for certain parties depends on characteristics of the voters themselves. Let us first consider the covariate unemployment. With respect to unemployment, the parties can be divided into two main clusters. The Left party and the Greens in one cluster, CDU, SPD and FDP in another cluster. As a global tendency one sees that unemployed persons tend to prefer the younger parties (Greens and Left Party) while the tendency to the more established parties (SPD, CDU, FDP) is reduced. In the optimal model, the second cluster of parties can be further divided into a cluster of SPD and FDP and a cluster only consisting of CDU. For gender, four different clusters are identified in the final model. The Greens are much more attractive for female than for male voters and form a cluster of their own. The SPD and the Left party seem almost equally attractive for males and females while the CDU and the FDP are more attractive for males. For the variable school leaving certificate, a very sparse solution with only two clusters (Greens vs. all other parties) emerged confirming the reputation of the Greens to be a party for academics. The German citizenship was completely eliminated from the model, naturalized citizens do not systematically prefer other parties

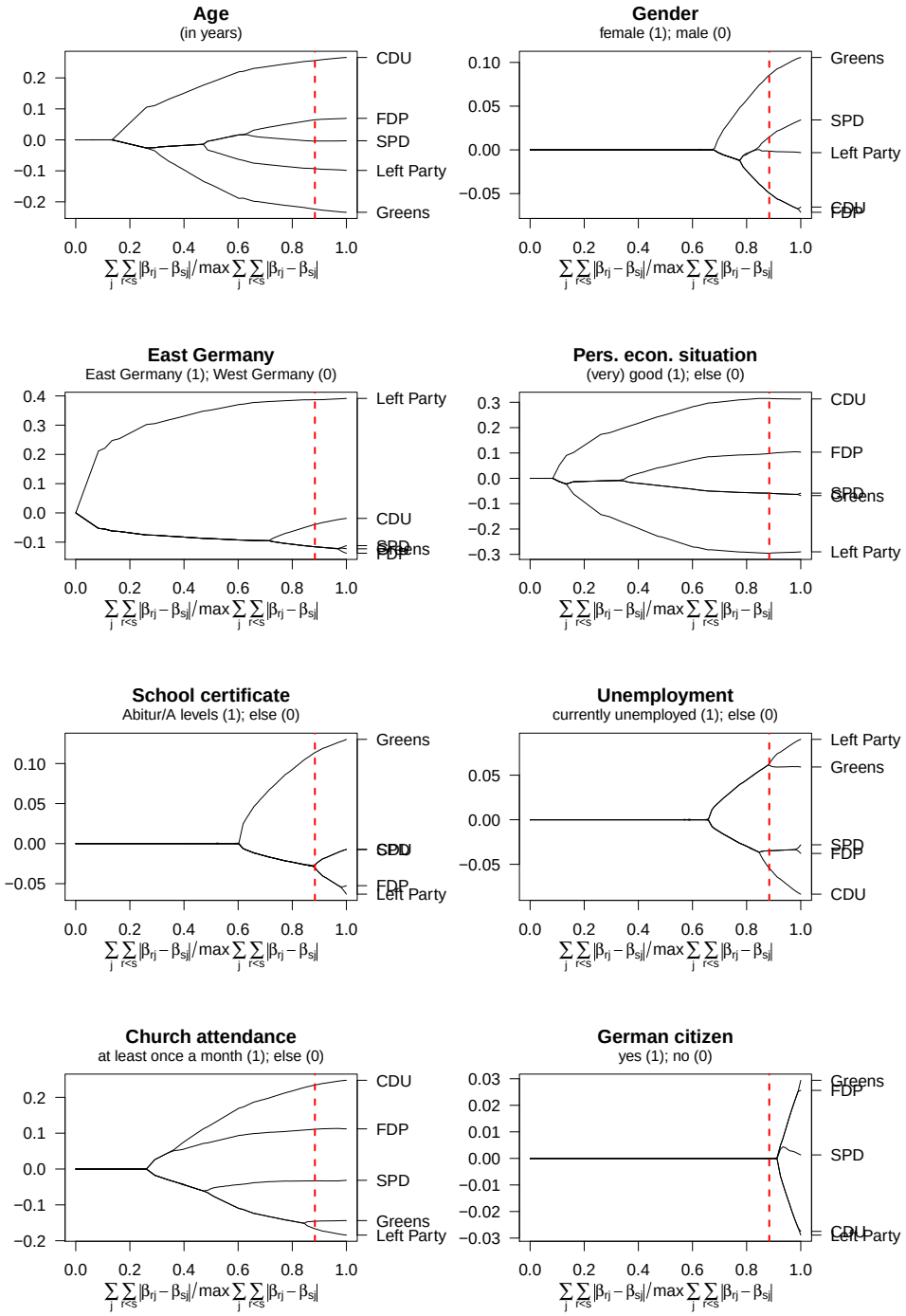

Figure 4: Coefficient paths separately for all eight covariates. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

than citizens that were German citizens since birth. The variables age and church attendance have a specific impact on the preference of parties and every party forms a cluster of its own.

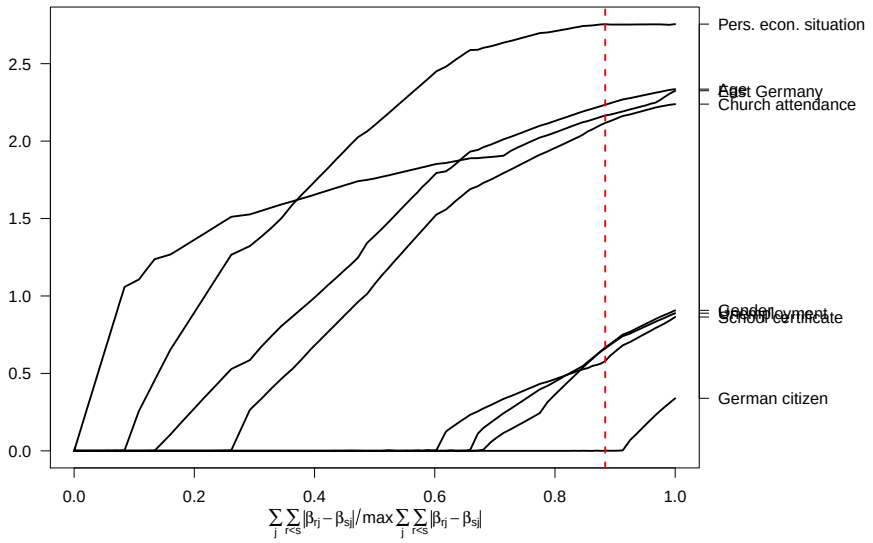

Figure 5: Paths representing the sums of absolute differences for all eight covariates. Dashed vertical line represents optimal model according to 10-fold cross-validation.

Figure 5 shows the paths for whole covariates represented by the sum of absolute differences between all parameters corresponding to one covariate. Every covariate is represented by a single path. With the used penalty term, the sum of the absolute differences between all parameters corresponding to one covariate can be seen as a measure of effect strength for this covariate. Again, one has to keep in mind that all covariates have been standardized. It can be seen that, not very surprisingly, the personal economic situation of the voters is the most important modifier of the preference of a party in the data set. Yet, the first covariate that is included (for decreasing tuning parameter λ) is the covariate East Germany. Even 23 years after the German reunification, the differences between the former GDR and the former FRG were still extremely relevant in 2013. Also the covariates age and church attendance have very strong effects. Again, it can be seen that the variable German citizenship since birth is eliminated from the model. Figure 5 can provide valuable additional information on the paths depicted in Figure 4 where the variable importance is harder to recognize due to the different scales in the single plots.

Finally, $B = 500$ bootstrap iterations were performed to receive confidence intervals. Figure 6 depicts the estimates of all (penalized) parameters together

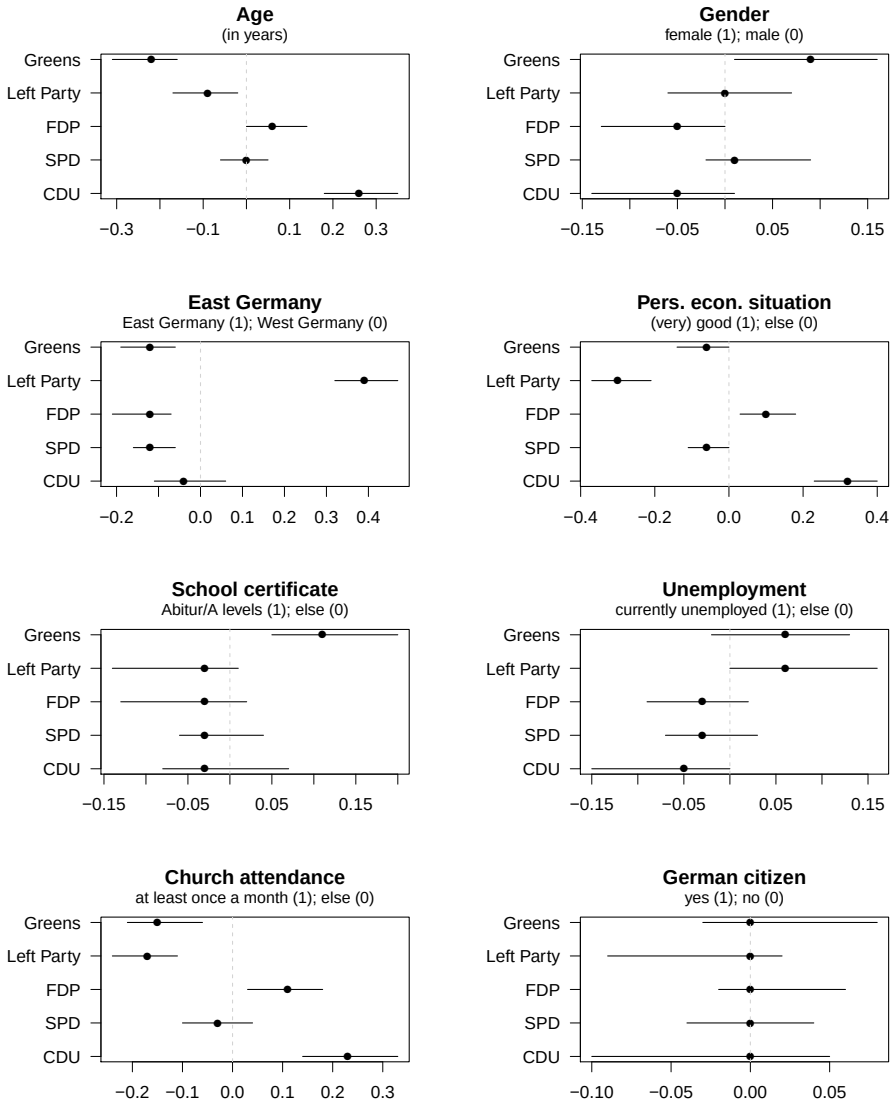

Figure 6: Parameter estimates and 95% bootstrap confidence intervals separately for all eight covariates.

with the corresponding 95% bootstrap confidence intervalls. It can be seen if two clusters differ significantly from each other. For example, the parameters for the Left party and the Greens are not significantly different for church attendance although they are splitted into two different clusters. For the covariate unemployment, no parameter is significantly different from zero although three different clusters were estimated. Except for the covariates unemployment and German citizenship, for all other covariates at least one coefficient differs significantly from zero.

5 Concluding Remarks

A model that explicitly accounts for heterogeneity in (possibly ordered) paired comparison models is proposed. The heterogeneity is modeled by the incorporation of subject-specific covariates. The model is estimated using a specific L_1-type penalty. The penalty has two main features: First, the penalty clusters items with regard to certain covariates. Therefore, one can identify clusters of items whose preferences are equally affected by a covariate. Second, the penalty can eliminate whole covariates from the model indicating that the respective covariates do not affect the preference for one or another item. Bootstrap intervals can be calculated which can be used to check if certain parameters differ significantly.

In particular the ability to select and cluster distinguishes the method from the few methods that are able to include covariates in paired comparison models. Francis et al. (2010) and Francis et al. (2002) include covariates but do not select the relevant ones, Casalicchio et al. (2015) presented a boosting approach that is able to select explanatory variables but is unable to detect clusters. Moreover, an advantage of penalty methods over boosting approaches is that the structure of the regularization is more clearly defined. In contrast to Strobl et al. (2011), where the underlying structure is searched for by recursive partitioning techniques, we consider a parametric model that allows for easy interpretation of parameters and clustering.

The proposed method could be extended in various ways. First, the restriction of the covariate effects to linear terms could be weakened by allowing for smooth covariate effects. A big challenge with such an approach would be to find an appropriate penalty term to have a similar cluster effect as for the linear terms. Second, the model could be extended by item-specific covariates similar to Tutz and Schauberger (2014). For the application to the data from the GLES in this work, this would correspond to the inclusion of party-specific covariates, for example the popularity of the respective leading candidates.

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics 41*(2), 287–297.

Akaike, H. (1973). In B. Petrov and F. Caski (Eds.), *Information Theory and the Extension of the Maximum Likelihood Principle*, Second International Symposium on Information Theory, Budapest. Akademia Kiado.

Archer, K. J. (2014a). *glmnetcr: Fit a penalized constrained continuation ratio model for predicting an ordinal response.* R package version 1.0.2.

Archer, K. J. (2014b). *glmpathcr: Fit a penalized continuation ratio model for predicting an ordinal response.* R package version 1.0.3.

Archer, K. J. and A. A. A. Williams (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine 31*(14), 1464–1474.

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics 65*, 169–177.

Bradley, R. A. (1976). Science, statistics, and paired comparison. *Biometrics 32*, 213–232.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs, I: The method of pair comparisons. *Biometrika 39*, 324–345.

Casalicchio, G., G. Tutz, and G. Schauberger (2015). Subject-specific bradley-terry-luce models with implicit variable selection. *Statistical Modelling*.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science 27*(3), 412–433.

David, H. A. (1988). *The method of paired comparisons, 2nd ed.* Griffin's Statistical Monographs & Courses 41, Griffin, London.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2007). A paired comparison approach for the analysis of sets of Likert-scale responses. *Statistical Modelling 7*(1), 3–28.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser (2004). A log-linear approach for modelling ordinal paired comparison data on motives to start a phd programme. *Statistical Modelling 4*(3), 181–193.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp.* Springer.

Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software 40*(8), 1–18.

Eddelbuettel, D. and C. Sanderson (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis 71*, 1054–1063.

Francis, B., R. Dittrich, and R. Hatzinger (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do europeans get their scientific knowledge? *The Annals of Applied Statistics 4*(4), 2181–2202.

Francis, B., R. Dittrich, R. Hatzinger, and R. Penn (2002). Analysing partial ranks by using smoothed paired comparison methods: an investigation of value orientation in europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 51*, 319–336.

Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorial explanatory variables. *Annals of Applied Statistics 4*, 2150–2180.

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

LeCessie (1992). Ridge estimators in logistic regression. *Applied Statistics 41*(1), 191–201.

Luce, R. D. (1959). *Individual Choice Behaviour*. New York: Wiley.

Masarotto, G. and C. Varin (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics 6*(4), 1949–1970.

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

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., J. Gertheiss, and G. Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling 14*(2), 157–177.

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.

Plass, J., P. Fink, N. Schöning, and T. Augustin (2015). Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision - extended version.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rattinger, H., S. Roßteutscher, R. Schmitt-Beck, B. Weßels, and C. Wolf (2014). Pre-election cross section (GLES 2013). *GESIS Data Archive, Cologne ZA5700 Data file Version 2.0.0*.

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

Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Communications in Statistics – Theory and Methods 21*, 2227–2246.

Strobl, C., F. Wickelmaier, and A. Zeileis (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics 36*(2), 135–153.

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

Turner, H. and D. Firth (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software 48*(9), 1–21.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology 30*, 306–316.

Tutz, G. and G. Schauberger (2014). Extended ordered paired comparison models with application to football data from german bundesliga. *AStA Advances in Statistical Analysis*, 1–19.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.