LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz & Gunther Schauberger

# Extended Ordered Paired Comparison Models with Application to Football Data from German Bundesliga

Technical Report Number 151, 2014
Department of Statistics
University of Munich

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

# Extended Ordered Paired Comparison Models with Application to Football Data from German Bundesliga

Gerhard Tutz & Gunther Schauberger

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

{gerhard.tutz}@stat.uni-muenchen.de

January 7, 2014

**Abstract**

A general paired comparison model for the evaluation of sports 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. It is shown that regularization techniques allow to identify the contribution of explanatory variables to the success of teams. The usefulness of the methods is demonstrated by investigating the performance and its dependence on the budget for football teams of the German Bundesliga.

**Keywords:** Paired comparison systems, Penalized Estimation, Bradley-Terry model

## 1 Introduction

Bayern Munich has been the dominating team in the last season of the German football league Deutsche Bundesliga. The dominance can be seen from the ranking according to the final points order. In the Bundesliga the winning team gains 3 points, the loosing team receives nothing, and both teams gain 1 point if the match is drawn. This scheme of distributing points according to the outcome of the match can be seen as an ad hoc measure of the strengths of teams. But

it is not without problems. In particular, if a team wins it is irrelevant if the adversary was a weak or a strong team. In the same way, each team gains one point in a draw, although, if the difference in strengths is large, the performance is weak for the stronger team but strong for the weaker team. A more elaborate way to measure the strength of teams is by considering the strength of a team as a latent trait and the performance, that is, the observable results, as determined by the latent traits of both teams. Models of this type have some tradition in statistics, in particular Bradley-Terry (BT-) models have been used to model competitions. Proposed by Bradley and Terry (1952), the model has been widely used to measure underlying strength in sport competitions. Dynamic models were considered, for example, by Fahrmeir and Tutz (1994), Knorr-Held (2000), Glickman and Stern (1998), and, more recently, by Cattelan et al. (2013).

In this paper, we analyse the results of the German Bundesliga. We use a general latent trait model that does not only account for draws but allows for ordinal response categories that represent the competition results, thereby aiming at the efficient use of the information in the data. The model also includes an effect that represents the advantage in playing at home, which can also vary over teams. Aspects of the model have been already proposed in the literature. Models that allow for a draw were proposed by Rao and Kupper (1967), Davidson (1970) and used to model sports tournaments by Cattelan et al. (2013), models that allow for any number of ordered response categories were proposed by Tutz (1986) and Agresti (1992). Heterogeneity of the home advantage has been considered by Kuk (1995), Knorr-Held (2000), and Glickman and Stern (1998), but only for models with a draw. An approach to find clusters of teams that can not be distinguished has been proposed by Masarotto and Varin (2012). Here, it is extended to work in the general model and also to find clusters of teams with the same home advantage.

In a second step it is investigated how much of the variation in the strengths of the teams is explained by team-specific covariates. It is especially interesting how much of the strength of a team is explained by the budget. Is Bayern Munich the best team because it is the richest club in Germany? For the analysis the estimated strength parameters are used and a model that includes effects of covariates is proposed. Estimation is based on penalization methods that allow to group the abilities of teams. We analyse the German Bundesliga data and demonstrate that the model with explanatory variables yields useful estimates.

In Section 2 we briefly describe the data. In Section 3 we introduce the general ordinal model and give results for the Bundesliga. Section 5 is devoted to the inclusion of team-specific explanatory variables.

## 2 German Bundesliga

Before defining latent trait models, which will be quite general for the modelling of competition results, we briefly describe the system German Bundesliga. The tournament comprises $m = 18$ teams, we analyse the matches played in the 50th season of the Bundesliga from August 24, 2012 to May 18, 2013 . The tournament structure is that of a double round-robin, each team competes twice against all the other teams, once on home ground and once away. On average, 42.5% of the matches were won by the home team, 25.5% of the matches ended with a draw and 32% of the matches were won by the away team. Table 1 shows the results ranked according to the final points order.

| | **Points** | **Home** | **Away** | **Ability** | **QSE** | **Rank** |
|---|---|---|---|---|---|---|
| FC Bayern München | 91 | 44 | 47 | 2.562 | 0.377 | 1 |
| Borussia Dortmund | 66 | 33 | 33 | 1.361 | 0.314 | 2 |
| Bayer 04 Leverkusen | 65 | 39 | 26 | 0.983 | 0.306 | 3 |
| FC Schalke 04 | 55 | 33 | 22 | 0.460 | 0.300 | 4 |
| Eintracht Frankfurt | 51 | 31 | 20 | 0.350 | 0.300 | 6 |
| Sport-Club Freiburg | 51 | 28 | 23 | 0.409 | 0.300 | 5 |
| Hamburger SV | 48 | 26 | 22 | 0.023 | 0.300 | 11 |
| Borussia Mönchengladbach | 47 | 29 | 18 | 0.235 | 0.300 | 7 |
| Hannover 96 | 45 | 32 | 13 | 0.074 | 0.300 | 9 |
| 1. FC Nürnberg | 44 | 27 | 17 | 0.057 | 0.300 | 10 |
| VfB Stuttgart | 43 | 19 | 24 | -0.183 | 0.302 | 13 |
| VfL Wolfsburg | 43 | 17 | 26 | 0.000 | 0.300 | 12 |
| 1. FSV Mainz 05 | 42 | 26 | 16 | 0.084 | 0.300 | 8 |
| SV Werder Bremen | 34 | 20 | 14 | -0.272 | 0.303 | 14 |
| FC Augsburg | 33 | 20 | 13 | -0.562 | 0.307 | 16 |
| 1899 Hoffenheim | 31 | 19 | 12 | -0.616 | 0.308 | 17 |
| Fortuna Düsseldorf | 30 | 21 | 9 | -0.287 | 0.303 | 15 |
| SpVgg Greuther Fürth | 21 | 4 | 17 | -0.956 | 0.315 | 18 |

TABLE 1: *Final ranking of the German Bundesliga 2012/2013 including points in home matches and away matches; the last three columns show the estimated abilities, quasi standard errors and the ranking corresponding to the estimated abilities for the ordered model including a home advantage parameter*

Two aspects from the final ranking are unique occurences in the history of the Bundesliga. Bayern Munich was the dominating team for the season and set several new records. For example, Bayern Munich gained the highest number of points and victories for a team in one season. For the Spielvereinigung Greuther Fürth, it was the first participation in the German Bundesliga, they were the first team without a victory on home ground for a whole season.

# 3 Ordered Paired Comparison Model with Home Advantage

In the following, latent trait models are considered. The basic concept is that winning or loosing is the result of the underlying strengths of teams. While the strengths are fixed the result of a competition is a random variable. The models can be used in all competitions where two teams or players compete in a tournament like tennis, football, and chess. In some sports there is a clear winner, in others draws can occur. Another feature that depends on the form of competition is that home effects can occur. In particular, in football playing at the home ground seems to be advantageous. We will consider a general model that can account for all these effects.

## 3.1 The Basic Binary Bradley-Terry Model

Let $\{a_1, \ldots, a_m\}$ denote the set of teams or players that compete. In the simplest case when a team can only win or loose the relation between the underlying strengths of the teams and the outcome can be modeled by the Bradley-Terry model (Bradley and Terry, 1952), which specifies for the probability that $a_r$ dominates $a_s$

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

The parameters $\gamma_r, r = 1, \ldots, m$, can be interpreted as the strengths of the teams $\{a_1, \ldots, a_m\}$. For $\gamma_r = \gamma_s$ the probability that $a_r$ wins against $a_s$ is 0.5, for growing distance $\gamma_r - \gamma_s$ the probability increases accordingly.

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

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

where the conditioning on the given pair $(a_r, a_s)$ is suppressed and the dependence on the teams is contained in the subscript of $Y_{rs}$. The model in this form is not identifiable because strengths parameters $\gamma_r + c$ for fixed value $c$ yield the same probabilities. Therefore, a constraint is needed. We choose to fix one parameter, that is, $\gamma_m$ is set to zero. In our case the reference team is Wolfsburg.

## 3.2 Ordinal Models including the Advantage in Playing at Home

Let now the success of team $r$ in a competition between team $r$ and $s$ be measured on an ordinal scale represented by $Y_{rs} \in \{1, \ldots, k\}$, for odd $k$, where low numbers denote dominance of team $r$ and high numbers dominance of team $s$. The scale is assumed to be symmetric regarding the two teams. That means the numbers 1 to $k$ represent categories like "strong dominance of team $r$", "weak dominance of team $r$" "draw", "weak dominance of team $s$", "strong dominance of team $s$". In

the simplest case, where $k = 3$, the responses are "team $r$ wins", "draw", "team $s$ wins". But to exploit the information contained in the results of matches one might also consider the differences in scored goals as indicators of dominance. In the application we use a difference of at least 2 goals as an indicator for strong dominance and work with a 5-point scale. A model that allows for ordered responses is the cumulative type model

$$P(Y_{rs} \leq t) = F(\eta_{rst}), \quad \eta_{rst} = \theta_t + \gamma_r - \gamma_s, \tag{1}$$

where $F(.)$ is a symmetric distribution function, which in Bradley-Terry type models is the logistic distribution function. The linear predictor $\eta_{rst}$ contains the difference in strengths $\gamma_r - \gamma_s$ and so-called threshold parameters that account for the frequency of the response categories. The symmetry of the response categories entails the restrictions $\theta_t = -\theta_{k-t}$, $t = 1, \ldots, [k/2]$. That means, in particular, that for teams with identical strengths, $\gamma_s = \gamma_r$, one obtains $P(Y_{rs} = t) = P(Y_{rs} = k + 1 - t)$. For the most important case $k = 3$ one obtains $P(Y_{rs} = 1) = P(Y_{rs} = 3)$, that means that the probability of winning is the same for both teams. Similar restrictions are needed if the number of response categories $k$ is even, which is relevant only in competitions that do not allow for a draw (see Tutz (1986)).

The cumulative model (1) is able to use the information contained in ordered responses; with more categories better estimates are to be expected. In the literature alternative models have been proposed. In particular, the adjacent category models, proposed by Agresti (1992) is an alternative that also uses the full information in ordinal data. It is an extension of the three category model of Davidson (1970), which can also be estimated within a log linear model framework (Dittrich et al., 2004). Further applications of the adjacent categories model are found in Dittrich et al. (2000), Böckenholt and Dillon (1997a) and Böckenholt and Dillon (1997b).

### Home Effects

When modelling competitions one also has to account for the advantage deriving from playing at home. Therefore, the linear predictor is extended to

$$\eta_{rst} = \alpha + \theta_t + \gamma_r - \gamma_s,$$

where $\alpha > 0$ represents the home effect. It increases the probability for low response categories that correspond to the dominance of team $r$. It is easily derived that for $k = 3$ and equal strength, $\gamma_r = \gamma_s$, $\alpha$ reflects the proportion of odds for winning of team $r$ and winning of team $s$,

$$\alpha = \frac{1}{2} \log \frac{P(Y_{rs} = 1)/(1 - P(Y_{rs} = 1))}{P(Y_{rs} = 3)/(1 - P(Y_{rs} = 3)}.$$

However, it is questionable that the home effect is the same for each team. Some teams may profit more from playing at home than others. A team-specific home effect is obtained by using the predictor

$$\eta_{rst} = \alpha_r + \theta_t + \gamma_r - \gamma_s.$$

In this general model the $\gamma$-parameters do not represent the strengths of teams per se because performance depends on whether playing at home or not. Again, for $k = 3$ and equal strength, $\gamma_r = \gamma_s$, the home effect when playing at the home ground of team $r$ is given by the proportion of odds for winning (of team $r$) against loosing

$$\alpha_r = \frac{1}{2} \log \frac{P(Y_{rs} = 1)/(1 - P(Y_{rs} = 1))}{P(Y_{rs} = 3)/(1 - P(Y_{rs} = 3)}.$$

But in the general model, the proportion of odds for winning (of team $r$) against loosing when playing at the home ground of the second team $s$ are not just the inverse of the proportion when playing at the home of team $r$ as in the model with constant home effect.

By defining $\tilde{\gamma}_r = \alpha_r + \gamma_r$, the predictor obtains the form $\eta_{rst} = \theta_t + \tilde{\gamma}_r - \gamma_s$. As in the basic model (1), the result of a match is determined by the difference of strength, but now it is $\tilde{\gamma}_r - \gamma_s$. Therefore, $\tilde{\gamma}_r$ represents the strength when playing at home and $\gamma_r$ the strength when not playing at home.

## 3.3 Fitting the Model

Estimation of the cumulative model can be embedded into the framework of generalized linear models (GLMs), which were thoroughly investigated by McCullagh and Nelder (1989). For data $Y_{rs} \in \{1, \ldots, k\}$, $r, s \in \{1, \ldots, m\}$ the linear predictor can be written as

$$\eta_{rst} = \alpha_r + \theta_t + \gamma_r - \gamma_s = \alpha_r + \theta_t + x_2^{(r,s)}\gamma_2 + \cdots + x_m^{(r,s)}\gamma_m = \alpha_r + \theta_t + (\boldsymbol{x}^{(r,s)})^T\boldsymbol{\gamma},$$

where the components of the $(m-1)$-vector $\boldsymbol{x}^{(r,s)}$ are given by

$$x_j^{(r,s)} = \begin{cases} 1 & j = r \\ -1 & j = s \\ 0 & \text{otherwise.} \end{cases}$$

Thus, it is a cumulative model with threshold $\theta_t$, the additional parameter $\alpha_r$ and "predictor" $\boldsymbol{x}^{(r,s)}$. The predictor can also be given by $\boldsymbol{x}^{(r,s)} = \mathbf{1}_r - \mathbf{1}_s$, where $\mathbf{1}_r = (0, \ldots, 0, 1, 0, \ldots, 0)$ has length $m - 1$ with 1 at position $r$. Cumulative models have been considered in particular by McCullagh (1980), estimation within the framework of multivariate GLMs was considered by Fahrmeir and Tutz (2001), Tutz (2012). The embedding into this framework allows to use the familiar goodness-of-fit statistics as well as likelihood ratio statistics to test hypotheses.

## 3.4 Football Data

We first consider the modelling of the football data under the assumption that the home advantage is global, that is, it does not depend on the team. Then, one obtains one strength parameter for each team and has not to distinguish between the strength when playing at home or away. In the following, we try to use the available information by using a 5-point scale to evaluate the performance in a competition. The categories refer to "winning with a difference of at least two goals", "winning with a difference of less than two goals" and "draw" as the middle category.

### Global Home Effect Model

The estimated home advantage is $\hat{\alpha} = 0.293$; for the threshold parameters one obtains $\hat{\theta}_1 = -\hat{\theta}_4 = -1.66$ and $\hat{\theta}_2 = -\hat{\theta}_3 = -0.65$. If one assumes that two teams have equal abilities, the threshold parameters correspond to probabilities of 0.41 for a victory of the home team, 0.31 for a draw and 0.28 for a victory of the away team. Thus the home advantage can definitely not be ignored. The tendency is also seen from the averages over all games, because 42.5% of the matches were won by the home team, 25.5% of the matches ended with a draw and 32% of the matches were won by the away team. But these numbers are averages over games played by teams with differing abilities. The strength of the latent trait model is that the home advantage takes this variation of abilities into account when estimating the home advantage. Table 1 shows the estimated abilities together with the ranks according to the final points. It is seen that for the best teams the rank is in accordance with the estimated abilities but in the middle part of the table there are some permutations. However, quasi standard errors, computed following Firth and De Menezes (2004) suggest that the permutations are not to be taken too seriously. This will be investigated in more detail in Section 4.

### Team-Specific Home Effects

The question if home effects are team-specific is investigated by computing the likelihood ratio test for the hypothesis that all effects are equal, yielding a value of 24.69 on 17 degrees of freedom, which corresponds to a $p$-value of 0.102. Therefore it is not significant when using significance level 0.05, but nevertheless it is small. If one uses a 3-point scale that only distinguishes between "winning", "draw" and "loosing", the $p$-value is 0.022, which is definitely smaller. In Table 2 the estimates and the corresponding ranks are given when one distinguishes between home and away strength. As always in the applications we use the more informative 5-point scale. It is seen that for the best performers the order is very stable. It is the same when playing at home or away or when not distinguishing between the two. But one also finds large differences. For example, Hannover has rank 4 at home, but rank 17 when playing away with a difference of 1.167 in abilities. For Wolfsburg

the ranks are just the opposite, it has rank 17 at home and rank 4 when playing away.

| | Overall | | Home | | Away | |
|---|---|---|---|---|---|---|
| | **Ability** | **Rank** | **Ability** | **Rank** | **Ability** | **Rank** |
| FC Bayern München | 2.562 | 1 | 1.871 | 1 | 2.220 | 1 |
| Borussia Dortmund | 1.361 | 2 | 0.901 | 2 | 0.729 | 2 |
| Bayer 04 Leverkusen | 0.983 | 3 | 0.851 | 3 | 0.013 | 3 |
| FC Schalke 04 | 0.460 | 4 | 0.191 | 6 | -0.505 | 6 |
| Eintracht Frankfurt | 0.350 | 6 | 0.258 | 5 | -0.782 | 11 |
| Sport-Club Freiburg | 0.409 | 5 | -0.068 | 8 | -0.334 | 5 |
| Hamburger SV | 0.023 | 11 | -0.490 | 11 | -0.708 | 8 |
| Borussia Mönchengladbach | 0.235 | 7 | -0.020 | 7 | -0.722 | 9 |
| Hannover 96 | 0.074 | 9 | 0.338 | 4 | -1.505 | 17 |
| 1. FC Nürnberg | 0.057 | 10 | -0.140 | 9 | -0.999 | 14 |
| VfB Stuttgart | -0.183 | 13 | -1.024 | 16 | -0.564 | 7 |
| VfL Wolfsburg | 0.000 | 12 | -1.262 | 17 | 0.000 | 4 |
| 1. FSV Mainz 05 | 0.084 | 8 | -0.293 | 10 | -0.782 | 10 |
| SV Werder Bremen | -0.272 | 14 | -1.014 | 15 | -0.803 | 12 |
| FC Augsburg | -0.562 | 16 | -0.881 | 13 | -1.541 | 18 |
| 1899 Hoffenheim | -0.616 | 17 | -0.966 | 14 | -1.486 | 16 |
| Fortuna Düsseldorf | -0.287 | 15 | -0.695 | 12 | -1.204 | 15 |
| SpVgg Greuther Fürth | -0.956 | 18 | -2.278 | 18 | -0.906 | 13 |

TABLE 2: *Comparison of the estimated abilities from the model with a global home advantage to the estimated abilities from the model with team-specific home advantages*

### Ranks and Abilities

The traditional measure for the performance of teams is the number of gained points summarized over all games. It is interesting to investigate, how this measure that is defined by the association of the football league is related to the abilities found by the fitting of a latent trait model. To our surprise, we found that the correlation is quite high. For the 50th season we obtained a correlation of 0.982, which means that gained points and abilities measure almost the same. One may wonder if this is an effect of the specific scheme, which gives winning team 3 points, the loosing team nothing, and both teams 1 point if the match is drawn. Is this scheme appropriate under the assumption that the latent trait model is an adequate representation of the link between the observations and the latent abilities? Therefore, we shortly investigate how the scheme of distributing points influences the correlation between number of points and estimated abilities. In a general scheme, the winning team gains $w > 0$ points, the loosing team nothing and both teams $d > 0$ points if the match is drawn. It is easily derived that for constant proportion $w/d$ one obtains up to a scaling factor the same number of points. Because a scaling factor is irrelevant when

computing the correlation, it suffices to vary only one of the two parameters $w$ and $d$. Without loss of generality we set $d = 1$. Figure 1 shows the dependence of the correlation on the gained points for winning $w$ (bold faced curve). It is seen that the maximum is obtained for $w = 2.2$, which is not far from the 3 points fixed in the regulations. The surprise is in the slow decrease of the curve beyond its maximum. Given that the estimated abilities measure the strength of a team also much higher points could be given to the winning team and still the number of points is in strong accordance with the abilities. The strong correlation found for the data could be related to the double round robin structure of the tournament. In pair comparisons, where not all pairs are evaluated, we expect lower correlations. To investigate the effects we have drawn sub samples of the pair comparisons containing 50% of the pairs. Two specific sub samples are the results of the first round and the second round. Figure 1 shows the corresponding correlations. It is seen that, depending on the sample, correlations can be much smaller. That means, in particular, for an ongoing season, when not all matches have been played, the ranking by points and abilities are less strongly connected.

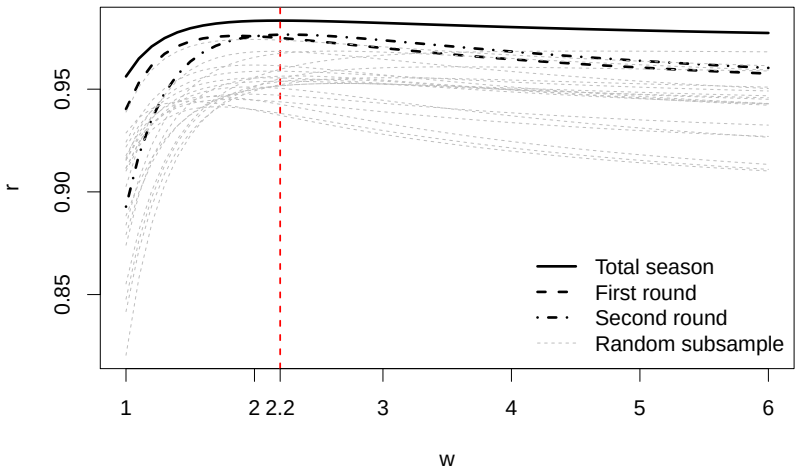


FIGURE 1: *Correlations plotted against the points gained when winning for the whole season (bold faced curve), first round (dashed), second round (dashed dotted) and several sub samples.*

## 4 Identification of Clusters

A disadvantage of simply measuring the performance of teams by points is that there is no information on the precision of this measurement tool. In contrast the

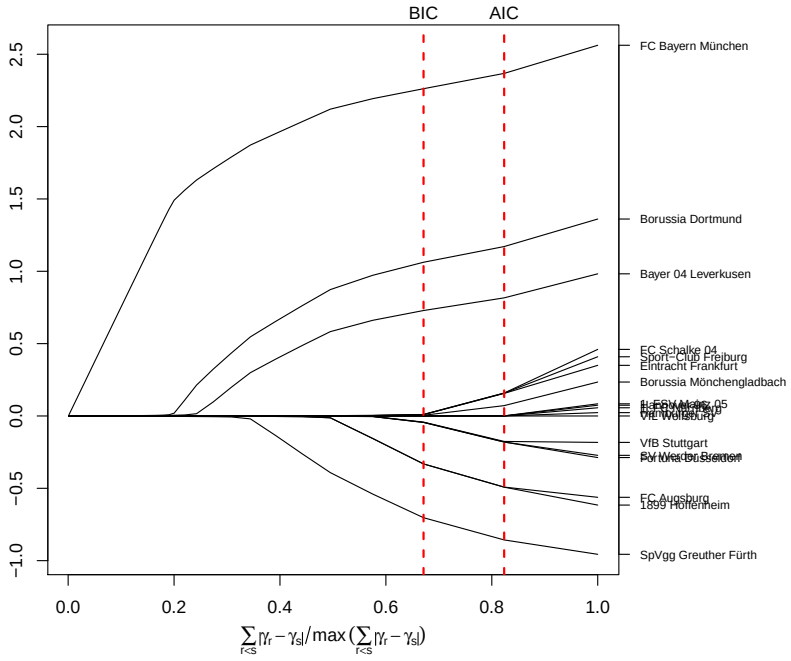


FIGURE 2: *Coefficient paths for ability parameters in the model with a global home advantage using an adaptive $L_1$-penalty.*

| **Cluster** | | **Ability** |
|---|---|---|
| 1 | FC Bayern München | 2.26 |
| 2 | Borussia Dortmund | 1.06 |
| 3 | Bayer 04 Leverkusen | 0.73 |
| 4 | FC Schalke 04; Sport-Club Freiburg; Eintracht Frankfurt | 0.01 |
| 5 | Borussia Mönchengladbach; 1. FSV Mainz; Hannover 96; 1. FC Nürnberg; Hamburger SV; VfL Wolfsburg | 0.00 |
| 6 | VfB Stuttgart; SV Werder Bremen; Fortuna Düsseldorf | -0.04 |
| 7 | FC Augsburg; 1899 Hoffenheim | -0.33 |
| 8 | SpVgg Greuther Fürth | -0.70 |

TABLE 3: *Clusters of teams with corresponding abilities.*

latent trait model allows to evaluate which teams are really to be distinguished. One way is to consider the standard errors, which contain the information about the relevance of differences between the estimated abilities. An alternative approach is to explicitly aim at finding clusters of teams which share the same ability by using regularization techniques. Clustering techniques proposed by Bondell and Reich (2009) and Gertheiss and Tutz (2010) have been used by Masarotto

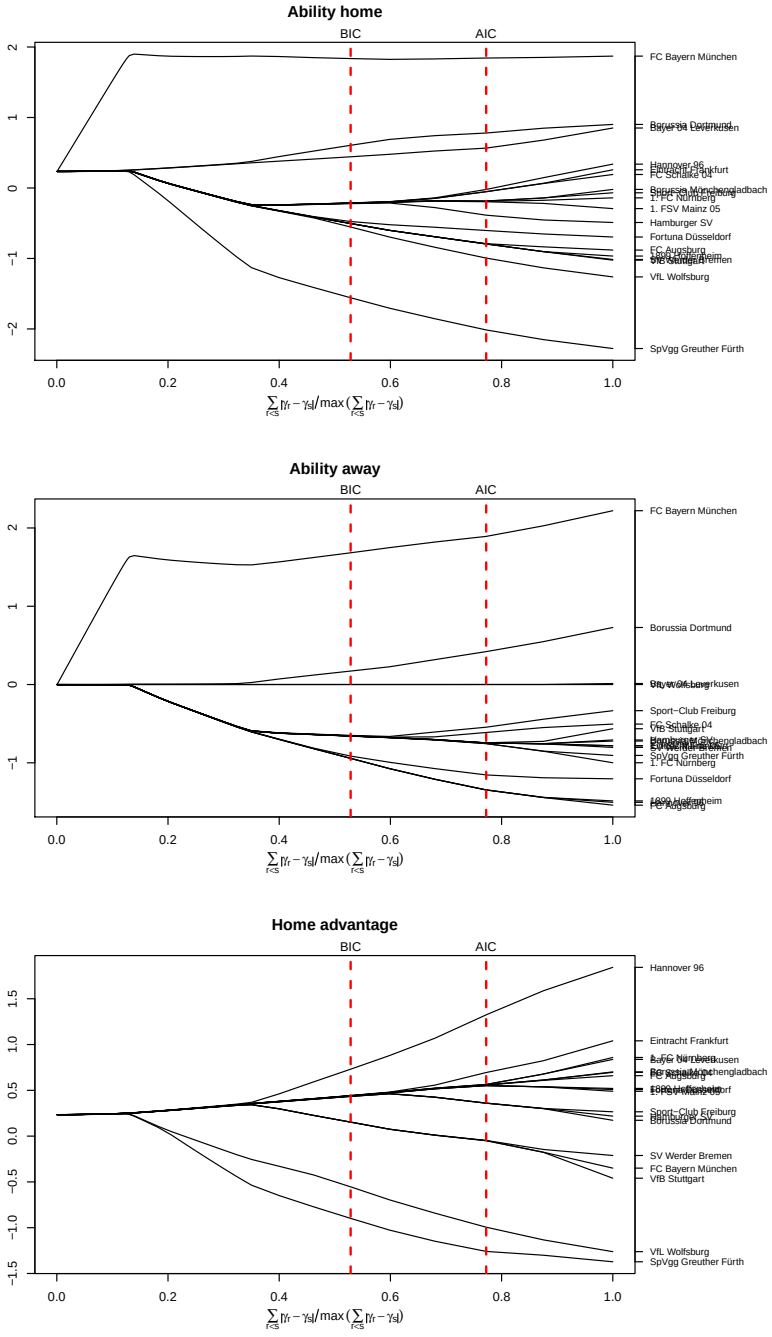


FIGURE 3: *Coefficient paths for home abilities, away abilities and home advantages in the model with team specific home advantages using an adaptive $L_1$-penalty*

and Varin (2012) to cluster abilities in a paired comparison model which allows for draws. In the next section we will use these techniques in the general case of ordinal response data. In Section 4.2 the method is extended to find clusters of abilities as well as clusters of home advantages.

## 4.1 Clustering of Teams

One way of obtaining regularized estimates is to use penalty terms that yield structured estimates. Instead of maximizing the log-likelihood, one maximizes the penalized log-likelihood

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

where $l(\boldsymbol{\beta})$ denotes the familiar un-penalized log-likelihood, $\lambda$ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a penalty term. A specific penalty term, which enforces the clustering of abilities and which will also be useful later, is given by

$$J(\boldsymbol{\beta}) = \sum_{r<s} w_{rs}|\gamma_r - \gamma_s|, \tag{2}$$

where $w_{rs}$ are specific weights. The penalty is a fusion type penalty, which enforces the fusion of abilities. By using the $L_1$-norm it enforces, in particular, that for growing $\lambda$ abilities are set equal. The effect of the penalty is also seen by looking at extreme values of the tuning parameter $\lambda$. If $\lambda \to \infty$, all strength parameters $\gamma_r$ are estimated as identical.

In the case of a global home advantage the procedure typically yields distinct clusters. Figure 2 shows the coefficient paths with the weights given by $w_{rs} = |\hat{\gamma}_r^{(\mathrm{ML})} - \hat{\gamma}_s^{(\mathrm{ML})}|$, where $\hat{\gamma}_r^{(\mathrm{ML})}$ denotes the maximum likelihood estimate of team $r$. For details of this weighting scheme, which yields more stable coefficient paths than un-weighted fusion penalties, see Gertheiss and Tutz (2010) and Masarotto and Varin (2012). The straight lines in Figure 2 represent the BIC (Schwarz, 1978) and the AIC (Akaike, 1974) criterion. Based on the BIC criterion one finds that the 18 teams are divided into eight clusters with abilities being identical within clusters. Table 3 shows the clusters and the corresponding estimated abilities. It is seen that the three best teams and the two worst teams form clusters of their own. All other teams are collected in three big clusters, which have rather similar abilities. In fact, if one measures abilities only up to one digit, they form just one big cluster.

## 4.2 Clustering of Teams and Home Effects

Clustering becomes much more difficult if one suspects team-specific home advantages because then one has to distinguish the strength when playing at home

| Cluster (ability home) | | Ability |
|---|---|---|
| 1 | FC Bayern München | 1.84 |
| 2 | Borussia Dortmund | 0.61 |
| 3 | Bayer 04 Leverkusen | 0.44 |
| 4 | Hannover 96; FC Schalke 04; Eintracht Frankfurt; Sport-Club Freiburg | -0.21 |
| 5 | Borussia Mönchengladbach; 1. FC Nürnberg; 1. FSV Mainz 05; Hamburger SV | -0.22 |
| 6 | Fortuna Düsseldorf | -0.48 |
| 7 | FC Augsburg; SV Werder Bremen; VfB Stuttgart; 1899 Hoffenheim | -0.50 |
| 8 | VfL Wolfsburg | -0.55 |
| 9 | SpVgg Greuther Fürth | -1.56 |
| **Cluster (ability away)** | | **Ability** |
| 1 | FC Bayern München | 1.68 |
| 2 | Borussia Dortmund | 0.17 |
| 3 | Bayer 04 Leverkusen; VfL Wolfsburg | 0.00 |
| 4 | Sport-Club Freiburg; FC Schalke 04 | -0.65 |
| 5 | Borussia Mönchengladbach; VfB Stuttgart; Hamburger SV; Eintracht Frankfurt; 1. FSV Mainz 05; SV Werder Bremen; 1. FC Nürnberg; SpVgg Greuther Fürth | -0.66 |
| 6 | Fortuna Düsseldorf | -0.91 |
| 7 | Hannover 96; 1899 Hoffenheim;FC Augsburg | -0.94 |
| **Cluster (home advantage)** | | **Ability** |
| 1 | Hannover 96 | 0.73 |
| 2 | Eintracht Frankfurt; Bayer 04 Leverkusen; 1. FC Nürnberg; Borussia Mönchengladbach; FC Schalke 04; FC Augsburg; 1899 Hoffenheim; 1. FSV Mainz 05; Fortuna Düsseldorf | 0.44 |
| 3 | Sport-Club Freiburg; Hamburger SV; Borussia Dortmund | 0.43 |
| 4 | SV Werder Bremen; FC Bayern München; VfB Stuttgart | 0.15 |
| 5 | VfL Wolfsburg | -0.55 |
| 6 | SpVgg Greuther Fürth | -0.90 |

TABLE 4: *Clusters of teams when distinguishing between abilities when playing at home and playing not at home, and clusters of home advantages.*

and the strength when playing away. A penalty term that clusters the home advantage, $\alpha_r$, the abilities when playing at home, $\gamma_r$, as well as the abilities when playing away, $\gamma_r + \alpha_r$, is

$$J(\boldsymbol{\beta}) = \sum_{r<s} w_{rs}|\gamma_r - \gamma_s| + \sum_{r<s} u_{rs}|\gamma_r - \gamma_s + \alpha_r - \alpha_s| + \sum_{r<s} v_{rs}|\alpha_r - \alpha_s|.$$

with $w_{rs} = |\gamma_r^{(\mathrm{ML})} - \gamma_s^{(\mathrm{ML})}|$, $u_{rs} = |\gamma_r^{(\mathrm{ML})} - \gamma_s^{(\mathrm{ML})} + \alpha_r^{(\mathrm{ML})} - \alpha_s^{(\mathrm{ML})}|$, and $v_{rs} = |\alpha_r^{(\mathrm{ML})} - \alpha_s^{(\mathrm{ML})}|$. It enforces clustering of both abilities and the home advantage. For the selection of the optimal tuning paramater $\lambda$, we again use the BIC criterion

$$BIC(\lambda) = -2 \cdot l(\boldsymbol{\beta}) + df(\lambda) \cdot \log(n),$$

where $n$ is the number of observations. It depends on the degrees of freedom $df(\lambda)$ of the respective model. For penalized models, the degrees of freedom

do not equal the number of parameters in the model because of the effects of shrinkage and variable selection. Therefore, following Buja et al. (1989), the degrees of freedom are calculated by $\text{tr}(2\mathbf{H} - \mathbf{H}^T\mathbf{H})$. Here, $\mathbf{H}$ represents the hat matrix obtained in the last Fisher scoring step in the penalized iteratively re-weighted least squares (PIRLS) algorithm that is used. The algorithm and the corresponding hat matrix are described in more detail by Oelker and Tutz (2013).

Figure 3 shows the coefficient build-ups and Table 4 the corresponding clusters. For the strong teams one obtains very similar classes, but in particular in the middle different clusters are found when playing at home and away. Clustering of the home effect yields essentially 5 classes; Hannover is a class of its own, the big clusters 2 and 3 are hardly different and there are even two clusters with negative home advantage.

# 5 Accounting for Explanatory Variables

Scaling of teams by use of paired comparison models yields estimated abilities but does not explain why some teams are better than others. If one wants to explain the variation in abilities, a natural way is to include covariates in the model. The most interesting variables are variables that characterize the clubs and, therefore, the teams, in contrast to variables that are shared by both teams like day of the week or weather when playing. Explanatory variables of the latter type are more interesting when items are compared and preference is to be modeled as a function of characteristics of the person that chooses. Explanatory variables of this type have been considered, for example, by Dittrich et al. (1998) when modeling the preference for European universities.

## 5.1 A Model with Team-Specific Explanatory Variables

Let the data be given by $(Y_{rs}, r, s \in \{1, \ldots, m\}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ where $Y_{rs} \in \{1, \ldots, m\}$ denotes the ordinal response and $\boldsymbol{x}_r$ is a vector of explanatory variables linked to team $a_r$. Exemplarily, we will consider the budget of a club, which should be influential because the budget determines if a club is able to get the best and most expensive players.

In a general model that accounts for team-specific variables, the strength of the teams, $\gamma_r$, is replaced by $\gamma_r + \boldsymbol{x}_r^T\boldsymbol{\beta}$ yielding the linear predictor

$$\eta_{rst} = \alpha_r + \theta_t + \gamma_r - \gamma_s + (\boldsymbol{x}_r - \boldsymbol{x}_s)^T\boldsymbol{\beta}.$$

In this model, parameters are not identifiable because the parameters $\gamma_r$ can not be distinguished from the parameters $\tilde{\gamma}_r = \gamma_r + \boldsymbol{x}_r^T\boldsymbol{\beta}$. Therefore, additional constraints are needed to obtain unique estimates. A very restrictive model that is identifiable has been proposed by Springall (1973). He obtains identifiability

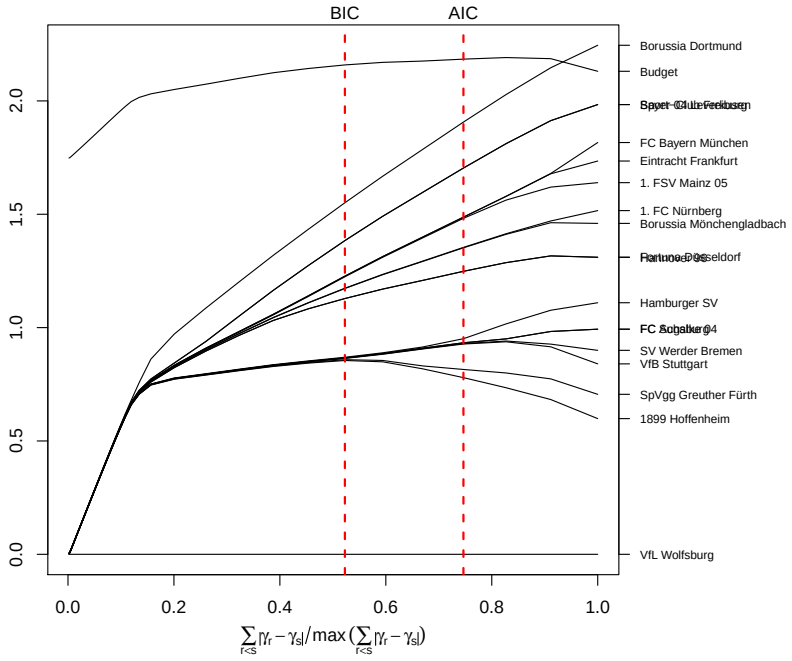


FIGURE 4: *Coefficient paths for ability parameters in the model with a global home advantage and the budget (in* 100 *millions) using an adaptive* $L_1$*-penalty*

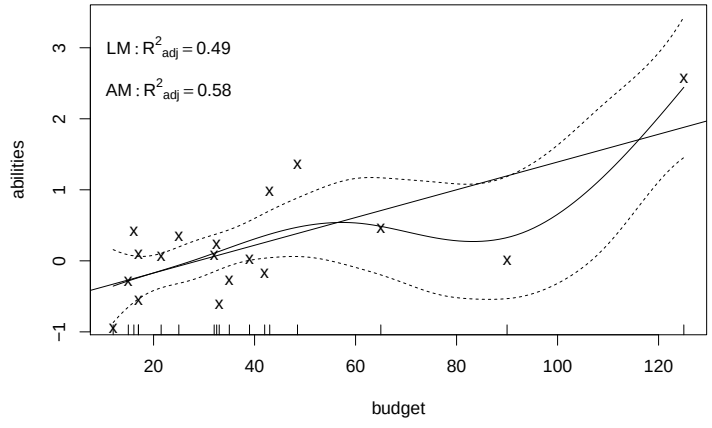


FIGURE 5: *Budgets (in millions) versus estimated abilities for all teams from the Bundesliga season 2012/2013; lines represent linear and additive model fit*

by setting $\gamma_r = 0$, $r = 1, \ldots, m$. The corresponding model assumes that the explanatory variables totally determine the abilities. It is hardly appropriate when a limited number of explanatory variables is available.

An alternative way to constrain estimates is to use a random effects model instead of a fixed effect model. By assuming that the strengths are random effects, for example, by assuming $\gamma_r \sim N(0, \sigma^2)$, parameters can be estimated within a random effects model, see Firth (2005) and Turner and Firth (2012) who used random effects models to account for correlations between responses. A disadvantage of random effects models is that they assume that random effects and covariates are uncorrelated, certainly not realistic in football if the covariates contain the budget of teams, because it might be the main source of the strength of a team. The assumption that random effects and covariates have to be uncorrelated in random effects models has been widely discussed. An early reference is **?**, more recently the topic was discussed, for example, by **?**, **?** and **?**.

An alternative approach that is advocated here is to use penalized estimation procedures. Assuming that teams are clustered one can use the penalty (2). It penalizes the abilities that are not explained by covariates, $\gamma_r$, $r = 1, \ldots, m$, but not the parameter $\boldsymbol{\beta}$. If the tuning parameter gets large, $\lambda \to \infty$, all strength parameters $\gamma_r$ are estimated as identical and the total strength is determined solely by $\boldsymbol{x}_r^T \boldsymbol{\beta}$ as in the model proposed by Springall (1973). By using a regularization term with positive tuning parameter the parameters are defined and estimable, compare also Friedman et al. (2010), where this procedure has been used in overparameterized multinomial regression models.

In Section 5.2, we will show that the procedure works. But first we show how the performance of football teams in the German Bundesliga can be explained by the budget. We use budgets as published by the German sports magazine Kicker (Kicker, August 20, 2012) given in millions. Figure 4 shows the coefficient paths for the coefficients plotted against varying strength of the constraints. Here, we use budget in 100 millions for better visibility of the coefficient path. It is seen that the effect of budget is very stable across constraints. As expected, when including the budget different clusters are found because now the $\gamma$-parameters represent the abilities that are not explained by the budget. For example, now Borussia Dortmund forms a cluster of its own, whereas Bayern München is in a cluster together with Eintracht Frankfurt and Mainz.

The estimated parameter $\hat{\beta} = 2.16$, obtained for $\lambda$ chosen by BIC, implies strong dependence on the budget. In order to get an impression on the reliability of the parameter estimate of the budget at the BIC-optimal $\lambda$, we conducted a parametric bootstrap analysis. The corresponding bootstrap confidence interval for $\hat{\beta}$ is $[1.55; 2.77]$; it supports that budget does have an influence on the team abilities that is not to be neglected.

The effect of the budget can also be tackled in a different way. In Figure 5 the estimated abilities are plotted against the budget. In addition, it shows the fit of a linear regression model and a smoothed version. The smooth model was fitted

by use of penalized B-splines (also called P-splines), see Eilers and Marx (1996), with the smoothing parameter chosen by the generalized cross-validation (GCV) criterion. Up to about 70, the linear model fits well, beyond 70 the fit of the non-linear model is determined by just two observations, Wolfsburg and Bayern München. The adjusted R-squared of the linear model is 0.49, that means almost 50% of the variation in abilities is explained by the budget. For the non-linear model the value increases to 0.58. When accepting the linear model as a simple model that shows almost the same explanatory strength as the non-linear model, one can infer that Wolfsburg (with a budget of 90) is an underachiever. Given the high budget, which is partly due to the fact that the city of Wolfsburg is the home of Volkswagen, the ability is rather low. This holds even in the non-linear model. Bayern München, the club with the highest budget, still shows a positive deviation from the fitted expectation, which is strong for the linear and weak for the the non-linear model. A distinct overachiever is Dortmund (budget of 48.5), which shows one of the strongest deviations from both models. Beyond the identification of over- and underachievers, it is seen that budget is a strong explanatory variable for the ability of a team. Thus, a strong part of the success of Bayern München seems to be related to the high budget of the club.

## 5.2 Evaluation of Penalized Estimation

In this section, we investigate in a small simulation study how well the penalized estimation procedure works for the model with explanatory variables. As true coefficients, we chose values derived from the coefficient estimates of the model fit for the real data from the Bundesliga. We used the thresholds $\theta_1 = -1.66$, $\theta_2 = -0.65$, $\alpha = 0.29$ and the budget parameter $\beta = 2.13$. The team abilities were divided into 5 groups with the coefficients $\gamma_1 = \gamma_2 = \gamma_3 = 2.07$, $\gamma_4 = \gamma_5 = \gamma_6 = 1.73$, $\gamma_7 = \gamma_8 = \gamma_9 = \gamma_{10} = 1.40$, $\gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} = \gamma_{15} = \gamma_{16} = \gamma_{17} = 0.88$, $\gamma_{18} = 0$.

Figure 6 shows the box plots for 100 simulations. Stars denote the true parameter values. In particular, the threshold parameters and the home advantage parameter are estimated with high accuracy. As expected, the variation of estimates is stronger for the abilities. But, and most important, the parameter of the explanatory variable is estimated rather well.

# 6 Concluding Remarks

All calculations in this paper have been conducted by using the statistical software `R` (R Core Team, 2013). Most of the available add-on packages for paired comparison models in `R` are restricted to the case of binary response and cannot deal with ordered response. The most popular packages are `prefmod` (Hatzinger and Dittrich, 2012) and `BradleyTerry2` (Turner and Firth, 2012). The former

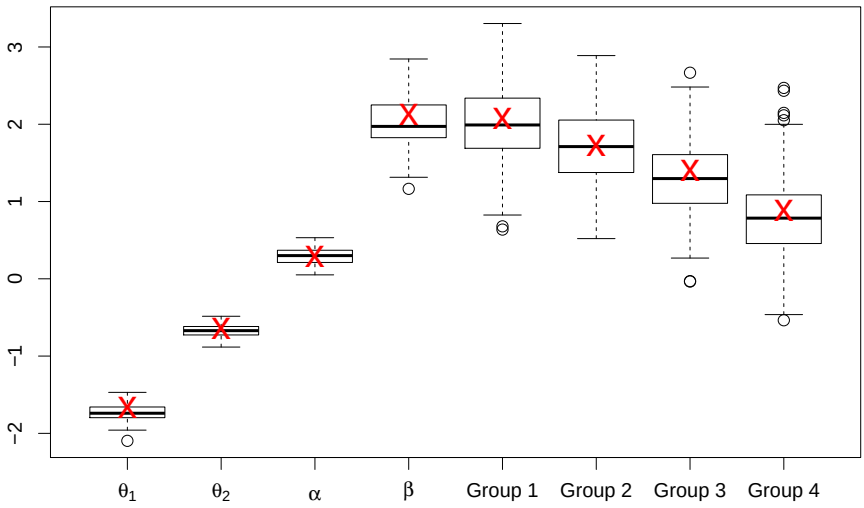


FIGURE 6: *Box plots of coefficient estimates for 100 simulation iterations; estimates for teams with equal abilities are collected in one box; stars denote true values*

uses the log linear representation of BT-models and can handle draws in the response variable. The latter can also handle covariates by assuming random effects for the ability parameters but only in the case of binary responses.

Here we favor a direct approach to the fitting of ordinal paired comparison models (without regularization) that is based on the embedding into the framework of generalized linear models. By including the restrictions on the thresholds and the construction of specific design matrices that include the effect of home advantages BT models for ordered response can be fitted by using the add-on package `VGAM` (Yee, 2010). It also allows to use alternative link functions. The procedure, but without team-specific covariates and regularization, has been implemented in the package `ordBTL` (Casalicchio, 2013).

In our extended framework we have to also include penalty terms. A very general approach that allows to combine a variety of different penalties in univariate GLMs has been proposed by Oelker and Tutz (2013), and is available in the package `gvcm.cat` (Oelker, 2013). With the help of Margret Oelker it has been adapted such that also cumulative logit models can be fitted. It is available from the authors.

## Acknowledgement

We thank Margret Oelker for her help when adapting her package `gvcm.cat` to include cumulative model fits.

## References


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

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Böckenholt, U. and W. R. Dillon (1997a). Modeling within-subject dependencies in ordinal paired comparison data. *Psychometrika 62*(3), 411–434.

Böckenholt, U. and W. R. Dillon (1997b). Some new methods for an old problem: Modeling preference changes and competitive market structures in pretest market data. *Journal of Marketing Research*, 130–142.

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

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

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Annals of Statistics 17*, 453–510.

Casalicchio, G. (2013). *ordBTL: Modelling comparison data with ordinal response.* R package version 0.7.

Cattelan, M., C. Varin, and D. Firth (2013). Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(1), 135–150.

Davidson, R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association 65*, 317–328.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*(4), 511–525.

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.

Dittrich, R., W. Katzenbeisser, and H. Reisinger (2000). The analysis of rank ordered preference data based on bradley-terry type models. *OR-Spektrum 22*(1), 117–134.

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

Fahrmeir, L. and G. Tutz (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association 89*, 1438–1449.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Firth, D. (2005). Bradley-terry models in r. *Journal of Statistical Software 12*(1), 1–12.

Firth, D. and R. De Menezes (2004). Quasi-variances. *Biometrika 91*, 65.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

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

Glickman, M. E. and H. S. Stern (1998). A state-space model for national football league scores. *Journal of the American Statistical Association 93*(441), 25–35.

Hatzinger, R. and R. Dittrich (2012). prefmod: An r package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software 48*(10), 1–31.

Knorr-Held, L. (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(2), 261–276.

Kuk, A. Y. C. (1995). Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *Journal of the Royal Statistical Society. Series D (The Statistician) 44*(4), pp. 523–528.

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

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*, 109–127.

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

Oelker, M.-R. (2013). *gvcm.cat: Regularized Categorial Effects/Categorial Effect Modifiers in GLMs*. R package version 1.6.

Oelker, M.-R. and G. Tutz (2013). A general family of penalties for combining differing types of penalties in generalized structured models. Technical Report 139, LMU, Department of Statistics.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, P. and L. Kupper (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association 62*, 194–204.

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

Springall, A. (1973). Response surface fitting using a generalization of the bradley-terry paired comparison model. *Applied Statistics*, 59–68.

Turner, H. and D. Firth (2012, 5). 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. (2012). *Regression for Categorical Data*. Cambridge University Press.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software 32*(10), 1–34.