A paired comparison model for the analysis of on-field variables in football matches

Gunther Schauberger and Andreas Groll

Abstract We use on-field variables from football matches in the German Bundesliga and connect them to the sportive success or failure of the single teams in a paired comparison model where each match in a Bundesliga season is treated as a paired comparison between the two competing teams. We propose an extended paired comparison model that extends the classical Bradley-Terry model to ordinal response variables and includes different types of covariates. We apply penalized likelihood estimation and use specific L_1 penalty terms for fusion and selection in order to reduce the complexity of the model and to find clusters of teams with equal covariate effects. The proposed model is a very general one and can easily be applied to other sports data or to data from different research fields. We apply the model to data from the latest season of the German Bundesliga.

Key words: Bundesliga, Paired Comparison, BTLLasso, Penalization

1 Introduction

In modern football, various variables as, for example, the distance a team runs or its percentage of ball possession, are collected throughout a match. However, there is a lack of methods to make use of these on-field variables simultaneously and to connect them with the final result of the match. We propose to treat each football match as a paired comparison between the two competing teams and to analyse the results of football data by an extended paired comparison model.

Gunther Schauberger
Chair of Epidemiology, Technical University Munich, e-mail: gunther.schauberger@tum.de,
Department of Statistics, LMU Munich

Andreas Groll
Faculty of Statistics, Technische Universität Dortmund e-mail: groll@statistik.tu-dortmund.de

Paired comparisons occur if two objects out of a set of objects are compared with respect to an underlying latent trait. In the case of football matches in national leagues all teams from the respective league are considered to be these objects. Football matches can be treated as paired comparisons between two teams where the playing abilities of the teams represent the underlying latent traits that are compared.

Our main goal is to set up a paired comparison model that is able to incorporate so-called on-field variables as covariates. In general, if covariates are to be considered in paired comparison, one has to distinguish between subjects and objects of the paired comparisons. A covariate can either vary across the subjects or the objects of a paired comparison, or, as in our case, both over subjects and objects. In football matches, the teams are the objects while a single match can be considered to be the subject that conducts the comparison between the two objects/teams. If one considers a variable like the percentage of ball possession a team has in a specific match, this variable varies both from team to team and from match to match. Therefore, in our application subject-object-specific covariates are considered. After all, the proposed model could in principle consider all three types of variables simultaneously.

The Bradley-Terry model (Bradley and Terry, 1952) is the standard model for paired comparison data. Assuming a set of objects $\{a_1,\ldots,a_m\}$, in its most simple form the Bradley-Terry model is given by

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

One models the probability that a certain object a_r dominates or is preferred over another object a_s, $a_r \succ a_s$. The random variable $Y_{(r,s)}$ is defined to be $Y_{(r,s)} = 1$ if a_r dominates a_s and $Y_{(r,s)} = 0$ otherwise. The parameters γ_r represent the attractiveness or strength of the respective objects. In football matches, the random variable $Y_{(r,s)}$ which represents the paired comparison between a_r and a_s needs to have at least $K = 3$ possible categories instead of two because in football one needs to account for the possibility of draws. However, if one distinguishes, for example, clear wins and losses from wins and losses with only one goal difference one could also use $K = 5$ categories. In general, for the case of ordered responses $Y_{(r,s)} \in \{1,\ldots,K\}$ the model is extended accordingly to

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1+\exp(\theta_k + \gamma_r - \gamma_s)}, \quad k = 1,\ldots,K, \tag{2}$$

which essentially corresponds to the generalization from a binary logistic regression model to a cumulative logistic regression model. In our application, the strength parameters γ_r represent the playing abilities of the teams.

In general, for the ordinal paired comparison model (2) it is assumed that the response categories have a symmetric interpretation so that $P(Y_{(r,s)} = k) = P(Y_{(s,r)} = K-k+1)$ holds. Therefore, the threshold parameters should be restricted by $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$ to guarantee for symmetric probabilities. The threshold for the last category is fixed to $\theta_K = \infty$ so that $P(Y_{(r,s)} \leq K) = 1$ will hold.

The probability for a single response category can be derived from the difference between two adjacent categories, $P(Y_{(r,s)} = k) = P(Y_{(r,s)} \leq k) - P(Y_{(r,s)} \leq k-1)$. To guarantee for non-negative probabilities of the single response categories one restricts $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$.

When football matches are considered as paired comparisons one has to consider that so-called order effects are possible. To some extent, order effects contradict the assumption of symmetric response categories which was described above. When order effects are present, the order of the two objects (i.e. which object is the first-named object and which is the second-named object) is not random and possibly has an influence on the outcome. In football (especially in national leagues) this is obviously the case as the first-named object is the home team and, hence, usually has a home advantage over the (second-named) away team. To include such an order effect in model (2) we extend the model to

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\delta + \theta_k + \gamma_r - \gamma_s)}{1 + \exp(\delta + \theta_k + \gamma_r - \gamma_s)}\,, \quad k = 1, \ldots, 5\,, \tag{3}$$

where δ represents an order effect. In football matches, this parameter represents the home effect (or home advantage if positive). It is possible to assume a global home effect δ which is equal for all teams or team-specific home effects δ_r.

2 Bundesliga Data

The main goal of this work is to analyze if (and which) on-field variables that are collected throughout a match are associated to the final result of football matches. In total, our data set contains all the following variables separately for each team and each match:

Distance	Total amount of km run
BallPossession	Percentage of ball possession
TacklingRate	Rate of tacklings won
ShotsonGoal	Total number of shots on goal
Passes	Total number of passes
CompletionRate	Percentage of passes reaching teammates
FoulsSuffered	Number of fouls suffered
Offside	Number of offsides (in attack)

The data were collected from the website of the German football magazin kicker (http://www.kicker.de/). Exemplarily, Table 1 shows the collected data for the opening match of the season 2016/17 between Bayern München and Hamburger SV.

Bayern München		Hamburger SV
Goals	5 : 0	Goals
Shots on goal	23 : 5	Shots on goal
Distance	108.54 : 111.28	Distance
Completion rate	90 : 64	Completion rate
Ball possession	77 : 23	Ball possession
Tackling rate	52 : 48	Tackling rate
Fouls	10 : 12	Fouls
Offside	3 : 0	Offside

Table 1 Illustrating table for original data situation showing data for the opening match in season 2016/17 between Bayern München and Hamburger SV. Source: http://www.kicker.de/

3 A Paired Comparison Model for Football Matches Including On-field Variables

When using a paired comparison model for football matches the standard Bradley-Terry model needs to be extended in several ways. In model (3) we already extended the Bradley-Terry model to handle both an ordinal response (in particular draws) and home effects. Now the model is further extended to incorporate on-field variables, which in the context of paired comparisons are considered as subject-object-specific variables. We propose to use the general model for ordinal response data $Y_{i(r,s)} \in \{1,\dots,K\}$ denoted by

$$\begin{aligned} P(Y_{i(r,s)} \leq k) &= \frac{\exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + z_{ir}^T \alpha_r - z_{is}^T \alpha_s)}{1 + \exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + z_{ir}^T \alpha_r - z_{is}^T \alpha_s)} \,. \end{aligned} \tag{4}$$

The model allows for the inclusion of so-called subject-object-specific covariates z_{ir}. It belongs to the general model family proposed by Schauberger and Tutz (2017a) for the inclusion of different types of covariates in paired comparison models. Within this framework, Tutz and Schauberger (2015) present a model including object-specific covariates z_r and Schauberger and Tutz (2017b) present a model including subject-specific covariates z_i. In Schauberger et al. (2017), the presented model is applied to data from the Bundesliga season 2015/16.

The response $Y_{i(r,s)}$ encodes an ordered response with K categories (including a category for draws) for a match between team a_r and team a_s on matchday i where a_r plays at its home ground. The linear predictor of the model contains the following terms:

δ_r team-specific home effects of team a_r

θ_k category-specific threshold parameters
β_{r0} team-specific intercepts
z_{ir} p-dimensional covariate vector that varies over teams and matches
α_r p-dimensional parameter vector that varies over teams.

Instead of fixed abilities γ_r, the teams have abilities $\gamma_{ir} = \beta_{r0} + z_{ir}^T \alpha_r$ which differ for each matchday depending on the covariates of team a_r on matchday i. In its general form, the model has a lot of parameters that need to be estimated. It could, for example, be simplified if both the home effect and the covariate effects were included with global instead of team-specific parameters. For this purpose, we use penalty terms to decide whether the home effect or single covariate effects should be considered with team-specific or global parameters. In particular, the absolute values of all pairwise differences between the team-specific home advantages are penalized using the L_1 penalty term

$$P(\delta_1, \ldots, \delta_m) = \sum_{r<s} |\delta_r - \delta_s| \,. \tag{5}$$

The penalty term enforces the clustering of teams with equal home effects as it is able to set differences between parameters to exactly zero. Therefore, the penalty could for example produce three clusters of teams where each of the clusters has a different home effect. As an extreme case, the penalty leads to one global home effect if all differences are set zero.

Also the team-specific covariate effects are penalized. The respective penalty term penalizes the absolute values of all pairwise differences of the covariate parameters and of the parameters themselves, i.e.

$$J(\alpha_1, \ldots, \alpha_m) = \sum_{j=1}^{p} \sum_{r<s} |\alpha_{rj} - \alpha_{sj}| + \sum_{j=1}^{p} \sum_{r=1}^{m} |\alpha_{rj}| \,. \tag{6}$$

The penalty enforces clustering of teams with respect to certain on-field variables, possibly leading to global effects instead of team-specific effects. Moreover, due to the penalization of the absolute values, covariates can be eliminated completely from the model. For comparability of the penalties and the resulting effects, all covariates have to be transformed to a joint scale.

Finally, both penalty terms are combined and the respective penalized likelihood

$$l_p(\cdot) = l(\cdot) - \lambda \left(P(\delta_1, \ldots, \delta_m) + J(\alpha_1, \ldots, \alpha_m) \right)$$

is maximized, $l(\cdot)$ denoting the (unpenalized) likelihood. The tuning parameter λ is chosen by 10-fold cross-validation with respect to the so-called ranked probability score (RPS) proposed by Gneiting and Raftery (2007). The RPS for ordinal response $y \in \{1, \ldots, K\}$ can be denoted by

$$RPS(y, \hat{\pi}(k)) = \sum_{k=1}^{K} (\hat{\pi}(k) - \mathbb{1}(y \leq k))^2,$$

where $\pi(k)$ represents the cumulative probability $\pi(k) = P(y \leq k)$. In contrast to other possible error measures (e.g. the deviance or the Brier score), it takes the ordinal structure of the response into account.

4 Application to Bundesliga Season 2016/17

We now apply the model to data from the Bundesliga season 2016/17. The data contain each of the 306 macthes of this season on the 34 matchdays. For easier interpretation of the intercepts, the covariates were centered (per team around the team-specific means). Centering of covariates only changes the paths (and interpretation) of the team-specific intercepts. Now, the intercepts represent the ability of a team when every covariate is set to the team-specific mean. Beside that, the paths and the interpretation of the covariate effects are not affected by the centering of the covariates. They represent the effect of a covariate on the ability of a team when the respective covariate deviates from the team-specific mean.

Figure 1 illustrates the parameters' paths for the proposed model, separately for each covariate along the tuning parameter λ. The dashed vertical line indicates the model that was selected by 10-fold cross-validation. In contrast to the home effects and all covariate effects, the team-specific intercepts are not penalized and, consequently, do not show any particular clusters of teams. Bayern München clearly dominated the league in this season which is also represented by a very high team-specific intercept.

The paths of the home and the covariate effects clearly illustrate the clustering effect of the penalty terms. It can be seen that the home effect seems to be equal for all teams. The home effect is positive and, therefore, represents an actual home advantage for all teams as it was expected. The greatest effect of all covariates can be seen for *Distance*. It has a strong positive effect for all teams. The teams gain better results in matches where they had a good running performance. Interestingly, the covariate *BallPossession* has negative effects for all teams. Here, only Darmstadt 98 is separate from the other teams with an even more negative effect while all other teams form a big cluster for this variable. None of the variables is eliminated completely from the model, each variable has effects for at least two of the teams. *TacklingRate* and *ShotsonGoal* have (small) positive effects for all teams. Figure 2 shows the RPS of the cross-validation along the tuning parameter λ.

5 Concluding remarks

This work deals with data from the German Bundesliga from the season 2016/17 and considers several on-field variables in a paired comparison model. We propose a model that is able to make use of the big amount of data that is collected in modern football and to simultaneously connect the corresponding variables to the outcome

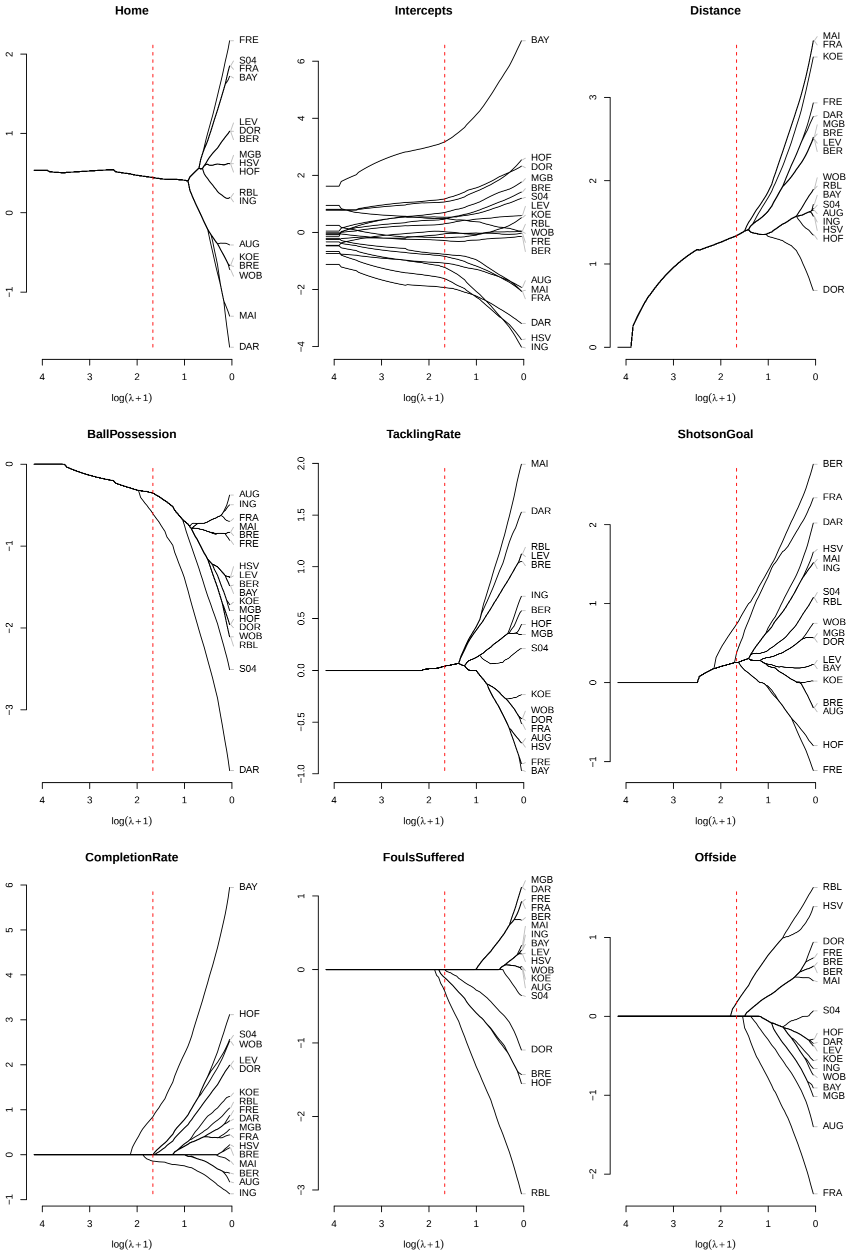

Fig. 1 Parameter paths, separately for home effect, intercepts and all (centered) covariates. Dashed vertical line represents the optimal model according to 10-fold cross-validation.

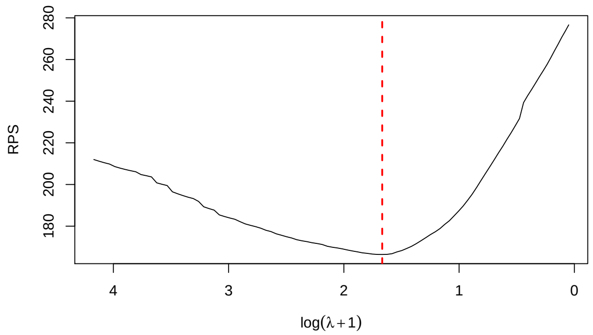

Fig. 2 Ranked probability score (RPS) for cross-validation along tuning parameter λ for model (4). Dashed vertical line represents optimal model according to 10-fold cross-validation.

of the matches. Complex modeling approaches are rather scarce in this area. The model incorporates football matches into the framework of paired comparisons and uses the general model proposed by Schauberger and Tutz (2017a) for the incorporation of different types of variables into paired comparison models.

In contrast to standard paired comparison models, the model offers a much more flexible and less restricted approach. Each team is assigned with individual strengths per matchday, depending on the on-field covariates of the team. This extension of the simple Bradley-Terry model allows for a much better discrimination between the different match outcomes and, therefore, for a better predictive performance.

References

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

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–376.

Schauberger, G., A. Groll, and G. Tutz (2017). Analysis of the importance of on-field covariates in the German Bundesliga. *Journal of Applied Statistics published online*, 1–18.

Schauberger, G. and G. Tutz (2017a). BTLLasso - A common framework and software package for the inclusion and selection of covariates in Bradley-Terry models. Technical Report 202, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

Schauberger, G. and G. Tutz (2017b). Subject-specific modelling of paired comparison data - a lasso-type penalty approach. *Statistical Modelling 17*(3), 223–243.

Tutz, G. and G. Schauberger (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis 99*(2), 209–227.