



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Gunther Schauberger & Andreas Groll & Gerhard Tutz

# Modeling Football Results in the German Bundesliga Using Match-specific Covariates

Technical Report Number 197, 2016  
Department of Statistics  
University of Munich

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



# Modeling Football Results in the German Bundesliga Using Match-specific Covariates

Gunther Schaubberger, Andreas Groll & Gerhard Tutz  
*Department of Statistics, LMU Munich*

August 3, 2016

## Abstract

In modern football, various variables as, for example, the distances the teams run or the percentages of ball possession, are collected throughout a match. However, there is a lack of methods to make use of these variables simultaneously and to connect them with the final result of the match. This paper considers data from the German Bundesliga season 2015/16. The objective is to identify the variables that are connected to the sportive success or failure of the single teams. A paired comparison model for football matches is proposed that is able to take into account match-specific covariates. The model extends the Bradley-Terry model in many different ways. In addition to the inclusion of covariates, it uses ordered response values and includes (possibly team-specific) home effects. Penalty terms are used to reduce the complexity of the model and to find clusters of teams with equal covariate effects.

**Keywords:** Paired Comparison, Bradley-Terry, penalization, BTLLasso.

## 1 Introduction

Traditionally, discussions about football (and football tactics in particular) are very controversial, both amongst professional and (us) non-professional football experts. After all, most football enthusiasts generally agree on platitudes like the importance of winning tackles or running more than the opponent. This work aims at contributing to these discussions from a scientific point of view and to the examination of the validity of football platitudes. In modern football, several match-specific variables as, for example, the running performance of teams or the tackling rate are measured and are publicly available from several online media. We will consider a specific regression model incorporating a set of covariates of that kind.

From a statistical point of view, a football match between two competing teams can be seen as a paired comparison. In paired comparisons, two objects are compared and it is observed, which of the objects dominates. It is assumed, that the dominance is generated by an unobserved latent trait. In football matches, the latent traits are the playing abilities of both teams. The standard model for paired comparisons is the Bradley-Terry model (Bradley and Terry, 1952), which has been extended in several ways. An extensive overview on different paired comparison models is found in Cattelan (2012). Only few publications address the issue of including covariates in paired comparison models. Francis et al. (2010) and Turner and Firth (2012) use (subject-specific) covariates characterizing the persons that perform the respective comparison. The incorporation of covariates into paired comparison models leads to more complex models. Therefore, regularization methods can be applied to reduce the complexity of the final models. Casalicchio et al. (2015) presented a boosting approach while Tutz and Schaubberger (2015) and Schaubberger and Tutz (2015) use  $L_1$ -type penalties.

There is a wide range of literature on modeling football match outcomes considering football matches in international tournaments or national football leagues. Part of the literature concentrates on models to predict the match outcomes. Therefore, these approaches can only use covariates which are known before a match takes place. Examples of predictive approaches focusing on the prediction of the exact scores of a match can be found in Dixon and Coles (1997); Karlis and Ntzoufras (2003); Dyte and Clarke (2000); Groll et al. (2015). Another (but smaller) part of the literature focuses on the post hoc analysis of football matches. Here, the goal is to detect which variables influence the observed outcomes. A popular field of interest is the influence of the ball possession on the success of teams. There is an ongoing debate of whether direct play or possession play is preferable (Collet, 2013; Hughes and Franks, 2005; Vogelbein et al., 2014). Simultaneous analyses of several match-specific covariates are rather rare. Castellano et al. (2012) perform a multivariate discriminant analysis to discriminate between winning, drawing and losing teams in FIFA World Cups. A model-based approach can be found in Carmichael et al. (2000). There, a linear model for the difference of goals is proposed considering a set of match-specific covariates, such as the number of shots, the percentage of successful passes or the number of tackles, and the model is applied on data from the English Premier League.

The goal of this work is to determine, which match-specific variables are related to the success or failure of teams in the German Bundesliga. Furthermore, we are interested in possible differences between the teams or if there are clusters of teams with similar effects of variables, as for example the percentage of ball possession. For that purpose, we include such match-specific covariates into a paired comparison model. The effects of the covariates can be parametrized in the form of global or team-specific effects. The ordinary Bradley-Terry model is extended in various ways. For the estimation, a penalty term is proposed

that is able to detect clusters of teams with respect to certain covariate effects and reduces the complexity of the final model. The model is estimated using R-Code extending the package `BTLasso` (Schauberger, 2015) from the statistical environment R (R Core Team, 2015) which is available from the authors.

The paper is structured as follows. In Section 2 some basic models for paired comparisons, especially for the case of football data, are introduced. Section 3 gives an introduction into the data with a special focus on the variables we are interested in. A paired comparison model including the variables of interest in our data set is introduced in Section 4. A penalized estimation approach is proposed and the results are presented. In Section 5 the results of an alternative modeling approach are shown. The predictive performance of all proposed models is assessed in Section 6.

## 2 Modeling Paired Comparisons

In the following, different models for paired comparisons are introduced, beginning with the Bradley-Terry model. The Bradley-Terry model (Bradley and Terry, 1952) is the standard model for paired comparisons. It does not consider covariates and, in general, does not pay any attention to heterogeneity caused by the subjects of paired comparisons.

### 2.1 The Bradley-Terry Model

Assuming a set of objects  $\{a_1, \dots, a_m\}$ , in its most simple form the (binary) Bradley-Terry model is given by

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

The response of the model represents the probability that a certain object  $a_r$  is preferred over another object  $a_s$ , denoted by  $a_r > a_s$ . This response can be formalized in the dichotomous random variable  $Y_{(r,s)}$  which is defined to be  $Y_{(r,s)} = 1$  if  $a_r$  is preferred over  $a_s$  and  $Y_{(r,s)} = 0$  otherwise. The parameters  $\gamma_r$ ,  $r = 1, \dots, m$ , represent the attractiveness or strength of the respective objects. For identifiability, a restriction on the parameters is needed, for example  $\sum_{r=1}^m \gamma_r = 0$  or  $\gamma_m = 0$ . In the following, we will use the symmetric side constraint  $\sum_{r=1}^m \gamma_r = 0$ .

### 2.2 The Bradley–Terry Model with Ordered Response Categories

In many applications the dominance of one of the objects is quite naturally observed on an ordered scale. In our current application of paired comparisons to football matches, it is mandatory to account for draws. Therefore, the model has

to account for at least three ordered response categories. Early extensions of the BTL-model include at least the possibility of ties, see Rao and Kupper (1967), Glenn and David (1960) and Davidson (1970). General models for ordered responses, for example to allow for a general number of  $K$  categories were proposed by Tutz (1986) and Agresti (1992). In a natural extension of the binary Bradley-Terry model to  $K$  response categories, the model parametrizes the cumulative probabilities in the form

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

with  $k = 1, \dots, K$  denoting the possible response categories. The parameters  $\theta_k$  represent the so-called threshold parameters for the single response categories, they determine the preference for specific categories. In particular,  $Y_{(r,s)} = 1$  represents the maximal preference for object  $a_r$  over  $a_s$  and  $Y_{(r,s)} = K$  represents the maximal preference for object  $a_s$  over  $a_r$ . To be able to efficiently use the information contained in the result of a football match, we will consider a response variable on a 5-point scale.

In general, for ordinal paired comparisons it can be 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 for the single response categories one restricts  $\theta_1 \leq \theta_2 \leq \dots \leq \theta_K$ . The ordinal Bradley-Terry model corresponds to a cumulative logit model and can be estimated using methods from this general framework (Agresti, 2002).

### 2.3 The Bradley–Terry Models Including Order Effects

After all, the symmetry of the response categories, guaranteed by the restrictions on the threshold parameters  $\theta_k$ , is not appropriate for all data situations. Sometimes, the order of the objects can be decisive. In particular, in sport competitions as in our application to football matches the order matters. In our data structure, the first team represents the team playing at its home ground where it might have a (home) advantage over its opponent. Therefore, the assumption that the response categories are symmetric does not hold anymore and the model needs to be adapted accordingly. Extending the basic models by an additional parameter  $\delta$  yields the binary Bradley-Terry model

$$P(Y_{(r,s)} = 1) = \frac{\exp(\delta + \gamma_r - \gamma_s)}{1 + \exp(\delta + \gamma_r - \gamma_s)}$$

and the ordinal model

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

Here,  $\delta$  denotes the order effect which is simply incorporated into the design matrix by an additional intercept column. If  $\delta > 0$ , it increases the probability of the first-named object  $a_r$  to win the comparison or, in the case of an ordinal response, to achieve a superior result. Given the order effect, the symmetry assumption for the response categories still holds. When applied to football matches,  $\delta$  represents a home effect which, as  $\delta$  does not depend on team  $a_r$ , is assumed to be equal for all teams.

### 3 Bundesliga Data 2015/2016

The data we consider are data from the season 2015/2016 of the German Bundesliga. The German Bundesliga is played as a double round robin between 18 teams. Table 1 shows the final table of the season 2015/16. As in all three previous seasons, Bayern München won the championship. VfB Stuttgart and Hannover 96 were relegated to the second division.

Position	Team	Goals For	Goals Against	Points
1	 Bayern München	80	17	88
2	 Borussia Dortmund	82	34	78
3	 Bayer 04 Leverkusen	56	40	60
4	 Bor. Mönchengladbach	67	50	55
5	 FC Schalke 04	51	49	52
6	 1. FSV Mainz 05	46	42	50
7	 Hertha BSC	42	42	50
8	 VfL Wolfsburg	47	49	45
9	 1. FC Köln	38	42	43
10	 Hamburger SV	40	46	41
11	 FC Ingolstadt 04	33	42	40
12	 FC Augsburg	42	52	38
13	 Werder Bremen	50	65	38
14	 SV Darmstadt 98	38	53	38
15	 TSG Hoffenheim	39	54	37
16	 Eintracht Frankfurt	34	52	36
17	 VfB Stuttgart	50	75	33
18	 Hannover 96	31	62	25

Table 1: Final table of the German Bundesliga in the season 2015/2016.

All 306 matches played on the 34 match-days of this season will be considered as the observations in the data set. We will treat a match as a paired comparison of both teams with respect to their playing abilities. The response variables  $Y_{i(r,s)}$  represent the outcome of a match between team  $a_r$  (as the home team) and team  $a_s$  on matchday  $i$ . We use a 5-point scale defined by

$$Y_{i(r,s)} = \begin{cases} 1 & \text{if team } a_r \text{ wins by at least 2 goals difference,} \\ 2 & \text{if team } a_r \text{ wins by 1 goal difference,} \\ 3 & \text{if the match ends with a draw,} \\ 4 & \text{if team } a_s \text{ wins by 1 goal difference,} \\ 5 & \text{if team } a_s \text{ wins by at least 2 goals difference.} \end{cases}$$

Model (1) is able to handle such ordinal response categories. Fitted to the Bundesliga data of the season 2015/16, the model yields ability estimates as presented in Table 2. Additionally, estimates of the threshold parameters  $\hat{\theta}_1 = -\hat{\theta}_4 = -1.591$  and  $\hat{\theta}_2 = -\hat{\theta}_3 = -0.576$  and the home effect  $\hat{\delta} = 0.265$  were obtained.

Position		Team	$\gamma_r$	Rank
1		BAY Bayern München	1.899	1
2		DOR Borussia Dortmund	1.598	2
3		LEV Bayer 04 Leverkusen	0.433	4
4		MGB Bor. Mönchengladbach	0.475	3
5		S04 FC Schalke 04	0.133	5
6		MAI 1. FSV Mainz 05	0.088	6
7		BER Hertha BSC	-0.001	7
8		WOB VfL Wolfsburg	-0.142	9
9		KOE 1. FC Köln	-0.045	8
10		HSV Hamburger SV	-0.183	10
11		ING FC Ingolstadt 04	-0.228	11
12		AUG FC Augsburg	-0.363	13
13		BRE Werder Bremen	-0.361	12
14		DAR SV Darmstadt 98	-0.467	15
15		HOF TSG Hoffenheim	-0.448	14
16		FRA Eintracht Frankfurt	-0.623	16
17		STU VfB Stuttgart	-0.699	17
18		HAN Hannover 96	-1.068	18

Table 2: Ability estimates for single teams considering model (1)

The ranking of the estimated abilities more or less coincides with the rankings of the final table. After all, there are a few interesting differences. For example, Borussia Mönchengladbach is assessed to be on rank 3 according to the estimated

ability 0.475. However, the team finished the season on position 4 and, in contrast to the third-placed Bayer 04 Leverkusen, has to play qualification matches for the participation in the UEFA Champions League.

Nowadays, in professional football matches a huge amount of variables is collected. For example, for every team it is known what distance the team ran in a certain match or its number of shots on goal. The main goal of this work is to determine the influence of these match-specific variables. In the German Bundesliga, the data supplier opta (<http://www.optasports.com/>) provides interesting data collections. The data we use are freely available from the website of the German football magazine kicker (<http://www.kicker.de/>), Table 3 shows a short excerpt of the data including the first three matches of the season.


Match	Goals	Home	Team	Distance	Shots on Goal	...
1	5	yes	 Bayern München	109	23	...
1	0	no	 Hamburger SV	111	5	...
2	2	yes	 Bayer 04 Leverkusen	116	25	...
2	1	no	 TSG Hoffenheim	116	6	...
3	0	yes	 FC Augsburg	106	20	...
3	1	no	 Hertha BSC 04	111	11	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3: Exemplary extract of the Bundesliga data basis from <http://www.kicker.de/>

From these (original) data, the ordinal responses for the paired comparisons (as described above) were derived. In detail, the following variables are available (per team and per match):

*Home* Dummy variable for home team

*Distance* Total amount of km run

*BallPossession* Percentage of ball possession

*TacklingRate* Rate of won tacklings

*ShotsonGoal* Total number of shots on goal

*Passes* Total number of passes

*Misplaced* Total number of misplaced passes (not reaching teammates)

*CompletionRate* Percentage of passes reaching teammates



	Home	Distance	BallPossession	TacklingRate	ShotsonGoal	CompletionRate	FoulsSuffered	Offside
Home	1.000							
Distance	0.035	1.000						
BallPossession	0.102	-0.113	1.000					
TacklingRate	0.102	-0.082	0.186	1.000				
ShotsonGoal	0.230	0.042	0.519	0.261	1.000			
CompletionRate	0.068	0.103	0.717	0.118	0.422	1.000		
FoulsSuffered	0.067	-0.200	0.089	0.236	0.035	-0.160	1.000	
Offside	0.038	-0.037	0.091	0.088	0.055	0.042	-0.011	1.000

Table 4: Correlation matrix for all used variables and home effect

*Fouls* Number of fouls or hands

*FoulsSuffered* Number of fouls suffered

*Offside* Number of offsides (in attack)

Obviously, some of these variables are correlated or even simple transformations of each other and, therefore, not all of the variables should be included into a regression analysis. As the variable *Fouls* is equal to the variable *FoulsSuffered* of the respective opponent (except for hands), only *FoulsSuffered* will be used. *CompletionRate* can be calculated as the ratio  $\frac{Passes - Misplaced}{Passes}$  and seems to be a sensible and very informative variable for the passing behavior of a team. Also, *Passes* is highly correlated with *BallPossession* with an overall correlation of 0.88 and team-specific correlations up to 0.92 for Hertha BSC Berlin. Therefore, *Passes* and *Misplaced* were excluded from the analysis. Table 4 contains the correlation matrix for all remaining variables from the data set. Note that Table 4 was generated by considering the overall pairwise correlations of the variables, over all the 34 matches and the 18 teams. It can be seen that, due to the high correlation between *Passes* and *BallPossession* also *CompletionRate* and *BallPossession* are correlated, but not too strongly.

## 4 A Paired Comparison Model Including Match-specific Covariates

In general, in paired comparison data one has to distinguish between objects and subjects. The objects in paired comparisons are the entities that are compared with respect to a certain underlying latent (or non-observable) trait. In football matches, the objects are the teams that are compared with respect to their playing abilities. The subjects are the entities that perform the respective comparison. For example, in marketing studies one often tries to determine the attractiveness

of several products by presenting pairs of the products to participants. Then, the participant (who is the subject of the paired comparison) has to decide which product is more attractive to him. In football matches, a single match itself or a match-day, respectively, can be seen as the subject that performs the comparison. The distinction between objects and subjects is particularly important when it comes to the inclusion of covariates. Covariates in paired comparisons can vary

- only over the subjects (subject-specific)
- only over the objects (object-specific)
- both over the subjects and the objects (subject-object-specific).

For each type of covariates, different modeling strategies are necessary. The variables that will be considered in the following (and are introduced in Section 3) vary both over the teams and the matchdays and, therefore, can be regarded as subject-object-specific covariates.

## 4.1 Model Specification

In the following, a model is proposed that is able to include match-specific covariates. The starting point is the basic model (1), which is able to handle ordered response values (including draws) together with a global order effect. In the context of football matches the order effect is considered as the home effect, i.e. the (possible) advantage a team has over its opponent if playing at the home ground. The order effect  $\delta$  in model (1) is a global order effect which does not vary across objects. In our extended model,  $\delta$  is replaced by  $\delta_r$  so that home effects are team-specific instead of being global effects equal for all teams.

Another and more important extension is the inclusion of match-specific, or, more technical, subject-object-specific covariates  $z_{ir}$ . The covariates are incorporated into the model with object-specific parameters  $\boldsymbol{\alpha}_r$ . As the covariates vary over the subjects (matches)  $i$ , also the playing abilities  $\gamma_{ir}$  and the response  $Y_{i(r,s)}$  now have to depend on the specific match. For that purpose, we propose to use the general model for ordinal response data  $Y_{i(r,s)} \in \{1, \dots, K\}$

$$\begin{aligned} P(Y_{i(r,s)} \leq k) &= \frac{\exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)}{1 + \exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)}, \end{aligned} \quad (2)$$

assuming that the abilities in match  $i$  are given by  $\gamma_{ir} = \beta_{r0} + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r$ . Altogether, the linear predictor of the model contains the following terms:

$\delta_r$  team-specific home effects of team  $a_r$

$\theta_k$  category-specific threshold parameters

$\beta_{r0}$  team-specific intercepts

$\mathbf{z}_{ir}$   $p$ -dimensional covariate vector that varies over teams and matches

$\boldsymbol{\alpha}_r$   $p$ -dimensional parameter vector that varies over teams.

In contrast to the playing abilities  $\gamma_r$  from model (1) the playing abilities are now extended by covariate effects.

## 4.2 Estimation and Penalization

The team-specific home effects and the inclusion of team-match-specific covariates leads to a huge increase of the model complexity. Therefore, it is reasonable to include penalty terms into the estimation procedures. The goal is to end up with a model with a moderate complexity only using the parameters that are really needed. In general, a penalized version  $l_p(\cdot) = l(\cdot) - \lambda J(\cdot)$  of the likelihood  $l(\cdot)$  will be maximized considering a general penalty term  $J(\cdot)$  controlled by a tuning parameter  $\lambda$ . In particular,  $L_1$ -type penalties on differences of coefficients will be used.

Both the home effect and the covariate effects could also be included as global parameters instead of team-specific parameters. To decide, whether the home effect or single covariate effects should be considered with team-specific or global parameters, penalty terms with respect to all pairwise differences of the respective parameters will be used. Such a penalty is able to set differences between parameters to exactly zero and, therefore, to find clusters of teams with equal effects. Furthermore, it is also possible that single covariates have no effect at all and are excluded from the model completely.

First, the penalty term for the home effects is considered. Penalizing all pairwise absolute differences leads to the penalty term

$$P_\delta(\delta_1, \dots, \delta_m) = \sum_{r < s} |\delta_r - \delta_s|. \quad (3)$$

As stated before, the penalty can lead to differences of exactly zero so that  $\delta_r = \delta_s$  for  $r, s \in \{1, \dots, m\}$ . If several differences are set zero, one gets clusters of teams with equal home effects. In the most extreme case ( $\lambda \rightarrow \infty$ ), all differences are estimated to be zero which leads to a model with a global home effect  $\delta = \delta_1 = \dots = \delta_m$  equal across all teams. As there is no doubt about the general presence of a home effect in national league football, no additional penalty on the absolute values of the home effects is applied.

Second, the penalty term for the covariate effects is considered. Here, in addition to all pairwise absolute differences between the parameters that correspond

to one covariate, also all absolute values are penalized using the penalty term

$$P_{\alpha}(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m) = \sum_{j=1}^p \sum_{r<s} |\alpha_{rj} - \alpha_{sj}| + \sum_{j=1}^p \sum_{r=1}^m |\alpha_{rj}|. \quad (4)$$

In contrast to the home effect, for the other covariates it is not known in advance if a certain covariate is influential at all. Therefore, an additional penalty on the absolute values is introduced. Now, in the most extreme case ( $\lambda \rightarrow \infty$ ) all covariates are excluded completely from the model. With a decreasing tuning parameter  $\lambda$ , single covariates enter the model, either with equal effects for all teams or with different clusters of teams.

Both penalties are combined resulting in a joint penalty term  $J(\cdot) = P_{\delta}(\cdot) + P_{\alpha}(\cdot)$ . In general, the tuning parameter  $\lambda$  bridges between two extreme models, namely model (1) and model (2). While model (1) contains a global home effect and no covariate effects at all, model (2) contains (different) team-specific home and covariate effects. Starting from model (1), the team-specific playing abilities  $\gamma_r$  coincide with the team-specific intercepts  $\beta_{r0}$  from model (2). With decreasing tuning parameter  $\lambda$ , additional covariate effects enter the model. In general, it can be assumed that there is a strong correlation between some covariates and the team-specific intercepts. For example, stronger teams certainly have (on average) higher values for the shots on goal than weaker teams. As, in contrast to the covariate effects, the intercepts are not penalized, in such a case the effect of the shots on goal is already covered by the regular team-specific intercepts. Therefore, the covariate effects can be seen as extensions of the playing abilities, containing *additional* effects that are not yet covered by those. In that sense, the covariate effects can help to explain (unexpected) match results which can not fully be explained solely by the team-specific intercepts. As the (unpenalized) team-specific intercepts can be expected to cover most of the abilities of the teams, the covariates only become relevant if teams over- (or under-)perform in certain matches. In order to investigate how the team abilities depend on the single covariates, in Section 5 a second model without team-specific intercepts is applied to the data.

In order to achieve comparable effects of the different penalty terms on the parameters of the different covariates, the covariates have to be transformed into a common scale. For that purpose, all values corresponding to the home effect and the covariates (across all matches and all teams) are scaled to a variance of one. Consequently, due to the scaling the magnitude of parameter estimates is comparable between different covariates.

In general, for regularization techniques a crucial point is the determination of the optimal tuning parameter. Mostly, two different strategies can be applied, namely model selection criteria (e.g. AIC or BIC) or cross-validation. While AIC or BIC use the models complexity in terms of the degrees of freedom of the models, cross-validation is solely based on out-of-sample prediction. While

the determination of the degrees of freedom is lively discussed for different models (and regularization techniques), cross-validation is applicable in almost all circumstances. Therefore, in this work the optimal tuning parameter  $\lambda$  is determined by 10-fold cross-validation with respect to the so-called ranked probability score (RPS). The RPS for ordinal response  $y \in \{1, \dots, K\}$  (Gneiting and Raftery, 2007) 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), it takes the ordinal structure of the response into account.

### 4.3 Results

For easier interpretation of the intercepts, the covariates were centered (per team around the team-specific means). Centering the covariates only changes the paths (and interpretation) of the team-specific intercepts. Now, a team-specific intercept represents the ability of a team if all covariates are assumed to be equal to the team-specific means. The paths and the interpretation of the covariate effects remain unchanged, representing the effect of a covariate for the team ability when the respective covariate changes.

Figure 1 shows the coefficient paths for all parameter estimates (except the threshold parameters  $\theta_k$ ) along (a transformation of) the tuning parameter  $\lambda$ . All paths corresponding to one covariate are collected in a separate plot. For a large tuning parameter  $\lambda$ , the home effects start with one joint cluster of all teams and (with decreasing  $\lambda$ ) end up with separate home effects for all teams. Similarly, all covariate effects start with an effect of zero and end up with separate effects for all teams. As the intercept parameters are not penalized, they only vary due to the changes of the covariate effects. Consequently, for large  $\lambda$  model (1) and for  $\lambda = 0$  the unpenalized model (2) is obtained, respectively. In general, the clustering effect of the penalty becomes obvious. For example, for the covariate *Distance* a joint cluster of all teams is formed with decreasing tuning parameters. In contrast, for *CompletionRate* only single teams like Bayern München and Borussia Dortmund form clusters of their own. The dashed vertical lines represent the optimal model according to the 10-fold cross-validation. Compared to the most complex model possible, the complexity of the final model found by the cross-validation is clearly reduced. Figure 2 shows the results of the cross-validation along the tuning parameter  $\lambda$ .

Table 5 shows all final parameter estimates (for the model chosen by cross-validation) separately per team and per covariate. *Distance* has the largest effects among all covariates, it takes the value 1.01 for all teams. Therefore, in general

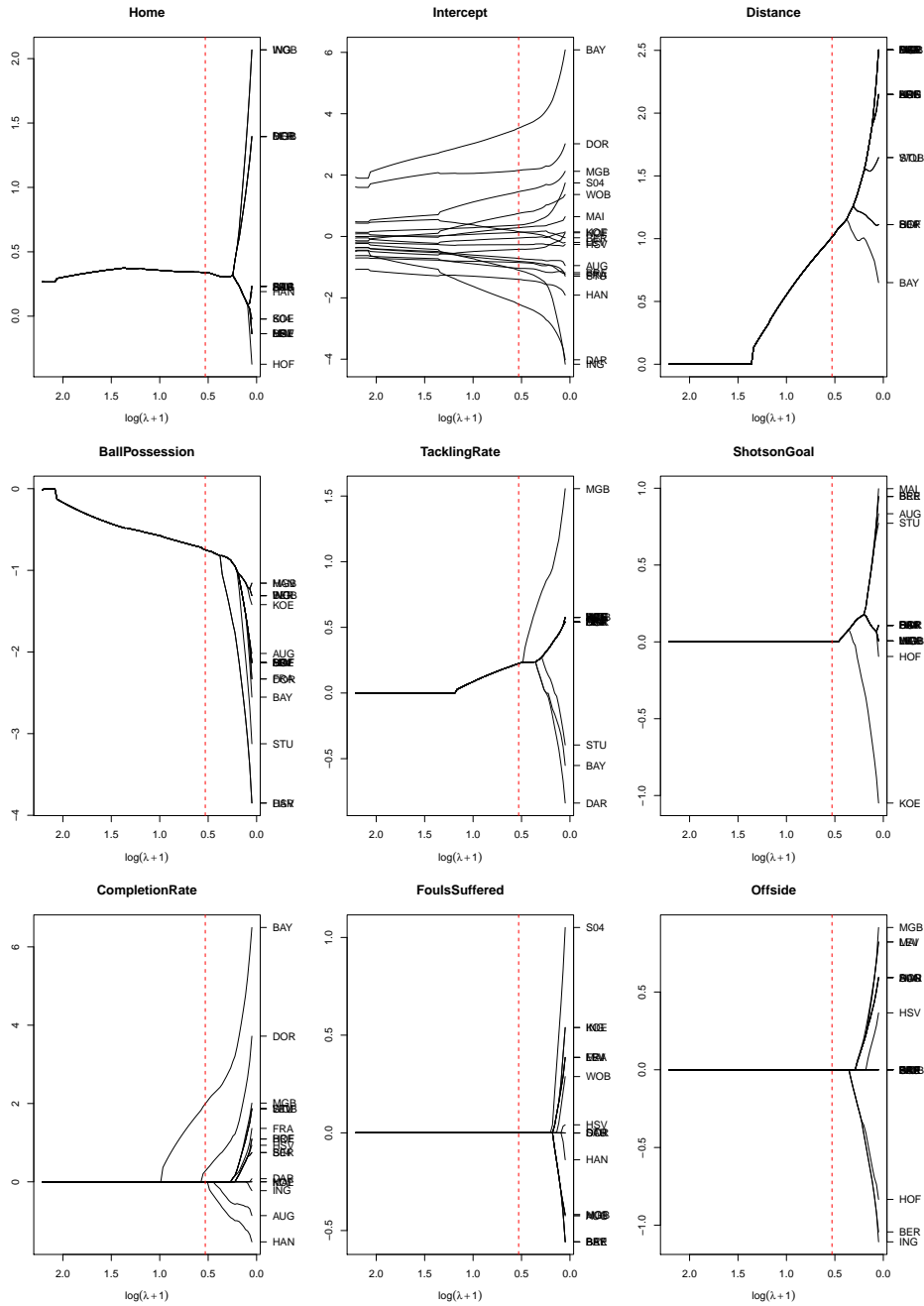


Figure 1: Coefficient paths (along sequence of  $\lambda$ ) for model (2) separately for all covariate effects. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

a better (worse) running performance of a team clearly improves (diminishes) the chances of the team for a good result. The second largest effect corresponds

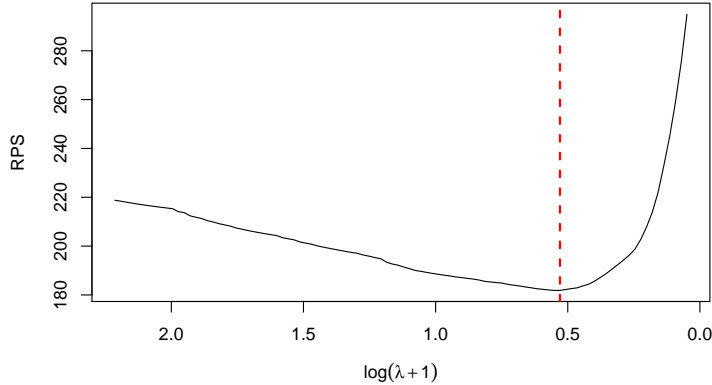


Figure 2: Cross-validation error along tuning parameter  $\lambda$  for model (2). Dashed vertical line represents optimal model according to 10-fold cross-validation.

to the covariate *BallPossession*. Interestingly, it has a negative effect for all teams. However, one has to keep in mind that there are correlations between some covariates. In particular, the variables *BallPossession* and *CompletionRate* are fairly correlated. For Borussia Dortmund and, especially for Bayern München, the team-specific effects of the *CompletionRate* are positive and, therefore, act contrarily to the negative effect of *BallPossession*. Furthermore, a positive home effect and a positive effect for the *TacklingRate* are estimated for all teams. The covariates *ShotsonGoal*, *FoulsSuffered* and *Offside* are excluded completely from the final model. While this result might have been expected for the latter two variables, it may seem somewhat surprising for *ShotsonGoal*. However, one has to keep in mind that all covariate effects are additional effects to the general abilities represented by the (unpenalized) team-specific intercepts.

In order to illustrate the overall importance of the single covariate effects, Figure 3 displays the paths of the  $L_2$ -norms

$$\|(\alpha_{1j}^\lambda, \dots, \alpha_{18j}^\lambda)\|$$

of the single covariates  $j$  along the tuning parameter  $\lambda$ . In contrast to Figure 1, in Figure 3 it is easier to compare the magnitude of the different covariate effects. *Distance* is by far the most influential variable followed by *BallPossession*.

The covariate effects of model (2) have a very specific interpretation. Every team has an (unpenalized) intercept that reflects the average ability of the team over the season. Therefore, the intercepts already cover the mean covariate effects of all teams. Accordingly, the covariate effects captured in the respective parameter vectors  $\alpha_r$  represent effects where covariates can explain deviations of the performance of a team from its average performance. This fact has to be kept in mind for the interpretation of the covariate effects from model (2).



















		Home	Intercept	Distance	BallPossession	TacklingRate	ShotsonGoal	CompletionRate	FoulsSuffered	Offside
AUG		0.34	-0.71	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
BAY		0.34	3.53	1.01	-0.75	0.22	0.00	1.99	0.00	0.00
BER		0.34	0.14	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
BRE		0.34	-0.81	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
DAR		0.34	-2.21	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
DOR		0.34	2.15	1.01	-0.75	0.22	0.00	0.27	0.00	0.00
FRA		0.34	-1.03	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
HAN		0.34	-1.40	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
HOF		0.34	-0.42	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
HSV		0.34	-0.27	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
ING		0.34	-1.10	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
KOE		0.34	-0.05	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
LEV		0.34	0.15	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
MAI		0.34	0.28	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
MGB		0.34	1.46	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
S04		0.34	0.37	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
STU		0.34	-0.85	1.01	-0.75	0.22	0.00	0.00	0.00	0.00
WOB		0.34	0.75	1.01	-0.75	0.22	0.00	0.00	0.00	0.00

Table 5: Parameter estimates of Model (2) at optimal tuning parameter according to 10-fold cross-validation.

## 5 Alternative Modeling Approach for Covariate Effects

If one is interested in the total effect of a covariate on the performance of single teams, a different parameterization seems appropriate. In an alternative approach, the team-specific intercepts are simply eliminated from the model. In this parametrization, the specific ability of team  $a_r$  on matchday  $i$  is specified by  $\gamma_{ir} = \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r$  instead of  $\gamma_{ir} = \beta_{r0} + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r$  as in model (2). Therefore, with this alternative parameterization the model can be denoted by

$$\begin{aligned}
P(Y_{i(r,s)} \leq k) &= \frac{\exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})} \\
&= \frac{\exp(\delta_r + \theta_k + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)}{1 + \exp(\delta_r + \theta_k + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)}. \tag{5}
\end{aligned}$$



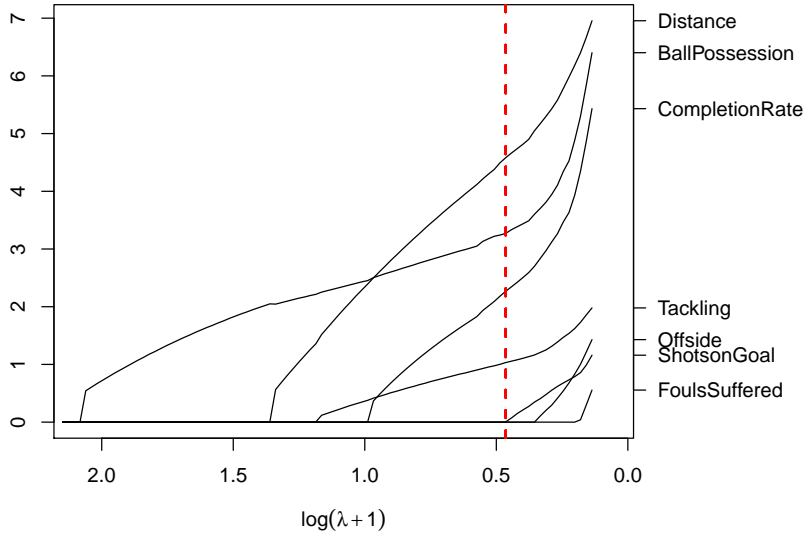


Figure 3: Variable importance with respect to the  $L_2$  norms of the variable-specific parameter vectors for model (2) along tuning parameter  $\lambda$ .

In this alternative approach, the mean abilities of the teams cannot be covered by the team-specific intercepts and have to be replaced by covariate effects. This also implies that in this alternative model the average values of the covariates for each team are relevant and, hence, the covariates are not centered per team and covariate but only per covariate. Although the team-specific intercepts  $\beta_{r0}$  are now eliminated, model (5) can still become highly complex if for each team and covariate separate effects are estimated. Therefore, again the penalty terms (3) and (4) are used for estimation.

Figure 4 shows the coefficient paths of model (5) along the tuning parameter  $\lambda$ , the dashed vertical lines represent the optimal model according to 10-fold cross-validation. Similar to the effects estimated for model (2), positive effects for **Distance** and (mostly) negative effects for **BallPossession** are found. Now, for both covariates we see different clusters of teams with equal effects. For example for **Distance**, Bayern München and Hannover 96 have slightly smaller effects than all the other teams. For **CompletionRate**, again Bayern München and Borussia Dortmund stand out. Only **FoulsSuffered** is eliminated completely from the model, all other covariates have effects for at least some of the teams.

In contrast to model (2), here the covariates were not centered per team in advance to the analyses, but only globally per covariate. This is due to the fact that in model (2) team-specific differences would all be captured in the intercepts. In model (5), no intercepts exist and differences between the teams with respect to the absolute levels of the covariates do matter. In Figure 5, the mean

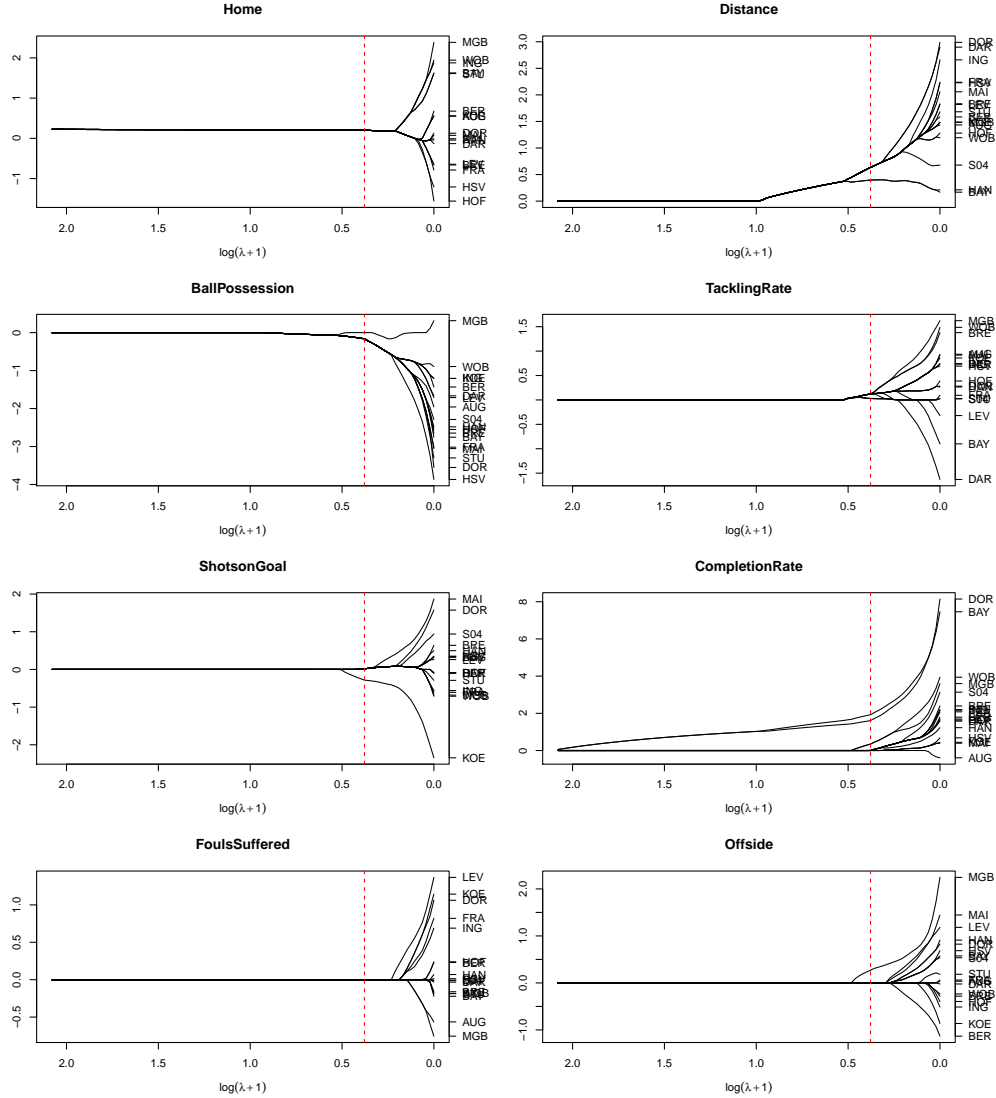


Figure 4: Coefficient paths (along relevant sequence of  $\lambda$ ) for alternative model (5), separately for all covariate effects. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

effect of the respective covariates together with the single parameter estimates is illustrated. In these so-called effect stars (Tutz and Schauburger, 2013) one can see the average covariate values (per team, per covariate) multiplied by the respective parameter estimates. Therefore, these values represent the average contribution of a covariate to the ability of a single team. More precisely, the effect stars show the exponentials of the product of average covariate values and parameter estimates. Per effect star, a circle with radius  $\exp(0) = 1$  is drawn representing the no-effect case. Values within the circle represent negative (average)

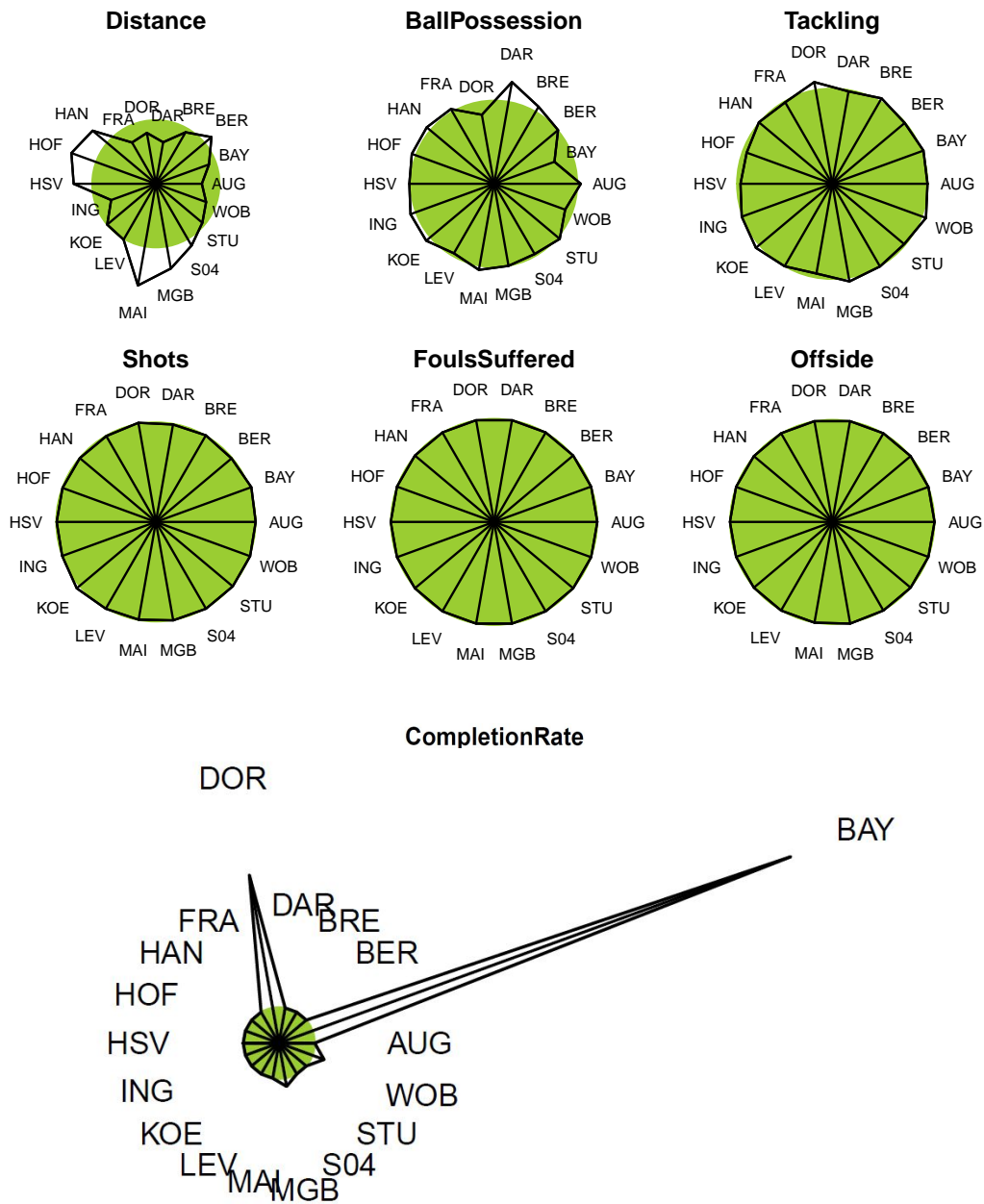


Figure 5: Effect stars for (average) covariate effects for model (5)

effects for the team ability, values beyond the circle represent positive (average) effects for the team ability. The effect star for completion rate is displayed on a different scale for better visibility. It can be seen that (together with its high mean value of `CompletionRate`) Bayern München has a huge positive effect of `CompletionRate` for the team ability, it is the variable that at most distinguishes Bayern München from the rest of the league. Also Borussia Dortmund has a big effect of `CompletionRate`. Compared to these two effects, all other effects seems negligible at first sight, but there are some noticeable effects for `BallPossession` and especially for `Distance`.

## 6 Assessment of Model Performances

Finally, it is desirable to assess the performance of the basic model (1) and the two proposed models (2) and (5). Beside comparing the models with each other, it can also be interesting to see if the models can compete with bookmakers' odds. Bookmakers' odds are known before the respective matches and aggregate most of the information that is known in advance of a match (including information not available in our data like injuries or presumable team line-ups). The Website <http://www.football-data.co.uk> provides odds averaged over different bookmakers. After eliminating the bookmakers margins, these odds are easily transformed into probabilities. In contrast to the bookmakers' odds, models (2) and (5) use covariate data which are only known after the match. Therefore, the bookmakers' odds can serve as a good benchmark. If the proposed models outperform the bookmakers' odds, this is a clear hint that the covariate information is used in a sensible manner to gain more knowledge. Of course, in practice the models can not be used for prediction as the respective covariate information is only available after a match.

The comparison of the different match predictions is performed in the following manner. To prevent effects of overfitting, the predictive power of the three models is assessed using a leave-one-out strategy. Step by step, the models are fitted (and optimized) on training data consisting of 33 matchdays. One matchday at a time is left out of the training data and the corresponding nine matches are used for prediction. To make our predictions (5 categories) compatible to the bookmakers' odds (3 categories for victory home team, draw and victory away team), the predictions are reduced to the respective 3 categories merging categories 1 and 2 and categories 4 and 5 of the response variables. In total, one ends up with 3 probabilities for each of the 306 matches of the season, separately for the three models and the bookmakers' odds. Per match, the probability of the true match outcome is stored. Table 6 contains the mean probabilities for a correct (out-of-sample) prediction for the three approaches and the bookmakers' odds:

Moreover, Figure 6 illustrates boxplots of the differences of the single pre-

Model (1)	Model (2)	Model (5)	Bookmakers
42.0%	49.8%	43.3%	41.9%

Table 6: Average (out-of-sample) probabilities for a correct match prediction

differences of the three models compared to the bookmakers odds. Here, positive values represent matches with a better prediction compared to the bookmakers' odds.

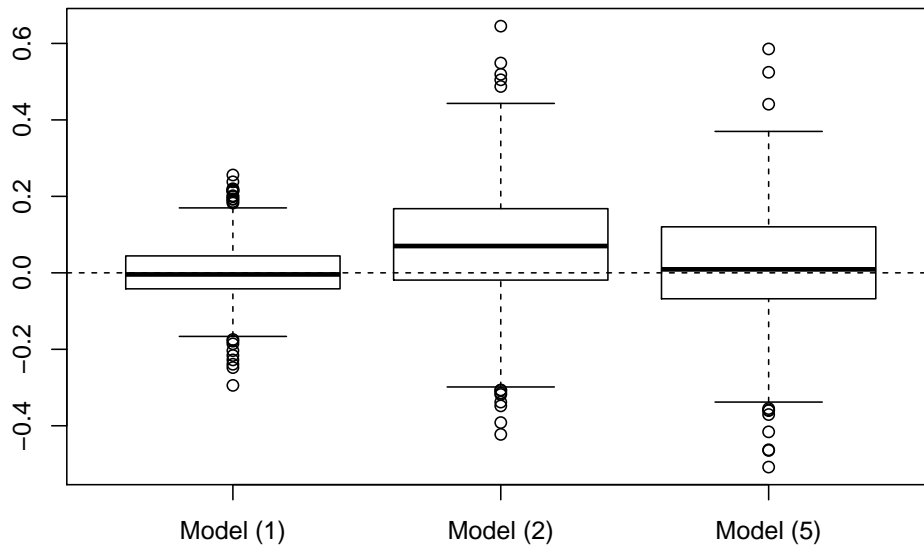


Figure 6: Boxplots of differences comparing out-of-sample predictions for realized match results with betting odds

As pointed out before, the bookmakers odds and the models use different information. After all, it can be seen that the predictive performance of model (2) outperforms all other approaches. Also models (1) and (5) can compete with the bookmakers' odds. This is surprising, as model (1) only uses mean abilities over the whole season (except for the predicted matchday) and, therefore, does not contain any match-specific information, in contrast to the bookmakers. On the other hand, the ability estimates of model (5) are solely based on match-specific covariate information without any component regarding the overall strength of a team. Also this model competes well with the bookmakers odds.

## 7 Concluding Remarks

This work deals with data from the German Bundesliga from the season 2015/16 and considers several match-specific variables in a paired comparison model. The proposed model is an attempt 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 of the matches. Due to the fact that the used covariates are correlated, a simultaneous modeling approach seems sensible. After all, complex modeling approaches are rather scarce in this area. The model treats the matches of the respective season of the German Bundesliga as paired comparisons between the competing teams, comparing the playing abilities of the teams using ordinal responses. The variables are, in a linear way, incorporated into the playing abilities of the teams in specific matches. In future work also non-linear effects might be worth considering. The model can easily be applied to data from other leagues or other types of sport.

Overall, the variable **Distance** turned out to be the most important variable among all considered variables. This finding endorses the widely spread belief that a good running performance of a team is the most important premise for a successful match. The variable **TacklingRate** also turned out to have the expected positive effect, although it is much smaller than the effect of **Distance**. In contrast, the finding of a negative effect of **BallPossession** seems to be rather counter-intuitive. Maybe, this finding reflects a new trend in the German Bundesliga (started by Borussia Dortmund and Jürgen Klopp) to focus on fast counter attacks rather than on long (and rather slow) periods of permanent ball possession. After all, for Bayern München one might also argue that they have a strong positive effect of **CompletionRate** which is quite strongly (positively) correlated to **BallPossession**. Therefore, this finding might in fact also represent a positive effect of **BallPossession** for Bayern München.

The comparison of the model performance to bookmakers' odds shows that the variables actually carry information and clearly improve the model in contrast to a model without covariate effects and that the model also outperforms the bookmakers odds. Therefore, the model seems to be a promising approach to make use of the big amount of data available in football and to better understand the effect of the single covariates for the success of teams.

## References

- Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics* 41(2), 287–297.
- Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

- Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs, I: The method of pair comparisons. *Biometrika* 39, 324–345.
- Carmichael, F., D. Thomas, and R. Ward (2000). Team performance: the case of English Premiership football. *Managerial and Decision Economics* 21(1), 31–45.
- Casalicchio, G., G. Tutz, and G. Schaubberger (2015). Subject-specific Bradley-Terry-Luce models with implicit variable selection. *Statistical Modelling* 15(6), 526–547.
- Castellano, J., D. Casamichana, and C. Lago (2012). The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of human kinetics* 31, 137–147.
- Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science* 27(3), 412–433.
- Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences* 31(2), 123–136.
- 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.
- Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46(2), 265–280.
- Dyte, D. and S. R. Clarke (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society* 51 (8), 993–998.
- 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.
- Glenn, W. and H. David (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics* 16(1), 86–109.
- Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477), 359–376.

- Groll, A., G. Schauburger, and G. Tutz (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports* 11(2), 97–115.
- Hughes, M. and I. Franks (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences* 23(5), 509–514.
- Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician* 52, 381–393.
- R Core Team (2015). *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.
- Schauburger, G. (2015). *BTLLasso: Modelling Heterogeneity in Paired Comparison Data*. R package version 0.1-2.
- Schauburger, G. and G. Tutz (2015). Modelling heterogeneity in paired comparison data - an L1 penalty approach with an application to party preference data. Technical Report 183, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- 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. Schauburger (2013). Visualization of categorical response models: From data glyphs to parameter glyphs. *Journal of Computational and Graphical Statistics* 22(1), 156–177.
- Tutz, G. and G. Schauburger (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis* 99(2), 209–227.
- Vogelbein, M., S. Nopp, and A. Hökelmann (2014). Defensive transition in soccer - are prompt possession regains a measure of success? A quantitative analysis of German Fußball-Bundesliga 2010/2011. *Journal of Sports Sciences* 32(11), 1076–1083. PMID: 24506111.