Spain retains its title and sets a new record - generalized linear mixed models on European football championships

Andreas Groll * Jasmin Abedieh †

July 2, 2012

Abstract Nowadays many approaches that analyze and predict the results of soccer matches are based on bookmakers' ratings. It is commonly accepted that the models used by the bookmakers contain a lot of expertise as the bookmakers' profits and losses depend on the performance of their models. One objective of this article is to analyze the explanatory power of bookmakers' odds together with many additional, potentially influental covariates with respect to a national team's success at European football championships. Therefore a pairwise Poisson model for the number of goals scored by national teams competing in European football championship matches is used. Moreover, the generalized linear mixed model (GLMM) approach, which is a widely used tool for modeling cluster data, allows to incorporate team-specific random effects. Two different approaches to the fitting of GLMMs incorporating variable selection are used, subset selection as well as a LASSO-type technique, including an L_1-penalty term that enforces variable selection and shrinkage simultaneously. Based on the two preceeding European football championships a sparse model is obtained that is used to predict all matches of the current tournament resulting in a possible course of the European football championship (EURO) 2012.

Keywords Soccer, EURO 2012, Sports tournaments, Generalized linear mixed model, Lasso, Variable selection.

1 Introduction

In the last years more and more attention has been devoted on the statistical analysis of major soccer events as for example the Union of European Football Associations (UEFA) Champions League (CL, see Leitner et al., 2011, Eugster et al., 2011), the European football championship (see Leitner et al., 2008, Leitner et al., 2010a or Zeileis et al., 2012) or the Fédération Internationale de Football Association (FIFA) World Cup (see Leitner et al., 2010b, Stoy et al., 2010, Dyte and Clarke, 2000).

Most of these articles deal with the challenging task of forecasting the winner of the respective tournament. The aforementioned approaches can be divided into two major

*Department of Mathematics, Workgroup Financial Mathematics, Ludwig-Maximilians-University Munich, Theresienstr. 39, 80333 Munich, Germany, *andreas.groll@math.lmu.de*

†*jasmin.abedieh@hotmail.de*

categories: The first ones are based on the easily available source of "prospective" information contained in bookmakers' odds (compare Leitner et al., 2008, Leitner et al., 2010b and Zeileis et al., 2012) and already correctly predicted the final of the EURO 2008 as well as Spain as the 2010 FIFA World Champion. But at this point one has to keep in mind that in both years Spain was rated as one of the main favorites[1]. If a tournament is instead won by a clear underdog, as for example Greece at the EURO 2004[2], a method based solely on bookmakers' odds (or solely on the market value instead) probably would have great difficulties to provide good prediction results. Though it can be expected, that the bookmakers' odds are based on complex models that cover already a huge part of the relevant information about the success of a soccer team, it would be a great benefit if additional relevant influence variables could be determined that provide further information.

This task leads to the second category of approaches that are based on regression models. In Stoy et al. (2010) a standard linear regression approach is used to analyze the success of national teams at FIFA World Cups. The success of a team at a World Cup is measured by a defined point scale that is supposed to be normally distributed. Besides some sport-specific covariates also political-economic, socio-geographic as well as some religious and psychological influence variables are considered. Based on this model, a prediction for the FIFA World Cup 2010 is obtained. Looking back, the predicted tournament outcome fits quite poorly to the true one, with already seven "wrong" teams among those who qualified for the round of sixteen. Furthermore, the predicted 2010 FIFA World Champion Brazil was already eliminated in the quater-finals. This indicates that a more sophisticated model is needed and that some additional covariates, such as bookmakers' odds, need to be considered.

An alternative approach by Dyte and Clarke (2000) predicts the distribution of scores in international soccer matches, treating each team's goals scored as independent Poisson variables dependent on two influence variables, the FIFA world ranking of each team and the match venue. Poisson regression is used to estimate parameters for the model and based on these parameters the matches played during the 1998 FIFA World Cup can be simulated.

The approach that we propose here is based on a similar model. We focus on European championships and use a pairwise Poisson model for the number of goals scored by national teams in the single matches of the tournaments. Several potential influence variables are considered and team-specific random effects are included, resulting in a flexible GLMM. The 62 matches of the European championships 2004 and 2008 serve as basis for our analysis[3], whereas each match occurs in the data set in the form of two different rows, one for each team, containing both the variables corresponding to the team whose goals are considered as well as those of its opponent. Comparing two

[1]The German state betting agency ODDSET ranked Spain on place three among the favorites for the EURO 2008 with an odd of 6.50 behind Germany (4.50) and Italy (5.50). Before the FIFA World Cup 2010 Spain was ranked on the first place among the favorites with an odd of 5.00 together with Brazil.

[2]The German state betting agency ODDSET ranked Greece on place twelve among the favorites for the EURO 2004 with an odd of 45.00.

[3]Though this represents a quite small basis of data, we abstain from using earlier European championships, as one of our main objects is to analyze the explanatory power of bookmakers' odds together with many additional, potentially influental covariates. Unfortunately, the possibility of betting on the overall cup winner before the start of the tournament is quite novel. The German state betting agency ODDSET e.g. offered the bet for the first time at the EURO 2004.

different methods for the selection of relevant predictors, we obtain a sparse solution for our model.

The first approach of variable selection is based on L_1-penalization techniques and works by combining gradient ascent optimization with the Fisher scoring algorithm and is presented in detail in Groll and Tutz (2011). It is implemented in the `glmmLasso` function of the corresponding R-package (Groll, 2011a; publicly available via CRAN, see http:// www.r-project.org). The Lasso proposed by Tibshirani (1996) has become a very popular approach to regression that uses an L_1-penalty on the regression coefficients. This has the effect that all coefficients are shrunken towards zero and some are set exactly to zero. The basic idea is to maximize the log-likelihood $l(\boldsymbol{\beta})$ of the model while constraining the L_1-norm of the parameter vector $\boldsymbol{\beta}$. Thus one obtains the Lasso estimate

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ l(\boldsymbol{\beta}), \qquad \text{subject to } ||\boldsymbol{\beta}||_1 \leq s,$$

with $s \geq 0$ and with $||\cdot||_1$ denoting the L_1-norm. Equivalently the Lasso estimate $\hat{\boldsymbol{\beta}}$ can be derived by solving the optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ \left[l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1\right], \tag{1}$$

with $\lambda \geq 0$. Both s and λ are tuning parameters that have to be determined, for example by information criteria or cross-validation. A similar approach to ours, based on a LASSO-type regularization with a cyclic coordinate descent optimization which is interesting both from an algorithmic and theoretical perspective, was proposed by Schelldorfer and Bühlmann (2011). An overview of other possible regularization methods for GLMMs such as boosting techniques can be found in Groll (2011b). A wide class of variable selection procedures for GLMMs with a focus on longitudinal data analysis is studied in Yang (2007).

A second, classical approach for the selection of predictors is subset selection. In general, the R-functions `glmmPQL` (Venables and Ripley, 2002), `glmmML` (Broström, 2009) and `glmer` (Bates and Maechler, 2010) are able to fit the underlying model. The `glmmPQL` routine is supplied by the `MASS` library. It operates by iteratively calling the R-function `lme` from the `nlme` library and returns the fitted `lme` model object for the working model at convergence. For more details about the `lme` function, see Pinheiro and Bates (2000). The `glmmML` function is supplied with the `glmmML` package (Broström, 2009) and features two different methods of approximating the integrals in the log-likelihood function, Laplace and Gauss-Hermite, whereas for the first method the results coincide with the results of the `glmmPQL` routine. Unfortunately, for both functions no model testing methods are available, thus no subset selection procedures can be performed. However, the `glmer` function from the `lme4` package (Bates and Maechler, 2010) provides model testing based on an analysis of deviance. We restrict consideration to forward procedures because forward/backward procedures imply huge computational costs. It should be mentioned that the `glmer` function also features two different methods of approximating the integrals in the log-likelihood function, Laplace and adaptive Gauss-Hermite. We focused on the former and call the corresponding forward selection procedure `glmer-select`. The results serve as a control for our L_1-penalization technique.

Finally, we compare the results of both regularization approaches in order to determine a final model, which is then used to predict the current EURO 2012. Note here,

that in contrast to other team sports, such as basketball, icehockey or handball, in soccer pure chance plays a larger role. A major reason for this is, that, compared to other sports, in soccer fewer points (goals) are scored and thus single game situations can have a tremendous effect on the outcome of the match. The consequence is that time and time again alleged underdogs win tournaments[4]. This makes predictions of soccer tournaments especially hard, so that we get the notion, that tournament wins of extreme underdogs are almost impossible to be predicted correctly by any statistical model, no matter how sophisticated the model might be. Nevertheless, we find it highly worthwile to investigate the relationship and dependency structure between different potentially influental covariates and the success of soccer teams (in our case in terms of the number of goals they score). Besides, we hope to get a little more insight into which covariates are already covered by bookmakers' odds and which covariates may give some additional useful information.

The rest of the article is structured as follows. In Section 2 we introduce the GLMM. Next, we present a list of several possible influence variables in Section 3 that will be considered in our regression analysis. The pairwise Poisson model for the number of goals is used in Section 4 to determine the covariates of a final model, which is then used in Section 5 for the prediction of the EURO 2012.

2 Generalized Linear Mixed Models - GLMMs

Let y_{it} denote observation t in cluster i, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$, collected in $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{iT_i})$. In our case, i represents a specific national team and T_i is the total number of games played by team i at the European championships under consideration. Furthermore, let $\mathbf{x}_{it}^T = (1, x_{it1}, \ldots, x_{itp})$ be the covariate vector associated with fixed effects and $\mathbf{z}_{it}^T = (z_{it1}, \ldots, z_{itq})$ be the covariate vector associated with random effects. It is assumed that the observations y_{it} are conditionally independent with means $\mu_{it} = E(y_{it}|\mathbf{b}_i, \mathbf{x}_{it}, \mathbf{z}_{it})$ and variances $var(y_{it}|\mathbf{b}_i) = \phi v(\mu_{it})$, where $v(.)$ is a known variance function and ϕ is a scale parameter. The GLMM that we consider in the following has the form

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i = \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}}, \tag{2}$$

where g is a monotonic and continuously differentiable link function, $\eta_{it}^{\text{par}} = \mathbf{x}_{it}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ including intercept and $\eta_{it}^{\text{rand}} = \mathbf{z}_{it}^T\mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{Q})$, with $q \times q$ covariance matrix $\mathbf{Q}$. An alternative form that we also use is

$$\mu_{it} = h(\eta_{it}), \quad \eta_{it} = \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}},$$

where $h = g^{-1}$ is the inverse link function. In the case of Poisson regression, which we will use in the following, the mean μ_{it} corresponds to the Poisson parameter λ_{it} and the standard link function is the natural logarithm $log(\lambda_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i$.

A closed representation of model (2) is obtained by using matrix notation. By collecting observations within one cluster, the model has the form

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i,$$

[4]There are countless examples in history for such events, throughout all competitions. We want to mention only some of the most famous ones: Germany's first World Cup success in Switzerland 1954, known as the "miracle from Bern"; Greece's victory at the EURO 2004 (compare footnote 2); FC Porto's triumph in the UEFA CL season 2003/2004.

where $\mathbf{X}_i^T = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i})$ denotes the design matrix of the i-th cluster and $\mathbf{Z}_i^T = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT_i})$. For all observations one obtains

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},$$

with $\mathbf{X}^T = [\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T]$ and block-diagonal matrix $\mathbf{Z} = diag(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. For the random effects vector $\mathbf{b}^T = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)$ one has a normal distribution with block-diagonal covariance matrix $\mathbf{Q}_\mathbf{b} = diag(\mathbf{Q}, \ldots, \mathbf{Q})$.

Focusing on GLMMs we assume that the conditional density of y_{it}, given explanatory variables and the random effect $\mathbf{b}_i$, is of exponential family type

$$f(y_{it}|\mathbf{x}_{it}, \mathbf{b}_i) = \exp\left\{\frac{(y_{it}\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi)\right\},$$

where $\theta_{it} = \theta(\mu_{it})$ denotes the natural parameter, $\kappa(\theta_{it})$ is a specific function corresponding to the type of exponential family, $c(.)$ the log-normalization constant and ϕ the dispersion parameter (compare Fahrmeir and Tutz, 2001).

One popular method to fit GLMMs is penalized quasi-likelihood (PQL), which has been suggested by Breslow and Clayton (1993), Lin and Breslow (1996) and Breslow and Lin (1995). Typically the covariance matrix $\mathbf{Q}(\boldsymbol{\varrho})$ of the random effects $\mathbf{b}_i$ depends on an unknown parameter vector $\boldsymbol{\varrho}$. In penalization-based concepts the joint likelihood-function is specified by the parameter vector of the covariance structure $\boldsymbol{\varrho}$ together with the dispersion parameter ϕ, which are collected in $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\varrho}^T)$, and parameter vector $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \mathbf{b}^T)$. The corresponding log-likelihood is

$$l(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma}) p(\mathbf{b}_i, \boldsymbol{\gamma}) d\mathbf{b}_i\right),$$

where $p(\mathbf{b}_i, \boldsymbol{\gamma})$ denotes the density of the random effects. Breslow and Clayton (1993) derived the approximation

$$l^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma})) - \frac{1}{2}\mathbf{b}^T\mathbf{Q}(\boldsymbol{\varrho})^{-1}\mathbf{b}, \tag{3}$$

where the penalty term $\mathbf{b}^T\mathbf{Q}(\boldsymbol{\varrho})^{-1}\mathbf{b}$ is due to the approximation based on the Laplace method.

PQL usually works within the profile likelihood concept. It is distinguished between the estimation of $\boldsymbol{\delta}$, given the plugged-in estimate $\hat{\boldsymbol{\gamma}}$, resulting in the profile-likelihood $l^{\text{app}}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}})$, and the estimation of $\boldsymbol{\gamma}$. The PQL method is implemented for example in the `glmmPQL` function, whereas the `glmer` and `glmmML` functions use Laplace approximation or Gauss-Hermite quadrature.

Regularization in GLMMs

The `glmmLasso` function also uses PQL and is based on the log-likelihood (3) that is expanded to include the penalty term $\lambda \sum_{i=1}^{p} |\beta_i|$. Approximation along the lines of Breslow and Clayton (1993) yields the penalized log-likelihood

$$l^{\text{pen}}(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\gamma}) = l^{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = l^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i|. \tag{4}$$

For given $\hat{\boldsymbol{\gamma}}$ the optimization problem reduces to

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\ l^{\text{pen}}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\ \left[l^{\text{app}}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) - \lambda \sum_{i=1}^{p} |\beta_i| \right]. \tag{5}$$

Our `glmmLasso` technique uses a full gradient algorithm that is based on the algorithm of Goeman (2010), for details see Groll and Tutz (2011). It can easily be amended to situations in which some parameters should not be penalized. In this case the penalty term from the optimization problem of equation (1) is replaced by $\sum_{i=1}^{p} \lambda_i |\beta_i|$, where $\lambda_i = 0$ is chosen for unpenalized parameters. The penalty used in (4) and (5) can be seen as a partially penalized approach if the whole parameter vector $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \mathbf{b}^T)$ is considered.

The gradient algorithm can automatically switch to a Fisher scoring procedure when it gets close to the optimum and therefore avoids the tendency to slow convergence which is typical for gradient ascent algorithms. An additional step is needed to estimate the variance-covariance components $\mathbf{Q}$ of the random effects. Here, two methods can be chosen, an EM-type estimate and an REML-type estimate. After convergence, a model that includes only the variables corresponding to non-zero parameters of $\hat{\boldsymbol{\beta}}$ is fitted in a final re-estimation step. A simple Fisher scoring, resulting in the final estimates $\hat{\boldsymbol{\delta}}, \hat{\mathbf{Q}}$ is used. Note, that by a small modification in the spirit of the group Lasso the `glmmLasso` function has been extended to account for grouped variables in the form of dummy-coded factors which is the case for categorical predictors.

3 Possible Influence Variables

In this section a detailed description of the covariates is given, that are used in the regression models in Section 4. Most of the variables contain information about strength and recent sportive success of national teams, as it is reasonable to assume that the current shape of a national team at the start of an European championship has an influence on the team's success in the tournament, and thus on the goals the team will score. Besides these sportive (and quite obvious) variables, also economic factors, such as a country's GDP and population size, are taken into account. Furthermore, variables are incorporated that describe the structure of a team's squad. The correlation matrix for all considered metric covariates together with the response variable *goals* is presented in Table 6 in Appendix A.

Economic Factors:

- **GDP[5] per capita:** The GDP per capita represents the economic power and welfare of a nation. It is to be expected, that countries with great prosperity will tend to focus more on sports training and promotion programs than poorer countries. In the context of success at olympic games the effect of the GDP has already been analyzed in Bernard and Busse (2004), whereas Stoy et al. (2010) investigated its influence on national teams' success at the FIFA World Cup. The GDP per capita (in US Dollar) is publicly available on the website of The World Bank (see `http://data.worldbank.org/indicator/NY.GDP.PCAP.CD`).

[5]The GDP per capita is the gross domestic product divided by midyear population. The GDP is the sum of gross values added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.

- **Population**[6]**:** The idea is, that larger countries have a deeper pool of talented soccer players from which a national coach can recruit the national team squad. Similar to Bernard and Busse (2004), we use the logarithm of the quantity, because this effect might not hold in a linear relationship for arbitrarily large numbers of populations and instead might diminish. Furthermore, national teams cannot send players in proportion to their populations, as the squads are restricted to 23 players by the UEFA standing orders for European championships.

Sportive Factors:

- **Fairness:** In modern soccer it is an inevitable strategic feature to occasionally commit a tactical foul. On the other hand, too many fouls or hard tackles are punished by yellow or red cards, resulting in disqualifications of players and match suspensions and thus having a negative effect on the team strenght. The fairness is measured by the average number of unfairness points per game (1 point for yellow card, 3 points for second yellow card, 5 points for red card) and can be found on the webpage `http://www.transfermarkt.de`.

- **Home advantage:** The existence of home advantage in soccer has often been analyzed in recent years, compare for example Pollard and Pollard (2005); Pollard (2008). Data from the FIFA World Cup (Brown et al., 2002) as well as from various soccer competitions in Europe, e.g. the English Premier league (Clarke and Norman, 1995), were used to assess the effects on home advantage. Many different factors such as crowd support, travel fatigue, familiarity, referee bias, tactics and psychology have been investigated from a sociological perspective, see for example Dawson and Dobson (2010) and Nevill et al. (1999). A dummy is used, indicating if a national team belongs to the organizing countries.

- **ODDSET odd:** For the EURO 2004 and 2008 we got the 16 odds of all possible tournament winners before the start of the corresponding tournament from the German state betting agency ODDSET. As already mentioned, one can assume that these odds contain a lot of expertise and cover big parts of the team specific information and market appreciation with respect to the tournament's favourites. Consequently, this variable plays a key role in our regression analysis. We show, that the odds can be used to determine those covariates, that are already covered by it or included, and those, that may give some additional information.

- **Market value:** For each national team participating in a European championship the average market value (in Euro) before the start of the tournament is estimated. In the last years, the market value has gained increasing importance and newly approaches for the prediction of the most renowned soccer events have been based on it (see for example Gerhards and Wagner, 2008, 2010; Gerhards et al., 2012). Estimates of market values can be found on the webpage `http://www.transfermarkt.de`[7]. The registered users of the website rate the

[6]We had to resort to different sources in order to collect data for all participating countries at the EURO 2004, 2008 and 2012. Amongst the most useful ones are `http://www.wko.at`, `http://www.statista.com/` and `http://epp.eurostat.ec.europa.eu`. For some years the populations of Russia and Ukraine had to be searched individually.

[7]Unfortunately, the archive of the webpage was established not until 4th October 2004, so the average market values of the national teams that we used for the EURO 2004 can only be seen as a

market values of single international players, and a player's market value then essentially results as an average of these ratings. Besides the transfer value of a player, the user ratings also cover aspects such as experience, future perspective or prestige of a player. Hence, a national team's average market value is a good indicator for the quality of a national team's squad.

- **FIFA points:** The current number of FIFA points of a national team accounts for all games of the team during the last four years. For each game the result, the importance of the game, the strength of the opponent, the strength of the continental associations of both teams as well as time-dependent weight factors are regarded. Thus, the FIFA points reflect a lot of information about the current strength of a national team in a world-wide comparison. The FIFA point ranking can be found on the official FIFA website `http://de.fifa.com/worldranking/rankingtable/index.html`.

- **UEFA points:** The success of clubs of the associations in the UEFA CL and the UEFA Europa League is awarded, two points for each win and one for a draw (qualifying and playoff rounds are down-weighted, penalty shootouts are not assessed)[8]. Furthermore, bonus points are allocated for the qualification into latter rounds. For the total number of points, the last five seasons are taken into account. Based on the UEFA club coefficient the number of clubs from a country's national league is determined, that participate in the UEFA CL and UEFA Europa League in the next season. Thus, the UEFA points represent the strength and success of a national league in comparison to other European national leagues. Besides, the more teams of a national league participate in the UEFA CL and the UEFA Europa League, the more experience the players from that national league are able to earn on an international level. As usually a relationship between the level of a national league and the level of the national team of that country is supposed, the UEFA points could also affect the performance of the corresponding national team. The data is available on the UEFA European cup coefficients database (see `http://kassiesa.home.xs4all.nl/bert/uefa/data/index.html`).

Factors describing the team's structure:

- **Maximum number of teammates:**[9] For each team's squad the maximum number of players, that play at the same club, has been derived. On the one hand one may argue, that it may have a positive effect on a national team's strength, if many players come from the same club and are thus experienced

rough approximation, as market values certainly changed after the EURO 2004.

[8]Note, that European national teams also gain UEFA team points. For each game played in the most recently completed full cycle (a full cycle is defined as all qualifying games and final tournament games, whereas a half cycle is defined as all games played in the latest qualifying round only) of both the latest FIFA World Cup and European championship, with addition of points for each game played at the latest completed half cycle. Similar to the FIFA points a time-dependent weight-adjustment is used, allocating to both the latest full and half cycle double the weight as to the older full cycle. Thus, the UEFA team points would reflect a lot of information about the current strength of a national team in a European-wide comparison, but as the UEFA changed the coefficient ranking system in 2008, we focused on the UEFA club ranking.

[9]Note, that this variable is not available by any soccer data provider and thus had to be counted "by hand".

and attuned to playing together. On the other hand, if a nation has many top players, these are usually scattered all over Europe's top clubs. Besides, it could also be an advantage to unite players with different experiences from all over the world. Altogether, the effect of this variable is not yet clear and needs to be further investigated in the following regression analysis.

- **Second maximum number of teammates:**[9] Similar to the previous variable, for each team's squad also the second largest number of players, that play at the same club, has been derived.[10]

- **Average age:** In general, younger soccer players are associated with qualities such as strong fitness, dynamics and rapidness, whereas older players can rely on better game experience and routine. Furthermore, it is of course depending on the player's specific position within the team, which of these abilities are essential for him. This indicates, that the "optimal" age of soccer players lies somewhere in between. For further investigation, we incorporate the average team age into the regression models, which is also available on the webpage `http://www.transfermarkt.de`.

- **Number of CL players:**[9] For each national team the number of players is derived, who reached at least the half-final with their club in the UEFA CL season preceeding the European championship under consideration. As the UEFA CL line-up consists of teams of similar high quality as a European championship (at least in the final rounds) and as its final rounds take place relatively short before the start of an European championship, this number could have a positive effect on the success of a national team. Indeed, Frohwein (2010) has already pointed out, that there is a connection between the final rounds of the UEFA CL and the FIFA World Cup final.

- **Number of Europa League players:**[9] Analogously, for each national team also the number of players is derived, who reached at least the preceeding UEFA Europa League half-final with their club. The final rounds of both UEFA CL and UEFA Europa League take place at about the same time, hence the same arguments as for the previous variable can be sustained. The main difference is, that the UEFA Europa League line-up generally consists of teams of less quality, with the consequence that the positive effect on the corresponding national team may be less pronounced.

- **Age of the national coach:**[11] This variable is used as an indicator of a national coach's experience and knowledge, which is supposed to advance with increasing age. On the other hand, the gap between coach and players may become too big at a certain age of the coach, with the consequence that mutual understanding and communication may suffer. For further investigation, we incorporate the age of the coach into the regression model.

[10]The two variables "Maximum number of teammates" and "Second Maximum number of teammates" are highly (negatively) correlated with the number of different clubs, where the players are under contract, and hence also include information about the structure of the teams' squads. Therefore, we did not consider the number of different clubs as a separate variable.

[11]This variable is available on several soccer data providers, see for example `http://www.kicker.de/`.

- **Nationality of the national coach:**[11] National affiliations always have to decide whether to choose a national coach of their country or a foreigner. The former choice would naturally have a positive effect on the communication between coach and team, but may at the same time limit the range of available coaches. Thus, the nationality of the national coach might be a relevant variable, which is to be further inspected. A dummy is used, indicating if the nationality of the coach coincides with the one of the team under his responsibility.

- **Number of players abroad:**[9] Similar to the variable "UEFA points", the number of players that are under contract in a club of a foreign national league, so-called "legionnaires", can be seen as another indicator for a national team's international experience. Besides, it could be a positive feature if players are used to be away from home for some time, because at European championships participating teams usually stay in training camps for several weeks (including the tournament itself). Sometimes this might cause psychological troubles, which was demonstrated at the current EURO 2012 by the Spanish player Jesus Navas, who produced headlines by suffering chronical home sickness.

4 Poisson Regression on the EURO 2004 and 2008

The following regression analysis is based on a mixed Poisson model with the covariates from Section 3 and the number of goals scored by national teams in the single matches of the tournaments as response variable. Team-specific random intercepts are included in order to adequately account for different basis levels of the national teams. We use two different approaches that are both able to perform variable selection, a L_1-penalization technique, which is implemented in the `glmmLasso` function, and forward subset selection based on the `glmer` function, denoted by `glmer-select`.

For the Lasso approach we obtain different levels of sparseness by changing the determination procedure of the optimal tuning parameter λ from equation (4) or (5), respectively. In the following we consider three techniques, namely Akaike's information criterion (AIC, see Akaike, 1973), the Bayesian information criterion (BIC, see Schwarz, 1978), also known as Schwarz's information criterion, as well as leave-one-out cross-validation (LOOCV). The BIC leads to the sparsest models, followed by LOOCV, whereas the AIC yields models that include several covariates, see Table 1. The sparseness of the models obtained by the forward selection procedure `glmer-select` can be controlled directly by specification of the level of significance α in the corresponding model testing, which is based on an analysis of deviance. We specify $\alpha \in \{0.01, 0.05, 0.1\}$ and show the corresponding results in Table 1.

With regard to our objectives we consider three different models, decreasing step by step the number of given influence variables. The three models are explained in detail in the following and the corresponding results, based on different levels of sparseness, can be found in Table 1. Note, that all covariates have been standardized to having an empirical mean of zero and a variance of one.

Model 1: A model containing all covariates from Section 3 is fitted. Depending on the different degree of sparseness, either the only variable selected is the *ODDSET odd* or the *ODDSET odd* together with the *fairness* of both competing teams. This indicates, that the fairness of competing teams offers some additional information, that is not yet fully covered by the bookmakers' odds. The identification of such variables was

one of our major objectives. Nevertheless, the bookmakers' odds seem to have already strong explanatory potential with respect to a national team's success at the EURO 2004 and 2008.

	`glmmLasso`			`glmer-select`		
	BIC	LOOCV	AIC	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Model 1	ODDSET[12]	ODDSET	ODDSET	ODDSET	ODDSET	ODDSET
	-	fairness	fairness	-	-	fairness
	-	fairness opp.	fairness opp.	-	-	fairness opp.
Model 2	-	fairness	fairness	fairness	fairness	fairness
	-	market value	fairness opp.	-	fairness opp.	fairness opp.
	-	-	market value	-	-	population
	-	-	max. # teamm.	-	-	-
	-	-	average age	-	-	-
	-	-	UEFA points opp.	-	-	-
Model 3	market value	market value	market value	market value	market value	market value
	-	-	max. # teamm.	-	max. # teamm.	max. # teamm.
	-	-	UEFA points opp.	-	-	-

Table 1: Selected variables for `glmmLasso` and `glmer-select` for Model 1-3 and different levels of sparseness.

Model 2: A model containing all covariates from Section 3 except for the variable *ODDSET odd* is fitted. For the settings that correspond to sparse solutions, solely the variable *fairness* (either of the team whose goals are considered only or of both competing teams) is selected. For the `glmmLasso` approach based on BIC not a single variable is detected[13] In contrast, the `glmmLasso` approach based on LOOCV also chooses the variable *market value* and `glmmLasso` based on AIC additionally chooses the variables *average age*, *UEFA points opponent* and *maximum number of teammates*. `glmer-select` with $\alpha = 0.1$ also includes the *population*. A comparison between Model 1 and Model 2 allows some conclusions concerning the construction of bookmakers' odds and gives some insight, which covariates may also be relevant for the bookmakers. As in Model 2 no information about the ODDSET odds is available to the model, all variables that newly supervene, compared to Model 1, are possible candidates that may be integrated in the bookmakers' odds. Here, especially the variable *market value* turns out to explain partial information covered by the odds, but also *average age*, *UEFA points opponent* and *maximum number of teammates* may be integrated in the odds to some extent. If one compares Model 1 and Model 2 for `glmer-select` with $\alpha \in \{0.01, 0.05\}$, the results indicate, that also the variables *fairness* and *population* correlate with the odds. This was a second major objective of our analysis. Note here, that fairness was already identified in Model 1 to contain some additional information in comparison to the odds. This can be explained by a closer look on the correlation structure in Table 6 in Appendix A. On the one hand, the variables *ODDSET odd* (V6, $cor_{V1,V6} = -0.31$) and *fairness* (V2, $cor_{V1,V2} = -0.25$) manifest the two largest correlations with the response variable (V1), while on the other hand yielding a high correlation ($cor_{V2,V6} = 0.33$) between themselves.

[12]For reasons of clearness we simply write "ODDSET" for the variable *ODDSET odds*, abbreviate opponent typing "opp." and abbreviate *maximum number of teammates* typing "max. # teamm."

[13]A closer look on the coefficient paths of this model shows, that the variable *fairness* (V2) is included shortly before the *market value* (V9), with a little larger correlation with the response variable *goals* (V1; $cor_{V1,V2} = -0.25$ and $cor_{V1,V9} = 0.24$, see Table 6 in Appendix A), but in terms of BIC the incorporation of *fairness* already deteriorates the model fit. Note, that in Model 3, where the variable *fairness* is omitted, now the BIC-based approach includes the *market value*.

Model 3: As the variable *fairness* is not available for the prediction of future European championships, finally we fit a model containing all covariates from Section 3 except for the variables *ODDSET odd* and *fairness*. While for the `glmmLasso` approach based on BIC and LOOCV only the variable *market value* is detected, `glmmLasso` based on AIC chooses also the variables *UEFA points opponent* and *maximum number of teammates*[14] .

In general the results for `glmer-select`, which serves as a control for our L_1-penalization approach, agree with those obtained by the `glmmLasso` function, but are somewhat sparser. In Figure 1 the BIC, the LOOCV score and the AIC for the Lasso approach are plotted against the penalty parameter λ on a fine grid, exemplarily for Model 3. The corresponding coefficient built-ups are illustrated in Figure 2, the colored paths representing the selected variables and the grey paths representing omitted variables.

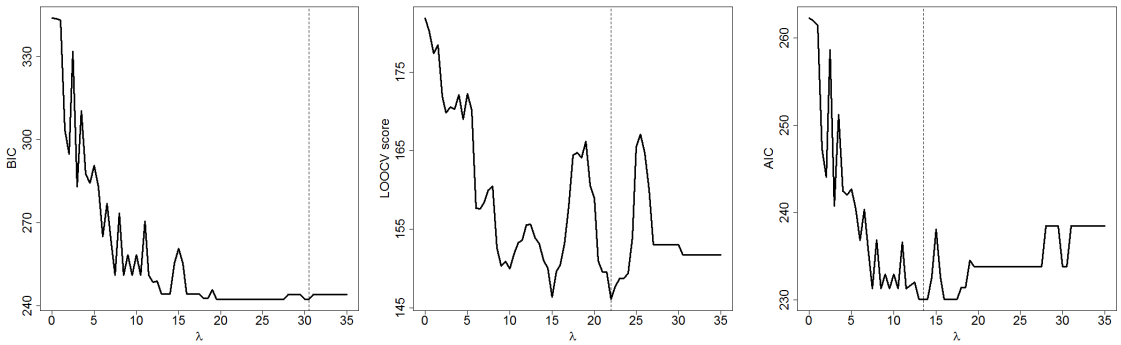

Figure 1: Results for BIC (left), LOOCV (middle) and AIC (right) for the `glmmLasso` as function of penalty parameter λ for Model 3; the optimal value of the penalty parameter λ is shown by the vertical lines.

In order to assess the performance of our models, we explain a possible goodness-of-fit criterion. In addition to the 16 odds corresponding to all possible tournament winners, which are fixed before the start of the tournament, we also got the "three-way" odds[15] from the German state betting agency ODDSET for all 62 games of the EURO 2004 and 2008. By taking the three quantities $\tilde{p}_i = 1/\text{odd}_i, i \in I := \{1, 2, 3\}$ and by normalizing with $c := \sum_{i \in I}$ in order to adjust for the bookmakers' margins, the odds can be directly transformed into probabilities using $\hat{p}_i = \tilde{p}_i / c$[16]. On the other hand, let G_k denote the random variables representing the number of goals scored by Team k in a certain match and G_l the goals of its opponent, respectively. Then we can compute the same probabilities by approximating $\hat{p}_1 = P(G_k > G_l), \hat{p}_2 = P(G_k = G_l)$ and $\hat{p}_3 = P(G_k < G_l)$ for each of the 62 matches using the corresponding Poisson distributions $G_k \sim Poisson(\hat{\lambda}_k)$, $G_l \sim Poisson(\hat{\lambda}_l)$, whereas the estimates $\hat{\lambda}_k$ and

[14] In comparison to Model 2, for `glmmLasso` based on AIC now the *average age* (V8) is not selected anymore, when the variable *fairness* (V2) is excluded. This may be due to the considerable correlation between these two variables ($cor_{V2,V8} = -0.18$, see Table 6 in Appendix A).

[15] Three-way odds consider only the tendency of a match with the possible results *winning of Team 1*, *draw* or *defeat of Team 1* and are usually fixed some days before the corresponding match takes place.

[16] The transformed probabilities only serve as an approximation, based on the assumption, that the bookmakers' margins follow a discrete uniform distribution on the three possible match tendencies.

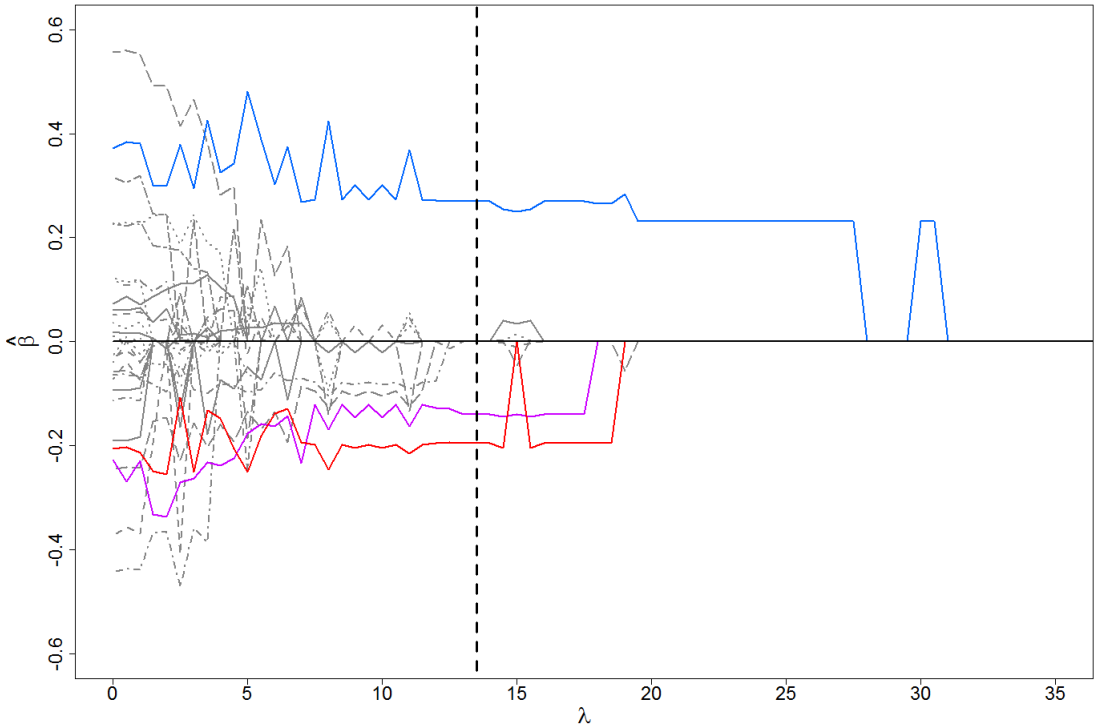

Figure 2: Coefficient built-ups for the `glmmLasso` for Model 3; colored paths represent selected variables, grey paths represent omitted variables; the optimal value of the penalty parameter λ, according to AIC, is shown by the vertical line

$\hat{\lambda}_l$[17] are obtained by our regression models. Hence, we can provide a goodness-of-fit criterion by comparing the values of the log-likelihood of the 62 matches for the ODDSET odds with those obtained for our regression models. For $\omega_i \in I, i = 1, \ldots, 62$, the likelihood is given by the product $\prod_{i=1}^{62} \hat{p}_1^{\delta_{1\omega_i}} \hat{p}_2^{\delta_{2\omega_i}} \hat{p}_3^{\delta_{3\omega_i}}$, with δ_{ij} denoting Kronecker's delta. The log-likelihood scores for `glmmLasso` and `glmer-select` corresponding to Model 1-3 and different levels of sparseness can be found in Table 2. In general, the regression models should be able to produce higher log-likelihood scores compared to the log-likelihood score corresponding to the ODDSET odds (which yields -63.81), indicating a better fit to the realized "three-way" tendencies. If the fits obtained by our models would not even be able to beat the bookmakers' odds "in sample", the whole regression analysis would be useless. That would mean, that one would achieve a better fit just by following the bookmakers' odds, which are usually publicly available shortly before the matches and thus are "out-of-sample". The results in Table 2 show, that for all settings that account for covariates, the fit obtained by our regression models outperforms the log-likelihood score corresponding to the ODDSET odds and hence, the models seem reasonable at all.

Uniting the results of all three models, we are now able to determine the final model that is used for the prediction of the EURO 2012 in Section 5. Though the bookmaker's odd for the tournament victory of a national team seems to have the

[17]For convenience we suppress the index t for both teams here, which indicates the number of the game for a team. As the match under consideration could have a different number in the individual match numbering of each team, one should correctly write $\hat{\lambda}^{(l)}_{kt_k}$ and $\hat{\lambda}^{(k)}_{lt_l}$, if Team k and Team l are facing each other in a certain match, where the superscript indicates, that the estimate is also depending on the opponent's covariates.

	glmmLasso			glmer-select		
	BIC	LOOCV	AIC	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Model 1	-62.04	-60.07	-60.07	-62.04	-62.04	-60.07
Model 2	-65.71	-61.69	-58.76	-63.57	-62.96	-61.05
Model 3	-62.76	-62.76	-60.41	-63.56	-62.74	-62.74

Table 2: Log-likelihood scores for `glmmLasso` and `glmer-select` for Model 1-3 and different levels of sparseness.

biggest influence on the number of goals scored by that team in single matches at the EURO 2004 and 2008, nevertheless we believe that not all relevant information can be covered by it. Consequently, we focus on the contribution of several, single covariates that seem to be able to adequately replace the bookmakers' odds and reflect about the same information, maybe even more. Besides, we prefer such a model, because a method based solely on bookmakers' odds probably would have great difficulties in providing good prediction results for the whole tournament development, because underdogs could hardly create a surprise, not even in single matches. Not all results in Table 1 are perfectly plausible, for example `glmmLasso` based on BIC selects not a single variable, when the variable *fairness* is incorporated (Model 2), but selects the *market value*, when *fairness* is omitted (Model 3). A similar manner is observed for the variable *average age* for `glmmLasso` based on AIC in Model 2 and 3. This may also be related to the small size of our data, but can be partly explained by the correlation structure of the variables. Therefore we decide to concern all covariates from Model 2 and 3 that have been selected in any of the aforementioned settings, except for the variable *fairness*, as it cannot be observed before the start of the tournament and thus cannot be used for prediction. This yields the following predictor:

$$\begin{aligned} \log(\lambda_{it}) &= \beta_0 + (\text{market value})_{it}\beta_1 + (\text{average age})_{it}\beta_2 \\ &\quad + (\text{population})_{it}\beta_3 + (\text{UEFA points opponent})_{it}\beta_4 \\ &\quad + (\text{maximum number of teammates})_{it}\beta_5 + b_i, \end{aligned} \tag{6}$$

where λ_{it} denotes the expected number of goals scored by team i at game t and $b_i \sim N(0, \sigma_b^2)$ represent team-specific random intercepts. The corresponding fit is easily obtained by using e.g. the `glmmPQL` function, the results are presented in Table 3. As expected, the variable *market value* has a clear positive effect on the number of goals a national team scores. Also the *population* has a slightly positive effect, while

	Coefficients	Standard errors
Intercept	0.149	0.089
market value	0.266	0.097
average age	-0.091	0.093
population	0.012	0.115
UEFA points opp.	-0.135	0.080
max. number of teammates	-0.201	0.093
$\hat{\sigma}_b$	0.170	-

Table 3: Estimates for the final model from equation (6) with `glmmPQL`.

the variable *UEFA points opponent* has a negative effect. What is more remarkable is the negative effect of the variable *maximum number of teammates*. Thus, the positive

effect of having players that are experienced and attuned to playing together seems to be predominated by a lack of top players, who are usually scattered all over Europe's top clubs, and by a lack of players with foreign experiences. Also the *average age* has a moderate negative effect, indicating that players' fitness, dynamics and rapidness have become more decisive in modern soccer as qualities like game experience and routine. The standard errors in Table 3 show that most covariates are significant, except for *average age* and especially *population*, which is far from significance. The final model from equation (6) yields a rather respectable fit with an "in sample" log-likelihood score of -59.86.

In addition we show the estimated random intercepts for the 20 different national teams, that participated in the EURO 2004 and 2008. They can be seen as representing the team-specific playing ability that is not covered by the explanatory variables (see Table 4). For example the Netherlands were rather successful (and scored many goals

Team	$\hat{b}_i$
NED	0.161
TUR	0.128
CZE	0.065
ENG	0.064
SWE	0.062
POR	0.059
GRE	0.057
RUS	0.042
CRO	0.030
GER	-0.005
LVA	-0.012
FRA	-0.040
DEN	-0.045
ROU	-0.049
BUL	-0.055
SUI	-0.070
AUT	-0.071
POL	-0.083
ESP	-0.093
ITA	-0.146

Table 4: Estimated random intercepts for national teams using `glmer`.

in their matches) both at the EURO 2004 (half-final) and at the EURO 2008 (quarter-final), although they had e.g. a medium population size, a medium average team market value and a quite high average age. Thus, the mixed model takes this into account by allocating to the Netherlands the biggest estimated random effect amongst all teams, closely followed by the Turkish team, which reached the half-finals at the EURO 2008 (Turkey did not qualify for the EURO 2004), although having a quite low average team market value. The reverse effect can be observed for Italy, which had a huge population size and a high average team market value at both tournaments, but nevertheless was not very successful (at the EURO 2004 the Italian national team

failed at group stage, at the EURO 2008 at the quarter finals). Hence the mixed model allocates a large negative random intercept to Italy.

5 Prediction of the EURO 2012

In this section we use the estimates obtained from the model in equation (6), which are based on the EURO 2004 and 2008, for the prediction of the EURO 2012. For each game of the grouping stage we compute forecasts of the number of goals scored by both teams and are thus able to forecast the whole tournament outcome. If Team k and Team l are facing each other according to the tournament schedule, we use the corresponding predictions $\hat{\lambda}_k$ and $\hat{\lambda}_l$ for the forecast of the match result, which are both depending on the covariates of both teams[18]. We suggest two different methods how $\hat{\lambda}_k$ and $\hat{\lambda}_l$ can be used in order to obtain the result for the match between Team k and Team l. Let again G_k and G_l denote the random variables representing the number of goals for the considered teams, with predicted distributions $G_k \sim Poisson(\hat{\lambda}_k)$ and $G_l \sim Poisson(\hat{\lambda}_l)$.

Method (a): The goals g_k, g_l of both teams are computed using the modes of both distributions, $g_k = mode(G_k)$ and $g_l = mode(G_l)$. The mode of a discrete random variable is defined as the realization that appears with the highest probability. Thus, this method can be seen as yielding the match results that are most likely with respect to both Poisson distributions of the team's scored goals.

Method (b): First, the difference $d = [\hat{\lambda}_l - \hat{\lambda}_k]$ is derived, where the square brackets $[\cdot]$ indicate, that the quantity is rounded to the nearest integer. Then the number of goals g_k, g_l is computed as follows:

$$g_k = \begin{cases} [\hat{\lambda}_k] & \text{if } |[\hat{\lambda}_k] - \hat{\lambda}_k| < |[\hat{\lambda}_l] - \hat{\lambda}_l| \\ [\hat{\lambda}_l] - d & \text{else ,} \end{cases}$$

$$g_l = \begin{cases} [\hat{\lambda}_k] + d & \text{if } |[\hat{\lambda}_k] - \hat{\lambda}_k| < |[\hat{\lambda}_l] - \hat{\lambda}_l| \\ [\hat{\lambda}_l] & \text{else .} \end{cases}$$

This means, that if the absolute value of the difference between $\hat{\lambda}_k$ and $\hat{\lambda}_l$ is smaller than 0.5, the game results in a draw, or more general, if the absolute value of the difference between $\hat{\lambda}_k$ and $\hat{\lambda}_l$ is smaller than $m + 0.5, m \in \mathbb{Z}$, the goal difference yields m. In a second step, for the determination of the precise match result, the individual absolute distances between $[\hat{\lambda}_k]$ and $\hat{\lambda}_k$ and between $[\hat{\lambda}_l]$ and $\hat{\lambda}_l$, respectively, are taken into account.

The results for the group stage as well as for the knockout stage of the EURO 2012 based on Method (a) can be found in Table 5 and Figure 3, the results based on Method (b) are presented in Appendix B.

The UEFA standing orders of European championships constitute, that if teams have the same number of points in the group stage, the second criterion for the determination of the final group standings is the direct comparison of these teams. The third criterion is then the goal difference. But, for example in Group A based on Method

[18]Similar to footnote 17, in the following we suppress the index for the match numbering as well as the superscripts for both teams, in order to keep the notation simple. Note here, that for the two teams of Ireland and Ukraine, that did not qualify for either EURO 2004 or 2008, no random effects estimates exist and thus their random effects are set to zero.

(a) (compare Table 5), the national teams of Greece, Russia and Czech Republic are indistinguishable with respect to these three criteria. Using Method (b), such situations even occur in all of the four groups (compare Table 7). In such situations we determine the final group standings by having a closer look on the goal differences of the "equal" teams. For each "equal" team $j, j \in \{1,2,3,4\}$, we aggregate the exact goal differences $(\hat{\lambda}_j - \hat{\lambda}_l) \in \mathbb{R}, l \in \{1,2,3,4\}, l \neq j$, resulting from its three matches against the remaining teams of the group and finally order "equal" teams with respect to the exact goal differences.

Group A			Group B			Group C			Group D		
POL	GRE	1:0	NED	DEN	2:0	ESP	ITA	2:0	FRA	ENG	1:1
RUS	CZE	0:0	GER	POR	1:1	IRL	CRO	1:1	UKR	SWE	0:1
GRE	CZE	0:0	DEN	POR	0:2	ITA	CRO	0:0	UKR	FRA	0:1
POL	RUS	1:0	NED	GER	1:1	ESP	IRL	3:0	SWE	ENG	0:2
CZE	POL	0:1	POR	NED	2:1	CRO	ESP	0:2	ENG	UKR	1 :0
GRE	RUS	0:0	DEN	GER	0:1	ITA	IRL	1:0	SWE	FRA	1:2

	Points	Goals		Points	Goals		Points	Goals		Points	Goals
POL	9	3:0	POR	7	5:2	ESP	9	7:0	ENG	7	4:2
GRE	2	0:1	GER	5	3:2	ITA	4	1:2	FRA	7	4:1
RUS	2	0:1	NED	4	3:3	CRO	2	1:3	SWE	3	2:4
CZE	2	0:1	DEN	0	0:5	IRL	1	1:5	UKR	0	0:3

Table 5: Estimated group stage results together with final group standings for the EURO 2012 using prediction method (a)

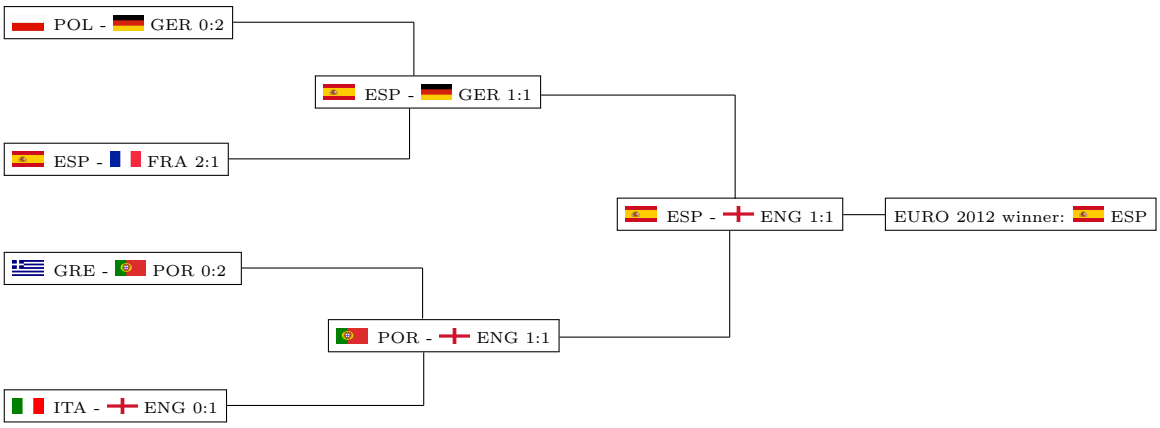

Figure 3: Estimated results of the knockout stage for the EURO 2012 using prediction method (a)

Due to the UEFA standing orders in matches of the knockout stage no draws are possible and finally a winner has to be determined, if necessary, after penalty shootouts. So if a match in the knockout stage between two Teams k and l ends in a draw, as for example the half-finals and the final in Figure 3, we just compare $\hat{\lambda}_k$ and $\hat{\lambda}_l$ and state,

that the team with the larger quantity wins the match. This can be interpreted e.g. as a narrow victory in a penalty shootout.

In general, Method (a) and (b) produce the same tournament outcome, just with different results in some single matches, but the final group standings seem much more realistic for Method (a). The biggest difference between Method (a) and Method (b) is obtained for Group A. This indicates, that here the qualities of the four teams are very similar and hence the competition in this group may be most exciting and may be settled by a few brilliant moments or decision of individual players, some essential referee decisions or simply by luck.

What is remarkable is, that, in comparison to the true tournament outcome of the EURO 2012, the model predicts seven of the eight teams correctly that have qualified for the knockout stage. Furthermore, it predicts three of the four teams correctly that have qualified for the half-finals. In addition, in the forecast both teams of Group A (even though in the forecast Poland wrongly qualifies instead of Czech Republic) are eliminated immediately during quater-finals. Note here again, that our final model from equation (6) does not use any information about bookmakers' odds. A model based solely on bookmakers' odds would have been hardly able to forecast Portugal (ODDSET-odd on EURO 2012 victory: 18) to qualify for the knockout stage in Group B instead of the Netherlands (ODDSET-odd on EURO 2012 victory: 7) or that Greece, which had the second largest odd among all sixteen participants (together with Denmark; ODDSET-odd on EURO 2012 victory: 60), succeeds to qualify for the knockout stage in Group A.

6 Concluding Remarks

A pairwise generalized linear mixed Poisson model for the number of goals scored by national teams facing each other in European football championship matches is used on data of the EURO 2004 and 2008 to analyse the influence of several covariates on the success of national teams in terms of the number of goals they score in single matches. A procedure for variable selection based on a L_1-penalty, implemented in the R-package `glmmLasso`, is used and compared to a forward subset selection approach based on the `glmer` R-function.

The major objective of this article was to analyse the explanatory power of bookmakers' odds in this context and, by incorporation of additional covariates, to get some insight into which covariates may give some information exceeding the information given by odds, and second, which covariates are already covered by bookmakers' odds. In a first regression model the fairness of national teams could be identified to contain such additional information. Nevertheless, the bookmakers' odds seem to have already strong explanatory potential with respect to a national team's success.

By a comparison of two different regression models the second task is addressed. Besides the variable *market value*, also the *average age*, the *UEFA points* of the opponent and the *maximum number of teammates* turn out to explain partial information covered by the odds. Also the variables *fairness* and *population* correlate with the odds to some extent, but fairness can not be used for prediction. Based on the other five variables a final regression model is specified and estimates for the covariate effects are derived, without further consideration of bookmakers' odds. An "in-sample" performance measure is introduced, that is based on the log-likelihood corresponding to the three-way tendencies of the considered matches.

Two methods are proposed that use these estimates for the prediction of the EURO 2012. Compared to the true tournament outcome of the EURO 2012, surprisingly many accordances are recognized. Seven of the eight teams that have qualified for the knockout stage are predicted correctly, as well as three of the four teams that have qualified for the half-finals. In contrast to methods that are strongly connected to bookmakers' odds, the model also permits some surprises by underdogs, such as for example the unexpected qualification of Greece and Portugal for the knockout stage.

Though our model could not identify the variable *number of CL players* to be influental, we believe that it has a strong explanatory potential in modern soccer. For the half-final of the current EURO 2012, with the national teams of Spain, Germany and Portugal, exactly those three teams have qualified that have the largest proportion of players amongst their squad that reached at least the half-finals of the UEFA CL 2012: Spain with 14, Germany with 10 and Portugal with 4 players. All other national teams, except for France with 3 players, have only 2 or fewer players that reached at least the half-finals of the preceeding UEFA CL season. In our view, the positive correlation between a national team's success at the European championship and the number of its players that have been successful in the preceeding UEFA CL season seems too distinct to be just a matter of chance. Therefore, we are planning to incorporate the data of the EURO 2012 into our analysis. On the one hand, the data basis is then in general more reliable, compared to the quite small data basis given by just the two tournaments from 2004 and 2008, on the other hand, we want to check again for a possible effect of the variable *number of CL players* and also revise all other results obtained in this article by an analysis based on the EUROs 2004 - 2012.

We also plan to adopt the approach presented here for the analysis of FIFA World Cups in our future work. For this tournament an even wider range of possible influence variables is available. Moreover, the cultural and geographical distances between the nations participating in the FIFA World Cup are more pronounced than for European championships, which offers lots of new aspects that are worth consideration.

Acknowledgement

We are grateful to Falk Barth from the ODDSET-Team for providing us all necessary odds data. The articel has strongly benefitted from a methodical and statistical perspective by suggestions from Christian Groll, Jan Gertheiss, Felix Heinzl and Gunther Schauberger. The insightful discussions with the hobby football experts Ludwig Weigert and Tim Frohwein also helped a lot to improve the article.

Appendix

A Correlation structure of the EURO 2004 and 2008 data

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1	1.00	-0.25	0.08	-0.15	0.03	-0.31	0.09	-0.11	0.24	0.12	0.18	0.18	0.02	-0.01	0.03
V2	-0.25	1.00	-0.31	0.37	0.21	0.33	0.22	-0.18	-0.30	-0.29	-0.16	-0.16	0.01	0.01	-0.22
V3	0.08	-0.31	1.00	-0.30	-0.36	-0.25	-0.01	0.08	0.29	0.15	0.16	0.09	0.00	-0.10	0.05
V4	-0.15	0.37	-0.30	1.00	0.63	0.10	0.33	0.00	0.15	0.07	0.22	0.17	0.22	-0.14	-0.48
V5	0.03	0.21	-0.36	0.63	1.00	-0.19	0.55	-0.13	0.33	0.21	0.48	0.14	0.23	-0.13	-0.71
V6	-0.31	0.33	-0.25	0.10	-0.19	1.00	-0.42	0.14	-0.74	-0.52	-0.61	-0.46	-0.25	0.08	0.11
V7	0.09	0.22	-0.01	0.33	0.55	-0.42	1.00	-0.37	0.47	0.36	0.64	0.20	0.62	-0.08	-0.75
V8	-0.11	-0.18	0.08	0.00	-0.13	0.14	-0.37	1.00	-0.05	0.08	-0.28	-0.31	-0.30	-0.14	0.38
V9	0.24	-0.30	0.29	0.15	0.33	-0.74	0.47	-0.05	1.00	0.54	0.84	0.71	0.20	-0.08	-0.39
V10	0.12	-0.29	0.15	0.07	0.21	-0.52	0.36	0.08	0.54	1.00	0.45	0.32	0.31	-0.15	-0.18
V11	0.18	-0.16	0.16	0.22	0.48	-0.61	0.64	-0.28	0.84	0.45	1.00	0.64	0.42	0.07	-0.65
V12	0.18	-0.16	0.09	0.17	0.14	-0.46	0.20	-0.31	0.71	0.32	0.64	1.00	0.08	0.15	-0.19
V13	0.02	0.01	0.00	0.22	0.23	-0.25	0.62	-0.30	0.20	0.31	0.42	0.08	1.00	-0.17	-0.48
V14	-0.01	0.01	-0.10	-0.14	-0.13	0.08	-0.08	-0.14	-0.08	-0.15	0.07	0.15	-0.17	1.00	0.05
V15	0.03	-0.22	0.05	-0.48	-0.71	0.11	-0.75	0.38	-0.39	-0.18	-0.65	-0.19	-0.48	0.05	1.00

Table 6: Correlation matrix of the considered metric variables for the EURO 2004 and 2008; V1=goals, V2=fairness, V3=GDP per capita,V4=maximum number of teammates, V5=second maximum number of teammates, V6=ODDSET odds, V7=population, V8=average age, V9=market value, V10=FIFA points, V11=UEFA points, V12=number of CL players, V13=number of Europa League players, V14=age of the national coach, V15=number of players abroad.

B Alternative Predictions of the EURO 2012

Group A			Group B			Group C			Group D		
POL	GRE	1:1	NED	DEN	2:1	ESP	ITA	2:1	FRA	ENG	2:2
RUS	CZE	1:1	GER	POR	2:2	IRL	CRO	1:1	UKR	SWE	0:1
GRE	CZE	1:1	DEN	POR	1:2	ITA	CRO	1:1	UKR	FRA	1:2
POL	RUS	1:1	NED	GER	2:2	ESP	IRL	3:1	SWE	ENG	1:2
CZE	POL	1:1	POR	NED	2:2	CRO	ESP	1:3	ENG	UKR	2 :1
GRE	RUS	1:1	DEN	GER	1:2	ITA	IRL	1:1	SWE	FRA	1:2

	Points	Goals		Points	Goals		Points	Goals		Points	Goals
POL	3	3:3	POR	5	6:4	ESP	9	8:3	ENG	7	6:4
GRE	3	3:3	GER	5	6:4	ITA	2	3:4	FRA	7	6:4
RUS	3	3:3	NED	5	6:4	CRO	2	3:5	SWE	3	3:4
CZE	3	3:3	DEN	0	3:6	IRL	2	3:5	UKR	0	2:5

Table 7: Estimated group stage results together with final group standings for the EURO 2012 using prediction method (b)

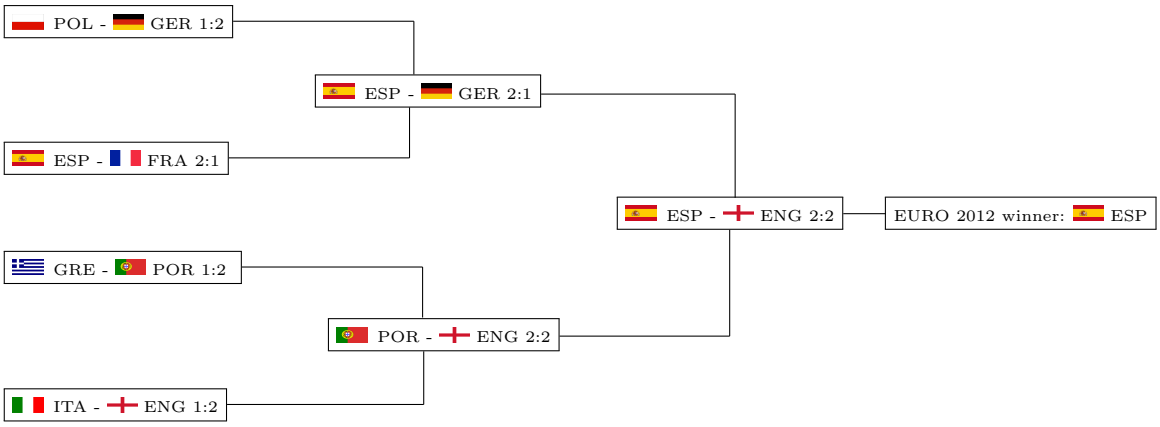

Figure 4: Estimated results of the knockout stage for the EURO 2012 using prediction method (b)

References

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267–281.

Bates, D. and M. Maechler (2010). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-34.

Bernard, A. B. and M. R. Busse (2004). Who wins the olympic games: Economic developement and medall totals. *The Review of Economics and Statistics 68*(1).

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

Broström, G. (2009). *glmmML: Generalized linear models with clustering*. R package version 0.81-6.

Brown, T. D., J. L. V. Raalte, B. W. Brewer, C. R. Winter, A. E. Cornelius, and M. B. Andersen (2002). World cup soccer home advantage. *Journal of Sport Behavior 25*, 134–144.

Clarke, S. R. and J. M. Norman (1995). Home ground advantage of individual clubs in English soccer. *The Statistician 44*, 509–521.

Dawson, P. and S. Dobson (2010). The influence of social pressure and nationality on individual decisions. evidence from the behaviour of referees. *Journal of Economic Psychology 31*, 181–191.

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)*.

Eugster, M. J. A., J. Gertheiss, and S. Kaiser (2011). Having the second leg at home - advantage in the UEFA Champions League knockout phase? *Journal of Quantitative Analysis in Sports 7*(1).

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer-Verlag.

Frohwein, T. (2010, June). Die falschen Pferde. In: e-politik.de (08.06.2010), available at: `http://www.e-politik.de/lesen/artikel/2010/die-falschen-pferde/` (12.06.2012).

Gerhards, J., M. Mutz, and G. G. Wagner (2012). Keiner kommt an Spanien vorbei - außer dem Zufall. *DIW-Wochenbericht 24*, 14–20.

Gerhards, J. and G. G. Wagner (2008). Market value versus accident - who becomes European soccer champion? *DIW-Wochenbericht 24*, 236–328.

Gerhards, J. and G. G. Wagner (2010). Money and a little bit of chance: Spain was odds-on favourite of the football worldcup. *DIW-Wochenbericht 29*, 12–15.

Goeman, J. J. (2010). L_1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal 52*, 70–84.

Groll, A. (2011a). *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L_1-Penalized Estimation*. R package version 1.0.3.

Groll, A. (2011b). *Variable selection by regularization methods for generalized mixed models*. Ph. D. thesis, University of Munich, Göttingen. Cuvillier Verlag.

Groll, A. and G. Tutz (2011). Variable selection for generalized linear mixed models by L_1-penalized estimation. Technical Report 108, Ludwig-Maximilians-University.

Leitner, C., A. Zeileis, and K. Hornik (2008). Who is Going to Win the EURO 2008? (A Statistical Investigation of Bookmakers Odds). Research report series, Department of Statistics and Mathematics, University of Vienna.

Leitner, C., A. Zeileis, and K. Hornik (2010a). Forecasting Sports Tournaments by Ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting 26 (3)*, 471–481.

Leitner, C., A. Zeileis, and K. Hornik (2010b). Forecasting the Winner of the FIFA World Cup 2010. Research report series, Department of Statistics and Mathematics, University of Vienna.

Leitner, C., A. Zeileis, and K. Hornik (2011). Bookmaker Concensus and Agreement for the UEFA Champions League 2008/09. *IMA Journal of Management Mathematics 22 (2)*, 183–194.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

Nevill, A., N. Balmer, and M. Williams (1999). Crowd influence on decisions in association football. *The Lancet 353*(9162), 1416.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal 1*, 12–14.

Pollard, R. and G. Pollard (2005). Home advantage in soccer: A review of its existence and causes. *International Journal of Soccer and Science Journal 3*(1), 25–33.

Schelldorfer, J. and P. Bühlmann (2011). GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using l1-penalization. Preprint, ETH Zurich. http://stat.ethz.ch/people/schell.

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

Stoy, V., R. Frankenberger, D. Buhr, L. Haug, B. Springer, and J. Schmid (2010). Das ganze ist mehr als die Summe seiner Lichtgestalten. Eine ganzheitliche Analyse der Erfolgschancen bei der Fußballweltmeisterschaft 2010. Working Paper 46, Eberhard Karls University, Tübingen, Germany.

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

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Yang, H. (2007). *Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis*. Ph. D. thesis, North Carolina State University.

Zeileis, A., C. Leitner, and K. Hornik (2012). History Repeating: Spain Beats Germany in the EURO 2012 Final. Working paper, Faculty of Economics and Statistics, University of Innsbruck.