IMA Journal of Management Mathematics (2011) **22**, 171–182
doi:10.1093/imaman/dpq013
Advance Access publication on October 21, 2010

Robust fitting of football prediction models

DIMITRIS KARLIS* AND IOANNIS NTZOUFRAS
Department of Statistics, Athens University of Economics and Business, 76 Patission Street, 10434 Athens, Greece
*Corresponding author: karlis@aueb.gr

[Received on 31 July 2009; accepted on 25 August 2010]

Existing methods for the prediction of the final scores in football games focus on modelling the numbers of goals scored by the two competitors with parameter estimation of the assumed model usually based on the maximum likelihood approach. Although this approach allows for sufficiently accurate prediction of the final score, it does not account for large or surprising final scores than may deteriorate parameter estimates. This is especially the case in competitions with insufficient number of games compared to the participating teams (e.g. World Cup or Champions League). In this paper, we propose a weighted likelihood approach which allows the modeller to underweight a specific football score if it is felt that the result was not typical and falsifies (in any way) the parameter estimates. The imposed game weights can be defined subjectively or by assuming a model-based structure where the parameters can be estimated by iterative algorithms. The weight structure usually reflects deviations from the assumed model. Hence, scores that have low probability under the assumed model will be underweighted. This procedure may provide robust estimates even if surprising (under the assumed model) scores are observed. Champions League data are used to demonstrate the potential of the proposed approach.

Keywords: weighted maximum likelihood; outliers; model deviation; Poisson regression models.

1. Introduction

Over the last year, increasing interest on betting industry has led to a significant demand for models that predict the outcome of football games. Since the variety of supplied bets becomes wider and their complexity increases, more sophisticated models are needed.

A series of statistical models have been proposed in the literature for the prediction of football outcomes. They can be divided in two broad categories. The first one models directly the probability of a game outcome (win/loss/draw), while the second one focusses on the match score. In this paper, we use the second category of models while our approach can be extended to the first one.

There are several models for such purpose. For example, Lee (1997) used a double Poisson model, assuming that the number of goals scored by each team can be independently modelled by Poisson regression models. Maher (1982) proposed a bivariate Poisson model, Dixon & Coles (1997) and Karlis & Ntzoufras (2003) extended the bivariate Poisson model, while McHale & Scarf (2007) proposed copulas-based models. In all the models, the issue of robustness has been overlooked.

Robustness is a major issue in statistical modelling; despite this fact, it is often not taken seriously when creating and fitting models. For example, while maximum likelihood (ML) methods are well known to be highly efficient, they are highly vulnerable to outliers and leverage points. In soccer modelling, this corresponds to some unexpectedly high scores that can influence considerably the team's estimated ability.

© The authors 2010. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

Following Grunert da Fonseca & Fieller (2006), there are two kind of achieved robustness that one would like to consider. The first one refers to contamination from outlier observations or, better, from observations that are not expected under a certain model. The second one refers to model deviation, i.e. a researcher would like to fit the model with such a method that even if the model is not correct the method would protect from deriving inconsistent results. For the soccer modelling concept such an approach would protect against selecting a wrong model, as e.g. one that assumes independence between the score of the two teams while dependence exists.

Robust methods are usually cumbersome and more computationally demanding than standard likelihood–based approaches. This is an important reason that lead to their limited practical implementation. Moreover, robust methods sacrifice a part of efficiency in order to achieve the desired degree of robustness. Hence, an appropriate trade-off between efficiency and robustness must be found.

In this paper, we propose the use of a weighted likelihood approach in order to improve robustness of the estimated model parameters. Our approach is based on creating weights for each match which can be defined as fixed prespecified quantities or can be defined through the assumed model (fixed vs. model-based weights). With fixed weights, one might wish to down weight some matches with large score difference. For example, football games with scores 3–0 and 4–0 possibly convey similar information for the attacking and defensive abilities of the two opposing teams. For this reason, we may assume them as similar by providing a fixed lower weight to the latter scores. On the other hand, model-based weights can be used to down weight observed values with low probability under the assumed model. Following this direction, we adopt the approach of Windham (1995) for robustifying a statistical model. Recall that models are ideal approximations to reality, and deviations from the assumed distribution can have important effects on classical estimators. Robustifying a model implies that we derive estimates of certain quantities of interest that are more resistant to deviations from the true model (which, in practice, is not known).

Using either type of weights, the weighted likelihood approach can be also used to specify the importance of each game depending on time sequence. Hence, older games can be given lower weights than more recent games. This approach has been used by Dixon & Coles (1997), see also Zidek & Hu (2004) for another sport related application.

The paper continues by introducing a motivating example in Section 2 which demonstrates drawbacks of standard model–based methods. The weighted likelihood approach is introduced in Section 3. In Section 4, we apply the proposed methodology on Champions League 2008–2009 data. A small simulation experiment in Section 5 examines the behaviour of the proposed approach to model misspecification issue. The paper closes with a discussion and concluding remarks in Section 6.

2. Motivating example

Let us consider the data of the first group from the group stage of UEFA (Union of European Football Associations) Champions League for season 2008–2009. Each group is composed of four teams which compete in a round-robin scheme with each couple playing twice, once in each home field.

Table 1 presents the expected number of points, the goals scored for and against each team and the probabilities for each team to end up in the first two places (1–2) qualifying in the next round, to the third place (which allows the team to continue in the UEFA cup) and the last place according to Lee (1997) double Poisson model as defined later by formulation (4.1). The actual number of points for each team are also provided within brackets. The finally observed goals are the same as the expected models under the assumed model and for this reason are omitted.

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

TABLE 1 *Simulated standings using Lee's double Poisson model for the first group of UEFA Champions League* 2008–2009 *(using* 1000 *generated leagues). Observed points are indicated within brackets*

Team	Avg.[1] points	Avg.[1] GF[2,4]	Avg.[1] GA[3,4]	Probability (%) 1st	1-2	2.5-3	3.5-4
AS Roma	11.8 (12)	11.9	6.0	49.2	87.1	10.4	2.5
Chelsea	11.1 (11)	9.0	5.1	36.9	82.7	13.7	3.6
Bordeaux	4.7 (7)	5.0	11.0	1.0	7.4	33.0	59.6
CFR 1907 Cluj	5.6 (4)	5.0	8.9	1.8	13.5	42.6	43.9

[1]Avg: Average. [2]GF: Goals for. [3]GA: Goals against. [4]Total observed goals are the same as the total fitted (property of the Lee's model).

TABLE 2 *Simulated standings using Lee's model for the first group of UEFA Champions League* 2008–2009 *(using* 1000 *generated leagues) after changing the outcome of one game. Observed points are indicated within brackets*

Team	Avg.[1] points	Avg.[1] GF[2,4]	Avg.[1] GA[3,4]	Probability (%) 1st	1-2	2.5-3	3.5-4
AS Roma	11.7 (12)	12.2	6.0	40.4	90.1	8.9	1.0
Chelsea	12.4 (11)	11.9	5.0	50.1	92.1	7.2	0.7
Bordeaux	4.9 (7)	5.0	11.1	0.8	6.7	43.4	49.9
CFR 1907 Cluj	4.6 (4)	5.1	12.1	0.5	4.6	34.7	60.7

[1]Avg: Average. [2]GF: Goals for.[3]GA: Goals against. [4]Total observed goals are the same as the total fitted (property of the Lee's model).

Let us now consider that the match of Chelsea against Cluj did not ended 2-1 (which is the true score) but 5–1. The selection of this score is indicative of the strength difference between the two teams (according to their monetary budget and the UEFA point standings) and was selected just for illustration purposes. Any other score with large goal difference will also contribute to the same direction.

This score change does not have any effect in the final group points and rankings and for this reason one would expect that it will also have small impact on model-based simulated standings. Unfortunately, this is not the case as presented in Table 2. This minor change leads to very large changes in the predictions of the model announcing now that Chelsea should be the first team in the group.

Comparing Tables 1 and 2, one can see that the probabilities for each position are affected for all participating teams. Hence, the impact of just one game is large and none of the current approaches takes into account this important aspect. The impact of just one match score is much larger for smaller leagues such as the ones formed in Champions League competition, UEFA's European Cup and FIFA's World Cup. For larger competitions (e.g. full national championships with 16–20 teams), the effect is much smaller (but still existent). Even in such competitions, the effect of single scores might be large in the first fixtures resulting to large differences of the estimated attacking and defensive abilities from week to week. The results of the example of this section clearly show that the existing approaches fail to account unexpected large scores since they model just the number of goals. Although goal scoring in football is one of the most prominent components of the sport, in terms of prediction large number

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

of scored goals seems to add little to our knowledge for the scoring ability of a team and therefore for predictive ability of the assumed model.

For this reason, in Section 3, we propose a weighted likelihood approach which enables us to assign a weight to each observation so as to consider with decreased importance surprising scores.

3. The weighted likelihood approach

In this section, we describe a weighted likelihood approach where a weight is attached to each observation. These weights aim at the reduction of the effect of outlying observations on the estimates. Other robust approaches can be based on M-estimators (see, Cantoni & Ronchetti, 2001) and minimum distance estimators (see, Lindsay, 1994). Here, we adopt the weighted likelihood due to its computational convenience. No special software is needed to produce estimates using such methods. We have implemented everything in R. However, we acknowledge that more sophisticated and computationally demanding methods can provide more robust estimates.

Suppose that we have n matches at hand. Let us assume that we observe n scores, with X_i, Y_i, $i = 1, \ldots, n$ be the number of goals scored by the home and away team, respectively. Let us further denote by $\boldsymbol{\theta}_i$ the game-specific parameters needed to calculate the joint probability $f(x_i, y_i; \boldsymbol{\theta}_i)$ of an observed score $x_i - y_i$. The game-specific parameters $\boldsymbol{\theta}_i$ are usually a function of a reduced set of parameters $\boldsymbol{\vartheta}$ which are the model parameters and are common for all data and a set of game-specific covariates $\boldsymbol{z}_i$, i.e. $\boldsymbol{\theta}_i = g(\boldsymbol{\vartheta}, \boldsymbol{z}_i)$. Typical ML approach maximizes

$$L = \sum_{i=1}^{n} \log f_i(x_i, y_i; \boldsymbol{\theta}_i) = \sum_{i=1}^{n} \log f_i(x_i, y_i; g(\boldsymbol{\vartheta}, \boldsymbol{z}_i))$$

with respect to $\boldsymbol{\vartheta}$. In this paper, we introduce the weighted likelihood which takes the form

$$L_w = \sum w_i \log f_i(x_i, y_i; \boldsymbol{\theta}_i), \tag{3.1}$$

where w_i is the weight given in the ith game. From the above, it is clear that for $w_i = 1$ for all $i = 1, \ldots, n$, then we end up to the standard ML approach.

In practice, the weights w_i can be used to give less (or more) weight to certain games. From a statistical point of view, the weights can represent the volume of information that the researcher believes that each observation carries according to some prespecified model. As an illustration, we describe two indicative types of model weights which can be certainly improved in the future but they are used here as a starting point for our research.

A first set of model weights can be specified by setting w_i to be a fixed prespecified function of the responses and the covariates

$$w_i = w(x_i, y_i, \boldsymbol{z}_i), \quad i = 1, \ldots, n.$$

Using the above set-up, the weight depends on the score and possibly on some covariate values (e.g. if a player was out or not) or any other information related to the game such as the teams motivation for that game (what do they earn or loose on that game?). Based on the experts opinion (e.g. players, managers, bookies and sport journalists), it is relatively easy to create such weights which may represent some external information which is intuitively available prior to the game.

The second approach is more complicated and assumes that the weights also depend on the unknown and under estimation model parameters $\boldsymbol{\vartheta}$. Hence, they can be written as

$$w_i = w(x_i, y_i, \boldsymbol{z}_i, \boldsymbol{\vartheta}), \quad i = 1, \ldots, n,$$

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

i.e the weight of each observation depends on the assumed model. Such 'model-based' weights usually reflect how much we trust a particular observed score after fitting a hypothesized model.

The above model-based weights can be defined once after fitting the assumed model for reasons that we will explain later on. So, this approach assumes that one fits the assumed model and then calculates the weights. The calculated weights are then used in order to refit the model based on weighted likelihood approach. The weights and the corresponding weighted likelihood estimates can be calculated iteratively (by calculating the weighted likelihood estimates for given weights and then recalculate the weights) until the global maximum is identified but this approach can be computationally demanding without improving much the robustness of the estimates. For this reason, only one step is recommended since it gives sufficient estimates for our purpose here.

Two sets of weights are proposed here for usage with football data. The first weighting scheme is simple assuming fixed predefined weights that reduce the importance of scores with large differences. A natural way to define such weights is based on the score difference. Hence, we propose to use weights with the following structure:

$$w_i = \begin{cases} 1 & \text{if } |x_i - y_i| < m_0, \\ p & \text{otherwise} \end{cases}$$

for $p < 1$. The above weighting scheme assumes that a game with score difference larger than m_0 should be down weighted and account $100p\%$ of a usual observation. The reason of reducing the importance of such score has been already discussed and assumes that for such games, the winning team plays in a more enthusiastic way than usually, while the opponent team looses any motivation due to disappointment.

For the second weighting scheme, we propose to use model-based weights that can be based on one step weighted likelihood estimates. The idea is to simply fit the model with standard ML approach and then down weight observations that had small probability to occur.

Let $f(x_i, y_i; \widehat{\boldsymbol{\theta}}_i)$ be the estimated probability for a match based on the maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\theta}}_i = g(\boldsymbol{z}_i, \widehat{\boldsymbol{\vartheta}})$ derived in the usual way. We propose to define the weights as a function of this probability. Hence, we may use

$$w_i(x_i, y_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\vartheta}}) = h\left(f(x_i, y_i; \widehat{\boldsymbol{\theta}}_i)\right).$$

A simple possible choice is $h(f) = f^q$ for $q \geqslant 0$ as proposed also by Windham (1995) in order to provide robustified versions of existing models.

By this approach, observations not relevant to the model are down weighted. Parameter q controls the volume of weighting. For $q = 0$, we have no weighting (all weights equal to one resulting in the usual MLE). As q increases we tend to give more weight to central values and less to outliers resulting in a robust estimate.

Finally, a more advanced weighting scheme may also facilitate the observed frequencies in order to down weight observations that occur more frequently than expected.

To illustrate and understand how the above weighting schemes work in practice, we briefly present two simple examples.

EXAMPLE 1 To provide some insight on the proposed methodology, let us consider the following toy example with the following nine observations:

$$0, 0, 0, 1, 1, 1, 2, 3, 10.$$

The last observation clearly looks as an outlier. Thus considering it during the estimation will inflate any estimate (especially if we focus on the mean). If we further consider the simple Poisson model then

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

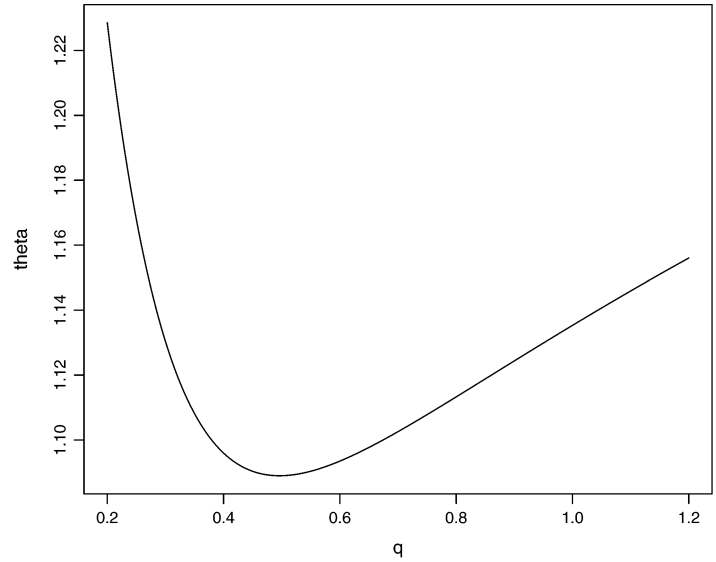

FIG. 1. One-step weighted likelihood estimate for different values of q for Example 1.

the MLE is merely the sample mean. For the full data (including the outlier), the estimated sample mean is equal to 2, while when we remove the last observation we obtain mean equal to 1.

Our approach implies obtaining as an initial estimate $\hat{\vartheta} = 2$. Then we calculate weights $w_i = f(x_i; \vartheta = 2)^q$ and we derive the weighted likelihood estimate. By this way, observations with low probability under the estimated Poisson model with mean equal to 2 will be down weighted. As we have already mentioned, the volume of down weight is controlled by q with $q = 0$ providing zero down weight and returning the usual MLEs. As q increases we tend to give more weight to central values resulting in a robust set of estimates. Figure 1 depicts the behaviour of the estimated $\hat{\vartheta}_q = \sum w_i x_i / \sum w_i$ for each value of q. For values of $q \in (0.4, 0.6)$, we obtain estimates for the mean close to the sample mean calculated when we excluded the outlier.

EXAMPLE 2 Let us now reconsider the data from the motivating example of Section 2 which refers to the data of the first group of 2008–2009 Champions League. We have calculated the weighted maximum likelihood estimates (WMLEs) for various values of q. Figure 2 depicts the changes in the probabilities of qualifying in the next round. For values close to $q = 0.8$, we achieve full robustness in the sense that the large change in the score of Chelsea against Cluj has a minor effect on the qualifying probabilities.

Computationally, WMLEs can be obtained in a straightforward manner using standard packages, like R, that allows for Poisson regression or bivariate Poisson regression by simply assigning weights to the observations. In fact, each package allowing for specifying weights for generalized linear models can be used. This makes the approach plausible for a wide audience since no special tools are needed.

4. Application

4.1 *The data*

In this section, we consider the application of the proposed methodology in the data set from UEFA Champions League for period 2008–2009. We use weighted maximum likelihood (WML) approach at

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

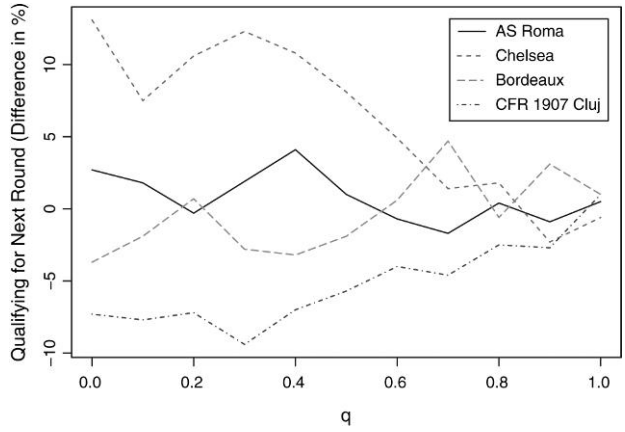

FIG. 2. Difference in probabilities of qualifying to the next round (%) against different value of q [difference ≈ 0 for $q \geqslant 0.8$].

each stage in order to predict the results of the next stage. At the first round, there are eight groups of four teams each and they play in a round-robin form. The first two teams in each group qualify to the next round and continue in a knockout type of competition. In the following section, we define a different model to allow for combining information from different groups when teams from different groups have not played one against each other.

4.2 *The model*

The Lee (1997) double Poisson model assumes, for the goals G_{1i} and G_{2i} scored by the home and away team respectively in game i, the following model structure:

$$\begin{aligned} &G_{1i} \sim \text{Poisson}(\lambda_{1i}) \text{ and } G_{2i} \sim \text{Poisson}(\lambda_{2i}), \\ &\log \lambda_{1i} = \mu + \text{home} + \text{att}_{\text{HT}_i} + \text{def}_{\text{AT}_i}, \\ &\log \lambda_{2i} = \mu + \qquad\quad \text{att}_{\text{AT}_i} + \text{def}_{\text{HT}_i}. \end{aligned} \tag{4.1}$$

The design of the Champions League competition does not allow to fit the model above since the teams in the groups stage play in distinct-isolated groups. This model will be appropriate only for each group separately but estimates are not comparable across different groups (since they express relative strength) and they cannot be used for prediction in the next knockout stage.

In order to obtain identifiable parameters, we need covariates that will connect teams competing in different groups and will not produce ill-conditioned data/design matrix. One way to achieve that is to assume common attacking and defensive parameters for teams of same countries (possibly including random effects to separate the strengths of different teams). To additionally discriminate between teams (especially of the same country), we can facilitate the UEFA ranking and scores as covariates. Such an approach makes the model identifiable since it carries information across different groups (teams of the same countries play in different groups making now the parameters comparable and the UEFA scores are numerical covariates which are common for all groups). A drawback of the model used here is that the model uses the scores of the previous years. This can be avoided if an additional covariate with the corresponding scores earned and updated within the current season is used as an additional covariate (unfortunately the data were not available in this dataset). Hence, the model accounting the above has

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

the following form: For ith game with home team HT_i playing against AT_i, we have expected counts λ_{1i} and λ_{2i} given by

$$\begin{aligned} \log \lambda_{1i} &= \mu + \text{home} + \text{co.att}_{\mathrm{CH}_i} + \text{co.def}_{\mathrm{CA}_i} + \beta_1 \mathrm{UEFA}_{\mathrm{HT}_i} + \beta_2 \mathrm{UEFA}_{\mathrm{AT}_i} \\ \log \lambda_{2i} &= \mu + \qquad\quad \text{co.att}_{\mathrm{CA}_i} + \text{co.def}_{\mathrm{CH}_i} + \beta_1 \mathrm{UEFA}_{\mathrm{AT}_i} + \beta_2 \mathrm{UEFA}_{\mathrm{HT}_i}, \end{aligned} \tag{4.2}$$

where co.att_k and co.def_k are the team and defensive parameters for teams coming from country k, CH_i, CA_i is the country origin of the home and away team (respectively) in game i, UEFA_ℓ is the UEFA score ranking for team ℓ, while β_1 and β_2 are the attacking and defensive parameters related to UEFA scores.

The proposed model is more parsimonious than than Lee's double Poisson model (4.1) since the attacking and the defensive parameters are now reduced to the number of countries from which originate the participating teams. Moreover, we have compared Lee's model with the new proposed model using the full set of data using akaike information criterion (AIC), bayes information criterion (BIC) and a 'pseudo-R^2' measure calculated by

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_i - \lambda_{i1})^2 + \sum_{i=1}^n (y_i - \lambda_{i2})^2}{\sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (y_i - \overline{y})^2}.$$

Table 3 provides these measures for the 2008–2009 Champions League data. As we can observe, both Akaike and Bayes information criteria (AIC and BIC respectively) clearly indicate that the proposed model is better. R^2 type of statistic shows that goodness of fit is similar for the two models although our proposed model uses a considerably lower number of parameters (36 instead of 64 for the independence model).

For illustration and comparison reasons, we have also fitted the bivariate model of Karlis & Ntzoufras (2003) to account for the correlation between the home and away goals. Under this model, we assume that the goals (G_{1i}, G_{2i}) scored in game i follow the bivariate Poisson distribution with parameters λ_{1i}, λ_{2i} and λ_{3i}. Recall that the marginal means for this model are $\lambda_{1i} + \lambda_{3i}$ and $\lambda_{2i} + \lambda_{3i}$, respectively, and that λ_{3i} reflect the covariance. The structure for λ_{1i} and λ_{2i} is assumed to be the same as in (4.2) while we assume the covariance parameter λ_{3i} constant across all games.

We used the ML as well as the WML approach with the two types of weights described in Section 3. All models were fitted using R software and especially the bivpois package (Karlis & Ntzoufras, 2005) with minor amendments. Both double and bivariate Poisson models provide similar results. Model 1 refers to the standard ML approach. Models 2 and 3 refer to fixed weights reducing the importance of observations from games with large score differences. The model weights are summarized by the following equations:

$$w_i = 0.5 + 0.5 * I(|d_i| \leqslant 3) \qquad \text{(model 2)}, \tag{4.3}$$

$$w_i = 0.5 + 0.25 * I(|d_i| \leqslant 2) + 0.25 * I(|d_i| = 3) \qquad \text{(model 3)}, \tag{4.4}$$

TABLE 3 *Comparison of between Lee's and the proposed double Poisson models (see equations [4.1] and [4.2], respectively) for Champions League data of season* 2008–2009. *(Lee's model assumes parameters at the level of each team, while the proposed model assumes parameters at the country level)*

	Independence model	Proposed model
AIC	761.9	733.6
BIC	981.2	858.9
R^2	38.6%	30.2%

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

TABLE 4 *Predicted results under the bivariate Poisson model for semi-finals using the data from the previous rounds*

Game	Score	Models 1–3			Model 4		
Barcelona–Chelsea	0–0	37.7	23.3	39.0			
		36.4	23.0	40.5			
		36.2	22.7	41.1	35.7	27.9	36.4
Manchester United–Arsenal	1–0	38.5	31.4	30.1			
		37.9	32.0	30.0			
		38.3	31.8	29.9	33.4	40.4	26.2
Chelsea–Barcelona	1–1	55.1	21.2	23.7			
		56.3	20.8	22.9			
		57.8	20.1	22.1	54.5	24.8	20.6
Arsenal–Manchester United	1–3	39.9	31.2	28.9			
		39.1	31.9	29.0			
		39.4	31.7	28.9	34.3	40.3	25.4

TABLE 5 *Estimated outcome probabilities for the final score using the data from the previous rounds*

	Model	Barcelona	Draw	Und	$\log \lambda_3$
1.	ML	34.5	26.8	38.7	−1.189
2.	WML1 ($w_i = 0.75, 0.5, 1$)	33.1	26.6	40.3	−1.092
3.	WML2 ($w_i = 0.5, 1$)	32.5	26.5	41.0	−1.073
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	30.6	32.7	36.7	−1.340

where $d_i = x_i - y_i = \text{goals}_{\text{HT}_i} - \text{goals}_{\text{AT}_i}$ is the goal difference in the ith game and $I(A)$ is the indicator function taking the value of one when A true and zero otherwise. In more detail, model 2 assumes weights equal to 0.5 for games with observed differences equal or higher than 3 (and one otherwise). Model 3 is more conservative since for games with difference of 3 goals then the weight is set equal to 0.75. Finally, model 4 refers to model-based weights of type (3) with $h(f) = f^{1/2}$.

Predicted outcome probabilities under the bivariate Poisson model (Karlis & Ntzoufras, 2003) are presented in Table 4 for the semi-finals using previous data (from the group stage, round of 16 and the quarter-finals). Similarly, outcome probabilities are presented for the final in Table 5. Note that Barcelona finally won united by 2-0. Additional results under the double Poisson model and comparisons are available in the second author web page.[1] From these tables, we observe minor differences ML and WML with fixed weights. For the model-based approach, we observe a considerable increase in the probability of draw in all games.

Finally in Table 6, we present the out-of-sample success rate of each model measured by $\sum_i \sum_{k=1}^{3} p_{ik} I(k = y_i)$ which is equal to the sum of the probabilities of the observed scores. If a model predicts all true scores with probability one, then this will be equal to the number of games under consideration while it will be zero if all true scores have zero probability under the assumed model. All models are compared with the double Poisson saturated model which assumes that the expected number of goals are equal to the observed ones. According to the presented results, for the the first knockout phase (phase of sixteen teams), the success rate is slightly reduced when weighted likelihood methods are used.

[1] http://stat-athens.aueb.gr/~jbn/papers/presentations/2009_IMA_Sport2_Groningen.pdf.

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

TABLE 6 *Out-of-sample success rate for the first knockout phase (of sixteen teams) and the quarter finals*

		Phase of 16		Quarter finals	
Model		Success rate	% Relative to saturated	Success rate	% Relative to saturated
1.	ML	6.26	57.8	2.72	63.1
2.	WML1 ($w_i = 0.75, 0.5, 1$)	5.96	55.1	2.71	62.8
3.	WML2 ($w_i = 0.5, 1$)	5.92	54.7	2.70	62.6
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	5.72	52.8	3.05	70.7
6.	Saturated	10.82	100.0	4.31	100.0

Nevertheless, the loss in prediction seems to be low relative to the robustness we gain. For the quarter finals, the success rates are again similar but now the model-based weighted likelihood provides slightly higher success rates (∼71% of the saturated model).

5. Accounting for model misspecification: some simulation-based evidence

The weighted likelihood approach can account for model misspecification. This is important for the modelling of football outcomes since one of the main controversies in such models is whether the assumption of independent number of goals in a game has large effect on prediction. For this reason, we conducted a small simulation experiment. The findings support that the robust approach proposed here can correct model misspecification by providing estimated probabilities closer to the true ones.

Let us consider some data generated from a bivariate Poisson model with parameters $(\lambda_1, \lambda_2, \lambda_3) = (1, 1.4, \lambda_3)$, $\lambda_3 = 0.15, 0.30$. So, in the following illustration, we falsely assume a double Poisson model ignoring the underlying correlation (which we know that exists).

Boxplots in Fig. 3 represent sum of the squared differences between the true probabilities π_{ij} and the estimated ones $\hat{\pi}_{ij}$ from the wrongly assumed double Poisson model for 1000 simulated data sets. Therefore, the sum of squared differences is given by $\sum_{i=0}^{5}\sum_{j=0}^{5}(\pi_{ij} - \hat{\pi}_{ij})^2$, where $\pi_{ij} = f_{\mathrm{BP}}(x_i, y_i; \lambda_1 = 1, \lambda_2 = 1.4, \lambda_3)$ and $\hat{\pi}_{ij} = f_{\mathrm{P}}(x_i; \hat{\lambda}_1) f_P(y_i; \hat{\lambda}_2)$ with $f_{\mathrm{BP}}(x, y; \lambda_1, \lambda_2, \lambda_3)$ denoting the probability function of the bivariate Poisson with parameters λ_k $(k = 1, 2, 3)$ and $f_{\mathrm{P}}(x; \lambda)$ denoting the probability function of the Poisson with parameter λ. $\hat{\lambda}$ implies estimated values using either the ML method or WML. Two sample sizes were used, namely $n = 100, 1000$. We also considered two different values for q, the parameter in the WML approach using weight of the form $f(x, y)^q$.

According to the generated plots, WML seems that provides probabilities much closer to the correct ones correcting by this way for the model misspecification. Hence, the WML corrects for the ignored correlation and provides better probability estimates for the outcome, implying that it is much more robust. Also a cautionary note refers to the higher variance of the WML values. This is due to the fact that each sample may have different aspects of model deviation in the sense, outliers, inliers, etc. and hence the weighting differs from sample to sample.

6. Closing remarks

In this paper, we presented the implementation of the weighted likelihood approach using easy to derive weighting schemes. The proposed methodology offers an efficient protection against outliers, i.e. games with high or unexpected scores. The selection of weights is an important issue that needs to be further

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

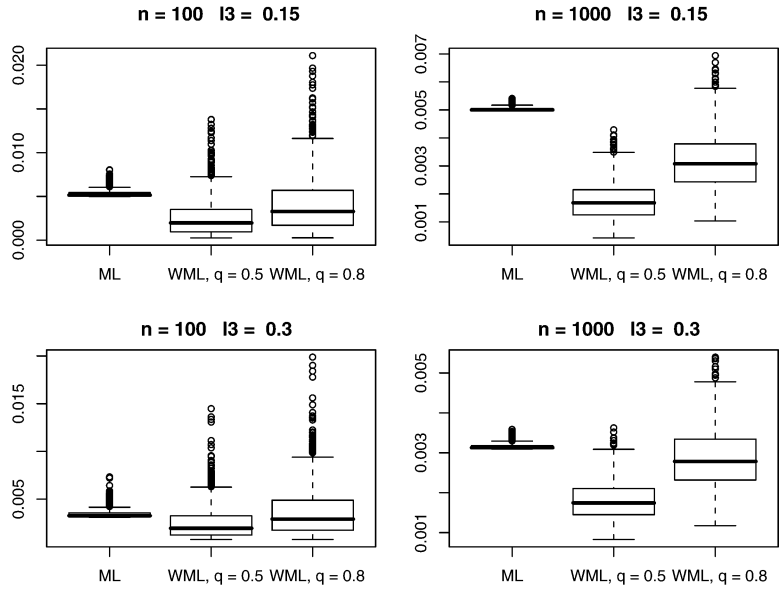

FIG. 3. Boxplots of absolute squared differences between the probabilities of the true underlying bivariate Poisson and the double Poisson model.

investigated. Fixed weights, like the ones proposed in this paper, can reasonably scale down large scores and can be incorporated in our fitting procedure in straightforward manner. Nevertheless, the selection of such weights is subjective. This subjectiveness can be avoided by model-based approaches, where weights are selected in an automatic way based on the assumed model. The function that will be used in the model weights is also a topic for further investigation. The choice of the square root presented here seems to provide more robust estimates than the corresponding proposed fixed weights. Further properties of the proposed estimates are currently under investigation by the authors. More sophisticated weights can be derived based on the work of Lindsay (1994) and Markatou *et al.* (1997). In these papers, the key idea is to compare the observed frequencies with the expected ones and create weighting schemes based on their relative differences. However, such approaches, since covariates are present, imply that we have very few observations for each combination of covariates in order to be able to make the empirical frequency a good estimator and hence compare to the theoretical one.

Returning back to our proposed methodology, a variety of different functions $h(\cdot)$ can be used to down weight outlying observations. This can be possibly a step function in a similar fashion to M-estimators. Concerning the power probability function $f(x)^q$ used in this paper, an intriguing problem is the appropriate specification or efficient estimation of the power parameter q. One possible solution might be offered by considering it as a parameter under estimation in the weighted likelihood function. This will complicate the optimization algorithm but it will possibly provide a solution under a reasonable computational burden.

Finally, another important aspect of robustness is related with model deviations. There is a debate on football modelling whether the marginal distributions are Poisson or not and whether bivariate models must be used instead. The weighted likelihood approach can provide a compromise in such a debate in the sense that it tries to correct for a model that ignores some of the data features. As we have shown in a small simulation experiment if one falsely assumes a double Poisson model while the true

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011

underlying model is a bivariate Poisson one, then the WML approach can provide estimates closer to reality protecting against this model misspecification.

REFERENCES

CANTONI, E. & RONCHETTI, E. (2001) Robust inference for generalized linear models. *J. Am. Stat. Assoc.*, **96**, 1022–1030.

DIXON, M. J. & COLES, S. G. (1997) Modelling association football scores and inefficiencies in football betting market. *Appl. Stat.*, **46**, 265–280.

GRUNERT DA FONSECA, V. & FIELLER, N. R. J. (2006) Distortion in statistical inference: the distinction between data contamination and model deviation. *Metrika*, **63**, 169–190.

KARLIS, D. & NTZOUFRAS, I. (2003) Analysis of sports data using bivariate Poisson models. *J. R. Stat. Soc. D*, **52**, 381–393.

KARLIS, D. & NTZOUFRAS, I. (2005) Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *J. Stat. Softw.*, **10**, 1–36.

LEE, A. J. (1997) Modelling scores in the Premier League: is Manchester United really the best? *Chance*, **10**, 15–19.

LINDSAY, B. G. (1994) Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.*, **22**, 1018–1114.

MAHER, M. J. (1982) Modelling association football scores. *Stat. Neer.*, **36**, 109–118.

MARKATOU, M., BASU, A. & LINDSAY, B. (1997) Weighted likelihood estimating equations: the discrete case with applications to logistic regression. *J. Stat. Plan. Inference*, **57**, 215–232.

MCHALE, I. G. & SCARF, P. A. (2007) Modelling soccer matches using bivariate discrete distributions. *Stat. Neer.*, **61**, 432–445.

WINDHAM, M. P. (1995) Robustifying model fitting. *J. R. Stat. Soc. B*, **57**, 599–609.

ZIDEK, J. & HU, F. (2004) Forecasting NBA basketball playoff outcomes using the weighted likelihood. *A Festschrift for Herman Rubin* (A. DasGupta ed.). Beachwood, OH: Institute of Mathematical Statistics, pp. 385–395.

Downloaded from imaman.oxfordjournals.org at Athens University of Economics and Business on March 16, 2011