Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference

Dimitris Karlis and Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business, Athens, GREECE; e-mails: {karlis, ntzoufras}@aueb.gr .

ABSTRACT

Modelling football match outcomes is becoming increasingly popular nowadays for both team managers and betting funs. Most of the existing literature deals with modelling the number of goals scored by each team. In the present paper we work in a different direction. Instead of modelling the number of goals directly, we focus on the difference of the number of goals, i.e. the margin of victory. We recast interest in the so-called Skellam distribution. Modelling the differences instead of the scores themselves has some major advantages. Firstly, we eliminate correlation imposed by the fact that the two opponent teams compete each other and secondly we do not assume that the scored goals by each team are marginally Poisson distributed. Application of the Bayesian methodology for the Skellam's distribution using covariates is discussed. Illustrations using real data from the English Premiership for the season 2006-2007 are provided. The advantages of the proposed approach are also discussed.

Key Words: Goal difference, Overdispersion, Poisson difference, Skellam's Distribution, Soccer, Zero Inflated Models.

1 Introduction

In the recent years, an increasing interest has been observed concerning models related to football (soccer). The increasing popularity of football modelling and prediction is mainly

due to two distinct reasons. Firstly, the market related to football has increased considerably the last years; modern football teams are profitable companies usually with large investments and budgets, while the interest of the sports funs concerning football is extremely large. The second reason is betting. The amount spent on bets have increased dramatically in Europe. As a result, the demand for models which provide good predictions for the outcome of a football game arises. Since bets are becoming more complicated, more complicated and more refined models are needed.

The statistical literature contains a series of models for this purpose. Some of them model directly the probability of a game outcome (win/loss/draw) while other formulate and predict the match score. On the other hand, the type of models applied vary according to the geographical location of each football league or tournament indicating that the characteristics of the game may be influenced by local features.

The Poisson distribution has been widely used as a simple modelling approach for describing the number of goals in football (see, for example, Lee, 1997). This assumption can be questionable in certain leagues where overdispersion (sample variance exceeds the sample mean) has been observed in the number of goals. In addition, empirical evidence has shown a (relatively low) correlation between the goals in a football game. This correlation must be incorporated in the model.

Maher (1982) discussed this issue, while Dixon and Coles (1997) extended the independent Poisson model introducing indirectly a type of dependence. Moreover, Karlis and Ntzoufras (2003) extended the bivariate Poisson model with diagonal inflation in order to account for the increased (relative to the simple bivariate Poisson model) draws observed in certain leagues. This inflation produces non-Poisson marginal distributions that can be overdispersed.

The present paper moves in a different direction. Instead of modelling jointly the number of goals scored by each team, we focus on the goal difference. By this way, we remove the effect of the correlation between the scoring performance of the two competing teams, while the proposed model does not assumes Poisson marginals. The model can be used to predict

the outcome of the game as well as for betting purposes related to the Asian handicap. However, the model cannot predict the final score.

We proceed using the Bayesian approach concerning the estimation for the model parameters along the lines of Karlis and Ntzoufras (2006). The Bayesian approach is suitable for modelling sports outcomes in general, since it allows the user to incorporate any available information about each game via the prior distribution. Information that can be incorporated in the model can be based on historical knowledge or data, weather conditions or the fitness of a team. Finally, the Bayesian approach naturally allows for predictions via the predictive distribution. This allows to predict future games and produce a posterior predictive distribution for future scores, outcomes or even reproduce the whole tournament and produce quantitative measures concerning the performance of each team.

The remaining of the paper proceeds as follows: Section 2 describes the proposed model and provides some properties and interesting points that will assist us to understand the behavior of the model and its interpretation. A zero-inflated model is also described in order to capture the (possible) excess of draws in a league. In section 3, we describe the Bayesian approach used to estimate the model. A real data application using the results of the English Premier League for the season 2006-2007 is described in section 4. Finally concluding remarks and further work is discussed in section 5.

2 The Model

2.1 Derivation

Consider two discrete random variables X and Y and their difference $Z = X - Y$. The probability function of the difference Z is a discrete distribution defined on the set of integer numbers $\mathcal{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Although publications concerning distributions defined on $\mathcal{Z}$ are rare, the difference of two independent Poisson random variables has been discussed by Irwin (1937) for the case of equal means and Skellam (1946) for the case of different means.

The Skellam's distribution (or Poisson difference distribution) is defined as the distribution of a random variable Z with probability function

$$f_{PD}(z|\lambda_1,\lambda_2) = P(Z=z|\lambda_1,\lambda_2) = e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}\left(2\sqrt{\lambda_1\lambda_2}\right). \tag{2.1}$$

for all $z \in \mathcal{Z}$, $\lambda_1, \lambda_2 > 0$, where $I_r(x)$ is the modified Bessel function of order r (see Abramowitz and Stegun, 1974, pp. 375) given by

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{k=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^k}{k!\Gamma(r+k+1)}.$$

We will denote as $PD(\lambda_1,\lambda_2)$ the distribution with probability function given in (2.1). Although the Skellam's distribution was originally derived as the difference of two independent Poisson random variables, it can be also derived as the difference of distributions which have a specific trivariate latent variable structure.

Lemma: For any pair of variables (X,Y) that can be written as $X = W_1+W_3$, $Y = W_2+W_3$ with $W_1 \sim Poisson(\lambda_1)$, $W_2 \sim Poisson(\lambda_2)$ and W_3 follows any distribution with parameter vector $\boldsymbol{\theta}_3$ then $Z = X - Y$ follows a $PD(\lambda_1,\lambda_2)$ distribution.

The proof of the above Lemma is straightforward. It is however interesting that the joint distribution of X, Y is a bivariate distribution with correlation induced by the common stochastic component in both variables W_3. For example, if W_3 follows a Poisson distribution then the joint distribution is the bivariate Poisson which has been used for modelling scores (see Karlis and Ntzoufras, 2003). In addition, the marginal distributions for X and Y will be Poisson distributed only in the case where W_3 is degenerate at zero or a Poisson distributed random variable. Therefore, in the general formulation, the marginal distributions of X and Y are not any more Poisson but they are defined as the convolution of a Poisson random variable with another discrete random variable of any distributional form. Thus the marginal distribution can be overdispersed or even underdispersed relative to a Poisson distribution and, hence, a large portion of the distributional assumptions concerning the number of goals scored by each team is removed. This underlines the efficiency and the flexibility of our proposed model.

Although the above lemma implies that the type of the goal difference distribution will be the same regardless the existence or the type of association between the two variables, this does not implies that the parameter estimates and their interpretation will be the same. Finally, the trivariate reduction scheme used to define the PD distribution provides a suitable data augmentation scheme that can be used efficiently for constructing the estimation algorithm (see section 2.2 for details).

The expected value of the $PD(\lambda_1, \lambda_2)$ distribution is given by $E(Z) = \lambda_1 - \lambda_2$ while the variance is $Var(Z) = \lambda_1 + \lambda_2$. Note that, for the range of mean values observed in football games, the distribution cannot be sufficiently approximated by the normal distribution and, hence, inference based on simple normal regression can be misleading. Additional properties of the distribution are described in Karlis and Ntzoufras (2006).

2.2 A model for the goal difference.

We can use the Poisson difference distribution to model goal difference by specifying as the response variable the goal difference in game i. Hence we specify

$$Z_i = X_i - Y_i \sim PD(\lambda_{1i}, \lambda_{2i})$$

for $i = 1, 2, \ldots, n$; where n is the number of games, X_i and Y_i are the number of goals scored by the home and away team respectively in i game. Concerning the model parameters λ_{1i}, λ_{2i}, we adopt the same structure as in simple or bivariate Poisson models used for the number of goals scored by each team (see Lee, 1997, Karlis and Ntzoufras, 2003 respectively). Therefore, we set

$$\log(\lambda_{1i}) = \mu + H + A_{HT_i} + D_{AT_i} \tag{2.2}$$

$$\log(\lambda_{2i}) = \mu + A_{AT_i} + D_{HT_i} \tag{2.3}$$

where μ is a constant parameter, H is the home effect, A_k and D_k are the 'net' attacking and defensive parameters of team k after removing correlation, HT_i and AT_i are the home and away team competing each other in game i.

Note that for parameters A_k and D_k we propose to use the sum to zero constraints in order to make the model identifiable. Therefore we need to impose the constraints

$$\sum_{k=1}^{K} A_k = 0 \quad \text{and} \quad \sum_{k=1}^{K} D_k = 0 \tag{2.4}$$

where K is the number of the different teams competing each other in the available dataset. Under the above parametrization all parameters have a straightforward interpretation since H is the expected goal difference in a game where two opponent teams have the same attacking and defensive skills, μ is a constant parameter corresponding to the PD parameter for the away team in the same case, while A_k and D_k can be interpreted as deviations of the 'net' attacking and defensive abilities from a team of moderate performance.

2.3 Zero-Inflated version of the model.

As we have already mentioned, the number of draws (and the corresponding probability) are under estimated by Poisson based models used for the number of goals scored by each team. For this reason, we can specify the zero inflated version of Skellam's distribution to model the possible excess of draws. Hence, we can define the zero inflated Poisson difference (ZPD) distribution as the one with probability function

$$f_{ZPD}(0|\ p, \lambda_1, \lambda_2) = \quad p \quad + \quad (1-p) f_{PD}(0|\ \lambda_1, \lambda_2) \tag{2.5}$$

$$f_{ZPD}(z|\ p, \lambda_1, \lambda_2) = \qquad\quad (1-p) f_{PD}(z|\ \lambda_1, \lambda_2), \quad \text{for} \quad z \in \mathcal{Z} \setminus \{0\}, \tag{2.6}$$

where $p \in (0,1)$ and $f_{PD}(z|\ \lambda_1, \lambda_2)$ is given by (2.1).

3 Bayesian Inference

3.1 The prior distributions

To fully specify a Bayesian model, we need to specify the prior distribution. When no information is available, we propose to use normal prior distributions for the parameters

of the PD model with mean equal to zero and large variance (e.g. 10^4) to express prior ignorance. For the mixing proportion p used in the zero inflated version of the proposed model, we propose a uniform distribution defined in the $(0, 1)$ interval. This set of priors was used in the analysis of the Premier league which follows.

Nevertheless, the Bayesian approach offers the ability to incorporate external information to our inference via our prior distribution. For example, when a last minute injury is reported or the weather conditions support one of the two competing teams, then this information can be easily used to specify our prior distributions. Also, prior elicitation techniques can be employed in order to produce a prior distribution by extracting information by experts on the topic such as sport analysts and bookmakers. In this case, more general prior distributions can be used. For example we can use normal prior distributions with small variance centered at a certain value for the parameters of the PD model and a Beta prior for the mixing proportion for the zero inflated model.

Finally, the Bayesian approach can be used sequentially by using the previous fixture posterior distribution as a prior distribution and update by this way much faster our model.

3.2 The posterior distributions

In the Bayesian approach, the inference is based on the posterior distribution of the model parameters $\boldsymbol{\theta}$. In the PD model we consider the parameter vector

$$\boldsymbol{\theta} = (\mu, H, A_2, \ldots, A_K, B_2, \ldots, B_K)$$

and we need to calculate the posterior distribution

$$f(\boldsymbol{\theta}|\boldsymbol{z}) = \frac{f_{PD}(\boldsymbol{z}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int f_{PD}(\boldsymbol{z}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

where $\boldsymbol{z}$ is a $n \times 1$ vector with the observed goal differences, $f(\boldsymbol{\theta})$ is the joint prior distribution which is here defined as the product of independent normal distributions and $f_{PD}(\boldsymbol{z}|\boldsymbol{\theta})$ is the model likelihood

$$f_{PD}(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{PD}(z_i|\lambda_{1i}, \lambda_{2i})$$

with $f_{PD}(z|\lambda_1, \lambda_2)$ given by (2.1) and $\lambda_{1i}, \lambda_{2i}$ by (2.2) and (2.3) respectively. The attacking and defensive abilities of the omitted team is simply calculated via the constraints (2.4) and therefore A_1 and B_1 will be substituted in the likelihood by

$$A_1 = -\sum_{k=2}^{K} A_k \quad \text{and} \quad B_1 = -\sum_{k=2}^{K} B_k \ .$$

Note that the approach is similar for the zero inflated version but, in this case, we have to additionally estimate the mixing proportion p. Hence the posterior is given by

$$f(\boldsymbol{\theta}, p|\boldsymbol{z}) = \frac{\prod_{i=1}^{n} f_{ZPD}(z_i|\lambda_{1i}, \lambda_{2i}, p) f(\boldsymbol{\theta}) f(p)}{\int\int \prod_{i=1}^{n} f_{ZPD}(z_i|\lambda_{1i}, \lambda_{2i}, p) f(\boldsymbol{\theta}) f(p) d\boldsymbol{\theta} dp}$$

where $f(p)$ is the prior of the mixing proportion p and $f_{ZPD}(z|\lambda_1, \lambda_2, p)$ is given by (2.5) and (2.6).

Inference concerning the components of the parameter vector $\boldsymbol{\theta}$ (and p) can be based on the posterior summaries of the marginal posterior distribution (mean, median, standard deviation and quantiles). The above posterior distribution is not analytically tractable. For this reason, we use Markov chain Monte Carlo (MCMC) algorithms to generate values from the posterior distribution and hence estimate the posterior distribution of interest and their corresponding measures of fit. In the next section, we provide brief details on how to implement MCMC in our proposed model.

3.3 The Markov chain Monte Carlo algorithm

Our approach is based on the sampling augmentation scheme proposed by Karlis and Ntzoufras (2006). Hence, a key element for constructing an MCMC algorithm for the proposed PD and ZPD models is to generate the w_{1i} and w_{2i} augmented data for PD model and additionally the latent binary indicators δ_i for the ZPD model. The first set, will be used to specify the observed data z_i as a difference of two Poisson distributed variables while the latter will be used in the ZPD model to identify from which component we get the observed difference z_i (i.e. from the PD component or from the inflated one).

Hence, in each iteration of the MCMC algorithm we

- Generate latent data w_{1i} and w_{2i} from

$$f(w_{1i}, w_{2i}|z_i = w_{1i} - w_{2i}, \lambda_{1i}, \lambda_{2i}) \propto \frac{\lambda_{1i}^{w_{1i}}}{w_{1i}!} \frac{\lambda_{2i}^{w_{2i}}}{w_{2i}!} I(z_i = w_{1i} - w_{2i})$$

  where $I(E) = 1$ if E is true and zero otherwise.

- Generate latent binary indicators δ_i from

$$f(\delta_i|z_i, \lambda_{1i}, \lambda_{2i}) \cong Bernoulli(\tilde{p}_i) \text{ with } \tilde{p}_i = \frac{p}{p + (1-p) f_{PD}(z_i|\ \lambda_{1i}, \lambda_{2i})} .$$

Concerning the simulation of the augmented data (w_{1i}, w_{2i}) used in the PD we propose to use the following Metropolis Hastings step:

- If $z_i < 0$ and (w_{1i}, w_{2i}) the current values of the augmented data then
  - Propose $w'_{1i} \sim Poisson(\lambda_{1i})$ and $w'_{2i} = w'_{1i} - z_i$.
  - Accept the proposed move with probability $\alpha = \min\left\{1, \lambda_{2i}^{(w'_{1i} - w_{1i})} \frac{(w_{1i} - z_i)!}{(w'_{1i} - z_i)!}\right\}$.
- If $z_i \geq 0$ and (w_{1i}, w_{2i}) the current values of the augmented data then
  - Propose $w'_{2i} \sim Poisson(\lambda_{2i})$ and $w'_{1i} = w'_{2i} + z_i$.
  - Accept the proposed move with probability $\alpha = \min\left\{1, \lambda_{1i}^{(w'_{2i} - w_{2i})} \frac{(w_{2i} + z_i)!}{(w'_{2i} + z_i)!}\right\}$.

Given the augmented data, $(w_{1i}, w_{2i}, \delta_i)$ the parameters $\boldsymbol{\theta}$ can be generated as in simple Poisson log-models with data $\boldsymbol{y} = (\boldsymbol{w}_1^{PD}, \boldsymbol{w}_2^{PD})$; where $\boldsymbol{w}_1^{PD}$, $\boldsymbol{w}_2^{PD}$ are vectors with elements the w_{1i} and w_{2i} for which $\delta_i = 0$. The conditional posterior distributions will be given by

$$f(\boldsymbol{\theta}|p, \boldsymbol{\delta}, \boldsymbol{w}_1, \boldsymbol{w}_2) \propto \prod_{i=1}^{n} \left[f_P(w_{1i}|\lambda_{1i}) f_P(w_{2i}|\lambda_{2i})\right]^{1-\delta_i} f(\boldsymbol{\theta})$$

and

$$f(p|\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{w}_1, \boldsymbol{w}_2) \propto p^{\sum_{i=1}^n \delta_i} (1-p)^{n - \sum_{i=1}^n \delta_i} f(p).$$

Note that in the PD model, is similar to setting all $\delta_i = 0$ for all observations (and $p = 0$ respectively). In the case, that we use a Beta prior distribution with parameters a and b

for p then the above conditional posterior will be also beta with parameters $\sum_{i=1}^{n}\delta_i + a$ and $n - \sum_{i=1}^{n}\delta_i + b$. When we wish to impose additional covariates on the mixing proportion then the parameters can be generated as in the case of a simple logistic regression model having as a response the latent binary indicators $\boldsymbol{\delta}$.

The above algorithm can be implemented in any programming language or more statistical friendly programming software (such as `R` and `Matlab`). Alternatively we can directly use `WinBUGS` (Spiegelhaler et al., 2003), a statistical tool for the implementation of Bayesian models using MCMC methodology. Results presented in this article, have been reproduced using both `R` and `WinBUGS`. The latter is available by the authors upon request.

3.4 Simulating future games and leagues from the predictive distribution

An important feature of Bayesian inference is the predictive distribution. Consider a future game between the home team h and away team a. Hence we wish to predict a future goal difference $z_{(h,a)}^{pred}$. This can be done directly using the posterior predictive distribution

$$f\big(z_{(h,a)}^{pred}\big|\boldsymbol{z}\big) = \int f\big(z_{(h,a)}^{pred}\big|\boldsymbol{\theta}\big) f(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta}. \tag{3.7}$$

Note that in the ZPD model $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}' = (\boldsymbol{\theta}, p)$. Moreover, $f\big(z_{(h,a)}^{pred}\big|\boldsymbol{\theta}\big)$ depends only on parameters μ, H, (A_h, D_h) and (A_a, D_a) (and p in the ZPD model) which are related to the teams competing each other in the game we wish to predict. When we wish to predict goal differences $\boldsymbol{z}^{pred}$ for $n^{pred} > 1$ games in which the home team HT_k^{pred} competes with the away team AT_k^{pred} in k-th game (for $k = 1, \ldots, n^{pred}$), then the resulting posterior predictive distribution is given by

$$f\big(\boldsymbol{z}^{pred}\big|\boldsymbol{HT}^{pred}, \boldsymbol{AT}^{pred}, \boldsymbol{z}\big) = \int f\big(\boldsymbol{z}^{pred}\big|\boldsymbol{HT}^{pred}, \boldsymbol{AT}^{pred}, \boldsymbol{\theta}\big) f(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta}$$

$$= \int \prod_{k=1}^{n^{pred}} f(z_k^{pred}|\mu, H, A_{HT_k^{pred}}, D_{HT_k^{pred}}, A_{AT_k^{pred}}, D_{AT_k^{pred}}) f(\mu, H, A_2, \ldots, A_K, D_2, \ldots, D_K|\boldsymbol{z}) d\boldsymbol{\theta}.$$

where $\boldsymbol{HT}^{pred}$ and $\boldsymbol{AT}^{pred}$ are the vectors of length n^{pred} with the competing teams in the future games we wish to predict.

When using an MCMC algorithm, it is straight forward to generate values of the $z_{(h,a)}^{pred}$ from the corresponding predictive distribution (3.7) by simply adding the following steps in the MCMC sampler used in the PD model

- Calculate $\lambda_1^{pred} = \mu + A_h + D_a + H$ and $\lambda_2^{pred} = \mu + A_a + D_h$
- Generate w_1^{pred} and w_2^{pred} from a Poisson distribution with parameters λ_1^{pred} and λ_2^{pred} respectively.
- Set $z_{(h,a)}^{pred} = w_1^{pred} - w_2^{pred}$ as the generated value from the predictive distribution of interest.

In the above procedure, parameters μ, H, (A_h, D_h) and (A_a, D_a) will be equal to their corresponding values generated in each iteration of the MCMC algorithm.

For the ZPD model we need add the following steps in our sampling algorithm

- Generate δ^{pred} from a Bernoulli with probability p.
- If $\delta^{pred} = 1$ then set $z_{(h,a)}^{pred} = 0$ otherwise proceed as in the PD model.

A usual practice when modelling sports outcomes is to reproduce the predictive distribution of the ranking table in a league or a tournament. This procedure is very useful and was initially introduced by Lee (1997) using plug-in maximum likelihood estimates from a simple Poisson model. It can be used for two reasons. Firstly, to probabilistically quantify the final outcome of a league or a tournament. This can be used to assess the goodness of fit of the model and the overall performance of specific teams in the league. For example, if the ranking resulting from the predictive distribution is in general agreement with the observed data, this implies a good fit of our model. On the other hand, specific deviations may indicate that specific teams performed better or worse than expected in specific games and hence ended up in a different ranking. Secondly, it can be used to estimate the ranking

distribution if the competition had a different structure. For example, when we use data from knock-out tournaments to estimate the rankings if the teams were competing in full season in round robin (league) system. Of course this procedure, with the current model, assumes that the teams' performance is constant across the whole competition.

In order to generate the posterior predictive distribution of each league, in each iteration of the algorithm we need to generate $\boldsymbol{z}^{pred}$. In the case of a foul season league, $\boldsymbol{z}^{pred} = \boldsymbol{z}$, $HT_i^{pred} = HT_i$ and $AT_i^{pred} = AT_i$ for all $i = 1, 2, \ldots, n$. Having $\boldsymbol{z}^{pred}$ generated within a single iteration of the MCMC algorithm, we further

- Calculate the points P_k^{pred} of each team for $k = 1, 2, \ldots, K$ (usually giving three points for a win, one point for a draw and zero points for a loss in each team per game).
- Calculate the ranking R_k^{pred} in descending order (giving one to the team with highest number of points) for each team.

After completing the procedure we end up with posterior samples for the points gained by each team as well as for the rankings in the final league table. We can directly estimate the number of points that each team was a-posteriori expected to earn (using simple means of the sampled P_k values), and the probability distribution of the ranking for each team (from simple frequency tables of R_k's).

4 Application: The English Premier League 2006-2007

The data refer to the English Premierhsip for 2006-2007 season. Data were downloaded by the web page `http://soccernet-akamai.espn.go.com`.

For all the parameters of the PD component we have used normal prior distributions with zero mean and low precision equal to 10^{-4} (i.e. large variance equal to 10^4) to express our prior ignorance. A uniform prior distribution was used for the mixing proportion p of the ZPD model. All results were produced using 10,000 iterations after discarding additional 1,000 iterations as a burn-in period.

A plot of the 95% posterior intervals for all 'net' attacking and defensive parameters for all teams is provided in Figures 1 and 2 respectively. According to this plot, Manchester United, Liverpool, Arsenal, Blackburn and Chelsea had the highest 'net' attacking parameters. Concerning the 'net' defensive parameters, Chelsea had a considerably higher parameter than the rest of the teams followed by Manchester United and Arsenal. If we look at the actual goals scored and conceded by each team, we see that Chelsea and Arsenal had scored more goals than Liverpool. Nevertheless, Liverpool 'attacking' parameter is much higher than the corresponding parameters of the other teams. This is due to the properties of the Poisson difference distribution. Since focus is given in the differences, 'net' attacking and defensive parameter must be interpreted altogether. Hence, we can see that Chelsea overall has much better performance since its 'net' defensive parameters are much higher than Liverpool. In Figure 3 we provide the distribution of goals in favour of Liverpool and Chelsea for comparison reasons. In this Figure we observe that Liverpool has earned more games with differences higher than one goal margin (right part of the distribution), while it suffers considerably in games with differences of one goal, in draws and losses. These differences resulted to a better 'net' attacking parameters but much worse 'net' defensive parameters for Liverpool than Chelsea.

Figure 4 depicts the observed and predicted relative frequencies for the goal differences. It is evident that there is there is a close agreement between the observed and the predicted values. Concerning the number of draws, the PD models slightly over-estimates them which comes in contradiction with the results observed when using simple or bivariate Poisson models for the actual number of goals. However, note that the marginal distributions for our model are not necessarily Poisson and this can explain why our model fits better the draws. Nevertheless, the 95% posterior predictive interval contains the observed value, indicating minor deviations from the data. Since the PD model slightly overestimates draws, the ZPD model will not offer much in terms of goodness of fit. Indeed, fitting the ZPD model resulted to similar predictive values with no obvious differences.

Furthermore, we have reproduced the predictive distribution of the final table using the

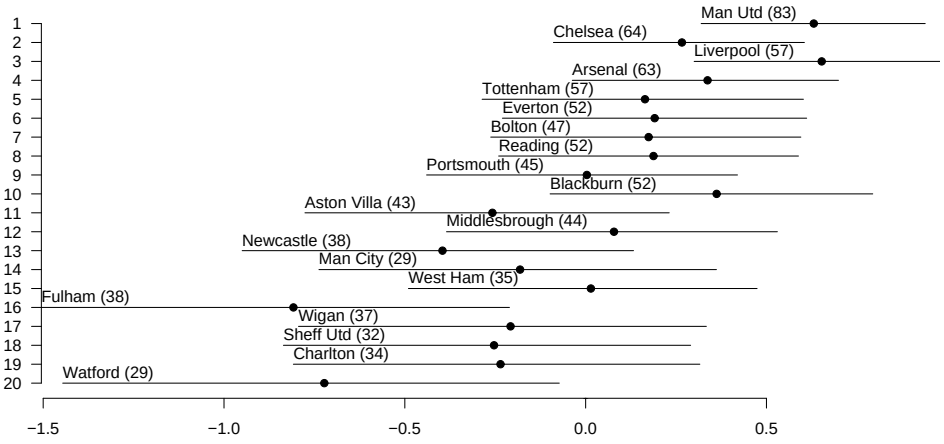

Figure 1: 95% Posterior intervals for 'net' attacking coefficients (A_i)

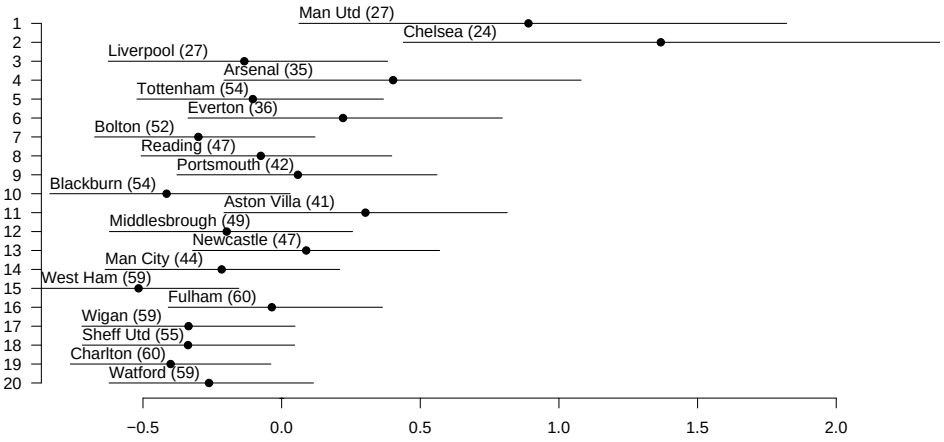

Figure 2: 95% Posterior intervals for 'net' defensive coefficients $[(-1) \times D_i]$; within brackets the observed scored goals.

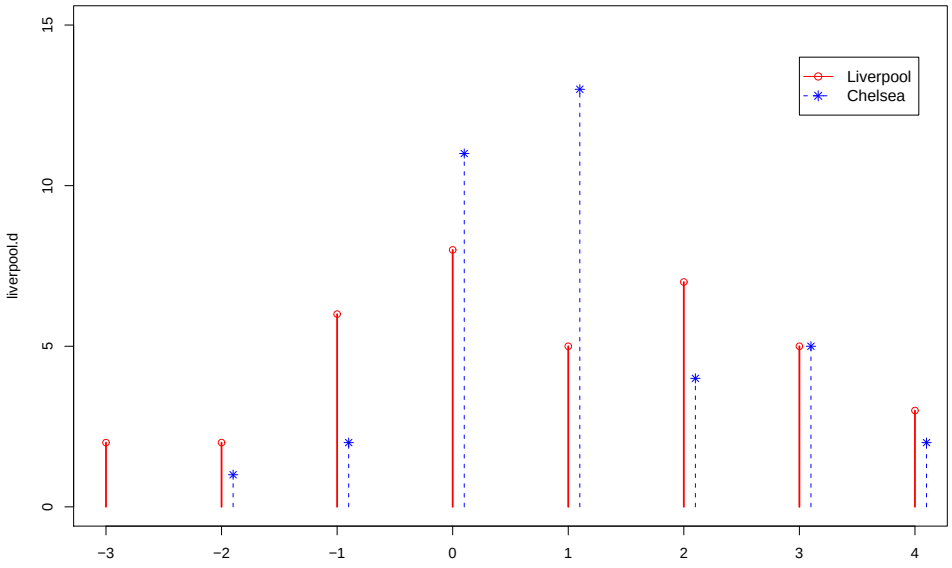

Figure 3: Distribution of goal differences for Liverpool and Chelsea; within brackets the observed conceded goals.

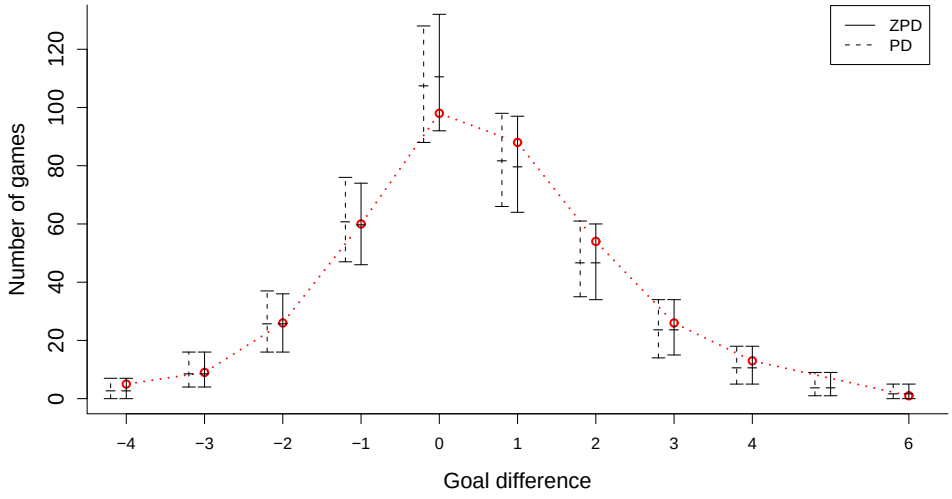

Figure 4: 95% Posterior intervals for predicted differences (in red=observed differences; '-' indicates the posterior median) No major differences between the two models are observed.

procedure described in section 3.4. Results are presented in Table 1 including details from the observed final Table. Minor differences were observed between expected and observed points. Predicted rankings were calculated via the expected number of points. For 60% and 80% of the competing teams, the absolute difference between predicted and observed points were found to be ≤ 2 and ≤ 3 points respectively. Only for 4 teams the corresponding difference was found between 3-7 points. Predictive and observed rankings are also close since, for 80% of the teams, the final observed and predicted ranking was the same or changed by only one position.

The largest differences are related to the performance of Bolton and Fulham. Bolton managed to earn seven (7) points additional to the ones expected (56 instead of 49) with final ranking 7 instead of the expected 12 while Fulham earned six (6) points additional to the ones expected (39 instead of 33.1) getting 16th position instead of the predicted 19th avoiding by this way the relegation to the next division.

In order to examine the behaviour of these two teams and compare them with the pre-

Posterior Predictive Table				Observed Final Table			
Pred.(Obs.)		Post. Expectations		Obs.		Obs. values	
Rank	Team	Pts	G.Dif.	Rank	Team	Pts	G.Dif.
1	(1) Man Utd	86.7	56.0	1	Man Utd	89	56
2	(2) Chelsea	81.0	40.0	2	Chelsea	83	40
3	(3) Arsenal	70.5	28.0	3	Liverpool	68	30
4	(3) Liverpool	69.4	30.2	4	Arsenal	68	28
5	(6) Everton	62.5	16.0	5	Tottenham	60	3
6	(8) Reading	55.5	5.4	6	Everton	58	16
7	(5) Tottenham	54.0	3.2	7	Bolton	56	-5
8	(9) Portsmouth	53.3	2.8	8	Reading	55	5
9	(10) Blackburn	51.8	-2.1	9	Portsmouth	54	3
10	(11) Aston Villa	51.5	1.6	10	Blackburn	52	-2
11	(12) Middlesbrough	49.0	-4.7	11	Aston Villa	50	2
12	(7) Bolton	49.0	-5.7	12	Middlesbrough	46	-5
13	(13) Newcastle	43.8	-8.8	13	Newcastle	43	-9
14	(14) Man City	41.8	-14.8	14	Man City	42	-15
15	(15) West Ham	38.6	-24.3	15	West Ham	41	-24
16	(17) Wigan	38.1	-21.6	16	Fulham	39	-22
17	(18) Sheff Utd	37.0	-23.0	17	Wigan	38	-22
18	(19) Charlton	35.7	-25.9	18	Sheff Utd	38	-23
19	(16) Fulham	33.1	-22.0	19	Charlton	34	-26
20	(20) Watford	29.7	-30.3	20	Watford	28	-30

Table 1: Observed and predicted under the model points and goals differences for all the teams; in the second column (within brackets) the actual ranking is provided.

dicted ones, we have calculated their outcome probabilities for each game and the points expected to be earned in each game. We have traced as outliers (surprising results according to our model), all games with absolute difference between the expected and observed number of points greater than 1.95 (first criterion) or games with probability of observed outcome lower that 20% (second criterion). Six and four games were traced as surprising results (or outliers) for Bolton and Fulham respectively and are presented in Table 2.

Concerning Bolton, we notice that it earned 32 points instead of the expected 30.6 in their home field (+1.4 points) and 24 instead of the expected 18.3 in their away games (+5.7 points). Hence, Bolton performance in away games was much higher than expected. Specifically, in their home field won Arsenal where the expected number of points was equal to one (and probability of loosing equal to 47%) but they lost by Wigan in which game the expected number of points were equal to 2.1 and the probability of loosing only 16% (probability of winning this game 61%). All four away games with surprising scores are in favour of Bolton. They surprisingly won Aston Villa, Blackburn and Portsmouth (with probabilities ranging from 20% to 24%) while they managed to draw with Chelsea (with probability of draw equal to 19%). The additional point difference in these four games is equal to seven (7) points which is the difference between the predicted and observed league tables.

Looking at Fullham's games, we observe four games with highly surprisingly results. Fulham, unlike Bolton, had much better performance than the expected one in home games. It managed to earn 28 points instead of the expected 20.6 (+7.4) in home games while it earned 11 points instead of the expected 12.6 (-1.6) in the away games. From the presented results, we see that Fulham managed to win Arsenal, Everton and Liverpool in its home field. All these three teams are much better in terms of performance and budget and the probability for Fulham winning these games was lower than 16%. Finally, Fulham managed to unexpectedly win Newcastle in the corresponding away game (probability of winning equal to 13%).

The same type of analysis can be performed for all teams. We have simply focused on

			Goal	Probabilities					
Home	Away	Final difference		Home		Away	Observed	Expected	Point
Team	Team	Score	(z_i)	Win	Draw	Win	Points	Points	Difference
Bolton	Arsenal	3-1	2	0.25	0.29	0.47	3	1.0	2.0
Bolton	Wigan	0-1	-1	0.61	0.23	0.16	0	2.1	-2.1
Aston Villa	Bolton	0-1	-1	0.48	0.32	0.20	3	0.9	2.1
Blackburn	Bolton	0-1	-1	0.56	0.21	0.24	3	0.9	2.1
Portsmouth	Bolton	0-1	-1	0.53	0.27	0.20	3	0.9	2.1
Chelsea	Bolton	2-2	0	0.77	0.19	0.04	1	0.3	0.7
Fulham	Arsenal	2-1	1	0.13	0.36	0.51	3	0.7	2.4
Fulham	Everton	1-0	1	0.16	0.39	0.46	3	0.9	2.1
Fulham	Liverpool	1-0	1	0.16	0.28	0.56	3	0.7	2.3
Newcastle	Fulham	1-2	-1	0.44	0.43	0.13	3	0.8	2.2

Table 2: Surprising Results for Bolton and Fulham; Games with expected absolute point difference > 1.95 or probability of observed outcome < 0.20.

these two teams in order to understand why these two teams performed in a way different than the one expected by the model.

Finally, Table 3 presents some goodness of fit measures based on the predictive distribution. Namely, for each of the quantity appearing in the table, denoted in a general form as Q, we have calculated the deviation given by

$$Deviation = \sqrt{\frac{1}{|\boldsymbol{Q}|}\sum_{i=1}^{|\boldsymbol{Q}|}\left(E(Q_i^{Pred}|\boldsymbol{y}) - Q_i^{obs}\right)^2},$$

where $|\boldsymbol{Q}|$ is the length of vector $\boldsymbol{Q}$. For the calculation of the deviations of the frequencies and the relative we used $|\boldsymbol{Q}| = 13$ to consider for differences from -6 to 6, while for the deviations of the expected points and the expected differences from the final table we set

	Deviation	
Comparison	PD	ZPD
1. Relative Frequencies (counts/games)	1.06%	1.32%
2. Frequencies (counts)	4.04	5.04
3. Relative Frequencies of win/draw/lose	2.20%	2.80%
4. Frequencies of win/draw/lose	8.30	10.65%
5. Expected points	3.02	3.07
6. Expected goal difference	0.28	0.40

Table 3: Deviations between observed and predictive measures.

$|\boldsymbol{Q}| = K = 20$ i.e. the number of teams in the league. So Q_i^{obs} is the observed quantity and $E(Q_i^{Pred}|\boldsymbol{y})$ the predicted quantity; the quantities used are reported in the first column of Table 3. The results show a satisfactory fit of the model to the actual data. In addition the zero-inflated model does not seem to improve the fit of the model for this data since we did not observed any excess of draws.

5 Concluding Remarks

In the present paper, we have proposed an innovative approach for modelling football data. The proposed model has some interesting advantages over the existing ones used for the same purpose. It is based on the goal differences in each game and its main feature is that it accounts for the correlation by eliminating any additive covariance. For this reason, we avoid modelling correlation or imposing assumptions about its structure as needed in any bivariate distribution (as for example in the bivariate Poisson model). The model has a straightforward Poisson latent variable interpretation although we do not need to make assumptions about the distributions of the actual goals scored by each team. Therefore, the proposed model is quite generic and applicable to data from a wide range of football

leagues in which teams with different behaviour compete. Furthermore, its parameters have a relatively easy interpretation while the parameter estimation is easier than the corresponding one for the bivariate Poisson model. The Poisson difference model is appropriate for bets like the Asian handicap where only the difference between the two teams is considered. On the other hand, by considering only the goal differences, it discards part of the available data and information and, hence, it cannot be used for modelling the final score of a game.

We have also considered a possible extension of the model by considering a zero-inflated component. Although, draws were under-estimated using other Poisson related models, for the 2006-7 English Premier league data the draws are slightly over-estimated using the proposed PD model. Therefore no zero inflation was needed.

We are currently working on further extensions of the proposed model. A first direction for extending the proposed model is to consider other distributions defined on $\mathcal{Z}$ (see for examples in Ong et al, 2007). Secondly, variable selection techniques may be implemented to identify variables with good predictive power and by this way construct a more precise but also parsimonious model. In addition, Bayesian model averaging techniques can be used to improve the predictive power. As far as the zero-inflated version, when excess of draws is observed, we may also incorporate covariates in the mixing proportion p in order to predict more precisely draws and further identify additional factors that increase the probability of a draw in a football game.

References

Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in football betting market. *Applied Statistics*, **46**, 265-280.

Irwin, W. (1937). The frequency distribution of the difference between two Poisson variates following the same poisson distribution. *Journal of the Royal Statistical Society, Series A*, **100**, 415-416.

Karlis D. and Ntzoufras, I. (2006). Bayesian analysis of the differences of count data.

Statistics in Medicine, **25**, 1885-1905.

Karlis, D. and Ntzoufras, I. (2005). Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R. *Journal of Statistical Software*, Volume **10**, Issue 10.

Karlis, D. and Ntzoufras, I. (2003). Analysis of Sports Data Using Bivariate Poisson Models. *Journal of the Royal Statistical Society*, D, (Statistician), **52**, 381 – 393.

Lee, A.J. (1997). Modeling scores in the Premier League: Is Manchester United really the best? *Chance*, **10**, 15-19.

Maher, M.J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109–118.

Ong, S.H., Shimizu, K. and Ng, C.M (2007). A Class of Discrete Distributions Arising from Difference of Two Random Variables, *Computational Statistics and Data Analysis* (to appear)

Skellam, J.G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society, Series A*, , **109**, 296.

Spiegelhalter, D. and Thomas, A. and Best, N. and Lunn, D. (2003). *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine, UK, available at `http://www.mrc-bsu.cam.ac.uk/bugs`.