An application of the generalised Poisson difference distribution to the Bayesian modelling of football scores

Golnaz Shahtahmassebi[a]*and Rana Moyeed[b]

[a] *Nottingham Trent University, Nottingham, UK, NG11 8NS*

[b] *Plymouth University, Plymouth, UK, PL4 8AA*

Abstract

The analysis of sports data, in particular football match outcomes, has always produced an immense interest among the statisticians. In this paper, we adopt the generalised Poisson difference distribution (GPDD) to model the goal difference of football matches. We discuss the advantages of the proposed model over the Poisson difference (PD) model which was also used for the same purpose. The GPDD model, like the PD model, is based on the goal difference in each game which allows us to account for the correlation without explicitly modelling it. The main advantage of the GPDD model is its flexibility in the tails by considering shorter as well as longer tails than the PD distribution. We carry out the analysis in a Bayesian framework in order to incorporate external information, such as historical knowledge or data, through the prior distributions. We model both the mean and the variance of the goal difference and show that such a model performs considerably better than a model with a fixed variance. Finally, the proposed model is fitted to the 2012-13 Italian Serie A football data and various model diagnostics are carried out to evaluate the performance of the model.

Keywords: Bayesian methods; Generalised Poisson difference distribution; goal difference; football data

1 Introduction

The statistical analysis of sports data, in particular football match outcomes, generates an enormous interest worldwide. Betting on the outcome of football matches has a long tradition. In addition, the football industry now attracts huge investments. As a result, the modelling and prediction of football match outcomes is becoming ever more popular (Karlis and Ntzoufras, 2009; Cattelan et al., 2013).

The outcome of football matches has been analysed from various perspectives such as modelling loss/draw/win (Cattelan et al., 2013; Koning, 2000), number of goals (Karlis and Ntzoufras, 2003; Maher, 1982; Dixon and Coles, 1997) and goal differences (Karlis and Ntzoufras,

*Corresponding author, Email: golnaz.shahtahmassebi@ntu.ac.uk

2009; Clarke and Norman, 1995). Among these outcomes, goal difference has attracted special attention because it eliminates the correlation between the teams and also we do not need to assume that the number of goals scored by each team is marginally Poisson distributed (Karlis and Ntzoufras, 2009).

Initially Clarke and Norman (1995) used the normal distribution to model the goal differences. However, since goal difference may take on a small range of integers, a normal approximation may not be valid (Karlis and Ntzoufras, 2006; Shahtahmassebi and Moyeed, 2014). Thus, techniques that are based on distributions defined over both negative and positive integer values may improve our inference of the goal difference. Karlis and Ntzoufras (2009) proposed the application of the Poisson difference (PD) distribution as an alternative to the normal approximation. The major drawback of the their model was the overestimation of number of draws. In addition, in double-round-robin structure tournaments, it can be seen that the presence of the home advantage effect results in a distribution with one or both tails being shorter or longer than the PD distribution.

To overcome these limitations of the PD distribution, we use the generalised Poisson difference distribution (GPDD) (Shahtahmassebi and Moyeed, 2014). In this way, we still remove the effect of the correlation between the scoring performance of the two competing teams, but the proposed model introduces more flexibility in the tails. Similar to the PD model, the GPDD model can be used to predict the outcome of the game, but cannot predict the final score.

We fit the model in a Bayesian framework, thus incorporating any available information about each game via the prior distribution. Information that can be incorporated into the model can be based on, e.g. historical knowledge or data, weather conditions and the fitness of a team. Finally, the Bayesian approach allows for prediction of match outcomes via the posterior predictive distribution. It can also produce quantitative measures relating to the performance of each team. We carried out all our analyses in R 3.0.1 (R Core Team, 2013). The R functions for evaluating and generating samples from the GPDD were written by the authors.

The remainder of the paper is organised as follows. In Section 2, we introduce the GPDD and obtain its properties. Section 3 describes the GPDD model proposed for the purpose of modelling goal differences. We describe the Bayesian approach in Section 4. A real data application based on the match outcomes of the Italian League Serie A for the season 2012−2013 is described in Section 5. Finally, we conclude and provide an overview of possible extensions in Section 6.

2 Generalised Poisson difference distribution

Let us assume X and Y are two non-negative integer random variables. The random variable $Z = X - Y$ follows a GPDD (Shahtahmassebi and Moyeed, 2014) if its probability density function has the following form

$$f_{\text{GPDD}}(Z = X - Y = z) = \; e^{-\lambda_1 - \lambda_2 - \theta_1 z} \sum_{y=0}^{\infty} (\lambda_1, \theta_1)_{z+y} \; (\lambda_2, \theta_2)_y \; e^{-(\theta_1 + \theta_2) y}, \tag{1}$$

for any value of $z \in \mathbb{Z}$, where

$$(\lambda, \theta)_x = \frac{\lambda(\lambda + x\theta)^{x-1}}{x!}.$$

Lower limits for θ_1 and θ_2 have been set to ensure that there are at least five classes of non-zero probabilities at both tails when $\theta_1 < 0$ or $\theta_2 < 0$:

$$\begin{aligned} \max(-1, -\lambda_1/m_1) < \theta_1 < 1, \\ \max(-1, -\lambda_2/m_2) < \theta_2 < 1, \end{aligned} \tag{2}$$

where $m_1,\ m_2 \geq 4$ are the largest positive integers in which $\lambda_1 + m_1\theta_1 > 0$ and $\lambda_2 + m_2\theta_2 > 0$. Therefore, for any $z > m_1$ when $\theta_1 < 0$, or $z < -m_2$ when $\theta_2 < 0$,

$$f_{\text{GPDD}}(z|\lambda_1, \lambda_2, \theta_1, \theta_2) = 0.$$

Furthermore, parameters λ_1 and θ_1 refer to the positive half and parameters λ_2 and θ_2 to the negative half of the GPDD. The GPDD$(\lambda_1, \lambda_2, \theta_1, \theta_2)$ reduces to the PD distribution (Skellam, 1946) when $\theta_1 = \theta_2 = 0$:

$$\text{GPDD}(\lambda_1, \lambda_2, \theta_1 = 0, \theta_2 = 0) = \text{PD}(\lambda_1, \lambda_2) = e^{-\lambda_1 - \lambda_2}(\frac{\lambda_1}{\lambda_2})^{z/2}\ \text{I}_{|z|}(2\sqrt{\lambda_1\lambda_2}),$$

where $\text{I}_r(x)$ is the modified Bessel function of order r Abramowitz and Stegun (1964). It may be noted that the GPDD of Consul (1986) is a special case of our GPDD when $\theta_1 = \theta_2$.

The cumulants of the probability distribution of the random variable Z can be derived using the following recurrence relation (see Shahtahmassebi and Moyeed (2014))

$$\begin{aligned} (1-\theta_1)(1-\theta_2)L_{k+1} =& (2-\theta_2)\lambda_1\theta_1\frac{\partial^2 L_k}{\partial\theta_1\partial\lambda_1} + (2-\theta_2)\lambda_1\frac{\partial L_k}{\partial\lambda_1} + \theta_1\theta_2\lambda_2\frac{\partial^2 L_k}{\partial\theta_2\partial\lambda_2} \\ &+ \theta_1\lambda_2\frac{\partial L_k}{\partial\lambda_2} - \theta_1\frac{\partial L_k}{\partial\theta_1} - \theta_2\frac{\partial L_k}{\partial\theta_2} - L_k, \end{aligned} \tag{3}$$

and the expression for the first four cumulants of the GPDD are obtained as follows

$$L_1 = \frac{\lambda_1}{(1-\theta_1)} - \frac{\lambda_2}{(1-\theta_2)}, \tag{4}$$

$$L_2 = \frac{\lambda_1}{(1-\theta_1)^3} + \frac{\lambda_2}{(1-\theta_2)^3}, \tag{5}$$

$$L_3 = \frac{\lambda_1\ (2\theta_1+1)}{(1-\theta_1)^5} - \frac{\lambda_2\ (2\theta_2+1)}{(1-\theta_2)^5}, \tag{6}$$

$$L_4 = \frac{\lambda_1\ (1+8\theta_1+6\theta_1^2)}{(1-\theta_1)^7} + \frac{\lambda_2\ (1+8\theta_2+6\theta_2^2)}{(1-\theta_2)^7}. \tag{7}$$

where $\mu = L_1$ and $L_2 = \sigma^2$ are the mean and the variance of the GPDD.

Similar to the generalised Poisson distribution, in a GPDD parameters θ_1 and θ_2 add more flexibility to the left and right tails. That is when $\theta_1 > 0$ and/or $\theta_2 > 0$, the GPDD will have a longer left and/or right tail than a standard PD distribution. On the other hand when $\theta_1 < 0$ and/or $\theta_2 < 0$ the GPDD will have a shorter left and/or right tail than a standard PD distribution, a property that had not been addressed by many alternative distributions to the Poisson difference distribution (for further details see Shahtahmassebi and Moyeed (2014)).

3 A model for the goal differences

The description of the proposed methodology for the analysis of a tournament is simplified by the assumption of an order for n matches among K teams that are involved in the tournament. A convenient choice is to arrange the matches in chronological order, with games played at the same time in alphabetical order of the home team. Let X_i and Y_i be the number of goals scored by the home and away team in the ith game, respectively and

$$Z_i = X_i - Y_i$$

be the goal difference of the ith match, $i = 1, \ldots, n$ played by the home team h_i against the visiting team v_i, with $h_i, v_i = 1, \ldots, K$ and $h_i \neq v_i$. We adopt a similar model set-up to the traditional paired comparison models (Cattelan et al., 2013) in which the outcome of interest can be described as a function of the difference in team abilities.

Let us assume the goal difference (an integer-valued variable) follows the GPDD with the probability mass function given in (1). Setting the goal difference as the outcome of interest, we model the mean and the variance of our GPDD model as follows

$$\mathsf{E}(Z_i) = \mu_i = H + a_{h_i} - a_{v_i} \tag{8}$$

$$\mathsf{Var}(Z_i) = \sigma_i^2 = \gamma_1 + |a_{h_i} - a_{v_i}| \tag{9}$$

where a_{h_i} and a_{v_i} are the parameters representing the abilities of the home and visiting teams in match i, H is a common home effect parameter for all teams and γ_1 is a positive constant for the variance. It can be seen that in this set-up the mean of the GPDD model is directly affected by the home effect as well as home and visiting team abilities. The variance is defined in a way that a larger difference between home and visiting team abilities results in a larger variance. Furthermore, θ_1 and θ_2 are assumed to be constant with respect to team abilities given by the following form

$$\log(1 - \theta_1) = \gamma_2, \tag{10}$$

$$\log(1 - \theta_2) = \gamma_3, \tag{11}$$

which allows us to have parameters over the real line. For the sake of parameter identifiability,

we require one constraint in the set of abilities, e.g. the sum constraint $\sum_{k=1}^{K} a_k = 0$. In the prediction step, the parameters of λ_1 and λ_2 can be obtained by rearranging (4) and (5) as follows

$$\begin{aligned}\lambda_{1,i} &= \frac{\left[(1-\theta_2)^2\,\sigma_i^2 + \mu_i\right](1-\theta_1)^3}{(1-\theta_1)^2 + (1-\theta_2)^2} \\ &= \frac{\left[\exp(\gamma_3)^2\,(\gamma_1 + |a_{h_i} - a_{v_i}|) + (H + a_{h_i} - a_{v_i})\right]\exp(\gamma_2)^3}{\exp(\gamma_2)^2 + \exp(\gamma_3)^2} \qquad (12)\end{aligned}$$

$$\begin{aligned}\lambda_{2,i} &= \frac{\left[(1-\theta_1)^2\,\sigma_i^2 - \mu_i\right](1-\theta_2)^3}{(1-\theta_1)^2 + (1-\theta_2)^2} \\ &= \frac{\left[\exp(\gamma_2)^2\,(\gamma_1 + |a_{h_i} - a_{v_i}|) - (H + a_{h_i} - a_{v_i})\right]\exp(\gamma_3)^3}{\exp(\gamma_2)^2 + \exp(\gamma_3)^2}. \qquad (13)\end{aligned}$$

4 Bayesian inference

4.1 Prior distribution

In the Bayesian framework, we need to specify the prior distributions. Here we include information from the previous year's tournament as prior information in our analysis. The previous season can be considered as a baseline for the following season's results as the teams generally at least intend to keep their current position in the table. Thus, it is realistic to consider these as prior information. Thus, the prior distributions of home effect and team abilities are as follows,

$$\begin{aligned}&H \sim N(\bar{z}, \sigma_z^2),\\ &a_k \sim N(\bar{z}_k, \sigma_{z_k}^2), \quad k = 1, \ldots, K,\end{aligned}$$

where $\bar{z}$ and σ_z^2 are the mean and the variance of Z_i, while $\bar{z}_k$ and $\sigma_{z_k}^2$ are the mean and the variance of the goal difference of team k in the previous year (Appendix A). For teams who are promoted and playing for the first time, we assign a non-informative normal prior distribution with mean zero and a large variance (e.g. 10^4) because by definition team abilities are measured relative to other teams in the same league. That is a top team in one league does not have the same ability as a top team or, for that matter, any other team in another league. Usually, the bottom three teams are relegated from a higher league and replaced by three teams coming up from the lower league to play in the following year's tournament. Therefore, we can say that the promoted teams start the new league with no prior information.
Finally, a non-informative gamma distribution is considered for γ_1 and truncated normal prior distributions, $\mathrm{N}(0, 10^4)$, are assigned to parameters $\gamma_2, \gamma_3 < \log 2$ (which we denote by $N_+(\mu, \sigma; \log(2))$), in order to satisfy the conditions in (2).

4.2 Posterior distribution

We obtain our inference based on the posterior distribution of the parameter vector

$$\boldsymbol{\psi} = (H, a_1, \ldots, a_{K-1}, \gamma_1, \gamma_2, \gamma_3).$$

As suggested above, we impose that average quality is 0, so all a_k's are in deviation from an hypothetical average team with 0 quality, thus set the ability of the K^{th} team as follows

$$a_K = -\sum_{k=1}^{K-1} a_k.$$

The posterior distribution of $\boldsymbol{\psi}$ can be obtained as

$$f_{\text{post}}(\boldsymbol{\psi}|z) \propto f_{\text{like}}(z|\boldsymbol{\psi}) f_{\text{prior}}(\boldsymbol{\psi}),$$

where the likelihood is given by

$$\begin{aligned} f_{\text{like}}(z|\boldsymbol{\psi}) &= \prod_{i=1}^{n} f_{\text{GPDD}}(z_i|\lambda_{1,i},\ \lambda_{2,i},\ \theta_1,\ \theta_2) \\ &= \prod_{i=1}^{n} f_{\text{GPDD}}\left(z_i|\ \psi\right), \end{aligned} \tag{14}$$

and $f_{\text{prior}}(\boldsymbol{\psi})$ is the prior distribution given by

$$f_{\text{prior}}(\boldsymbol{\psi}) = f_H(H) f_{a_1}(a_1) \cdot \ldots \cdot f_{a_k}(a_{k-1}) f_{\gamma_1}(\gamma_1) f_{\gamma_2}(\gamma_2) f_{\gamma_3}(\gamma_3).$$

It can be seen that the posterior distribution is known up to a normalising constant and not analytically tractable. Thus, in order to generate samples from the posterior distribution we use MCMC methods, more specifically, the random walk Metropolis-Hastings (M-H) algorithm.

4.3 Metropolis Hastings algorithm

In the M-H algorithm, for each of the parameters $H, a_1, \ldots, a_{K-1}$, we choose our proposal distribution to be normal, such that in the jth iteration, for $j = 1, \ldots, m$, where m is the number of iterations, the normal distribution is centred at the value from the previous iteration for some suitably chosen value of the variance parameter. The proposal density for γ_1 is $\gamma_1^{\text{cand}} \sim \Gamma(\alpha^{(j-1)}, \beta^{(j-1)})$, where

$$\alpha^{(j-1)} = \frac{{\gamma_1^{(j-1)}}^2}{\sigma_{\gamma_1}^2} \text{ and } \beta^{(j-1)} = \frac{\gamma_1^{(j-1)}}{\sigma_{\gamma_1}^2}$$

and those of γ_2 and γ_3 are truncated normals with the mean centred on the current values.

The choice of the value of the dispersion parameter of the proposal distribution is an important

point to consider in the case of random walk chains. A large value for the variance allows a greater variation from the previous value, but will lead to a very small acceptance rate. On the other hand, a small value of the variance results in draws which are close to the previous value with a high acceptance rate (Gamerman and Lopes, 2006). The optimal choice for the variance of the normal proposal is $\sigma^2 = c^2\Sigma$, where $c \approx 2.4/\sqrt{d}$ (d is the dimension of the parameter vector) and Σ is the variance-covariance matrix based on the curvature of the posterior at the mode (Tanner, 1998).

4.4 Simulating posterior predictive distribution

Let $Z^{\text{pred}} = (Z_1^{\text{pred}}, \ldots, Z_n^{\text{pred}})^{\text{T}}$ be the goal differences of n matches which we wish to predict. The posterior predictive distribution is defined as

$$f_{\text{pred}}(Z^{\text{pred}}|z) = \int f_{\text{like}}(Z^{\text{pred}}|\boldsymbol{\psi}) f_{\text{post}}(\boldsymbol{\psi}|z) d\boldsymbol{\psi}, \tag{15}$$

where $f_{\text{like}}(Z^{\text{pred}}|\boldsymbol{\psi})$ may be given as

$$f_{\text{like}}(Z^{\text{pred}}|\boldsymbol{\psi}) = \prod_{i=1}^{n} f(Z_i^{\text{pred}}|\boldsymbol{\psi}). \tag{16}$$

To obtain the prediction distribution, we add the following steps to the MCMC algorithm. At the jth iteration, $j = 1, \ldots, m$, where m is the number of MCMC iterations, repeat the following steps, for $i = 1, \ldots, n$:

1. Obtain

$$\lambda_{1,i}^{\text{pred},(j)} \text{ and } \lambda_{2,i}^{\text{pred},(j)},$$

$$\theta_1^{\text{pred},(j)} = 1 - \exp(\gamma_1^{\text{pred},(j)}) \text{ and } \theta_2^{\text{pred},(j)} = 1 - \exp(\gamma_2^{\text{pred},(j)})$$

   where $\lambda_{1,i}^{\text{pred},(j)}$, $\lambda_{2,i}^{\text{pred},(j)}$, $\theta_1^{\text{pred},(j)}$ and $\theta_2^{\text{pred},(j)}$ are the GPDD model parameters which are obtained by substituting the values of H, $a_1, \ldots, a_k$, γ_1, γ_2 and γ_3 at the j^{th} iteration in (10)-(13).

2. Randomly draw

$$Z_i^{\text{pred},(j)} \sim \text{GPDD}(\lambda_{1,i}^{\text{pred},(j)}, \lambda_{2,i}^{\text{pred},(j)}, \theta_1^{\text{pred},(j)}, \theta_2^{\text{pred},(j)}),$$

where a random sample from the GPDD is obtained as the difference of the two generalised Poisson random variables.

The posterior predictive distribution will be used to probabilistically quantify the response variable which enables us to assess the goodness of fit and the overall performance of the model. Therefore, if the predictive distribution, in general, is in agreement with the observed data, this implies a good fit to the model. This also can be used to estimate the ranking distribution if the competition had a different structure (Karlis and Ntzoufras, 2009; Lee, 1997).

	Mean	Median	2.50%	97.50%
H	0.394	0.394	0.248	0.553
γ_1	1.719	1.709	1.405	2.072
θ_1	-0.037	-0.032	-0.163	0.062
θ_2	-0.033	-0.028	-0.197	0.094

Table 1: Posterior mean, median and 95% credible intervals for the home effect (H), constant term in the variance (γ_1), θ_1 and θ_2.

5 Results

The GPDD model is fitted to a set of data from 2012-2013 Italian Serie A football league[1]. This tournament comprises 20 teams with matches played between August 2012 and May 2013. The tournament has a double-round-robin structure, so each team competes twice against all the other teams in the league. The total number of matches is $20(20-1) = 380$ played over 91 days. On average home and visiting teams scored 1.49 and 1.14 goals, respectively.

Informative prior distributions described in Section 4.1 were assigned to home effect and team abilities (Appendix A: Table 4). The teams Sampdoria, Torino and US Pescara were promoted to Serie A league in 2012-13, hence non-informative normal priors were assigned to their abilities. After running the M-H algorithm, the Gelman and Rubin's convergence diagnostic with the statistic value of 1.01 suggested that we can consider the convergence of the MCMC chains after 15000 iterations (Gelman and Rubin, 1992). A further 5000 samples were collected after burn-in.

Table 1 provided summary statistics of the posterior predictive distribution of H, γ_1, θ_1 and θ_2. A positive home effect can be observed indicating that the expected goal difference is positive when teams with equal abilities are playing. The small but negative values of θ_1 and θ_2 may suggest that the distribution of the goal difference tends to have slightly shorter tails than a PD distribution. It also suggests that though the left tail is slightly longer that the right tail, the values of θ_1 and θ_2 are quite similar.

A plot of the 95% posterior interval for team abilities is provided in Figure 1. According to this plot, Juventus, Napoli and Fiorentina have the highest abilities which is supported by the actual overall goal difference. Also, it can be seen that the model predicted the abilities of Sampdoria, Torino and US Pescara well.

Figure 2 depicts the observed and predicted frequencies for the overall goal differences. It is evident that there is a close agreement between the observed and predicted values. The 95% credible predictive intervals contains the observed value, indicating only minor deviation from the data. Under the GPDD model it is expected that approximately 60 matches played at home would finish with -1 goal difference, whereas in practice 50 games finished with -1 goal difference. This was compensated by the number of games finished with a goal difference of -2 (expected frequency 30 as against observed frequency 40).

Nevertheless, the posterior predicted distribution of the goal difference shows a strong agree-

[1] http://espnfc.com/?cc=5739

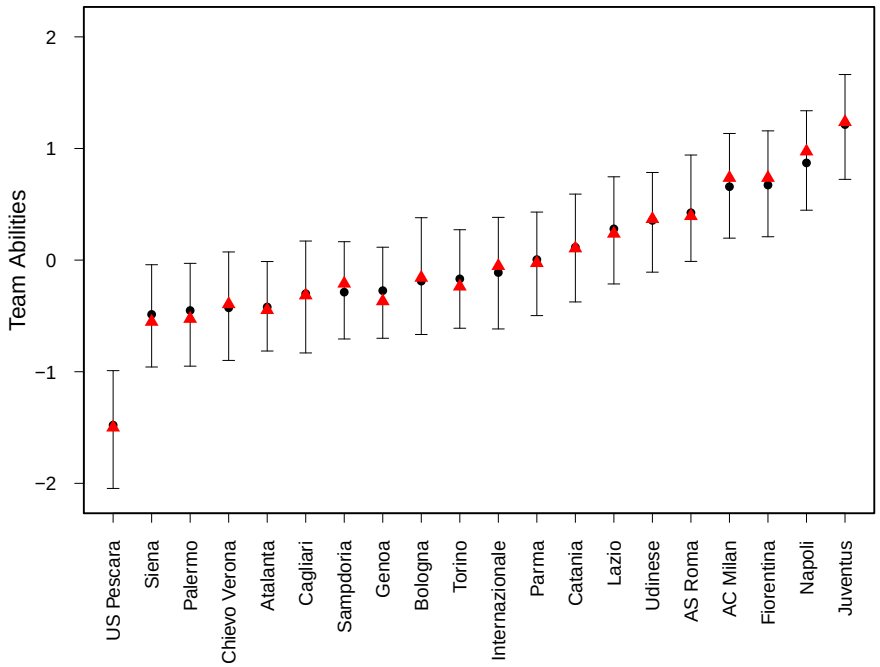

Figure 1: Posterior mean and 95% credible intervals of the team abilities $(a_1, \ldots, a_{20})$ and the observed mean goal difference illustrated by red triangles.

ment between predicted and observed goal differences for each team (Table 2). Minor differences were observed between the expected and observed goals. For 18 of the 20 teams, the difference between the expected and observed goal differences were no greater than 2. There are only two teams that had a difference of 3 goals. Furthermore, we ranked the teams based on the expected goal difference. It can be seen that the top 10 and the bottom 5 teams have the same expected and observed ranks in the final table. The predicted rank of teams with ranks 12-16 differ from their observed ranks by only one position and that of the team with rank 11 (Torino) has an observed rank of 13.

5.1 Model checking

The performance of the model is evaluated using a deviation score based on the predictive distribution used also by Karlis and Ntzoufras (2009) and compared with a GPDD model with a constant variance with respect to team abilities. We compared the deviation of the predicted values of the quantity of interest from their corresponding observed values. Adopting the notations of Karlis and Ntzoufras (2009), the deviation score was defined by

$$\text{Deviation} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(E(Q_l^{\text{pred}}|z) - Q_l^{\text{obs}}\right)^2} \tag{17}$$

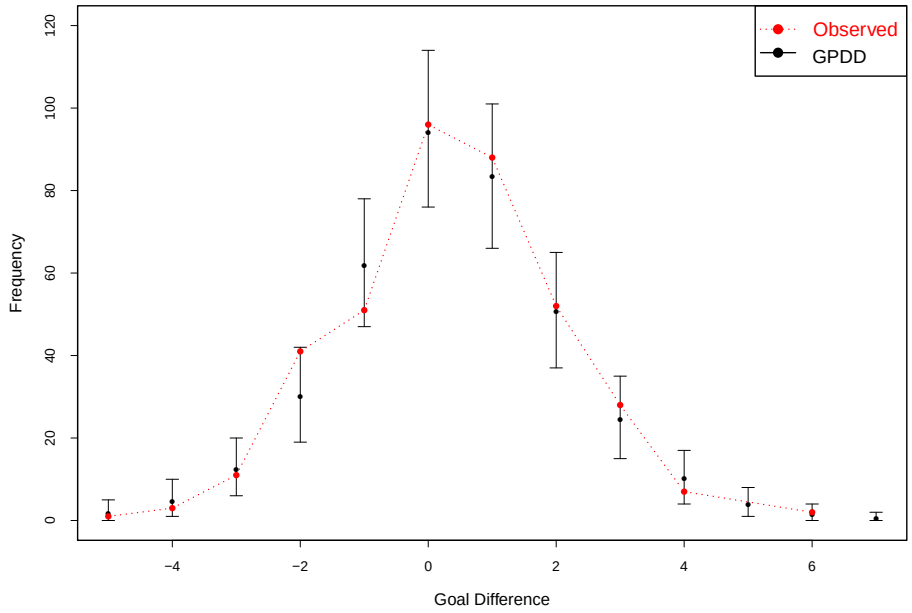

Figure 2: Posterior predictive distribution of the goal difference under the GPDD model along with the observed frequency.

where Q_l^{obs} is the observed quantity, $E(Q_l^{\text{pred}}|z)$ is the predicted quantity and L is the length of vector Q. For the calculation of the deviations of the frequencies and the relative frequencies, we used $L = 13$ in order to consider differences from -6 to 6, while for the deviations of the expected differences from the final table, we set $L = K = 20$, i.e. the number of teams in the league. The deviation scores were compared with a similar GPDD model with a fixed variance. Let M_1 be the GPDD model with a constant variance with respect to team abilities and M_2 be the GPDD model described in (8)-(9). It can be seen from Table 3 that modelling the variance improves the fit of the model for this set of data.

In order to show that model M_2 model outperforms model M_1, we use the Bayes factor (BF), which is defined as the ratio of marginal likelihood under one model to the marginal likelihood under the other model (Gelman et al., 2003; Kass and Raftery, 1995; Plummer et al., 2006):

$$\begin{aligned} \text{BF}(M_2, M_1) &= \frac{f_{like,M_2}(z|M_2)}{f_{like,M_1}(z|M_1)} \\ &= \frac{\int f_{prior,M_2}(\psi|M_2)\ f_{like,M_2}(z|\psi', M_2)\ d\psi'}{\int f_{prior,M_1}(\psi|M_1)\ f_{like,M_1}(z|\psi, M_1)\ d\psi}, \end{aligned} \tag{18}$$

where ψ and ψ' are vectors of parameters under the M_1 and M_2 distributions, respectively and $f_{like,M_1}(z|M_1)$ and $f_{like,M_2}(z|M_2)$ are marginal likelihoods under M_1 and M_2 models. A Bayes factor greater than one (or $\log(\text{BF}) > 0$) supports M_2 model over M_1 model. In our case, we obtain $\text{BF}(M_2, M_1) = 5.60$ which suggests that modelling the variance in addition to the mean

				95% Credible interval	
Precited (Obs) Rank	**Teams**	**Observed**	**Mean**	**2.5%**	**97.5%**
1(1)	Juventus	47	48	21	77
2(2)	Napoli	37	35	10	62
3(4)	Fiorentina	28	27	1	54
4(3)	AC Milan	28	26	0	54
5(5)	AS Roma	15	17	-9	44
6(6)	Udinese	14	14	-11	40
7(7)	Lazio	9	11	-15	38
8(8)	Catania	4	5	-22	31
9(9)	Parma	-1	0	-26	25
10(10)	Internazionale	-2	-4	-32	23
11(13)	Torino	-9	-7	-33	18
12(11)	Bologna	-6	-8	-35	21
13(12)	Sampdoria	-8	-11	-37	14
14(15)	Genoa	-14	-11	-36	13
15(14)	Cagliari	-12	-12	-39	14
16.5(16)	Chievo Verona	-15	-17	-44	9
16.5(17)	Atalanta	-17	-17	-41	7
18(18)	Palermo	-20	-18	-46	7
19(19)	Siena	-21	-19	-46	6
20(20)	US Pescara	-57	-59	-92	-30

Table 2: Expected (observed) rank, observed goal difference, posterior mean and 95% credible intervals of goal differences for the teams in the 2012-13 Italian Serie A football league.

	M1	M2
Frequency (counts/games)	3.72	3.32
Relative frequency (counts)	0.01	0.008
Frequency of win/draw/loss	6.79	2.93
Relative frequency of win/draw/loss	0.017	0.006
Expected goal difference	6.22	1.81

Table 3: Deviations between observed and predictive measures for models M1 and M2.

considerably improves the fit.

Finally, Figure 3 illustrates predictions for the total number of wins/draws/losses for the next season (2013-14 Italian Serie A football league) using the GPDD model. For the purpose of a uniform comparison, the horizontal axes were sorted by the total number of wins (ascending from left to right) for all the three panels. It can be seen that out of 60 observed wins/draws/losses, 56 values fall within the 95% credible intervals. In other words, in the top panel, the number of wins was underestimated for Juventus and AS Roma, and overestimated for Bologna. On the other hand, in the bottom panel, the number of losses was underestimated for Catania.

6 Discussion

In this paper, we proposed the application of the GPDD for modelling football data. The proposed model has some interesting advantages over a similar PD model used for the same purpose. The GPDD model, like the PD model, is based on the goal difference in each game which allows us to

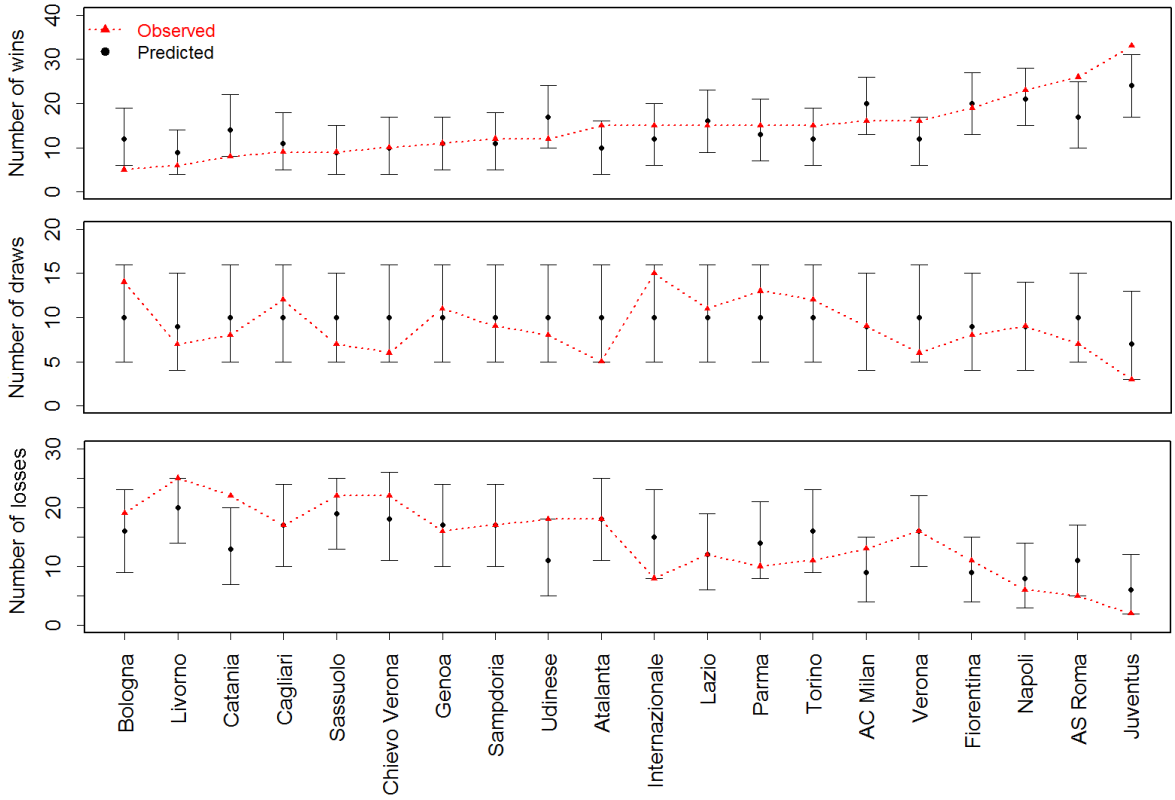

Figure 3: Predicted number of wins/draws/losses with median and 95% credible intervals and observed values for 2013-14 Italian Serie A football league.

account for the correlation without explicitly modelling it as needed in any bivariate distribution. The main advantage of the GPDD model to the PD model is its flexibility in the tails that allows for shorter as well as longer tails than the PD distribution.

We carried out the analysis in a Bayesian framework which allowed us to incorporate external information through the prior distributions. We chose the prior distributions of team abilities and the home effect based on the previous year's league table. We modelled both the mean and the variance of the goal difference such that the variance was defined as a function of the absolute difference in the team abilities. In this paper, we obtained small values for the estimates of the parameters θ_1 and θ_2. However, the possibility to model the variance of the distribution appears important. More importantly, the utilisation of the generalised Poisson difference distribution may be more relevant in situations in which the tails of the distribution of the difference in scores are longer, as for example with basketball data.

We demonstrated that such a model performs considerably better than a model with a fixed variance. We could also see that the proposed model did not overestimate the number of drawn games as was the case with the PD distribution. However, for the 2012-13 Italian Serie A football data, our GPDD model predicted more games with -1 goal difference than that was actually observed. This was, in fact, compensated by predicting fewer games with -2 goal difference. Finally, various model diagnostics were carried out that showed that goal differences were explained

very well by the fitted GPDD model.

Our model in its current form does not allow for time-varying team abilities. Thus, an extension would be to consider a dynamic model with varying team abilities and home effect. Another possible extension of the model would be time varying home effect and team specific θ_1 and θ_2.

References

Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (ninth Dover printing, tenth GPO printing ed.). New York: Dover.

Cattelan, M., C. Varin, and D. Firth (2013). Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(1), 135–150.

Clarke, S. R. and J. M. Norman (1995). Home ground advantage of individual clubs in english soccer. *Journal of the Royal Statistical Society. Series D (The Statistician) 44*(4), pp. 509–521.

Consul, P. C. (1986). On the differences of two generalized Poisson variates. *Communications in statistics. Simulation and computation 15*(3), 761–767.

Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and Inefficiencies in the football betting market. *Appl. Statist. 46*, 265 – 280.

Gamerman, D. and H. F. Lopes (2006). *Markov chain Monte Carlo: stochastic simulation for bayesian inference*. Boca Raton: Chapman&Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and R. B. Donald (2003, July). *Bayesian Data Analysis* (second ed.). London: Chapman & Hall/CRC.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393.

Karlis, D. and I. Ntzoufras (2006). Bayesian analysis of the differences of count data. *Statistics in Medicine 25*(11), 1885–1905.

Karlis, D. and I. Ntzoufras (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA J Management Math 20*(2), 133–145.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), pp. 773–795.

Koning, R. H. (2000). Balance in Competition in Dutch Soccer. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*(3), 419–431.

Lee, A. J. (1997). Modeling scores in the premier league: Is manchester united really the best? *CHANCE 10*(1), 15–19.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica 36*(3), 109–118.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shahtahmassebi, G. and R. Moyeed (2014). Bayesian modelling of integer data using the generalised poisson difference distribution. *International Journal of Statistics and Probability 3*(1), 35–48.

Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society 109*(3), 296.

Tanner, M. A. (1998). *Tools for statistical inference: models for the exploration of posterior distributions and likelihood functions*. New York: Springer.

A Prior mean and variance

Teams	$\bar{z}_j$	$\sigma^2_{z_j}$	Teams	$\bar{z}_j$	$\sigma^2_{z_j}$
AC Milan	1.079	2.886	Catania	-0.132	2.334
Napoli	0.526	2.959	Fiorentina	-0.158	2.731
Udinese	0.447	2.038	Cagliari	-0.237	2.834
Lazio	0.237	2.942	Chievo Verona	-0.263	1.767
AS Roma	0.158	3.38	Palermo	-0.263	3.118
Internazionale	0.079	3.264	Lecce	-0.421	1.818
Parma	0.026	3.053	Genoa	-0.5	2.797
Siena	0	3.297	Novara	-0.79	2.603
Atalanta	-0.053	2.159	Cesena	-0.947	1.835

Table 4: Mean and variance of the goal difference of teams which played in 2011-12 Italian Serie A league.