Bayesian hierarchical model for the prediction of football results

Article *in* Journal of Applied Statistics · February 2010

DOI: 10.1080/02664760802684177 · Source: RePEc

2 authors:

Gianluca Baio
University College London
278 PUBLICATIONS **7,160** CITATIONS
SEE PROFILE

Marta Blangiardo
Imperial College London
239 PUBLICATIONS **6,970** CITATIONS
SEE PROFILE

Bayesian hierarchical model for the prediction of football results

Gianluca Baio[1,2*] Marta A. Blangiardo[3]

[1] *University College London*
Department of Statistical Sciences,
Gower Street, London WC1 6BT
Tel.: +44(0)20 7679 1879, Fax: +44(0) 7383 4703,
email: gianluca@statistica.it

[2] *University of Milano Bicocca*
Department of Statistics
Via Bicocca degli Arcimboldi, 8 20126, Milano (Italy)
Tel.: +39(0)2 6448 5847, Fax: +39(0)2 647 3312 ,
email: gianluca.baio@unimib.it

[3] *Imperial College London*
Department of Epidemiology and Public Health
Norfolk Place, London W2 1PG
Tel.: +44(0)20 7594 1542 , Fax: +44(0)20 7402 2150,
email: m.blangiardo@imperial.ac.uk

Abstract

The problem of modelling football data has become increasingly popular in the last few years and many different models have been proposed with the aim of estimating the characteristics that bring a team to lose or win a game, or to predict the score of a particular match. We propose a Bayesian hierarchical model to address both these aims and test its predictive strength on data about the Italian Serie A championship 1991-1992. To overcome the issue of overshrinkage produced by the Bayesian hierarchical model, we specify a more complex mixture model that results in better fit to the observed data. We test its performance using an example about the Italian Serie A championship 2007-2008.

Keywords: Bayesian hierarchical models; overshrinkage; Football data; bivariate Poisson distribution.

1 Introduction

Statistical modelling of sport data is a popular topic and much research has been produced to this aim, also in reference to football. From the statistical point of view, the task is stimulating because it raises some interesting issues. One such

*To whom correspondence should be sent. However, this work is the result of equal participation of both authors in all the aspects of preparation, analysis and writing of the manuscript.

matter is related to the distributional form associated with the number of goals scored in a single game by the two opponents.

Although the Binomial or Negative Binomial have been proposed in the late 1970s (Pollard et al. 1977), the Poisson distribution has been widely accepted as a suitable model for these quantities; in particular, a simplifying assumption often used is that of independence between the goals scored by the home and the away team. For instance, Maher (1982) used a model with two independent Poisson variables where the relevant parameters are constructed as the product of the strength in the attack for one team and the weakness in defense for the other.

Despite that, some authors have shown empirical, although relatively low, levels of correlation between the two quantities (Lee 1997, Karlis & Ntzoufras 2000). Consequently, the use of more sophisticated models have been proposed, for instance by Dixon & Coles (1997), who applied a correction factor to the independent Poisson model to improve the performance in terms of prediction. More recently, Karlis & Ntzoufras (2000, 2003) advocated the use of a bivariate Poisson distribution that has a more complicated formulation for the likelihood function, and includes an additional parameter explicitly accounting for the covariance between the goals scored by the two competing teams. They specify the model in a frequentist framework (although extensions using the Bayesian approach have been described by Tsionas 2001), and their main purpose is the estimation of the effects used to explain the number of goals scored.

We propose in this paper a Bayesian herarchical model for the number of goals scored by the two teams in each match. Hierarchical models are widely used in many different fields as they are a natural way of taking into account relations between variables, by assuming a common distribution for a set of relevant parameters, thought to underlay the outcomes of interest (Congdon 2003).

Within the Bayesian framework, which naturally accommodates hierarchical models (Bernardo & Smith 1999), there is no need of the bivariate Poisson modelling. We show here that assuming two *conditionally independent* Poisson variables for the number of goals scored, correlation is taken into account, since the observable variables are mixed at an upper level. Moreover, as we are framed in a Bayesian context, prediction of a new game under the model is naturally accommodated by means the posterior predictive distribution.

The paper is structured as follow: first we describe in § 2 the model and the data used; § 3 describes the results in terms of parameter estimations as well as prediction of a new outcome. In § 4 we deal with the problem of overshrinkage and present a possible solution using a mixture model. Finally § 5 presents some issues and some possible extension that can be material for future work and in § 6 we include the `WinBUGS` code for our analysis.

2 The model

In order to allow direct comparison with Karlis & Ntzoufras (2003), we first consider the Italian Serie A for the season 1991-1992. The league is made by a total of $T = 18$ teams, playing each other twice in a season (one at home and one away). We indicate the number of goals scored by the home and by the away team in the $g-$th game of the season ($g = 1, \ldots, G = 306$) as y_{g1} and y_{g2}

respectively.

The vector of observed counts $\boldsymbol{y} = (y_{g1}, y_{g2})$ is modelled as independent Poisson:

$$y_{gj} \mid \theta_{gj} \sim \text{Poisson}(\theta_{gj}),$$

where the parameters $\boldsymbol{\theta} = (\theta_{g1}, \theta_{g2})$ represent the scoring intensity in the g-th game for the team playing at home ($j = 1$) and away ($j = 2$), respectively.

We model these parameters according to a formulation that has been used widely in the statistical literature (see Karlis & Ntzoufras 2003 and the reference therein), assuming a log-linear random effect model:

$$\begin{aligned} \log \theta_{g1} &= home + att_{h(g)} + def_{a(g)} \\ \log \theta_{g2} &= att_{a(g)} + def_{h(g)}. \end{aligned}$$

Poisson-logNormal models have been discussed and widely used in the literature — see for instance Aitchinson & Ho (1989), Chib & Winkelman (2001) and Tunaru (2002).

The parameter *home* represents the advantage for the team hosting the game and we assume that this effect is constant for all the teams and throughout the season. In addition, the scoring intensity is determined jointly by the attack and defense ability of the two teams involved, represented by the parameters *att* and *def*, respectively. The nested indexes $h(g), a(g) = 1, \ldots, T$ identify the team that is playing at home (away) in the g-th game of the season.

The data structure for the model is presented in Table 1 and consist of the name and code of the teams, and the number of goals scored for each game of the season. As is possible to see, the indexes $h(g)$ and $a(g)$ are uniquely associated with one of the 18 teams. For example, in Table 1 Sampdoria are always associated with the index 16, whether they play away, as for $a(4)$, or at home, as for $h(303)$.

g	*home team*	*away team*	$h(g)$	$a(g)$	y_{g1}	y_{g2}
1	Verona	Roma	18	15	0	1
2	Napoli	Atalanta	13	2	1	0
3	Lazio	Parma	11	14	1	1
4	Cagliari	Sampdoria	4	16	3	2
...	...	...	...	...	...	...
303	Sampdoria	Cremonese	16	5	2	2
304	Roma	Bari	15	3	2	0
305	Inter	Atalanta	9	2	0	0
306	Torino	Ascoli	17	1	5	2

Table 1: The data for the 'Serie A' 1991-1992

In line with the Bayesian approach, we have to specify some suitable prior distributions for all the random parameters in our model. The variable *home* is modelled as a fixed effect, assuming a standard flat prior distribution (notice that we use here the typical notation to describe the Normal distribution in terms of the mean and the precision):

$$home \sim \text{Normal}(0, 0.0001).$$

Conversely, for each $t = 1, \ldots, T$, the team-specific effects are modelled as exchangeable from a common distribution:

$$att_t \sim \text{Normal}(\mu_{att}, \tau_{att}), \qquad def_t \sim \text{Normal}(\mu_{def}, \tau_{def}).$$

As suggested by various works, we need to impose some identifiability constraints on the team-specific parameters. In line with Karlis & Ntzoufras (2003), we use a sum-to-zero constraint, that is

$$\sum_{t=1}^{T} att_t = 0, \qquad \sum_{t=1}^{T} def_t = 0.$$

However, we also assessed the performance of the model using a corner-constraint instead, in which the team-specific effect for only one team are set to 0, for instance $att_1 := 0$ and $def_1 := 0$. Even if this latter method is slightly faster to run, the interpretation of these coefficients is incremental with respect to the baseline identified by the team associated with an attacking and defending strength of 0 and therefore is less intuitive.

Finally, the hyper-priors of the attack and defense effects are modelled independently using again flat prior distributions:

$$\begin{aligned} \mu_{att} &\sim \text{Normal}(0, 0.0001), & \mu_{def} &\sim \text{Normal}(0, 0.0001), \\ \tau_{att} &\sim \text{Gamma}(0.1, 0.1), & \tau_{def} &\sim \text{Gamma}(0.1, 0.1). \end{aligned}$$

A graphical representation of the model is depicted in Figure 1. The inherent hierarchical nature implies a form of correlation between the observable variables y_{g1} and y_{g2} by means of the unobservable hyper-parameters $\boldsymbol{\eta} = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$. In fact, the components of $\boldsymbol{\eta}$ represent a latent structure that we assume to be common for all the games played in a season and that determine the average scoring rate.

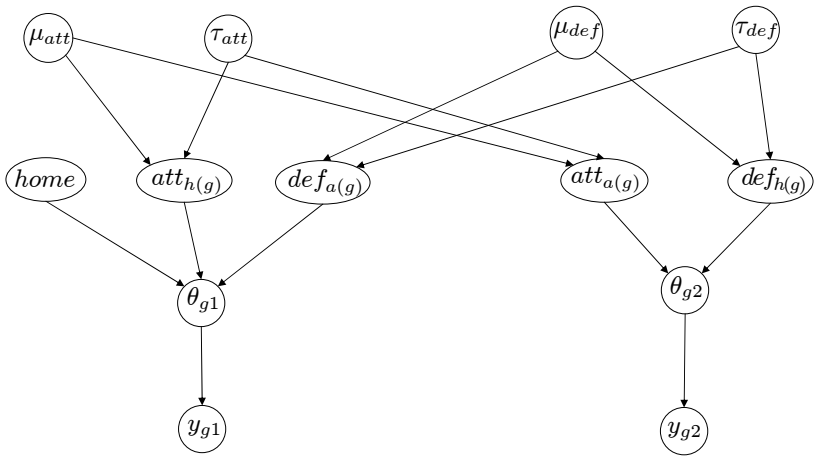


Figure 1: The DAG representation of the hierarchical model

Each game contributes to the estimation of these parameters, which in turn generate the main effects that explain the variations in the parameters $\boldsymbol{\theta}$ and therefore implying a form of correlation on the observed counts $\boldsymbol{y}$.

3 Results

According to the Bayesian approach, the objective of our modelling is two-fold: first, we wish to estimate the value of the main effects that we used to explain the scoring rates. This task is accomplished by entering the evidence provided by the observed results (the vector $\boldsymbol{y}$) and updating the prior distributions by means of the Bayes's theorem using an MCMC-based procedure.

teams	*attack effect*				*defense effect*			
	mean	*2.5%*	*median*	*97.5%*	*mean*	*2.5%*	*median*	*97.5%*
Ascoli	-0.2238	-0.5232	-0.2165	0.0595	0.4776	0.2344	0.4804	0.6987
Atalanta	-0.1288	-0.4050	-0.1232	0.1321	-0.0849	-0.3392	-0.0841	0.1743
Bari	-0.2199	-0.5098	-0.2213	0.0646	0.1719	-0.0823	0.1741	0.4168
Cagliari	-0.1468	-0.4246	-0.1453	0.1255	-0.0656	-0.3716	-0.0645	0.2109
Cremonese	-0.1974	-0.4915	-0.1983	0.0678	0.1915	-0.0758	0.1894	0.4557
Fiorentina	0.1173	-0.1397	0.1255	0.3451	0.0672	-0.1957	0.0656	0.3372
Foggia	0.3464	0.1077	0.3453	0.5811	0.3701	0.1207	0.3686	0.6186
Genoa	-0.0435	-0.3108	-0.0464	0.2149	0.1700	-0.0811	0.1685	0.4382
Inter	-0.2077	-0.4963	-0.2046	0.0980	-0.2061	-0.5041	-0.2049	0.0576
Juventus	0.1214	-0.1210	0.1205	0.3745	-0.3348	-0.6477	-0.3319	-0.0514
Lazio	0.0855	-0.1626	0.0826	0.3354	0.0722	-0.1991	0.0742	0.3145
Milan	0.5226	0.2765	0.5206	0.7466	-0.3349	-0.6788	-0.3300	-0.0280
Napoli	0.2982	0.0662	0.2956	0.5267	0.0668	-0.2125	0.0667	0.3283
Parma	-0.1208	-0.3975	-0.1200	0.1338	-0.2038	-0.5136	-0.2031	0.0859
Roma	-0.0224	-0.2999	-0.0182	0.2345	-0.1358	-0.4385	-0.1300	0.1253
Sampdoria	-0.0096	-0.2716	-0.0076	0.2436	-0.1333	-0.4484	-0.1317	0.1346
Torino	0.0824	-0.1821	0.0837	0.3408	-0.4141	-0.7886	-0.4043	-0.1181
Verona	-0.2532	-0.5601	-0.2459	0.0206	0.3259	0.1026	0.3254	0.5621
	mean	*2.5%*	*median*	*97.5%*				
home	0.2124	0.1056	0.2128	0.3213				

Table 2: Estimation of the main effects for the loglinear model

Table 2 shows some summary statistics for the posterior distributions of the coefficients for the log-linear model describing the scoring intensity.

Similarly to what found in other works, the home effect is positive (the posterior mean and 95% CI are 0.2124 and $[0.1056; 0.3213]$, respectively). AC Milan, the league winner in that year, have by far the highest propensity to score (as suggested by the posterior mean of 0.5226 for the effect *att*). The top three clubs (AC Milan, Juventus and Torino) performed better than any other club in terms of defence showing the lowest value for the parameter *def*, while Ascoli (who were actually relegated), Foggia and Verona showed the highest propensity to concede goals.

The second — and probably more interesting — objective of the model is that of prediction. We can use the results derived in the implied posterior distributions for the vector $\boldsymbol{\theta}$ to predict a future occurrence of a similar (exchangeable) game. In this case, we produced a vector of 1000 replications for the posterior predictive distribution of $\boldsymbol{y}$ that we used for the purpose of model checking.

Figure 2 shows the comparison between the observed results throughout the season (the black line) and the estimations provided by both the posterior predictions from our model (in blue) and the bivariate Poisson model of Karlis & Ntzoufras (2003) (in red). As one can appreciate, for most of the teams (Atalanta, Foggia, Genoa, Inter, Juventus, AC Milan, Napoli, Parma, Roma, Sampdoria, Torino and Verona), the Bayesian hierarchical model seem to pro-

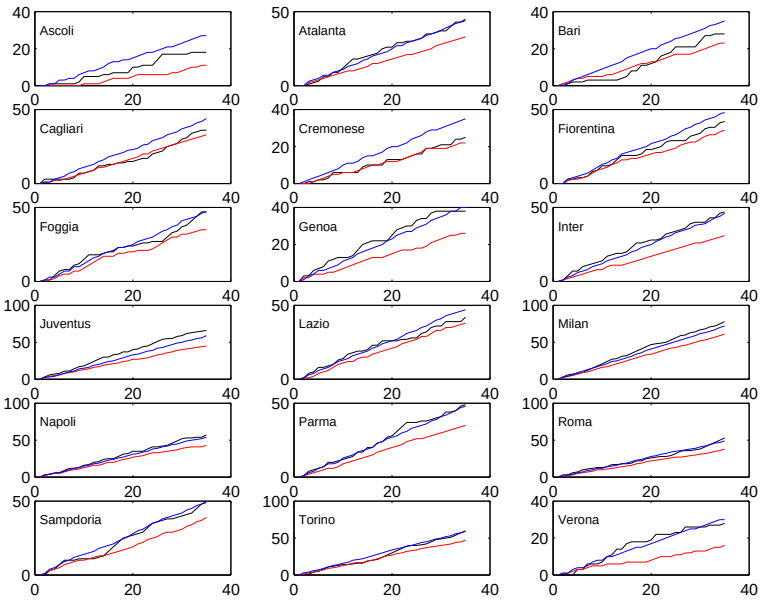


Figure 2: Posterior predictive validation of the model in comparison with Karlis and Ntzoufras (2003). Data are for the Italian Serie A 1991-92. For each team, the dark line represents the observed cumulative points through the season, while the blue and the red lines represent predictions for the Bayesian hierarchical and the Bivariate Poisson model (Karlis & Ntzoufras 2003), respectively

duce a better fit to the observed results. For a few teams, the red line is closer to the black one (Cagliari, Cremonese and, marginally, Lazio), while for Ascoli, Bari and Fiorentina the two estimations are equally poor (with the Bayesian hierarchical model generally overestimating the final result and the bivariate Poisson generally underestimating it). In general, it seems that our model performs better than the bivariate Poisson in terms of adapting to the observed dynamics throughout the season.

4 Reducing overshrinkage caused by hierarchical model

One possible well known drawback of Bayesian hierarchical models is the phenomenon of *overshrinkage*, under which some of the extreme occurrences tend to be pulled towards the grand mean of the overall observations. In the application of a hierarchical model for the prediction of football results this can be particularly relevant, as presumably a few teams will have very good performances (and therefore compete for the final title or the top positions), while some other teams will have very poor performance (struggling for relegation).

The model of § 2 assumes that all the attack and defense propensities be drawn by a common process, characterised by the common vector of hyperparameters $(\mu_{att}, \tau_{att}, \mu_{def}, \tau_{def})$; clearly, this might be not sufficient to capture the presence of different quality in the teams, therefore producing overshrinkage,

with the effect of: *a)* penalising extremely good teams; and *b)* overestimate the performance of poor teams.

One possible way to avoid this problem is to introduce a more complicated structure for the parameters of the model, in order to allow for three different generating mechanism, one for the top teams, one for the mid-table teams, and one for the bottom-table teams. Also, in line with Berger (1984), shrinkage can be limited by modelling the attack and defense parameters using a non central t (nct) distribution on $\nu = 4$ degrees of freedom instead of the normal of § 2.

Consequently, the model for the likelihood, and the prior specification for the θ_{gj} and for the hyper-parameter *home* is unchanged, while the other hyper-parameters are modelled as follows. First we define for each team t two latent (unobservable) variables $grp^{att}(t)$ and $grp^{def}(t)$, which take on the values 1, 2 or 3 identifying the bottom-, mid- or top-table performances in terms of attack (defense). These are given suitable categorical distributions, each depending on a vector of prior probabilities $\boldsymbol{\pi}^{att} = (\pi_{1t}^{att}, \pi_{2t}^{att}, \pi_{3t}^{att})$ and $\boldsymbol{\pi}^{def} = (\pi_{1t}^{def}, \pi_{2t}^{def}, \pi_{3t}^{def})$. We specify minimally informative models for both $\boldsymbol{\pi}^{att}$ and $\boldsymbol{\pi}^{def}$ in terms of a Dirichlet distribution with parameters $(1, 1, 1)$, but obvioulsy one can include (perhaps subjective) prior information on the vectors $\boldsymbol{\pi}^{att}$ and $\boldsymbol{\pi}^{def}$ to represent the prior chance that each team is in one of the three categories.

The attack and defense effects are then modelled for each team t as:

$$att_t \sim \text{nct}\left(\mu_{grp(t)}^{att}, \tau_{grp(t)}^{att}, \nu\right), \qquad def_t \sim \text{nct}\left(\mu_{grp(t)}^{def}, \tau_{grp(t)}^{def}, \nu\right).$$

In particular, since the values of $grp^{att}(t)$ and $grp^{def}(t)$ are unknown, this formulation essentially amounts to defining a mixture model on the attack and defense effects:

$$att_t = \sum_{k=1}^{3} \pi_{kt}^{att} \times \text{nct}\left(\mu_k^{att}, \tau_k^{att}, \nu\right), \quad def_t = \sum_{k=1}^{3} \pi_{kt}^{def} \times \text{nct}\left(\mu_k^{def}, \tau_k^{def}, \nu\right).$$

The location and scale of the nct distributions (as suggested, we use $\nu = 4$) depend on the probability that each team actually belongs in any of the three categories of $grp^{att}(t)$ and $grp^{def}(t)$.

The model for the location and scale parameters of the nct distributions is specified as follows. If a team have poor performance, then they are likely to show low (negative) propensity to score, and high (positive) propensity to concede goals. This can be represented using suitable truncated Normal distributions, such as

$$\begin{aligned} \mu_1^{att} &\sim \text{truncNormal}(0, 0.001, -3, 0) \\ \mu_1^{def} &\sim \text{truncNormal}(0, 0.001, 0, 3). \end{aligned}$$

For the top teams, we can immagine a symmetric situation, that is

$$\begin{aligned} \mu_3^{att} &\sim \text{truncNormal}(0, 0.001, 0, 3) \\ \mu_3^{def} &\sim \text{truncNormal}(0, 0.001, -3, 0) \end{aligned}$$

Finally, for the average teams we assume that the mean of the attack and defense effect have independent dispersed Normal distributions

$$\mu_2^{att} \sim \text{Normal}(0, \tau_2^{att}) \quad \mu_2^{def} \sim \text{Normal}(0, \tau_2^{def})$$

(that is, on average, the attack and defense effects are 0, but can take on both negative or positive values).

For all the groups $k = 1, 2, 3$, the precisions are modelled using independent minimally informative Gamma distributions

$$\tau_k^{att} \sim \text{Gamma}(0.01, 0.01), \qquad \tau_k^{def} \sim \text{Gamma}(0.01, 0.01).$$

4.1 Results for the Italian Serie A 2007-2008

We used the Italian Serie A 2007-2008 to test the model described above. A few major differences between this season and the one described in § 2 should be noticed. Firstly, starting from the season 1994-1995, in Serie A a win is worth 3 points (instead of just 2). Moreover, in the season 2003-2004, the number of teams in the league was increased to 20 (and therefore the total number of games played is now $G = 360$).

These two factors are likely to increase the gap between top and bottom teams and consequently to invalidate the assumption that the attack (defense) effects are drawn from a common distribution. In fact, when using the basic model of § 2 for the 2007-2008 data, a large overshrinkage was produced, penalising in particular Inter and Roma (the top two teams). These clubs performed very well, with over 80 points in the final table, while the estimated points were only 69 and 67, respectively (see Table 3). Similarly, the bottom-table teams were predicted to have significantly more points than observed.

team	*Observed results*			*Basic model (medians)*			*Mixture model (medians)*		
	points	*scored*	*conc'd*	*points*	*scored*	*conc'd*	*points*	*scored*	*conc'd*
Inter	85	69	26	69	62	38	76	65	30
Roma	82	72	37	67	64	42	70	68	40
Juventus	72	72	37	68	65	42	69	67	41
Fiorentina	66	55	39	59	52	43	58	51	43
Milan	64	66	38	64	60	42	66	62	42
Sampdoria	60	56	46	57	53	47	57	54	47
Udinese	57	48	53	50	47	50	50	45	50
Napoli	50	50	53	52	49	49	50	47	51
Genoa	48	44	52	52	50	51	48	43	50
Atalanta	48	52	56	49	45	49	52	50	52
Palermo	47	47	57	49	47	52	47	43	52
Lazio	46	47	51	50	46	49	49	44	50
Siena	44	40	45	49	42	46	48	41	48
Cagliari	42	40	56	46	41	51	45	41	52
Torino	40	36	49	44	39	51	45	39	49
Reggina	40	37	56	45	38	48	44	40	52
Catania	37	33	45	45	37	46	45	37	48
Empoli	36	29	52	41	34	50	40	34	51
Parma	34	42	62	45	42	54	44	42	55
Livorno	30	35	60	40	38	53	42	38	54

Table 3: Posterior predictive validation of the model. Observed and estimated league table (2007/2008)

Figure 3 shows the (average) attack versus defense effects for each of the 20 teams. The positioning of the teams on the plane fits well the observed values of scored and conceded goals (that are obviously the base upon which the attack and defense effects are calculated) — recall that good teams are associated with negative defense effect.

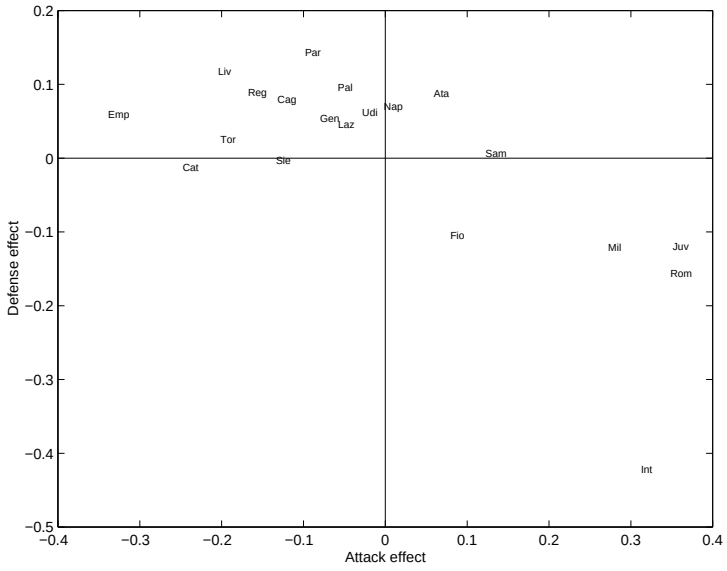


Figure 3: Attack vs defense average effect

As one can observe, several clusters of teams can be identified in the graph. Firstly, Inter has the smallest value for the defense effect (in line with the fact that the number of conceded goals is by far the minimum); Roma, Juventus and AC Milan perform well defensively (although not as good as Inter) all in a similar manner, while showing a very good attacking capacity. Fiorentina and Sampdoria are the best "mid-table" teams (doing slightly better in defense and in attack, respectively, in comparison with the rest of the teams). Empoli (who were actually relegated at the end of the campaign) have the poorest performance in terms of attack. Parma and Livorno have the highest defense values (that is the worst performances).

Figure 4 shows the posterior prediction of the entire season. As one can see, for many of the teams the dynamics are quite consistent with the observed results (see for example Atalanta, Genoa, Juventus, AC Milan, Napoli, Palermo and Sampdoria). The "extreme" observations are subject to some shrinkage (but to a much lower extent, as compared to the results from the standard model). In particular, while the estimation for the top-table teams (Inter and Roma in particular) is relatively in line with the observed results, the performances of Parma, Livorno and perhaps Reggina are slightly overestimated.

Table 3 shows the estimated final table in comparison with the one actually observed. As one can see, despite some differences in the overall number of points (particularly for the extreme teams), the estimates of the number of scored and conceded goals are generally very much in line with the observed results. The distribution of the *home* effect is characterised by a mean and 95% credible interval of 0.3578 and $[0.2748; 0.4413]$ respectively.

Finally, Figure 5 shows the posterior probability that each team belongs in one of the three groups (bottom-, mid- and top-table) — respectively for (a) the attack, and (b) the defense effects. Again, considering the final standings showed in Table 3, the results seem reasonable.

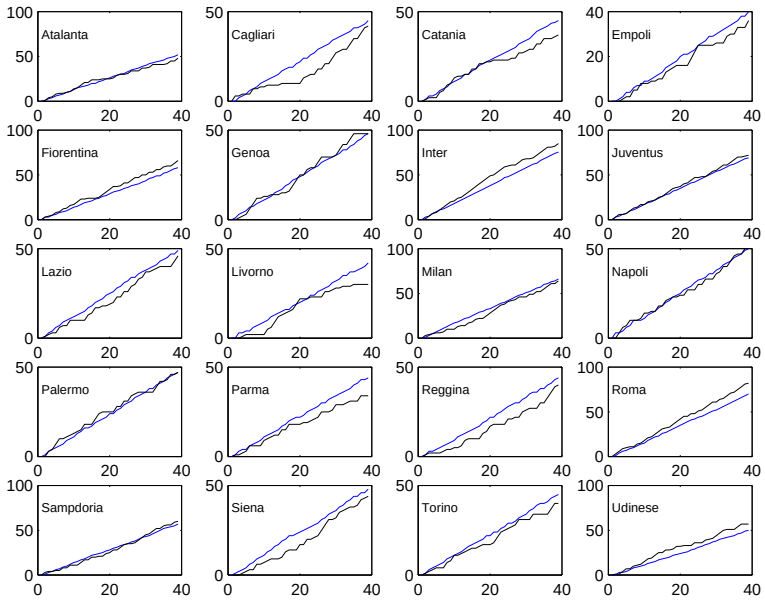
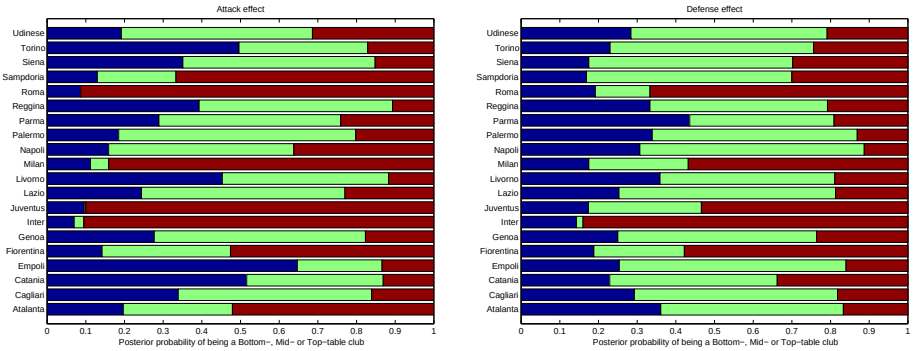


Figure 4: Posterior predictive validation of the mixture model: the black line represents the observed cumulative points through the season, while the blue line represents predictions for the Bayesian hierarchical mixture model



(a) (b)

Figure 5: Posterior probability that each team belongs in one of the three groups

5 Discussion

The model presented in this paper is a simple application of Bayesian hierarchical modelling. The basic structure presented in § 2 can be easily implemented and run using standard MCMC algorithms, such as the one provided for `WinBUGS` in § 6. The performance of this model is not inferior to the one used by Karlis & Ntzoufras (2003), which relies on a bivariate Poisson structure and requires a specific algorithm.

Moreover, the hierarchical model can be easily extended to include a mixture structure to account for the fact that teams show a different "*propensity*" to

score and concede goals, as represented by the attack and defense effects. In this case, the model becomes more complex and time consuming, but it can still be accommodated within standard MCMC algorithms (that we again developed in WinBUGS).

Sensitivity analysis has been performed on the he choice of the arbitrary cutoffs of ± 3 for the truncated Normal distributions used in the mixture model. When larger values were chosen, the model was not able to assign the teams to the three components of mixture, with almost all them being associated with the second category. This is intuitively due to the fact that when the truncated Normal distributions have a larger support, their density is too low, in comparison with the central component. On the other hand, when the cutoff is too small, then the densities of the extreme components are too high and therefore none of the teams is assigned to the second category.

One of the limitations of the results produced by the model (rather than of the model itself) is that, for the sake of simplicity, predictions are obtained in one batch, that is for all the G games of the season, using the observed results to estimate the parameters. An alternative, more complex approach would be to *dynamically* predict new istances of the games. One possibility would be to define the hyper-parameters $\boldsymbol{\eta}$ as "time"-specific, in order to account for periods of variable form of the teams (including injuries, suspensions, etc.). Moreover, prior information can be included at various level in the model, perhaps in the form of expert opinion about the strength of each team.

6 Appendix – WinBUGS model

The complete codes for the WinBUGS models presented in § 2 and § 4 are given below. Notice the use of the function `djl.dnorm.trunc` used to code the truncated Normal distribution. This is not available in the standard WinBUGS version, but can be freely downloaded from the internet (see Lunn 2008).

```
model {
# LIKELIHOOD AND RANDOM EFFECT MODEL FOR THE SCORING PROPENSITY
    for (g in 1:ngames) {
# Observed number of goals scored by each team
        y1[g] ~ dpois(theta[g,1])
        y2[g] ~ dpois(theta[g,2])
# Predictive distribution for the number of goals scored
        ynew[g,1] ~ dpois(theta[g,1])
        ynew[g,2] ~ dpois(theta[g,2])
# Average Scoring intensities (accounting for mixing components)
        log(theta[g,1]) <- home + att[hometeam[g]] + def[awayteam[g]]
        log(theta[g,2]) <- att[awayteam[g]] + def[hometeam[g]]
        }

# 1. BASIC MODEL FOR THE HYPERPARAMETERS
# prior on the home effect
    home ~ dnorm(0,0.0001)

# Trick to code the ''sum-to-zero'' constraint
    for (t in 1:nteams){
        att.star[t] ~ dnorm(mu.att,tau.att)
        def.star[t] ~ dnorm(mu.def,tau.def)
        att[t] <- att.star[t] - mean(att.star[])
```

```
        def[t] <- def.star[t] - mean(def.star[])
    }
# priors on the random effects
    mu.att ~ dnorm(0,0.0001)
    mu.def ~ dnorm(0,0.0001)
    tau.att ~ dgamma(.01,.01)
    tau.def ~ dgamma(.01,.01)

# 2. MIXTURE MODEL FOR THE HYPERPARAMETERS
# prior on the home effect
    home ~ dnorm(0,0.0001)

# Mixture parameters & components (``sum-to-zero'' constraint)
for (t in 1:nteams){
        grp.att[t] ~ dcat(p.att[t,])
        grp.def[t] ~ dcat(p.def[t,])
        att[t] ~ dt(mu.att[grp.att[t]],tau.att[grp.att[t]],4)
        def[t] ~ dt(mu.def[grp.def[t]],tau.def[grp.def[t]],4)
        att.star[t] <- att[t] - mean(att[])
        def.star[t] <- def[t] - mean(def[])

# Priors on the mixture parameter (team specific)
        p.att[t,1:3] ~ ddirch(prior.att[t,])
        p.def[t,1:3] ~ ddirch(prior.def[t,])
}

# Priors on the random effects
# group 1: bottom-table teams
    mu.att[1] ~ djl.dnorm.trunc(0,0.001,-3,0)
    mu.def[1] ~ djl.dnorm.trunc(0,0.001,0,3)
    tau.att[1] ~ dgamma(0.01,0.01)
    tau.def[1] ~ dgamma(0.01,0.01)
# group 2: mid-table teams
    mu.att[2] <-0
    mu.def[2] <-0
    tau.att[2] ~ dgamma(0.01,0.01)
    tau.def[2] ~ dgamma(0.01,0.01)
# group 3: top-table teams
    mu.att[3] ~ djl.dnorm.trunc(0,0.001,0,3)
    mu.def[3] ~ djl.dnorm.trunc(0,0.001,-3,0)
    tau.att[3] ~ dgamma(0.01,0.01)
    tau.def[3] ~ dgamma(0.01,0.01)
}
```

References

Aitchinson, J. & Ho, C. (1989), 'The multivariate poisson-log normal distribution', *Biometrika* **76**, 643–653.

Berger, J. (1984), The robust Bayesian point of view, *in* J. Kadane, ed., 'Robustness of Bayesian analysis', North Holland, Amsterdam, Netherlands.

Bernardo, J. M. & Smith, A. (1999), *Bayesian Theory*, John Wiley and Sons, New York, NY.

Chib, S. & Winkelman, R. (2001), 'Markov Chain Monte Carlo Analysis of Correlated Count Data', *Journal of Business and Economic Statistics* **4**, 428–435.

Congdon, P. (2003), *Applied Bayesian Modelling*, John Wiley and Sons, Chichester, UK.

Dixon, M. & Coles, S. (1997), 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society* C **46**, 265–280.

Karlis, D. & Ntzoufras, I. (2000), 'On modelling soccer data', *Student* **3**, 229–244.

Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate Poisson models', *Journal of the Royal Statistical Society* D **52**, 381–393.

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

Lunn, D. (2008), '`WinBUGS` code for the truncated normal distribution', Documentation and code available online.
URL: *http://www.winbugs-development.org.uk/shared.html*

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

Pollard, R., Benjamin, P. & Reep, C. (1977), Sport and the negative binomial distribution, *in* S. Ladany & R. Machol, eds, 'Optimal Strategies in Sports', North Holland, New York, NY, pp. 188–195.

Tsionas, E. (2001), 'Bayesian Multivariate Poisson Regression', *Communications in Statistics – Theory and Methodology* **30(2)**, 243–255.

Tunaru, R. (2002), 'Hierarchical bayesian models for multiple count data', *Austrian Journal of Statistics* **31**, 221–229.