Bayesian Sports Modelling Handout (Final Year MMath Mathematics Project)

Ryan Chan

Overview

I consider the task of predicting football results in the Premier League and propose a Bayesian hierarchical model. The model will be implemented using R and the Bayesian inference software Stan, which automatically implements the Hamiltonian Monte Carlo algorithm. Further, we look at different techniques to assess model performance and use these to compare our model with Baio & Blangiardo's model (2010).

Negative Binomial Model

The use of the negative binomial distribution to model the number of goals scored by a team has been proposed by several authors but has been largely ignored in relevant literature.

In many models that have been proposed previously, an independent Poisson distribution was assumed to model goals scored by the home and away team. But by doing some exploratory analysis, there were other distributions that had a better fit to the data and in the end I chose to use the negative binomial.

To model the number of goals scored by two teams in a match, this model will use the negative binomial distribution. The negative binomial distribution is the distribution of the number of successes in a sequence of independent Bernoulli trials before n failures, with probability mass function defined as

$$p(x) = \frac{\Gamma(x+n)}{\Gamma(n)x!}p^n(1-p)^x = \frac{(x+n-1)!}{(n-1)!\,x!}p^n(1-p)^x$$

for $x = 0, 1, 2, ..., n$, $n > 0$, $0 < p \leq 1$ and where, $\Gamma(x)$ is the Gamma function, defined by $\Gamma(x) = (x-1)!$, where x is an integer. If a random variable X follows a negative binomial distribution with size n and probability p, then we write $X \sim NB(n, p)$.

In order to propose prior distributions on random variables, I use an alternative parametrisation of the negative binomial distribution, in terms of its mean μ and size n, which is used in Stan, as seems more natural to view the mean for modelling goals, whereas to try and propose prior distributions for the probability p of scoring a goal is less intuitive.

The negative binomial distribution has mean $\mu = \frac{n(1-p)}{p}$, then by rearrangement, one can find that $p = \frac{n}{n+\mu}$ and hence the probability mass function is now given by

$$p(x) = \frac{(x+n-1)!}{(n-1)!\,x!}\Big(\frac{n}{n+\mu}\Big)^n\Big(\frac{\mu}{n+\mu}\Big)^x.$$

Let y_{g1} and y_{g2} to denote the number of goals scored by the home and away team in the g-th game of the season, respectively. Here, the vector of observed goals, $\boldsymbol{y} = (y_{g1}, y_{g2})$ are modelled using a independent negative binomial distribution,

$$y_{gj} \mid \mu_{gj}, n_j \sim \mathrm{NB}(\mu_{gj}, n_j),$$

where $\boldsymbol{\mu} = (\mu_{g1}, \mu_{g2})$ represents the mean number of goals expected to be scored by the home team ($j = 1$) and the away team ($j = 2$) in the g-th game of the season. We assume a log-linear random effect model, as it allows for the condition that the mean number of goals must be positive:

$$\log \mu_{g1} = home\_att_{h(g)} + away\_def_{a(g)}$$
$$\log \mu_{g2} = away\_att_{a(g)} + home\_def_{h(g)}$$

These parameters are indexed by $h(g)$ and $a(g)$, which identify the team that is playing home or away in the g-th game of the season. In this model, the prior distributions for the home and away parameters for the attacking and defensive strengths of each team, $t = 1, ..., T$, where T is the number of teams, are

$$home\_att_t \sim \text{Normal}(\mu_{h\_att}, \sigma^2_{att}),$$
$$away\_att_t \sim \text{Normal}(\mu_{a\_att}, \sigma^2_{att}),$$
$$home\_def_t \sim \text{Normal}(\mu_{h\_def}, \sigma^2_{def}),$$
$$away\_def_t \sim \text{Normal}(\mu_{a\_def}, \sigma^2_{def}).$$

To impose identifiability constraints on these parameters, we use a sum-to-zero constraint, that is,

$$\sum_{t=1}^{T} home\_att_t = 0, \sum_{t=1}^{T} away\_att_t = 0, \sum_{t=1}^{T} home\_def_t = 0, \sum_{t=1}^{T} away\_def_t = 0.$$

Then the prior distributions for the hyperparameters are given as follows:

$$\mu_{h\_att} \sim \text{Normal}(0.2, 1),$$
$$\mu_{a\_att} \sim \text{Normal}(0, 1),$$
$$\mu_{h\_def} \sim \text{Normal}(-0.2, 1),$$
$$\mu_{a\_def} \sim \text{Normal}(0, 1).$$

where the slight difference in means for the home parameters are used to try to encode a belief that teams tend to play better at home. The prior distributions for the variance of the attack and defence parameters of the model are

$$\sigma^2_{att} \sim \text{Gamma}(10, 10),$$
$$\sigma^2_{def} \sim \text{Gamma}(10, 10).$$

Lastly, the prior distributions for the size n in the model are,

$$n_{home} \sim \text{Gamma}(2.5, 0.05),$$
$$n_{away} \sim \text{Gamma}(2.5, 0.05).$$

A graphical representation of this model is shown in Figure 1.

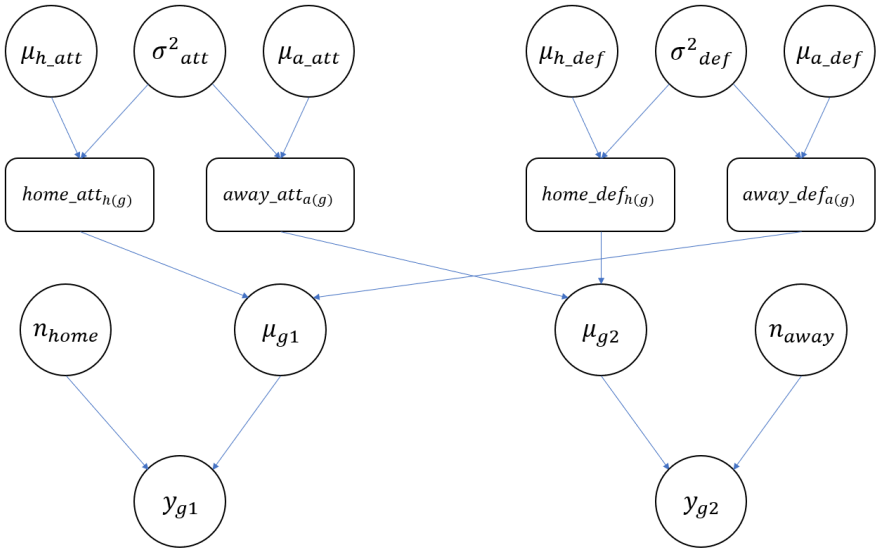

Figure 1: The DAG representation of the Negative-Binomial Model

Results

Using the data from the 2017/18 premier league season, we can get estimates for the attack and defence parameters for each team.

Higher attack parameter $\implies$ better attacking ability.
Higher defence parameter $\implies$ worse defending ability.

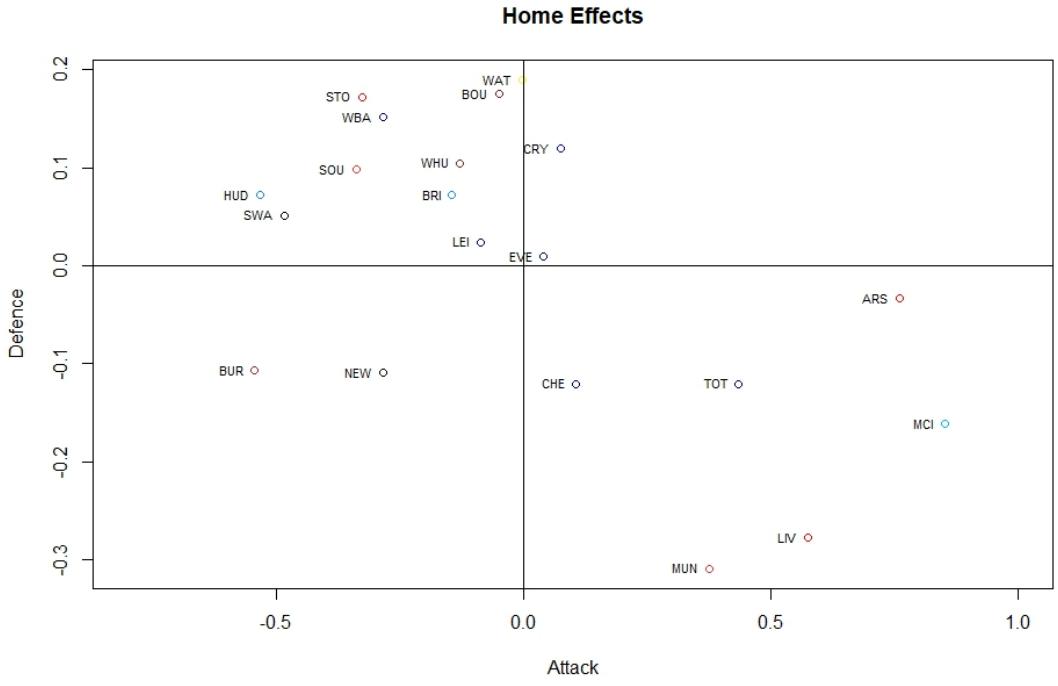

Figure 2: A plot of the posterior means of the home effects for each team

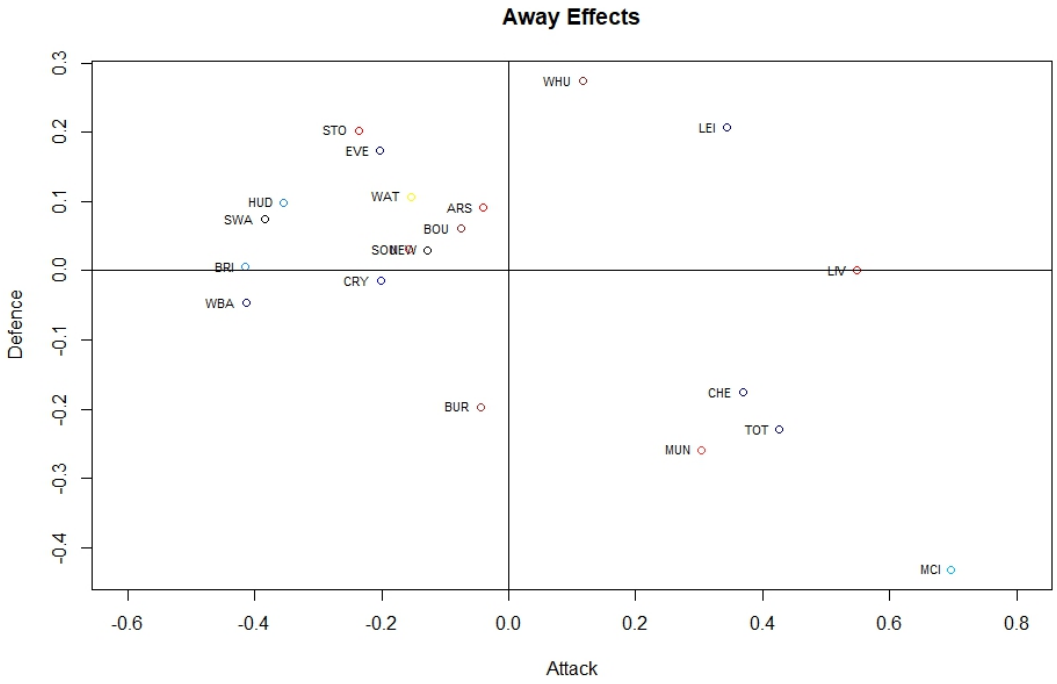

Figure 3: A plot of the posterior means of the away effects for each team

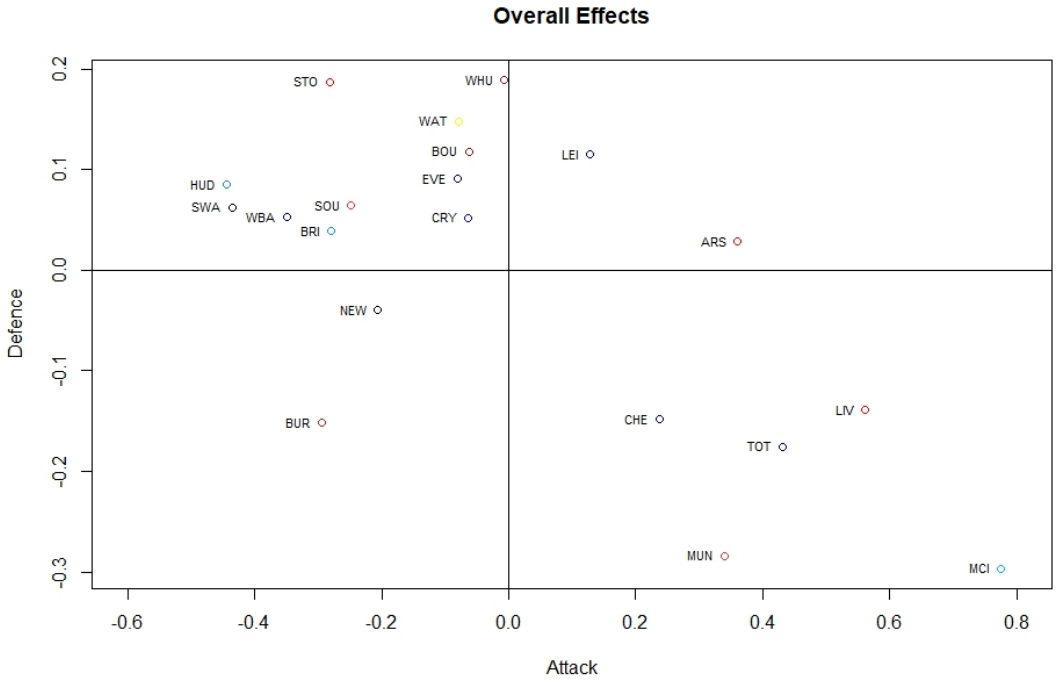

Figure 4: A plot of the averages of posterior means of the home and away effects for each team

Predicting Tottenham vs. Leicester

I use the Stan programming language and R to implement the model. Stan uses Hamiltonian Monte Carlo to obtain a sample from a target density (a posterior density in our case).

Once we have a sample from the posterior distribution, $\pi(\theta \mid y)$, we can draw from predictive distribution of unobserved data or future data, y^*. For each draw of θ, we draw a sample y^* from the predictive distribution $\pi(y^* \mid \theta)$.

In the case of one football match, we take our samples for the attack and defence parameters for each team and n_j for $j = 1, 2$, and then simulate from a Negative Binomial distribution for each sample.

Using the example for predicting the game between Tottenham and Leicester, we first remove this game from our data and obtain parameter estimates. Then we obtain μ_1 and μ_2 by

$$\begin{aligned} \log \mu_1 &= home\_att_{TOT} + away\_def_{LEI}, \\ \log \mu_2 &= away\_att_{LEI} + home\_def_{TOT}. \end{aligned}$$

Next, we obtain draws from our likelihood, $\pi(y_j^* \mid \mu_j, n_j)$, for $j = 1, 2$, (from a negative binomial distribution) and then we estimate the probabilities as

$$\begin{aligned} \Pr(\text{Tottenham Win}) &= \frac{\text{Number of times } y_1^* > y_2^*}{\text{Number of samples}}, \\ \Pr(\text{Draw}) &= \frac{\text{Number of times } y_1^* = y_2^*}{\text{Number of samples}}, \\ \Pr(\text{Leicester Win}) &= \frac{\text{Number of times } y_1^* < y_2^*}{\text{Number of samples}}. \end{aligned}$$

From using the data from the 2017/18 season, the estimate of the probabilities were

$$\begin{aligned} \Pr(\text{Tottenham Win}) &= 0.509, \\ \Pr(\text{Draw}) &= 0.235, \\ \Pr(\text{Leicester Win}) &= 0.256. \end{aligned}$$

Also we can calculate the probabilities for each score line and obtain a table such as:

		Tottenham						
		0	1	2	3	4	5	6+
Leicester	0	0.048	0.081	0.076	0.049	0.024	0.011	0.006
	1	0.055	0.093	0.088	0.056	0.028	0.012	0.007
	2	0.036	0.060	0.056	0.036	0.018	0.008	0.005
	3	0.016	0.027	0.025	0.016	0.008	0.004	0.002
	4	0.006	0.009	0.009	0.006	0.003	0.001	0.001
	5	0.002	0.003	0.003	0.002	0.001	0.000	0.000
	6+	0.000	0.001	0.001	0.000	0.000	0.000	0.000

Table 1: Score probabilities for Tottenham vs. Leicester

And so the most probable scoreline for this game according to this model was a 1-1 draw.

Outcome: Tottenham 5-4 Leicester.

Model Assessment

We calculate the cross-validation score, Brier score, rank probability score and the profit/loss from betting £10 on the most probable outcome for the two models for the 2017/18 Premier League season.

Model	Cross-Validation	Brier score	Average RPS	Profit/Loss
BB	57.8%	0.532	0.173	£449.9
NB	59.7%	0.540	0.177	£873.3

Table 2: Results and comparison of the Negative Binomial model to Baio & Blangiardo's model (2010)

The breakdown of the profit/loss returned by the games for each model is shown in the following frequency table:

Profit/Loss (PL)	Frequency
-10.00 (lost bet)	156
$0 \leq PL < 10$	132
$10 \leq PL < 20$	63
$20 \leq PL < 30$	13
$30 \leq PL < 40$	2
$40 \leq PL < 50$	3
$PL \geq 50$	1

(a) Baio & Blangiardo's model

Profit/Loss	Frequency
-10.00 (lost bet)	149
$0 \leq PL < 10$	133
$10 \leq PL < 20$	49
$20 \leq PL < 30$	27
$30 \leq PL < 40$	7
$40 \leq PL < 50$	3
$PL \geq 50$	2

(b) Negative Binomial model

Table 3: Frequency of each profit/loss for each model in £s

Discussion

Strengths

- By simply just using previous goals data, we are able to achieve a good model for prediction of football matches.
- By assessing the model's usefulness as a basis of a decision rule for betting, it was able to turn a profit - has real world applications.
- By splitting up the attack and defence parameters for home and away and not using a constant home-advantage parameter as Baio & Blangiardo (2010), Dixon & Coles (1997), Lee (1997), Maher (1982), we are able to encode more information on each team's performances.

Weaknesses

- Although the model was able to turn a profit, there was still 149 games where the model incorrectly predicted the outcome of the game ($\approx 40.3\%$), so there is still a lot of room for improvement.
- The model only uses goals to obtain estimates for parameters for each team.

  - Goals may not the best indicator for how well a team is performing - teams can be lucky or unlucky.
  - Possibly by incorporating more data, we can obtain more accurate estimates for the attack and defence parameters for each team.

- Model ignores other possible factors that can affect team performance, for example:
  - injury/resting of star players
  - fatigue of players / number of days rest between games
  - distance travelled for away team
  - effect of managerial changes

Conclusion

To summarise:

- We built a Bayesian hierarchical model for prediction of football results, which used a negative binomial distribution to model the goals scored by each team.
- By using several techniques, there was not much difference between the negative binomial model and Baio & Blangiardo's model in terms of accuracy for predicting the 2017/18 Premier League season games.
  - But the model was far superior when using it as a basis for a betting decision rule and gave a much higher profit return.
- In the future, I wish to carry on with this model by incorporating more covariates into the model in the hope that they can help produce better estimates for the performance levels of each team.