MASH Department

Bayesian modelling of football outcomes

Professors:

Robin Ryder

Julien Stoehr

Guillaume Kon Kam King

Student:

Gian Mario Sangiovanni

Academic Year 2023/2024

Contents

1 Introduction

In the ever-evolving realm of sports analytics, the application of **Bayesian modeling** has emerged as a powerful tool for unraveling the intricate tapestry of football outcomes [1], [5], [11], [8]. Originating in the fields of Victorian England [2], football has since evolved into a global spectacle, captivating audiences with its unpredictability, which has always fascinated enthusiasts, analysts, and statisticians alike. Leveraging Bayesian inference techniques, this project delves into the depths of football dynamics, aiming to demystify the uncertainties surrounding match predictions and player performances. As the sporting world adopts increasingly sophisticated approaches to gain a competitive edge, Bayesian modeling [17] offers a unique perspective by seamlessly incorporating prior knowledge and updating it with new data.

Football, with its multifaceted variables and inherent unpredictability, poses a fascinating challenge for statistical modeling. Traditional frequentist models often fall short in capturing the nuanced interactions within a game. Here, we turn to the Bayesian paradigm, acknowledging and embracing uncertainty as an integral part of the footballing narrative. While my focus lies in the mathematical intricacies of Bayesian modeling, it is indeed possible to recognize the richness of football as a human experience. This project transcends the numerical realm to appreciate the emotional narratives, tactical innovations, and the sheer unpredictability that makes every match a spectacle. As you navigate through the complexities of football outcomes, I invite you to join me on this exploration, where equations and distributions meet the roar of the crowd, and where probabilities dance in tandem with the elegant movements of the "bomber" on the pitch. In the pages that follow, I unveil the methodology used, discuss the implementation (Appendix B), and share the insights gained from applying Bayesian modeling to football. In Section 2, a descriptive analysis of the selected dataset is provided, along with general ideas about the distributions that can be used to fit the number of goals scored by home and away teams in a match. A question arises:

Why do i choose **English Premier League**?

As specified before, the very roots of football trace back to the green fields of Victorian England, where the working class, fueled by a desire for recreation, gave birth to an unstructured chaotic version of the sport. The formalization of rules, notably the Cambridge Rules of 1848 [3], marked a pivotal moment in the evolution of football. The establishment of standardized regulations facilitated the spread of the game, transforming it from a local pastime into an organized sport. English football clubs, many of which boast a heritage dating back to these formative years, became the crucibles of innovation, refining tactics and shaping the beautiful game into the structured sport recognized today.

Section 3 introduces useful tools for analysing MCMC algorithms and the theory related to the "static" models that will be used. The term "static" refers to the assumption that time is not taken into account and all matches are considered exchangeable, which is a strong assumption. Section 4 is dedicated to presenting the results obtained by applying the models defined in the previous section. Specifically, I will attempt to answer some open research questions regarding the parameters involved in the model. Section 5 is expressly allocated to the theoretical elucidation and exposition of the results associated with the Dynamic model. This comprehensive treatment encompasses considerations within the realm of discrete time, rather than continuous time. Lastly, Section 6 tries to change perspective with a huge focus on the goals difference modelling.

2 Descriptive analysis

2.1 Main ideas

The dataset pertains to the **English Premier League** (EPL) data from the 2018/2019 season, sourced from the public repository *footystats*. A generic match is organized as follows:

Ht	**At**	**Attend**	**GW**	**Htg**	**Atg**	**Htc**	**Atc**	**Htyc**
Manchester City	Liverpool	54511	21	2	1	2	1	4
Atyc	**Htrc**	**Atrc**	**Htf**	**Atf**	**Htp**	**Atp**	**Hts**	**Ats**
2	0	0	12	7	49	51	9	8

Table 2.1: Generic row of the English Premier league dataset.

where:

- *Ht* and *At* represent respectively the home and the away teams;
- *Attend* represents the number of spectators to the match;
- *GW* represents the week of the game;
- *Htg* and *Atg* represent respectively the goals scored by the home and away teams;
- *Htc* and *Atc* represent respectively the number of corners taken by the home and away teams;
- *Htyc* (*Htrc*) and *Atyc* (*Atrc*) represent the number of yellow (red) cards taken by the home and away teams;
- *Htf* and *Atf* represent respectively the number of fouls committed by the home and away teams;
- *Htp* and *Atp* represent the percentage of ball possession for the home and away teams;
- *Hts* and *Ats* represent the number of shots of the the home and away teams.

In addition to these factors, two additional variables have been determined indicating the rank of the respective teams prior to the match. It is of particular interest to consider this information as a significant and impressive confrontation occurred between Manchester City and Liverpool during the season, whereby the former finished with 98 points and the latter with 97. In the league, there are twenty teams that have played for a total of 380 games over thirty-eight non-consecutive weeks. The first match, which involved Manchester United and Leicester City, kicked off on 10th August, whereas the last match between Watford and West Ham United was played

on 12th May. It is noteworthy that the top scorers were Aubameyang (A), Manè (L), and Salah (L) who have each scored twenty-two goals. A final comment pertains to the observation that LIVERPOOL has lost solely one match (Table 2.1).

Pos	Team	Pld	W	D	L	GF	GA	Pts
1	Manchester City (C)	38	32	2	4	95	23	98
2	Liverpool	38	30	7	1	89	22	97
3	Chelsea	38	21	9	8	63	39	72
4	Tottenham Hotspur	38	23	2	13	67	39	71
5	Arsenal	38	21	7	10	73	51	70
6	Manchester United	38	19	9	10	65	54	66
7	Wolverhampton Wanderers	38	16	9	13	47	46	57
8	Everton	38	15	9	14	54	46	54
9	Leicester City	38	15	7	16	51	48	52
10	West Ham United	38	15	7	16	52	55	52
11	Watford	38	14	8	16	52	59	50
12	Crystal Palace	38	14	7	17	51	53	49
13	Newcastle United	38	12	9	17	42	48	45
14	Bournemouth	38	13	6	19	56	70	45
15	Burnley	38	11	7	20	45	68	40
16	Southampton	38	9	12	17	45	65	39
17	Brighton & Hove Albion	38	9	9	20	35	60	36
18	Cardiff City (R)	38	10	4	24	34	69	34
19	Fulham (R)	38	7	5	26	34	81	26
20	Huddersfield Town (R)	38	3	7	28	22	76	16

Table 2.2: English Premier League Table 2018/2019. The first four teams have qualified for the Champions League group stage, the next three teams have qualified for the Europa league group stage, while the last three teams were relegated to the Football League Championship.

Despite of those interesting stats, my goals are:

1. to understand which are the factors that can influence the outcome of an English Premier League football match;

2. to understand whether or not it is possible to predict match outcomes with a certain degree of certainty.

First of all, let's examine the number of goals scored by home and away teams. Based on the two graphs presented, it is evident that the number of goals scored during a football match ranges from 0 to 6, however, teams are unlikely to score over 3 goals per game. Additionally, away teams are more inclined to score 0 goals as opposed to home teams. The occurrence of the '{*team scores 0 goals*}' event is particularly intriguing to examine. A comparison of team performances during both home and away matches for goals scored can be inferred from the information presented in the two graphs below (Figure 2.1). Objective data pertaining to the distribution of goals scored in the top five leagues at home and away locations is provided in the appendix A. The data was sourced from the website Sports-Statistics.

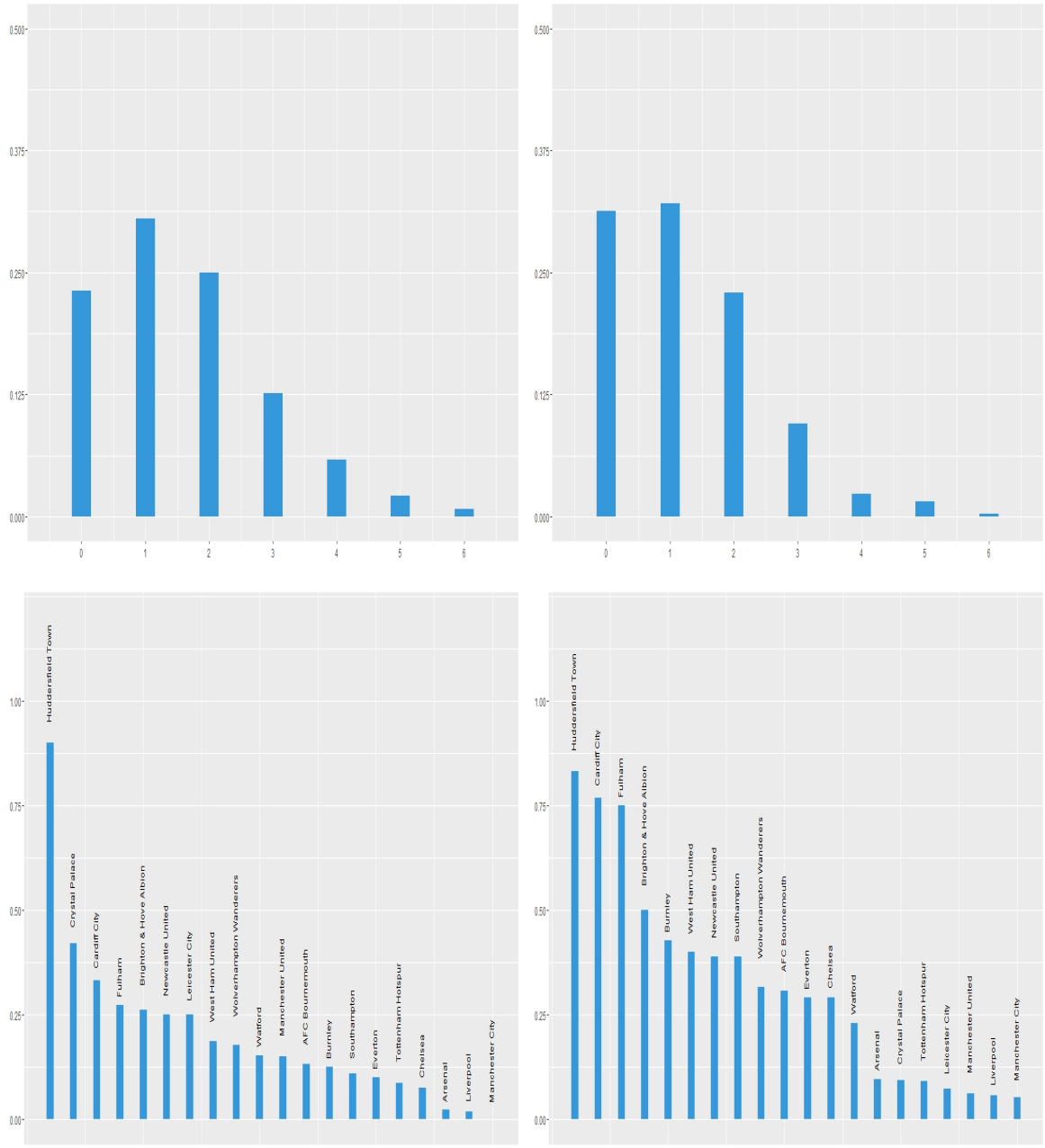


Figure 2.1: Relative frequency of the goals scored by the home (top left) and away (top right) teams. Relative frequency of zero scored goals for the home team (bottom left) and for the away team (bottom right).

As a side note, it is worth mentioning that Huddersfield Town, Cardiff City and Fulham are the teams that have the most frequent occurrence of no goals being scored during home and away matches, whereas Manchester City and Liverpool have the least. Additionally, Figure 2.2 displays a clear division in stadium attendance between top teams who possess large contemporary stadiums and middle to lower teams with smaller ones. It is unclear whether the capacity of a stadium or the passion of fans has a greater impact on the number of goals scored [21]. Passion refers to the emotional attachment that fans feel for their team. This phenomenon is typically associated with small and medium-sized clubs, although England is a notable exception

due to the prevalence of hooliganism (though this is not the case for Manchester City and Newcastle United). This observation suggests that a "home" effect may exist, which could affect the scoring rates of both the home and away teams.

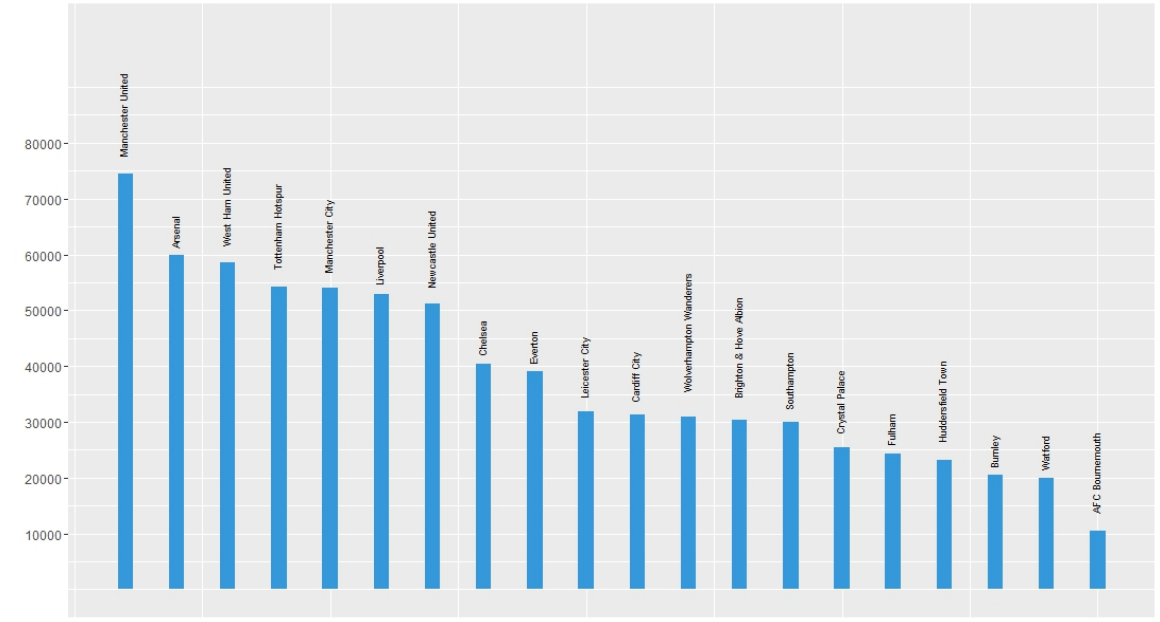


Figure 2.2: Average Attendance per team during the English Premier League season 2018/2019. Manchester United is the team with the highest average attendance while AFC Bournemouth is the one with the lowest average attendance

Based on the previous analysis, the stadium attendance was divided into two classes using the quantile at level 0.66. The first class includes lower and middle clubs with small stadiums, while the second class includes bigger teams. A **Kruskal** test was conducted, resulting in a p-value of 1.034×10^{-06}, indicating a significant difference in the median of goals scored by home teams based on stadium attendance.

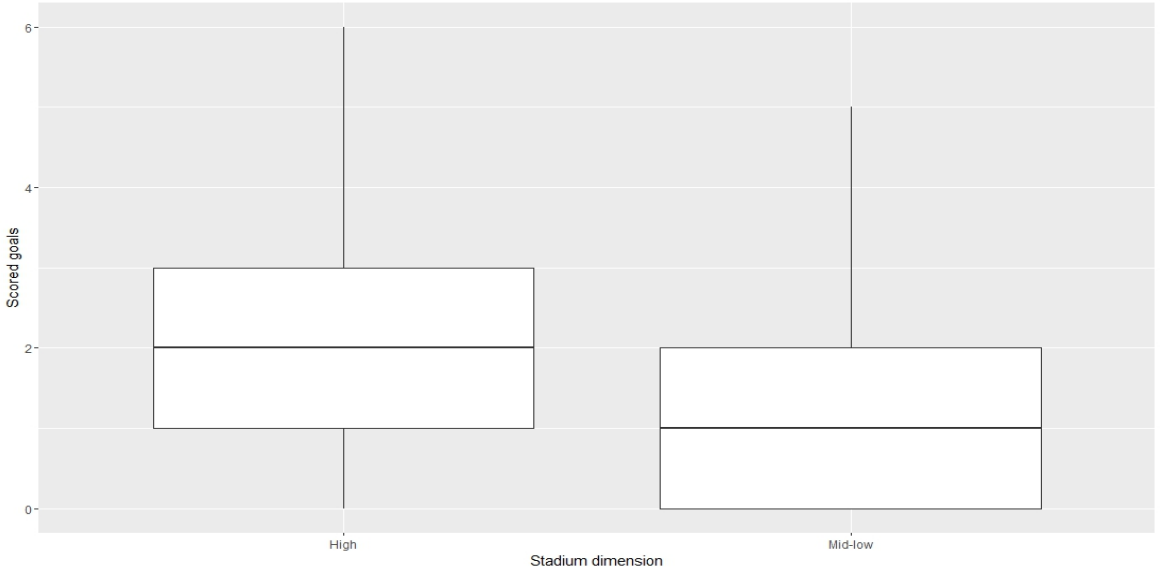


Figure 2.3: Goals scored by the home teams with respect to the size of their stadium.

To gain a deeper understanding of this effect, it is essential to conduct a thorough analysis of the home and away performances of various teams. As suggested by [22], the home and away strength can be measured using the following method:

1. *Home attacking strength = average home scored goals / league average*;
2. *Home defensive strength = average home conceded goals / league average*;
3. *Away attacking strength = average away scored goals / league average*;
4. *Away defensive strength = average away conceded goals / league average.*

Define the general strength as the vector which takes into account the four indexes. For instance, giving a look to the Figure 2.4, it is possible to observe two different situations:

1. on one hand, Manchester City strength is $(1.9127517, 0.5042017, 1.5966387, 0.3691275)$. Offensively, it seems that Manchester City tends to performer better at home, while defensively it seems that it performs better when away from **Etihad**;
2. on the contrary, the strength of Huddersfield town is $(0.3355705, 1.3025210, 0.5042017$ $1.5100671)$. It is noteworthy that the team is inclined towards scoring better on their away games, while manifesting a stronger home defense (conquering **Kirklees** is no mean feat).

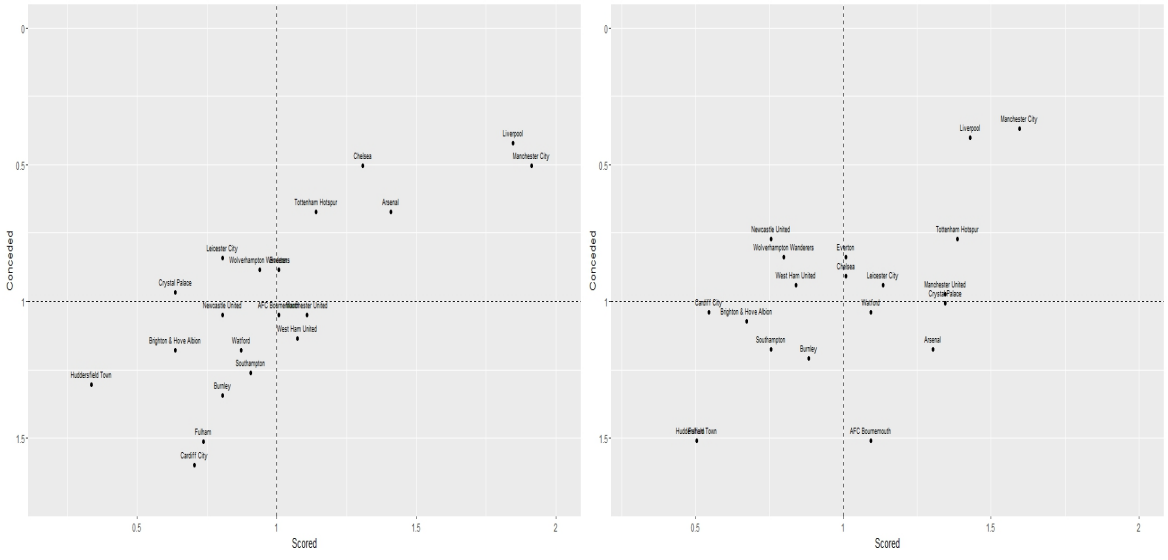


Figure 2.4: Home strength (left). Away strength (right). On the x-axis is plotted the offensive one, while on the y-axis is plotted the defensive.

In predicting the outcome of a football match, there are several factors that may influence it, beyond the home advantage as previously mentioned [19]. One among those factors is the team's playing style, which is dependent on the coach's ideas and

may vary throughout the season. On one hand, a team that adopts a defensive strategy is more likely to secure a draw or a narrow victory, whereas an offensive team may score more goals but at the risk of conceding more. On the other hand, a team that prefers ball possession over allowing opponents to have it can create more goal-scoring opportunities and be more dangerous. Additionally, this strategy limits the opponents' ability to express their own style of play, resulting in fewer defensive risks.

Using ball possession as an indicator of playing style, Figure 2.5 analysis shows that top-ranked teams typically prefer to maintain possession of the ball and assert their dominance over the game. Conversely, lower-ranked teams tend to adopt a more defensive approach. Supporting this theory, there is a correlation of approximately 0.22 between the goals scored by the home team (away team) and the percentage of ball possession maintained during the game by the home team (away team). [1].

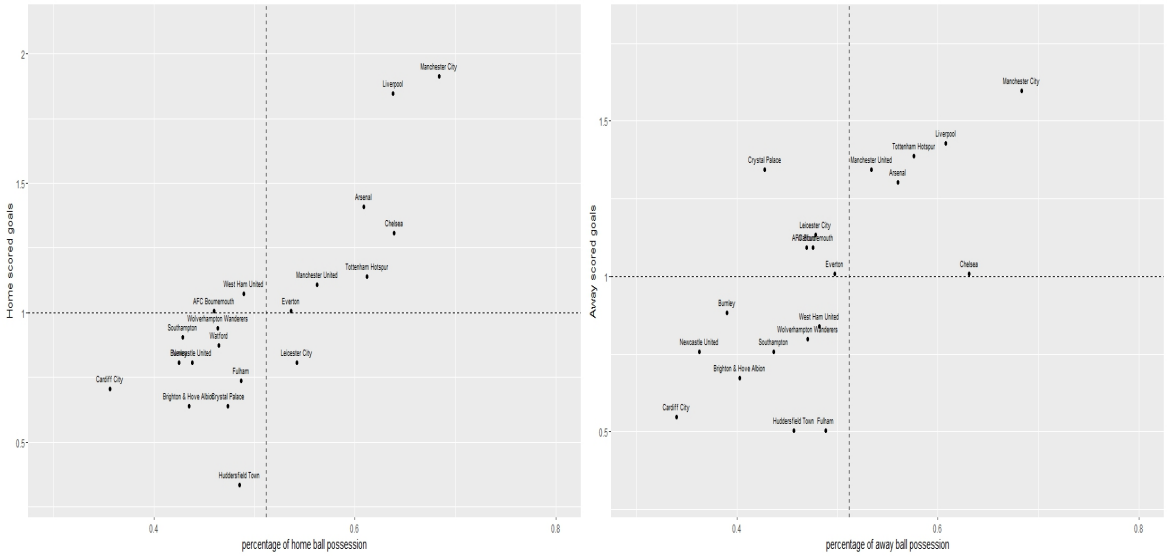


Figure 2.5: Average percentage of ball possession at home vs average home scored goals (left). Average percentage of ball possession away vs average away scored goals (right)

By using the number of corners and shots taken as indicators of a team's performance, it is possible to gain insight into the situation of different teams (Figure 2.6). Generally, higher-ranked teams take more corners and shots towards the goal, resulting in fewer goals conceded. Conversely, lower-ranked teams tend to take fewer corners and shots, leading to more goals conceded. The data shows a correlation of approximately 0.37 between the number of goals scored by the home team (away team) and the number of shots. However, there is only a weak correlation of approximately 0.05 between the number of goals scored by the home team (away team) and the number of taken corners.

[1] In the appendix, it is possible to find the *correlogram*

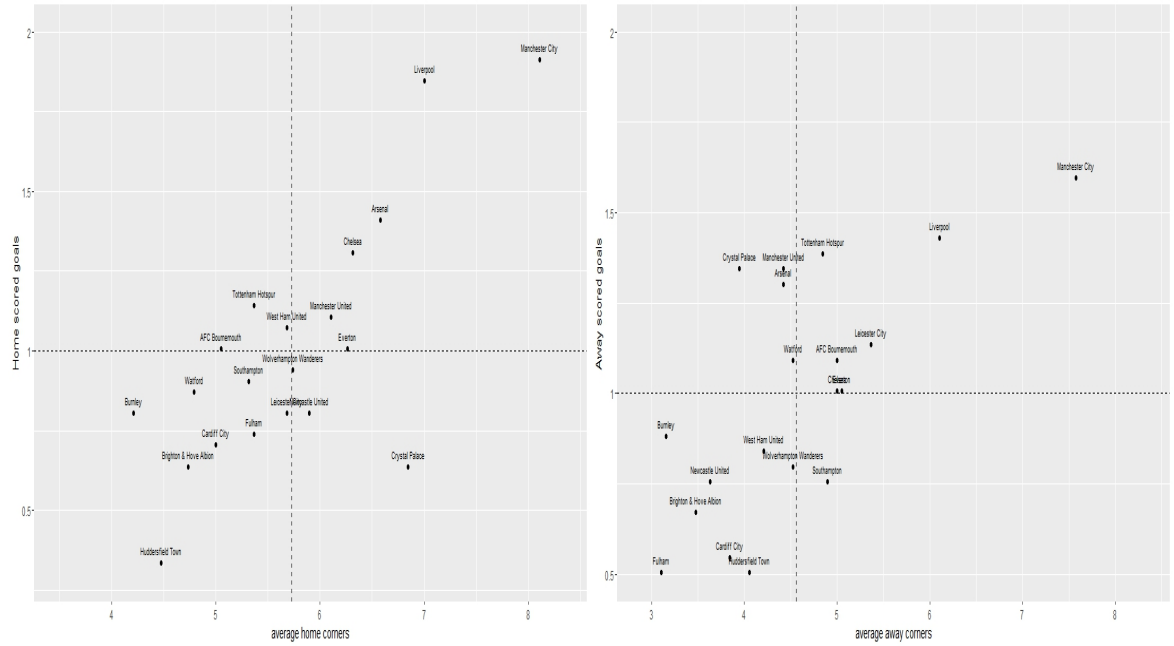


Figure 2.6: Average home taken corners vs average home scored goals per team (left). Average away taken corners vs average away scored goals per team (right).

Discipline can have a significant impact on a team's performance, both individually and collectively. To measure this, the number of yellow and red cards, as well as the number of fouls committed by the team during the match, should be considered. The Table 2.3 demonstrates a low negative correlation between a team's goals scored and its discipline.

goals/card	yellow	red
home	−0.115	−0.037
away	0.0264	−0.051

goals	fouls
home	−0.061
away	−0.029

Table 2.3: Correlation between the number of taken cards (either yellow or red) and the number of scored goals (either scored by the home or away team (left). Correlation between the number of fouls committed by the team and the number of goals scored by the team (right).

As can be seen from Appendix A, the situation is now a little more complicated to interpret. For example, if I consider only the top 5 teams, it can be seen that while Liverpool and Manchester City concede and commit few fouls, Chelsea tends to concede more and still commit few while Arsenal tends to concede and commit more. Tottenham stands at the league average for both items. As far as cards are concerned, the trend remains valid. Note how *Arsenal* tends to be an undisciplined team.

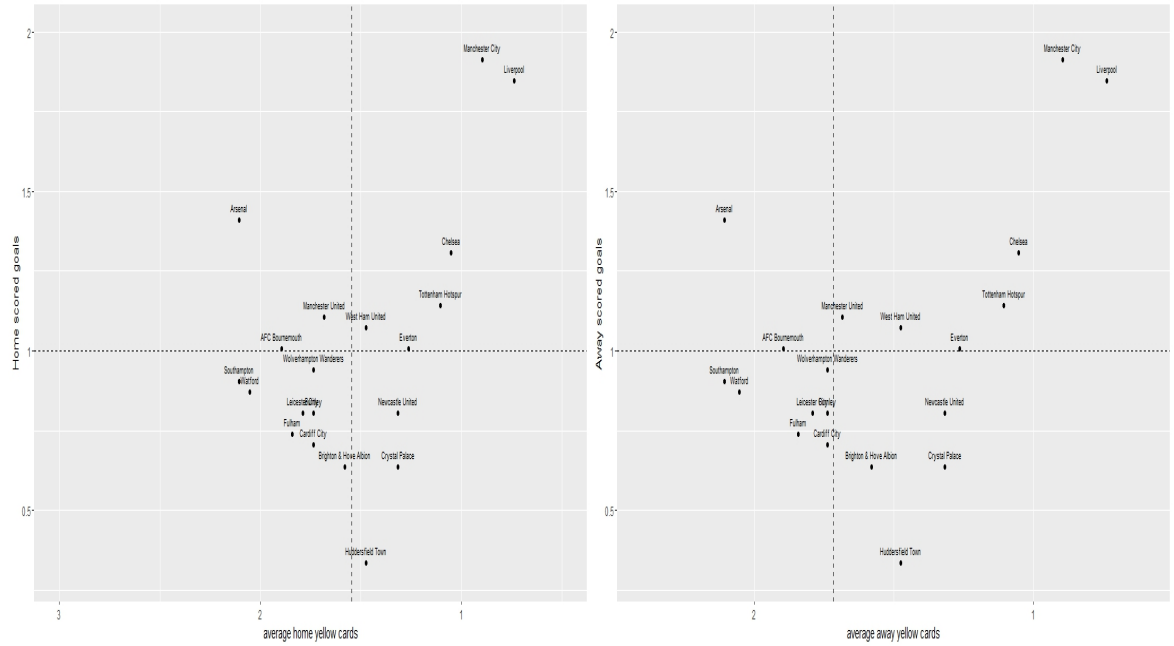


Figure 2.7: Average home yellow cards vs average home scored goals (left). Average away yellow cards vs average away score goals (right).

One last factor that can be considered is the **overall shape** of a team, which may in turn be influenced by several factors regarding *injuries*, *team mentality*, *team tiredness*, *changes in team composition*, etc. However due to the lack of data, it is not so easy to define it into a proper way but it can be regarded as a sort of latent effect which affects the performances of teams. Several indicators can be defined to measure a team's condition at the time of a match [2].

- *Recent scoring condition of a team* (RSC). Average goals scored by a team over the past 5 weeks;

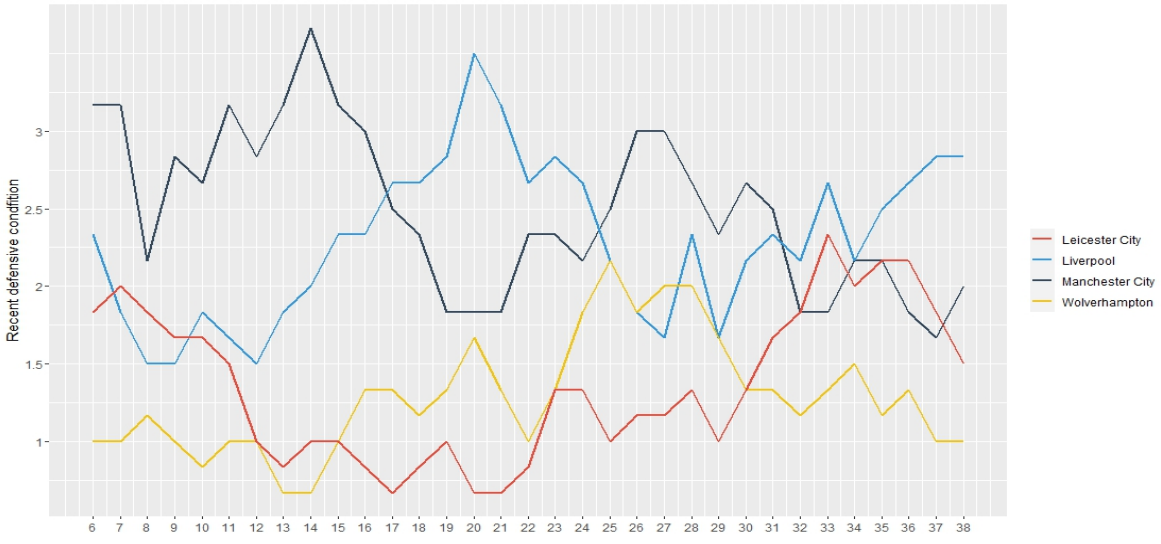


Figure 2.8: Each line represents the rolling mean of order 5 of the scored goals by a team during the season.

[2] usually 5 is chosen as an upper bound to reflect the condition of a team.

- *Recent defensive condition of a team* (RDC). Average conceded goals by a team over the past 5 weeks.

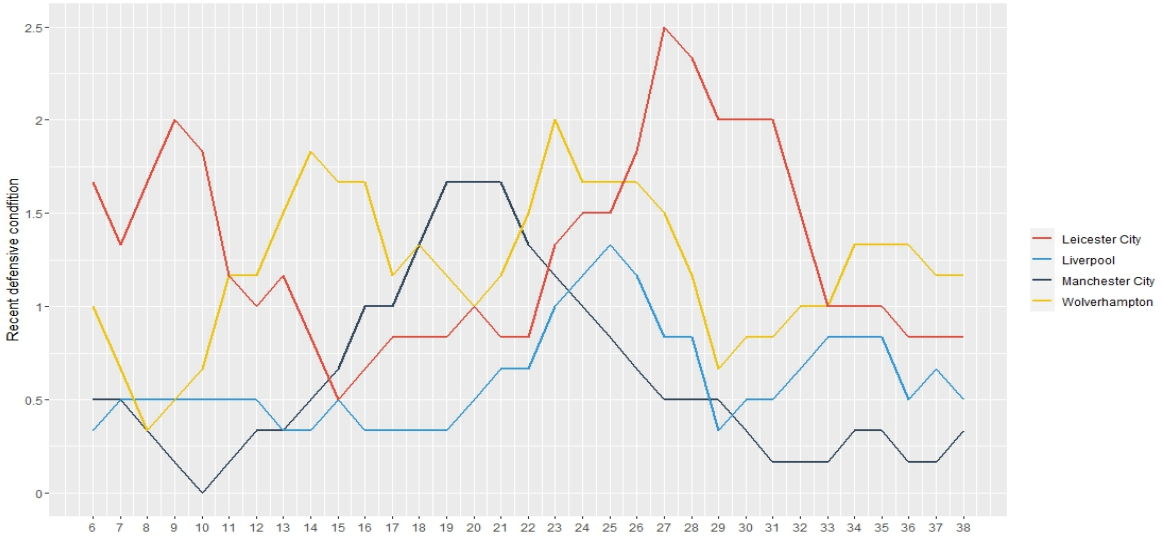


Figure 2.9: Each line represents the rolling mean of order 5 of the conceded goals by a team during the season.

- *Quality of opponents* (QO). Average obtained points by all the opponents of a team in the past 5 weeks.

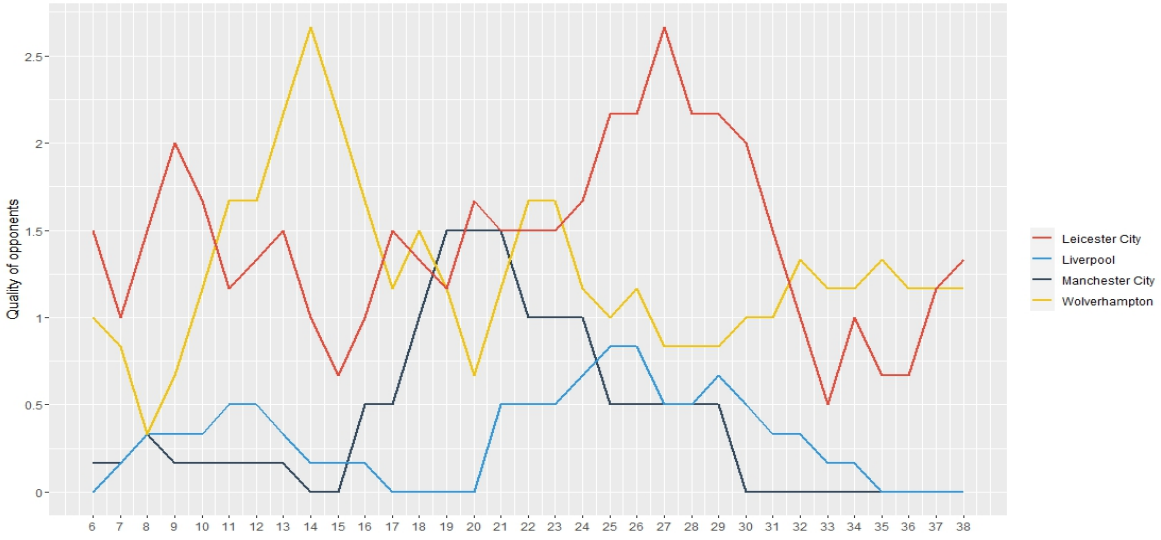


Figure 2.10: Each line represents the rolling mean of order 5 of the points obtained by the opponents of a team during the season.

- *Elo rating of a team* (ELO)[9]. It is based on a generalization of the Elo system used by the FIDE[3] in the chess world[6]. The authors of the paper said that:

  > We do not make use of the FIFA ranking (which is simply Elo ranking since July 2018), because the calculation of the FIFA ranking changed

[3]Fédération Internationale des Échecs

over time and the Elo ranking is more widely used in football forecast models.

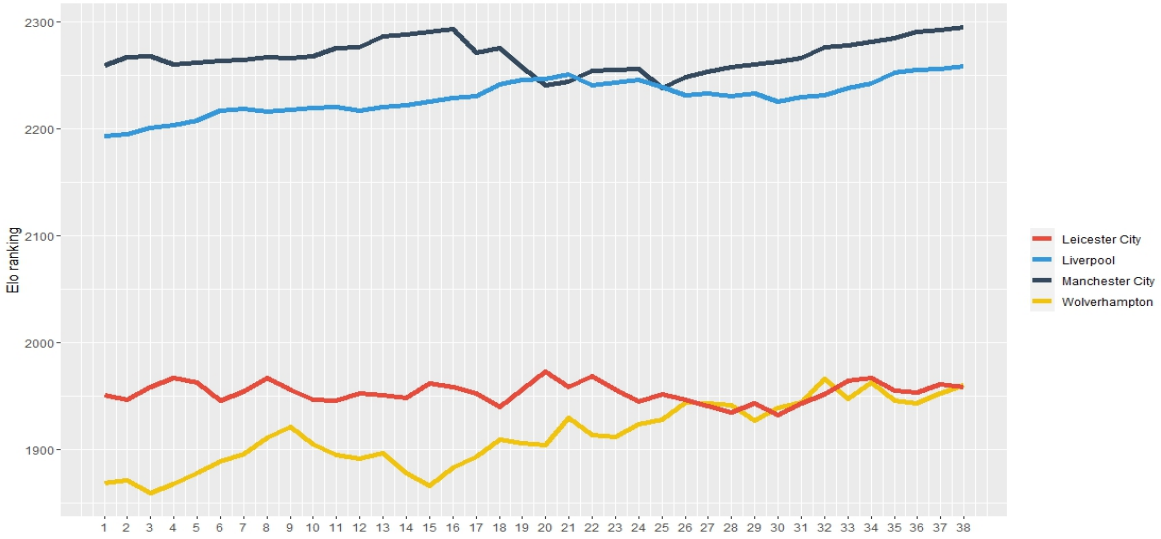


Figure 2.11: Each line represents the elo points before each game for a team during the season.

To adapt it to sports needs, Elo rating can be reformulated in this new version:

$$\Delta ELO = K \cdot G\,(W - W_\epsilon)$$

where:

- $\Delta ELO = ELO_{after} - ELO_{before}$.
- K is a weight which affects the magnitude of the variation of the Elo. *ClubElo* suggests as value of 20 in order to mitigate the possible variation.
- $G = \begin{cases} 1 & \text{draw or 1 goal difference between the two teams} \\ \frac{3}{2} & \text{goal difference between the two teams is 2} \\ \frac{1+f}{8} & \text{goal difference between the two teams is } f \end{cases}$
- $W = \begin{cases} 1 & \text{win} \\ 0.5 & \text{draw} \\ 0 & \text{defeat} \end{cases}$
- $W_\epsilon = \left(10^{-\frac{(ELO_{before} - ELO_{opp})}{400}} + 1\right)^{-1}$. It considers the Elo of the two teams before the match.

I have taken the Elo ratings for all the teams involved in the competition from elofootball considering as reference year the season 2017/2018. The Elo points were

updated dynamically based on the match score. For example, when considering Leicester City, their performance throughout the season appeared to fluctuate, with a notably negative period between the 22nd and 27th game week. During this period, they only gained one point, scoring an average of 1.2 goals and conceding an average of 2.6 goals. However, they experienced a particularly positive period between the 29th and 33rd game week.

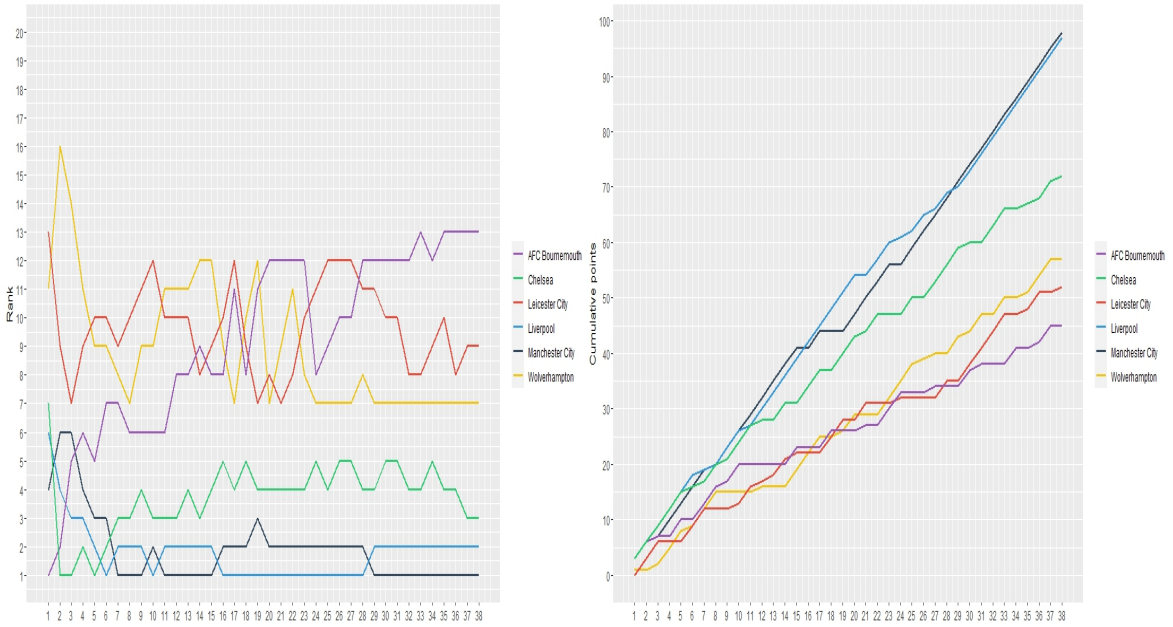


Figure 2.12: Time series of the ranking positions (left) and time series of the gained points (right) for Leicester City, Manchester City, AFC Bournemouth, Wolverhampton Wanderers, Chelsea, Liverpool

2.2 Heuristic Ideas

Taking into account the existing literature [17], **Poisson** distribution (as well as **Negative Binomial** have been considered as a coherent distribution to describe and model the number of goals. Assuming that [4]:

- $y_{h[g]}|\theta \sim Poisson(\theta) \quad g = 1, \dots, 380;$
- $y_{a[g]}|\lambda \sim Poisson(\lambda) \quad g = 1, \dots, 380$

it is well known that the mean and the variance of each random variable are equal to the corresponding Poisson parameter, namely $\mathbb{E}[y_{h[g]}|\theta] = \theta$ and $\mathbb{V}[y_{h[g]}|\theta] = \theta$. Furthermore, the maximum likelihood estimator for the Poisson parameter is the sample mean. It seems that the Poisson distribution can be a good choice to model the number of goals, even though in general it tends to underestimate the probability of observing a zero and overestimate the probability of observing a one.
In order to solve the first problem, a more recent approach [9] has used the **Zero-inflated**

[4]From now on $h[g]$ and $a[g]$ represent respectively the home and the away team in the g-th game.

Poisson distribution to model the scored goals. The basic idea is that the event of scoring 0 goals can be regarded as a sort of `special event` and so a probability mass in zero is added. Formally, if $X|\theta,\omega \sim \textbf{ZIP}(\theta,\omega)$ [5], then:

$$\mathcal{P}\left(X = x|\theta,\omega\right) = \omega\mathbb{I}\{x=0\} + (1-\omega)\frac{e^{-\theta}\theta^{x}}{x!} \quad \omega \in [0,1], \theta \in \Re^{+} \tag{1}$$

where ω is known as the inflated parameter and it has an influence also on the mean and on the variance of a Poisson random variable. In fact, $\mathbb{E}\left[X|\theta,\omega\right] = \theta\left(1-\omega\right)$ and $\mathbb{V}\left[X|\theta,\omega\right] = \theta(1-\omega)(1+\theta\omega)$.

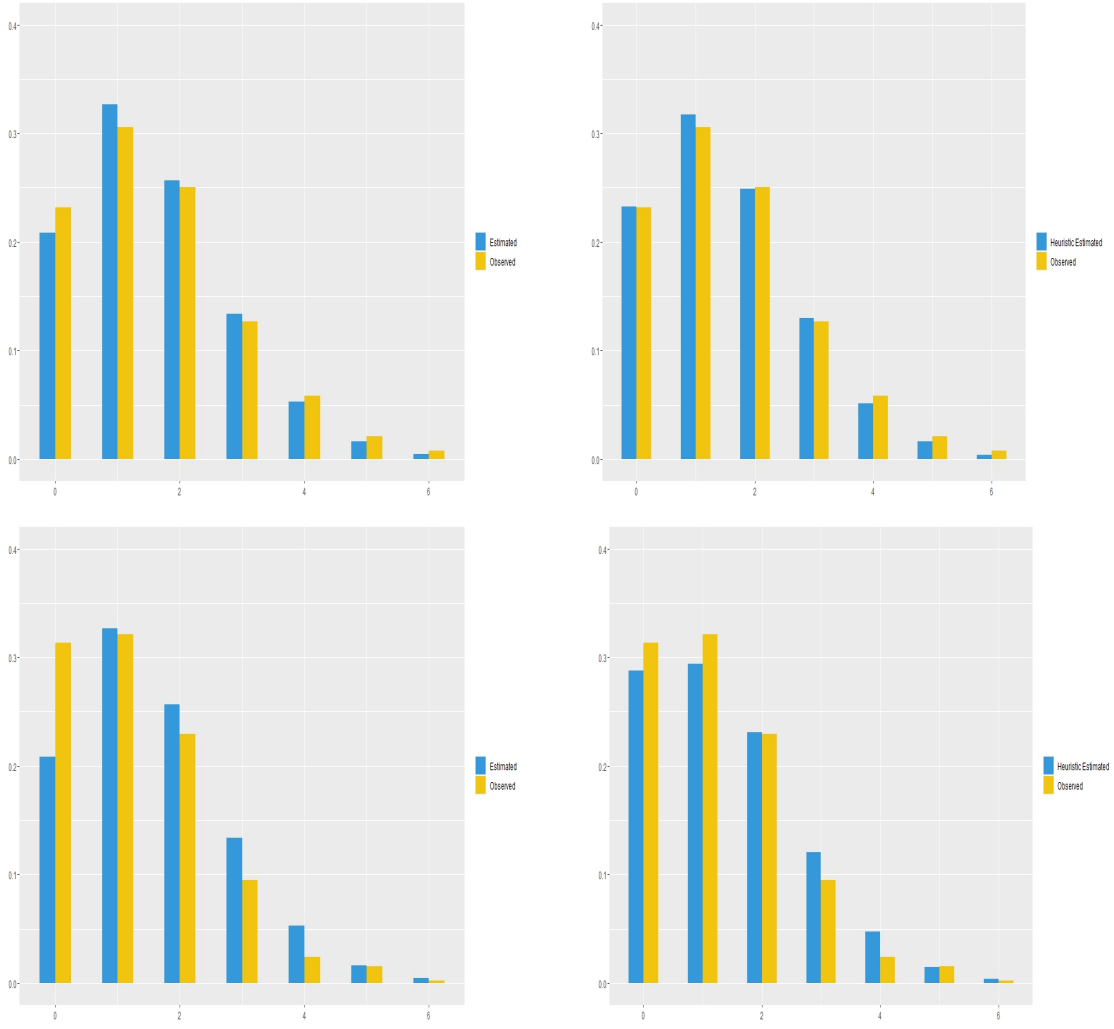


Figure 2.13: On the first row, it is placed the relative frequency of Home team goals compared to the maximum likelihood estimation under a Poisson distribution assumption (top left) and to an euristic estimation under a ZIP distribution (top right) with $\omega = 0.03$. On the second row, it is placed the relative frequency of Home team goals compared to the maximum likelihood estimation under a Poisson distribution assumption (bottom left) and to an euristic estimation under a ZIP distribution (bottom right) with $\omega = 0.04$.

[5]It does not exist a closed form for the maximum likelihood estimator, so I decide to use the sample mean for the parameter θ and fixed values for the parameter ω.

	$\hat{\theta}$	$\hat{\lambda}$
MLE	1.568	1.253

Table 2.4: Maximum likelihood estimates for the parameters of two **Poisson** distributions used to model the home and away scored goals.

As shown in Figure 2.13, the use of the modified Poisson distribution helps to mitigate the issue of underestimating 0 and overestimating 1. However, it is important to address another issue: the possibility of a correlation between the number of goals scored by the two teams. If a team focuses solely on the offensive side, they may be less attentive to their defensive play, resulting in more goals being scored by their opponents. Conversely, if a team is highly focused on defensive play, it can negatively impact the offensive performance of both teams. In this framework, the correlation value is -0.178, indicating that when the home team scores a goal, it is more likely that the away team will not score. Furthermore, a correlation test was conducted using Pearson's product moment correlation coefficient. The obtained p-value of 0.0004 allows us to reject the hypothesis of the absence of correlation. It is important to consider this result as it could significantly impact the validity of the analysis.

After considering these factors, it is important to focus on the joint analysis of the two teams, which can be a challenging task. Several attempts have been made to estimate the joint distribution of scored goals. For example, [10] uses a **Bivariate Poisson** distribution that explicitly takes into account a possible correlation effect.

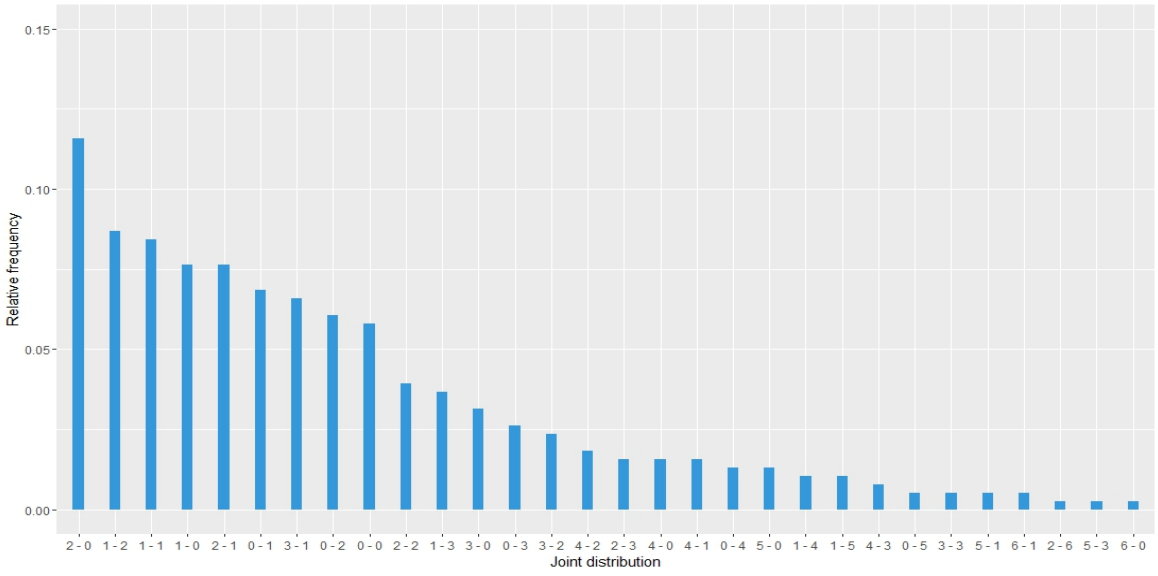


Figure 2.14: Relative frequency of the games' outcomes

Assume that $X_k \sim Poisson(\theta_k), k = 1, 2, 3$, and define $X = X_1 + X_3$ and $Y =$

$X_2 + X_3$. Then, (X, Y) follows a *Bivariate Poisson* distribution $\mathbf{BP}(\theta_1, \theta_2, \theta_3)$.

$$\mathbb{P}(X = x, Y = y) = e^{-(\theta_1+\theta_2+\theta_3)} \frac{\theta_1^x \theta_2^y}{x!y!} \sum_{l=0}^{\min(x,y)} \binom{x}{l} \binom{y}{l} l! \left(\frac{\theta_3}{\theta_1\theta_2}\right)^l \tag{2}$$

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
MLE	1.568	1.253	6.5×10^{-6}

Table 2.5: Maximum likelihood estimates for the parameters of the **Bivariate Poisson** distribution used to model the home and away scored goals.

The previous formulation was introduced by [10] and it is particularly effective for two reasons:

- marginally, the two random variables are Poisson distributions with $\mathbb{E}[X] = \theta_1 + \theta_3$ and $\mathbb{E}[Y] = \theta_2 + \theta_3$;
- θ_3 is a measure of the covariance between the X and Y. In particular $\rho(X, Y) = \frac{\theta_3}{\sqrt{(\theta_1+\theta_3)(\theta_2+\theta_3)}}$.

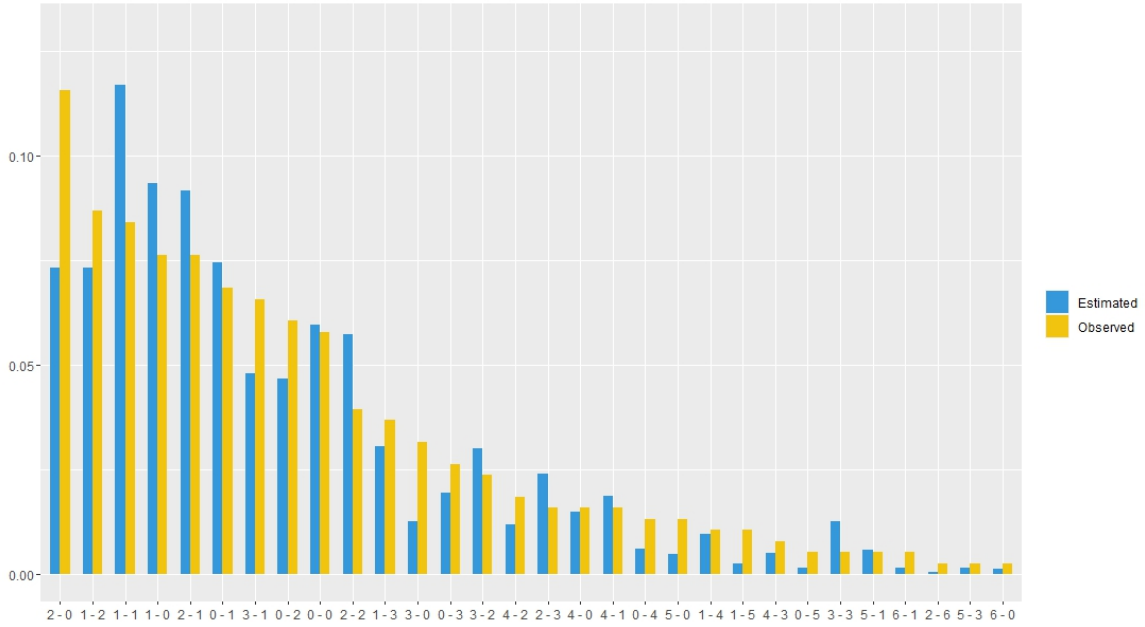


Figure 2.15: Relative frequency of the games' outcomes compared to the probabilities of the same events where the parameters are obtained via the maximum likelihood estimator.

3 Static Models: theory and acknowledgments

In the following section, several models will be discussed. All the different models use a Gibbs Sampler in order to draw samples from the target posterior density. The R code can be found in the appendix B. I have decided to use common values for the hyper-parameters of the MCMC method. In particular:

1. the number of iteration is 10000;
2. the number of discarded samples for the burn-in is 1000. It is used to ensure that the samples are drawn under the target posterior distribution;
3. the ratio of thinning is 1 accepted sample over 10. Thinning is used in order to obtain independent samples;
4. the number of run chains is 2. I have chosen to run multiple chains in order to check for convergence and stationarity.

In order to check for the convergence and the stationarity of the simulations, i have used the **Gelman Rubin** statistic. Knowing that 2 chains are run for each model, denote $\bar{\theta}_j$ the average value for the parameter in the j-th chain. Define:

- $B = N \sum_{j=1}^{2} (\bar{\theta}_j - \bar{\theta})^2$ as the **Between variance**;
- $W = \frac{1}{2} \sum_{j=1}^{2} \left(\frac{1}{N-1} \sum_{i=1}^{N} (\theta_i^{(j)} - \bar{\theta}_j)^2 \right)$ as the **Within variance**;
- $R = \frac{\frac{N-1}{N} W + \frac{1}{N} B}{W}$ as the Gelman Rubin statistic. When the sample tends to increase, the between variance tends to 0 and R tends to 1.

It might seem obvious, but it is important that the samples obtained from the simulation are independent, otherwise it is not possible to use them to rely on the estimation of important features due to the lack of convergence. A tool that can be used to assess the "power" of an MCMC method is the **Effective sample size**:

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

where N is the sample size and ρ_k is the correlation at lag k. ESS represents the number of independent samples with the same estimation power as the N correlated samples.

To add another level of control, traceplots and running mean plots of the parameters are checked, while to study the among samples correlation, autocorrelation plots are

proposed . Given the large number of parameters within the different models, it would be impossible to report all the diagnostics in the results subsections. For each model, a table showing summary statics, ESS, and $\hat{R}$ for the involved parameters is placed in the Appendix A.

The comparison of different models will be pursued through the **Deviance Information Criteria**, also known as DIC. It can be considered as a generalization of the Akaike Information Criteria. More formally, define:

- P as the number of variables involved in the model;
- y as the observed values;
- θ as the parameter of the model;
- $f(y|\theta)$ as the likelihood of the model;
- $D(y, \theta) = -2\log(f(y|\theta))$ as the deviance of the model.

Then, the introduction of other two quantities is needed [18]:

1. Posterior mean deviance [6]

$$D_{AVG} = \mathbb{E}[D(y|\theta)] = \int_{\Theta} D(y|\theta)\pi(\theta|y)d\theta = \int_{\Theta} -2\log(f(y|\theta))\pi(\theta|y)d\theta$$

2. $D(y, \hat{\theta})$, which represents the deviance evaluated at a particular θ value (usually the posterior mean).

The difference between the two quantities (**PD**) represents the *effective number of parameters*, which is used to measure the model complexity. Furthermore, the DIC can be computed as follows:

$$DIC = D_{AVG} + PD = 2D_{AVG} - D(y, \hat{\theta})$$

It has a close resemblance with the AIC, in fact they are both penalized likelihood criteria. Roughly speaking, when two or more models are compared, the one with the lower DIC is chosen because that model has a good balance between goodness of fit and number of involved parameters.

Another tool that can be used to assess the goodness of fit of a model is the Posterior Predictive distribution. Denoting $\tilde{y}$ as a new value, the definition of the previous distribution is straightforward:

$$f(\tilde{y}|y) = \int_{\Theta} f(\tilde{y}, \theta|y)d\theta = \int_{\Theta} f(\tilde{y}|\theta)\pi(\theta|y)d\theta \tag{3}$$

[6]Note that it is pretty easy to estimate this quantity. In fact, when an MCMC method is used, it can be approximated by the average of the simulated values.

3.1 Bayesian Hierarchical model without covariates

In the paper [1], the authors suggest to use of a Bayesian hierarchical structure to model the number of goals scored during a match by two teams because it allows to take into account relations between variables using common distributions for a set of relevant parameters. Furthermore, they affirm that:

> Within the Bayesian framework the use of the more complex bivariate structure is not essential to allow for correlation. [...] Correlation is taken into account since the observable variables are mixed at an upper level.

Denoting $g = 1, \dots, 380$ as a generic game and $w = 1, \dots, 20$ as a generic team, the model can be formalized as follows:

- $y_{g,1}$ represents the number of goals scored by the home team in the g-th game;
- $y_{g,2}$ represents the number of goals scored by the away team in the g-th game;
- $y_{g,1}, y_{g,2}|\theta_{g,1}, \theta_{g,2} \overset{C.I}{=} y_{g,1}|\theta_{g,1} \cdot y_{g,2}|\theta_{g,2}$;
- $y_{g,1}|\theta_{g,1} \sim Poisson(\theta_{g,1})$ and $y_{g,2}|\theta_{g,2} \sim Poisson(\theta_{g,2})$;
- $\theta_{g,1}$ and $\theta_{g,2}$ represent the scoring intensities of the home and away teams;
- $$\begin{cases} log(\theta_{g,1}) = home + att_{h[g]} + def_{a[g]} \\ log(\theta_{g,2}) = att_{a[g]} + def_{h[g]} \end{cases}$$

They suggest to use a **Poisson-log normal** with random effects. The log transformation is used to ensure the positiveness of the scoring intensities. The involved parameters have an easy interpretation:

- **home** represents the advantage for a team hosting the game;
- **att** represents the attacking ability of a team. In particular, it can be interpreted as the difference in a team's propensity to score compared to the average effect in the league. A high positive value indicates a team with a high attacking ability, while a high negative value indicates a team with a low attacking ability;
- **def** is the defensive ability of a team. As before, it can be interpreted as the differential defensive propensity of a team relative to the average effect in the league. Due to the structure of the regression equations, negative values imply a better defensive attitude than positive values.

Those abilities, as well as the home effects, cannot properly be measured, so they can also be interpreted as random latent effects which affect the scoring propensity of a team [7]. The main hypothesis of the model is that they remain constant over the entire

[7] Note that a_g and h_g allow to uniquely determine each team.

season for all the teams.

The likelihood of the model can be written:

$$\begin{aligned}\mathbb{L}(\underline{\theta}) &= f\left(y_{1,1}, y_{1,2}, \ldots, y_{380,1}, y_{380,2} | \theta_{1,1}, \theta_{1,2}, \ldots, \theta_{380,1}, \theta_{380,2}\right) \\ &\overset{c.i}{=} f\left(y_{1,1}, \ldots, y_{380,1} | \theta_{1,1}, \ldots, \theta_{380,1}\right) \cdot f\left(y_{1,2}, \ldots, y_{380,2} | \theta_{1,2}, \ldots, \theta_{380,2}\right) \\ &\overset{i.d}{=} \prod_{g=1}^{380} f\left(y_{g,1} | \theta_{g,1}\right) \cdot f\left(y_{g,2} | \theta_{g,2}\right)\end{aligned}$$

Priors for the parameter have been chosen in a suitable way, i.e. to allow for conjugacy. For simplicity's sake, those random variables are assumed to be independent from each other.

$$\begin{cases} home \sim N(0, 0.0001) \\ att_w \sim N(\mu_{att}, \tau_{att}) \quad w = 1, \ldots, 20 \\ def_w \sim N(\mu_{def}, \tau_{def}) \quad w = 1, \ldots, 20 \\ \pi\left(home, att_1, \ldots, att_{20}, def_1, \ldots, def_{20}\right) \overset{i}{=} \pi\left(home\right) \prod_{w=1}^{20} \pi\left(att_w\right) \pi\left(def_w\right) \end{cases}$$

As pointed out by the authors, in order to avoid identifiability issues, a *reference point* parametrization is used:

$$\sum_{w=1}^{20} att_w = 0 \Leftrightarrow att_1 = -\sum_{w=2}^{20} att_w \tag{4}$$

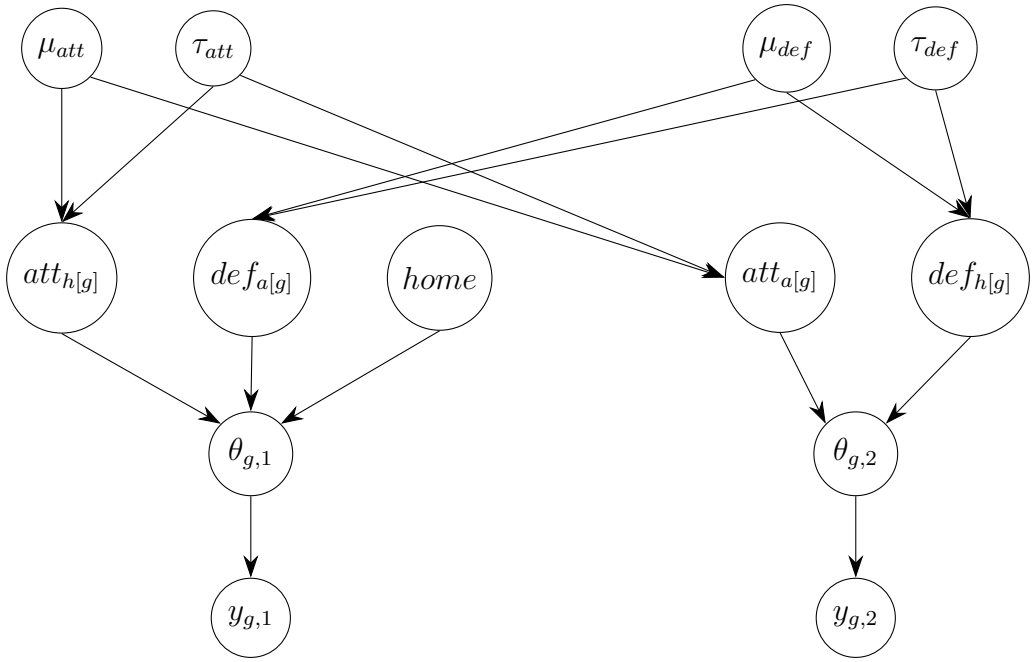
$$\sum_{w=1}^{20} def_w = 0 \Leftrightarrow def_1 = -\sum_{w=2}^{20} def_w \tag{5}$$

The hyperpriors are modelled independently using non-informative distributions. They are also assumed to be constant over the entire season:

$$\begin{cases} \mu_{att} \sim N(0, 0.0001) \\ \mu_{def} \sim N(0, 0.0001) \\ \tau_{att} \sim Gamma(0.1, 0.1) \\ \tau_{def} \sim Gamma(0.1, 0.1) \end{cases}$$

The entire hierarchical structure can be represented using the following DAG [8].

[8] Directed acyclic graph



The target posterior distribution of the model can be written using the usual bayesian updating rule:

$$\pi\left(\underline{\theta}|\underline{y}\right) \propto \pi\left(\underline{\theta}\right)\mathbb{L}\left(\underline{\theta}\right) = \pi\left(home\right)\prod_{w=1}^{20}\pi\left(att_w\right)\pi\left(def_w\right)\prod_{g=1}^{380}f\left(y_{g,1}|\theta_{g,1}\right)\cdot f\left(y_{g,2}|\theta_{g,2}\right) =$$

$$= N(0, 0.0001)\prod_{w=1}^{20}N(\mu_{att},\tau_{att})N(\mu_{def},\tau_{def})\prod_{g=1}^{380}Poisson\left(\theta_{g,1}\right)Poisson\left(\theta_{g,2}\right)$$

The previous posterior distribution is not analytically tractable and so a Monte Carlo Markov Chain method is used in order to generate samples from it. Due to the complex structure of the posterior, a Gibbs Sampler is adopted.

Given the hierarchical structure of the model I used, two objectives can be defined. The first is to estimate the main effects influencing the scoring rates. I initialised the parameters of the model involved in the Gibbs sampler with two different vectors of values. In the case of the offensive and defensive parameters, in one case they were randomly generated with a Gaussian distribution centered on 0 and with a small variance for each position. The other initial values were obtained by subtracting the league average from the offensive and defensive propensities (Figure 3.1). As mentioned above, these are not the observed values that represent attacking and defensive ability, as these effects cannot be observed.

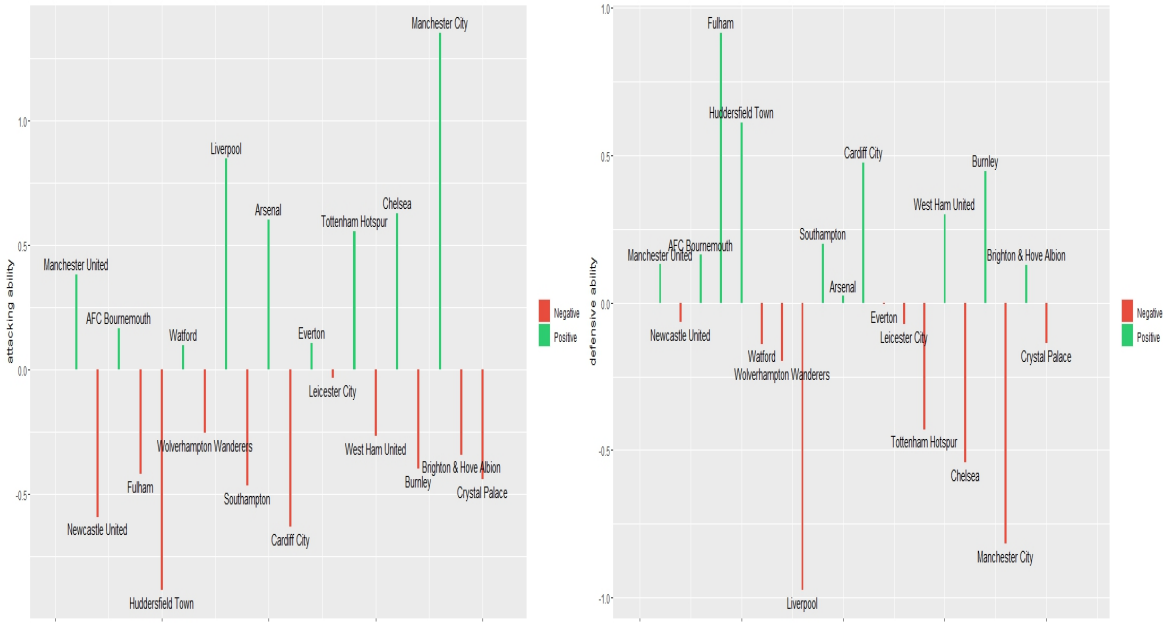


Figure 3.1: General differential Attacking ability w.r.t the average effect in the league (left). General differential Defensive ability w.r.t the average effect in the league (right)

3.2 Bayesian Hierarchical model with covariates

Taking into consideration the theories and the beliefs developed during the descriptive part, the previous model can be modified as follows. First of all, it is possible to add covariates because it has been proven that there are other factors which have an influence on the number of goals scored by a team. For the time being, only those that are useful to describe and represent the shape of a team will be considered.

Moreover, the **home effect** will be not considered to be the same for all the teams. As has been demonstrated above, the number of goals scored at home turns out to be higher for teams that can accommodate a larger number of people inside their stadium. So, the home effect will be partitioned into two clusters:

- teams that "own" stadiums with a large capacity, say cluster A;
- teams that "own" stadiums with a medium-small capacity, say cluster B.

More formally, the home effect has a different influence on the number of goals scored depending on which cluster the home team is in. The introduction of the indicator function $attendance_w$ is required. In particular:

$$attendance_w = \begin{cases} 1 & w \in A \\ 0 & w \in B \end{cases}$$

Let's try to formalize the main differences in the likelihood w.r.t the previous model:

1. $\log(\theta_{g,1}) = b_0 attendance_{h[g]} + b_1(1 - attendance_{h[g]}) + \beta_1 RSC_{h[g]} + \beta_2 RDC_{a[g]} + \beta_3 QO_{h[g]} + \beta_4 \Delta ELO_{h[g]} + att_{h[g]} + def_{a[g]}$;
2. $\log(\theta_{g,2}) = \beta_1 RSC_{a[g]} + \beta_2 RDC_{h[g]} + \beta_3 QO_{a[g]} + \beta_4 \Delta ELO_{a[g]} + att_{a[g]} + def_{h[g]}$;
3. as shown by the two equations above, I have opted to use the difference between the Elo points of both competitors as the independent variable;
4. the form of the likelihood is the same as before, even if the two scoring propensities depend on more parameters rather than before.

Regarding the new parameters:

- the random effects b_0 and b_1 are independent and they follow a Gaussian distribution centered in 0 and with the same variance σ_0^2. In turn, σ_0^2 follows a **Gamma** distribution;
- $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ represent a vector of fixed effects and they are assumed to be independent between each other. Furthermore, they are assumed to be the same among the two Poisson regressions. Each β follows a Gaussian distribution centered in 0 and with variance equal to 0.0001.

The prior distribution for all the parameters can now be written as:

$$\pi\left(b_0, b_1, \underline{att}, \underline{def}, \underline{\beta}\right) \overset{i}{=} \prod_{d=1}^{2} \pi\left(b_d\right) \prod_{p=1}^{4} \pi(\beta_p) \prod_{w=1}^{20} \pi\left(att_w\right) \pi\left(def_w\right)$$

The posterior target distribution is:

$$\pi\left(\underline{\theta}|\underline{y}\right) \propto \pi\left(\underline{\theta}\right) \mathbb{L}\left(\underline{\theta}\right) = \prod_{d=1}^{2} \pi\left(b_d\right) \prod_{p=0}^{4} \pi(\beta_p) \prod_{w=1}^{20} \pi\left(att_w\right) \pi\left(def_w\right) \prod_{g=1}^{380} f\left(y_{g,1}|\theta_{g,1}\right) \cdot f\left(y_{g,2}|\theta_{g,2}\right)$$

Also in this case, due to the complex structure of the model, a Gibbs Sampler is required to generate samples from the target posterior density.

As before, two chains have been initialized. For the attacking and defensive parameter i have chosen the same initialization as the previous model, while all the other parameter have been randomly initialized due to the lack of prior information.

3.3 Zero-Inflated Bayesian Hierarchical model

In the paper [11], the authors propose to introduce a factor in order to take into account the under prediction of zeros. As pointed out in the descriptive part, the number of goals scored by home and away teams are influenced by different inflation factors. A natural extension of the previous model is possible [9]:

- $y_{g,1}, y_{g,2}|\theta_{g,1}, \theta_{g,2}, p_1, p_2 \overset{C.I}{=} y_{g,1}|\theta_{g,1}, p_1 \cdot y_{g,2}|\theta_{g,2}, p_2$;
- $y_{g,1}|\theta_{g,1}, p_1 \sim \mathbf{ZIP}(\theta_{g,1}, p_1)$ and $y_{g,2}|\theta_{g,2}, p_2 \sim \mathbf{ZIP}(\theta_{g,2}, p_2)$;
- p_1 and p_2 denote the inflation factors for home and away goals scored, respectively. Referring to the definition of the Zero-inflated Poisson distribution 1, p_1 (alternatively p_2) can be interpreted as the mixing proportion of a mixture distribution involving a random variable degenerate at 0 and a Poisson distribution. These effects are assumed to remain constant throughout the season and are not impacted by any covariates;
- introduce two indicator functions $\delta_{g,1}$ and $\delta_{g,2}$, where $\delta_{g,1} = 1$ indicated the zero-inflated component for the home goals and $\delta_{g,2} = 1$ indicated the zero-inflated component for the away goals. In particular, $\delta_{g,1}|p_1 \sim \mathbf{Bern}(p_1)$ and $\delta_{g,2}|p_2 \sim \mathbf{Bern}(p_2)$. Recall that:

$$f(y_{g,1} = 0|p_1, \theta_{g,1}, \delta_{g,1} = 1) = 1 \quad f(y_{g,1} = k|p_1, \theta_{g,1}, \delta_{g,1} = 0) = \mathbf{Pois}(k|\theta_{g,1})$$
$$f(y_{g,2} = 0|p_2, \theta_{g,2}, \delta_{g,2} = 1) = 1 \quad f(y_{g,2} = k|p_2, \theta_{g,2}, \delta_{g,2} = 0) = \mathbf{Pois}(k|\theta_{g,2})$$

- $y_{g,1}, y_{g,2}, \delta_{g,1}, \delta_{g,2}|\theta_{g,1}, \theta_{g,2}, p_1, p_2 \overset{C.I}{=} y_{g,1}|\theta_{g,1}, p_1, \delta_{g,1} \cdot \delta_{g,1}|p_1 \cdot y_{g,2}|\theta_{g,2}, p_2, \delta_{g,2} \cdot \delta_{g,2}|p_2$ [10].

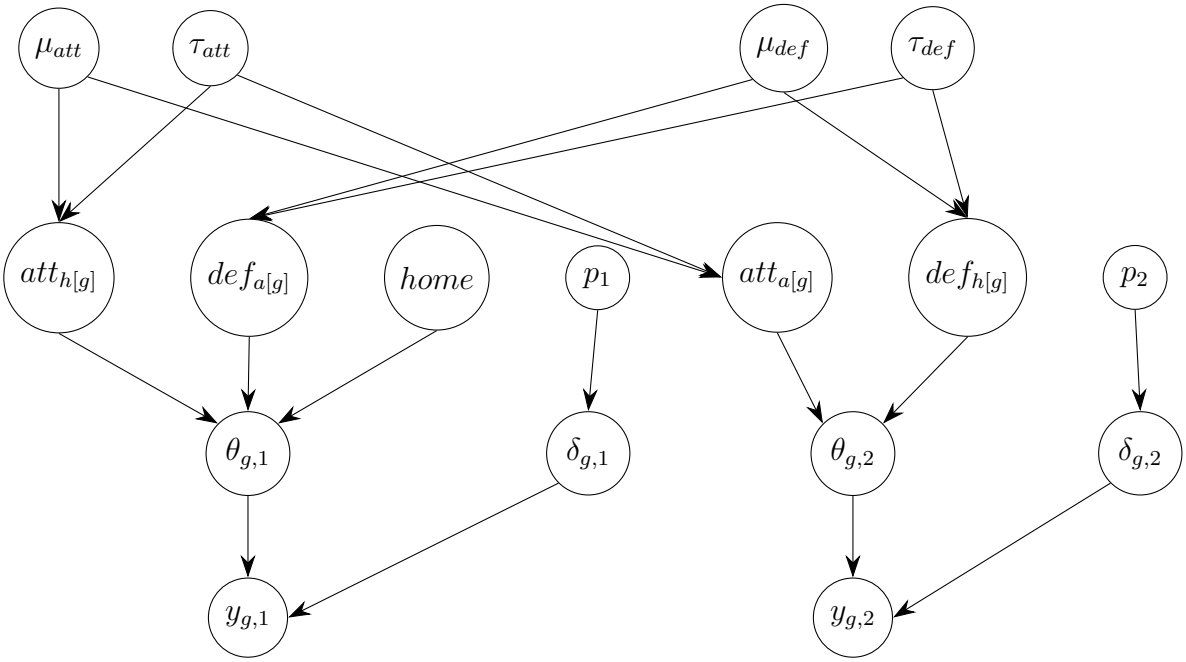
Under the new model specification, the likelihood can be written as:

$$\begin{aligned}
\mathbb{L}(\underline{\theta}, p_1, p_2) &= f\left(y_{1,1}, \delta_{1,1}, y_{1,2}, \delta_{1,2}, \dots, y_{380,1}, \delta_{380,1}, y_{380,2}, \delta_{380,2}|\theta_{1,1}, \theta_{1,2}, \dots, \theta_{380,1}, \theta_{380,2}, p_1, p_2\right) \\
&\overset{c.i.i.d}{=} \prod_{g=1}^{380} f(y_{g,1}|\delta_{g,1}, \theta_{g,1}, p_1) f(\delta_{g,1}|p_1) f(y_{g,2}|\delta_{g,2}, \theta_{g,2}, p_2) f(\delta_{g,2}|p_2) \\
&\propto \prod_{g=1}^{380} \left(p_1 \mathbb{I}\{y_{g,1} = 0\}\right)^{\delta_{g,1}} \left((1-p_1)\mathbf{Pois}(y_{g,1}|\theta_{g,1})\right)^{1-\delta_{g,1}} \left(p_2 \mathbb{I}\{y_{g,2} = 0\}\right)^{\delta_{g,2}} \\
&\left((1-p_2)\mathbf{Pois}(y_{g,2}|\theta_{g,2})\right)^{1-\delta_{g,2}}
\end{aligned}$$

Following the same reasoning as before, it remains only to specify prior distributions for the new parameters. In particular, $p_1 \sim \mathbf{Beta}(a_0, b_0)$ and $p_2 \sim \mathbf{Beta}(a_1, b_1)$. A common choice is to allow $a_0 = b_0 = a_1 = b_1 = 1$ in order to use non-informative priors

[9] Note that in the following bullet list, i will just report the differences w.r.t the previous case.
[10] This technique is known as **Data Augmentation.**

[11]. The new hierarchical structure can be represented as follows:



The target posterior distribution can be written as:

$$\pi\left(\underline{\theta}, p_1, p_2, \underline{\delta}|\underline{y}\right) \propto \pi\left(\underline{\theta}, p_1, p_2\right) \mathbb{L}\left(\underline{\theta}, p_1, p_2\right) = \pi\left(home\right) \pi(p_1)\pi(p_2) \prod_{w=1}^{20} \pi\left(att_w\right) \pi\left(def_w\right) \mathbb{L}\left(\underline{\theta}, p_1, p_2\right)$$

$$= \mathbf{N}(0, 0.00001)\mathbf{Beta}(a_0, b_0)\mathbf{Beta}(a_1, b_1) \prod_{w=1}^{20} N(\mu_{att}, \tau_{att})N(\mu_{def}, \tau_{def})\mathbb{L}\left(\underline{\theta}, p_1, p_2\right)$$

As before, the complex structure of the model does not allow the analytical computation of the target posterior and a Gibbs Sampler is adopted again. From both an interpretive and statistical point of view, it turns out to be interesting to formalize the full conditional distribution for the parameter $\delta_{g,1}$ ($\delta_{g,2}$).

$$\pi(\delta_{g,1}|\underline{\theta}, \underline{y}, p_1, p_2, \delta_{g,2}) \propto \prod_{g=1}^{380} \left(p_1\mathbb{I}\{y_{g,1}=0\}\right)^{\delta_{g,1}} \left((1-p_1)e^{-\theta_{g,1}}\frac{\theta_{g,1}^{y_{g,1}}}{y_{g,1}!}\right)^{1-\delta_{g,1}}$$

$$\propto \left(p_1\mathbb{I}\{y_{g,1}=0\}\right)^{\delta_{g,1}} \left((1-p_1)e^{-\theta_{g,1}}\frac{\theta_{g,1}^{y_{g,1}}}{y_{g,1}!}\right)^{1-\delta_{g,1}}$$

$$= q_1^{\delta_{g,1}} \left(\frac{p_1\mathbb{I}\{y_{g,1}=0\}}{q_1}\right)^{\delta_{g,1}} \left(\frac{(1-p_1)}{q_1}e^{-\theta_{g,1}}\frac{\theta_{g,1}^{y_{g,1}}}{y_{g,1}!}\right)^{1-\delta_{g,1}} q_1^{1-\delta_{g,1}}$$

$$\propto \left(\frac{p_1\mathbb{I}\{y_{g,1}=0\}}{q_1}\right)^{\delta_{g,1}} \left(\frac{(1-p_1)}{q_1}e^{-\theta_{g,1}}\frac{\theta_{g,1}^{y_{g,1}}}{y_{g,1}!}\right)^{1-\delta_{g,1}}$$

Where $q_1 = p_1\mathbb{I}\{y_{g,1}=0\} + (1-p_1)\mathbf{Pois}(y_{g,1}|\theta_{g,1})$. In particular, each argument of the distribution is multiplied and divided by q_1 in order to obtain a probability measure.

[11]Remember that a $\mathbf{Beta}(1,1)$ is a $\mathbf{Uniform}[0,1]$

Moreover, $\delta_{g,1}|p_1, \theta_{g,1}, y_{g,1} \sim \mathbf{Bern}(\tilde{p_1})$, where:

$$\tilde{p_1} = \mathbb{P}\left(\delta_{g,1} = 1|\theta_{g,1}, p_1, y_{g,1}\right) = \frac{p_1 \mathbb{I}\{y_{g,1} = 0\}}{p_1 \mathbb{I}\{y_{g,1} = 0\} + (1 - p_1)\mathbf{Pois}(y_{g,1}|\theta_{g,1})}$$

In my opinion, the way the parameter p_1 updates from one iteration to another is emblematic since it weights the information coming from the beta with the information coming from the mixture distribution. It is possible to make the same considerations for the random variable $\delta_{g,2}$.

As with all other scenarios, two chains have been run. The offensive and defensive coefficients have been set up in the same manner as illustrated in Figure 3.1, while the two inflation parameters, along with the home random effect, have been initialized randomly.

3.4 Modified Zero-Inflated Bayesian Hierarchical model

Taking into account [10] and considering the results showed in Section 4, it was not possible to accurately predict the probability of the event $\{$*team scores 0 goals*$\}$ using the previous models. One potential explanation is that they only consider a single parameter shared by all teams, which represents their tendency to score 0 goals at home or away. Note that this does not constitute a formal interpretation of the inflation parameter as it functions as a mixing weight within a mixture distribution. As shown in Figure 2.2, the teams had varying percentages of goals scored equal to zero throughout the season. Therefore, relying on a single parameter is too strong as an assumption and so it will be relaxed. In greater detail, the magnitude of this effect will be measured using a parameter for each team, one for when they play at home and one for when they play away. It is possible to extend the previous model as follows:

- the vector $\underline{p}_1 = \left(p_{1,1}, \ldots, p_{1,20}\right)$ represents the mixing proportion for the **ZIP** distribution used to model the home goals, while the vector $\underline{p}_2 = \left(p_{2,1}, \ldots, p_{2,20}\right)$ refers to **ZIP** distribution for the away goals. Again they are assumed to be constant throughout the season;
- $y_{g,1}, y_{g,2}|\theta_{g,1}, \theta_{g,2}, p_{1_{[g]}}, p_{2_{[g]}} \overset{C.I}{=} y_{g,1}|\theta_{g,1}, p_{1_{[g]}} \cdot y_{g,2}|\theta_{g,2}, p_{2_{[g]}}$. Note that by utilizing $1_{[g]}$ and $2_{[g]}$, it is feasible to distinctly determine a single coefficient $\forall g = 1, \ldots 20$;
- $y_{g,1}|\theta_{g,1}, p_{1_{[g]}} \sim \mathbf{ZIP}(\theta_{g,1}, p_{1_{[g]}})$ and $y_{g,2}|\theta_{g,2}, p_{2_{[g]}} \sim \mathbf{ZIP}(\theta_{g,2}, p_{2_{[g]}})$;
- the parameters retain their previous interpretation, and I additionally require the inclusion of the indicator functions $\delta_{g,1}$ and $\delta_{g,2}$.

Given the new specifications, the likelihood of the model has a similar form than before:

$$\begin{aligned}\mathbb{L}\left(\underline{\theta}, \underline{p}_1, \underline{p}_2\right) &= f\left(y_{1,1}, \delta_{1,1}, y_{1,2}, \delta_{1,2}, \ldots, y_{380,1}, \delta_{380,1}, y_{380,2}, \delta_{380,2} | \theta_{1,1}, \theta_{1,2}, \ldots, \theta_{380,1}, \theta_{380,2}, \underline{p}_1, \underline{p}_2\right) \\ &\overset{c.i.i.d}{=} \prod_{g=1}^{380} f(y_{g,1}|\delta_{g,1}, \theta_{g,1}, p_{1_{[g]}}) f(\delta_{g,1}|p_{1_{[g]}}) f(y_{g,2}|\delta_{g,2}, \theta_{g,2}, p_{2_{[g]}}) f(\delta_{g,2}|p_{2_{[g]}}) \\ &\propto \prod_{g=1}^{380} \left(p_{1_{[g]}} \mathbb{I}\{y_{g,1} = 0\}\right)^{\delta_{g,1}} \left((1 - p_{1_{[g]}}) \mathbf{Pois}(y_{g,1}|\theta_{g,1})\right)^{1-\delta_{g,1}} \left(p_{2_{[g]}} \mathbb{I}\{y_{g,2} = 0\}\right)^{\delta_{g,2}} \\ &\left((1 - p_{2_{[g]}}) \mathbf{Pois}(y_{g,2}|\theta_{g,2})\right)^{1-\delta_{g,2}}\end{aligned}$$

Appropriate prior distributions are required for the latest inflation factors. Adopting the same approach as the random effects specification, it is conceivable that the inflation factors for home and away teams are linked, originating from a shared population distribution. In order to achieve this outcome, it is possible to assume that the parameters are independent within each other as well as with respect to the other parameters of the model. The prior distributions for the parameters can be expressed as:

$$\begin{cases} p_{1,w} \sim N(\mu_{p_1}, \tau_{p_1}) T(0,1) \quad w = 1, \ldots, 20 \\ p_{2,w} \sim N(\mu_{p_2}, \tau_{p_2}) T(0,1) \quad w = 1, \ldots, 20 \\ \pi(\underline{p}_1, \underline{p}_2) \overset{=}{i} \prod_{w=1}^{20} N(\mu_{p_1}, \tau_{p_1}) T(0,1) N(\mu_{p_2}, \tau_{p_2}) T(0,1) \end{cases}$$

Two remarks can be made:

- a truncated normal distribution within the range of $[0, 1]$ is selected because the inflation factors represent proportions, and their value is in logit form;
- the population parameters μ_{p1} (μ_{p2}) and τ_{p1} (τ_{p2}) correspond to the population mean and variance of the inflation factors of home (away) goals scored. It is crucial to choose a probability distribution that fits these hyperparameters appropriately. The key principle to adhere to is to center the mean parameter on zero, as the variance parameter has a significant impact on the differences in coefficient values. Namely, if the common variance approaches zero, it is highly probable that the parameters will be concentrated around a single value (which is not significantly different from having only one parameter). However, a larger common variance permits the coefficients to take on distinct values. Without any loss of generality, it can be argued the common variance is acting as a **shrinkage hyperprior**.

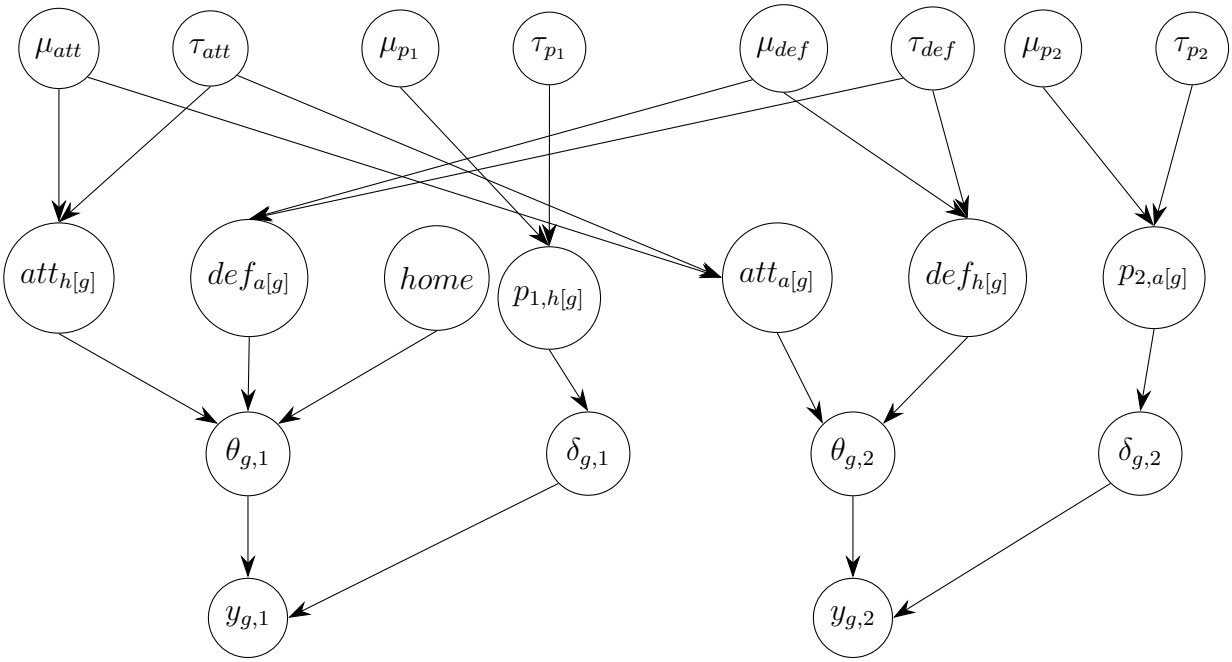
  Taking into account the ideas of [8], two weakly informative priors distribution will be used for the variance hyperparameter.

$$\begin{cases} \mu_{p_1} \sim Beta(1,1) \\ \tau_{p_1} \sim Gamma(0.001, 0.001) \\ \mu_{p_2} \sim Beta(1,1) \\ \tau_{p_2} \sim Gamma(0.001, 0.001) \end{cases} \quad \begin{cases} \mu_{p_1} \sim Beta(1,1) \\ \tau_{p_1} \sim Cauchy^{+}(0, c_0) \\ \mu_{p_2} \sim Beta(1,1) \\ \tau_{p_2} \sim Cauchy^{+}(0, c_1) \end{cases}$$

Note that the Half-Cauchy distribution, which is a truncated Cauchy distribution, only has non-zero probability density for values greater than or equal to 0. It is noteworthy that this is a heavy-tailed distribution. Moreover, it is imperative to carefully select the hyperparameters c_0 and c_1, as the prior distribution will become a uniform over the domain if $c_0 \to \infty$ (and likewise for c_1). The fundamental concept behind selecting priors that assign a substantial probability mass to 0 is that coefficients will be distinct from zero only if the data permits.
The prior distribution for all the parameters involved in the model can be expressed as follows:

$$\pi(\underline{\theta}, \underline{p}_1, \underline{p}_2) = \pi(home) \prod_{w=1}^{20} \pi(p_{1,w}) pi(p_{2,w}) \pi(att_w) \pi(def_w)$$

Taking into account the new model specification, the new hierarchical structure can be expressed in the following way.



The posterior distribution of the target model can be expressed via the commonplace Bayesian updating rule. However, it is not analytically tractable, necessitating the

implementation of the Gibbs sampler.

$$\pi\left(\underline{\theta}, \underline{p_1}, \underline{p_2}\middle|\underline{y}\right) \propto \pi\left(\underline{\theta}\right)\mathbb{L}\left(\underline{\theta}\right) = \pi\left(home\right)\prod_{w=1}^{20}\pi(p_{1,w})pi(p_{2,w})\pi\left(att_w\right)\pi\left(def_w\right)$$

$$\prod_{g=1}^{380} f\left(y_{g,1}|\theta_{g,1}, p_{1_{[g]}}\right)\cdot f\left(y_{g,2}|\theta_{g,2}, p_{2_{[g]}}\right) =$$

$$= N(0, 0.0001)\prod_{w=1}^{20} N(\mu_{p_1}, \tau_{p_1})T(0,1)N(\mu_{p_2}, \tau_{p_2})T(0,1)N(\mu_{att}, \tau_{att})N(\mu_{def}, \tau_{def})$$

$$\prod_{g=1}^{380} ZIP\left(\theta_{g,1}, p_{1_{[g]}}\right) ZIP\left(\theta_{g,2}, p_{2_{[g]}}\right)$$

To provide a comprehensive analysis, I have also considered the extreme case where the inflation factors do not originate from the same population. In this scenario, it is assumed that each parameter is independent and originates from a different population. To maintain consistency with the nature of the parameters, a non-informative distribution, specifically a Beta(1,1), is used. The model's structure remains largely unchanged, but hyperpriors linked to the inflation parameters for home and away scored goals have been removed.

Following the idea and the methodology of the previous experiments, two chains have been run. The offensive and defensive random effects in the model were initialized similarly to all the other models (refer to Figure 3.1), while the other parameters were randomly initialized. As a note, I selected values of c_0 and c_1 both equal to 30 to establish a weakly informative distribution. [8] suggests using a higher value for the scale parameter of the Half Cauchy distribution, in order to be as non-informative as possible and to allow the data to have a greater influence on the magnitude of the parameters. Figure 3.2 shows that despite using two different hyperpriors for the common variance, the estimates of the inflation factors and their posterior marginal distribution are quite similar. However, upon closer inspection of their similarity and conducting a **Kolmogorov-Smirnoff** test on their posterior cumulative distribution, it is consistently feasible to refute the null hypothesis that the two samples are drawn from the same distribution. Moreover, going forward, I will present only the outcome for the model that applies a Gamma hyperprior as it shows a lower DIC. Furthermore, it is evident that the coefficients have similar and minute values, leading to the conclusion that the current situation is not noticeably different from the earlier model with only two coefficients.

Figure 3.2: Marginal posterior densities for $p1, w$ using Gamma hyperprior (top left) and Half Cauchy hyperprior (top right). Marginal posterior densities for $p_{2,w}$ using Gamma hyperprior (bottom left) and Half Cauchy hyperprior (bottom right).

3.5 Extra: Explanatory Model

Following the ideas contained in [7], the first step I would like to take is to assess which are the most important variables capable of influencing the outcome of the match. The structure of the following subsection will be as follows, first I will introduce the chosen model with a suitable notation and then a comment on the results will be shown. In particular, the results will be presented here and not in the following section because the objectives are different.

Starting from the description of the model, the notation is the same as the others. The main difference with [1] is that no random effects are used because, as can be seen from section 4, their effect is taken over by the covariates. Furthermore, an intercept has been included in the equations of the scoring intensities related to the number of goals scored by the away team. It is important to note that there is a direct connection between this new parameterization and the conventional one. In fact, the home advantage can be seen as the disparity between the two intercepts of the model. The log-linear equations can be written as follows:

$$\begin{cases} log(\theta_{g,1}) = \underline{x}_{g,h}^T \underline{\beta}_h \\ log(\theta_{g,2}) = \underline{x}_{g,a}^T \underline{\tilde{\beta}}_a \end{cases}$$

where:

- $\underline{x}_{g,h} = [1\ x_{g,2}^h \dots\ x_{g,12}^h]$ and $\underline{x}_{g,a} = [1, x_{g,2}^a \dots\ x_{g,12}^a]$ represent the covariates vector associated to the home and away team for the g-th game;
- $\underline{\beta}_h = [u_h, \beta_1, \dots, \beta_{11}]$ and $\underline{\tilde{\beta}}_a = [u_a, \tilde{\beta}_1, \dots, \tilde{\beta_{10}}]$ are the set of coefficients associated to the home and away variables. The regression coefficients are assumed to be different because the covariates can have a different impact on the number of scored goals if the match is played at home or away.

The likelihood of the model can be written in the following way:

$$\begin{aligned} \mathbb{L}(\underline{\theta}) &= f\left(y_{1,1}, y_{1,2}, \dots, y_{380,1}, y_{380,2} | \theta_{1,1}, \theta_{1,2}, \dots, \theta_{380,1}, \theta_{380,2}\right) \\ &\overset{c.i}{=} f\left(y_{1,1}, \dots, y_{380,1} | \theta_{1,1}, \dots, \theta_{380,1}\right) \cdot f\left(y_{1,2}, \dots, y_{380,2} | \theta_{1,2}, \dots, \theta_{380,2}\right) \\ &\overset{i.d}{=} \prod_{g=1}^{380} f\left(y_{g,1} | \theta_{g,1}\right) \cdot f\left(y_{g,2} | \theta_{g,2}\right) = \prod_{g=1}^{380} Poisson(\exp(\underline{x}_{g,h}^T \underline{\beta}_h)) \cdot Poisson(\exp(\underline{x}_{g,a}^T \underline{\beta}_a)) \end{aligned}$$

In the context of regression analysis, the selection of appropriate prior distributions for regression coefficients is a crucial aspect, and one notable choice is the **Zellner G-prior** as introduced by [23]. This particular prior offers a distinctive approach by circumventing the need to explicitly specify the variance-covariance matrix, allowing it to adapt to the data's characteristics. The parameter denoted as g in the G-prior holds significance as it is inversely proportional to the information incorporated into the prior relative to the sample. In the conventional framework, the selection of g is guided by the principle of equating it to the number of observations in the dataset. This choice ensures that the prior is accorded a weight comparable to that of an individual observation.

It is noteworthy that, in a theoretical context, the Zellner G-prior also allows for the incorporation of weak assumptions regarding the location of the parameters. However,

for the present discussion, such assumptions are not explicitly considered:

$$\begin{cases} \underline{\beta}_h \sim \mathcal{MVN}_{12}\left(\underline{0}, g\sigma^2\left(X_1^T X_1\right)^{-1}\right) \\ \underline{\tilde{\beta}}_a \sim \mathcal{MVN}_{11}\left(\underline{0}, g\tilde{\sigma}^2\left(X_2^T X_2\right)^{-1}\right) \\ \sigma^2 \sim \mathcal{IG}(a,b) \quad \tilde{\sigma}^2 \sim \mathcal{IG}(a,b) \\ \pi\left(\underline{\beta}_h, \underline{\tilde{\beta}}_a, \sigma^2, \tilde{\sigma}^2\right) = \pi\left(\underline{\beta}_h|\sigma^2\right)\pi\left(\underline{\tilde{\beta}}_a|\tilde{\sigma}^2\right)\pi(\sigma^2)\pi(\tilde{\sigma}^2) \end{cases}$$

Following the usual Bayesian updating rule, it is possible to write the posterior distribution in the following way:

$$\pi\left(\underline{\theta}|\underline{y}\right) \propto \pi\left(\underline{\theta}\right)\mathbb{L}\left(\underline{\theta}\right) = \pi\left(\underline{\beta}_h|\sigma^2\right)\pi\left(\underline{\tilde{\beta}}_a|\tilde{\sigma}^2\right)\pi(\sigma^2)\pi(\tilde{\sigma}^2)\prod_{g=1}^{380} f\left(y_{g,1}|\theta_{g,1}\right)\cdot f\left(y_{g,2}|\theta_{g,2}\right) =$$

$$= \mathcal{MVN}_{12}\left(\underline{0}, g\sigma^2\left(X_1^T X_1\right)^{-1}\right)\mathcal{MVN}_{11}\left(\underline{0}, g\tilde{\sigma}^2\left(X_2^T X_2\right)^{-1}\right)\mathcal{IG}(a,b)\mathcal{IG}(a,b)$$

$$\prod_{g=1}^{380} Poisson(\exp(\underline{x}_{g,h}^T\underline{\beta}_h))\cdot Poisson(\exp(\underline{x}_{g,a}^T\underline{\beta}_a))$$

In addressing the intricate nature of the posterior distribution, a Gibbs sampler was employed as the computational methodology. Building upon the principles elucidated in Section 3, a dual-chain configuration was adopted, and the model parameters were initialized randomly. The computational process encompassed a total of 10^5 iterations, with half of these iterations dedicated to the burn-in phase. It is noteworthy that the independent variables within the model underwent a standardization process. This standardization not only enhances interpretability but also serves to facilitate convergence. Specifically, the variables were standardized to bring them to a comparable scale, thereby contributing to a more efficient convergence of the Gibbs sampler.

Based on the values derived from the simulated marginal posterior distributions, the estimated median values and 95% Highest Posterior Density (HPD) intervals are presented in Table 3.1. Notably, the parameter denoted as *home*, representing the disparity between u_h and u_a, exhibits a positive influence on the number of goals scored, characterized by an average estimate of 0.23. The associated 95% HPD interval is delineated as $[0.09, 0.36]$, providing a range within which the true parameter value is likely to reside. It is imperative to highlight that it is the lower value attained by the home effect until now, even though it is positive.

For a nuanced understanding of the factors influencing the number of goals scored, both for the home and away teams, a 'rule of thumb' is applied. Specifically, if the 95% HPD interval includes the value 0, any assertion regarding the existence of a discernible effect becomes untenable. From this perspective, it is arguable that solely variables capturing the overall shape and the dangerousness of a team has a

substantive influence on goal-scoring outcomes. Conversely, variables associated with team discipline appear to exert an insignificant impact on the goal-scoring process. Moreover, the aforementioned inference leads to the conjecture that time may emerge as a pivotal factor in this context, given the temporal dependence of team dynamics. The inference alludes to the notion that the overall team shape, being a time-dependent variable, plays a crucial role in determining goal-scoring patterns.

$\underline{\beta}_{\mathbf{h}}$	**Mean**	**2.5%**	**97.5%**
$\mathbf{u_h}$	0.32	0.225	0.404
RSC	0.17	0.063	0.271
RDC	0.25	0.061	0.450
QO	-0.39	-0.602	-0.172
DeltaElo	0.08	-0.079	0.227
Corners	-0.14	-0.253	-0.040
Yellow Cards	-0.05	-0.147	0.035
Red Cards	-0.01	-0.109	0.071
Shots	0.33	0.231	0.432
Fouls	0.07	-0.019	0.153
Possession	-0.07	-0.195	0.045
Attendance	-0.03	-0.143	0.072

$\underline{\beta}_{\mathbf{a}}$	**Mean**	**2.5%**	**97.5%**
$\mathbf{u_a}$	0.09	0.049	0.185
RSC	0.22	0.061	0.338
RDC	0.13	-0.069	0.377
QO	-0.42	-0.652	-0.193
DeltaElo	-0.08	-0.223	0.061
Corners	-0.15	-0.252	-0.054
Yellow Cards	0.03	-0.057	0.130
Red Cards	-0.02	-0.119	0.077
Shots	0.29	0.194	0.395
Fouls	0.02	-0.071	0.121
Possession	0.004	-0.123	0.135

Table 3.1: Average value attained by the regression coefficient times the standard deviation of the associated independent variable.

In a more formal approach, the selection of the best subset of variables for a model involves addressing a model selection problem, particularly within the Bayesian framework. The goal is to estimate the marginal posterior probability that a variable should be included in the model. Specifically, in the context of your problem, this involves exploring $2^{p_1+p_2}$ different models, where p_1 and p_2 represent the number of covariates used to model home and away scored goals, respectively. It's important to note that the two intercepts are always considered part of the model.

One common method, and arguably the simplest [12], is to assign an indicator function for each variable, say $i_j \ \forall j = 1, \ldots, p_1 + p_2$, where $i_j = 1$ implies that the j-th covariate is included in the model. Let $I_{h,g}$ and $I_{a,g}$ be two diagonals matrices, where the diagonals reflect the values of the corresponding indicator functions, and the first element is always one. The log-linear equations can now be generalized in the following way:

$$\begin{cases} log(\theta_{g,1}) = \underline{x}_{g,h}^T \left(\mathcal{I}_h \underline{\beta}_h \right) \\ log(\theta_{g,2}) = \underline{x}_{g,a}^T \left(\mathcal{I}_a \underline{\tilde{\beta}}_a \right) \end{cases}$$

The most important assumption of the above method is that the fixed effects and the indicators are independent. More complex techniques, such as the **Gibbs variables selection** and the **Stochastic search variable selection**, remove this assumption. So, the prior of the new model can be specified as:

$$\begin{cases} \underline{\beta}_h \sim \mathcal{MVN}_{12}\left(\underline{0}, g\sigma^2\left(X_1^T X_1\right)^{-1}\right) \\ \underline{\tilde{\beta}}_a \sim \mathcal{MVN}_{11}\left(\underline{0}, g\tilde{\sigma}^2\left(X_2^T X_2\right)^{-1}\right) \\ \sigma^2 \sim \mathcal{IG}(a,b) \quad \tilde{\sigma}^2 \sim \mathcal{IG}(a,b) \\ \pi\left(\mathcal{I}_h\right) = \prod_{j=1}^{11} \mathbf{Bern}(p_j) \quad \pi\left(\mathcal{I}_a\right) = \prod_{j=1}^{10} \mathbf{Bern}(\tilde{p}_j) \\ \pi\left(\underline{\beta}_h, \underline{\tilde{\beta}}_a, \sigma^2, \tilde{\sigma}^2, \mathcal{I}_h, \mathcal{I}_a\right) = \pi\left(\underline{\beta}_h|\sigma^2\right)\pi\left(\underline{\tilde{\beta}}_a|\tilde{\sigma}^2\right)\pi(\sigma^2)\pi(\tilde{\sigma}^2)\pi((\mathcal{I}_h)\,\pi((\mathcal{I}_a) \end{cases}$$

In the aforementioned method, the interest is devoted to the marginal posterior distribution associated to the indicators functions. More into detail, the focus is on the posterior inclusion probabilities of the variables. In fact, the number of runs of 0s and 1s in the chains is recorded for each i_j because it s a measure of mixing, where more runs indicate better mixing. An additional noteworthy consideration is the selection of hyperprior values for the inclusion probabilities associated to the indicators, and the existing literature reflects a degree of division on this matter. Optimal hyperprior values play a crucial role, influencing the sparsity or complexity of the resulting models. If set too low, sparser models may be favored, whereas excessively high values could lead to overly complex models. Notably, assigning a value of 0.5 is discouraged as it may, as demonstrated by [14], result in selecting approximately half of the covariates involved in the model. Thus, the determination of an appropriate value demands careful investigation. In particular, after some trials, I have decided to fix the value of the prior inclusion probabilities equal to 0.3 due to the fact that it ensures a good mixing and a relative sparse solution.

Upon scrutinizing the analysis depicted in Figure 3.3, the estimated posterior inclusion probabilities for various variables associated with home and away scored goals come to the forefront. Evidently, it is discernible that the crucial variables influencing both home and away scored goals remain consistent, with the notable exception of the *recent defensive condition*, which appears to be excluded from the consideration in the context of away scored goals. This observation implies a distinct influence of this particular variable in the home scoring dynamics. The analysis strongly suggests that the selected variables are intricately linked to the overall configuration of a team, with a noteworthy observation being the relative insignificance of disciplinary factors. This insight underscores the notion that team dynamics and structure play a more prominent role in influencing scoring outcomes compared to disciplinary considerations.

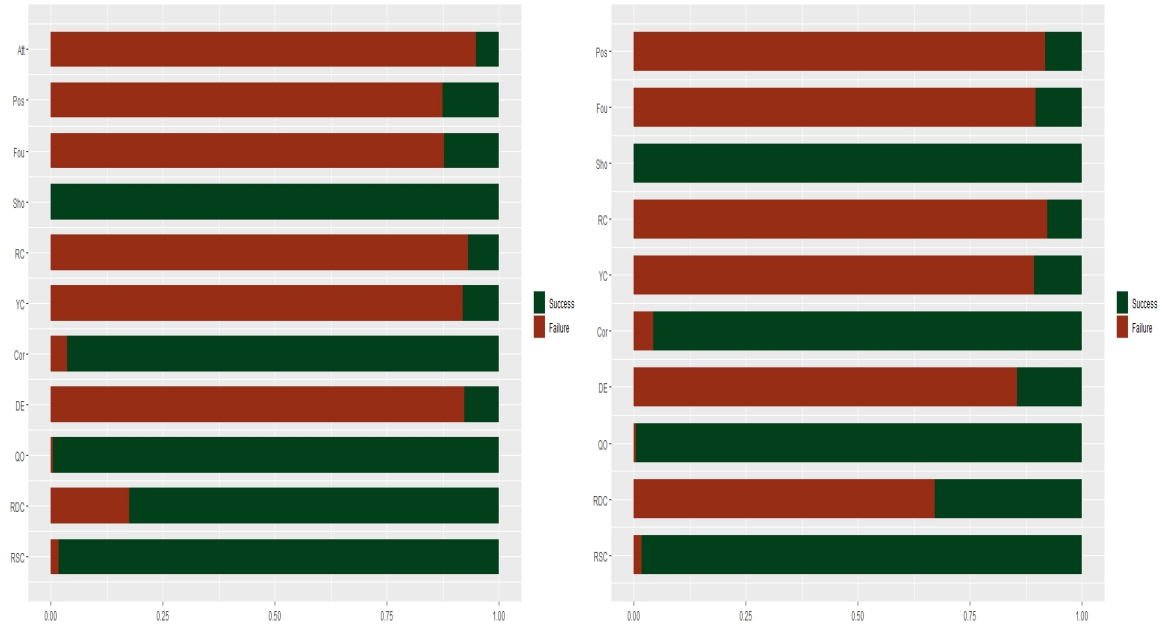


Figure 3.3: Posterior inclusion probabilities related to the variables used to model home scored goals (left) and away scored goals (right).

Once the important variables have been selected, a new model is constructed based on the configuration derived from the previous analysis. Specifically, the modeling of home scored goals will consider only five variables, while for away scored goals, four variables will be considered, each accompanied by the intercept term. By focusing on a reduced set of influential factors, the model is poised to offer insights into the essential elements governing the scoring dynamics in both home and away contexts. In comparing the posterior estimation of the regression coefficients (Figure 3.2), a notable concordance is discernible in relation to those obtained from the previous model. However, an anomalous observation pertains to the variable "corner", exhibiting a tendency towards negative values, contrary to intuitive expectations.

Initially, a hypothesis was entertained suggesting a potential latent effect influencing the relationship between the number of shots and the count of corners. Subsequently, I endeavored to reanalyze the data, excluding shots from the model but the sign of the coefficient for the variable "corner" remained negative, contradicting intuitive assumptions. Furthermore, upon the removal of shots from consideration, the number of corners no longer exhibited statistical significance based on posterior inclusion probability. It would be of interest to deeply analyse this strange behaviour to understand more.

In conclusion, It is noteworthy that the DIC for the new model registers at 2089.2, representing the lowest value encountered thus far. Notably, the posterior distributions associated with σ^2 and $\tilde{\sigma}^2$ exhibit comparability, suggesting stability in the variance

components. This observation underscores the potential efficacy of the new model in capturing the underlying data structure and warrants a comprehensive examination of the "corner" variable's impact.

$\underline{\beta}_{\mathbf{h}}$	**Mean**	**2.5%**	**97.5%**
$\mathbf{u_h}$	0.328	0.236	0.416
RSC	0.172	0.086	0.258
RDC	0.255	0.060	0.443
QO	-0.408	-0.606	-0.200
Corners	-0.163	-0.257	-0.066
Shots	0.318	0.223	0.411

$\underline{\beta}_{\mathbf{a}}$	**Mean**	**2.5%**	**97.5%**
$\mathbf{u_a}$	0.100	0.001	0.200
RSC	0.213	0.133	0.298
QO	-0.252	-0.349	-0.150
Corners	-0.173	-0.264	-0.081
Shots	0.299	0.199	0.399

Table 3.2: Average value attained by the regression coefficient times the standard deviation of the associated independent variable.

4 Static Models: results and research questions

The following section compares the results of previous models and attempts to answer some interesting questions. The section is organized as follows: for each research topic, the performance of different models will be compared (when possible) to understand which one is better suited to answer it. Generally, each question could be classified under two main categories: 'Estimation' or 'Prediction'. The estimation folder displays an examination related to the posterior distribution of the parameters and the model's deviance. The prediction subsection displays questions related to the posterior predictive distributions, including predictive checks. These two folders are useful for understanding the behaviour of different models in various aspects, which can help in defining a direction for building a new model.

4.1 Estimation

It is important to remark one point before moving on. All the parameters that are analyzed below have an influence on the propensity of a team to score (if the team plays home $\theta_{g,1}$, otherwise $\theta_{g,2}$). Without any loss of generality, I can claim that the prior distribution for the induced parameter is non-informative at all, while the posterior distribution tends to assume a shorter range of values, say between zero and six. Due to the high number of parameters (760), it is not possible to conduct a thorough analysis or comparison of the different models for the induced parameters, but it is possible to observe one prior to posterior update. The log-scale representation was chosen due to the nature of the values assumed by the prior distribution.

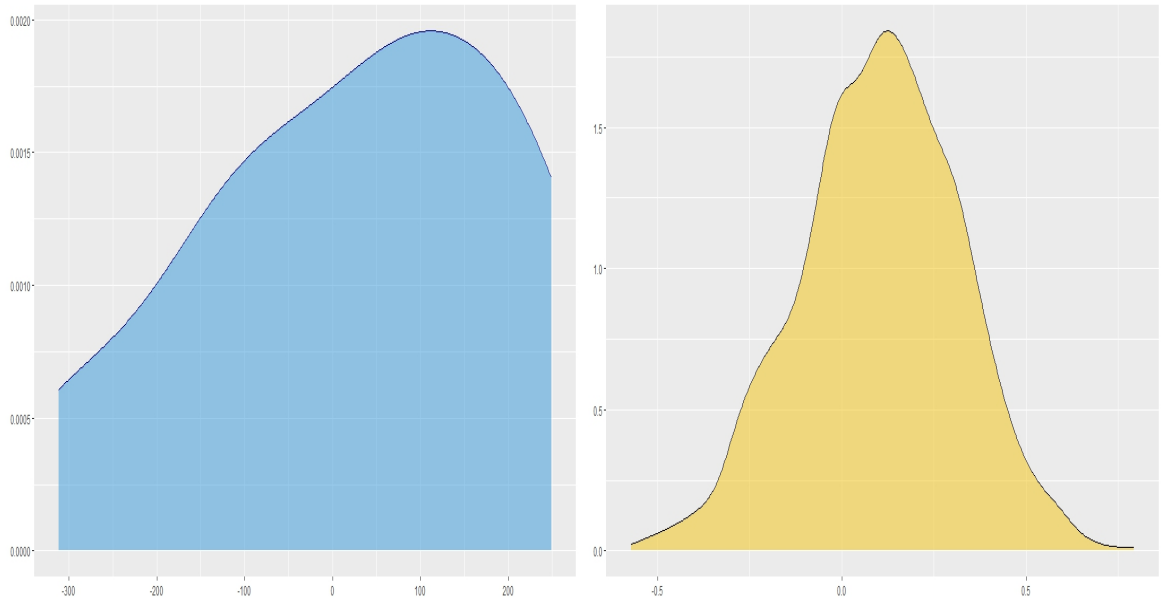


Figure 4.1: Prior to posterior update of the home theta parameter for 10_{th} game of the season in log scale.

Figure 4.1 illustrates the prior-to-posterior update of $\theta_{g,1}$ and $\theta_{g,2}$ for the tenth game of the season. This demonstrates the point I was trying to make earlier. In conclusion, the narrow range of values in the prior distribution is unusual. However, it is unlikely to affect the rest of the analysis and it worths to mention it.

4.1.1 Which is the impact of the offensive and defensive parameter?

Prior to estimating the random effects, it is crucial to comprehend the behaviour of the linked hyperprior parameters, particularly the mean and variance. It is assumed that the offensive and defensive random effects for all models originate from the same hyperprior distribution with the same structure. Regarding the mean, I have tried to different approaches, i.e. fixing it directly to zero or using a Gaussian distribution. Without any loss of generality, it can be argued that the estimated random effects are the same. Regarding the hyperprior variance of the offensive and defensive random effects, it can be proved that it tends to take values between zero and one in the posterior simulation.
After drawing attention to the higher level of the structure, let's now move on to analysing the offensive and defensive parameters. One important point to note is that the prior distribution of both offensive and defensive parameters is centered around zero with a high variance in order to ensure non-informativeness:

- taking into account the hierarchical model without covariates, analysing the net offensive and defensive effects (Figure 4.2), Manchester City (the league winner) has the highest offensive propensity (posterior mean of 0.56), followed by Liverpool (posterior mean of 0.49) and Arsenal (posterior mean of 0.32). The top four clubs (Manchester City, Liverpool, Chelsea and Totthenam) have the lowest propensity to concede goals, while Fulham, Huddersfield Town and Bournemouth have the highest. Note that the first two teams were actually relegated at the end of the season;

- concerning the offensive and defensive random effects, when covariates are considered, much of their effectiveness has been lost, reducing to very small values (Figure 4.3). In particular, the highest offensive propensity is owned by Manchester City and it is equal to 0.027, which is quite different from all the other results. The worst defending teams are Huddersfield Town, Fulham and Cardiff City, which present respectively an average effect of 0.043, 0.032 and 0.016. However, these values are not highly reliable since, by studying the confidence intervals and the traceplot, it can be seen that 0 is always a plausible value. Heuristically, the explanation is that the previously possessed effect was somehow incorporate into the other variables involved in the model;

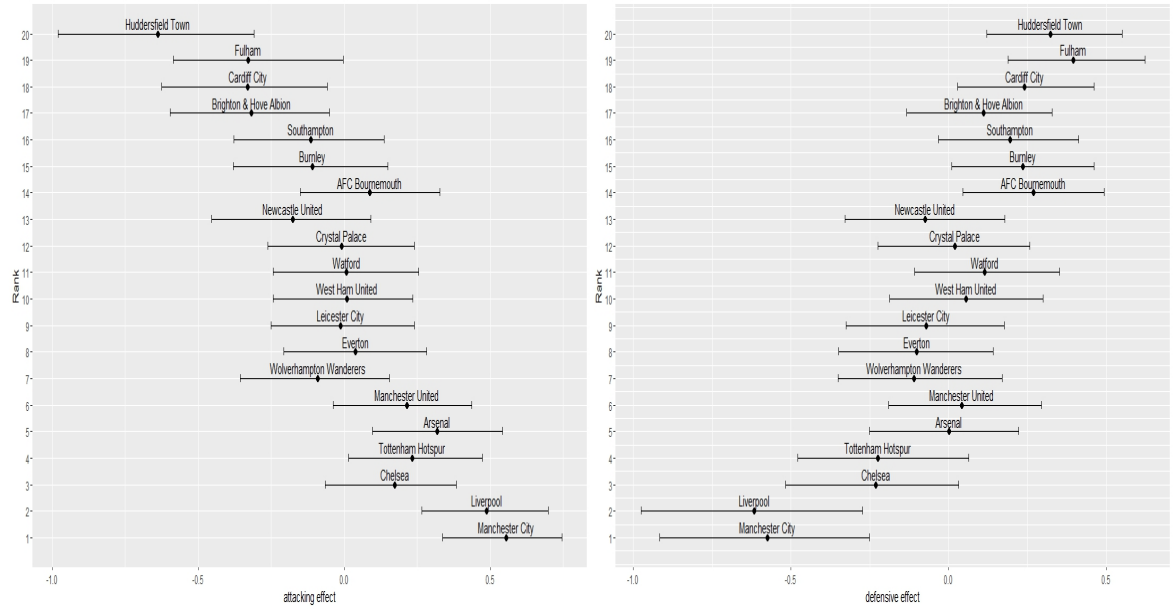


Figure 4.2: 95% high posterior density intervals for the attacking effect (left). 95% high posterior density intervals for the defensive effect (right). Dots represent the average effect.

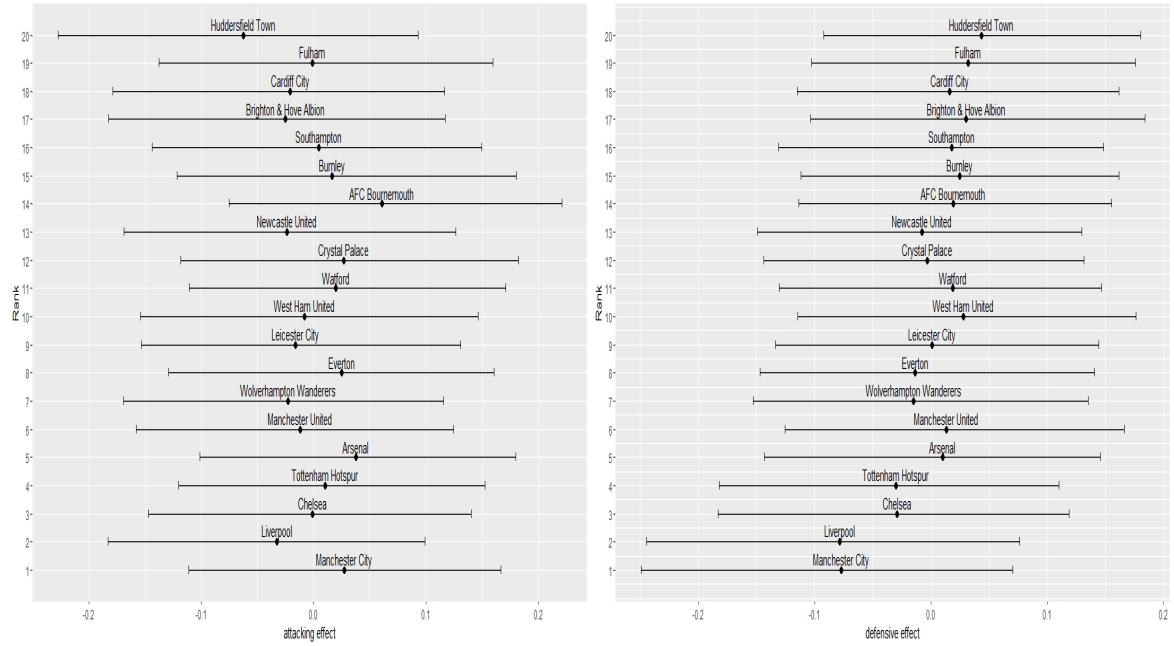


Figure 4.3: 95% High posterior density intervals for the attacking effect (left). 95% High posterior density intervals for the defensive effect (right). Dots represent the average effect.

- for the remaining models, the offensive and defensive parameters display almost identical average values, alongside 95% confidence intervals, compared to those observed in the first model (see Figure 4.2). Due to the numerous similarities between those models, this outcome is intuitive. Actually, Manchester City and Liverpool maintain the most potent attacking and defensive capabilities, whereas Huddersfield Town, Cardiff City, and Fulham continue to show the weakest performances.

In conclusion, the offensive and defensive parameters represent the team's actual strength. The use of covariates can capture the pattern described earlier. This means that observable variables can provide the same contribution, albeit with a different interpretation, even if random effects can be used.

4.1.2 Does a "home" effect really exist? Is it linked to the attendance?

Determining the existence of a 'home' effect is a crucial question, as evidenced by several papers [1] [11] [5]. Additionally, I investigated whether this effect is correlated with stadium size. Upon closer examination, it became apparent that every model tested revealed a positive effect when teams played in their home stadium. To proceed straightforwardly, let us first visualise the prior distribution for the home parameter. From Figure 4.4, it is possible to understand which are the values attained by the prior distributions associated to the home effect.

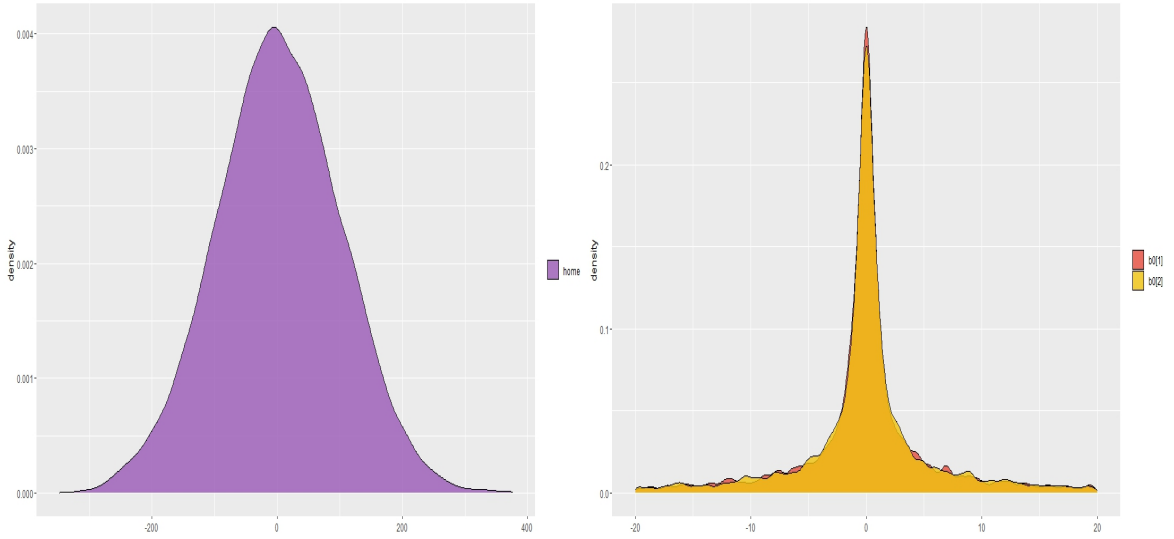


Figure 4.4: Prior distribution for the home parameter (left) and for the random effects associated to the attendance in the model with covariates (right)

In particular:

- considering the hierarchical model, the home effect attains an average value of 0.36 and the 95% highest posterior density interval is $[0.28, 0.45]$;
- in the model with covariates, the two parameters which represent the home effect are still positive. In particular, the one related to teams with a smaller stadium is on average 0.186 (95% HPD interval $[0.05, 0.32]$), while the one related to teams with a bigger stadium is 0.23 (95% HPD interval $[0.08, 0.40]$). After observing these values, two considerations are possible. The first concerns the fact that indeed there is a different, though not large, effect associated with the type of possessed stadium. Unfortunately, the value for the home effect is smaller than

in the first model. However, this is due to other circumstances, such as the inclusion of covariates;

- when two inflation factors are considered, the home impact has average value of 0.37 and a 95% HPD interval of $[0.28, 0.46]$;
- once the number of inflation factors is increased, it is evident that playing at home still has a positive impact with an average of 0.41 (95% HPD intervals $[0.32, 0.49]$);
- in the last model, it has an average value of 0.45 with 95% HPD intervals of $[0.36, 0.55]$, which is the highest value observed in the model to date.

Figure 4.5 shows the posterior distribution of the parameters mentioned above. The variance of the posterior is lower than the prior, as expected, and they are no longer centered at zero. In conclusion, it can be claimed that there is a home advantage effect that benefits the team playing in its own stadium. Additionally, the size of the stadium, which is linked to attendance, appears to have an impact on this effect.

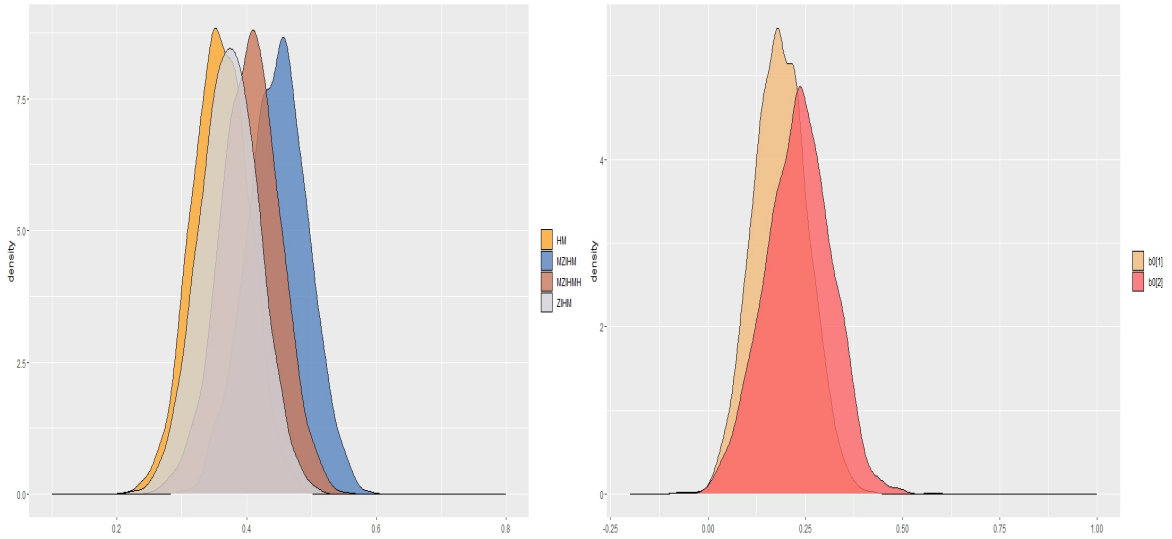


Figure 4.5: Posterior distribution for the home parameter (left) and for the random effects associated to the attendance in the model with covariates (right).

4.1.3 Are covariates really useful to understand the outcome of a football match?

To answer the question, only one model will be considered as it is the only one that directly uses covariates that affect the number of scored goals. As all parameters have the same prior distribution, only one simulation will be displayed. This distribution is also associated with the home effect, and a simulation can be seen in Figure 4.4. Regarding the fixed effects of the model:

- the parameter associated to the *Recent scoring condition* of the team has a positive average value of 0.11 (95% HDI $[0.01, 0.20]$). This is in agreement with what was said in the descriptive part, i.e. the number of goals scored in the past games by a team has a positive influence on the number of goals scored in the next game;

- going on, the parameter associated to the *Recent defensive condition* of the opponent has a positive average value of 0.42 (95% HPD interval $[0.31, 0.53]$). Again, this makes a lot of sense with what has been said previously. The more the opposing team has conceded goals in previous games, the greater the team's propensity to score;

- the parameter associated to the *Quality of the opponent* has a negative average value of -0.44 (95% HPD interval $[-0.53, -0.35]$). The greater the number of points accumulated by the opponents in the previous games, the lower the offensive propensity of the team is;

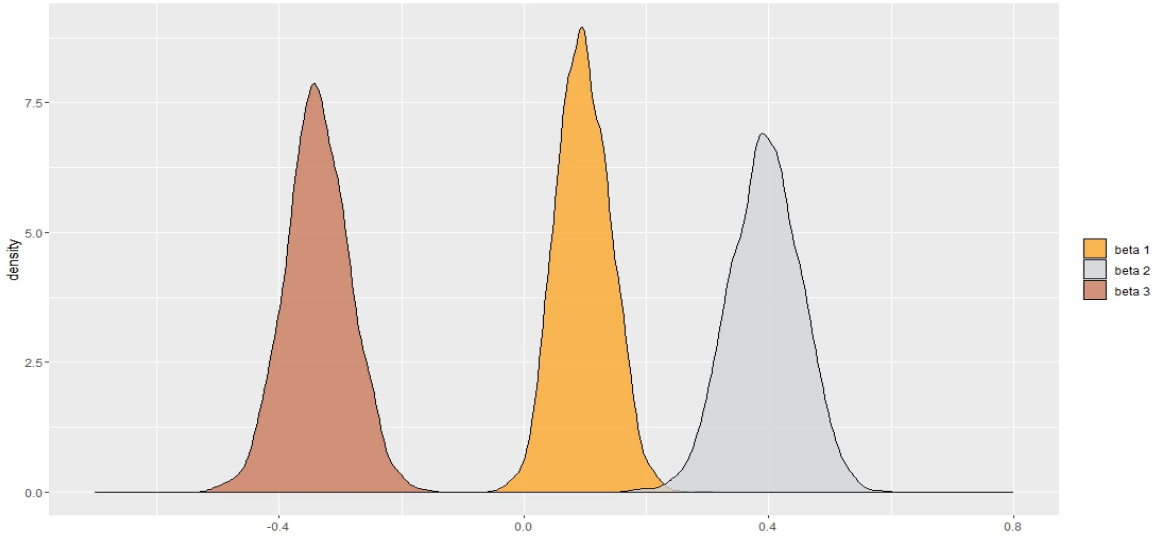


Figure 4.6: Posterior distribution associated to the coefficients of the three analyzed covariates

- the parameter linked to the difference between the *Elo rankings* of the two teams has a small positive average worth of 0.00035. It's estimated value does not represent at all what is shown in Figure 2.11, In reality, it can confidently be said that it is almost zero. This assertion is supported by Figure 4.7's traceplot, which indicates that the simulated values cross frequently zero.

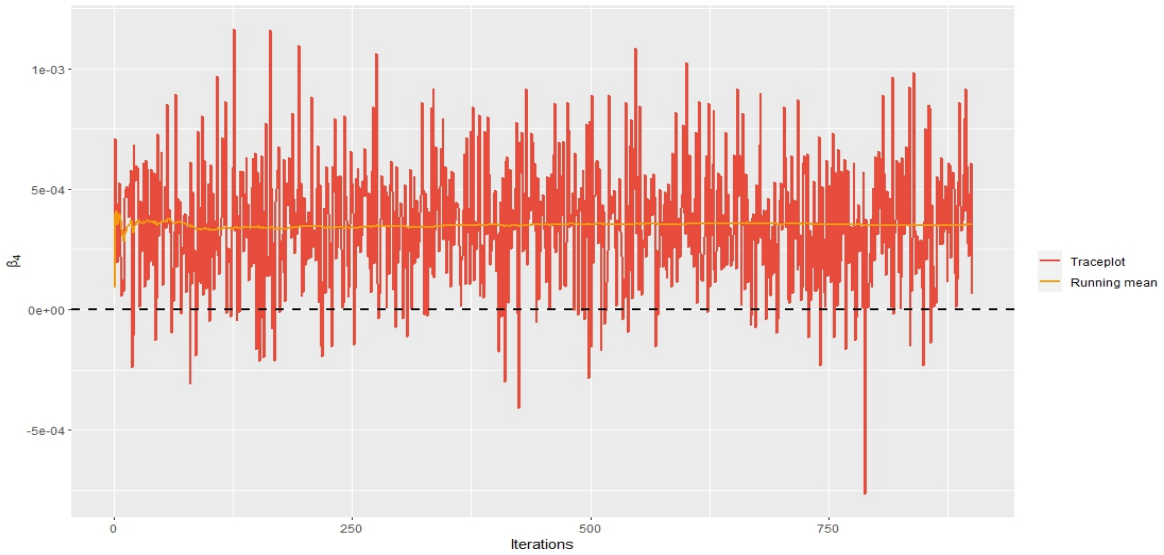


Figure 4.7: Traceplot and running mean for the parameter β_4. The dotted line comes from the equation $y = 0$

In Figure 4.6 the posterior distributions associated to the first three coefficients are placed. I have decided not to include the coefficients linked to the difference of Elo rankings because it is on a completely different scale and furthermore it can be regarded as a sort of degenerate posterior distribution in 0.
In conclusion, it can be argued that covariates are useful in understanding the outcome of a football match, despite their limitations. The variables associated with the first three coefficients have a zero value for the first 50 games due to being moving averages. This may introduce bias in the analysis that cannot be compensated for by any other variable, as they are not currently being considered. Improvements can be made in this area.

4.1.4 Inflation factors: a comparison.

Only [11] has considered the use of a zero-inflated Poisson (ZIP) distribution to model the number of scored goals. As explained in the theoretical section, three models directly use the inflation factors:

- considering the model with two inflation factors, in Figure 4.8 it is possible to visualize the prior-to-posterior update of the involved parameters. Going into more detail, the mean value of p_1 is 0.016, with a 95% HPD interval of [0,0.04], while the mean value of p_2 is 0.012, with a 95% HPD interval of $[0, 0.03]$. Though the effects are marginal, they appear to impact the number of goals scored, albeit it cannot be ascertained if they are distinguishable from zero;

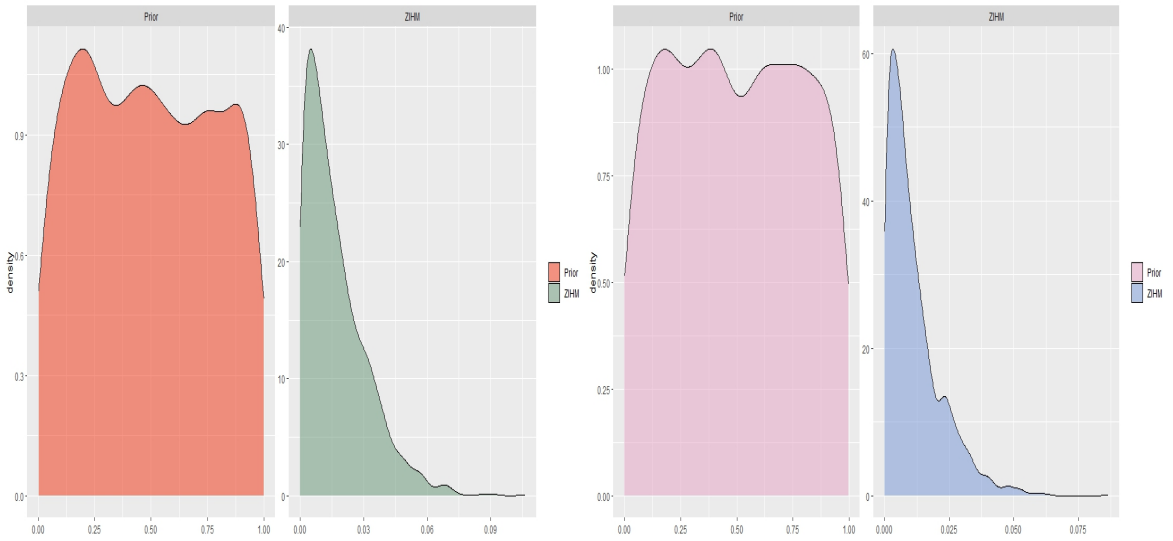


Figure 4.8: Prior to posterior update of the inflation factors linked to the home (left) and away (right) scored goals.

- concerning the model which considers the inflation factors coming from a common population, let's first of all visualize the prior to posterior update of the hyperparameters involved. Analysing Figure 4.9, it is quite easy to understand that the value attained by the population mean tends to be concentrated at zero. This will have a huge influence on the values assumed by the inflation parameters because they will be shrank toward zero. Furthermore, also the hyperposterior variance is quite low, concentrated around 0.1.

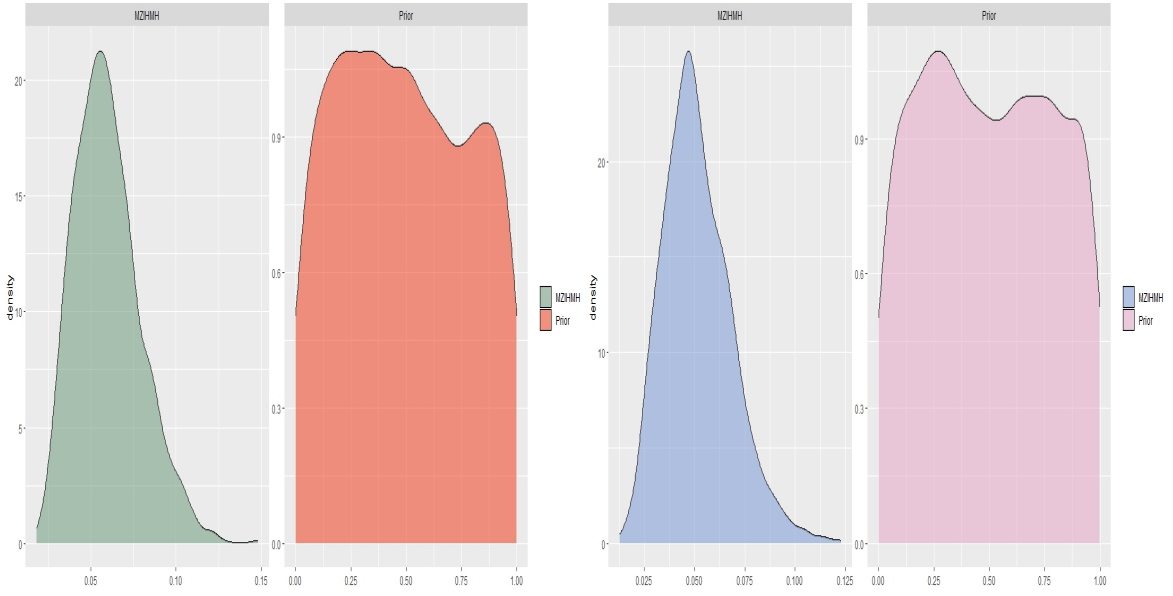


Figure 4.9: Prior to posterior update of the common population mean used to model to the home (left) and away (right) scored goals.

Considering the results in Figure 4.10, as expected, the values attained by the different inflation parameters, both for home and away scored goals, are very

similar and concentrated around 0. Without any loss of generality, it is not possible to capture the propensity of a team to score 0 goals assuming a common population;

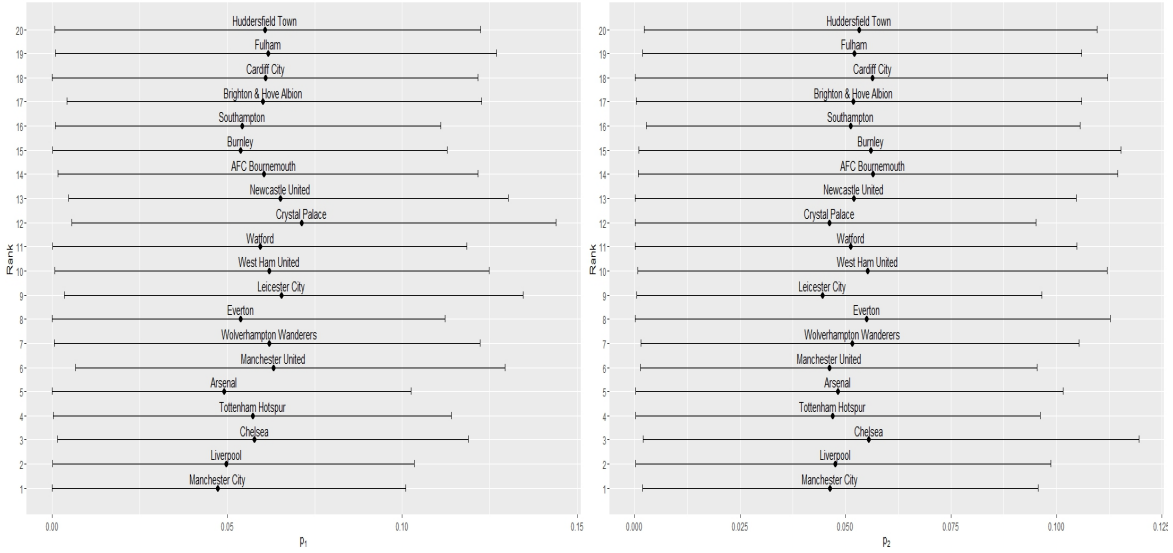


Figure 4.10: 95% HPD intervals for the inflation parameters linked to home (left) and away (right) scored goals.

- the inflation factors displayed in Figure 4.11 showcase a broad range of values, spanning from 0.032 to 0.3. Notably, the coefficients relating to Manchester City, Arsenal, and Liverpool assume the lowest values during both home and away matches. Nevertheless, the highest coefficient values introduce a somewhat intricate situation. Consider the case of Crystal Palace as an example. The team has the largest inflation factor for home goals scored and one of the smallest for away goals scored. The coefficients' size in Figure 2.1 shows the team's likelihood of scoring zero goals in a match is quite different w.r.t the considered team. It could be contended that the values of the two coefficients in the earlier model were markedly influenced by variations in propensities of different teams. Conversely, a more precise depiction of the scenario can be attained here.

In conclusion, it seems that letting the coefficients "free", i.e. not assuming a common underlying population, has allowed to capture the different propensities of not scoring. So, without any loss of generality, it can be claimed that the best choice is to use 20 different coefficients for home (away) scored goals.

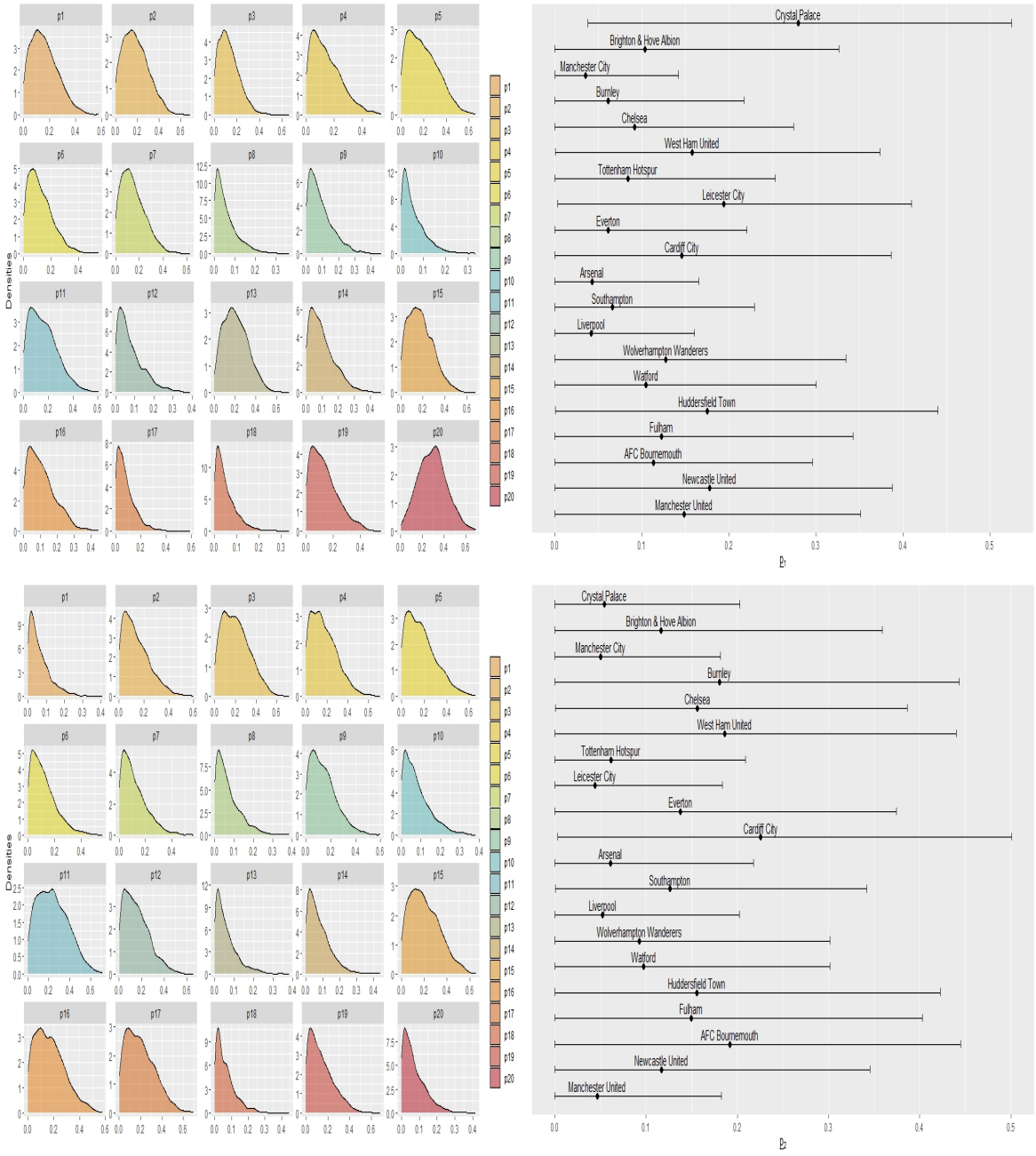


Figure 4.11: Marginal posterior densities for $p1, w$ (top left) and95% HPD intervals (top right). Marginal posterior densities for $p_{2,w}$ considering (bottom left) and 95% HPD intervals (bottom right). Dots represent the average effect.

4.1.5 Deviance of the models: a comparison among different trials.

The Appendix A displays the deviance of the five models. The first line shows the deviance of the Bayesian Hierarchical model (left) and the model with covariates (right). The second line shows the deviance of the Zero-Inflated model (left) and its modified version using hyperpriors (right). The last line shows the deviance of the modified Zero-Inflated model without the use of hyperpriors. Table 4.1 provides additional information:

- the hierarchical model with covariates has the lowest DIC, indicating that it is

the optimal model in terms of balancing optimality and complexity;

- the model with the worst performance is the Modified Zero-Inflated model with Gamma hyperpriors, as indicated by its DIC value of 2354.9;
- if the inflation factors are not assumed to come from the same population, a lower DIC of 2260.35 is obtained.

	H.M	H.M.C	Z.I.H.M	M.Z.I.H.M.H	M.Z.I.H.M
DIC	2220.65	2114.37	2322.28	2354.9	2260.35

Table 4.1: Deviance Information Criteria (DIC) for the different models.

4.2 Prediction

4.2.1 Which is the behaviour of our model during the season?

To answer the question, this subsection will display posterior predictive checks. For each team, it will show an estimated rank and their performance in some interesting games. The median will be used as a sufficient statistic to represent the entire posterior distribution.

- **Hierarchical model**
  Comparing Table 2.2 and Table 4.12 (top left), it is possible to affirm that a general underestimation effect is acting behind the scene. The scored and conceded goals tend to be concentrated around the average value, i.e. the variance is quite low. This has a huge effect also on the estimated points gained by a team, which are usually lower than the actual points. As pointed out by the authors:

  > 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 observations. [1]

  It is a well know phenomenon in Bayesian statistics and it is usually linked with the Normal distribution.
  Moving forward, I would like to focus on the performances of two teams, Cardiff City and Arsenal, which are the extreme cases. More into detail, the first team suffered the largest underestimation effect (13 points) while the other the largest overestimation one (4 points). Analysing the scored goals of the first team, I have noticed that it tends to score 0 goals in the 45% of the matches and 1 or

more goals in the remaining 55% of the games. Instead, the model predict that it has scored 0 goals in the 23% of the games while only 1 goals in the others. As can be seen, the model tends to overestimate the probability that Cardiff City scores 1 goal in a match, while it completely underestimates the other probability. In the table 4.2 two emblematic matches are shown with regard to the above situation.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Cardiff City	West Ham United	2	0	0.32	0.40	0.28	1	1
Manchester United	Cardiff City	0	2	0.71	0.12	0.17	2	1

Table 4.2: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

Following the same path also for Arsenal, it can be shown that the model tends to underestimate the event of scoring 0 goals or more than 2. On the other hand, it has the tendency to overestimate the event of scoring 2 or 3 goals. Unlike the previous case, it is apparent that the model can tell when Arsenal wins, loses or draws but still has difficulty predicting the exact outcome. From table 4.3, it seems that errors in goal estimation for one team, as is easily guessed, also affect the prediction of other teams.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Arsenal	Crystal Palace	2	3	0.61	0.18	0.21	2	1
Southampton	Arsenal	3	2	0.30	0.47	0.17	1	2

Table 4.3: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

As you can clearly see from Figure 4.12, the difference between the two cases is obvious. Regarding the Cardiff situation (left), a strong discrepancy can be seen between what is estimated and what is observed. On the other hand, Arsenal's situation seems to be more crystal clear. In fact, the estimates obtained turn out to be in agreement with what is actually observed.
The idea I developed after commenting on the results of this first model is that the worst teams are penalized more by the fact that our model has downward estimates for goal scored and conceded.

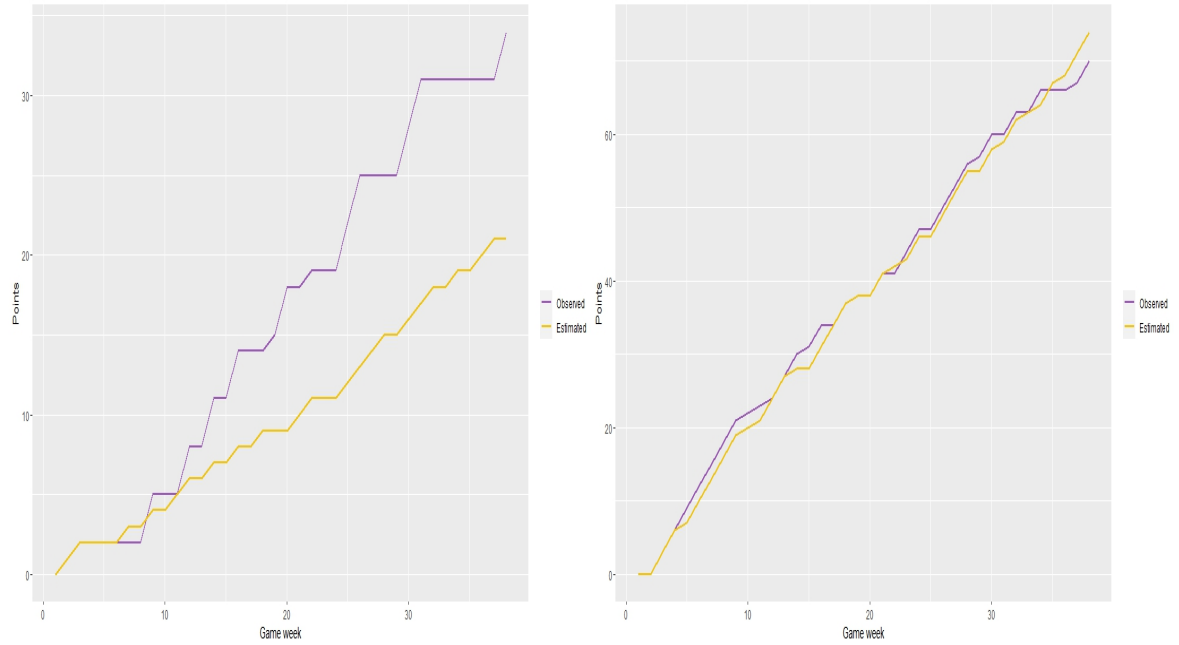


Figure 4.12: Observed vs Estimated points. On the left it is placed the plot for Cardiff City while on the right the Arsenal one.

- **Hierarchical model with covariates**

  Analysing Table 4.12 (top center) it is possible to affirm that, even if a general underestimation effect on the number of scored goals is still there, it has been mitigated. Teams whose gained points are always underestimated (see Wolverhampton, Cardiff City, West Ham) are ranked better by the model, which seems to be able to recognize on average the offensive and defensive strength of a team. Of course, far be it from me to claim that the situation is perfect, yet clear improvements can be discerned.

  Once again, I would like to focus on the performance of two teams which are somehow iconic. On one hand Manchester City, whose general effects have been underestimated. In particular, the general estimated points are 7 less than the observed. Comparing the relative frequencies of the observed and estimated scored goals, it is easy to notice an overestimation of the event {Scores 0 goals}. From Table 4.4, it is possible to notice the aforementioned effect. To be consistent with the discussion, these two games were played during the first 5 weeks, where the information I have got is "poor".

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Arsenal	Manchester City	0	2	0.39	0.29	0.32	1	1
Manchester City	Huddersfield Town	6	1	0.48	0.28	0.24	1	1

Table 4.4: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

On the other hand West Ham United, whose general effects have been estimated in a proper way. Unlike the previous case, the event {Scores 0 goals } has been heavily underestimated, as well as the event {Scores more than 2 goals}. Into the Table 4.5 are reported 2 interesting games, where the performances of West Ham United were overestimated.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
West Ham United	Watford	0	2	0.61	0.2	0.32	2	1
Burnely	West Ham United	2	0	0.22	0.22	0.56	1	2

Table 4.5: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

- **Zero-Inflated Hierarchical model**
  Looking at Table 4.12 (top right) and Table 2.2, there is a widespread underestimation of the goals scored and conceded by each team, which has a negative impact on their total points. There is a tendency for teams to be underestimated in terms of points, with the exception of the highest ranked teams. Comparing those results with the baseline model (top left), we can see that the number of goals is concentrated around the mean, indicating a low variance. To be coherent with the discussion, I want to focus on the performances of two teams. Firstly Arsenal, which is the team with the greater overestimation effect in terms of points won (+4). More specifically, the model seems to slightly overestimate the probability that the team will score 2 goals, while perfectly estimating the probability that they will score 1 goals. As can be clearly seen from the table 4.13, when Arsenal plays mid-low ranked teams, the model tends to give them a higher propensity to win. This can penalise other teams in the long run, making them look "poorer" (in terms of attacking and defensive strength) than they actually are.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Arsenal	Brighton & Hove Albion	1	1	0.72	0.17	0.1	2	0
Leicester City	Arsenal	3	0	0.38	0.34	0.28	1	1

Table 4.6: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

Secondly Watford, which is the team that suffers the greater underestimation effect in terms of points gained (-15). For this team, the model seems to be quite good when it has to predict 2 as the number of goals scored, while it has a general overestimation effect for 0 and 1. As you can see from table 4.7,

Watford has a prediction overestimate of $1-1$, which is the most predicted result for this team.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Watford	Tottenham	2	1	0.35	0.40	0.25	1	1
Cardiff City	Watford	1	5	0.30	0.41	0.30	1	1

Table 4.7: Comparison between observed result and estimated result (median of the samples). **W**, **L** and **D** represent respectively the expected probability to win, to lose and to draw.

- **Modified Zero-Inflated Hierarchical model**
  Figure 4.12 (bottom left) illustrates a consistent underestimation effect on the number of points gained and the number of goals scored. Additionally, it appears that the current situation is not significantly distinct from the previous model, suggesting that perhaps the enhancements have not been attained. Before progressing further in the argument, I would like to assess the performance of two teams, as usual. Arsenal displays an overestimation effect when it comes to the number of gained points (+4). The posterior marginal distributions of scored goals demonstrate an overestimation of the event where the team scores 2 goals during home matches, while the other events are underestimated. However, the posterior predictive for away goals suggest that the model favours the event where the team scores 0 goal over all others. Analysing Figure 4.13, one can observe the underestimation effect discussed earlier. However, the points gained remain coherent for all the games, except for the last four.

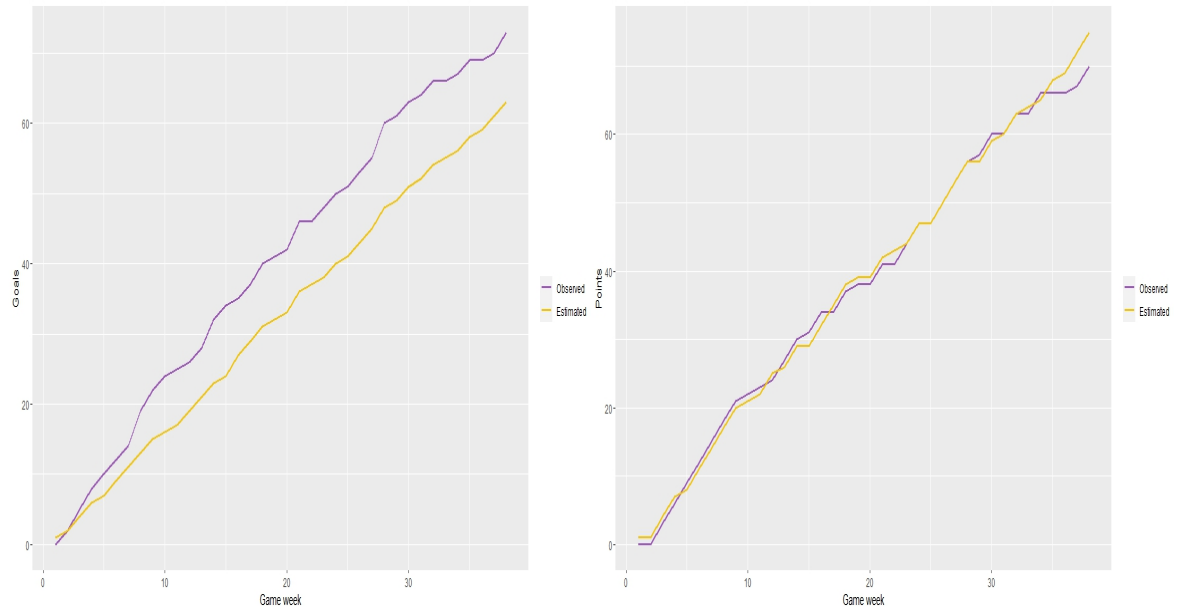


Figure 4.13: Estimated cumulative goals vs observed (left). Estimated cumulative points vs observed (right).

In Table 4.8, some intriguing games are presented that help comprehending my

prior argument. In the first game, Arsenal's ability may have been overestimated, whereas in the second game, it may have been underestimated.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Arsenal	Crystal Palace	2	3	0.60	0.16	0.24	2	1
Fulham	Arsenal	1	5	0.21	0.54	0.21	1	2

Table 4.8: Comparison between observed results and estimated results (median of the samples from the posterior predictive distribution). **W**, **L** and **D** represent respectively the expected probability to win, lose and to draw.

Conversely, Huddersfield Town exhibits a underestimation effect with regards to the points obtained (-6). In terms of the posterior predictive distribution for goals scored at home, it can be said that the model accurately predicts the club's goals. This is due to the fact that the team only scores either 0 or 1 goals when playing at their home stadium. However, the model has a tendency to overestimate the number of 0 goals scored when the team is playing away. Again, this is because the club highlighted in the observed games tends to score more when playing away from its home stadium. From Figure 4.13, it is apparent that the underestimation of goals scored, particularly when the team is playing away, has a negative impact on the estimation of the total points gained.

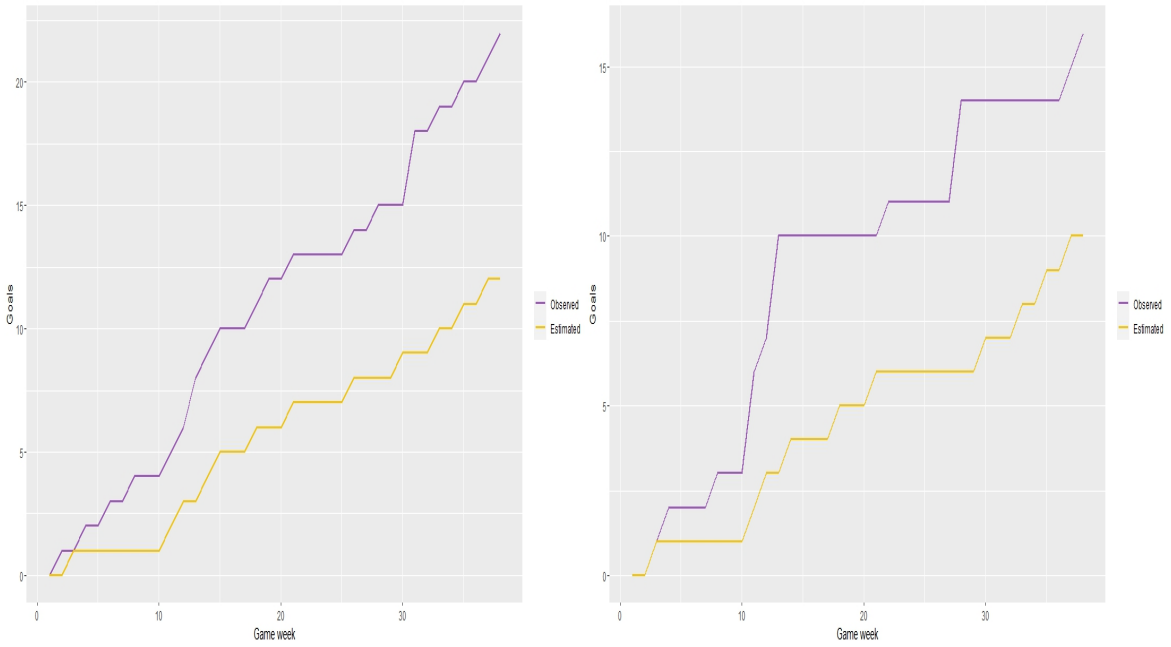


Figure 4.14: Estimated cumulative goals vs observed (left). Estimated cumulative points vs observed (right).

In Table 4.9, there are two noteworthy matches. Specifically, the first match shows a goodness of fit score of zero goals, while the second match indicates a reduced likelihood of scoring more than zero goals, as suggested by the preceding discussion.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Huddersfield Town	Tottenham Hotspur	0	2	0.11	0.65	0.24	0	2
Wolverhampton Wanderers	Huddersfield Town	0	2	0.50	0.17	0.33	1	0

Table 4.9: Comparison between observed results and estimated results (median of the samples from the posterior predictive distribution). **W**, **L** and **D** represent respectively the expected probability to win, lose and to draw.

- **Modified Zero-Inflated Hierarchical model without hyperpriors**
  Upon comparing Table 4.12 (bottom center) and Table 2.2, the model appears to adequately predict the final rankings of the English Premier League, with a consistent underestimation of scored goals. In particular, it fails to fully recognize the true final position of mid-ranking teams, which is quite understandable. As noted in the descriptive section, the majority of these teams share similar characteristics. In conclusion, despite the typical challenges, I can confirm that the situation has improved in comparison to the previous model. Once again i would like to focus on the specific case of two teams, which represent two far away situations. Tottenham, which experiences the largest overestimation effect in terms of points (+7), underwent underestimation in terms of the number of goals scored and conceded, as every other team did. The analysis of the posterior predictive distribution regarding the number of scored goals both at home and away demonstrates that there is a general overestimation of the impact when the team scores one goal, while there is an underestimation of all other events. Examining Figure 4.15, it is evident that the estimated cumulative distribution for the scored goals tends to underestimate the real effect. Additionally, the number of points appears to be less coherent with the observations after the first 27 matches.

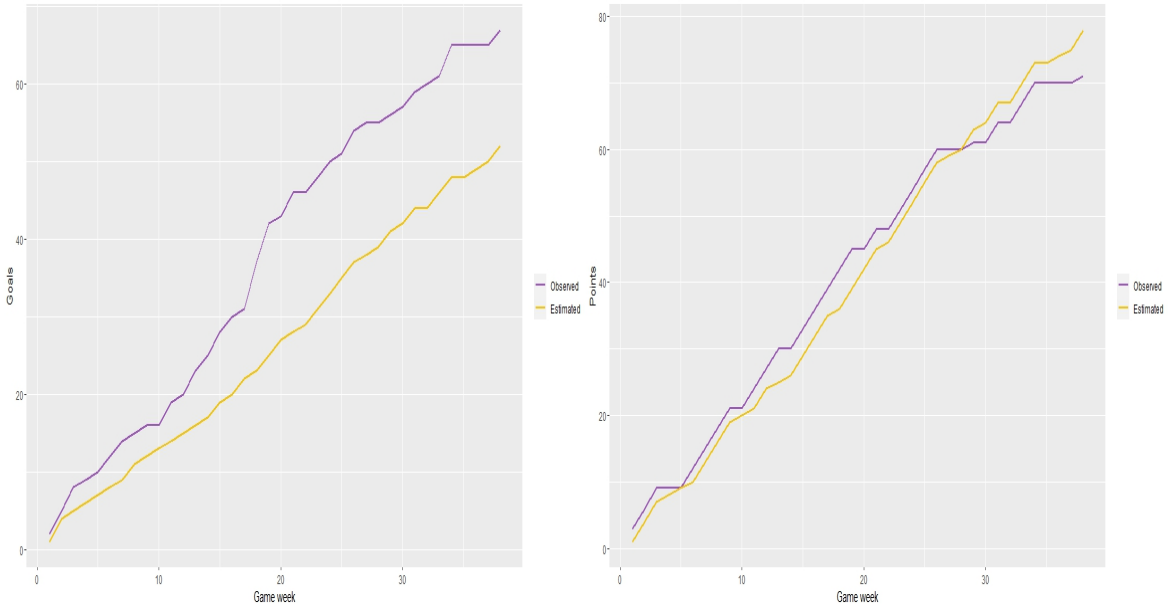


Figure 4.15: Estimated cumulative goals vs observed (left). Estimated cumulative points vs observed (right).

Once again, attention is drawn to two noteworthy matches that improve understanding of the situation at hand. Table 4.10 displays the aforementioned matches where the model has overestimated Tottenham's ability.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Tottenham	Wolverhampton Wanderers	1	3	0.61	0.16	0.23	2	1
Burnley	Tottenham	2	1	0.26	0.49	0.25	1	1

Table 4.10: Comparison between observed results and estimated results (median of the samples from the posterior predictive distribution). **W**, **L** and **D** represent respectively the expected probability to win, lose and to draw.

Wolverhampton Wanderers suffer the greatest underestimation in points, with a difference of -15. Upon examining the posterior predictive distribution of goals scored, it can be argued that there is a consistent overestimation of the occurrence of one goal being scored, a reasonable estimation for zero goals being scored and an underestimation of all other events. Observing Figure 4.16, it is evident that both the estimated cumulative distribution for the points and the scored goals are below the observed levels, leading to the conclusion that in this instance, the underestimation is stronger as compared to the previous case.

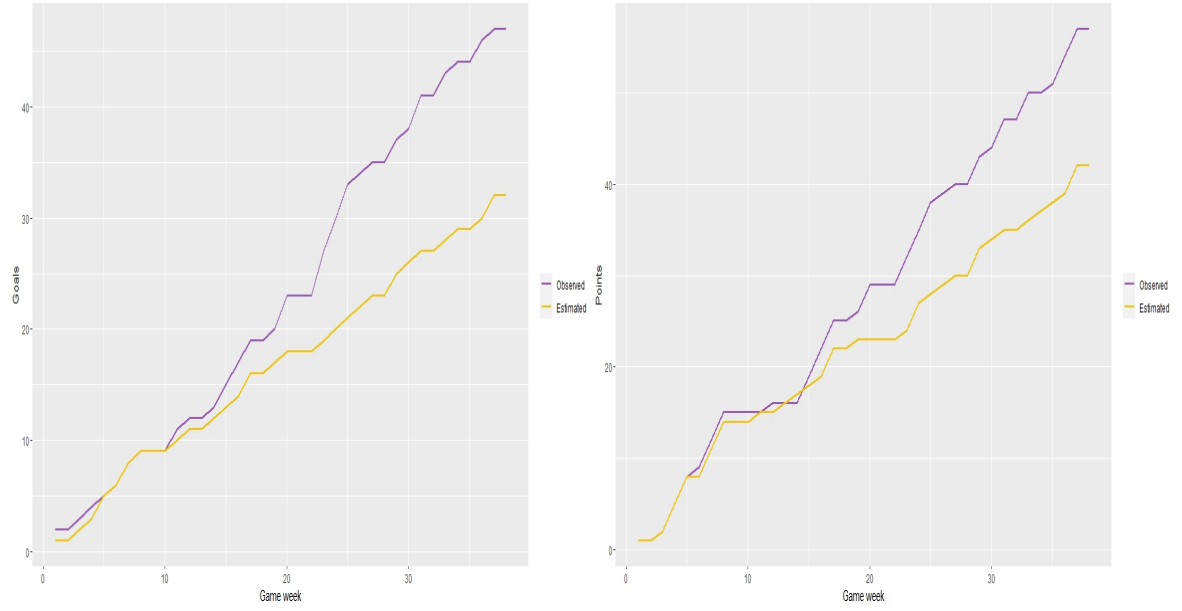


Figure 4.16: Estimated cumulative goals vs observed (left). Estimated cumulative points vs observed (right).

From Table 4.11, you can observe two intriguing matches played by Wolverhampton Wanderers, providing further evidence to support the ideas previously developed.

Home Team	Away Team	GH	GA	W	L	D	$\hat{GH}$	$\hat{GA}$
Wolverhampton Wanderers	Arsenal	3	1	0.42	0.32	0.26	1	1
Newcastle United	Wolverhampton Wanderers	1	2	0.45	0.25	0.30	1	1

Table 4.11: Comparison between observed results and estimated results (median of the samples from the posterior predictive distribution). **W**, **L** and **D** represent respectively the expected probability to win, lose and to draw.

In conclusion, it can be argued that the use of covariates and multiple inflation factors can aid in predicting a team's behaviour throughout a season. However, there are still unresolved issues, particularly with mid-ranked teams. Further improvements can be made in that direction.

Team	Pts	GF	GA	Pos
Manchester City	100	76	17	1
Liverpool	96	73	15	2
Chelsea	66	49	34	5
Tottenham Hotspur	70	55	34	4
Arsenal	74	60	43	3
Manchester United	59	50	44	6
Wolverhampton Wanderers	47	39	36	8
Everton	50	42	37	7
Leicester City	45	42	41	9
West Ham United	41	42	46	11
Watford	41	42	48	12
Crystal Palace	46	42	44	10
Newcastle United	37	34	40	14
Bournemouth	38	44	55	13
Burnley	29	39	54	15
Southampton	27	37	54	17
Brighton & Hove Albion	29	30	47	16
Cardiff City	21	29	55	18
Fulham	16	29	64	19
Huddersfield Town	13	15	61	20

Team	Pts	GF	GA	Pos
Manchester City	91	84	23	2
Liverpool	96	84	24	1
Chelsea	64	53	36	5
Tottenham Hotspur	70	59	35	3
Arsenal	67	61	45	4
Manchester United	64	60	54	6
Wolverhampton Wanderers	50	44	36	7
Everton	49	43	40	10
Leicester City	51	47	43	9
West Ham United	54	47	46	8
Watford	42	42	50	13
Crystal Palace	45	43	43	11
Newcastle United	43	39	45	12
Bournemouth	29	39	62	15
Burnley	29	35	59	16
Southampton	31	38	58	14
Brighton & Hove Albion	27	33	51	18
Cardiff City	29	32	58	17
Fulham	20	27	67	19
Huddersfield Town	15	19	64	20

Team	Pts	GF	GA	Pos
Manchester City	98	76	14	1
Liverpool	98	72	16	2
Chelsea	66	49	32	5
Tottenham Hotspur	68	54	34	4
Arsenal	74	61	42	3
Manchester United	58	50	43	6
Wolverhampton Wanderers	44	38	37	9
Everton	51	43	37	7
Leicester City	51	42	38	8
West Ham United	44	42	45	10
Watford	35	40	49	14
Crystal Palace	41	40	44	11
Newcastle United	37	34	39	13
Bournemouth	40	44	53	12
Burnley	30	36	53	15
Southampton	30	37	53	16
Brighton & Hove Albion	27	28	48	17
Cardiff City	21	28	55	18
Fulham	16	28	66	19
Huddersfield Town	13	15	60	20

Team	Pts	GF	GA	Pos
Manchester City	96	79	16	2
Liverpool	98	71	12	1
Chelsea	68	49	30	4
Tottenham Hotspur	67	53	33	5
Arsenal	75	63	42	3
Manchester United	58	51	44	6
Wolverhampton Wanderers	49	40	35	7
Everton	49	42	37	8
Leicester City	49	41	39	9
West Ham United	45	41	45	10
Watford	36	39	50	14
Crystal Palace	44	40	42	11
Newcastle United	37	35	38	13
Bournemouth	40	46	58	12
Burnley	30	34	53	16
Southampton	32	37	52	15
Brighton & Hove Albion	25	27	50	17
Cardiff City	22	27	55	18
Fulham	18	28	64	19
Huddersfield Town	10	12	60	20

Team	Pts	GF	GA	Pos
Manchester City	97	75	12	2
Liverpool	100	68	10	1
Chelsea	63	43	28	5
Tottenham Hotspur	78	52	26	3
Arsenal	73	62	37	4
Manchester United	58	46	41	6
Wolverhampton Wanderers	42	32	33	11
Everton	51	35	32	7
Leicester City	46	36	35	10
West Ham United	41	33	42	12
Watford	41	33	43	13
Crystal Palace	51	36	35	8
Newcastle United	37	28	34	16
Bournemouth	48	35	49	9
Burnley	39	32	46	15
Southampton	40	33	45	14
Brighton & Hove Albion	32	22	44	17
Cardiff City	32	20	47	18
Fulham	20	20	61	19
Huddersfield Town	16	11	54	20

Table 4.12: Estimated points, goals for and goals against considering the median as a summary statistics for the different parameters involved in the chain. Then, the final rank according to the estimated points have been built.

4.2.2 Are we able to approximate the marginal distributions?

Thanks to the posterior predictive distribution (3), it is possible to estimate the marginal distribution for the number of home/away scored goals. More into detail, given a posterior sample coming from the aforementioned distribution, the median is used as a sufficient statistic and comparisons can be made between the home and away

estimated number of goals for each new game. In particular:

- for the first model, looking at the Table 4.13, a very large discrepancy between estimated and observed values can be noticed. The model cannot reliably predict more than two goals (both for home and away teams) and predicts 1 very frequently. At the end of the day, this is something to put into consideration as they are rare events that happen in less than 1 percent of matches;

home	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**away**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.04	0.60	0.32	0.05	0	0	0	**Estimated**	0.17	0.74	0.09	0	0	0	0

Table 4.13: Observed vs estimated relative frequency of home and away scored goals.

- for the model with covariates, from Table 4.14 it is possible to notice that both for the event {0 scored goal} and {2 or more scored goals}, a strong underestimation effect is present. Reversely, for the event {1 scored goal}, an overestimation effect is acting behind this scene. Comparing these results with Table 4.13, it is feasible to argue that the situation has slightly improved due to the model's ability to predict higher values with greater probability. However, the overestimation of the event {1 scored goal} is still too high;

home	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**away**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.08	0.61	0.24	0.05	0.024	0	0.003	**Estimated**	0.14	0.67	0.15	0.03	0.008	0	0

Table 4.14: Observed vs estimated relative frequency of home and away scored goals.

- for the third model, with regards to the estimation of the number of goals scored by both home and away teams in the league, it is evident from table 4.15 that, once again, the model failed to accurately predict values beyond 2. Moreover, the purpose of implementing the **ZIP** distribution was to counteract the effect of predicting 1 and to increase the frequency of predicted values of 0. However, owing to the low values of the two inflation parameters, this objective was not satisfactorily met;

home	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**away**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.04	0.61	0.31	0.04	0.002	0	0	**Estimated**	0.18	0.73	0.08	0	0	0	0

Table 4.15: Observed vs estimated relative frequency of home and away scored goals.

home	0	1	2	3	4	5	6	away	0	1	2	3	4	5	6
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.06	0.56	0.32	0.06	0.003	0	0	**Estimated**	0.21	0.70	0.09	0	0	0	0

Table 4.16: Observed vs estimated relative frequency of home and away scored goals.

- for the fourth model, the Table 4.16 indicates that the situation has not improved in comparison to the baseline situation 4.13, and the strategy to increase the estimation of zero has failed.
- for the last model, Table 4.17 shows that, compared to Table 4.13, the model is now capable of more precisely predicting the probability of the event {team scores 0 goal} for both home and away matches, marking the first time this has been achieved. It is worth considering the introduction of several inflation factors and some improvements have been achieved.

home	0	1	2	3	4	5	6	away	0	1	2	3	4	5	6
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.21	0.38	0.35	0.06	0.003	0	0	**Estimated**	0.34	0.60	0.06	0	0	0	0

Table 4.17: Observed vs estimated relative frequency of home and away scored goals.

In conclusion, It can be affirm that a model which is able to estimate low number of goals has been achieved, while there is still a problem to estimate "rare" events.

4.2.3 Are we able to study the joint distribution? Are we able to understand if a team is going to win/lose/draw?

As a last point of discussion, [1], [11] offer the recommendation to examine the posterior predictive distribution of specific events, namely to study the joint distribution of the scored goals. Remember that the original aim of this report is to find a way to explain match results that are bivariate. Additionally, the distribution of goal differences is analysed to understand the frequency of draws predicted by the model.

Starting from the baseline model, Figure 4.17 shows that the model significantly overestimates the probability of the '1-1' event but fails to accurately predict '1-0' and '0-1' events. This is linked to the fact that it usually overestimates the likelihood of a team (whether home or away) scoring one goal. Additionally, it appears impartial regarding match results, as it cannot accurately predict whether a team will win or lose.

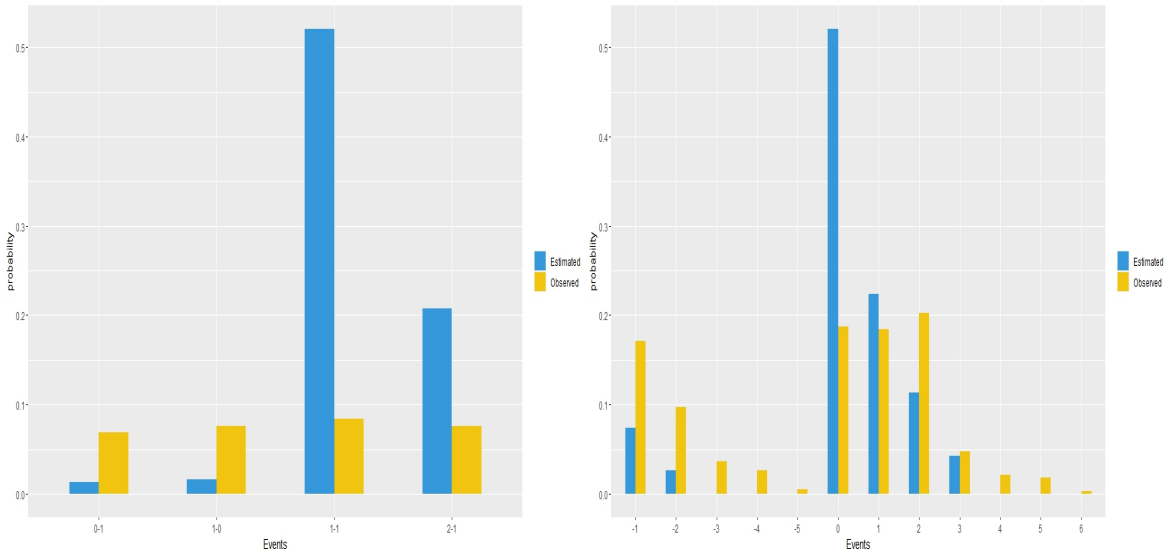


Figure 4.17: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

Once covariates have been added to the model, comparing Figure 4.12 and Figure 4.18, it is clear that the model still overestimates the event "1-1" and underestimates the events "1-0" and "0-1". Additionally, the amount of estimated draws (event "0") remains overestimated, but there is a gradual improvement in the fit for other events. In greater detail, the model can now assess matches with a higher number of scored goals with greater confidence, which it can be noticed that the goal difference takes higher absolute values. In conclusion, it can be affirmed that progress has been made towards developing an improved model. The next step is to define a model capable of predicting values with high confidence, not restricted to one. Additionally, efforts should be made to reduce underestimation of the prediction of zeros.

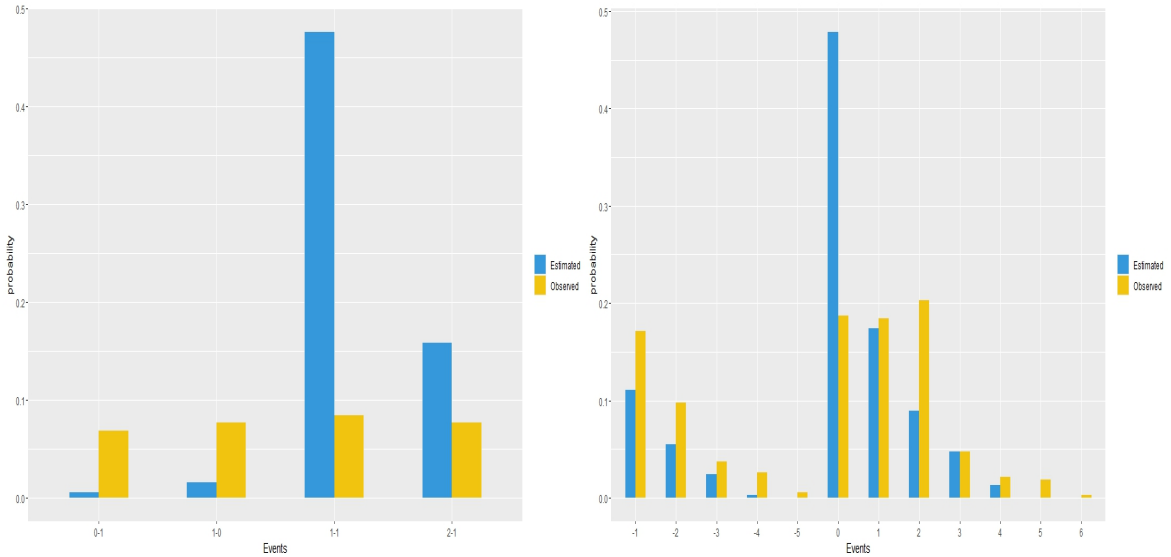


Figure 4.18: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

Taking into account two inflation factors, which has been used to reduce the underestimation of zeros, no significant disparities can be observed when scrutinising Figure 4.17 and Figure 4.19. The inclusion of inflation factors does not significantly affect the model's forecast. As previously mentioned, the model is inadequate at correcting the overestimation of 1 and underestimation of 0. This could be due to the assumption of one inflation parameter for all teams being too restrictive, as expected variations occur throughout the season. Considering that the inflation factor influences the predicted number of zeros, increasing the number of coefficients in the model may have a beneficial impact.

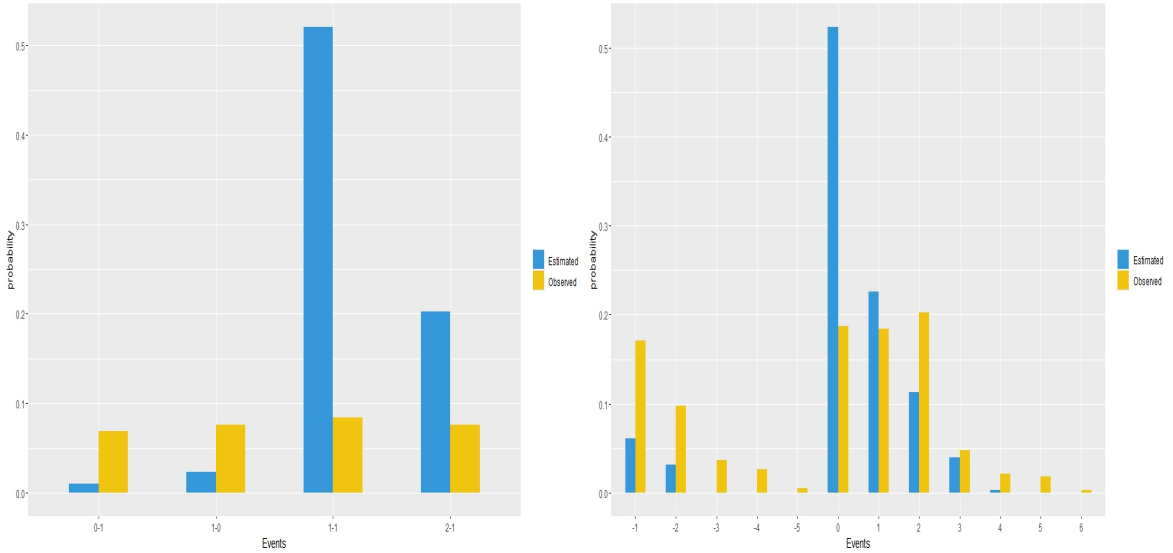


Figure 4.19: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

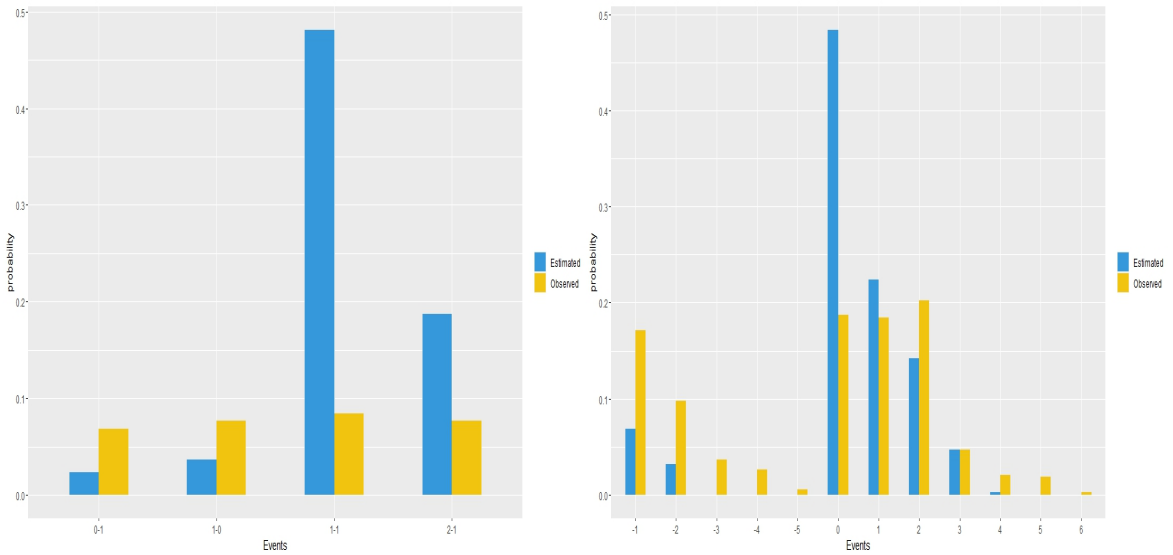


Figure 4.20: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

The addition of several linked inflation factors has not improved the situation.

Furthermore, Figure 4.20 illustrates that the model still overestimates the occurrence of the '1-1' event and the number of predicted draws. In general, assuming parameters from the same population has not improved the model and it is pretty far from the reality.

Comparing Figure 4.21 with the outcomes obtained by all other models, it can be claimed that although the model still tends to overestimate the event "1-1", this has been reduced (note that the estimated probability decreased from around 0.5 to around 0.3). It can be stated without any loss of generality that the model is also more confident about the results, as shown by the lowered estimated probability of draws. In conclusion, implementing one inflation factor for the team has resulted in some improvement.

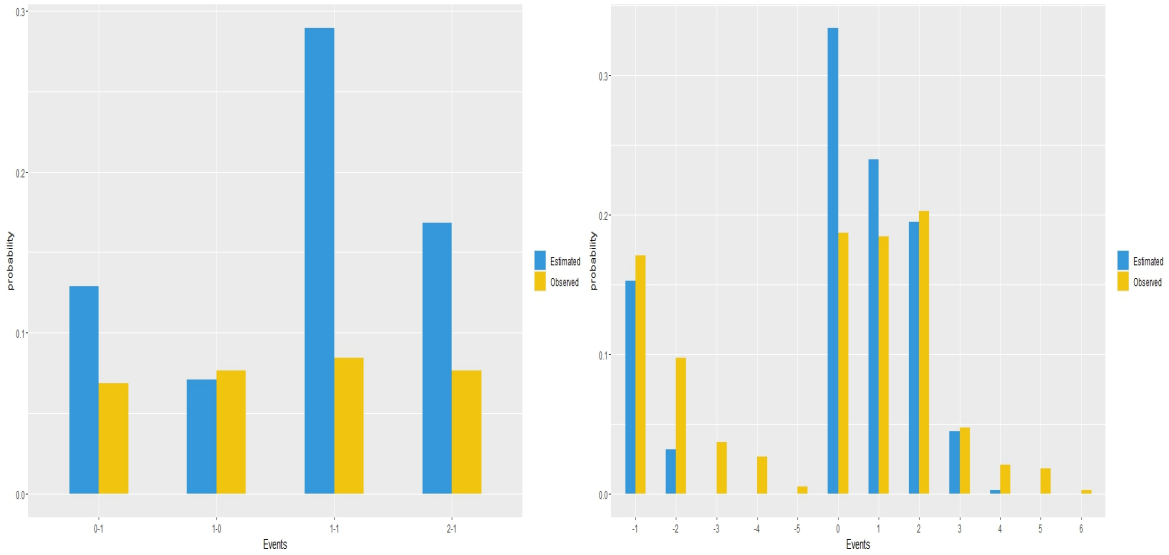


Figure 4.21: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

5 Dynamic Model

Previous models assumed that the season in which a match is played has no direct effect on the outcome of a match, meaning that the effect of time is negligible. The underlying random effects are constant throughout the season, which means that team strength does not change over time. However, using time-related variables, I have shown that time can have a significant impact on the outcome of a match, which is quite understandable. In fact, a team's conditions can vary greatly over the course of a season. In addition, I have decided not to go down the road of using a zero-inflated modelling, as the number of extra zeros was not correctly estimated anyway. The remainder of this section will be a presentation of the new model but, for the time being at least, the covariates are not directly part of the model.
Removing the non-time-dependence assumption requires a generalisation of the model introduced by [1]. In particular, the concept of **Dynamic Generalised Linear models** must be presented in detail, as described by [20], and the previous model adapted accordingly. In order to accomplish the previous task, a slight modification of the notation is required. Denoting $t = 1, \dots, 38$ as the game week and $j = 1, \dots, 10$ as one of the matches played during that week, it is possible to uniquely identify a game using those two indexes. Again $w = 1, \dots, 20$ [12] represents a generic team.
The model can be formalized in the following way. First of all, $y_{t,j,1}$ and $y_{t,j,2}$ represent respectively the number of goals scored by the home and away team in the j-th game of the t-th week. Then, following the usual Bayesian setup, $y_{t,j,1}, y_{t,j,2}|\theta_{t,j,1}, \theta_{t,g,2} \overset{C.I}{=} y_{t,j,1}|\theta_{t,j,1} \cdot y_{t,h,2}|\theta_{t,j,2}$, where $\theta_{t,j,1}$ and $\theta_{t,j,2}$ represent the scoring intensities of the home and away teams in the j-th match of the t-th week. Define:

- **Observation Model**: the structure of the likelihood is similar to the previous case, but with an increased number of parameters. The identifiable constraints on the parameters remain the same and are valid for each time. It is important to note that the parameterization with two intercepts is considered as in the Explanatory Model of Section 3.

  ⋆ $y_{t,j,1}|\theta_{t,j,1} \sim Poisson(\theta_{t,j,1})$ and $y_{t,j2}|\theta_{t,j,2} \sim Poisson(\theta_{t,j,2})$

  ⋆ $\begin{cases} log(\theta_{j,t,1}) = u_h + att_{h[g],t} + def_{a[g],t} \\ log(\theta_{j,t,2}) = u_a + att_{a[g],t} + def_{h[g],t} \end{cases}$

  ⋆ the likelihood can be written in the following way:

$$\begin{aligned} \mathbb{L}(\underline{\theta}) &= f\left(y_{1,1,1}, y_{1,1,2}, \dots, y_{38,10,1}, y_{38,10,2}|\theta_{1,1,1}, \theta_{1,1,2}, \dots, \theta_{38,10,1}, \theta_{38,10,1}\right) \\ &\overset{c.i}{=} f\left(y_{1,1,1}, \dots, y_{38,10,1}|\theta_{1,1,1}, \dots, \theta_{38,10,1}\right) \cdot f\left(y_{1,1,2}, \dots, y_{38,10,2}|\theta_{1,1,2}, \dots, \theta_{38,10,2}\right) \\ &\overset{i.d}{=} \prod_{t=1}^{38} \prod_{j=1}^{10} f\left(y_{t,j,1}|\theta_{t,j,1}\right) \cdot f\left(y_{t,j,2}|\theta_{t,j,2}\right) \end{aligned}$$

[12] from now on, the total number of teams will be represented by n.

$$\star \sum_{w=1}^{20} att_{w,t} = 0 \text{ and } \sum_{w=1}^{20} def_{w,t} = 0 \quad \forall t = 1, \ldots, 38$$

- **Evolution Equation**: a random walk stochastic process is employed to update the parameters associated with the attacking and defensive abilities. The underlying assumption is that a team's performance is not significantly different from its previous values, without taking into account possible 'shocks'. First of all, define $W = \sigma^2 \mathrm{I}_{20}$, where σ^2 is known as the **evolution variance**. The variance of matrix is assumed to be constant over time and equal for all the teams. Then, define $G \in 20 \times 20$ as the evolution matrix. For the purpose of simplicity, assume that only the offensive/defensive ability of the same team affects its future, which implies that the evolution matrix is equal to the identity matrix. The prior distributions can be specified as follows:

$$\star \begin{cases} \underline{att}_{\cdot,t} = G\underline{att}_{\cdot,t-1} + \underline{w}_t = \underline{att}_{\cdot,t-1} + \underline{w}_t & \underline{w}_t \sim \mathbb{MVN}_n(\underline{0}, W) \\ \underline{def}_{\cdot,t} = G\underline{def}_{\cdot,t-1} + \underline{\tilde{w}}_t = \underline{def}_{\cdot,t-1} + \underline{\tilde{w}}_t & \underline{\tilde{w}}_t \sim \mathbb{MVN}_n(\underline{0}, W) \\ \underline{att}_{\cdot,0} \sim \mathbb{MVN}_n(\underline{m}_{att}, W_0) & \\ \underline{def}_{\cdot,0} \sim \mathbb{MVN}_n\left(\underline{m}_{def}, W_0\right) & \\ u_h \sim N(0, 0.001) \quad u_a \sim N(0, 0.001) & \end{cases}$$

According to [15], the constraints are naturally satisfied if $1_n^T \underline{m}_{att} = 0$ and $1_n^T \underline{m}_{def} = 0$ [13] and W and W_0 are transformed into appropriate variance and covariance matrices. Specifically, the diagonal variances reflect the actual previously specified variance and a small negative covariance is used. [14].

$$R = \sigma^2 \frac{n}{n-1}\left(\mathrm{I}_n - \frac{1}{n} 1_n 1_n^T\right) = \sigma^2 \frac{n}{n-1}\left(\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} & \ldots & \ldots & \frac{1}{n} \\ \vdots & \ddots & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ \frac{1}{n} & \ldots & \frac{1}{n} & \frac{1}{n} \end{bmatrix}\right) \tag{6}$$

$$= \sigma^2 \frac{n}{n-1} \begin{bmatrix} \frac{n-1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \ldots & -\frac{1}{n} \\ \vdots & \ldots & \ddots & \vdots \\ -\frac{1}{n} & \ldots & -\frac{1}{n} & \frac{n-1}{n} \end{bmatrix} = \begin{bmatrix} \sigma^2 & -\frac{\sigma^2}{n-1} & \ldots & -\frac{\sigma^2}{n-1} \\ -\frac{\sigma^2}{n-1} & \sigma^2 & \ldots & -\frac{\sigma^2}{n-1} \\ \vdots & \ldots & \ddots & \vdots \\ -\frac{\sigma^2}{n-1} & \ldots & -\frac{\sigma^2}{n-1} & \sigma^2 \end{bmatrix} \tag{7}$$

Given all of these specifications, it is possible to derive a closed form for the prior distributions associated with the model:

[13] this means that the hyperprior values for the first random walk iteration must sum to zero.

[14] the same reasoning can be applied to the transformation of W_0 into R_0, but for the sake of brevity it will be omitted.

$$\pi\left(\sigma^2, u_h, u_a att, def\right) = \pi\left(home\right)\pi(\sigma^2)\prod_{t=1}^{38}\pi\left(\underline{att}_{,t}|\underline{att}_{,t-1}\right)\pi\left(\underline{def}_{,t}|\underline{def}_{,t-1}\right)\pi(\underline{att}_{,0})\pi(\underline{def}_{,0}) =$$

$$= logGamma(0.01, 0.01)\pi(\sigma^2)\prod_{t=2}^{38}\mathbb{MVN}_n\left(\underline{att}_{,t-1}, R\right)\mathbb{MVN}_n\left(\underline{def}_{,t-1}, R\right)\mathbb{MVN}_n$$

$$(\underline{m}_{att}, R_0)\,\mathbb{MVN}_n\left(\underline{m}_{def}, R_0\right)$$

Sampling from the distribution is required due to the complexity of the closed form of the posterior distribution. A Gibbs Sampler is adopted to approximate the desired distribution, as with previous settings. Considering the usual Bayesian updating rule:

$$\pi(\underline{\theta}|\underline{y}) \propto \mathbb{L}\left(\underline{\theta}\right)\pi(\underline{\theta}) = \prod_{t=1}^{38}\prod_{j=1}^{10} f\left(y_{t,j,1}|\theta_{t,j,1}\right)\cdot f\left(y_{t,j,2}|\theta_{t,j,2}\right)\pi\left(home\right)\pi(\sigma^2)$$
$$\prod_{t=1}^{38}\pi\left(\underline{att}_{,t}|\underline{att}_{,t-1}\right)\pi\left(\underline{def}_{,t}|\underline{def}_{,t-1}\right)\pi(\underline{att}_{,0})\pi(\underline{def}_{,0})$$

Before commenting on the obtained results, the following two subsections will present two remarks regarding the previous model specification.

5.1 Rjags setup

Although the model specification has been completed, a problem arises in the simulation process due to the lack of full rank in both R and R_0, making it impossible to sample from a multivariate normal distribution (the computation of the inverse of the variance-covariance matrix cannot be done). [15] proposed to make slight modifications to the simulation task to address the issue. A solution will be shown for the attacking parameters, but it can be applied in the same way for the defensive parameters.

The new approach consists in the sampling of $n-1$ unconstrained parameters $\underline{c}_t = [c_{t,1}, \ldots, c_{t,n-1}]^T$ from a Multivariate Normal distribution centered in the vector of zeros and with variance-covariance matrix equal to S_t, where:

$$S_t = \sigma^2\frac{n}{n-1}\left(\mathrm{I}_{n-1} + 1_{n-1}1_{n-1}^T\right) = \sigma^2\frac{n}{n-1}\left(\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & \ddots & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 1 & \ldots & \ldots & 1 \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{2\sigma^2 n}{n-1} & \frac{\sigma^2 n}{n-1} & \cdots & \frac{\sigma^2 n}{n-1} \\ \frac{\sigma^2 n}{n-1} & \frac{2\sigma^2 n}{n-1} & \cdots & \frac{\sigma^2 n}{n-1} \\ \vdots & \ldots & \ddots & \vdots \\ \frac{\sigma^2 n}{n-1} & \cdots & \frac{\sigma^2 n}{n-1} & \frac{2\sigma^2 n}{n-1} \end{bmatrix}$$

Define then a matrix J as follows:

$$J = \begin{bmatrix} \mathrm{I}_{n-1} - \frac{1}{n}1_{n-1}1_{n-1}^T & \frac{1}{n}1_{n-1} \\ -\frac{1}{n}1_{n-1}^T & \frac{1}{n} \end{bmatrix} = \begin{bmatrix} \frac{n-1}{n} & -\frac{1}{n} & \dots & \frac{1}{n} \\ -\frac{1}{n} & \ddots & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ -\frac{1}{n} & \dots & \dots & \frac{1}{n} \end{bmatrix}$$

After performing some calculations, it is possible to demonstrate the resulting values of $\underline{att}_{,t} = \underline{att}_{,t-1} + \underline{u}_t$ represent the required sampled values of the attack parameters with the required evolution structure and variance-covariance structure given by 6, and with the identifiability constraint holding for all t, where $\underline{u}_t = J\underline{\tilde{c}}_t$ [15]. Indeed, upon calculating the expected value and the variance-covariance matrix associated with the random $\underline{att}_{,t}$ utilizing the aforementioned methodology, it can be demonstrated that these computed statistical properties align precisely with those previously specified. The same reasoning can be applied to sample from the distribution of $\underline{att}_{,0}$ and for the defensive parameters.

The code reproducing the method described above can be found in the Appendix B.

5.2 Evolution Variance

As you may have noticed, no assumptions have been made about the distribution of the evolution variance so far. This is a crucial point as its value has a critical impact on the stochastic changes in the offensive and defensive parameters, making it a very sensitive choice. In order to decide which value is more accurate to use, an empirical study has to be done [15]. Specifically, a grid search is performed over different fixed values to find the one that maximises $\mathcal{P}_1$ and it is used as the mean of a low variance Gamma distribution to determine the values of the hyperparameters [16]. This measure reflects the short-term predictive ability of the model and it is defined as follows:

$$\mathcal{P}_1 = \left(\prod_{k=1}^{380} \mathbb{P}(O_k)\right)^{\frac{1}{380}} = \exp\left\{\frac{1}{380}\log\left(\prod_{k=1}^{380}\mathbb{P}(O_k)\right)\right\} = \exp\left\{\frac{1}{380}\sum_{k=1}^{380}\log\left(\mathbb{P}(O_k)\right)\right\}$$

where $\mathbb{P}(O_k)$ is the one match ahead predicted probability that match k will result in the final observed outcome, O_k, of either "home", "draw" or "away". The index thus represents the geometric expectation associated with the probability of the previously described event. More specifically, given the realisations of the posterior predictive distribution for the observed match, it is possible to easily estimate the probability by simply calculating the fraction of time that the corresponding event occurred.

[15] $\underline{\tilde{c}}_t = [c_{t,1}, \dots, c_{t,n-1}, 0]$

[16] an informative type of distribution is used once the optimal values have been found.

Subsequently, to account for the overall adequacy of the different models, the DIC is calculated, in conjunction with the metric $\mathcal{P}_1$.

To evaluate the parameter order, I gradually increased the value of the evolution variance from an initial value of 0.001 to an upper bound of 1000. Notably, I found that a value greater than 0.01 was unlikely due to the fact that despite of the apparent growth of $\mathcal{P}_1$, the DIC exhibited a disproportionately large increase in comparison to outcomes achieved with variances below 0.01. Subsequently, a more focused exploration was undertaken within the narrower interval spanning from 0.001 to 0.1 to discern the optimal value of the evolution variance.
Upon scrutinizing the data presented in Figure 5.1, it becomes apparent that a universally optimal value does not emerge. However, discernible trends suggest that a sub-optimal value for the evolution variance may be identified at 0.006. This assertion is substantiated by the minimization of the DIC at this specific value. Moreover, beyond this point, any subsequent increase in the $\mathcal{P}_1$ metric appears to be inconsequential. Therefore, it can be reasonably inferred that the evolution variance prior distribution is a Gamma with mean equal to 0.006, i.e. a $\mathbf{Gamma(3.6, 600)}$. The graphical representation in Figure 5.2 illustrates the prior-to-posterior update of the evolution variance, revealing a notable concentration around the value of 0.003 with a diminished level of uncertainty.

Figure 5.1: $\mathcal{P}_1$ measure (left) and DIC (right) for different values of the evolution variance.

Upon identifying the pertinent distribution associated with the parameter σ^2, it became evident that the selection of σ_0^2 exhibited a nominal impact on both DIC and $\mathcal{P}1$ when considering a judiciously chosen range of values. Consequently, a pragmatic decision was made to establish a fixed value for σ_0^2, with particular attention to

maintaining its influence within an acceptable range. The chosen value for σ_0^2 was set at 0.003 based on the previous reasoning.

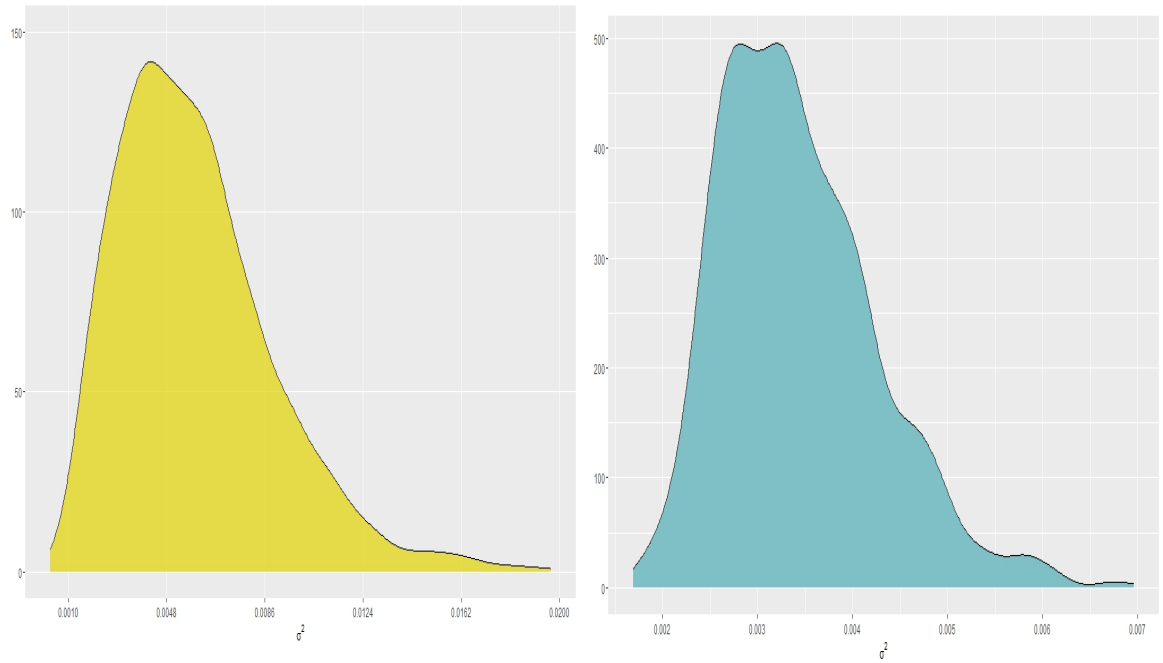


Figure 5.2: Prior (left) to posterior (right) update associated to the evolution variance σ^2.

5.3 Results

First of all, it is essential to provide a succinct overview of the experimental framework. Due to the considerable number of parameters involved and the substantial computational resources required, I opted to conduct a reduced number of simulations drawn from the posterior distribution. Specifically, after several preliminary trials, a decision was made to perform 5000 simulations utilizing the MCMC sampler. These simulations included a burn-in period of 1000 iterations and a thinning factor of 2. In this subsection, I will address the same research questions delineated in Section 4, even though more emphasise is put on the estimation side.

5.3.1 Which is the impact of the offensive and defensive parameter?

Prior to delving into the analysis, it is pertinent to examine the influence of $\underline{m}_{att}$ and $\underline{m}_{def}$ on the sequences of attacking and defensive capabilities. In principle, adhering to the conventional Bayesian framework, the choice of initial values for these parameters is expected to exert negligible impact. Nevertheless, considering the utilization of a random walk update, it is prudent to empirically investigate whether this choice indeed affects the underlying random effects. To explore this impact, various initializations were tested. One method involved obtaining initial values by subtracting the league average from the offensive and defensive propensities (see Figure 3.1). Additionally,

the others consider Gaussian distributions centered at zero with low variance. It is important to note that, to satisfy constraints, the sum of attacking and defensive parameters must equate to zero.
Surprisingly, the choice of the starting values had a substantial impact on both the values attained by the posterior distribution of the random effect and the goodness of fit of the entire model. Notably, randomly initialized models exhibited a greater DIC and a lower $\mathcal{P}_1$ compared to the model utilizing actual scoring propensities of the teams. Furthermore, the posterior distribution associated with the underlying random effects varied significantly. In summary, the choice of the starting values for the offensive and defensive abilities is fundamental and should not be underestimated [17].

This part delves into the analysis of random effects trends following the selection of initial values. Due to the high number of parameters (38 Game weeks, 20 teams, 2 teams for each game), it is not possible to report a deep study about all of them. I observe two different trends analysing the median attained at each time for the random effects. The first trend is characterized by a continuous increase or decrease, signaling potential issues with the initial value selection. Conversely, the second trend, marked by general stationarity, suggests stability in portraying team conditions (Figure 5.3). Additionally, it is noteworthy that the time series pertaining to both offensive and defensive capabilities exhibit a concentration around values approximated through non-dynamic models (Figure 4.2).

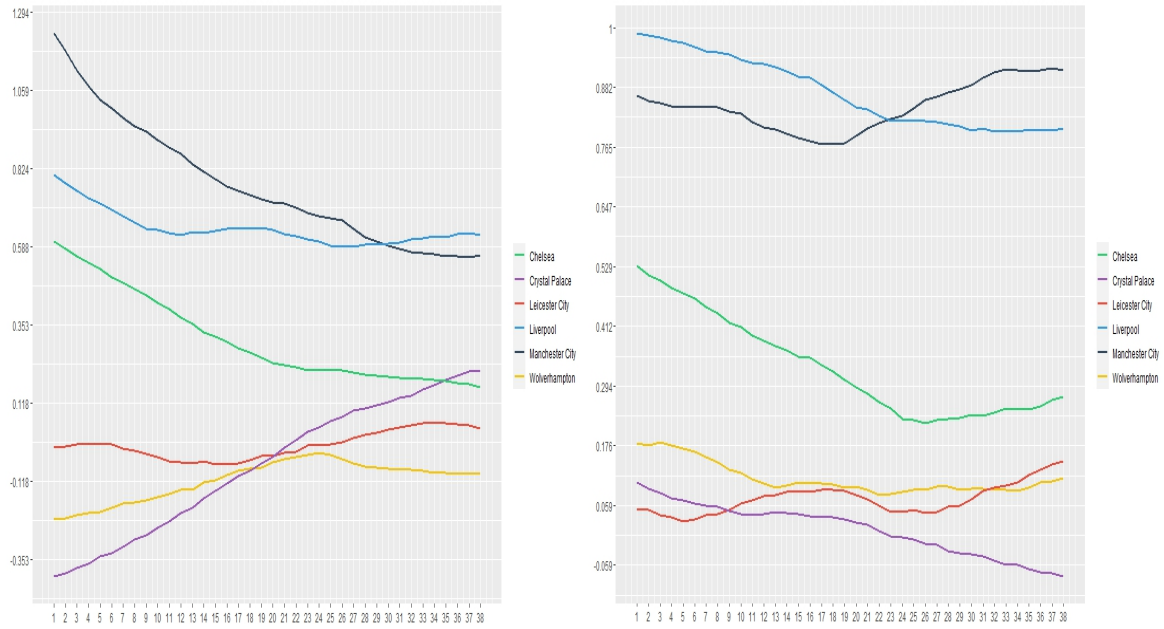


Figure 5.3: Time series of the offensive (left) and minus defensive (right) random effects for some team considering the median of the posterior distribution as a summary statistic.

[17]The author of the paper [15] suggests using the values estimated via the non-dynamic model in the previous season, but this is an arbitrary choice given that three teams were actually different.

To sum up, emphasizing the critical role of initial parameterization in shaping random effects trajectories, random effects remain effective in reflecting the intricate dynamics of offensive and defensive team conditions throughout a season. Nevertheless, it is essential to acknowledge that additional advancements are needed in mitigating the influence exerted by the starting values. Further research and development in this direction are imperative to enhance the robustness and reliability of random effects in modeling the nuanced fluctuations observed in team performance throughout a sporting season.

5.3.2 Does a "home" effect really exist?

Considering the adopted parametrization, remember that the home effect can be interpreted as the difference between the two intercepts involved in the Poisson regressions. Furthermore, it is assumed to be constant through the season, which is a reasonable hypothesis. In agreement with [1], [11] and my previous results, a positive home effect still exists. In particular, analysing the posterior distribution, it has a positive average value of 0.209 (95% HPD interval $[0.09, 0.33]$).

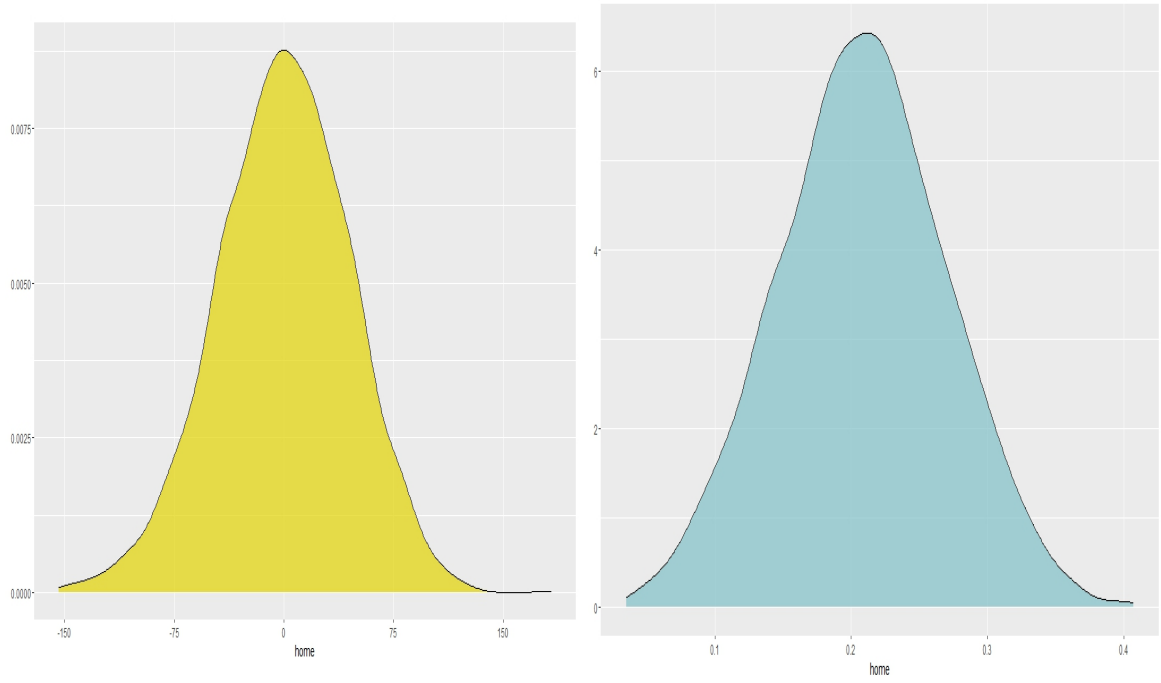


Figure 5.4: Prior (left) to posterior (right) update associated to the difference between the two intercepts.

In summary, the analysis in Section 4 and Figure 5.4 strongly supports the existence of a home advantage, consistently leading to a positive impact on the number of goals scored by the home team. This empirical evidence underscores the significant role of the home environment in shaping match outcomes, contributing valuable insights to sports dynamics.

5.3.3 Are we able to approximate the marginal/joint distributions?

The mandatory execution of predictive checks serves as a crucial step in empirically establishing the newfound methodology's reliability. A thorough comparison between the densities presented in Table 5.1 and those expounded upon in Section 4 elucidates a discernible refinement in the estimation of marginal distributions. The model exhibits a notable inclination toward predicting a broader range of values with heightened probabilities, indicative of an overall improvement in predictive accuracy. Notwithstanding these advancements, certain residual challenges persist in the predictive framework.

Upon close examination of Figure 5.5, it becomes apparent that a prevailing tendency towards overestimation in the generated draws is still observable, albeit with a degree of attenuation. It is pertinent to note that, despite these persisting challenges, the model evinces a commendable ability to effectively forecast match outcomes. The incorporation of a temporal dimension has been fundamental in providing a nuanced understanding of match results. While acknowledging the potential for further refinements, it can be asserted that the model, on the whole, demonstrates a proficient capacity to predict match outcomes. This substantiates the significance of temporal considerations in refining predictive models for enhanced accuracy and reliability.

home	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**away**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Observed	0.23	0.31	0.25	0.13	0.058	0.021	0.008	**Observed**	0.31	0.32	0.23	0.09	0.023	0.015	0.002
Estimated	0.14	0.51	0.25	0.06	0.002	0.005	0.005	**Estimated**	0.21	0.56	0.17	0.03	0.007	0	0

Table 5.1: Observed vs estimated relative frequency of home and away scored goals.

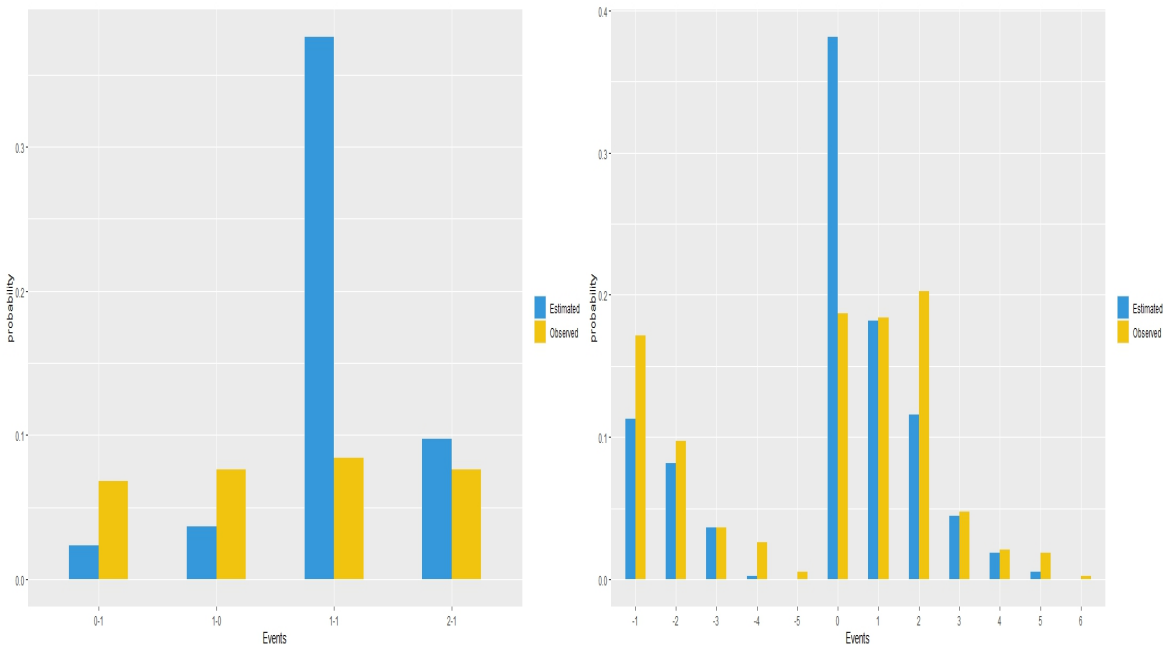


Figure 5.5: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

5.4 Is really useful to consider separately attacking and defensive abilities?

In the analysis of Figure 5.3, a noteworthy observation emerges, suggesting that certain teams exhibit a consistent trend in the relationship between offensive and defensive random effects throughout the entire season. This observation prompts consideration for a novel parameterization approach, wherein each team possesses a single random effect for each game week, encapsulating the overall team dynamics. To elaborate further, let $ov_{w,t}$ denote the vector of random effects associated with the overall team shape for a given week. In order to make this conceptual shift, a slight modification to the existing model is required:

- $\begin{cases} log(\theta_{j,t,1}) = u_h + (ov_{h[g],t} - ov_{a[g],t}) \\ log(\theta_{j,t,2}) = u_a + (ov_{a[g],t} - ov_{h[g],t}) \end{cases}$
- $\sum_{w=1}^{20} ov_{w,t} = 0 \quad \forall t = 1, \ldots, 38$
- $\begin{cases} \underline{ov}_{\cdot,t} = \underline{ov}_{\cdot,t-1} + \underline{w}_t \qquad \underline{w}_t \sim \mathbb{MVN}_n(\underline{0}, W) \\ \underline{ov}_{\cdot,0} \sim \mathbb{MVN}_n(\underline{m}_{ov}, W_0) \\ u_h \sim N(0, 0.001) \quad u_a \sim N(0, 0.001) \\ \sigma^2 \sim Gamma(3.6, 600) \end{cases}$

The likelihood and posterior distribution of the model exhibit a similar structure rather than before, whereas the prior distribution can be articulated as follows [18]:

$$\pi\left(\sigma^2, u_h, u_a ov, \right) = \pi\left(home\right)\pi(\sigma^2)\prod_{t=1}^{38}\pi\left(\underline{ov}_{,t}|\underline{ov}_{,t-1}\right)\pi(\underline{ov}_{,0})\pi(\underline{def}_{,0}) =$$

$$= N(0.01, 0.01)Gamma(3.6, 600)\prod_{t=2}^{38}\mathbb{MVN}_n\left(\underline{ov}_{,t-1}, R\right)\mathbb{MVN}_n\left(\underline{m}_{ov}, R_0\right)$$

In the initial analysis, it is evident that a home effect persists, as expected, with an average positive value of 0.21 and a 95% HPD interval of $[0.09, 0.34]$. The focus now shifts towards the posterior values obtained for the random effects. Without loss of generality, an examination of these parameters suggests their ability to encapsulate the overall condition of a team, where positive values denote an enhanced state and negative values indicate a less favorable condition.

Upon scrutinizing Figure 5.6, two notable observations emerge. Firstly, a discernible resemblance exists between the trends depicted in this model and those observed

[18] Note that also in this case, in order to fulfill the constraint, R and R_0 have to be modified into W and W_0 and the components of the vector $\underline{m}_{ov}$ sum up to 0.

in Figure 5.5. Secondly, it appears plausible that these values effectively capture the genuine state of a team during the tournament. This is exemplified in cases such as Liverpool, Manchester City, and Leicester City. The former two teams engaged in a tremendous battle throughout the tournament, as evidenced by the estimated mean coefficients crossing multiple times. Conversely, the latter team experienced a remarkable resurgence during the concluding games of the season.

Regrettably, despite the model's ability to provide a sort of aggregate measure, the DIC for the new model surpasses that of the previous one, indicative of a decreased goodness of fit. In conclusion, while the prospect of employing an aggregate measure remains interesting, a preference is retained for the model incorporating distinct random effects linked to attacking and defensive abilities. Furthermore, a re-examination of the estimated marginal and joint distributions for the scored goals has been undertaken, revealing outcomes of lesser quality compared to the previous case.

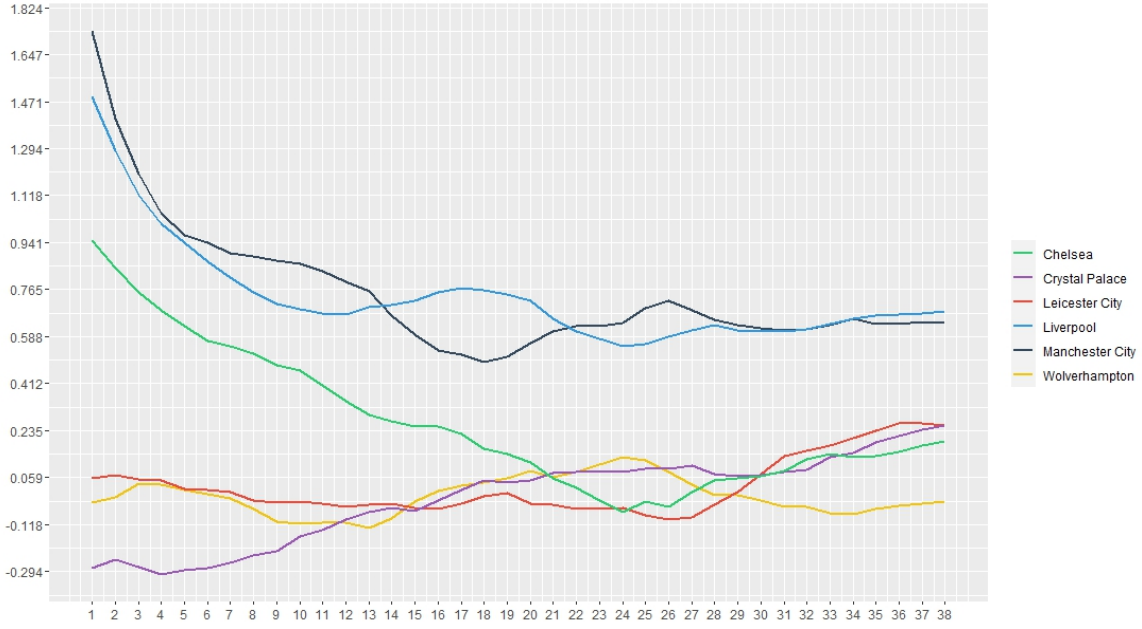


Figure 5.6: Observed vs Estimated probability of some interesting match outcomes (left). Observed vs Estimated probability for goals difference (right).

6 Bonus Section: Goals Difference

In the previous sections, my methodology primarily focused on directly modeling the number of goals scored by both the home and away teams. This approach aimed to discern the major influencing factors contributing to the goal-scoring dynamics. However, an alternative perspective, as proposed by [11], involves considering the variable of interest as the goal difference between the home and away teams. The authors advocate for the use of the **Skellam** distribution, conceptualized as arising from the difference between two random variables, each following a Poisson distribution. Initially, I hesitated to adopt this approach due to the authors' assertion that attempting to model the goal difference precludes the simultaneous modeling of the two marginal distributions, which constitutes a fundamental objective within the aim of this project. Nevertheless, I have resolved to continue the analysis to conclude a comprehensive exploration of various Bayesian techniques employed in modeling football outcomes [19].

Upon reflection, I discerned that an effective strategy involves the independent modeling of both home team scored goals and the goal difference. This helps for the derivation of the away team scored goals as the difference between the two aforementioned distributions. This refined methodological approach holds the potential to enhance the comprehensive understanding of the underlying dynamics influencing goal scoring patterns in the context of the studied football matches.

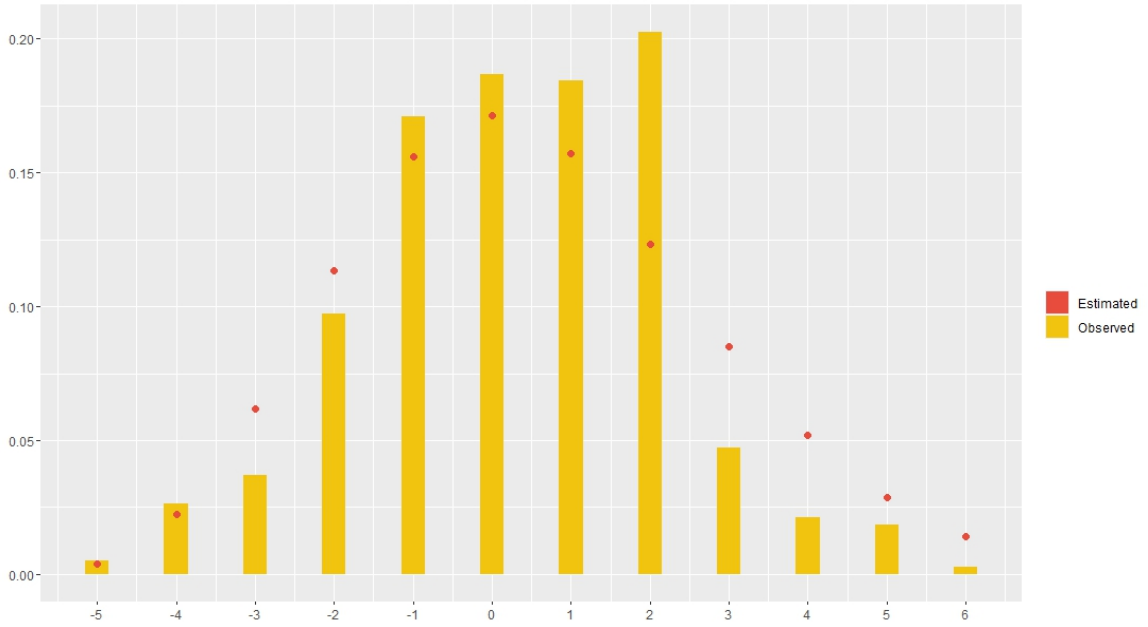


Figure 6.1: Observed vs Estimated probability of the goals difference events using a Shifted Poisson with $\lambda = 5.5$ and $u = 5$.

[19]It is worth noting that, in this report, I have not extensively discussed the potential application of a Bivariate Poisson model. However, it is essential to acknowledge that such a model represents a modest generalization of the frameworks elucidated in Section 3.

In the attempt to model the goal difference (D_g) between home and away scored goals, a departure from the approach proposed by [11] is contemplated. A fundamental observation lies in the distinctive nature of the distribution governing goal difference; while it bears resemblance to a Poisson distribution, it extends its support to negative values (which it is not exhibited by the conventional Poisson distribution). To address this peculiarity, a conceptualization is proposed: consider a transformed random variable, $Z_g = D_g + \mu$, wherein μ serves as a location-shifting parameter. This transformation is used to adhere to a Poisson distribution, denoted as $\mathcal{P}(\lambda_g)$. The introduction of μ is needed in order to shift the support of D_g on a positive scale (included zero). Furthermore, μ is not a proper parameter, but it can be fixed to a reasonable value such that the support of the random variable Z_g is proper defined. In particular, I have fixed $\mu = 5$, which is the minimum value attained by D_g through the entire season.

$$\mathcal{P}(Z_g = z_g) = \frac{e^{-\lambda_g}\lambda_g^{D_g+u}}{(D_g+u)!} \quad z_g = 0, 1, \dots$$

Starting from the previous considerations, it is straightforward to define a proper model to represent the situation, in fact a generalization of the model introduced by [1] is needed. Denoting $g = 1, \dots, 380$ as a generic game and $w = 1, \dots, 20$ as a generic team, the model can be formalized as follows:

- y_g represents the number of goals scored by the home team in the g-th game;
- D_g represents the difference of goals scored by the home and away team in the g-th game;
- $y_g, (D_g + \mu)|\theta_{g,1}, \theta_{g,2} \overset{C.I}{=} y_g|\theta_{g,1} \cdot Dg + \mu|\theta_{g,2}$;
- $y_{g,1}|\theta_{g,1} \sim Poisson(\theta_{g,1})$ and $D_g + \mu|\theta_{g,2} \sim Poisson(\theta_{g,2})$;
- $\theta_{g,1}$ represents the scoring intensities of the home team while $\theta_{g,2}$ can be interpreted as the difference between the scoring intensities of the two involved teams;
- $$\begin{cases} log(\theta_{g,1}) = \mu_h + (att_{h[g]} - def_{a[g]}) \\ log(\theta_{g,2}) = \mu_d + (att_{h[g]} - att_{a[g]}) + (def_{h[g]} - def_{a[g]}) \end{cases}$$

The likelihood of the model can be written:

$$\begin{aligned} \mathbb{L}(\underline{\theta}) &= f\left(y_1, \dots, y_{380}|\theta_{1,1}, \dots, \theta_{380,1}\right) \cdot f\left(D_1 + \mu, \dots, D_{380} + \mu|\theta_{1,2}, \dots, \theta_{380,2}\right) \\ &\overset{i.d}{=} \prod_{g=1}^{380} f\left(y_g|\theta_{g,1}\right) \cdot f\left(D_g + \mu|\theta_{g,2}\right) \end{aligned}$$

For the sake of brevity, priors distributions, as well as the posterior form, are not specified again due to the fact they are the same as the model in Section 3. Again, a

Gibbs Sampler is adopted due to the complex form of the joint posterior distribution of the parameters.

It is evident that the random effects persist in their ability to effectively encapsulate the effective ability of a team throughout the season, as delineated in Figure 6.2. Notably, however, their influence has marginally decrease in comparison to antecedent models. Furthermore, the discernible correlation between the aforementioned values of the random effects and the final ranking of the team has attenuated, albeit remaining subject to interpretation.
The evolution in parametrization precludes any discourse on a plausible home effect. Instead, a deliberation on a potential "away" effect can be done, encapsulating the discrepancy between the two intercepts within the model. Precisely, the mean value for this effect is computed as -1.28, accompanied by a 95% HPD interval of $[-1.29, -1.12]$. In particular, this is also in agreement with the ideas developed during the entire project.

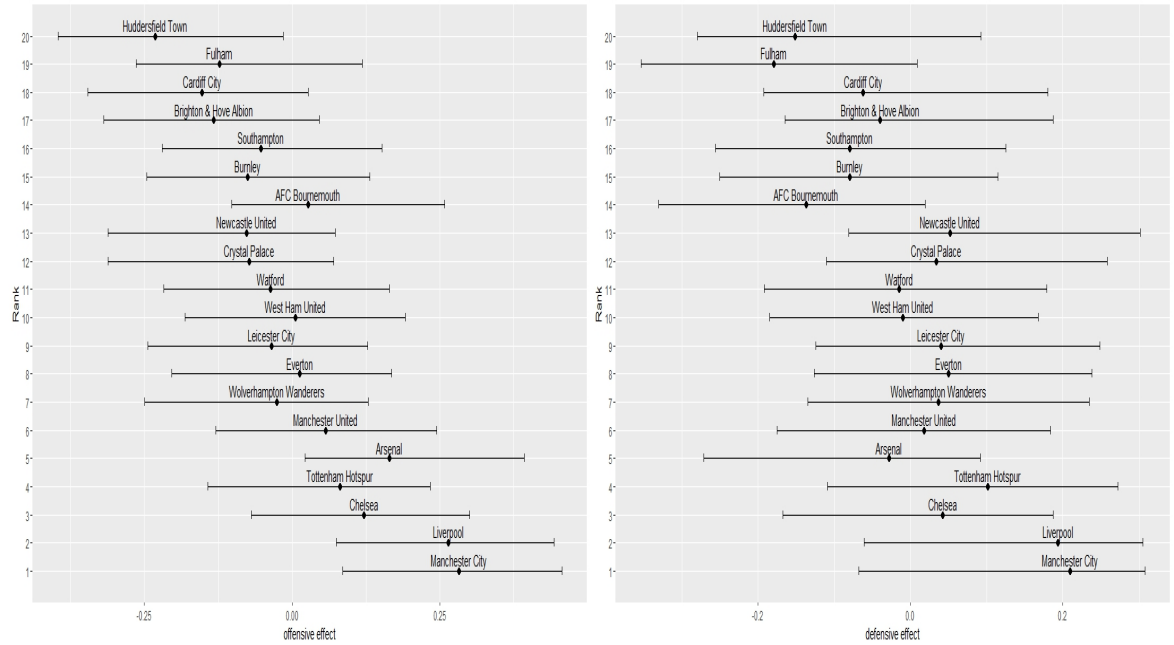


Figure 6.2: 95% HPD interval for offensive (left) and defensive random effects. Dots represent the average effect.

In addition to scrutinizing the estimated marginal distributions of home scored goals, which shares analogous challenges with its predecessor, an insightful examination pertains to elucidating the posterior predictive fit of goal differences. Notably, a recurring pattern emerges wherein there is a propensity for overestimation in the occurrence of draws. Despite this, the overall estimation demonstrates a decent precision, although temporal considerations are not explicitly incorporated. In fact, a comparative analysis between Figure 5.5 and 6.2 underscores the striking similarity in the estimated proportions. However, a salient limitation of this model lies in its assumption that the number of goals scored by the home team and the goal

difference are uncorrelated. This oversimplification becomes evident in the context of this modeling paradigm, in fact it is possible that the posterior predictive distribution for the away scored goals attains negative values.

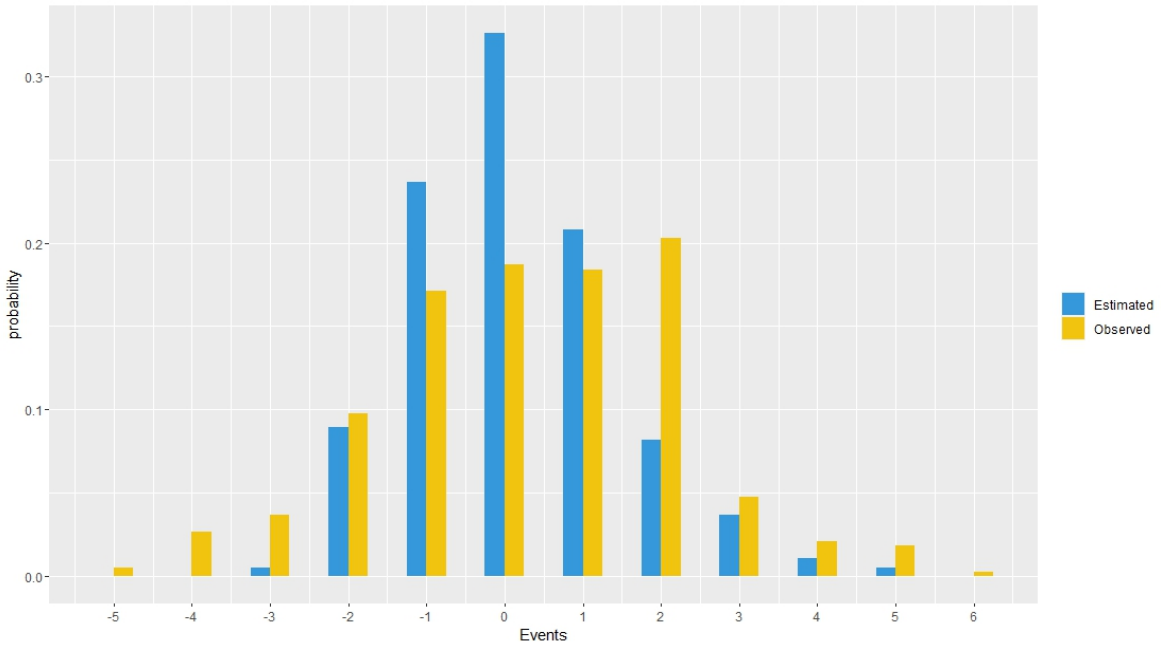


Figure 6.3: Observed vs Estimated probability of the goals difference events using the posterior predictive distribution.

While one might argue that the correlation between the number of goals scored by the home and away teams is negligible, it becomes apparent that such an oversimplification is untenable in this particular modeling framework. To address this inherent limitation, a novel model can be formulated based on the **Bivariate Poisson Distribution** 2.This alternative model introduces an additional parameter to explicitly account for the covariance between the two variables of interest, departing from the traditional independent Poisson regression. Assuming $log(\theta_{g,3}) = \gamma$, the joint distribution can be expressed as follows:

$$f(y_g, D_g + \mu | \theta_{g,1}, \theta_{g,2}, \theta_{g,3}) = e^{-(\theta_{g,1}+\theta_{g,2}+\theta_{g,3})} \frac{\theta_{g,1}^{y_g} \theta_{g,2}^{D_g+\mu}}{y_g!(D_g+\mu)!} \sum_{l=0}^{\min(y_g, D_g+\mu)} \binom{D_g+\mu}{l} \binom{y_g}{l} l! \left(\frac{\theta_{g,3}}{\theta_{g,1}\theta_{g,2}}\right)^l$$

It is straightforward the derivation of the likelihood function:

$$\begin{aligned}\mathbb{L}(\underline{\theta}) &= f\left(y_1, \ldots, y_{380}, D_1 + \mu, \ldots, D_{380} + \mu | \theta_{1,1}, \ldots, \theta_{380,1}, \theta_{1,2}, \ldots, \theta_{380,2}, \theta_{1,3}, \ldots, \theta_{380,3}\right) \\ &\overset{i.d}{=} \prod_{g=1}^{380} f\left(y_g, D_g + \mu | \theta_{g,1}, \theta_{g,2}, \theta_{g,3}\right)\end{aligned}$$

Assuming a Gaussian prior distribution centered at 0 with a large variance for the parameter γ, the joint posterior distribution can be easily established. For brevity, the detailed calculations are omitted.

Following the simulation, a discernible trend emerges, indicating a decrease significance of random effects. Notably, these effects converge towards values near zero, with small differentiations among them. Conversely, a noteworthy observation manifests in the model's tendency to estimate a correlation ranging approximately between 0.5 and 0.6 for the two variables of interest. This outcome implies that, within the formulated modeling framework, the stochastic variability introduced by the random effects become almost negligible, while the correlation parameter assumes an important role in characterizing the interdependence between the analyzed variables.

Despite successfully addressing the issue of negative values in estimated away scored goals, an examination of Figure 6.4 reveals a noteworthy observation: the posterior predictive distribution associated with the goals difference, along with that linked to home scored goals, demonstrates a tendency to adopt a more limited range of values. This phenomenon implies an improvement in the interpretability of the results, but a substantial reduction in predictive accuracy. In essence, the fundamental challenge that must be addressed to enable the joint modeling of home scored goals and goals difference revolves around striking a balance between the range of values assumed by the latter and the presence of negative estimates in away scored goals.

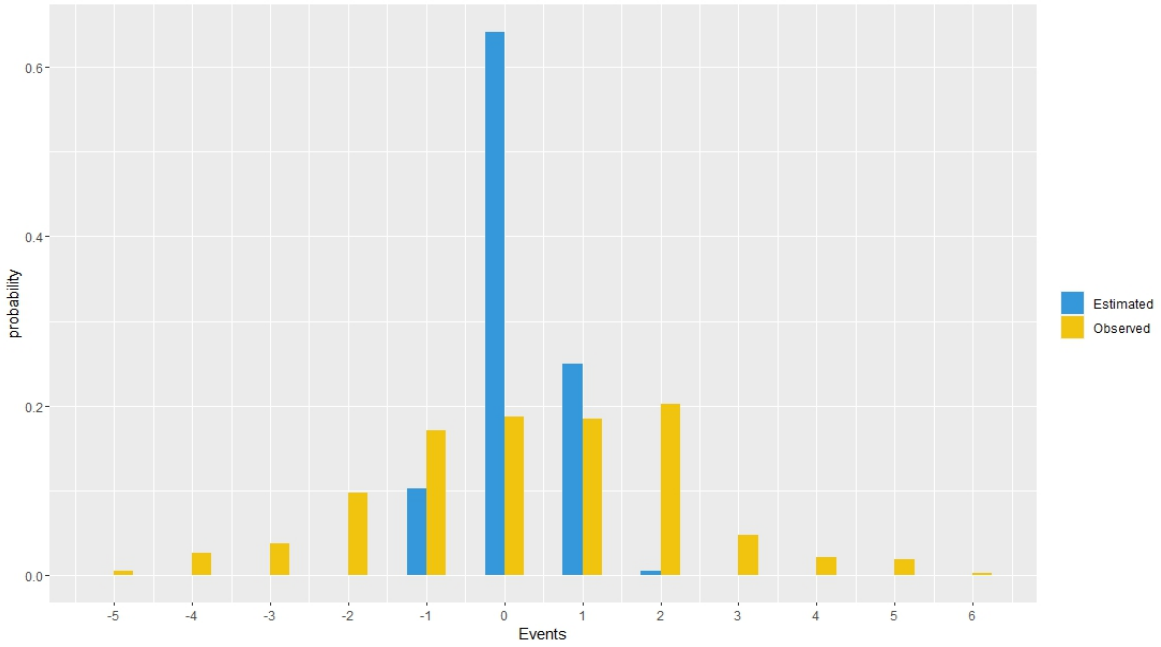


Figure 6.4: Observed vs Estimated probability of the goals difference events using the posterior predictive distribution.

7 Conclusion

Throughout the duration of this research project, various methodologies were explored to address the challenge of explaining the number of goals scored in a football match. Each model was crafted with unique characteristics, and a continuous effort was made to enhance both their explanatory power and predictive accuracy. A systematic progression was maintained, with each iteration building upon and refining its predecessor, trying to upgrade all the features that were inefficient in the previous model.

Initially, the investigation began by replicating the model introduced by [1], resulting in outcomes that were consistent and akin to the referenced work. Subsequently, an attempt was made to extend the previous model by incorporating a Zero-Inflated Poisson distribution to account for the count of 0 scored goals. Different combinations of inflation factors were explored to estimate a team's propensity to score zero goals, whether playing at home or away. Despite achieving some interpretability gains, these models proved to be excessively complex, leading to higher DIC values.

An intermediary phase in the investigation involved scrutinizing the variables that held genuine significance and exerted a substantive influence on the goals scored by the respective teams. Notably, my findings revealed that only temporal variables, indicative of the overall temporal dynamics, and variables associated with the perceived level of dangerousness possessed by offensive plays had an influence on the goal outcomes. This outcome aligns with expectations, underscoring the pivotal role played by temporal and dangerousness-related factors in shaping the goal-scoring dynamics of the two teams.

The final attempt involved the introduction of a Dynamic model, inspired by the concepts outlined in [20] and [15]. This approach aimed to capture the potential influence of time on the offensive and defensive abilities. The results obtained demonstrated the model's capability to replicate a team's offensive and defensive dynamics over the course of a season. Noteworthy, however, is the importance of careful parameter initialization, as it significantly influences their subsequent values during the time series and further improvements can be made not considering the evolution matrix G and the identity matrix.

An additional section was dedicated to the comprehensive analysis of goal differences, with a deliberate departure from the theoretical framework proposed by[11]. Notably, an empirical investigation was conducted, yielding intriguing outcomes. The primary emphasis was placed on the concurrent modeling of goals scored at home and goal differences, initially presupposing independence and subsequently incorporating dependencies. In conclusion, the findings garnered from this analytical endeavor proved to be noteworthy and contributory to the broader discourse.

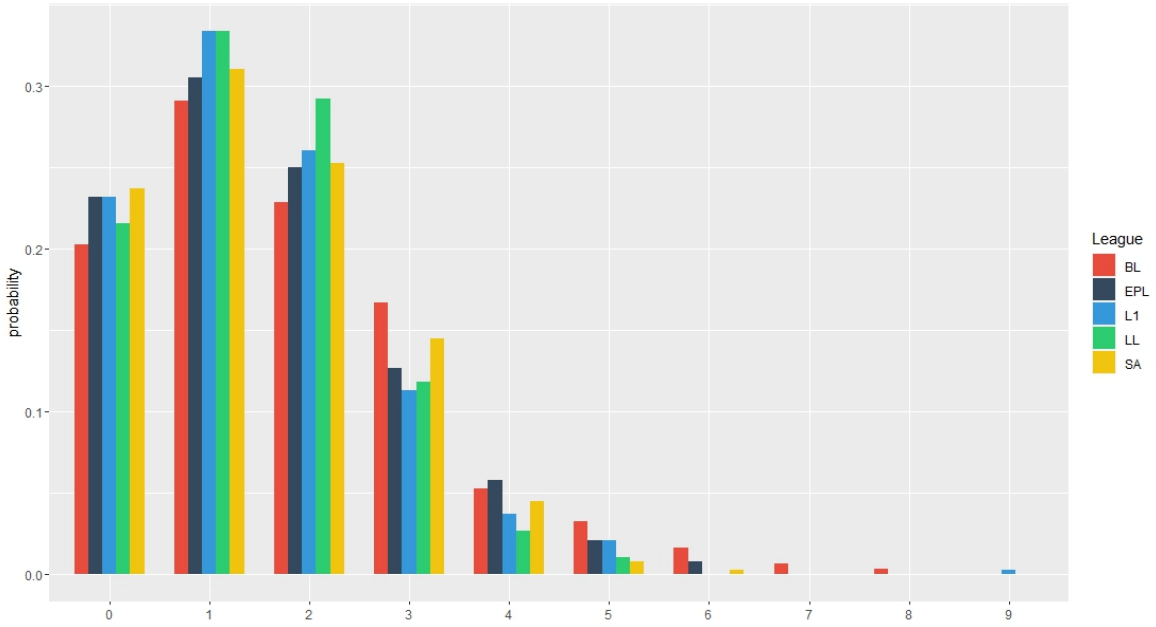
In summary, while there is space for further refinement in these models, significant insights were gained. The identification of crucial variables for explaining match outcomes was achieved, and an essential step forward was taken by reducing the reliance on random effects. Each research question now presents a distinct pathway for future exploration and improvement. Moreover, the presented project boasts complete reproducibility, providing the capability to customize the code outlined in Appendix B to suit different requirements. While it is noteworthy that the singular package enabling the utilization of Bayesian tools for football analysis is [4], it is arguable that my implementations offer greater flexibility for future applications.

7.1 AI Usage

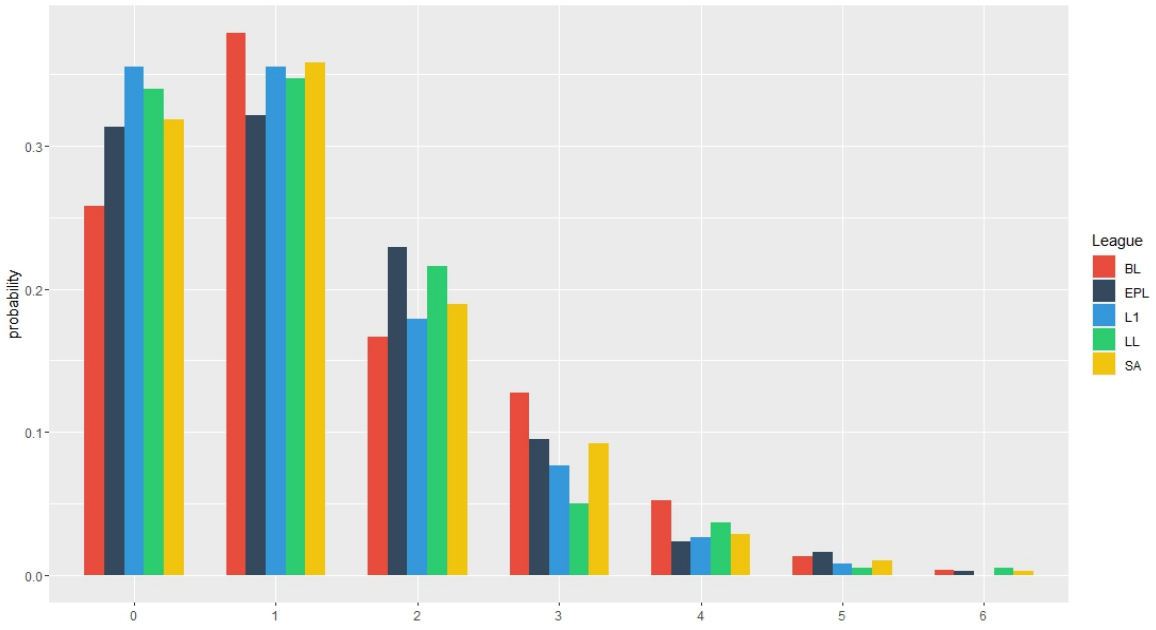
I used **DeepL** to improve the fluency of the text and to check for some errors, while I used **ChatGpt 3.5** to fix some RJags bugs. However, the use of AI in the whole project was minimal.

A Plots Appendix

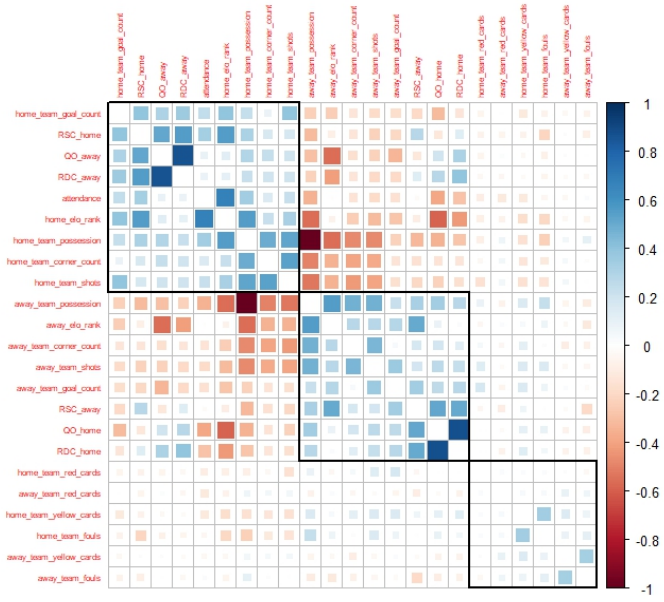
A.1 Frequency of goals scored at home in the top five leagues for the 2018/2019 season (Bundesliga, English Premier League, Ligue 1, Serie A, La Liga).



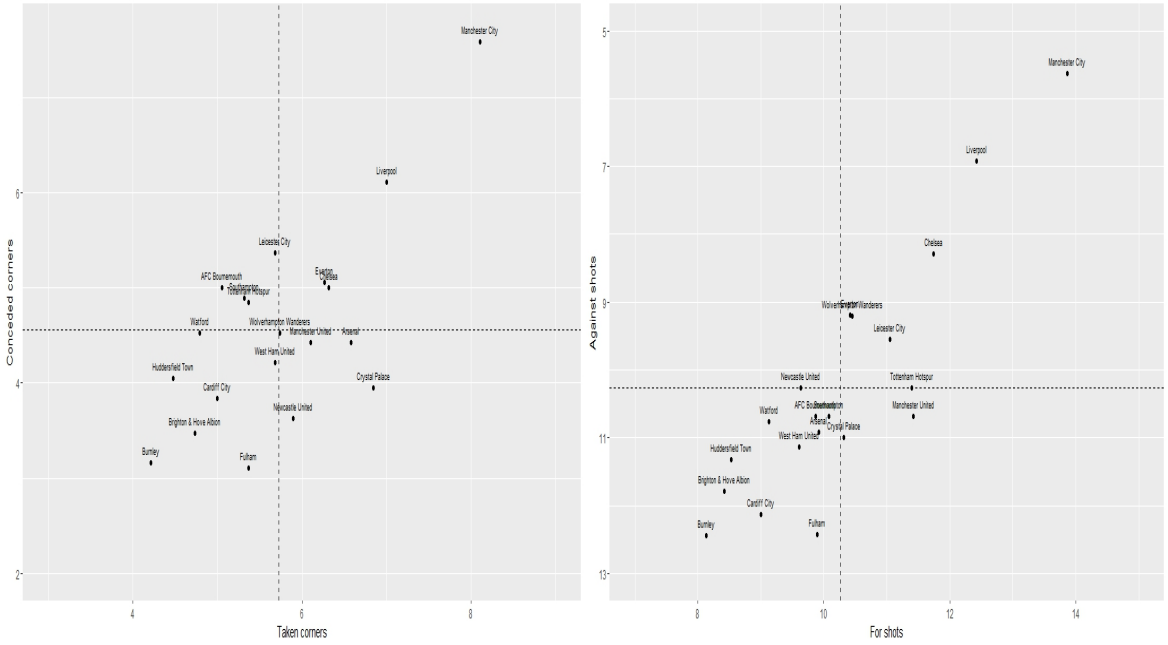
A.2 Frequency of goals scored away in the top five leagues for the 2018/2019 season.



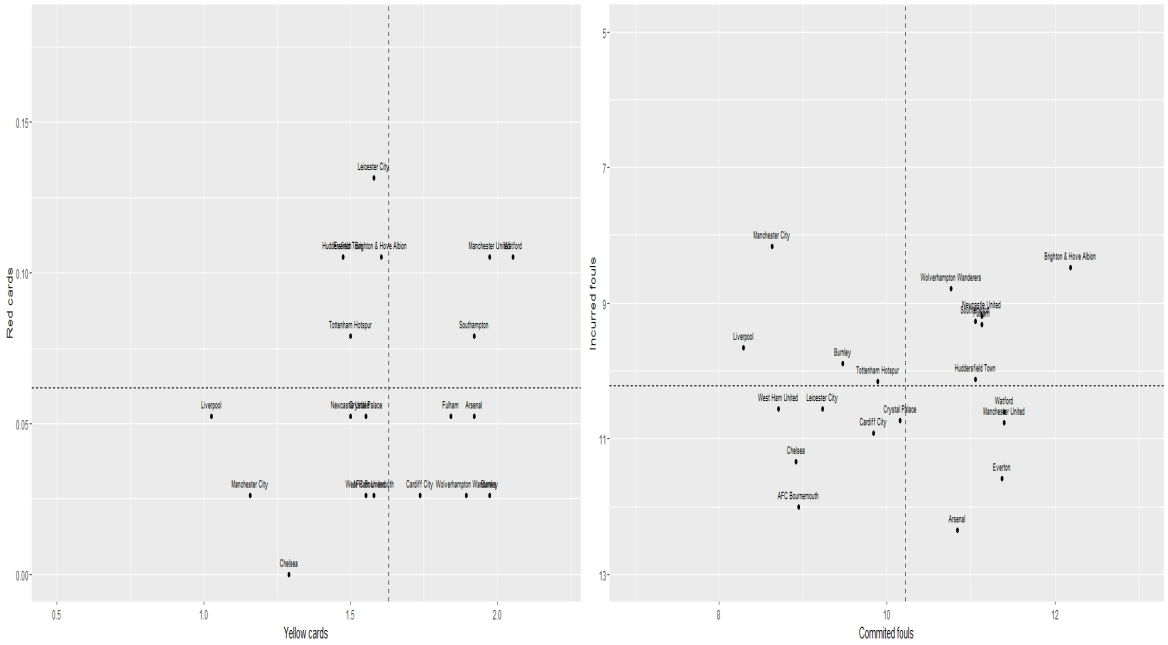
A.3 Correlogram associated to the quantitative variables of the dataset.



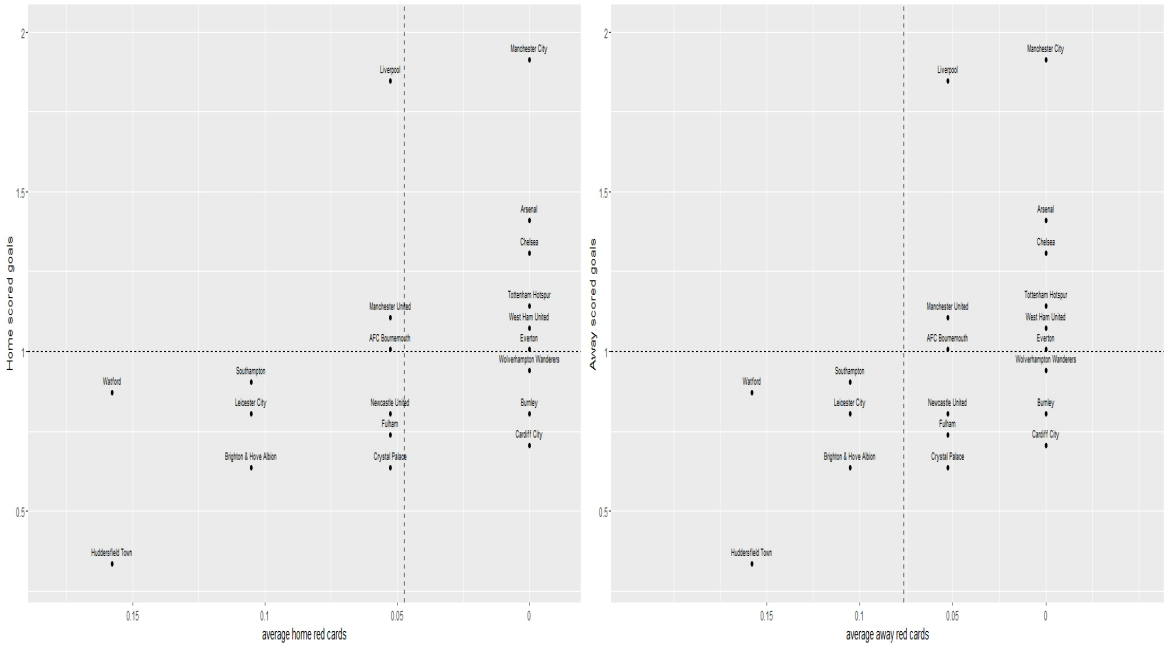
A.4 Average taken and conceded corners per team (left). Average taken and conceded shots per team (right). The two dotted lines represent the average effects between the teams.



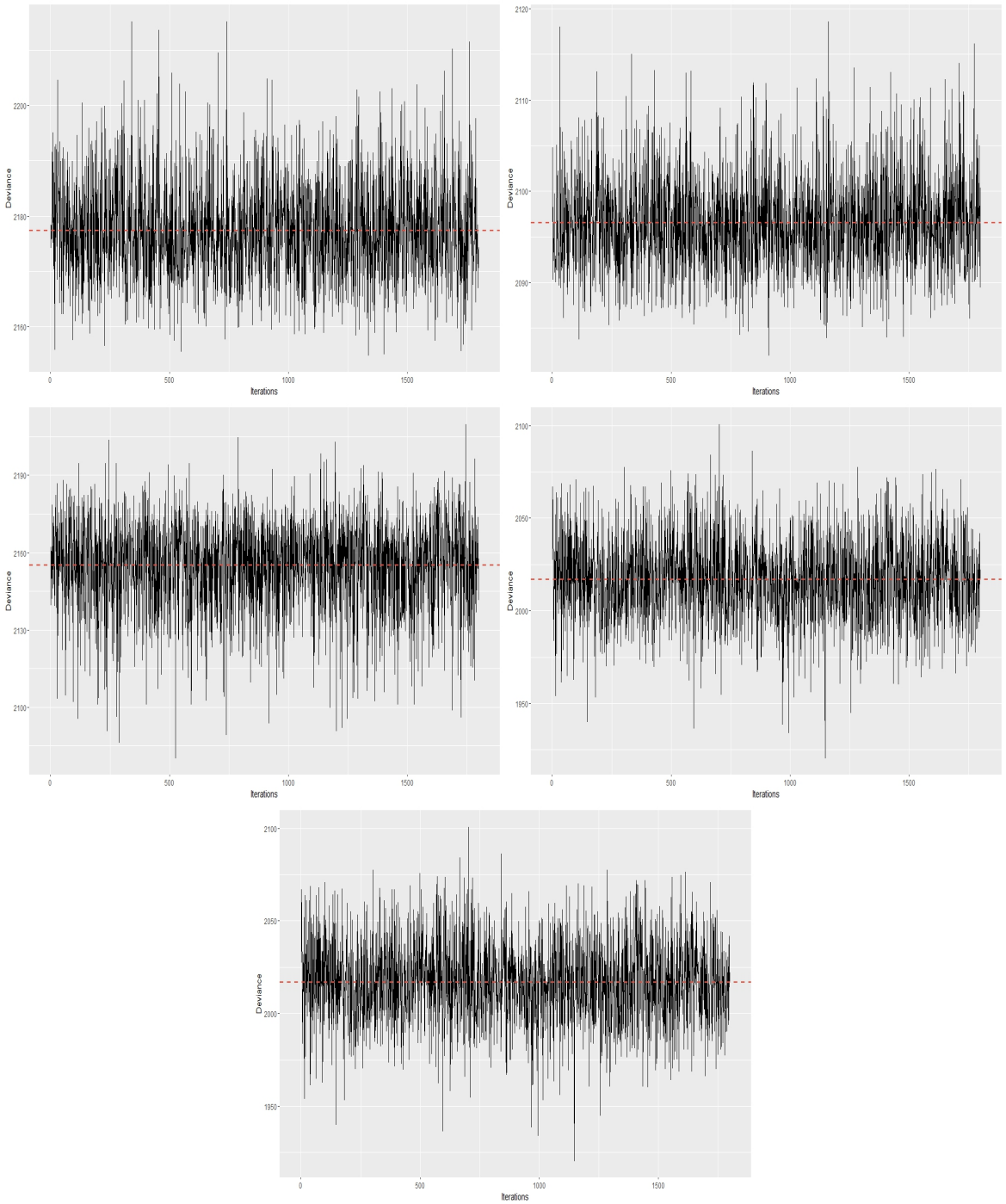
A.5 Average yellow and red cards per team (left). Average committed and incurred fouls per team (right). The two dotted lines represent the average effect between the teams.



A.6 Average home red cards vs average home scored goals (left). Average away red cards vs average away scored goals (right).



A.7 Traceplot of the deviance for all the models. The dotted red line represents the average value of the deviance.



A.8 Summary statistics for the parameters of the first model plus convergence diagnostics.

Param	Median	LB	UB	$\hat{R}$	ESS1	ESS2
$\mathbf{att_1}$	0.21	−0.04	0.44	1.004	900	900
$\mathbf{att_2}$	−0.17	−0.45	0.09	1.001	723.9	900
$\mathbf{att_3}$	0.08	−0.15	0.33	1.008	900	900
$\mathbf{att_4}$	−0.32	−0.58	0.00	1.002	900	900
$\mathbf{att_5}$	−0.64	−0.98	−0.31	1.001	900	900
$\mathbf{att_6}$	0.01	−0.24	0.25	1.004	900	900
$\mathbf{att_7}$	−0.09	−0.36	0.16	1.001	900	900
$\mathbf{att_8}$	0.49	0.27	0.70	1.002	900	900
$\mathbf{att_9}$	−0.11	−0.38	0.14	1.004	900	900
$\mathbf{att_{10}}$	0.32	0.10	0.54	1.001	900	790.5
$\mathbf{att_{11}}$	−0.32	−0.63	−0.06	1.003	900	900
$\mathbf{att_{12}}$	0.04	−0.21	0.28	1.006	900	900
$\mathbf{att_{13}}$	−0.01	−0.25	0.24	1.002	900	900
$\mathbf{att_{14}}$	0.24	0.02	0.47	1.001	900	900
$\mathbf{att_{15}}$	0.01	−0.24	0.24	1.002	785.03	900
$\mathbf{att_{16}}$	0.17	−0.06	0.38	1.001	900	900
$\mathbf{att_{17}}$	−0.11	−0.38	0.15	1.001	900	900
$\mathbf{att_{18}}$	0.56	0.34	0.75	1.001	900	900
$\mathbf{att_{19}}$	−0.31	−0.60	−0.05	1.002	487.5	743.9
$\mathbf{att_{20}}$	−0.01	−0.26	0.24	1.004	900	631.8
$\mathbf{def_1}$	0.04	−0.19	0.29	1.001	900	900
$\mathbf{def_2}$	−0.07	−0.33	0.18	1.02	900	900
$\mathbf{def_3}$	0.27	0.05	0.49	1.003	900	900
$\mathbf{def_4}$	0.39	0.19	0.62	1.01	900	900
$\mathbf{def_5}$	0.32	0.12	0.55	1.002	900	900
$\mathbf{def_6}$	0.11	−0.11	0.35	1.004	900	900
$\mathbf{def_7}$	−0.10	−0.35	0.17	1.001	469.43	900
$\mathbf{def_8}$	−0.61	−0.98	−0.27	1.002	734.14	900
$\mathbf{def_9}$	0.19	−0.03	0.41	1.003	900	900
$\mathbf{def_{10}}$	0.003	−0.25	0.22	1.004	900	900
$\mathbf{def_{11}}$	0.24	0.03	0.46	1.003	900	900
$\mathbf{def_{12}}$	−0.10	−0.35	0.14	1.002	900	900
$\mathbf{def_{13}}$	−0.06	−0.32	0.18	1.001	900	900
$\mathbf{def_{14}}$	−0.22	−0.48	0.06	1.004	900	900
$\mathbf{def_{15}}$	0.05	−0.19	0.30	1.003	900	900
$\mathbf{def_{16}}$	−0.22	−0.52	0.03	1.005	900	900
$\mathbf{def_{17}}$	0.23	0.01	0.46	1.001	900	900
$\mathbf{def_{18}}$	−0.56	−0.92	−0.25	1.001	900	900
$\mathbf{def_{19}}$	0.11	−0.13	0.33	1.001	900	900
$\mathbf{def_{20}}$	0.02	−0.22	0.26	1.001	758.8	900
home	0.36	0.28	0.45	1.008	900	900

A.9 summary statistics for the parameters of the second model plus convergence diagnostics.

Param	Median	LB	UB	$\hat{R}$	ESS1	ESS2
$\mathbf{att}_1$	-0.0179	-0.1670	0.1127	1.0019	900	900
$\mathbf{att}_2$	-0.0120	-0.1487	0.1411	1.0013	900	900
$\mathbf{att}_3$	0.0552	-0.0735	0.2106	1.0047	678.58	900
$\mathbf{att}_4$	0.0015	-0.1601	0.1444	1.0012	900	750.11
$\mathbf{att}_5$	-0.0473	-0.2115	0.0889	1.0013	900	881.23
$\mathbf{att}_6$	0.0220	-0.1236	0.1596	1.0016	900	900
$\mathbf{att}_7$	-0.0169	-0.1524	0.1303	1.0025	900	900
$\mathbf{att}_8$	-0.0402	-0.1679	0.0877	1.0023	900	900
$\mathbf{att}_9$	0.0092	-0.1322	0.1556	1.0069	900	900
$\mathbf{att}_{10}$	0.0275	-0.1108	0.1642	1.0021	809.34	900
$\mathbf{att}_{11}$	-0.0137	-0.1952	0.1222	1.0016	900	900
$\mathbf{att}_{12}$	0.0253	-0.1136	0.1744	1.0030	900	900
$\mathbf{att}_{13}$	-0.0102	-0.1552	0.1189	1.0012	900	900
$\mathbf{att}_{14}$	-0.0039	-0.1385	0.1249	1.0025	797.57	900
$\mathbf{att}_{15}$	0.0004	-0.1538	0.1281	1.0096	900	900.
$\mathbf{att}_{16}$	-0.0096	-0.1477	0.1331	1.0018	900	900
$\mathbf{att}_{17}$	0.0180	-0.1199	0.1691	1.0071	900	900
$\mathbf{att}_{18}$	0.0079	-0.1299	0.1312	1.0081	900	900
$\mathbf{att}_{19}$	-0.0211	-0.1697	0.1272	1.0023	900	900
$\mathbf{att}_{20}$	0.0283	-0.1042	0.1807	1.0011	900	763.76
$\mathbf{def}_1$	0.0294	-0.1129	0.1710	1.0012	900	900
$\mathbf{def}_2$	-0.0127	-0.1544	0.1230	1.0028	900	900
$\mathbf{def}_3$	0.0168	-0.1170	0.1561	1.0011	593.75	900
$\mathbf{def}_4$	0.0246	-0.1138	0.1622	1.0011	900	900
$\mathbf{def}_5$	0.0271	-0.1096	0.1584	1.0019	900	900
$\mathbf{def}_6$	0.0147	-0.1230	0.1547	1.0052	810.78	900
$\mathbf{def}_7$	-0.0190	-0.1418	0.1286	1.0013	900	900
$\mathbf{def}_8$	-0.0591	-0.2378	0.0807	1.0037	900	900
$\mathbf{def}_9$	0.0129	-0.1215	0.1395	1.0025	900	900
$\mathbf{def}_{10}$	0.0178	-0.1223	0.1589	1.0025	900	900
$\mathbf{def}_{11}$	0.0062	-0.1351	0.1389	1.0039	900	900
$\mathbf{def}_{12}$	-0.0150	-0.1559	0.1142	1.0016	900	900
$\mathbf{def}_{13}$	-0.0043	-0.1392	0.1295	1.0011	900	900
$\mathbf{def}_{14}$	-0.0123	-0.1680	0.1231	1.0051	900	900
$\mathbf{def}_{15}$	0.0183	-0.1159	0.1583	1.0011	900	900
$\mathbf{def}_{16}$	-0.0141	-0.1643	0.1351	1.0046	900	792.04
$\mathbf{def}_{17}$	0.0164	-0.1183	0.1456	1.0017	615.5	900
$\mathbf{def}_{18}$	-0.05	-0.22	0.09	1.001	900	900
$\mathbf{def}_{19}$	0.0163	-0.1118	0.1557	1.0015	900	900
$\mathbf{def}_{20}$	-0.0078	-0.1424	0.1191	1.0019	900	900
$\mathbf{b_0^{(1)}}$	0.1998	0.0629	0.3263	1.0016	900	900
$\mathbf{b_0^{(2)}}$	0.2181	0.0699	0.3883	1.0076	900	900
β_1	0.11	0.01	0.20	1.002	900	737.4
β_2	0.42	0.31	0.53	1.001	900	900
β_3	-0.44	-0.53	-0.35	1.002	900	900
β_4	0.0003	-1.3×10^{-04}	8.6×10^{-04}	1.001	900	900

A.10 Summary statistics for the parameters of the third model plus convergence diagnostics.

Param	Median	LB	UB	$\hat{R}$	ESS1	ESS2
$\mathbf{att_1}$	0.216	-0.008	0.440	1.002	900	900
$\mathbf{att_2}$	-0.173	-0.445	0.115	1.001	900	900
$\mathbf{att_3}$	0.094	-0.140	0.359	1.002	900	900
$\mathbf{att_4}$	-0.326	-0.623	-0.038	1.002	900	900
$\mathbf{att_5}$	-0.634	-1.015	-0.298	1.006	900	900
$\mathbf{att_6}$	0.019	-0.241	0.259	1.002	900	766.1
$\mathbf{att_7}$	-0.078	-0.319	0.212	1.001	655.4	900
$\mathbf{att_8}$	0.489	0.268	0.680	1.003	900	900
$\mathbf{att_9}$	-0.108	-0.377	0.155	1.002	900	900
$\mathbf{att_{10}}$	0.314	0.094	0.527	1.014	900	900
$\mathbf{att_{11}}$	-0.334	-0.656	-0.046	1.002	677.7	900
$\mathbf{att_{12}}$	0.037	-0.216	0.275	1.002	900	849.7
$\mathbf{att_{13}}$	-0.010	-0.260	0.250	1.001	900	900
$\mathbf{att_{14}}$	0.229	-0.020	0.452	1.002	900	823.3
$\mathbf{att_{15}}$	0.015	-0.233	0.252	1.003	900	900
$\mathbf{att_{16}}$	0.171	-0.064	0.413	1.001	900	900
$\mathbf{att_{17}}$	-0.114	-0.349	0.177	1.003	900	900
$\mathbf{att_{18}}$	0.548	0.330	0.742	1.012	900	900
$1\mathbf{att_{19}}$	-0.318	-0.612	-0.038	1.001	900	900
$\mathbf{att_{20}}$	-0.007	-0.262	0.246	1.002	900	900
$\mathbf{def}_1$	0.047	-0.237	0.262	1.001	900	900
$\mathbf{def}_2$	-0.074	-0.344	0.174	1.011	900	900
$\mathbf{def}_3$	0.275	0.049	0.514	1.001	900	900
$\mathbf{def}_4$	0.390	0.165	0.588	1.002	900	900
$\mathbf{def}_5$	0.319	0.093	0.525	1.001	900	900
$\mathbf{def}_6$	0.111	-0.130	0.346	1.001	900	900
$\mathbf{def}_7$	-0.116	-0.362	0.152	1.003	900	900
$\mathbf{def}_8$	-0.606	-0.985	-0.301	1.002	900	900
$\mathbf{def}_9$	0.203	-0.025	0.444	1.002	725.3	900
$\mathbf{def}_{10}$	-0.005	-0.249	0.252	1.001	900	900
$\mathbf{def}_{11}$	0.248	0.023	0.489	1.001	900	900
$\mathbf{def}_{12}$	-0.098	-0.345	0.164	1.002	900	900
$\mathbf{def}_{13}$	-0.067	-0.313	0.176	1.001	900	900
$\mathbf{def}_{14}$	-0.213	-0.490	0.057	1.005	900	900
$\mathbf{def}_{15}$	0.046	-0.194	0.281	1.002	900	900
$\mathbf{def}_{16}$	-0.228	-0.500	0.060	1.006	816.7	900
$\mathbf{def}_{17}$	0.243	-0.013	0.448	1.002	900	900
$\mathbf{def}_{18}$	-0.565	-0.919	-0.252	1.001	765.5	900
$\mathbf{def}_{19}$	0.114	-0.113	0.356	1.004	900	900
$\mathbf{def}_{20}$	0.028	-0.217	0.282	1.001	900	900
$\mathbf{p}_1$	1.26×10^{-2}	6.2×10^{-6}	4.5×10^{-2}	1.001	801.9	900
$\mathbf{p}_2$	9.1×10^{-3}	9.4×10^{-6}	3.4×10^{2}	1.001	900	900
home	0.38	0.28	0.46	1.004	900	900

A.11 Summary statistics for the parameters of the modified zero-inflated model with **Gamma** hyperprior plus convergence diagnostics.

Param	Median	LB	UB	$\hat{R}$	ESS1	ESS2
$\mathbf{att_1}$	0.22	-0.04	0.43	1.003	900	900
$\mathbf{att_2}$	-0.15	-0.48	0.10	1.007	900	900
$\mathbf{att_3}$	0.11	-0.12	0.37	1.002	900	900
$\mathbf{att_4}$	-0.32	-0.63	-0.04	1.001	900	900
$\mathbf{att_5}$	-0.63	-0.98	-0.29	1.002	900	900
$\mathbf{att_6}$	0.012	-0.25	0.28	1.002	900	900
$\mathbf{att_7}$	-0.08	-0.34	0.19	1.001	900	900
$\mathbf{att_8}$	0.46	0.26	0.68	1.001	900	900
$\mathbf{att_9}$	-0.11	-0.39	0.14	1.003	900	900
$\mathbf{att_{10}}$	0.29	0.08	0.51	1.011	900	900
$\mathbf{att_{11}}$	-0.32	-0.62	-0.03	1.002	900	900
$\mathbf{att_{12}}$	0.03	-0.22	0.29	1.002	900	900
$\mathbf{att_{13}}$	0.001	-0.25	0.27	1.001	900	900
$\mathbf{att_{14}}$	0.23	-0.02	0.48	1.002	900	900
$\mathbf{att_{15}}$	0.02	-0.23	0.28	1.001	900	900
$\mathbf{att_{16}}$	0.18	-0.07	0.41	1.001	894.7	900
$\mathbf{att_{17}}$	-0.12	-0.38	0.15	1.003	706.6	900
$\mathbf{att_{18}}$	0.52	0.32	0.73	1.004	900	900
$\mathbf{att_{19}}$	-0.32	-0.63	-0.03	1.001	900	900
$\mathbf{att_{20}}$	0.01	-0.23	0.30	1.001	900	900
$\mathbf{def}_1$	0.046	-0.218	0.301	1.001	900	900
$\mathbf{def}_2$	-0.07	-0.369	0.162	1.002	900	900
$\mathbf{def}_3$	0.283	0.064	0.508	1.001	900	900
$\mathbf{def}_4$	0.388	0.179	0.628	1.004	839.587	900
$\mathbf{def}_5$	0.31	0.086	0.532	1.001	798.018	900
$\mathbf{def}_6$	0.099	-0.132	0.342	1.001	900	900
$\mathbf{def}_7$	-0.117	-0.367	0.157	1.001	900	691.952
$\mathbf{def}_8$	-0.61	-0.974	-0.273	1.003	900	900
$\mathbf{def}_9$	0.199	-0.03	0.437	1.002	900	900
$\mathbf{def}_{10}$	-0.017	-0.257	0.235	1.002	900	813.338
$\mathbf{def}_{11}$	0.262	0.028	0.493	1.002	900	900
$\mathbf{def}_{12}$	-0.093	-0.347	0.166	1.013	900	900
$\mathbf{def}_{13}$	-0.069	-0.315	0.205	1.001	900	900
$\mathbf{def}_{14}$	-0.221	-0.506	0.066	1.001	900	900
$\mathbf{def}_{15}$	0.053	-0.19	0.288	1.003	793.179	900
$\mathbf{def}_{16}$	-0.222	-0.514	0.05	1.001	900	900
$\mathbf{def}_{17}$	0.241	-0.009	0.467	1.001	900	900
$\mathbf{def}_{18}$	-0.564	-0.915	-0.219	1.003	777.209	900
$\mathbf{def}_{19}$	0.102	-0.125	0.336	1.004	900	900
$\mathbf{def}_{20}$	0.022	-0.244	0.267	1.003	900	900
$\mathbf{p}_{1,1}$	0.059	0.007	0.129	1.001	900	873.334
$\mathbf{p}_{1,2}$	0.062	0.005	0.13	1.001	900	900
$\mathbf{p}_{1,3}$	0.056	0.002	0.122	1.003	900	900
$\mathbf{p}_{1,4}$	0.057	0.001	0.127	1.004	900	900
$\mathbf{p}_{1,5}$	0.057	0.001	0.122	1.009	900	900
$\mathbf{p}_{1,6}$	0.056	0	0.118	1.005	900	900
$\mathbf{p}_{1,7}$	0.058	0.001	0.122	1.006	900	900

$\mathbf{p}_{1,8}$	0.046	0	0.103	1.013	900	900
$\mathbf{p}_{1,9}$	0.051	0.001	0.111	1.012	900	900
$\mathbf{p}_{1,10}$	0.045	0	0.103	1.003	900	900
$\mathbf{p}_{1,11}$	0.057	0	0.122	1.003	900	900
$\mathbf{p}_{1,12}$	0.049	0	0.112	1.002	900	900
$\mathbf{p}_{1,13}$	0.061	0.004	0.135	1.003	780.887	900
$\mathbf{p}_{1,14}$	0.054	0	0.114	1.001	900	900
$\mathbf{p}_{1,15}$	0.058	0.001	0.125	1.003	900	900
$\mathbf{p}_{1,16}$	0.054	0.002	0.119	1.01	900	799.374
$\mathbf{p}_{1,17}$	0.051	0	0.113	1.005	900	900
$\mathbf{p}_{1,18}$	0.044	0	0.101	1.017	900	900
$\mathbf{p}_{1,19}$	0.056	0.004	0.123	1.001	900	900
$\mathbf{p}_{1,20}$	0.066	0.006	0.144	1.001	900	900
$\mathbf{p}_{2,1}$	0.043	0.001	0.095	1.001	900	900
$\mathbf{p}_{2,2}$	0.048	0	0.105	1.003	900	900
$\mathbf{p}_{2,3}$	0.053	0.001	0.115	1.001	743.516	900
$\mathbf{p}_{2,4}$	0.049	0.002	0.106	1.006	900	900
$\mathbf{p}_{2,5}$	0.049	0.002	0.11	1.001	900	900
$\mathbf{p}_{2,6}$	0.047	0	0.105	1.002	900	900
$\mathbf{p}_{2,7}$	0.048	0.001	0.105	1.003	900	900
$\mathbf{p}_{2,8}$	0.044	0	0.099	1.003	900	900
$\mathbf{p}_{2,9}$	0.047	0.003	0.106	1.003	900	900
$\mathbf{p}_{2,10}$	0.044	0	0.102	1.001	900	900
$\mathbf{p}_{2,11}$	0.052	0	0.112	1.001	900	900
$\mathbf{p}_{2,12}$	0.051	0	0.113	1.001	900	900
$\mathbf{p}_{2,13}$	0.041	0	0.097	1.01	900	900
$\mathbf{p}_{2,14}$	0.044	0	0.096	1.001	900	726.16
$\mathbf{p}_{2,15}$	0.052	0.001	0.112	1.001	900	673.38
$\mathbf{p}_{2,16}$	0.051	0.002	0.12	1.005	900	670.529
$\mathbf{p}_{2,17}$	0.052	0.001	0.115	1.001	900	900
$\mathbf{p}_{2,18}$	0.043	0.002	0.096	1.001	815.97	900
$\mathbf{p}_{2,19}$	0.05	0	0.106	1.001	900	864.431
$\mathbf{p}_{2,20}$	0.043	0	0.095	1.001	900	900
home	0.405	0.313	0.498	1.002	900	900

A.12 Summary statistics for the parameters of the modified zero-inflated model without hyperprior plus convergence diagnostics.

Param	Median	LB	UB	$\hat{R}$	ESS1	ESS2
$\mathbf{att_1}$	0.211	-0.043	0.463	1.007	900	900
$\mathbf{att_2}$	-0.136	-0.422	0.161	1.001	900	900
$\mathbf{att_3}$	0.119	-0.15	0.395	1.006	812.388	900
$\mathbf{att_4}$	-0.311	-0.655	-0.03	1.002	900	900
$\mathbf{att_5}$	-0.602	-1.001	-0.254	1.003	900	900
$\mathbf{att_6}$	0.007	-0.272	0.244	1.001	547.34	900
$\mathbf{att_7}$	-0.076	-0.377	0.179	1.009	900	900
$\mathbf{att_8}$	0.432	0.215	0.649	1.008	734.728	900
$\mathbf{att_9}$	-0.138	-0.419	0.134	1.004	900	900
$\mathbf{att_{10}}$	0.262	0.043	0.494	1.001	900	817.845
$\mathbf{att_{11}}$	-0.282	-0.582	0.045	1.001	900	900
$\mathbf{att_{12}}$	0.018	-0.252	0.259	1.001	900	900
$\mathbf{att_{13}}$	0.003	-0.269	0.271	1.001	900	900
$\mathbf{att_{14}}$	0.193	-0.063	0.414	1.004	900	900
$\mathbf{att_{15}}$	0.046	-0.238	0.329	1.004	900	900
$\mathbf{att_{16}}$	0.171	-0.077	0.42	1.002	900	900
$\mathbf{att_{17}}$	–0.121	-0.416	0.136	1.004	900	900
$\mathbf{att_{18}}$	0.488	0.289	0.698	1.001	900	900
$\mathbf{att_{19}}$	-0.321	-0.613	-0.013	1.001	900	900
$\mathbf{att_{20}}$	0.056	-0.218	0.331	1.003	900	900
$\mathbf{def}_1$	0.037	-0.234	0.258	1.002	900	900
$\mathbf{def}_2$	-0.07	-0.328	0.223	1.001	727.183	900
$\mathbf{def}_3$	0.28	0.043	0.522	1.002	900	900
$\mathbf{def}_4$	0.377	0.142	0.592	1.001	900	900
$\mathbf{def}_5$	0.295	0.074	0.52	1.001	900	900
$\mathbf{def}_6$	0.098	-0.141	0.33	1.003	900	900
$\mathbf{def}_7$	-0.11	-0.367	0.163	1.001	900	900
$\mathbf{def}_8$	-0.586	-0.963	-0.251	1.001	900	900
$\mathbf{def}_9$	0.204	-0.035	0.456	1.006	900	900
$\mathbf{def}_{10}$	-0.03	-0.287	0.231	1.001	900	735.873
$\mathbf{def}_{11}$	0.286	0.045	0.541	1.005	900	900
$\mathbf{def}_{12}$	-0.076	-0.358	0.172	1.001	900	900
$\mathbf{def}_{13}$	-0.066	-0.368	0.18	1.001	900	900
$\mathbf{def}_{14}$	-0.212	-0.482	0.09	1.002	900	900
$\mathbf{def}_{15}$	0.03	-0.21	0.279	1.004	900	900
$\mathbf{def}_{16}$	-0.206	-0.495	0.079	1.003	900	900
$\mathbf{def}_{17}$	0.243	0.015	0.487	1.001	900	900
$\mathbf{def}_{18}$	-0.56	-0.938	-0.232	1.004	900	900
$\mathbf{def}_{19}$	0.099	-0.163	0.33	1.001	805.392	900
$\mathbf{def}_{20}$	0.03	-0.224	0.291	1.003	900	900
$\mathbf{p}_{1,1}$	0.129	0	0.352	1.004	900	900
$\mathbf{p}_{1,2}$	0.161	0	0.383	1.003	900	900
$\mathbf{p}_{1,3}$	0.108	0.001	0.298	1.01	900	900
$\mathbf{p}_{1,4}$	0.115	0	0.324	1.001	900	900
$\mathbf{p}_{1,5}$	00.16	0	0.445	1.001	900	792.48
$\mathbf{p}_{1,6}$	0.104	0.001	0.29	1.001	900	900
$\mathbf{p}_{1,7}$	0.122	0	0.325	1.001	900	900

$\mathbf{p}_{1,8}$	0.042	0	0.157	1.002	900	774.078
$\mathbf{p}_{1,9}$	0.063	0	0.221	1.001	900	814.564
$\mathbf{p}_{1,10}$	0.041	0	0.16	1.001	900	769.881
$\mathbf{p}_{1,11}$	0.13	0	0.356	1.003	900	900
$\mathbf{p}_{1,12}$	0.056	0	0.201	1.001	900	900
$\mathbf{p}_{1,13}$	0.185	0	0.411	1.001	900	900
$\mathbf{p}_{1,14}$	0.076	0	0.239	1.005	837.682	865.168
$\mathbf{p}_{1,15}$	0.148	0	0.371	1.003	900	695.521
$\mathbf{p}_{1,16}$	0.09	0	0.25	1.001	900	811.68
$\mathbf{p}_{1,17}$	0.056	0	0.202	1.006	723.493	781.879
$\mathbf{p}_{1,18}$	0.035	0	0.138	1.002	900	900
$\mathbf{p}_{1,19}$	0.096	0	0.298	1.001	844.614	900
$\mathbf{p}_{1,20}$	0.27	0.002	0.504	1.005	810.002	900
$\mathbf{p}_{2,1}$	0.048	0	0.179	1.005	499.579	900
$\mathbf{p}_{2,2}$	0.11	0	0.328	1.002	900	900
$\mathbf{p}_{2,3}$	0.18	0	0.419	1.001	900	867.362
$\mathbf{p}_{2,4}$	0.141	0	0.414	1.002	900	900
$\mathbf{p}_{2,5}$	0.138	0	0.407	1.002	900	900
$\mathbf{p}_{2,6}$	0.092	0	0.289	1.001	900	900
$\mathbf{p}_{2,7}$	0.091	0	0.301	1.002	900	900
$\mathbf{p}_{2,8}$	0.054	0	0.199	1.002	900	900
$\mathbf{p}_{2,9}$	0.109	0	0.346	1.007	900	900
$\mathbf{p}_{2,10}$	0.058	0	0.22	1.001	900	900
$\mathbf{p}_{2,11}$	0.209	0	0.489	1.001	900	900
$\mathbf{p}_{2,12}$	0.132	0	0.36	1.02	900	734.865
$\mathbf{p}_{2,13}$	0.042	0	0.175	1.008	900	850.262
$\mathbf{p}_{2,14}$	0.054	0	0.201	1.008	900	900
$\mathbf{p}_{2,15}$	0.173	0	0.42	1.003	803.51	900
$\mathbf{p}_{2,16}$	0.137	0	0.369	1.002	900	900
$\mathbf{p}_{2,17}$	0.168	0	0.435	1.001	900	900
$\mathbf{p}_{2,18}$	0.046	0	0.175	1.001	900	817.518
$\mathbf{p}_{2,19}$	0.107	0	0.337	1.004	900	900
$\mathbf{p}_{2,20}$	0.055	0	0.2	1.001	900	900
home	0.451	0.362	0.542	1.004	900	900

B Codes Appendix

- **Bayesian Hierarchical Model without covariates**

```
cat(model{

  for (g in 1:ngames) {

    y1[g] ~ dpois(theta[g,1])
    y2[g] ~ dpois(theta[g,2])
    # Predictive distribution
    ynew[g,1] ~ dpois(theta[g,1])
    ynew[g,2] ~ dpois(theta[g,2])

    log(theta[g,1]) <- home + att[home_team[g]] + def[away_team[g]]
    log(theta[g,2]) <- att[away_team[g]] + def[home_team[g]]
  }

  home ~ dnorm(0,0.0001)

  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[])
  }

  mu.att ~ dnorm(0,0.0001)
  mu.def ~ dnorm(0,0.0001)
  tau.att ~ dgamma(.01,.01)
  tau.def ~ dgamma(.01,.01)
}", file="project_model_1_jags.txt",fill=TRUE)
```

- **Bayesian Hierarchical Model with covariates**

```
cat("model{
  for (g in 1:ngames) {
    y1[g] ~ dpois(theta[g,1])
    y2[g] ~ dpois(theta[g,2])

    ynew[g,1] ~ dpois(theta[g,1])
    ynew[g,2] ~ dpois(theta[g,2])

    log(theta[g,1]) <-b0[attendance[g]] + beta1 * RSC_h[g]+ beta2*RDC_a[g] +
    beta3 * QO_h[g]+ beta4*ELO_h[g]+ att[home_team[g]] + def[away_team[g]]
    log(theta[g,2]) <- beta1*RSC_a[g] + beta2*RDC_h[g] + beta3*QO_a[g] +
    beta4 * ELO_a[g] + att[away_team[g]] + def[home_team[g]]
  }
  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[])
  }

  mu.att ~ dnorm(0,0.0001)
  mu.def ~ dnorm(0,0.0001)
  tau.att ~ dgamma(.01,.01)
```

```
  tau.def ~ dgamma(.01,.01)

  sigma_0  ~ dgamma(.01,.01)
  for (j in 1:2){
    b0[j] ~ dnorm(0,sigma_0)

  }
  beta1 ~ dnorm(0,0.0001)
  beta2 ~ dnorm(0,0.0001)
  beta3 ~ dnorm(0,0.0001)
  beta4 ~ dnorm(0,0.0001)
}", file="covariates_jags.txt",fill=TRUE)
```

- **Zero-Inflated Bayesian Hierarchical model**

```
cat("model{
  for (g in 1:ngames) {
    y1[g] ~ dpois(theta1[g,1])
    y2[g] ~ dpois(theta2[g,2])

    ynew[g,1] ~ dpois(theta1[g,1])
    ynew[g,2] ~ dpois(theta2[g,2])

    theta[g,1] <- exp(home + att[home_team[g]] + def[away_team[g]])
    theta[g,2] <- exp(att[away_team[g]] + def[home_team[g]])
    z[g] ~ dbern(psi)
    z2[g] ~ dbern(psi2)
    theta1[g,1] <- z[g] * theta[g, 1] + 0.00001
    theta2[g,2] <- z2[g] * theta[g, 2] + 0.00001
  }

  home ~ dnorm(0,0.0001)
  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[])
  }

  mu.att ~ dnorm(0,0.0001)
  mu.def ~ dnorm(0,0.0001)
  tau.att ~ dgamma(.01,.01)
  tau.def ~ dgamma(.01,.01)
  psi ~ dunif(0, 1)
  psi2 ~ dunif(0, 1)
}", file="project_model_3_jags.txt",fill=TRUE)
```

- **Modified Zero-Inflated Bayesian Hierarchical model using Gamma and Half Cauchy hyperpriors**

```
cat("model{

  for (g in 1:ngames) {

    y1[g] ~ dpois(theta1[g,1])
```

```
    y2[g] ~ dpois(theta2[g,2])

    ynew[g,1] ~ dpois(theta1[g,1])
    ynew[g,2] ~ dpois(theta2[g,2])

    theta[g,1] <- exp(home + att[home_team[g]] + def[away_team[g]])
    theta[g,2] <- exp(att[away_team[g]] + def[home_team[g]])
    z[g] ~ dbern(psi[home_team[g]])
    z2[g] ~ dbern(psi2[away_team[g]])
    theta1[g,1] <- (1 - z[g]) * theta[g, 1] + 0.00001
    theta2[g,2] <- (1 - z2[g]) * theta[g, 2] + 0.00001

  }
  home ~ dnorm(0,0.0001)
  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[])
    psi[t] ~ dnorm(mu.p1, tau.p1) T(0, 1)
    psi2[t] ~ dnorm(mu.p2, tau.p2) T(0,1)
  }

  mu.att ~ dnorm(0,0.0001)
  mu.def ~ dnorm(0,0.0001)
  tau.att ~ dgamma(.01,.01)
  tau.def ~ dgamma(.01,.01)
  mu.p1 ~ dbeta(1, 1)
  mu.p2 ~ dbeta(1 ,1)
  tau.p1 ~ dgamma(0.001,0.001)
  tau.p2 ~ dgamma(0.001,0.001)

  #### Half Cauchy parametrization #####
  #tau.p1 ~ dt(0, 30, 1)T(0, )
  #tau.p2 ~ dt(0, 30, 1)T(0, )

}", file="project_tanti_p.txt",fill=TRUE)
```

- **Modified Zero-Inflated Bayesian Hierarchical model results without hyperpriors**

```
cat("model{

  for (g in 1:ngames) {

    y1[g] ~ dpois(theta1[g,1])
    y2[g] ~ dpois(theta2[g,2])

    ynew[g,1] ~ dpois(theta1[g,1])
    ynew[g,2] ~ dpois(theta2[g,2])

    theta[g,1] <- exp(home + att[home_team[g]] + def[away_team[g]])
    theta[g,2] <- exp(att[away_team[g]] + def[home_team[g]])
    z[g] ~ dbern(psi[home_team[g]])
    z2[g] ~ dbern(psi2[away_team[g]])
    theta1[g,1] <- (1 - z[g]) * theta[g, 1] + 0.00001
    theta2[g,2] <- (1 - z2[g]) * theta[g, 2] + 0.00001

```

```
  }
  home ~ dnorm(0,0.0001)
  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[])
    psi[t] ~ dbeta(1, 1)
    psi2[t] ~ dbeta(1, 1)
  }
  mu.att ~ dnorm(0,0.0001)
  mu.def ~ dnorm(0,0.0001)
  tau.att ~ dgamma(.01,.01)
  tau.def ~ dgamma(.01,.01)
}", file="project_model_4_jags.txt",fill=TRUE)
```

- **Explanatory Model**

```
cat("model{
  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])

    log(theta[g, 1]) <- X1[g, ]%*%Beta1
    log(theta[g, 2]) <- X2[g, ]%*%Beta2

  }

  tau ~ dgamma(0.1, 0.1)
  tau2 ~ dgamma(0.1, 0.1)
  sigma <- 1/tau
  sigma2 <- 1/tau2
  V1 <- sigma*(n)*D1[1:12, 1:12]
  V2 <- sigma2*(n)*D2[1:11, 1:11]

  m <- rep(0, 12)
  m2 <- rep(0, 11)
  Beta1 ~ dmnorm.vcov(m, V1)
  Beta2 ~ dmnorm.vcov(m2, V2)

}", file = "p.txt", fill = TRUE)
```

- **Dynamic Model**

```
cat("model{

  for (t in 1:T) {
    for (g in 1:10){

      y1[t,g] ~ dpois(theta1[t, g])
      y2[t,g] ~ dpois(theta2[t, g])
```

```
8
9       ynew1[t, g] ~ dpois(theta1[t, g])
10      ynew2[t, g] ~ dpois(theta2[t, g])
11
12      log(theta1[t, g]) <- home + att[home_team[g + (t -1)*10], t]
13      + def[away_team[g + (t - 1)*10], t]
14      log(theta2[t, g]) <- away + att[away_team[g + (t - 1)*10], t]
15      + def[home_team[g + (t - 1)*10], t]
16    }
17  }
18
19  home ~ dnorm(0, 0.001)
20  away ~ dnorm(0, 0.001)
21
22
23  S0 <- ((n*sigma_0)/(n-1))*(I + i)
24  S00 <- inverse(S0)
25  m0 <- rep(0, n-1)
26
27  ########## attack parameters
28  c0 ~ dmnorm(m0, S00)
29  c1 <- c(c0, 0)
30  u0 <- J%*%c1
31  att0 <- m_att + u0
32
33  S1 <- ((n*sigma)/(n-1))*(I + i)
34  S11 <- inverse(S1)
35  c2 ~ dmnorm(m0, S11)
36  c3 <- c(c2, 0)
37  u1 <- J%*%c3
38  att[1:20,1] <- att0 + u1
39
40
41  def0 <- m_def + u0
42
43  d2 ~ dmnorm(m0, S11)
44  d3 <- c(d2, 0)
45  uu1 <- J%*%d3
46  def[1:20,1] <- def0 + uu1
47
48  for (t in 2:T){
49
50    c_t[1:19,t] ~ dmnorm(m0, S11)
51    c_2t[1:20, t] <- c(c_t[1:19, t], 0)
52    u[1:20, t] <- J %*% c_2t[1:20, t]
53    att[1:20,t] <- att[1:20,t-1] + u[1:20, t]
54
55    d_t[1:19,t] ~ dmnorm(m0, S11)
56    d_2t[1:20, t] <- c(d_t[1:19, t], 0)
57    uu3[1:20, t] <- J %*% d_2t[1:20, t]
58    def[1:20,t] <- def[1:20,t-1] + uu3[1:20, t]
59
60  }
61
62  sigma ~ dgamma(3.6, 600)
63  sigma_0 <- 0.003
64}", file="DGLM4.txt",fill=TRUE)
```

- **Bonus Section**

```
cat("model{
  for (g in 1:ngames) {

    y1[g] ~ dpois(theta[g,1])
    y2[g] ~ dpois(theta[g,2])

    ynew[g,1] ~ dpois(theta[g,1])
    ynew[g,2] ~ dpois(theta[g,2])

    ynew2[g,1] <- ynew[g,1]
    ynew2[g,2] <- ynew[g,2] - u


    log(theta[g,1]) <- home + att[home_team[g]] - def[away_team[g]]
    log(theta[g,2]) <- diff + (att[home_team[g]]- def[away_team[g]])
    + (-att[away_team[g]]+def[home_team[g]])
  }

  home ~ dnorm(0,0.0001)
  diff ~ dnorm(0,0.0001)
  for (t in 1:nteams){
    att.star[t] ~ dnorm(0, tau.att)
    def.star[t] ~ dnorm(0, tau.def)
    att[t] <- att.star[t] - mean(att.star[])
    def[t] <- def.star[t] - mean(def.star[])
  }

  tau.att ~ dgamma(0.1,0.1)
  tau.def ~ dgamma(0.1,0.1)
}", file="project_model_1_jags.txt",fill=TRUE)
```

- **Bonus Section: Bivariate Poisson**

```
cat("model{

  for (g in 1:ngames) {

    ynew[g,1] ~ dpois(theta[g,1])
    ynew[g,2] ~ dpois(theta[g,2])
    ynew[g,3] ~ dpois(theta[g,3])

    ynew2[g,1] <- ynew[g,1]
    ynew2[g,2] <- ynew[g,2] - u

    for (i in 1:(z[g] + 1)){
      summ[g, i] <- exp(logfact(y1[g]) - logfact(i- 1) - logfact(y1[g] - i + 1)
      + logfact(y2[g]) - logfact(i - 1) - logfact(y2[g] - i + 1) + (i - 1)
      *log(theta[g, 3]) - (i -1)*log(theta[g, 1]) - (i - 1)*log(theta[g, 2])
      + logfact(i- 1))
    }

    sum2[g] <- sum(summ[g,  1:(z[g] + 1)])
    L[g] <- exp(-(theta[g, 1] + theta[g, 2] + theta[g, 3]))
    * ((pow(theta[g, 1], y1[g]))/exp(logfact(y1[g])))
    * (pow(theta[g, 2], y2[g])/exp(logfact(y2[g]))) *sum2[g]
    ones[g] ~ dbern(L[g])
    log(theta[g,1]) <- mu + home + att[home_team[g]] - def[away_team[g]]
```

```
    log(theta[g,2]) <- mu + (att[home_team[g]] - att[away_team[g]])
    + (def[home_team[g]] - def[away_team[g]])
    log(theta[g, 3]) <- const
  }

  home ~ dnorm(0,0.0001)
  mu ~ dnorm(0,0.0001)
  const ~ dnorm(0,0.0001)

  for (t in 1:nteams){
    att.star[t] ~ dnorm(0, tau.att)
    def.star[t] ~ dnorm(0, tau.def)
    att[t] <- att.star[t] - mean(att.star[])
    def[t] <- def.star[t] - mean(def.star[])
  }
  tau.att ~ dgamma(0.1,0.1)
  tau.def ~ dgamma(0.1,0.1)
}", file="pro.txt",fill=TRUE)
```

References

Major Sources

[1] Gianluca Baio and Marta Blangiardo. "Bayesian hierarchical model for the prediction of football results". In: *Journal of Applied Statistics* 37.2 (2010), pp. 253–264.

[4] Leonardo Egidi. "Fitting football models and visualizing predictions with the footBayes package". In: (2022).

[5] Leonardo Egidi, Francesco Pauli, and Nicola Torelli. "Combining historical data and bookmakers' odds in modelling football scores". In: *Statistical Modelling* 18.5-6 (2018), pp. 436–459.

[8] Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[9] Lorenz A Gilch. "UEFA EURO 2020 Forecast via Nested Zero-Inflated Generalized Poisson Regression". In: *arXiv preprint arXiv:2106.05174* (2021).

[10] Dimitris Karlis and Ioannis Ntzoufras. "Analysis of sports data by using bivariate Poisson models". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3 (2003), pp. 381–393.

[11] Dimitris Karlis and Ioannis Ntzoufras. "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference". In: *IMA Journal of Management Mathematics* 20.2 (2009), pp. 133–145.

[12] Lynn Kuo and Bani Mallick. "Variable selection for regression models". In: *Sankhyā: The Indian Journal of Statistics, Series B* (1998), pp. 65–81.

[14] Robert B O'hara and Mikko J Sillanpää. "A review of Bayesian variable selection methods: what, how and which". In: (2009).

[15] Alun Owen. "Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter". In: *IMA Journal of Management Mathematics* 22.2 (2011), pp. 99–113.

[16] Martyn Plummer. "JAGS: Just another Gibbs sampler". In: (2004).

[20] Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

[23] Arnold Zellner. "Models, prior information, and Bayesian analysis". In: *Journal of Econometrics* 75.1 (1996), pp. 51–68.

Minor Sources

[2] Tony Collins. *How football began: a global history of how the world's football codes were born*. Routledge, 2018.

[6] Arpad E Elo. “The proposed uscf rating system, its development, theory, and applications”. In: *Chess Life* 22.8 (1967), pp. 242–247.

[7] J Fahey-Gilmour et al. “Multifactorial analysis of factors influencing elite Australian football match outcomes: a machine learning approach”. In: *Journal homepage: http://iacss. org/index. php? id* 18.3 (2019).

[13] Leslie Lamport. *LaTeX: a Document Preparation System*. 2nd ed. Massachusetts: Addison Wesley, 1994.

[17] Edgar Santos-Fernandez, Paul Wu, and Kerrie L Mengersen. “Bayesian statistics meets sports: a comprehensive review”. In: *Journal of Quantitative Analysis in Sports* 15.4 (2019), pp. 289–312.

[18] David J Spiegelhalter et al. “Bayesian measures of model complexity and fit”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.4 (2002), pp. 583–639.

[19] Joshua Trewin et al. “The influence of situational and environmental factors on match-running in soccer: a systematic review”. In: *Science and Medicine in Football* 1.2 (2017), pp. 183–194.

[21] John Williams and Nicola Vannucci. “English hooligans and Italian ultras sport, culture and national policy narratives”. In: *International Journal of Sport Policy and Politics* 12.1 (2020), pp. 73–89.

[22] Gheyath Mustafa Zebari et al. “Predicting Football Outcomes by Using Poisson Model: Applied to Spanish Primera División”. In: *Journal of Applied Science and Technology Trends* 2.04 (2021), pp. 105–112.