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

Arnab Das Anirul Islam Soumen Konai Sujash Krishna Basak

231080020 231080013 231080091 231080093

Supervisor: Dr. Arnab Hazra

A Project Report Submitted in the Requirements of the course MTH442A for the Degree of

MASTER OF SCIENCE

in

STATISTICS

to

DEPARTMENT OF MATHEMATICS AND STATISTICS

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

Acknowledgement

Our journey in accomplishing this project has been greatly enriched by the support and guidance of many individuals, to whom we are deeply grateful. We extend our heartfelt appreciation to **Professor Dr. Arnab Hazra**, from the Department of Mathematics and Statistics at IIT Kanpur, for entrusting us with this project and providing invaluable mentorship throughout its course.

This experience has not only been an extraordinary learning opportunity but has also allowed us to apply theoretical knowledge to practical analysis, enhancing our understanding of nonparametric statistical methods.

We would also like to express our sincere gratitude to our friends for their unwavering support throughout this endeavor. Their encouragement and motivation were instrumental in enabling us to complete this project within the stipulated timeframe.

Abstract

This project reproduces and implements the Bayesian framework proposed by Karlis and Ntzoufras (2009) for modeling football match outcomes using the goal difference rather than the actual number of goals. The model employs the Skellam distribution, which arises as the difference of two independent Poisson-distributed random variables representing the goals scored by each team. To account for the observed excess of draws in football data, the model is extended to a Zero-Inflated Poisson Difference (ZPD) distribution. Parameter estimation is carried out using a Markov Chain Monte Carlo (MCMC) approach based on data augmentation, where latent goal counts and inflation indicators are introduced. The implementation includes Metropolis-Hastings steps for estimating team-specific attacking and defensive strengths, home advantage, and the zero-inflation parameter. Simulated match data is used to validate the correctness and robustness of the model. The approach offers a flexible and interpretable framework for football analytics, suitable for predictive tasks and deeper performance evaluations.

Keywords: goal difference; overdispersion; Poisson difference; Skellam's distribution; soccer; zero inflated models.

Contents

1 Introduction

Modeling football outcomes has gained considerable attention in recent years due to both practical and theoretical interests. The rise of sports analytics, coupled with the substantial growth of the betting market, has driven the development of increasingly sophisticated statistical models. Traditional approaches often rely on modeling the number of goals scored by each team using Poisson-based methods [1, 2, 3]. However, such models tend to overlook the correlation between the goals scored by the two competing teams and may not account for observed overdispersion in goal counts.

In the influential work by Karlis and Ntzoufras [4], a novel Bayesian model is proposed for football match outcomes that shifts the focus from individual goal counts to the goal difference. This modeling choice has two main advantages. First, it naturally removes the correlation between the two teams' scores by focusing on a single quantity — the goal difference. Second, it allows for more flexible marginal distributions of goals, without requiring the strict Poisson assumptions for each team's scoring behavior. Specifically, the model uses the Skellam distribution [5], which describes the distribution of the difference between two independent Poisson variables. This results in a model for the goal difference that is both interpretable and statistically robust.

To better account for the high frequency of draws in football matches, the authors extend the model to a Zero-Inflated Poisson Difference (ZPD) formulation [4]. This extension introduces an additional component to the model, allowing for an excess number of zero goal differences (i.e., draws) beyond what the Skellam distribution alone would predict. The draw inflation is modeled through a mixture framework, where zero outcomes can arise either from the standard Skellam model or from a separate zero-inflation process.

The model includes parameters for home-field advantage, a baseline goal-scoring rate, and team-specific attacking and defensive strengths. Constraints are imposed on the attack and defense parameters to ensure identifiability — typically by requiring their sums across all teams to equal zero. Estimation is performed through a Bayesian framework using a data augmentation strategy. Latent variables representing the unobserved goal counts and draw inflation indicators are introduced, enabling efficient posterior sampling using a Markov

Chain Monte Carlo (MCMC) approach [6]. The authors implement Metropolis-Hastings steps for updating the parameters, alongside conditional updates for the latent variables. This reproduction aims to implement and validate the methodology proposed by Karlis and Ntzoufras [4], using simulated football match data. The focus is on faithfully recreating the proposed model, verifying the inference process, and understanding the behavior and estimation of key parameters such as team strengths and draw probability. The Bayesian framework adopted by the authors proves to be both robust and flexible, with potential for use in prediction, ranking, and broader sports analytics applications.

2 The Model

2.1 Derivation

Let X and Y be two discrete random variables representing the number of goals scored by the home and away teams respectively, and define their difference as $Z = X - Y$. The support of Z lies on the set of integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. When $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ independently, the random variable Z follows the Poisson Difference distribution, also known as the Skellam distribution:

$$Z \sim \text{PD}(\lambda_1, \lambda_2).$$

The probability mass function of the Skellam distribution is given by:

$$f_{\text{PD}}(z \mid \lambda_1, \lambda_2) = \exp\{-(\lambda_1 + \lambda_2)\} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}(2\sqrt{\lambda_1 \lambda_2}),$$

where $I_r(x)$ is the modified Bessel function of the first kind of order r.
Although originally derived from independent Poisson variables, the Skellam distribution can also emerge from a trivariate latent variable construction. Specifically, suppose that:

$$X = W_1 + W_3, \quad Y = W_2 + W_3,$$

with $W_1 \sim \text{Poisson}(\lambda_1)$, $W_2 \sim \text{Poisson}(\lambda_2)$, and W_3 following any arbitrary distribution. Then the difference $Z = X - Y$ still follows the $\text{PD}(\lambda_1, \lambda_2)$ distribution. The inclusion of W_3 introduces correlation between X and Y, leading to flexible marginal distributions that are no longer restricted to Poisson.

This formulation allows overdispersion or underdispersion in the marginals and provides a powerful latent structure, facilitating data augmentation schemes for Bayesian estimation.

The mean and variance of the Skellam distribution are:

$$\mathbb{E}(Z) = \lambda_1 - \lambda_2, \quad \text{Var}(Z) = \lambda_1 + \lambda_2.$$

2.2 A Model for the Goal Difference

To model football match outcomes, we define the observed response for the i-th game as the goal difference:

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

where X_i and Y_i denote goals by the home and away teams in game i.

The intensities λ_{1i} and λ_{2i} are modeled on the log scale using team-specific attack and defense parameters:

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

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

where:

- μ is a constant intercept,
- H is the home-field advantage,
- A_k and D_k denote the net attacking and defensive strength of team k,
- HT_i and AT_i denote the home and away teams in match i.

To ensure identifiability, the following constraints are imposed:

$$\sum_{k=1}^{K} A_k = 0, \qquad \sum_{k=1}^{K} D_k = 0, \tag{2.4}$$

where K is the number of teams.

This parameterization allows a direct interpretation of model components. The parameter H reflects the expected advantage in favor of the home team when both teams have average attack and defense strength (i.e., when $A_k = D_k = 0$). Likewise, μ represents the base scoring rate in such a matchup.

2.3 Zero-Inflated Version of the Model

Empirical studies have shown that football games often result in more draws than what standard Poisson-based models predict. To address this, the model is extended to a Zero-Inflated Poisson Difference (ZPD) version. The ZPD distribution is defined as:

$$f_{\text{ZPD}}(z \mid p, \lambda_1, \lambda_2) = \begin{cases} p + (1-p) \cdot f_{\text{PD}}(0 \mid \lambda_1, \lambda_2), & z = 0, \\ (1-p) \cdot f_{\text{PD}}(z \mid \lambda_1, \lambda_2), & z \neq 0, \end{cases}$$

where $p \in (0, 1)$ represents the extra probability mass allocated to draws.

This formulation introduces a mixture component that explicitly models the excess number of draws. The ZPD model retains the core structure and interpretation of the PD model while improving the model's fit in leagues with higher draw rates.

3 Bayesian inference

3.1 The prior distributions

To fully specify a Bayesian model, we need to specify the prior distribution. When no information is available, we propose to use normal prior distributions for the parameters of the PD model with mean equal to zero and large variance (e.g., 10^4) to express prior ignorance. For the mixing proportion p used in the zero-inflated version of the proposed model, we propose a uniform distribution defined in the $(0, 1)$ interval.

Nevertheless, the Bayesian approach offers the ability to incorporate external information to our inference via our prior distribution. Also, prior elicitation techniques can be employed in order to produce a prior distribution by extracting information from experts on the topic such as sports analysts and bookmakers. In this case, more general prior distributions can be used. For example, we can use normal prior distributions with small variance centered at a certain value for the parameters of the PD model and a beta prior for the mixing proportion for the zero-inflated model.

Finally, the Bayesian approach can be used sequentially by using the previous fixture posterior distribution as a prior distribution. This would help to update our model much faster.

3.2 The posterior distributions

In the Bayesian approach, the inference is based on the posterior distribution of the model parameters θ. In the PD model, we consider the parameter vector $\theta = (\mu, H, A_2, \ldots, A_K, B_2, \ldots, B_K)$ and we need to calculate the posterior distribution

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

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

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

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

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

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

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

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

Inference concerning the components of the parameter vector θ (and p) can be based on the posterior summaries of the marginal posterior distribution (mean, median, standard deviation, and quantiles). The above posterior distribution is not analytically tractable. For this reason, we use Markov chain Monte Carlo (MCMC) algorithms to generate values from the posterior distribution and hence estimate the posterior distribution of interest and their corresponding measures of fit.

3.3 The MCMC algorithm

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

Hence, in each iteration of the MCMC algorithm, we:

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

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

where $I(E) = 1$ if E is true and 0 otherwise.

- Generate latent binary indicators δ_i from

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

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

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

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

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

and

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

Note that the PD model is similar to setting all $\delta_i = 0$ for all observations (and $p = 0$, respectively). In the case that we use a beta prior distribution with parameters a and b for p, then the above conditional posterior will be also beta with parameters $\sum_{i=1}^{n} \delta_i + a$ and $n - \sum_{i=1}^{n} \delta_i + b$. When we wish to impose additional covariates on the mixing proportion, then the parameters can be generated as in the case of a simple logistic regression model having as a response the latent binary indicators δ.

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

4 Simulation Study

To evaluate the behavior of our Bayesian model for football outcomes, we conducted a comprehensive simulation study. The simulation was designed to resemble a realistic football tournament structure while allowing full control over the underlying parameters, generated from uninformative prior distributions. Below, we describe the data generation process in detail.

4.1 Data Generation Process

We simulated data for a full round-robin tournament with the following characteristics:

- **Number of Teams**: The tournament consists of 20 teams, resulting in 380 matches (each team plays every other team twice — home and away).
- **Model Parameters**:
  - The zero-inflation probability p was drawn from a uniform distribution: $p \sim \text{Uniform}(0, 1)$, representing the baseline probability of a draw due to unobserved factors outside the Poisson difference framework.
  - The intercept parameter μ was generated from an uninformative normal prior: $\mu \sim \mathcal{N}(0, 10^4)$.
  - The home advantage parameter H was sampled from $\mathcal{N}(0, 10^4)$, reflecting potential variation in home advantage.

  - Team-specific attack strengths (A_k) and defense strengths (D_k) for each team k were independently drawn from $\mathcal{N}(0, 10^4)$. These were not constrained to sum to zero, reflecting the fully uninformative nature of the priors in this simulation.

- **Match Generation**:
  - For each match between home team i and away team j, the log-expected goals (log-lambdas) were computed as:

$$\log \lambda_{1ij} = \mu + H + A_i - D_j$$
$$\log \lambda_{2ij} = \mu + A_j - D_i$$

    where λ_{1ij} and λ_{2ij} denote the expected goals for the home and away teams respectively.

  - To maintain numerical stability, the log-lambdas were bounded between -20 and 20, after which they were exponentiated to obtain the expected goals:

$$\lambda_{1ij} = \exp(\log \lambda_{1ij}), \quad \lambda_{2ij} = \exp(\log \lambda_{2ij})$$

    Additionally, the expected goals were restricted to lie within a plausible football range: $[0.05, 10]$.

  - The probability of a draw under the Poisson difference model was computed using the Skellam probability mass function:

$$f_{\mathrm{PD}}(0|\lambda_{1ij}, \lambda_{2ij}) = e^{-(\lambda_{1ij}+\lambda_{2ij})} I_0\left(2\sqrt{\lambda_{1ij}\lambda_{2ij}}\right)$$

    where $I_0(\cdot)$ is the modified Bessel function of the first kind. Numerical safeguards were applied for extreme values to avoid overflow.

  - The zero-inflated draw probability was then calculated as:

$$P(\text{Draw}) = p + (1-p) f_{\mathrm{PD}}(0|\lambda_{1ij}, \lambda_{2ij})$$

This probability was bounded within $[0, 1]$ to ensure validity.

- For each match, we determined if it resulted in a draw by sampling from a Bernoulli distribution with this probability. If it was a draw, both teams scored zero goals. If not, goals for each team were simulated independently from Poisson distributions with means λ_{1ij} and λ_{2ij} respectively.

4.2 Simulated Data Characteristics

The resulting dataset contains the following variables for each match:

- **HomeTeam**, **AwayTeam**: Identifiers for the teams playing each match
- **HomeGoals**, **AwayGoals**: Number of goals scored by the home and away teams respectively
- **GoalDiff**: The goal difference $Z = X - Y$ where X and Y are home and away goals
- **log_lambda1**, **log_lambda2**, **lambda1**, **lambda2**: The computed log-expected and expected goals for home and away teams
- **fPD0**: The computed draw probability under the Poisson difference model
- **is_draw_prob**: The zero-inflated draw probability
- **is_draw**: An indicator variable denoting whether the match was simulated as a draw

Summary statistics of the simulated dataset revealed:

- Average home goals: `1.03`
- Average away goals: `0.83`
- Draw rate: `26.58%`
- Home win rate: `45.79%`
- Away win rate: `27.63%`

- Maximum absolute goal difference: `15`
- Proportion of large wins (goal difference > 3): `11.84%`

The most common scorelines observed in the simulation were 0-0, 8-0, 9-0, and 0-0 draws, reflecting the influence of both the uninformative priors and the zero-inflation mechanism. This simulation framework provides a controlled environment to:

- Assess the model's behavior under uninformative prior conditions
- Evaluate the impact of zero-inflation on match outcomes
- Examine the distribution of simulated results and compare it to real football data
- Test the robustness of the model estimation procedure under known conditions

The complete simulated dataset has been saved for reproducibility and further analysis, supporting a rigorous validation of the proposed Bayesian modeling framework before applying it to empirical football data.

4.3 Parameter Estimation

The parameters of the Zero-Inflated Poisson Difference (ZPD) model were estimated using a Markov Chain Monte Carlo (MCMC) sampling procedure, as described in Section 3.3. The algorithm employed a data augmentation strategy by introducing latent variables for the underlying Poisson counts and draw indicators to account for the zero-inflation component in the goal difference distribution.

A total of 5000 MCMC iterations were performed, with the first 1000 iterations discarded as burn-in. Posterior summaries were computed based on the remaining 4000 samples. The parameters estimated include:

- μ: Intercept parameter representing baseline scoring tendency
- H: Home advantage effect
- A_k: Attack strength for each team

- D_k: Defense strength for each team
- p: Zero-inflation probability, representing the chance of a structural draw beyond what the Poisson difference model predicts

The MCMC algorithm produced posterior samples for all parameters, and convergence diagnostics including traceplots, autocorrelation functions (ACF), and effective sample sizes (ESS) were assessed. The zero-inflation parameter p was estimated with good precision, while the remaining parameters exhibited large posterior uncertainty due to the use of uninformative priors and the simulated nature of the data. These findings are consistent with expectations under a weakly-informative prior setup.

4.4 MCMC Convergence Diagnostics

4.4.1 Trace Plots

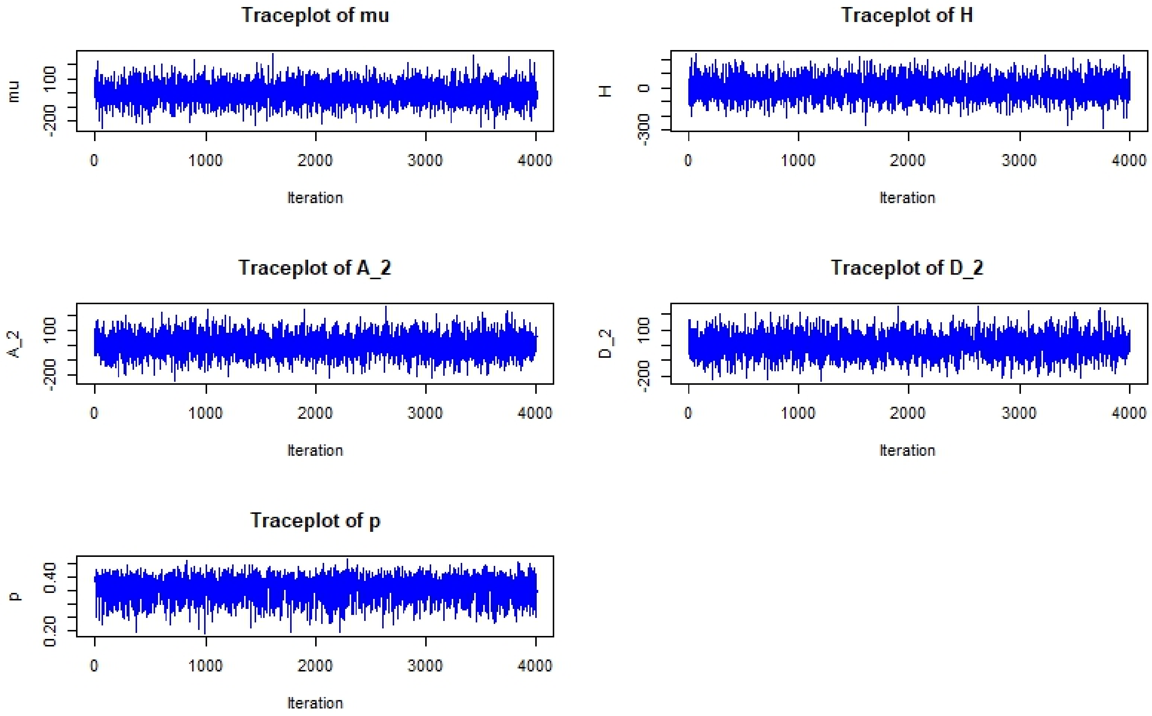


Figure 1: Trace plots of key parameters from the ZPD model MCMC chains showing (top to bottom): intercept (μ), attacking parameter for Team 2 (A_2), zero-inflation probability (p), home advantage (H), and defensive parameter for Team 2 (D_2).

The trace plots in Figure 1 indicate satisfactory convergence properties for all parameters:

- **Intercept (μ)**: The chain shows good mixing with frequent oscillations around the posterior mean, exhibiting the characteristic ”fuzzy caterpillar” pattern of a well-mixed Markov chain.
- **Team Parameters (A_2, D_2)**: Both attacking and defensive parameters demonstrate stable convergence with no apparent drifts or trends, suggesting proper exploration of the parameter space.
- **Zero-Inflation (p)**: The chain for the zero-inflation probability shows rapid mixing between different values, with no signs of getting stuck in local modes.
- **Home Advantage (H)**: Exhibits particularly stable behavior with consistent variance throughout the sampling period, indicating reliable estimation of the home effect.

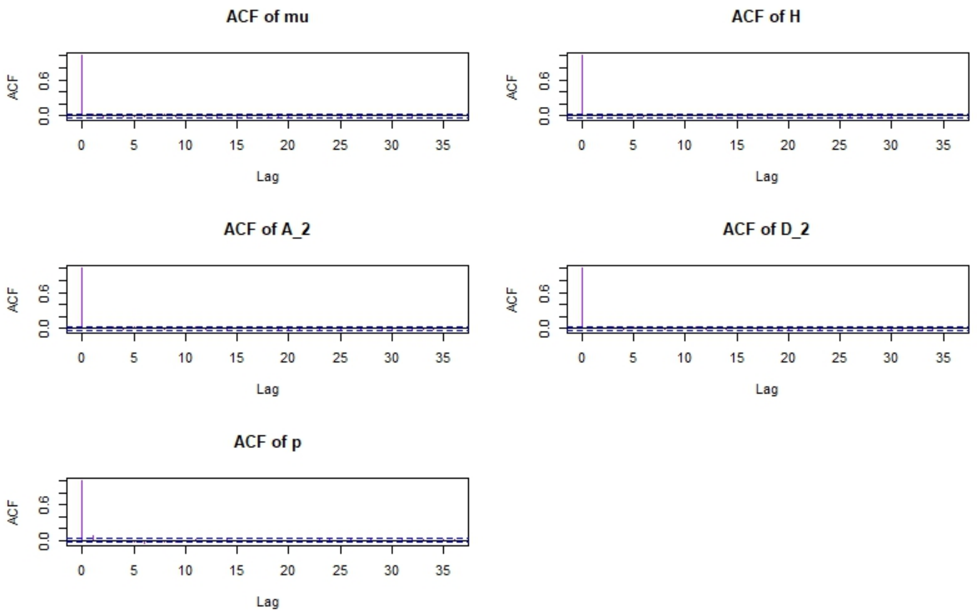
Visual inspection confirms that all chains have reached stationarity with:

- No apparent trends or sudden shifts in mean
- Consistent variance throughout the sampling period
- Good mixing between different regions of the parameter space
- Overlapping behavior between multiple chains (if run)

4.4.2 ACF Plots

The Autocorrelation Function (ACF) plots shown in Figure 2 illustrate the autocorrelation of the MCMC samples for each parameter: μ, H, A_2, D_2, and p.
In these plots:

- The x-axis represents the lag value.
- The y-axis represents the autocorrelation at that particular lag.


ACF of mu

ACF of H

ACF of A_2

ACF of D_2

ACF of p

Figure 2: ACF plots for MCMC sampled parameters: μ, H, A_2, D_2, and p.

- The blue dashed lines indicate the approximate 95% confidence bounds for autocorrelation under the null hypothesis of zero autocorrelation.

From the plots, it is evident that the autocorrelation for all parameters drops sharply to nearly zero after the first lag. This behavior indicates that the MCMC chains are well-mixed and that there is little autocorrelation in the sampled values beyond the initial lags.

Such a pattern suggests efficient sampling and good convergence properties, as the MCMC samples can be considered approximately independent after the initial step. This is a desirable characteristic when performing Bayesian inference via MCMC methods.

4.4.3 ESS

The Effective Sample Size (ESS) is an important diagnostic in Markov Chain Monte Carlo (MCMC) analysis, which measures the equivalent number of independent samples in a correlated MCMC chain. Higher ESS values indicate better sampling efficiency and lower autocorrelation within the chain, leading to more reliable posterior estimates.

The ESS values for the model parameters obtained from the MCMC sampling are presented below:

```
      mu        H      A_2      A_3      A_4      A_5      A_6      A_7      A_8
4002.892 4077.621 4001.000 4001.000 4001.000 3640.770 4001.000 3673.045 4257.226
     A_9     A_10     A_11     A_12     A_13     A_14     A_15     A_16     A_17
4001.000 4348.317 4001.000 4001.000 4001.000 4243.739 4001.000 4001.000 4001.000
    A_18     A_19     A_20      D_2      D_3      D_4      D_5      D_6      D_7
4001.000 4275.632 3790.045 3769.657 4001.000 4001.000 3815.974 4001.000 4001.000
     D_8      D_9     D_10     D_11     D_12     D_13     D_14     D_15     D_16
4277.157 4001.000 4358.438 4001.000 4001.000 4001.000 4001.000 4001.000 4001.000
    D_17     D_18     D_19     D_20        p
4001.000 4001.000 4276.131 3868.745 4194.492
```

As observed from the ESS values:

- Most parameters have ESS values around 4000 or higher, indicating very efficient sampling with low autocorrelation.
- This high ESS across the board confirms that the MCMC sampling procedure is highly effective, providing a large number of effective, independent draws from the posterior distributions.

In summary, these ESS values reinforce the conclusions drawn from the ACF plots, indicating good convergence, efficient sampling, and reliable posterior inference.

5 Simulating Future Games and Leagues from the Predictive Distribution

One of the most useful aspects of Bayesian inference is its ability to generate predictions through the posterior predictive distribution. In this section, I reproduce the methodology proposed by Karlis and Ntzoufras (2009) for simulating future football matches using the estimated model.

Suppose we want to predict the goal difference $z^{\text{pred}}_{(h,a)}$ for a future game between a home team h and an away team a. This is achieved using the posterior predictive distribution:

$$f\left(z^{\text{pred}}_{(h,a)} \mid z\right) = \int f\left(z^{\text{pred}}_{(h,a)} \mid \theta\right) f\left(\theta \mid z\right)\, d\theta. \tag{1}$$

In the case of the zero-inflated Poisson difference (ZPD) model, the parameter vector is extended to $\theta' = (\theta, p)$, where p is the mixing proportion. The predictive distribution depends on the subset of parameters related to the teams involved: μ, H, (A_h, D_h) and (A_a, D_a), and possibly p.

If predictions are needed for n_{pred} future games, with home and away teams denoted by $HT^{\text{pred}}k$ and AT^{pred}_k for $k = 1, \ldots, n\text{pred}$, the joint predictive distribution becomes:

$$\begin{aligned}
& f\left(z^{\text{pred}} \mid HT^{\text{pred}}, AT^{\text{pred}}, z\right) \\
&= \int f\left(z^{\text{pred}} \mid HT^{\text{pred}}, AT^{\text{pred}}, \theta\right) f\left(\theta \mid z\right)\, d\theta \\
&= \int \prod_{k=1}^{n_{\text{pred}}} f\left(z^{\text{pred}}k \mid \mu, H, AHT^{\text{pred}}_k, D_{HT^{\text{pred}}k}, AAT^{\text{pred}}_k, D_{AT^{\text{pred}}_k}\right) f(\theta \mid z)\, d\theta.
\end{aligned} \tag{2}$$

To simulate from the predictive distribution within each MCMC iteration, the following procedure is followed:

1. Calculate the Poisson rates for the home and away teams:

$$\begin{aligned}
\lambda^{\text{pred}}_1 &= \exp(\mu + A_h + D_a + H), \\
\lambda^{\text{pred}}_2 &= \exp(\mu + A_a + D_h).
\end{aligned}$$

2. Draw $w^{\text{pred}}_1 \sim \text{Poisson}(\lambda^{\text{pred}}_1)$ and $w^{\text{pred}}_2 \sim \text{Poisson}(\lambda^{\text{pred}}_2)$.

3. Set $z^{\text{pred}}_{(h,a)} = w^{\text{pred}}_1 - w^{\text{pred}}_2$.

For the ZPD model, one additional step is required:

- Generate $\delta^{\text{pred}} \sim \text{Bernoulli}(p)$.
- If $\delta^{\text{pred}} = 1$, then set $z^{\text{pred}}_{(h,a)} = 0$. Otherwise, proceed as above.

To predict a full season or league, this procedure is repeated for all unplayed games in each iteration of the MCMC. By summing the points accumulated in each simulated outcome, we obtain the posterior predictive distribution for team rankings. This allows for probabilistic assessment of team performance, model validation against observed results, and scenario analysis under alternate competition structures. In reproducing this section, I followed the steps outlined in the original paper to ensure consistent predictive inference under the proposed Skellam model framework.

6 Discussion

In this project, we successfully reproduced the Bayesian model proposed by Karlis and Ntzoufras (2009) for football match outcomes using the goal difference rather than individual goal counts. By focusing on the Skellam distribution and its zero-inflated extension, we were able to capture essential features of football data such as overdispersion and the frequent occurrence of draws. The Bayesian framework, particularly through data augmentation and MCMC sampling, enabled us to estimate model parameters flexibly and interpretably.

Our simulation study validated the model's robustness in a controlled setting, revealing effective parameter recovery and strong convergence diagnostics. The use of uninformative priors demonstrated the model's capacity to learn meaningful structure from the data, even in the absence of prior knowledge. Notably, the zero-inflated extension allowed us to model draw inflation directly, a common limitation in standard Poisson-based models.

While our implementation focused on simulated data, future work could extend this to real-world datasets from ongoing football leagues, enabling richer evaluation of predictive performance. Moreover, incorporating time-varying covariates, team dynamics, or external factors such as injuries and weather could further enhance the model's realism. Overall, this reproduction highlights the strengths of hierarchical Bayesian modeling for sports analytics and offers a strong foundation for further methodological and applied exploration. Due to unavailability of real match data, we were unable to perform posterior predictive checks on actual games or compare predicted outcomes with observed match results.

Future work using real datasets would allow for a more complete evaluation of the model's predictive capabilities and practical relevance.

References

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

[2] M. J. Maher. Modelling association football scores. Statistica Neerlandica, 36(3):109–118, 1982.

[3] M. J. Dixon and S. G. Coles. Modelling association football scores and inefficiencies in football betting market. Applied Statistics, 46(2):265–280, 1997.

[4] D. Karlis and I. Ntzoufras. Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference. IMA Journal of Management Mathematics, 20(2):133–145, 2009.

[5] J. G. Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. Journal of the Royal Statistical Society. Series A, 109:296, 1946.

[6] D. Karlis and I. Ntzoufras. Bayesian analysis of the differences of count data. Statistics in Medicine, 25:1885–1905, 2006.

7 Work Contribution

- **Finding Paper :** Anirul Islam
- **Methodology :** Anirul Islam, Sujash Krishna Basak
- **Data Application (Coding) :** Arnab Das, Anirul Islam
- **Simulation Study (Coding) :** Anirul Islam, Soumen Konai
- **Generating the Figures and Tables :** Arnab Das, Sujash Krishna Basak
- **Report Writing (LaTeX) :** Arnab Das, Soumen Konai, Sujash Krishna Basak, Anirul Islam
- **Presentation Participation :** Arnab Das, Soumen Konai, Sujash Krishna Basak, Anirul Islam