University of Missouri

STAT 8640

Analysis of Indian Super League Data Using Bivariate Poisson Model and Double Poisson Model

Souvik Bag

14421334

Introduction

Data has become ubiquitous, serving as a crucial element for making informed decisions in contemporary times. The realm of sports analytics is rapidly expanding. In this endeavor, I aim to construct a hierarchical Bayesian model to simulate the goals scored by football (soccer) teams in the Indian Super League.

The Poisson distribution is commonly employed by researchers to model the goal-scoring patterns of both home and away teams in soccer. While it's possible to use two independent Poisson models for the number of goals, there could be a correlation between the goals scored at home and those scored away. An approach to account for this dependency is to use bivariate poisson model.

In this project I implemented both double poisson model (two independent poisson models) and bivariate poisson model on the Indian Super League dataset from the $2021-22$ season and compared their posterior statistics.

The project unfolds in the following sequence. Initially, we delve into a discussion about our data, followed by the specification of our model. We establish our mean structure and define priors and hyperpriors. Subsequently, we assess model convergence, determine the appropriate number of iterations, and conduct other diagnostic checks.

Following this, our focus shifts to a detailed examination of the posterior results.

1 Data description

I used the 2021-22 season of Indian Super League data which comprises a total of 11 teams competing in a home-and-away format. Following the regular season, there are playoffs involving the top 4 teams, engaging in two-legged semifinals and a single final. Consequently, there are a total of 115 matches in the season.

In Figure (1), the aggregate goals scored and conceded by each team are depicted. Notably, Jamshedpur FC, Hyderabad FC, and Mohunbagan AC emerged with the highest goal-scoring records. Given the pivotal role of goals in determining match outcomes, these three teams occupied the top positions in the league. Conversely, Odisha FC and Northeast FC faced challenges, conceding the highest number of goals, which reflected in their lower standings in the league.

An insightful observation is that teams exhibiting strong offensive capabilities, reflected in higher goal-scoring figures, tend to have robust attacking strategies. Conversely, teams with

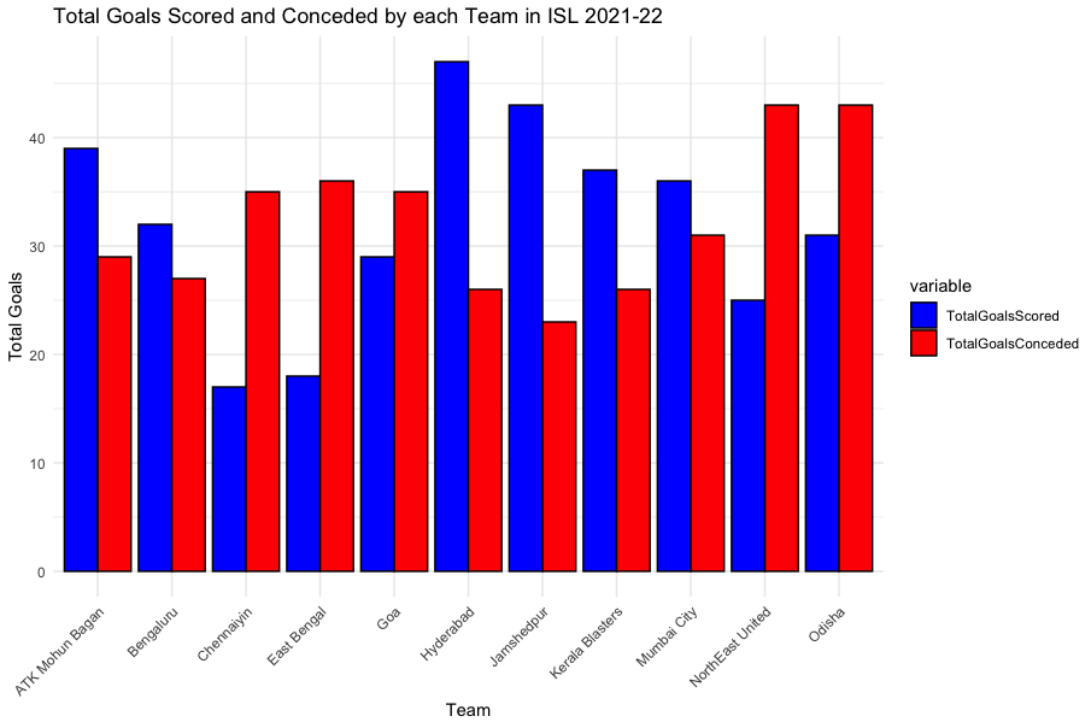

Figure 1: Total number of goals scored and conceded by each team

fewer goals conceded demonstrate effective defensive strategies.

2 Model description

We employ a modeling framework for predicting the number of goals scored by both home and away teams, utilizing two independent Poisson models alongside a bivariate Poisson model. The basis for our approach to modeling the mean parameter draws inspiration from the work of Baio & Blangiardo([1]).

There are total $T = 11$ teams and $G = 115$ games. Let y_{g1} and y_{g2} be the number of goals scored by home and away team in $g - th$ match respectively where $g = 1, 2, \ldots, 115$.

2.1 Double poisson model

The observed vector of number of goals $\mathbf{y} = (y_{g1}, y_{g2})$ is modeled as independent Poisson such that:

$$y_{gj}|\theta_{gj} \sim Poisson(\theta_{gj})$$

Where $\theta = (\theta_{g1}, \theta_{g2})$ represent the scoring intensity of the home and away team respectively in $g - th$ game.

2.2 Bivariate poisson model

It is more practical to model the number of goals scored by the two teams jointly rather independently. This motivates us to define a bivariate poisson model.

$$(y_{g1}, y_{g2}) \sim BP(\theta_{g1}, \theta_{g2}, \theta_{g3})$$

Where θ_{g3} denotes the covariance parameter. One natural interpretation of the parameters is that θ_{g1} and θ_{g2} reflect the net scoring ability of the home and away team and θ_{g3} reflect the gaming conditions such as speed of the game, weather situations etc.

2.3 Structure of mean and covarince function

Now as a hierarchical structure we model these parameters as used by Karlis & Ntzoufras ([2]). In the paper they used frequentist approach to estimate the parameters for English Premier League data. We will use hierarchical bayesian model using Rstan.

Our mean number of goals scored is modeled as the sum of a home effect, the attacking ability and defensive prowess. To make the θ_{gi}'s positive we model the natural log of θ_{gi} as follows,

$$log(\theta_{g1}) = home + att_{h(g)} + def_{a(g)}$$
$$log(\theta_{g2}) = att_{a(g)} + def_{h(g)}$$

That is, the scoring intensity of home team $log(\theta_{g1})$ depends on a "home" factor, the attack strength of home team and defensive capability of the away team.

Similarly, the scoring intensity of away team $log(\theta_{g2})$ depends only on the attack strength of away team and defensive capability of the home team.

In the bivariate poisson model The covariance parameter θ_{g3} is modeled as,

$$log(\theta_{g3}) = \beta^{con} + \gamma_1 \beta_{h(g)} + \gamma_2 \beta_{a(g)}$$

where, β^{con} is a constant parameter and $\gamma_1 \beta_{h(g)}$ and $\gamma_2 \beta_{a(g)}$ depends on home and away team respectively. also γ_1 and γ_2 are indicator variables taking values 0 or 1 depending on what kind of covariance model we choose. For example $\gamma_1 = 0$ and $\gamma_2 = 0$ reflect a constant covariance structure of the model.

The nested indexes $h(g), a(g) = 1, 2, \ldots, T = 11$ identify the team playing at home and away respectively in $g - th$ game.

2.4 Priors and hyper-priors

Now we need to define the priors and hyper-priors.

The hyperpriors are defined as,

$$home \sim N(0, 0.1)$$

where 0.1 is the standard deviation of the Gaussian distribution.

Conversely, for each $t = 1, 2, \ldots, T$ the team specific effects for example the attack ability of team t is drawn from a Gaussiaan centered at the population attack ability of all the team. Similarly the defensive ability of team t is drawn from a Gaussiaan centered at the population defensive ability of all the team. They are modeled as:

$$att_t \sim N(\mu_{att}, \tau_{att})$$
$$def_t \sim N(\mu_{def}, \tau_{def})$$

Finally, the hyperpriors of the attack and defense effects are modeled independently using wide priors. We use wide and flat priors because it will let the data drive the posterior and each team will have different posteriors based on their number of goals scored and conceded. I played with different distributions but Gaussians camee out to be most superior.

$$\mu_{att} \sim N(0, 0.1)$$
$$\mu_{def} \sim N(0, 0.1)$$
$$\tau_{att} \sim normal(0, 1)$$
$$\tau_{def} \sim normal(0, 1)$$

One challenge is to find suitable prior for the covariance parameter. We choose a vague prior as well for $\beta^{con} \sim N(0, 0.1)$. Subsequently for different covariance structures (either it is a constant covariance or covariance depends on home team only etc) we define our other priors accordingly.

2.5 Model diagnostics

We implemented the model in Stan ([3]). We used 6000 iterations where 3000 of those were burn-in period. We played with different numbers of iterations and numbers of chains ranging from 2000 to 20,000 and 4 to 10 respectively . We found that 6000 was good enough with 5

chains.

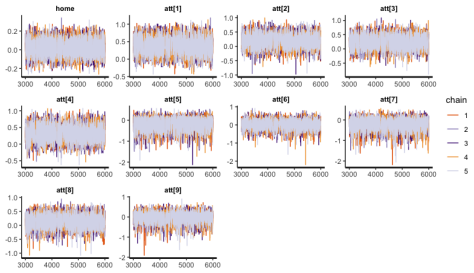

Figure 2: Traceplot for first 10 parameters in bivariate poisson modeling

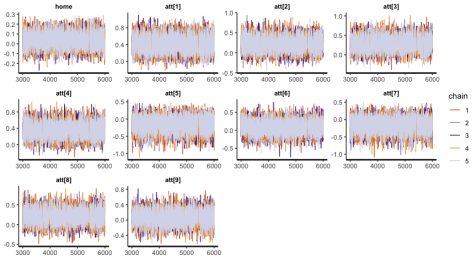

Figure 3: Traceplot for first 10 parameters in double poisson modeling

The traceplots (2, 3) shows that our posterior samples converged. Another way to check convergence is to see the Gelman-Rubin statistic. It compares the variance within chains to the variance between chains. All the statistic values are close to 1 which suggests that the samples converged (See R output).

2.6 Posterior estimates

Here I will present some of the posterior estimates from both the models.

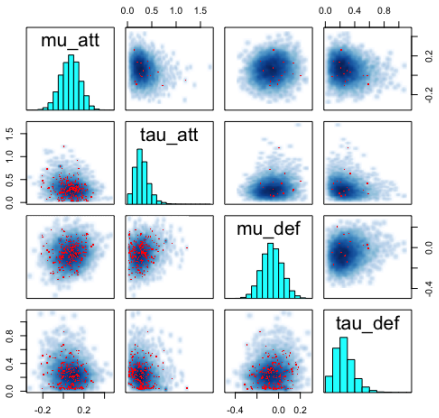

Figure 4: Posterior estimates of 4 parameters in bivariate poisson modeling

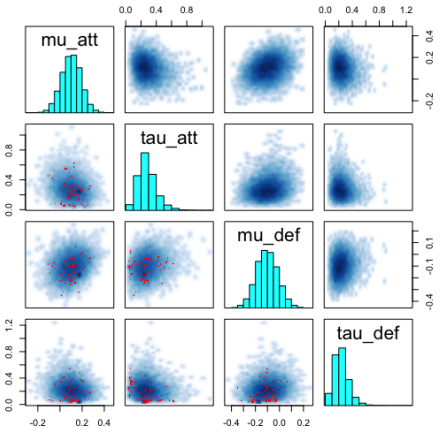

Figure 5: Posterior estimates of 4 parameters in double poisson modeling

The traceplots (4, 5) shows that there are slight differences between the posterior estimates between the two models.

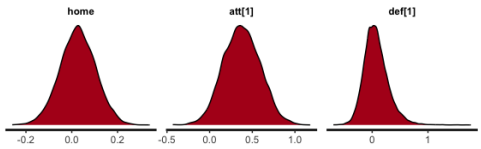

Figure 6: Posterior density of 3 parameters in bivariate poisson modeling

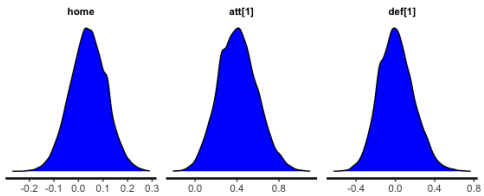

Figure 7: Posterior density of 3 parameters in double poisson modeling

The density plots (6, 7) shows that there are differences between the posterior estimates of parameters between the two models.

2.7 Posterior prediction and comparison between two models

In this section we will compare the posterior predictive estimates of two models. First we kept aside data of 5 matches for the prediction.

Figure 1 shows that there are slight difference in the predictive distribution of the two models.

Now we will compare the attack and defensive capabilities of 11 teams using two different

Figure 8: Red : Bivariate poisson, Blue : Double poisson

models and plot them.

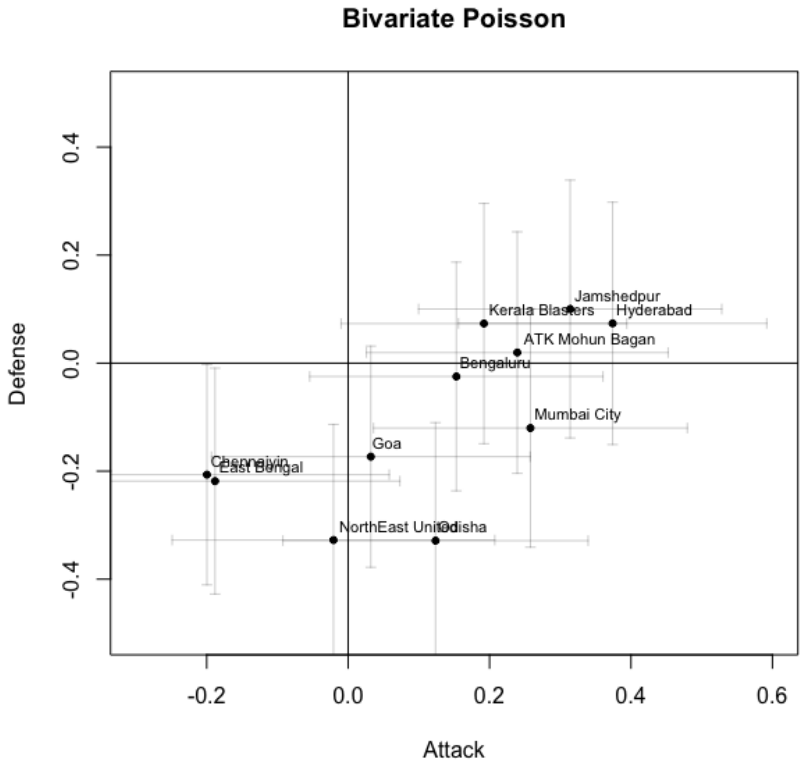

Figure 9: Attack and Defence capability of 11 ISL teams using Bivariate poisson model

Figures (9) and (10) offer intriguing insights. Teams exhibiting positive attributes in both defense and attack found themselves at the top of the league table, scoring more goals while conceding fewer. Conversely, teams such as East Bengal and Chennaiyan, manifesting negative attributes, struggled and occupied the lower ranks in the league.

Both models demonstrate proficiency in capturing these nuanced aspects of team performance. However, a nuanced comparison reveals distinctions between the two. In my assessment, both

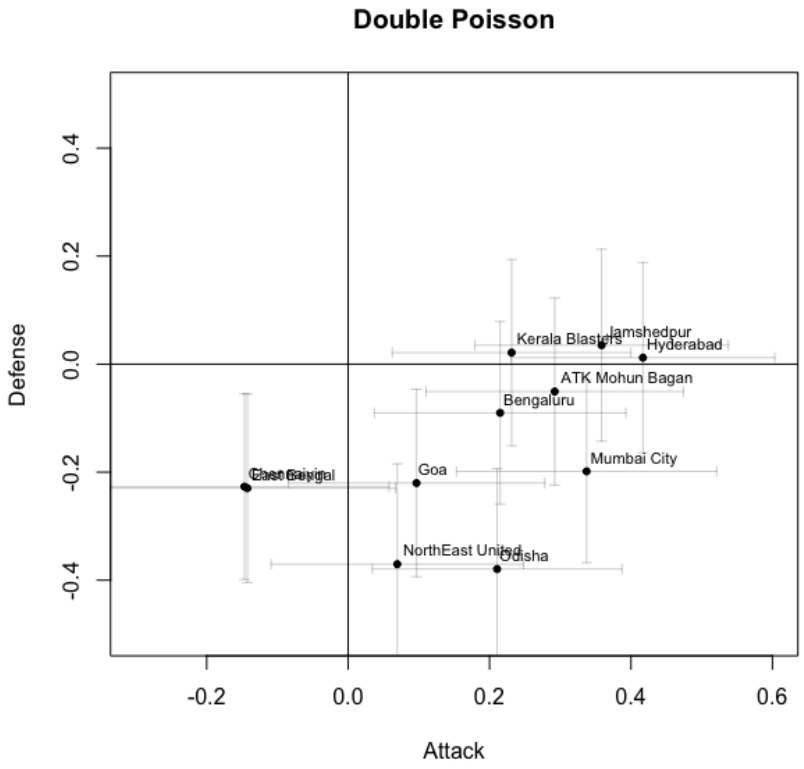

Figure 10: Attack and Defence capability of 11 ISL teams using Double poisson model

models exhibit similar overall performance, yet the bivariate Poisson model excels in depicting the defensive capabilities of the teams.

To further evaluate their effectiveness, I have juxtaposed the actual season-end results, attached for reference(11) alongside Figures (9) and (10). This comparison aims to enrich the understanding of how well the models align with the observed outcomes.

The alignment between the season-end standings and the visual representations in Figure (9) and Figure (10) is striking. The top-performing teams, namely Jamshedpur, Hyderabad, Mohunbagan, and Kerala, identified in Figure (11), are consistently portrayed in a favorable light in both of our defense-attack plots.

Furthermore, the defensive vulnerabilities of teams like East Bengal, Northeast, and Odisha, evident from their conceding the highest number of goals, are faithfully reflected in our figures (9) and (10). This concordance between observed outcomes and model predictions underscores the models' ability to capture and represent the nuances of team performance accurately.

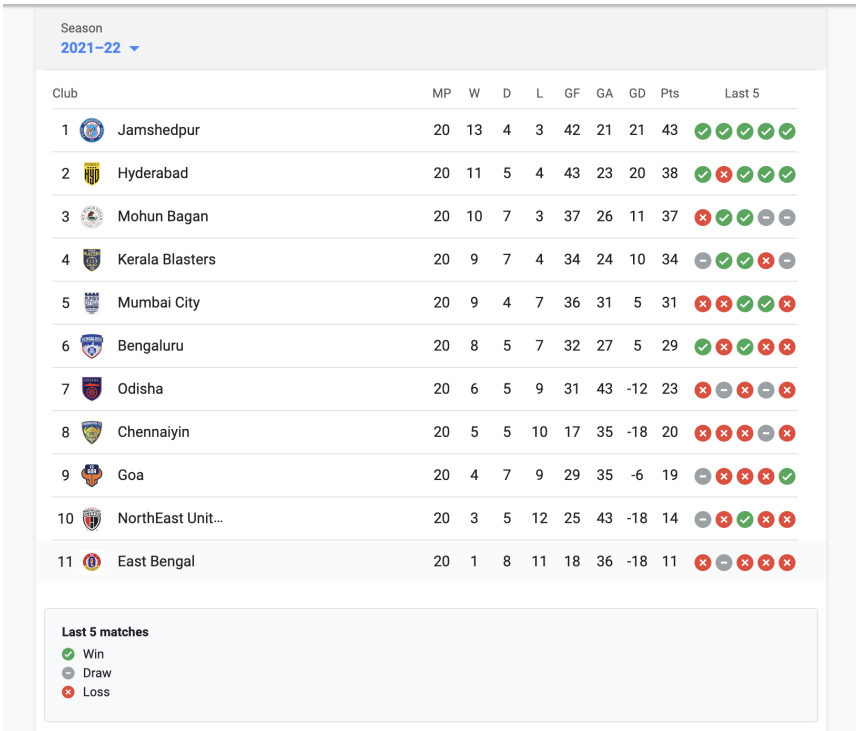

Figure 11: Attack and Defence capability of 11 ISL teams using Bivariate poisson model

3 Limitations and future work

There are several limitations of our study.

- **Data Limitations**: The analysis is constrained by the relatively small dataset from the Indian Super League, comprising only 11 teams and 115 matches. A more extensive dataset, particularly from larger leagues, would provide a broader context for assessing the discernible differences between the two models.

- **Covariance Structure in Bivariate Poisson Model**: The covariance structure employed in the bivariate Poisson model is contingent on a constant term, coupled with the influence of home and away teams. An interesting avenue for exploration involves evaluating how both models perform when relying solely on constant terms, thereby shedding light on the significance of these factors.

- **Exploration of Priors**: The current implementation utilizes normal priors; however, experimenting with more intricate priors could yield valuable insights. By employing more complex prior distributions, the models' sensitivity and responsiveness to various influences can be thoroughly examined, offering a deeper understanding of their behavior.

References

[1] G. Baio and M. Blangiardo, "Bayesian hierarchical model for the prediction of football results," *Journal of Applied Statistics*, vol. 37, no. 2, pp. 253–264, 2010.

[2] D. Karlis and I. Ntzoufras, "Analysis of sports data by using bivariate poisson models," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 3, pp. 381–393, 2003.

[3] J. Guo, J. Gabry, B. Goodrich, and S. Weber, "Package 'rstan'," *URL https://cran. r—project. org/web/packages/rstan/(2020 x y)*, 2020.

[4] OpenAI. ChatGPT: A large-scale generative language model. `https://www.openai.com/research/chatgpt`.

[5] Github. Initial pass at a bivariate Poisson model in Stan. `https://gist.github.com/mbjoseph/100a41d73901764a00a7`.