Bayesian analysis for football prediction

Francesco Cellitti

2023-01-16

Index

Introduction

In football, statistics are an important tool when we want to study a match and probabilities for one team to win a game (betting) and also for descriptive analysis in order to study the behavior of a single player during the game. Another type of statistics applied in football is about the prediction of the result of a single match. In order to make prediction about the goals scored by a team, we are going to use three different models:

- Double Poisson (independent components)
- Bivariate Poisson
- Zero inflated bivariate Poisson

Section 1

Descriptive analysis

Data

The main idea is estimating for each match of a football season some parameters in order to predict the number of goals scored. So at the end we will have for each model a matrix with rows equal to the number of games played in a season and the columns related to the goals scored in a match by two teams. For each match we want to estimate the parameters that we're going to use to make inference on the vector $(x_i, y_i) \; \forall \; i = 1...380$ where x_i are the goals scored by the home team of match i and y_i the goals scored by the away team in match i.

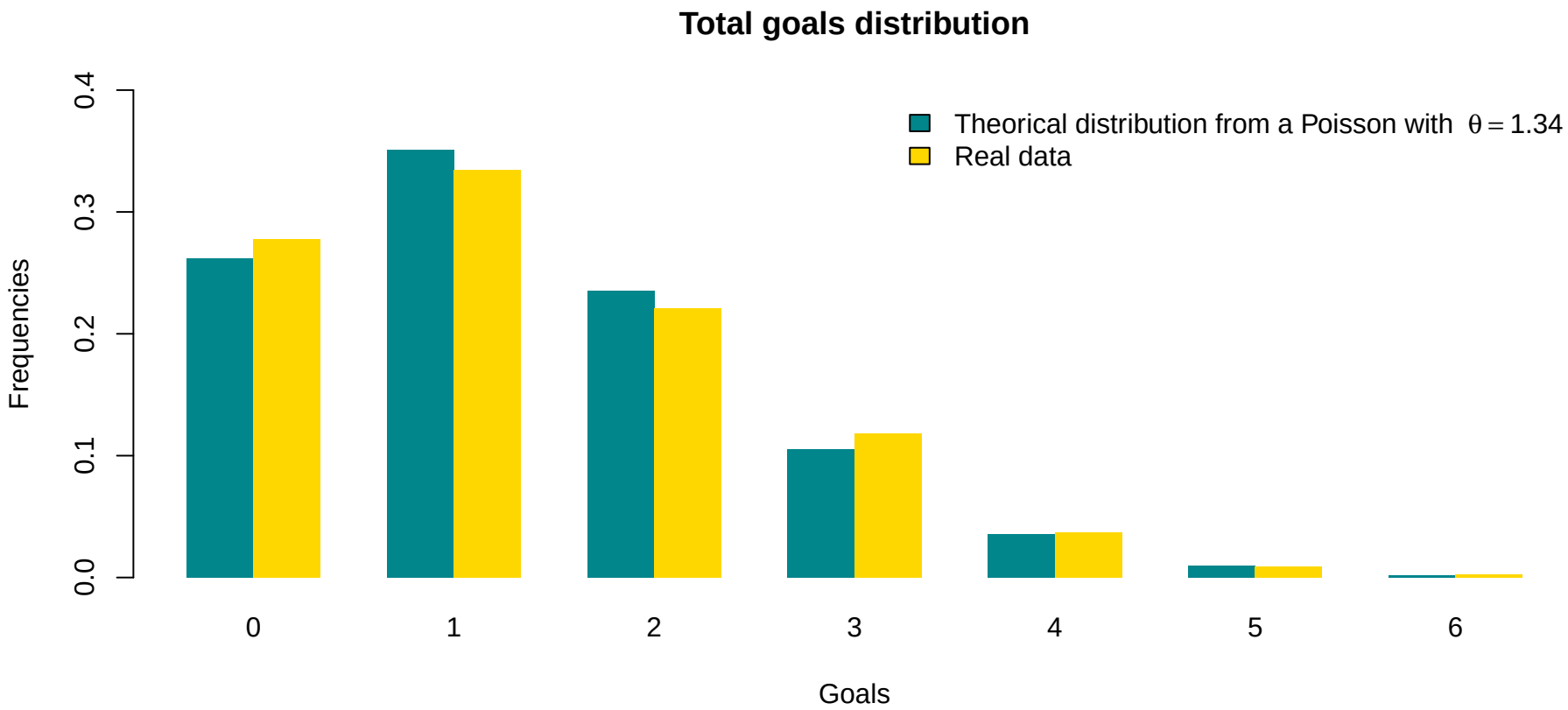
Data

In this project we will use a dataset from the Serie A 2018-2019 season with 20 teams (supplied by Datahub). For each week we have 10 games and so 380 different matches on the season. Each row has the name of two teams playing each other in a specific match and the goals scored in that game (FTHG for home and FTAG for away team).

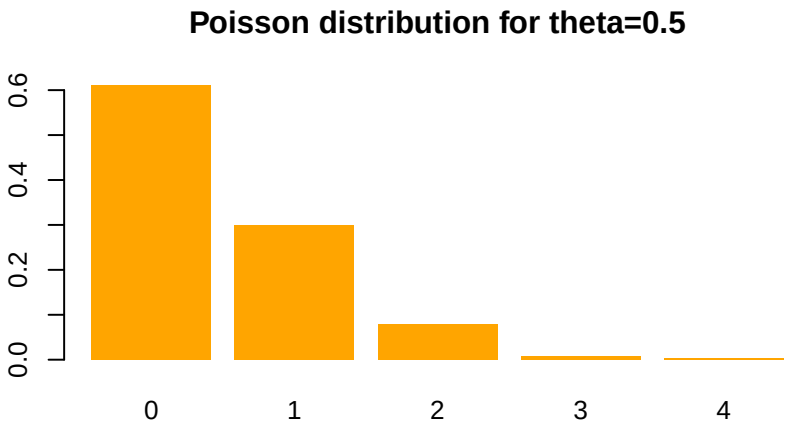
```
## # A tibble: 6 x 4
##   HomeTeam AwayTeam  FTHG  FTAG
##   <chr>    <chr>    <dbl> <dbl>
## 1 Chievo   Juventus     2     3
## 2 Lazio    Napoli       1     2
## 3 Bologna  Spal         0     1
## 4 Empoli   Cagliari     2     0
## 5 Parma    Udinese      2     2
## 6 Sassuolo Inter        1     0
```

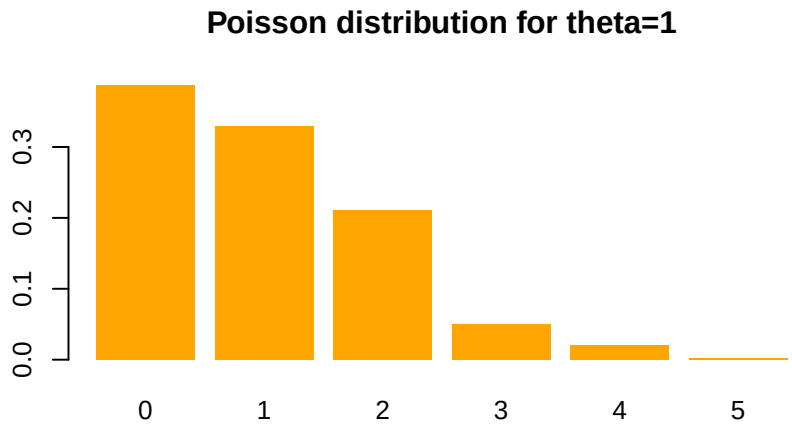
Total goals distribution

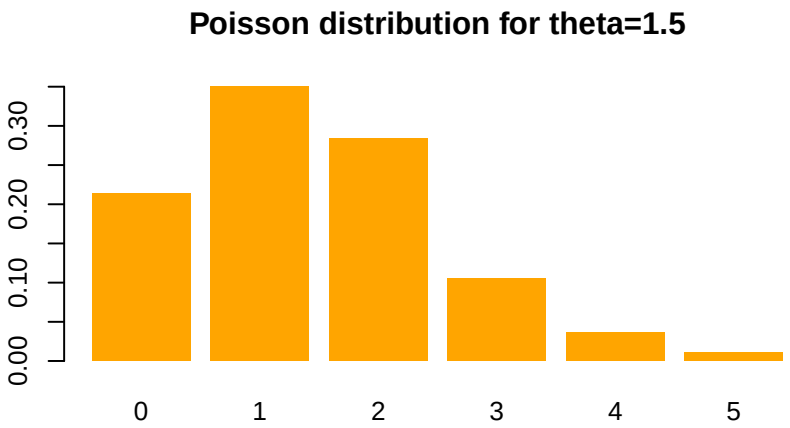
In this championship the distribution of all the goals scored during the tournament is the following. It looks like a Poisson distribution. Moreover we add an empirical distribution built estimating $\hat{\theta}_{MLE}$ and generating from a Poisson distribution with this parameter.

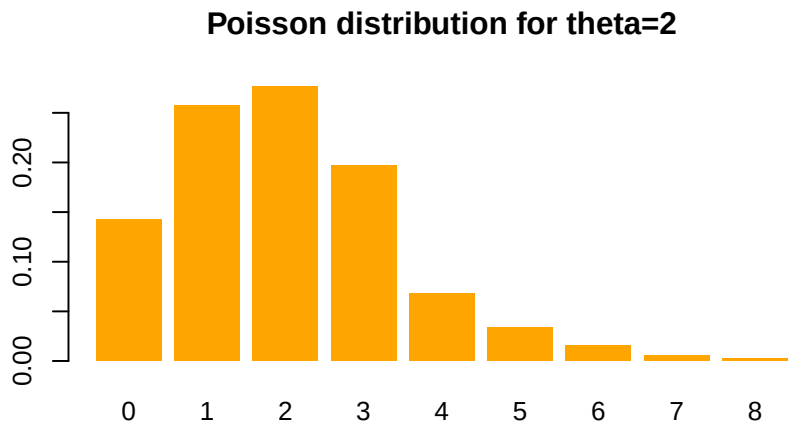


Poisson distributions



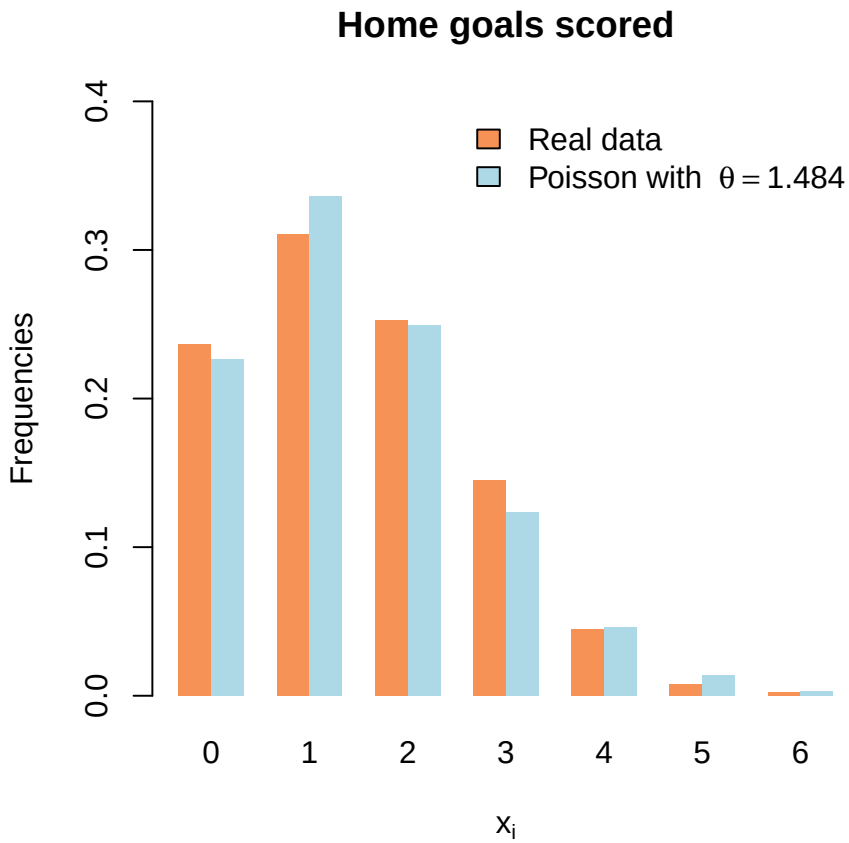


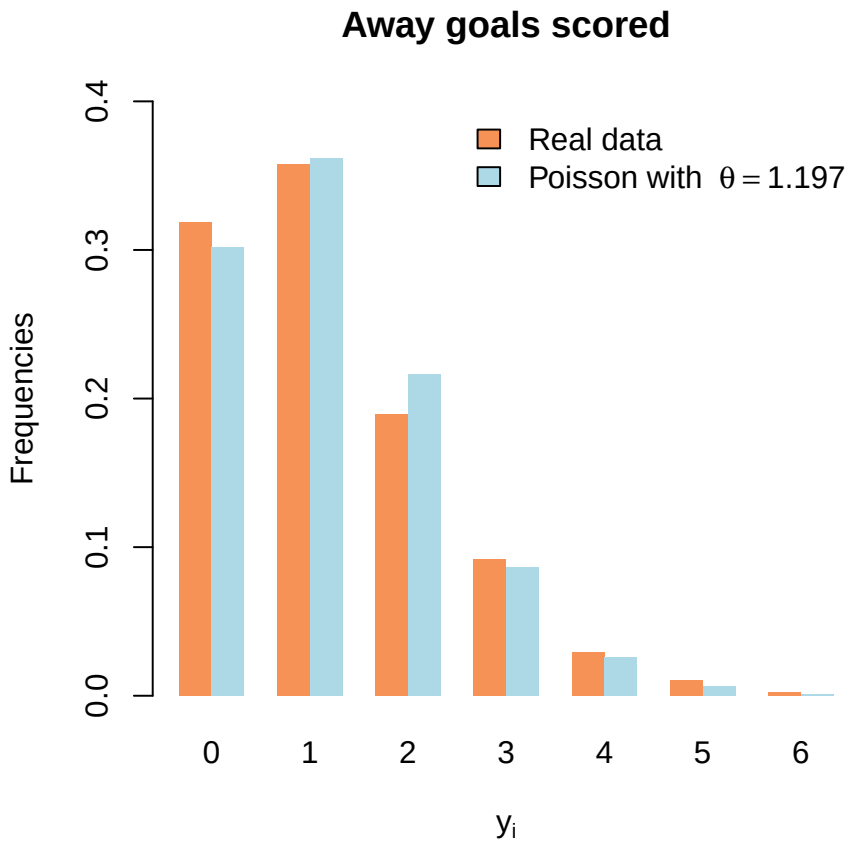




Goals analysis by home and away teams

We all know that a team playing at home in general would score more goals, how can we see from the barplot below.





Goals analysis by home and away

	Observed		Theorical		Observed cumulative	
Goals	Home	Away	Home	Away	Home	Away
0	0.237	0.318	0.223	0.303	0.237	0.318
1	0.311	0.358	0.335	0.357	0.547	0.676
2	0.253	0.189	0.252	0.219	0.800	0.866
3	0.145	0.092	0.125	0.085	0.945	0.958
4	0.045	0.029	0.046	0.025	0.989	0.987
5	0.009	0.011	0.014	0.006	0.997	0.997
6	0.003	0.003	0.0039	0.0014	1	1

Table 1: Relative frequencies of goals scored

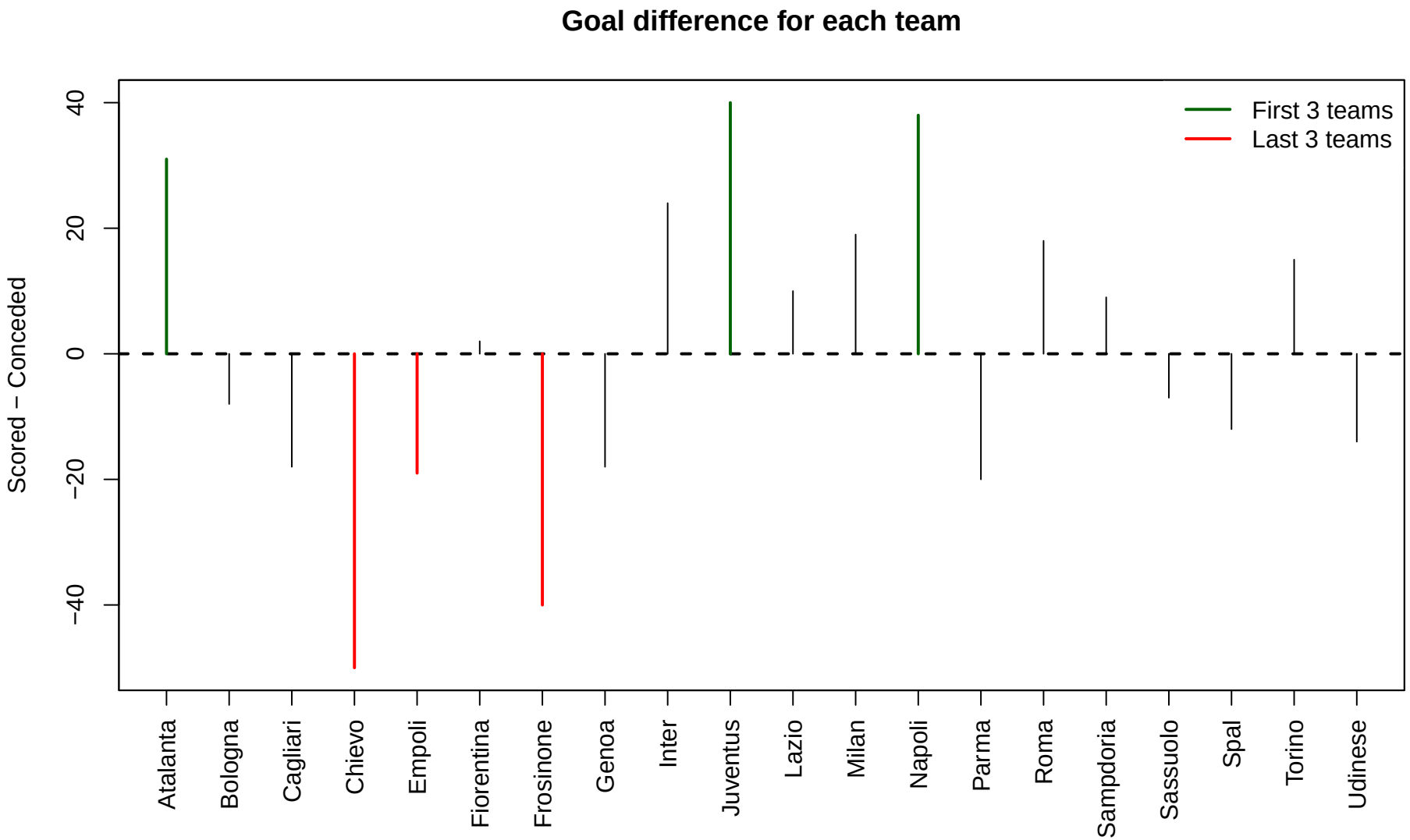
Goals analysis by team

In a football championship there are for sure teams that are stronger than others, able to score more goals or concede less. Is not sure that the team that score more goals will finish the championship in the first positions. There is a large positive correlation between the points of each team with their goals scored and also a negative one between points and conceded goals.

```
## [1] "Correlation between points & goals scored = 0.89"
```

```
## [1] "Correlation between points & goals conceded = -0.91"
```

Goals analysis by team



Section 2

Bayesian analysis

The distribution of home and away goals

In order to obtain the result of a match we have to use a bivariate distribution for the vector (x_i, y_i). In the previous section we understand that the distribution of goals is described by a Poisson distribution with a θ parameter. For (x_i, y_i) we can use a bivariate distribution in two different ways, with dependence or conditionally independent.

The Double Poisson distribution

In the literature, we can study the behavior of a vector representing goals scored in this way, so defining in case of conditionally independence:

$$X_i|Y_i, \theta_{hi} \sim Poisson(\theta_{hi})$$

$$Y_i|X_i, \theta_{ai} \sim Poisson(\theta_{ai})$$

Firstly we study the two results with conditionally independence between variables. We can identify that the value of x_i and y_i are depending only on θ_{hi} and θ_{ai} parameter that I want to estimate. The next step is find a model to estimate the 760 parameters using only the goals scored by each team.

The structure of the Poisson model

We have understood from slides 12-13 that there is a high correlation between goals scored and points made as normal as it is. We can use the information about goals to compute for each team some characteristic about attack and defense in order to use it in the model. In fact:

$$log(\theta_{hi}) = home + att_{ht_{(i)}} + def_{at_{(i)}}$$

$$log(\theta_{ai}) = att_{at_{(i)}} + def_{ht_{(i)}}$$

where *hi* and *ai* represent respectively the home and away attack and defense parameters for the game i.

The structure of the Poisson model

The bigger the θ_{hi} parameter is, the greater is attack effect of home team or less is the defense attack of away team and otherwise for the other estimate. In literature, when we have to use a Poisson model, we usually write it as a log-linear model. There is also a parameter for the home team about the effect that the squad has playing the game in front of their fans.

Choice of log-linear model

The choice of an exponential model is due to the possibility to have for the θ_{hi}, θ_{ai} values for sure greater than 0. In fact for the Poisson distribution we need the parameter non negative, so we exploit the exponential transformation.

Prior distributions

The main characteristics for estimating 760 θ are derived from the information about data. In inferential framework, we consider a hierarchical model in order to see how we can arrive at the estimation of the $380 \cdot 2$ θ. In this case we have to make assumption about the effects of each team. We write that:

$$att_t \sim Normal(\mu_{att}, \tau_{att})$$

$$def_t \sim Normal(\mu_{def}, \tau_{def})$$

$$home \sim Normal(0, 0.001)$$

Constraints about the model

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

$$\sum_{t=1}^{20} att_t = 0$$

$$\sum_{t=1}^{20} def_t = 0$$

Hyper-parameters distributions

$$\begin{cases} \mu_{att} \sim Normal(0, 0.001) \\ \tau_{att} \sim InvGamma(0.1, 0.1) \\ \mu_{def} \sim Normal(0, 0.001) \\ \tau_{def} \sim InvGamma(0.1, 0.1) \end{cases}$$

Hierarchical model DAG

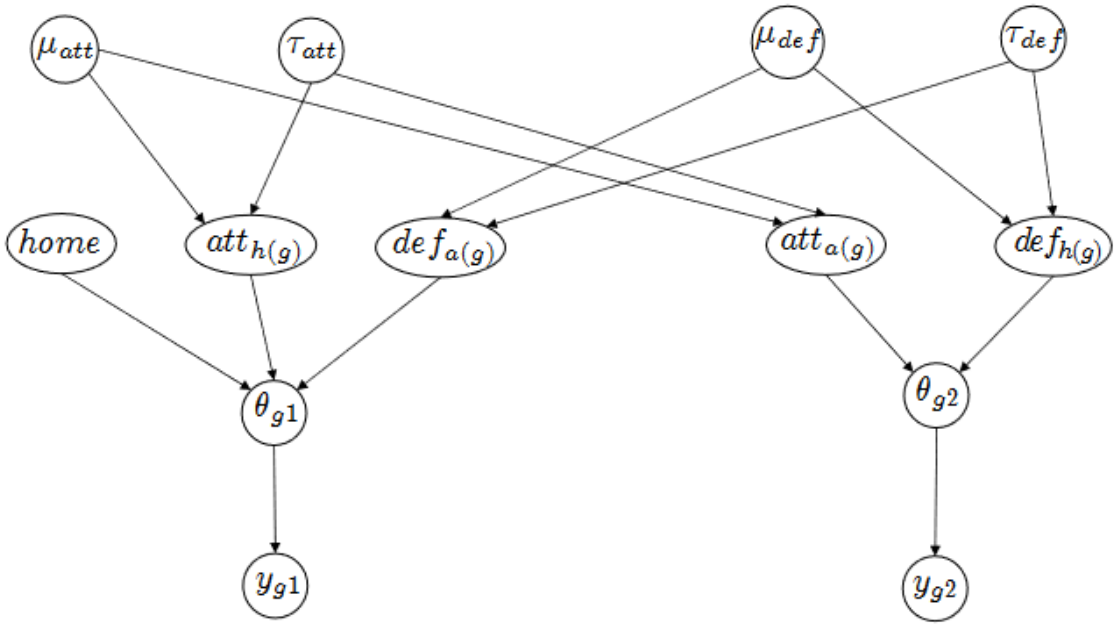


Figure 1: DAG of double Poisson model

Bayesian conjugate analysis

The starting point for making bayesian inference is to derive the posterior distribution using the fact that the joint distribution is equal to the product of the prior distribution times the likelihood.

$$j(\underline{x}, \underline{y}, \underline{att}, \underline{def}, home) = \pi(\underline{att}, \underline{def}, home) \cdot L(\underline{x}, \underline{y} \mid \underline{att}, \underline{def}, home)$$

where $\underline{x}$ and $\underline{y}$ are the home and away goals vectors. The hyperparameters are $(\mu_{att}, \tau_{att}, \mu_{def}, \tau_{def})$.

Bayesian conjugate analysis

I know that from this equation I can obtain the marginal posterior density of the parameters of interests:

$$\pi(\underline{att}|\underline{x}, \underline{y}) =$$

$$\int_{home} \int_{\underline{def}} \int_{\mu_{att}} \int_{\tau_{att}} \pi(\underline{att}, \underline{def}, home) \cdot L(\underline{x}, \underline{y}|home, \underline{att}, \underline{def})\, dhome\, d\underline{def}$$

From the DAG, we understand that att_t parameter depends only on the hyper-parameters μ_{att}, τ_{att}.

Bayesian conjugate analysis

I will estimate the other densities for def_t and *home* in a similar way:

$$\pi(\underline{def}|\underline{x},\underline{y}) =$$

$$\int_{\underline{att}}\int_{h}\int_{\mu_{def}}\int_{\tau_{def}} \pi(\underline{att},\underline{def},home)\cdot\prod_{i=1}^{380} f(x_i,y_i|home,\underline{att},\underline{def})\; d\underline{att}\; dhome$$

$$\pi(home|\underline{x},\underline{y}) =$$

$$\int_{\underline{att}}\int_{\underline{def}}\int_{\mu_{att}}\int_{\tau_{att}}\int_{\mu_{def}}\int_{\tau_{def}} \pi(\underline{att},\underline{def},home)\cdot\prod_{i=1}^{380} f(x_i,y_i|home,\underline{att},\underline{def})\; d\underline{att}$$

Bayesian conjugate analysis

The prior distribution is:

$$\pi(att, def, home) = \pi(home) \cdot \pi(\mu_{att}, \tau_{att}) \cdot \pi(\mu_{def}, \tau_{def})$$

In fact, from the DAG I can see that there is independence between the 3 parameters of interest. Moreover the hyperparameters distribution are independent each other, so:

$$\pi(\mu_{att}, \tau_{att}) = \pi(\mu_{att}) \cdot \pi(\tau_{att})$$

and:

$$\pi(\mu_{def}, \tau_{def}) = \pi(\mu_{def}) \cdot \pi(\tau_{def})$$

The bivariate Poisson model

The main idea of using the bivariate Poisson model is that, in accordance with some literature (see Ntzoufras & Karlis), the result of a football match (so the number of goals of the two teams which play against), has a sort of relationship which we have to measure. For this reason they considered a model for the score of a match using a bivariate Poisson model, with probability mass function defined as:

$$BP(x_i, y_i|\theta_{hi}, \theta_{ai}, \theta_{3i}) = e^{-(\theta_{hi}+\theta_{ai}+\theta_{3i})} \frac{\theta_{hi}^{x_i}}{x_i!} \frac{\theta_{ai}^{y_i}}{y_i!} \sum_{i=1}^{min(x_i,y_i)} \binom{x_i}{i} \binom{y_i}{i} i! \left(\frac{\theta_{3i}}{\theta_{hi}\theta_{ai}}\right)^i$$

The bivariate Poisson model

In this way θ_1 and θ_2 are the parameters in relationship with home goal and away goal while the third parameter is a measure of dependence of the random variable. If $\theta_3 = 0$, then the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions.

$$X_i|Y_i, \theta_{hi}, \theta_{3i} \sim Poisson(\theta_{hi} + \theta_{3i})$$

$$Y_i|X_i, \theta_{ai}, \theta_{3i} \sim Poisson(\theta_{ai} + \theta_{3i})$$

The bivariate Poisson model

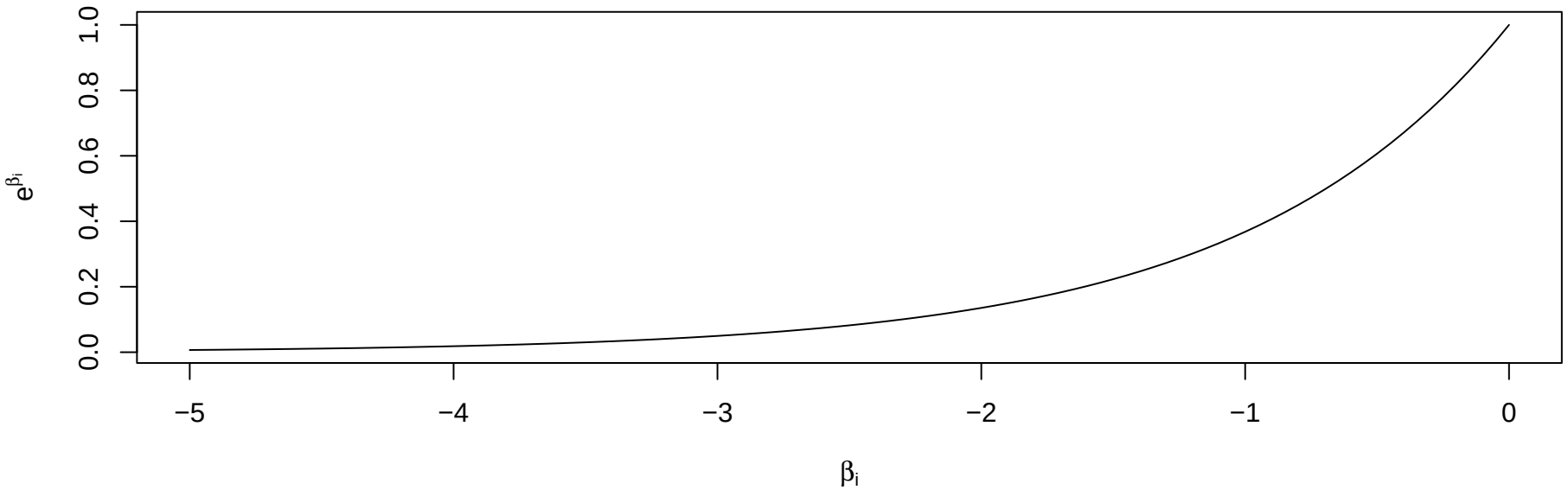
The structure of the model is the same as the double Poisson structure, with the only difference of adding a dependence term in the model. If there were some random influences in the match that affects x & y, I'm able to define the third equation of model:

$$log(\theta_{3i}) = \beta_i$$

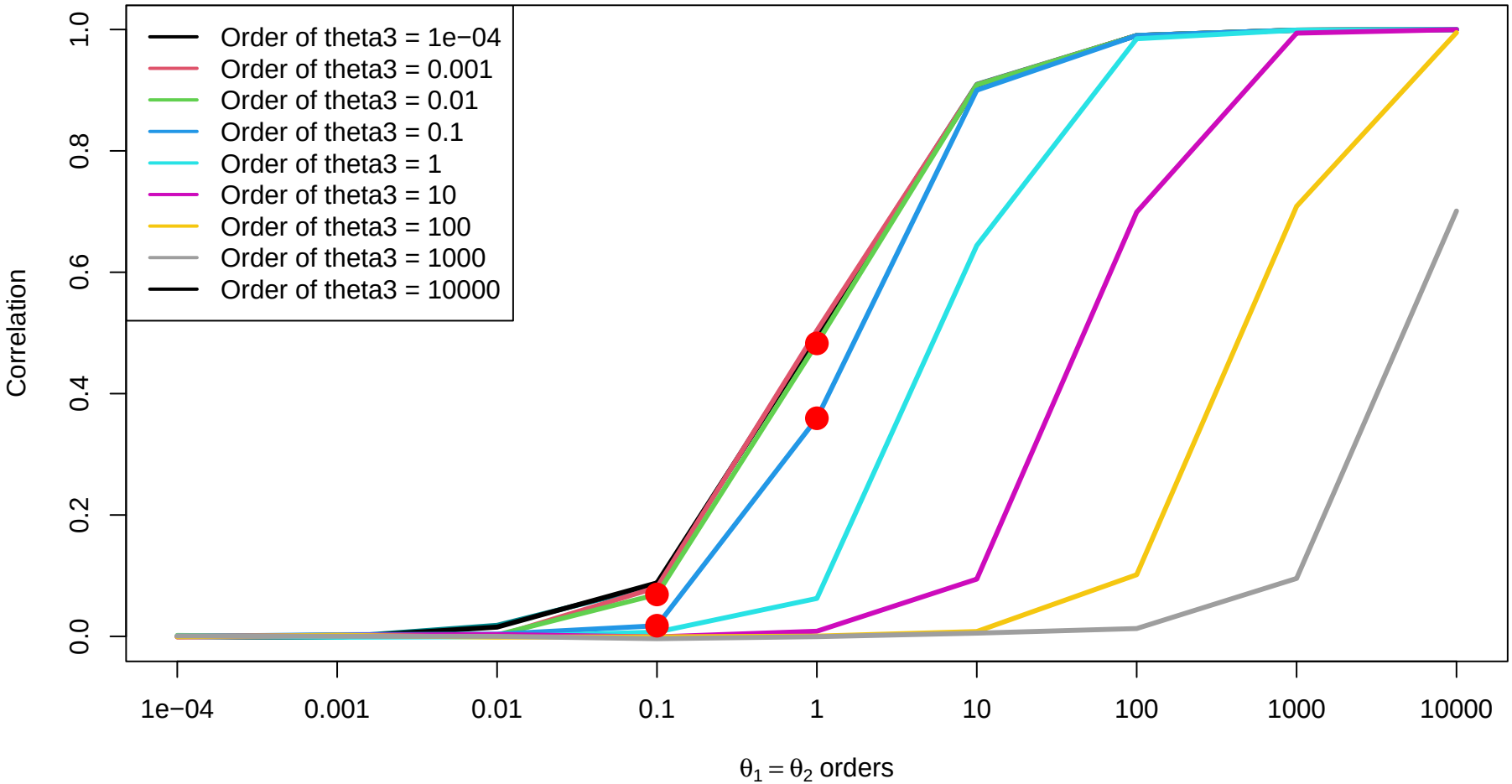
The bivariate Poisson model

We understand that this correlation must be a positive and small number in order to have the updated θ greater than 0 and not too large for having huge influence on data.

$$\theta_{3i} \in [0, 0.4] \implies \theta_{3i} = e^{\beta_i}$$



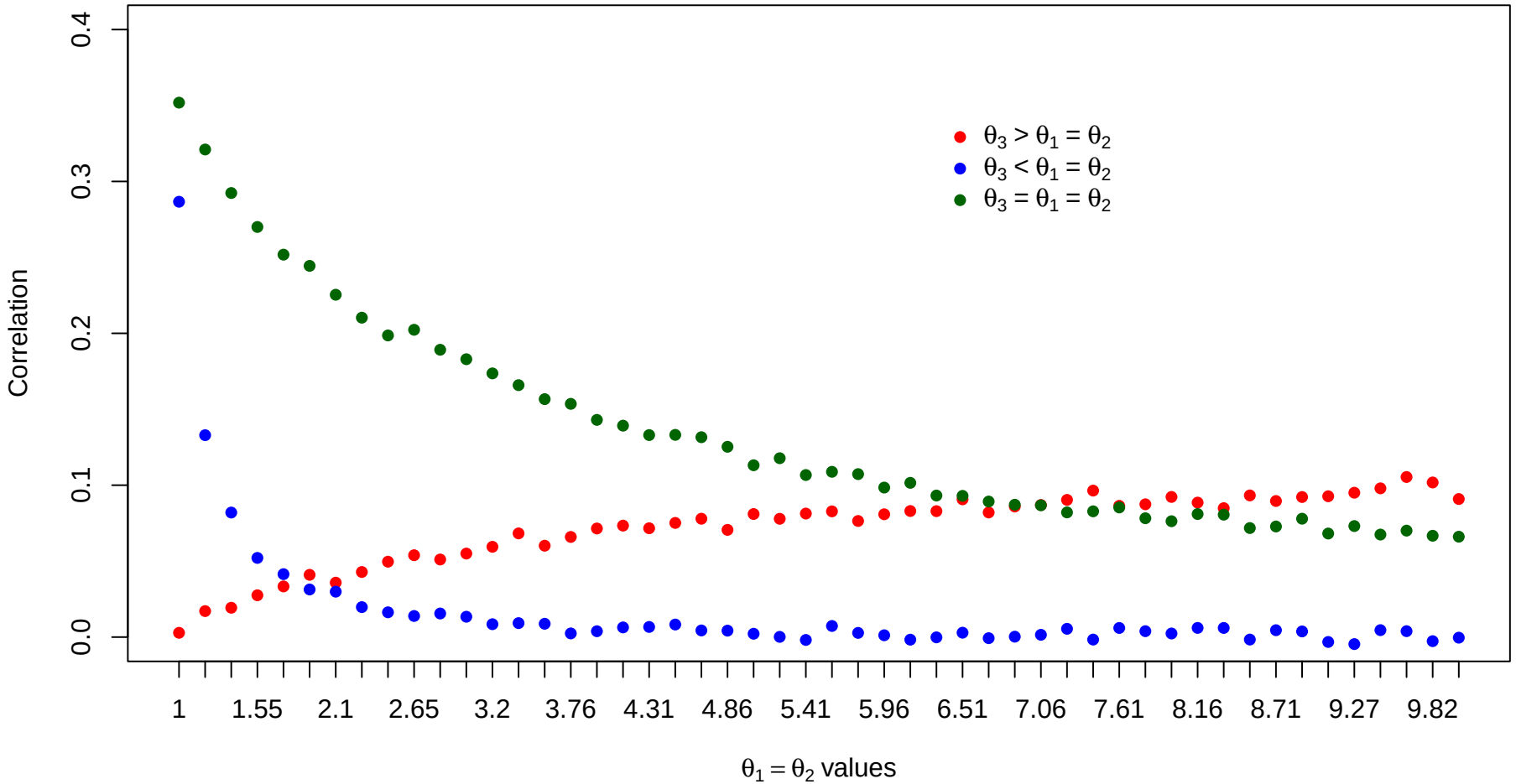
Empirical proof about correlation in the bivariate Poisson



Empirical proof about correlation in the bivariate Poisson

In the previous slide we can observe that changing the order of possible values of θ simulations, the correlation moves in that way. Our interest is related to the relation between order 1 and 0.1 where we can see a correlation less than 0.4 on average. The greater is the difference, the lower is the correlation while with values generated from an i.i.d distribution the correlation is similar when the other values are growing up.

Empirical proof about correlation in the bivariate Poisson



Zero inflated bivariate Poisson

As we have seen in the previous slides, a Poisson model with parameter in intervals from 1 to 2, will estimate a low probability to verify a zero occurrence. In this way, we can define a new kind of model's implementation using a zero-inflated model. This model is a mixture model where:

$$\begin{cases} p + (1-p) \cdot BP(x_i, y_i | \theta_1, \theta_2, \theta_3) & (x_i, y_i) = (0,0) \\ (1-p) \cdot BP(x_i, y_i | \theta_1, \theta_2, \theta_3) & (x_i, y_i) \neq (0,0) \end{cases}$$

A bivariate zero-inflated model can be constructed by increasing the probability of $(x_i, y_i) = (0,0)$ the event and decreasing the other joint probabilities.

Zero inflated bivariate Poisson

In this case I will estimate $\theta_{hi}, \theta_{ai}, \theta_{3i}$ and the correspondent results for all the teams, so (380x3 estimates), the attack and defense effects (20x2) and moreover the parameter p for the Zero-Inflated defined in its starting point as:

$$p \sim Bernoulli(\psi)$$

w here $\psi \sim Unif(0, 1)$ is the probability in the real data to have a 0-0 draw during the entire season. All the model's implementation are described due to R2Jags package in Rstudio.

MCMC

After knowing the posterior distribution of these parameters (attack and defense for each team, so 20 x 2) and parameter about the home effect due to a MCMC, we want to use them in order to estimate the θ for all matches and provide a prediction about the result of every match. At the end, after some iterations, We can see the convergence of the distribution. The number of iterations of the chain is 10.000. We report the number of iterations to keep (9.000) and the number to discard (burn-in = 1.000).

Comparison between models and checking diagnostics

In all the three models that I proposed in the previous slides, for understanding which model is better for estimating the number of goals in a season is useful to study a criterion for comparisons between that. For this comparison we use the DIC (Deviance Information Criteria):

$$DIC = D_{\hat{\theta}}(\underline{x}, \underline{y}) + 2p_D$$

Lower values of the penalized likelihood criteria are better.

Comparison between models

$$DIC_{doublePoisson} = 2243.448$$

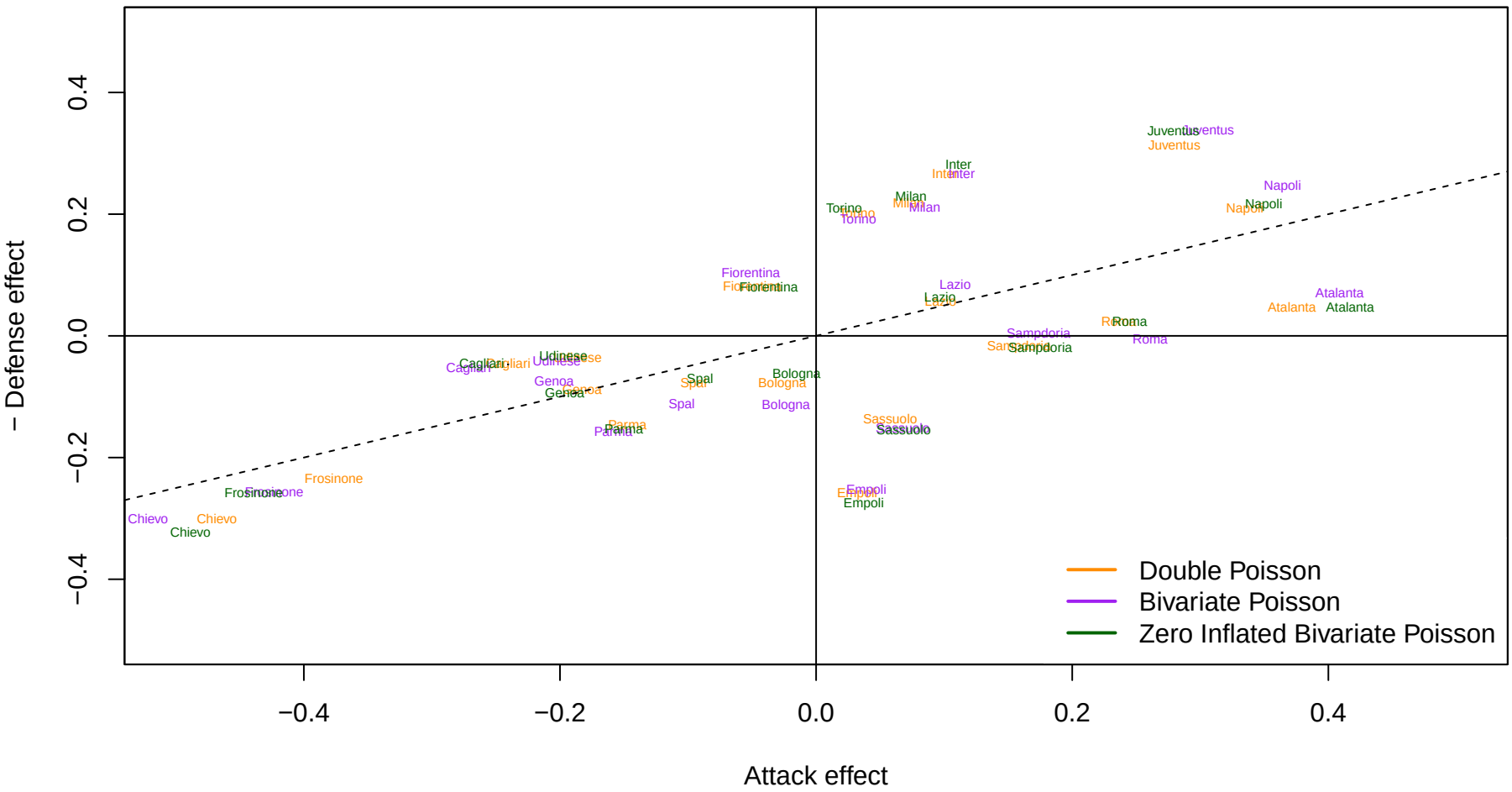
$$DIC_{bivariatePoisson} = 2209.953$$

$$DIC_{0InflatedPoisson} = 2262.35$$

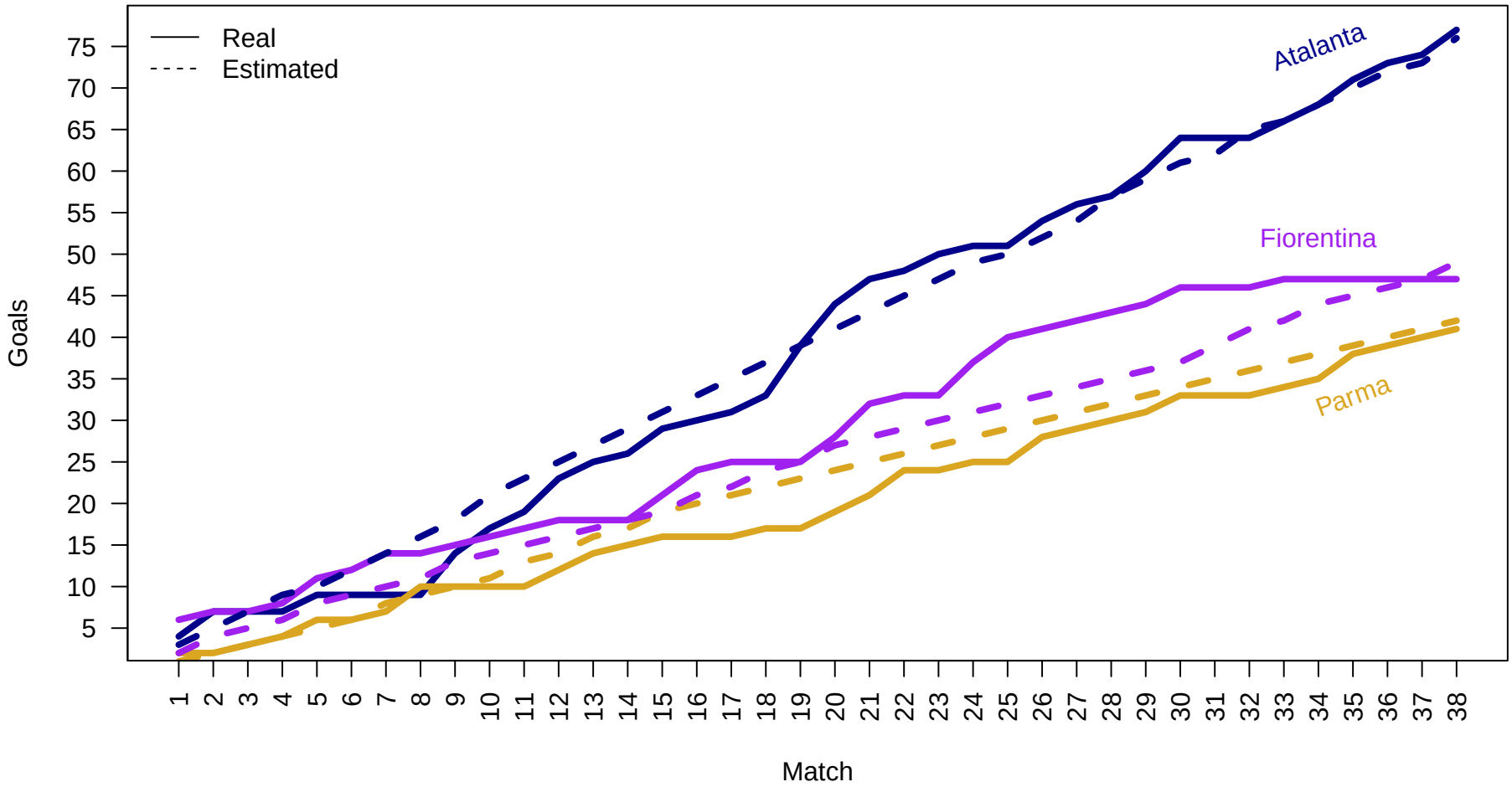
Section 3

Final results

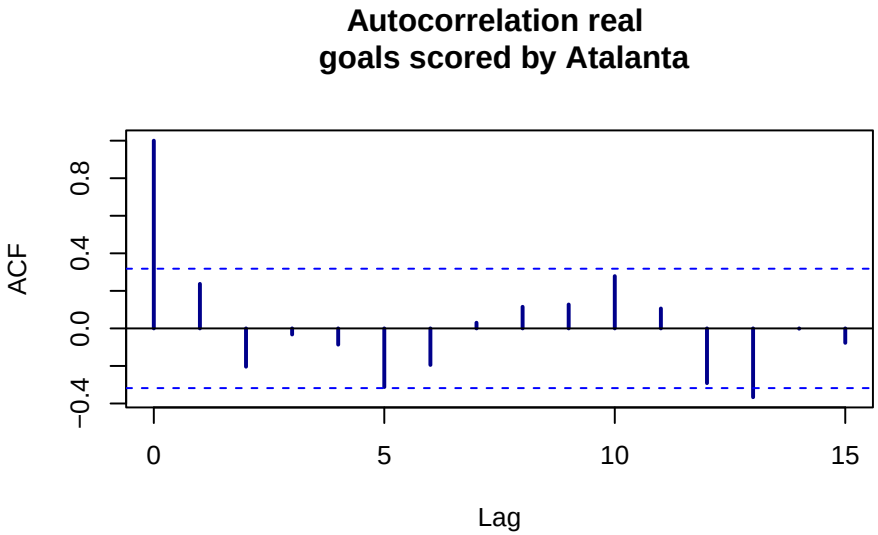
Comparison between attack and defense

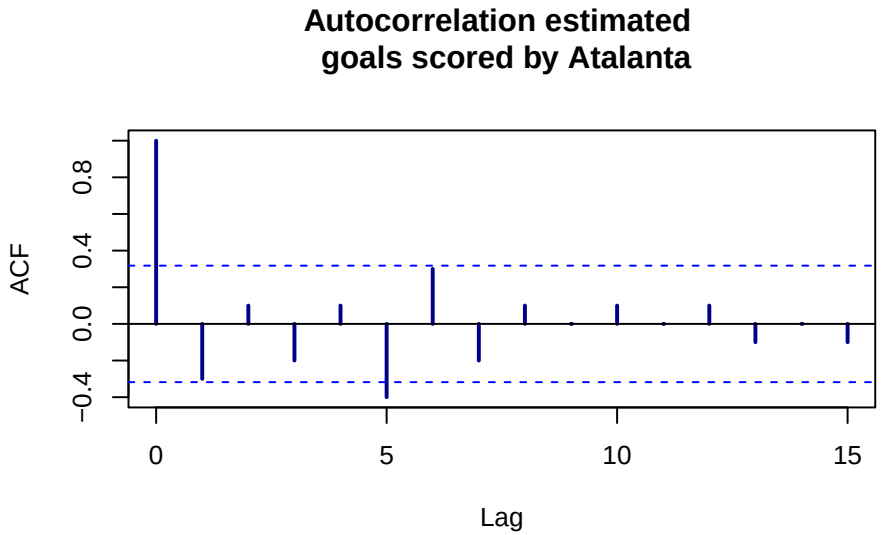


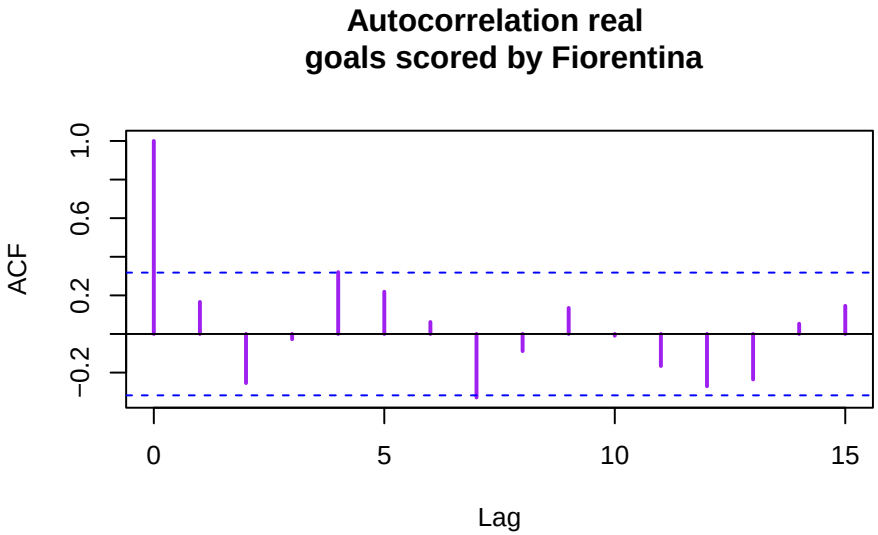
Real goals vs predicted one

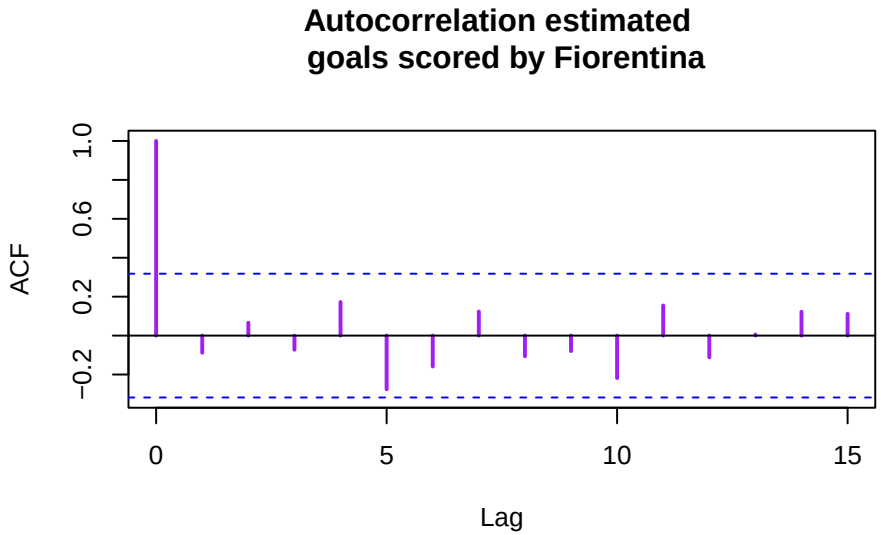


ACF


Autocorrelation real
goals scored by Atalanta
ACF
0.8
0.4
0.0
−0.4
0
5
10
15
Lag


Autocorrelation estimated
goals scored by Atalanta
ACF
0.8
0.4
0.0
−0.4
0
5
10
15
Lag


Autocorrelation real
goals scored by Fiorentina
ACF
1.0
0.6
0.2
−0.2
0
5
10
15
Lag


Autocorrelation estimated
goals scored by Fiorentina
ACF
1.0
0.6
0.2
−0.2
0
5
10
15
Lag

Real rank vs predicted

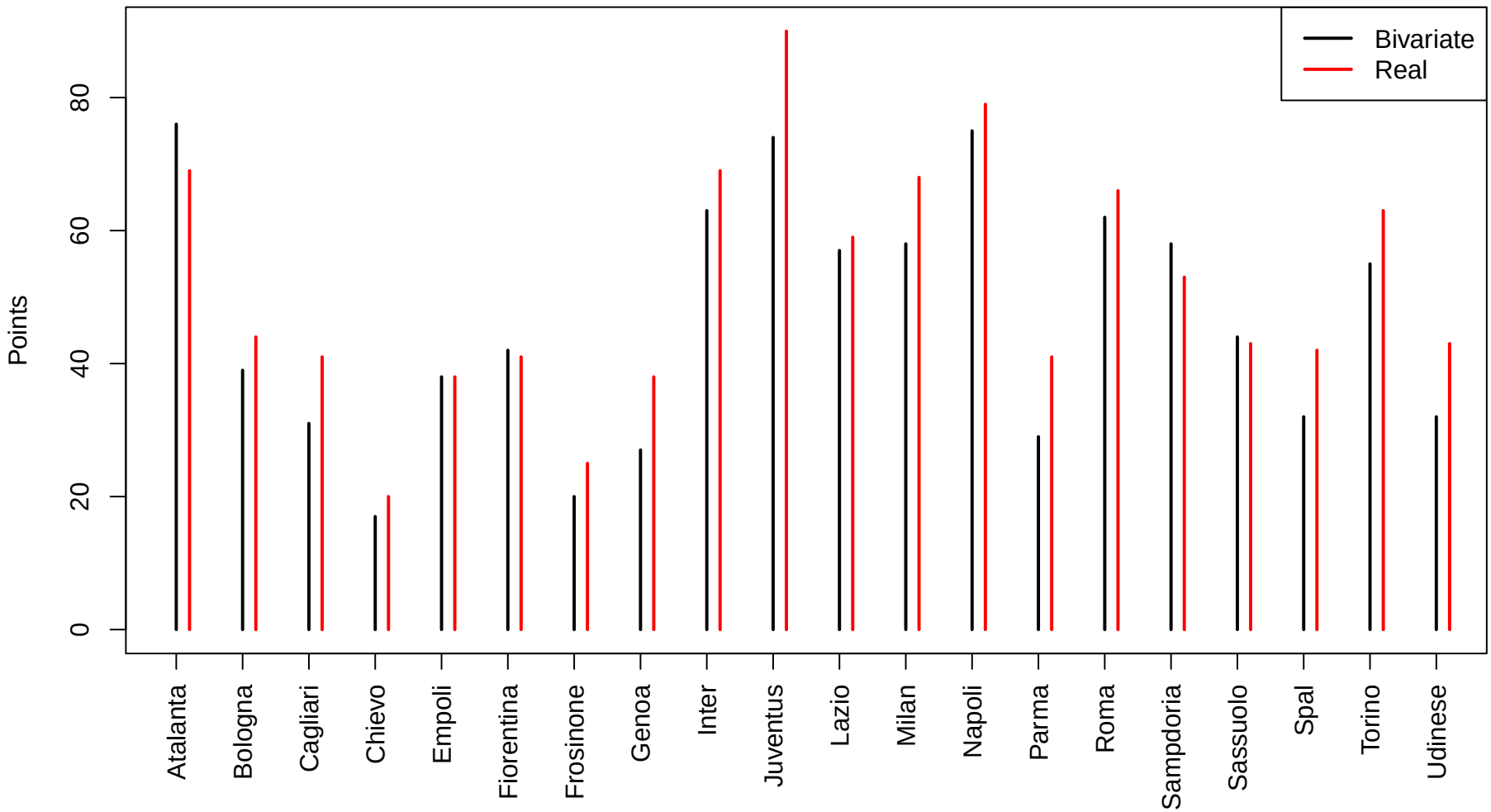
Team	Points	GS	GC
Juventus	90	70	30
Napoli	79	74	36
Atalanta	69	77	46
Inter	69	57	33
Milan	68	55	36
Roma	66	66	48
Torino	63	52	37
Lazio	59	56	46
Sampdoria	53	60	51
Bologna	44	48	56
Sassuolo	43	53	60
Udinese	43	39	53
Spal	42	44	56
Parma	41	41	61
Cagliari	41	36	54
Fiorentina	41	47	45
Genoa	38	39	57
Empoli	38	51	70
Frosinone	25	29	69
Chievo	17	25	75

Table 2: Real rank

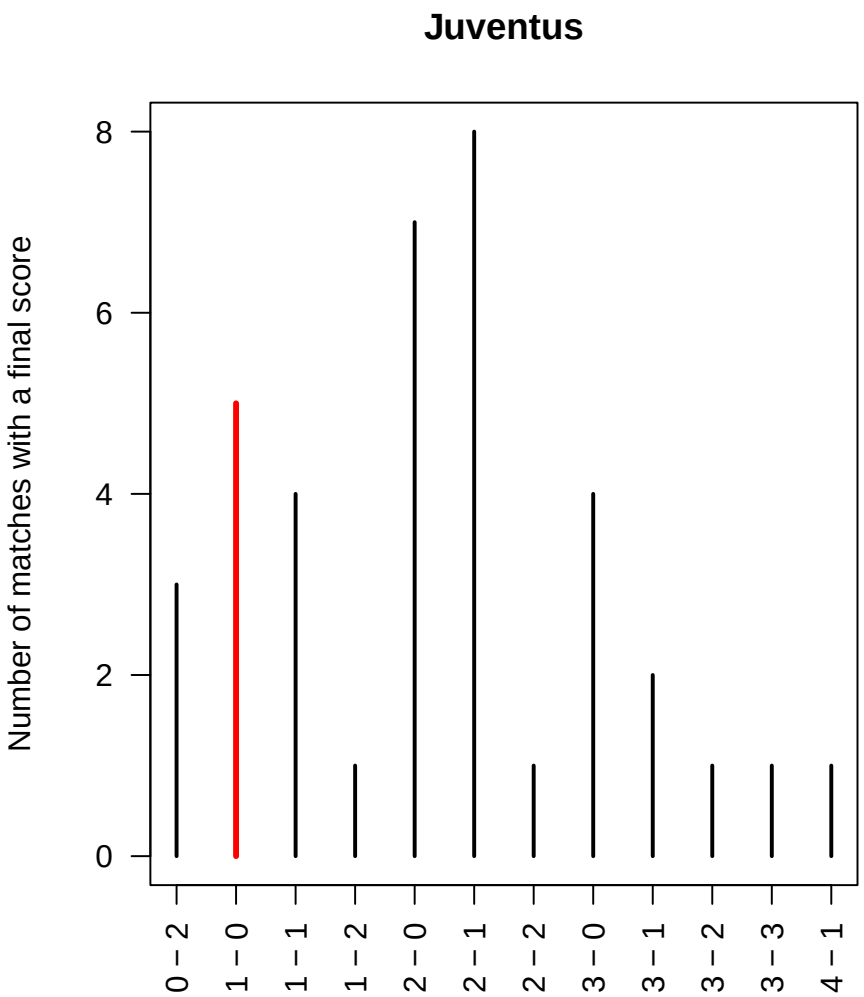
Team	Points	GS	GC
Atalanta	76	76	48
Napoli	75	72	41
Juventus	74	69	40
Inter	63	56	42
Roma	62	65	51
Milan	58	56	42
Sampdoria	58	57	52
Lazio	57	55	49
Torino	55	53	43
Sassuolo	44	55	59
Fiorentina	42	49	49
Bologna	39	49	55
Empoli	38	53	67
Udinese	32	43	54
Spal	32	46	55
Cagliari	31	41	55
Parma	29	42	60
Genoa	27	43	56
Frosinone	20	38	67
Chievo	17	38	71

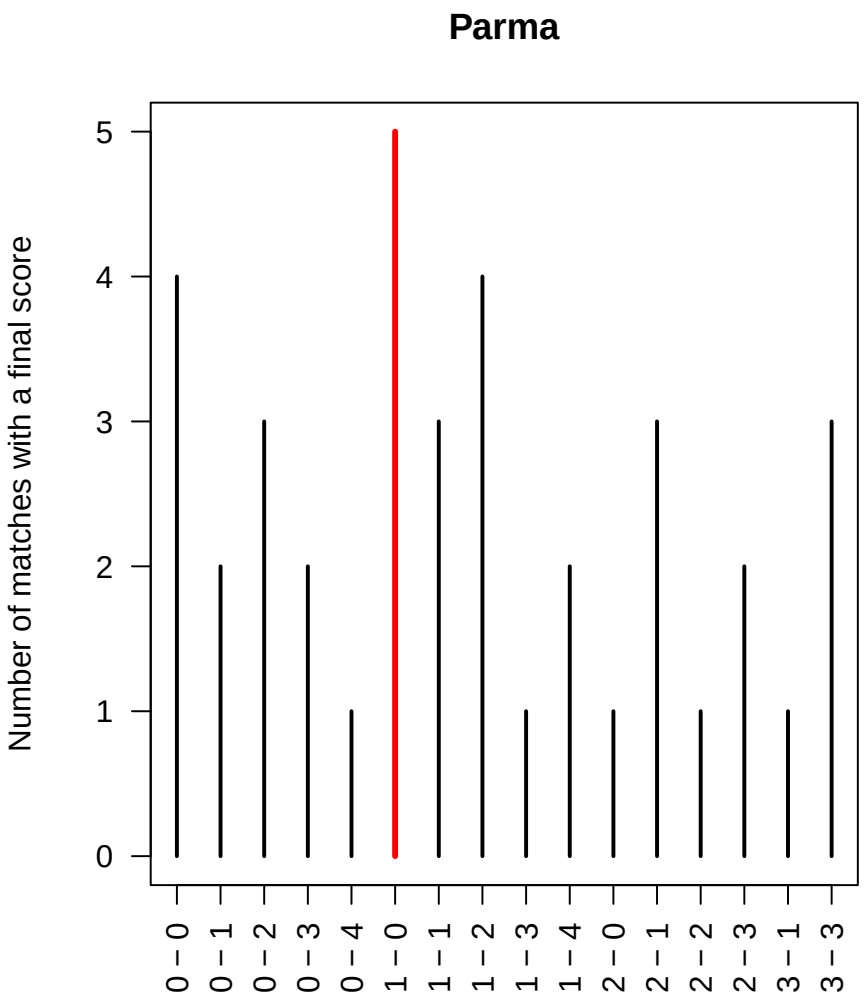
Table 3: Bivariate rank

Real rank vs predicted one

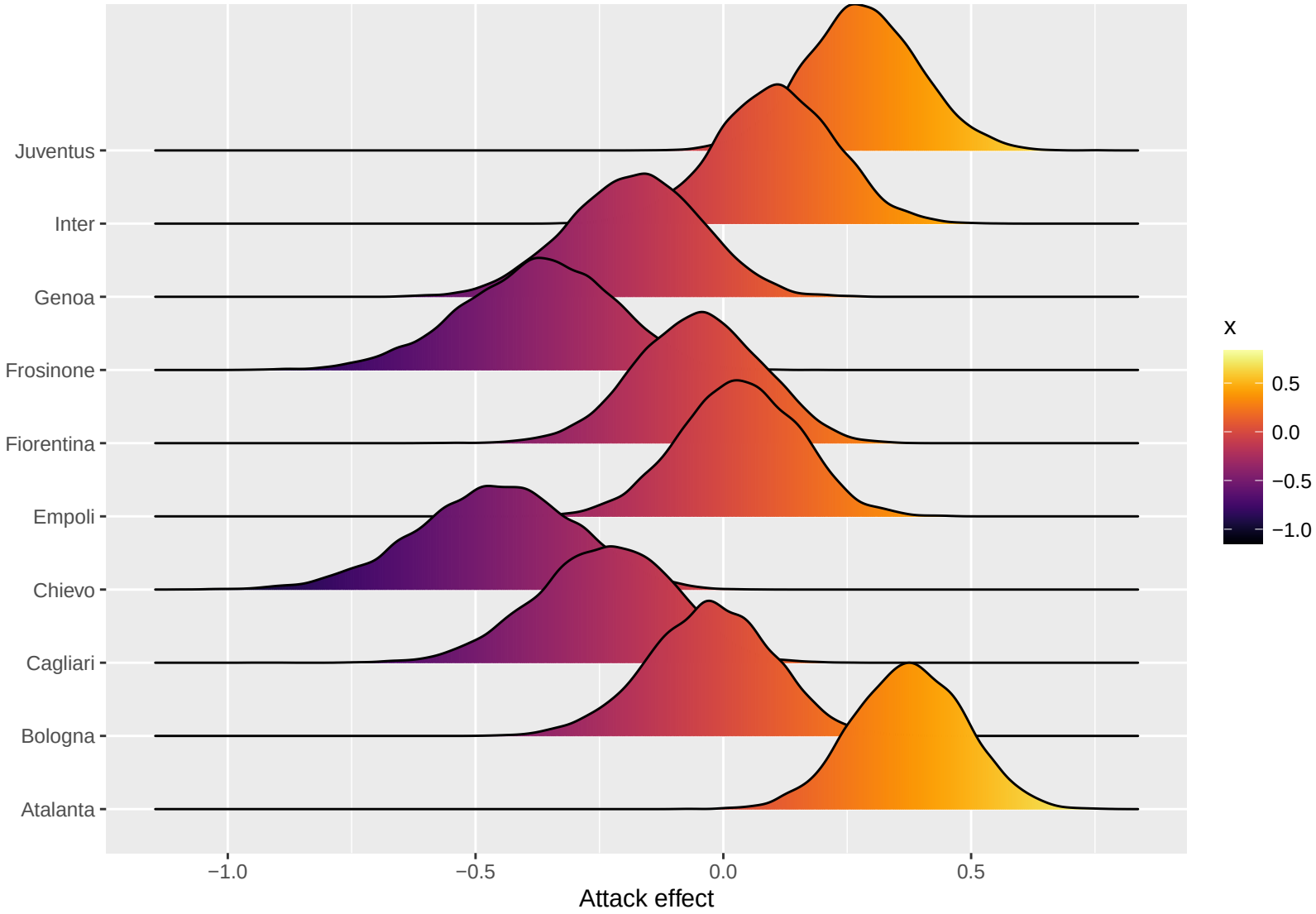


Main differences between rank of two teams

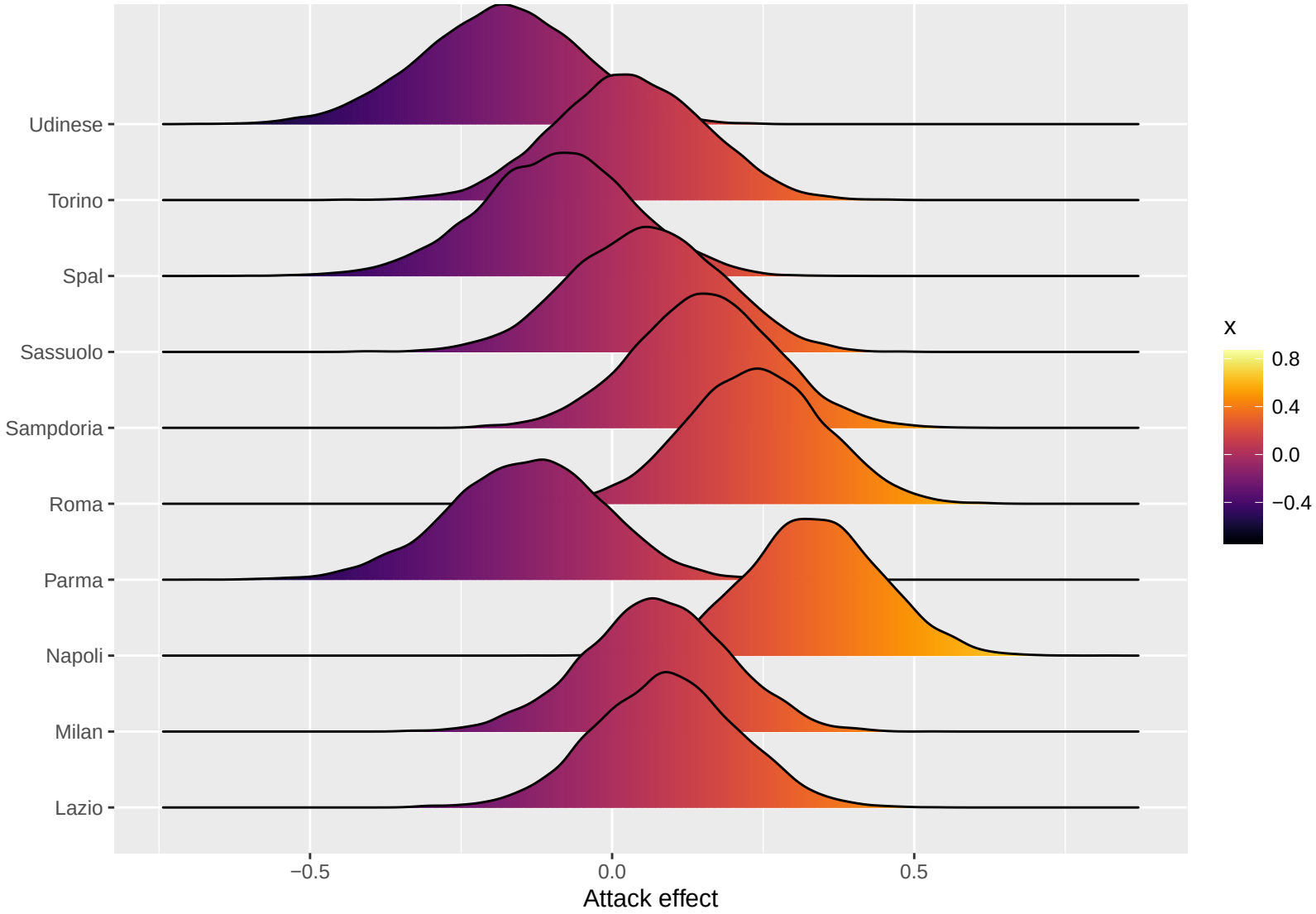




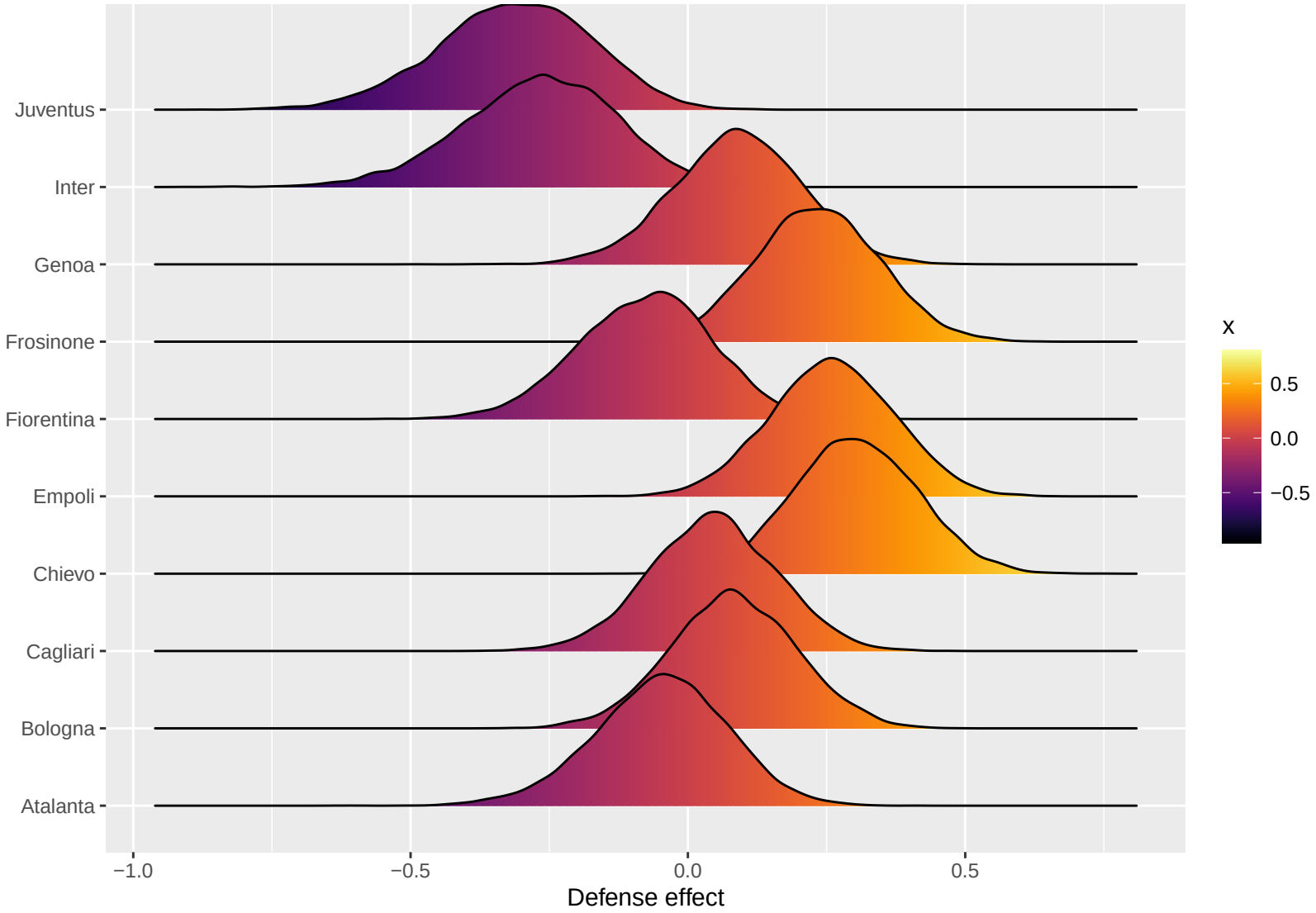
Attack results bivariate Poisson model



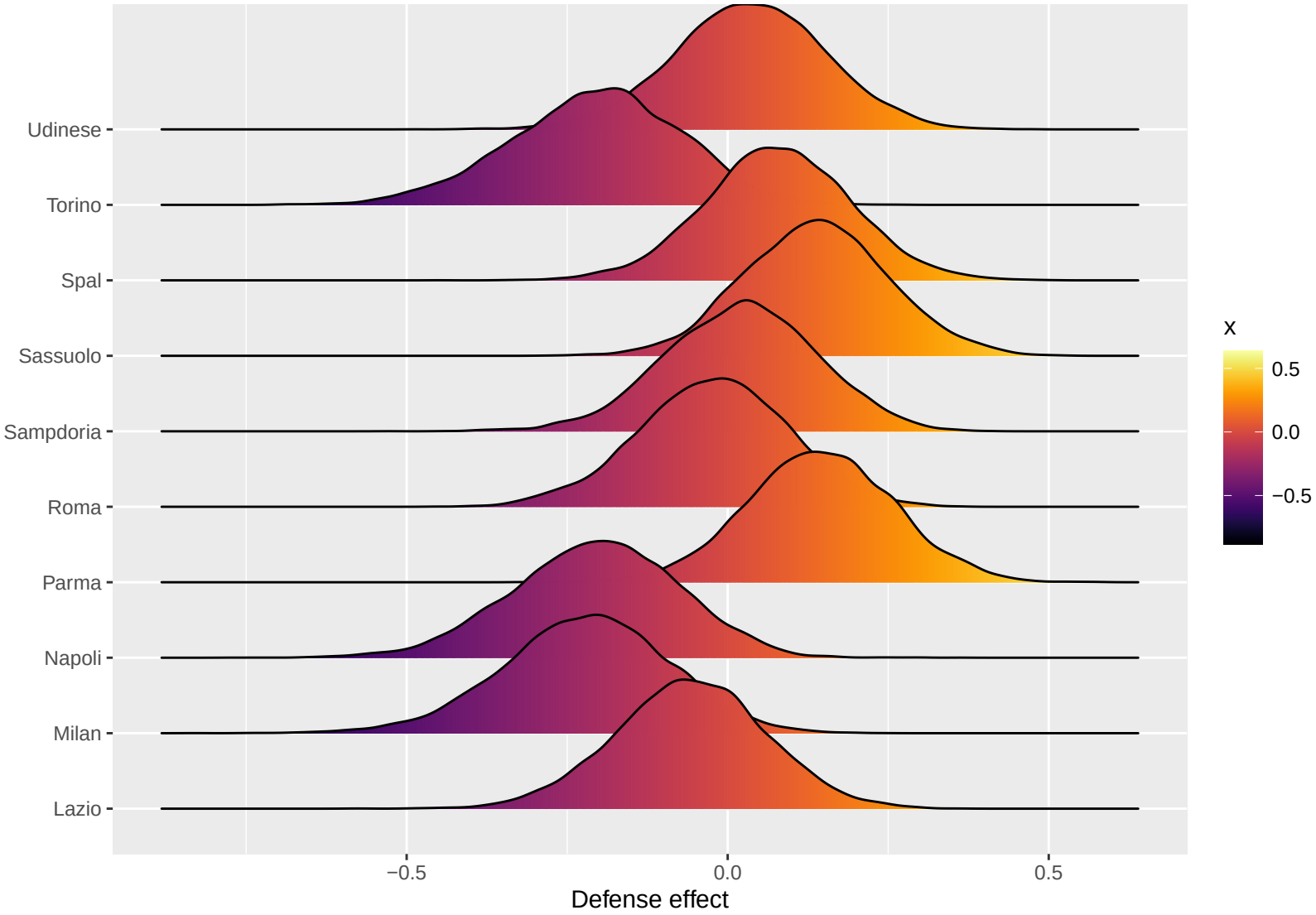
Attack results bivariate Poisson model



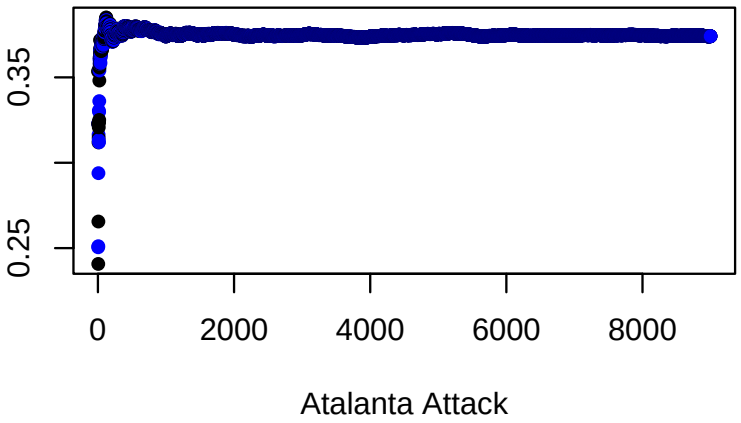
Defense results bivariate Poisson model



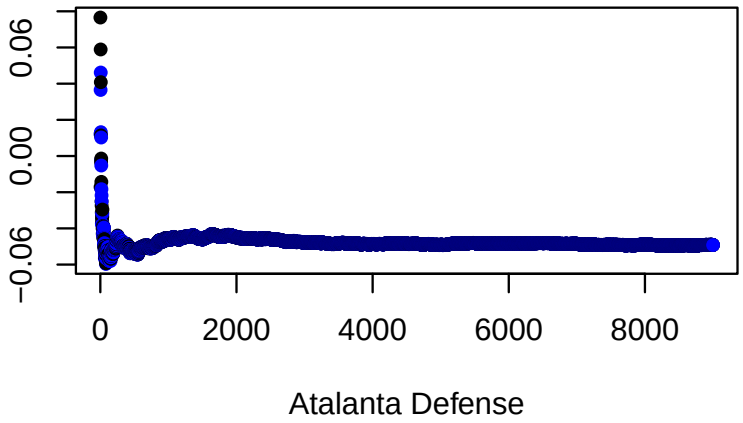
Defense results bivariate Poisson model



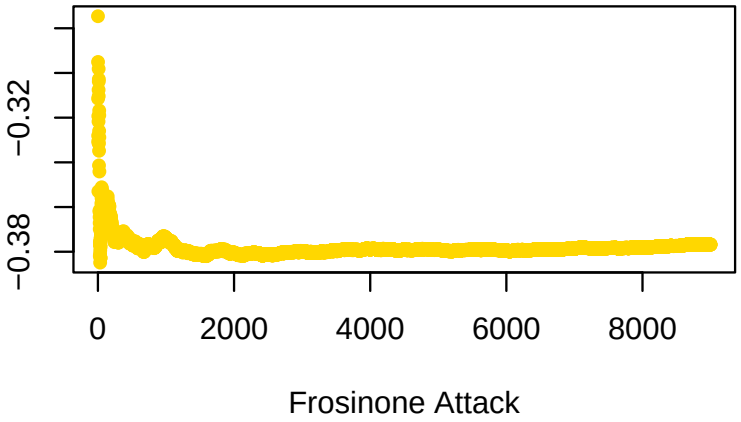
Convergence of parameters



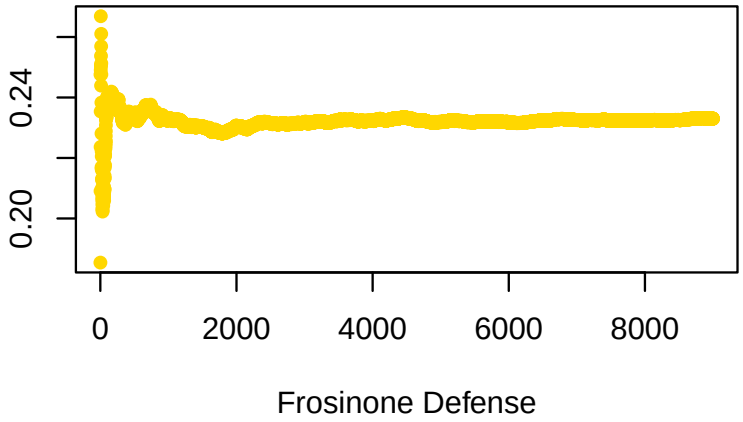
Atalanta Attack



Atalanta Defense

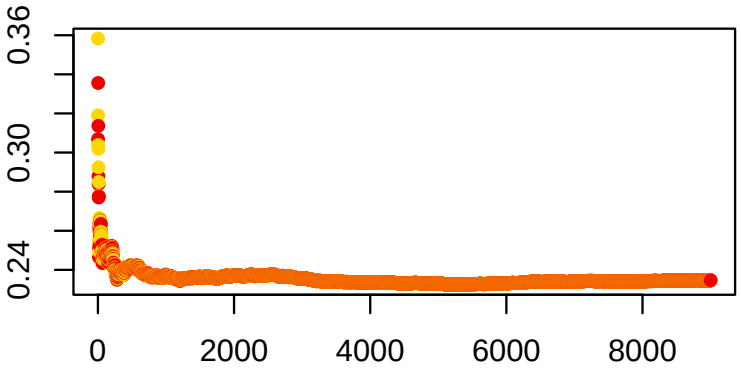


Frosinone Attack

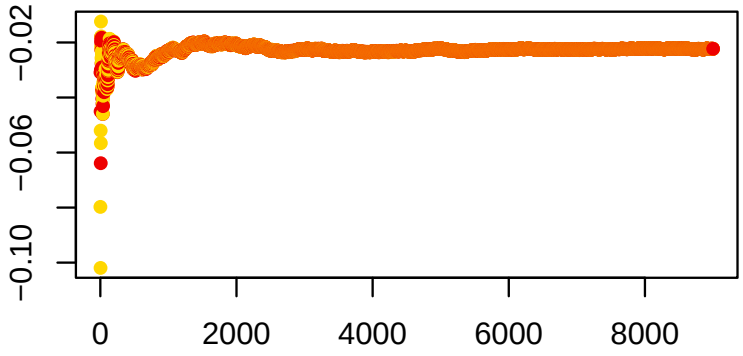


Frosinone Defense

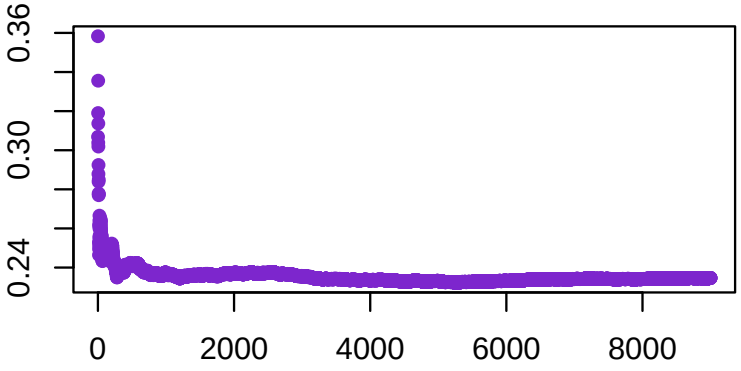
Convergence of parameters



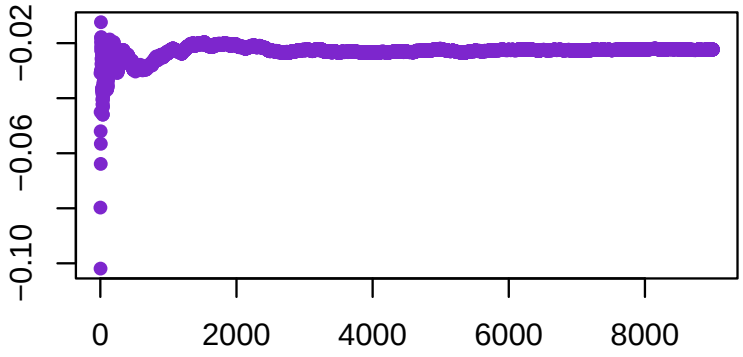
Roma Attack



Roma Defense

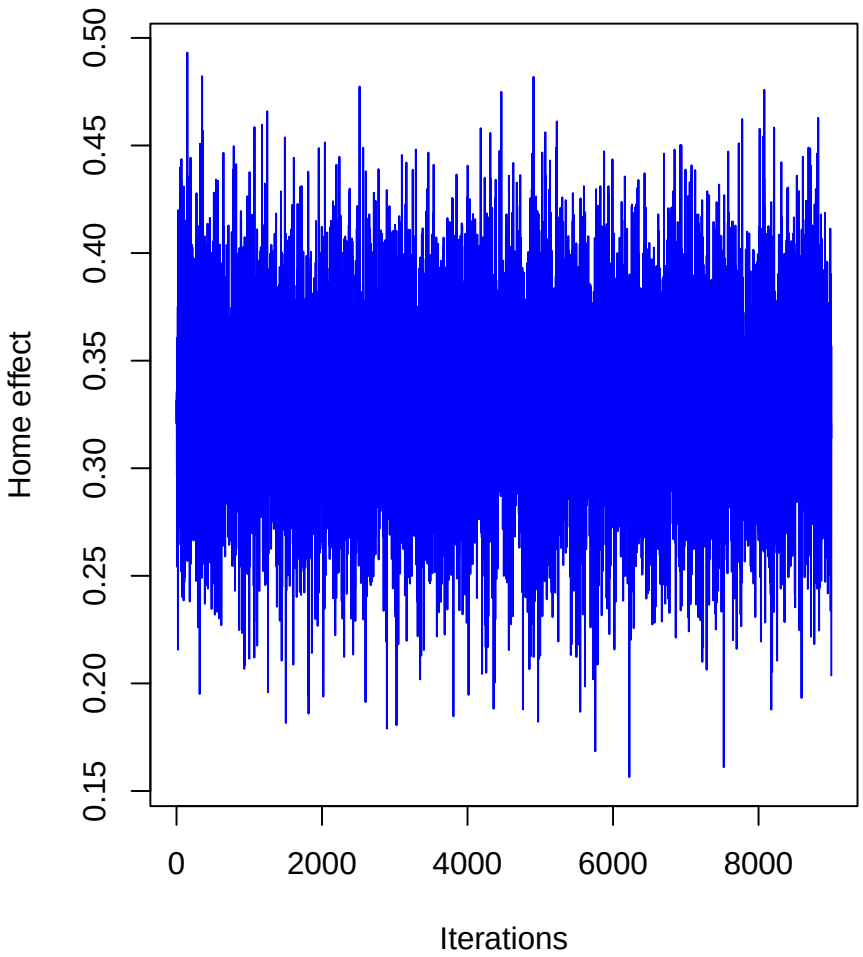


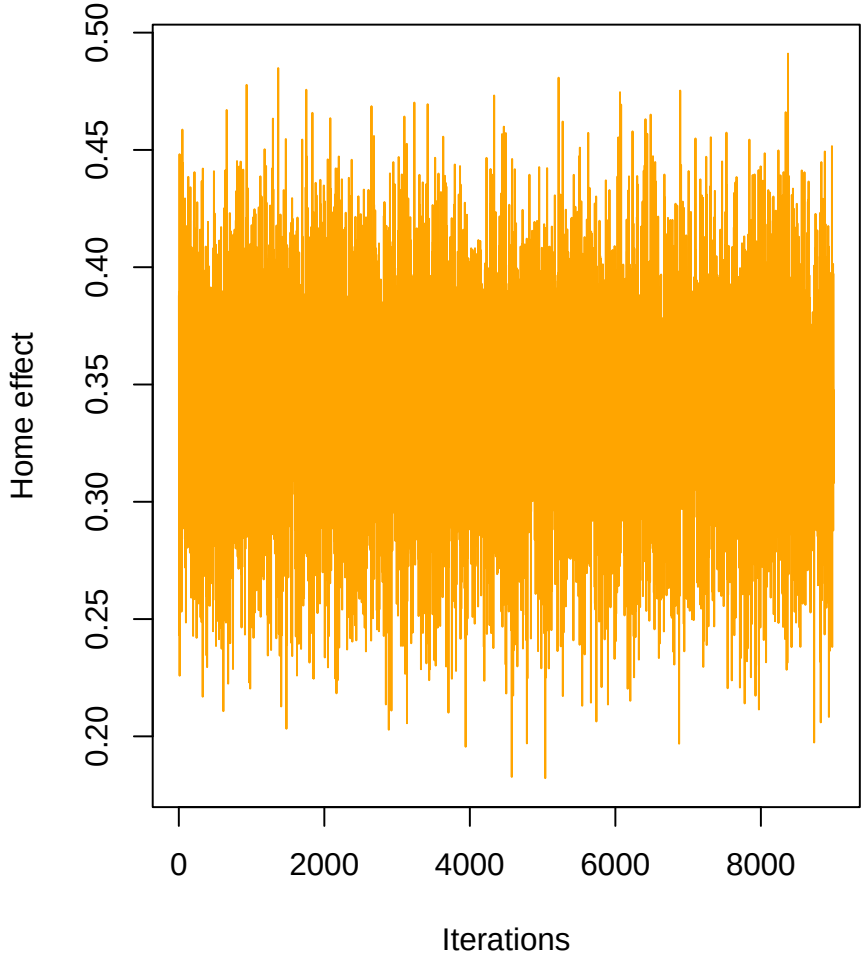
Fiorentina Attack



Fiorentina Defense

Traceplots


Home effect
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0
2000
4000
6000
8000
Iterations


Home effect
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0
2000
4000
6000
8000
Iterations

Bibliography

- Baio, Gianluca, and Marta Blangiardo. "Bayesian hierarchical model for the prediction of football results." Journal of Applied Statistics 37.2 (2010): 253-264.
- Karlis, Dimitris, and Ioannis Ntzoufras. "Analysis of sports data by using bivariate Poisson models." Journal of the Royal Statistical Society: Series D (The Statistician) 52.3 (2003): 381-393.
- Karlis, Dimitris, and Ioannis Ntzoufras. "Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R." Journal of Statistical Software 14 (2005): 1-36.
- AlMuhayfith, Fatimah E., Abdulhamid A. Alzaid, and Maha A. Omair. "On bivariate Poisson regression models." Journal of King Saud University-Science 28.2 (2016): 178-189.