Bayesian Hierarchical Model for prediction of football matches
Ryan Chan and Dr. John Paul Gosling

Negative Binomial Distribution

To model the number of goals scored by two teams in a match, this model will use the negative binomial distribution. The parametrisation that Stan uses for the negative binomial distribution is in terms of the mean μ and size n with probability mass function defined as

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

If a random variable Y follows a negative binomial distribution with mean μ and size n, then we write $Y \sim \text{NB}(\mu, n)$.

Negative Binomial Distribution

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

$$y_{gj} \mid \mu_{gj}, \sigma^2 \sim \text{NB}(\mu_{gj}, \sigma^2),$$

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

$$\log\,\mu_{g1} = \mathit{home\_att}_{h(g)} + \mathit{away\_def}_{a(g)}$$
$$\log\,\mu_{g2} = \mathit{away\_att}_{a(g)} + \mathit{home\_def}_{h(g)}$$

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

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

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

$$\sum_{t=1}^{T}\mathit{home\_att}_t = 0,\ \sum_{t=1}^{T}\mathit{away\_att}_t = 0,\ \sum_{t=1}^{T}\mathit{home\_def}_t = 0,\ \sum_{t=1}^{T}\mathit{away\_def}_t = 0.$$

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

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

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

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

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

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

The prior distributions for n_{home} and n_{away} were chosen to have a large variance, since we were uncertain about these parameters apriori.

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

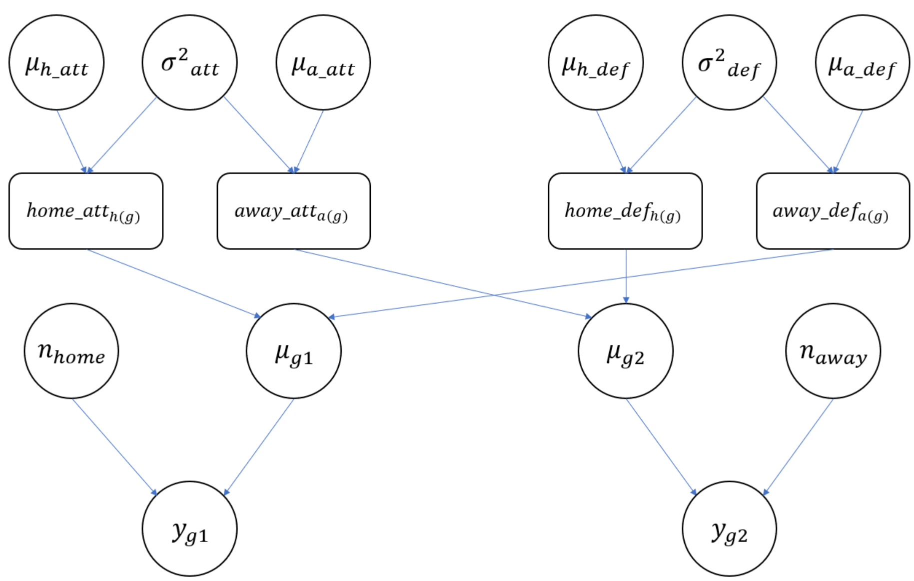

Figure 1: The DAG representation of the Negative-Binomial Model