The Dynamic Structured Bradley-Terry Model

Alan Chau James Thornton

April 26, 2019

Abstract

This report explores the Dynamic extension of the Bradley Terry model and how one may conduct inference using a data augmentation technique via Polya Gamma auxiliary variables. A data-augmented Gibbs sampler is derived and we present results on simulations and applications to the National Football League (NFL) 2007-2015 matches data .

1 Introduction

The Bradley-Terry model provides a popular model for tournaments of pairwise comparisons. The model assumes there exists a strength parameter λ_i for each player i in $S_N = [1 : N] = \{1, ..., N\}$. The comparison between player i and j is then modelled by the following probability of winning:

$$\mathbb{P}(i \ beats \ j) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \tag{1}$$

$$= \frac{e^{(\lambda_i - \lambda_j)}}{1 + e^{(\lambda_i - \lambda_j)}} \tag{2}$$

Setting $\psi_{i,j} = \lambda_i - \lambda_j$ to be the log odds of the above Bernoulli variable, we obtain the following likelihood:

$$L(\lambda) = \prod_{l \in L} \frac{(e^{\psi_l})^{y_l}}{(1 + e^{\psi_l})^{n_l}} \tag{3}$$

where $L = \{(i, j) \in S_N^2 | i < j\}$. For $l = (i, j)$, y_l is the number of times player i beat j. n_l is the number of times player i and j are compared.

We will consider a dynamic extension of the Bradley Terry model specified above and also incorporate covariates. A time index is introduced to the parameters in order to specify the dynamic model. We set $\lambda_i = x_{t,i}^T \beta_t$, where $x_{t,i}$ are the covariates of player i and β_t can be viewed as coefficients connecting the covariates to the strength variables, λ_i.

To summarise, the Dynamic Bradley Terry model can be used to model the following data generating process:

$$\begin{aligned}
\beta_0 &\sim \mathcal{N}(\mu, \Sigma) \\
\epsilon_t &\sim N(0, I) \\
\beta_{t+1} &= \rho \beta_t + \sigma \sqrt{1 - \rho^2} \epsilon_t \\
y_{t,i,j} &\sim Bin\Big(n_{t,i,j}, \frac{1}{1 + e^{(X_{t,j} - X_{t,i})\beta_t}}\Big) \\
&\forall t > 0, \ (i, j) \in L
\end{aligned}$$

for some parameters μ, Σ and ρ. The data generating process can visualised graphically in figure 1

This results in the following modified likelihood:

$$L(\{\beta_t\}_{t=1}^T) = \prod_{t=1}^{T} \prod_{l \in L} \frac{(e^{\psi_{t,l}})^{y_{t,l}}}{(1+e^{\psi_{t,l}})^{n_{t,l}}} \tag{4}$$

This analytically inconvenient form of the likelihood is what makes Bayesian inference on logistic models difficult. There are several workarounds for this; Albert and Chib (1993) proposed probit regression, a simple latent-variable method for posterior sampling, several extensions are also introduced afterwards. In this project, we will use the data augmentation strategy of Polson et al. (2013) with Polya-Gamma variables to perform inference on the above binomial likelihood.

This report is organised as follow; In section 2 we will go through the necessary tools required to simulate from the posterior and in section 3 we will derive the Gibbs sampler for the Dynamic Structured Bradley Terry model. Section 4 and 5 will contain simulation and experiment results.

1.1 Notation

To generalise notation let the following hold. For each $t \in [1:T]$:

$$\begin{aligned}
\psi_t &= \mathbf{D} X_t \beta_t \\
{}_t &= (\psi_{t_1}, \psi_{t_2}, \ldots) \in \mathbb{R}^{\frac{n(n-1)}{2}} \\
\beta_t &= (\beta_{t_1}, \beta_{t_2}, \ldots) \in \mathbb{R}^p \\
X_t &\in \mathbb{R}^{n \times p} \\
\mathbf{D} &\in \{0, 1, -1\}^{\frac{n(n-1)}{2} \times n}
\end{aligned}$$

X_t is the matrix of covariates, $\mathbf{D}$ is a linear difference operator matrix to obtain all pairwise-comparisons. Each row is unique with a single $+1$ and a single -1 entry per row, all other entries are null. The order of the rows encodes the mapping from (i, j) to $l \in L$.

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & \ldots \\ \vdots & \ddots & \\ \vdots & & -1 \end{bmatrix}$$

2 Background

In this section we will describe the necessary ingredients we need to derive an inference strategy for the Dynamic Structured Bradley Terry model.

2.1 The Polya-Gamma Trick

Data augmentation using the Polya-Gamma distribution is an efficient and simple method for inference in Bayesian Logistic Regression. It does not require the need for analytic approximations, numerical integration or Metropolis-Hastings and it has been seen to outperform other known data-augmentation techniques, Windle et al. (2013).

Definition 1. *A random variable X has a Polya-Gamma distribution with parameters $b > 0$ ad $c \in \mathbb{R}$, denoted by $X \sim PG(b, c)$ if:*

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + c^2/(4\pi^2)} \tag{5}$$

where the $g_k \sim Gamma(b, 1)$ are independent gamma distributed random variables. We will present here a useful property from Polya-Gamma distribution, which we will use it to derive our inference later. Let $w \sim PG(b, \psi)$, then the density is given by:

$$p(w|b, \psi) = cosh^b(\psi/2)exp(-w\psi^2/2)p(w|b, 0) \tag{6}$$

We will follow the method proposed by Windle et al. (2013) and demonstrate how one could utilise the Polya-Gamma trick to derive a posterior update for the Dynamic Bradley Terry model.

In short, we introduce a latent variable to augment the likelihood, and instead of working on a difficult-to-sample posterior, we sample from the conditional of a linear Gaussian state space model given the observation using simulation smoother. Inference can then be done iteratively using Gibbs Sampling.

2.2 Simulation Smoother

Durbin and Koopman (2002) proposed a simulation procedure for drawing samples from the conditional distribution of state vectors given the observations under the linear Gaussian state space model. A linear Gaussian state space model has the following form:

$$z_t = X_t\beta_t + V_t, \qquad V_t \sim N(0, H_t) \tag{7}$$

$$\beta_{t+1} = T_t\beta_t + R_t\epsilon_t, \qquad \epsilon_t \sim N(0, Q_t) \tag{8}$$

where z_t is a $p \times 1$ vector of observations, β_t is an $m \times 1$ state vector and V_t and ϵ_t are vectors of disturbances. Matrices X_t, T_t, R_t, H_t and Q_t are assumed to be known. To begin the process we assume that $\beta_1 \sim N(0, P_1)$, where P_1 are known.

The simulation smoother utilises the fact the conditional distributions, given the observations, are Gaussian and the conditional covariance matrix actually does not depend on the observations at all. This gives us the following algorithm to sample the states β_t from the conditional distribution:

Step 1: Draw β^+ and z^+ by means of recursion to Equation 7 and 8 where the recursion is initiated by a draw $\beta^+ \sim N(0, P_1)$

Step 2: Construct the artificial series $z^* = z - z^+$ and compute $\hat{\beta}^* = E(\beta|z^*)$ by putting z^* through the Kalman filter and smoother.

Step 3: Take $\tilde{\beta} = \beta^+ + \hat{\beta}^*$. $\tilde{\beta}$ is a draw from the distribution of β conditional on z.

This algorithm can be seen as a mean correction to the original sample β^+. For further details, please read Durbin and Koopman (2002) and Jarociński (2015). In the next section we will use the tools described above to derive a posterior update for the Dynamic Bradley Terry model.

3 Inference on Dynamic Structured Bradley Terry Model

Recall the likelihood we introduced in section 1 and let the prior of the βs' be denoted $p(\beta)$, then the posterior distribution of $\beta = \{\beta_t\}_{t=1}^T$ is:

$$P(\{\beta_t\}_{t=1}^T|y) = \prod_{t=1}^{T}\prod_{l\in L}\frac{(e^{\psi_{t,l}})^{y_{t,l}}}{(1+e^{\psi_{t,l}})^{n_{t,l}}}p(\beta) \tag{9}$$

$$= \frac{2^{-n_{t,l}}\exp\left\{\psi_{t,l}(y_{t,l} - \frac{n_{t,l}}{2})\right\}}{cosh^{n_{t,l}}(\frac{\psi_{t,l}}{2})} \tag{10}$$

To derive the sampling scheme for this posterior, we will follow the method proposed by Windle et al. (2013) and introduce latent Polya-Gamma variables.

We will now introduce the latent variable $w_{t,l} \sim PG(n_{t,l}, \psi_{t,l})$ to the system and and consider the joint distribution:

$$
\begin{aligned}
P(\beta, w|y) &= P(w|\beta, y)P(\beta|y) && (11)\\
&\propto \prod_{l\in L}\prod_{t=1}^{T}\Big[cosh^{n_{t,l}}\Big(\frac{\psi_{t,l}}{2}\Big)\exp\Big(-w_{t,l}\frac{\psi_{t,l}^2}{2}\Big)p(w_{t,l}|n_{t,l},0) && (12)\\
&\qquad\qquad \times \frac{(e^{\psi_{t,l}})^{y_{t,l}}}{(1+e^{\psi_{t,l}})^{n_{t,l}}}p(\beta)\Big]\\
&\propto \prod_{l\in L}\prod_{t=1}^{T}\exp\left\{-w_{t,l}\frac{\psi_{t,l}^2}{2}\right\}\exp\left\{\psi_{t,l}\Big(y_{t,l}-\frac{n_{t,l}}{2}\Big)\right\}p(w_{t,l}|n_{t,l},0)p(\beta)\\
&\propto \prod_{l\in L}\prod_{t=1}^{T}\exp\left\{\psi_{t,l}\Big(y_{t,l}-\frac{n_{t,l}}{2}\Big)-w_{t,l}\frac{\psi_{t,l}^2}{2}\right\}p(w_{t,l}|n_{t,l},0)p(\beta) && (13)
\end{aligned}
$$

The first step in the above uses equation (6) for the Polya-Gamma density and the second step uses (10) to remove the *cosh* terms. The third simply combined exponents.

Completing the square on the exponent in equation (13) leaves remaining terms related to β as the full conditional:

$$
\begin{aligned}
P(\beta|w,y) &\propto \prod_{l}\prod_{t=1}^{T}\exp\left\{-\frac{w_{t,l}}{2}\Big(\frac{1}{w_{t,l}}\big(y_{t,l}-\frac{n_{t,l}}{2}\big)-\psi_{t,l}\Big)^2\right\}p(\beta) && (14)\\
&\propto \prod_{l\in L}\prod_{t=1}^{T}\exp\left\{-\frac{w_{t,l}}{2}\Big(z_{t,l}-\psi_{t,l}\Big)^2\right\}p(\beta) && (15)
\end{aligned}
$$

If we set $z_{t,l} = \frac{1}{w_{t,l}}(y_{t,l} - \frac{n_{t,l}}{2})$, then clearly $z_{t,l} \sim N(\psi_{t,l}, \frac{1}{w_{t,j}})$ is a Gaussian distribution. Moreover, $p(\beta)$ itself is an autoregressive model. In other words, we can represent Equation (14) as the following Gaussian State space model, appealing to notation in section 1.1 for $\psi_t = \mathbf{D}X_t\beta_t$:

$$
\begin{aligned}
Z_t &= \mathbf{D}X_t\beta_t + V_t, \qquad V_t \sim N(0, H_t), \qquad H_t = Diag(\left\{\frac{1}{w_{t,l}}\right\}_l) && (16)\\
\beta_{t+1} &= \rho\beta_t + R_t\epsilon_t, \qquad \epsilon_t \sim N(0, I), \qquad R_t = \sigma\sqrt{1-\rho^2} && (17)
\end{aligned}
$$

The dynamic model with auxiliary variables can be visualised using figure 1.

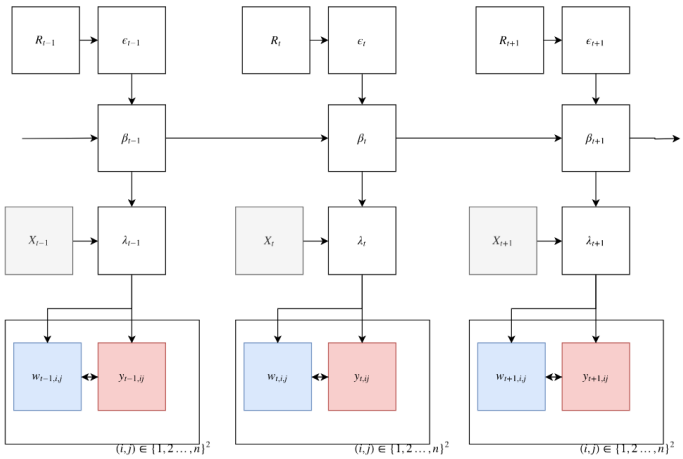

Figure 1: Plate diagram showing dependency structure of the Dynamic Bradley Terry model. Covariates are cloured grey, observations in red and the auxliary variables in blue. Note that the λ_t are a deterministic function of β_t given the covariates, but included for clarity.

This particular form allows us to use the simulation smoother described in the previous section and sample β from the full conditional. Finally, we arrive to the following Gibbs Sampler for the Dynamic Structured Bradley Terry model:

Step 1: For each time t and pairs l, simulate $w_{t,l} \sim PG(n_{t,l}, \psi_{t,l})$ and set up the Gaussian state space model.

Step 2: Sample the conditional $\{\beta_t\}_{t=1}^T$ from the state space model using simulation smoother.

Step 1 and 2 are repeated until convergence.

4 Simulation

In this section we present results from our simulation. At each of the 50 time points in the simulated tournament each of 10 players was compared with every other player 20 times. The initial vector β_0s are generated from Gaussian distribution with mean 1 and 3 standard deviation and we allow the state β_t to evolve as specified in the state space model with $\rho = 0.8$ and $\sigma = 1$. 3 static covariates are generated for each player from $N(1, 3^2)$. From Figure 2, we see that the Dynamic Structured BT model successfully recovers the true trends within the βs. The right-hand plot also shows the dynamic ranking through time based on an ordering of strength parameters.

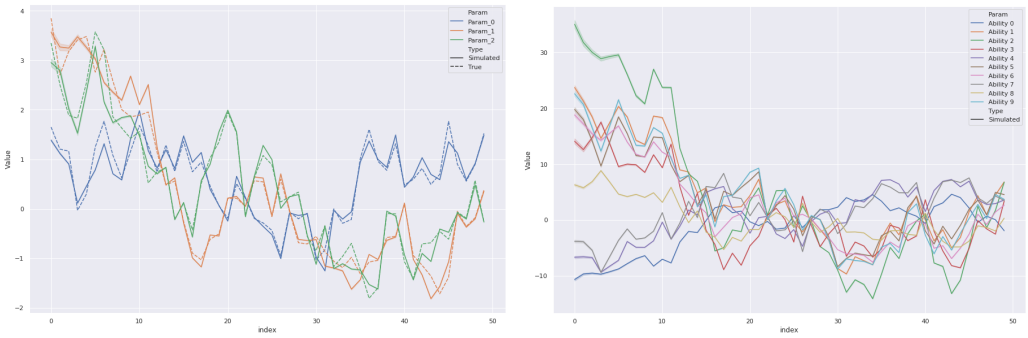

Figure 2:
Left: Progression of $\{\beta_t\}_{t=1}^3$ in simulated data with $\rho = 0.8$, $\sigma = 1$. Dotted lines are the true progression whereas the filled lines are the modelled progression.
Right: Ability/Strength parameters for the 10 simulated players

5 Application: NFL Football League 2007-2015

In this section we will apply the Dynamic structured Bradley Terry model to the NFL matches data from 2007 to 2015. The NFL is a professional American football league consisting of 32 teams, divided equally between the National Football Conference and the American Football Conference. The NFL is one of the four major professional sports leagues in North America, and the highest professional level of American football in the world.

To observe how the performance of the teams varies each year, we include the time element in a yearly base to our matches. Also, since not all teams had played against each other at least once every year, we borrow the idea from PageRank algorithm and include 2 artificial matches among all teams each year. To reduce the bias introduced, each team will lose and win exactly once in those artificial matches.

During the experiment, we realised it is difficult to elicit the correlation parameter ρ. Setting it below or above 1 will include a subjective trend to the skill vectors, while setting it to 1 will lose the autoregressive structure of the skills as the random noise is now deterministic (variance become 0). One way to counter this is to increase the noise level σ. Parameter elicitation must be carefully treated in practice.

To apply Dynamic Structured Bradley Terry to the NFL data, we picked the following dynamic features,

- Average winning score
- Average losing score
- Average winning yards
- Average losing yards

Looking into Figure 3, it is reassuring that the skill vector λ is negatively proportional to the average losing score, which is indicated by the negative values of the corresponding β parameter.

We could evaluate this model by comparing the predicted ranking with the overall league ranking online. Moreover, we could try to predict test set results using predicted ability parameter. This is not conducted due to time constraint of the project.

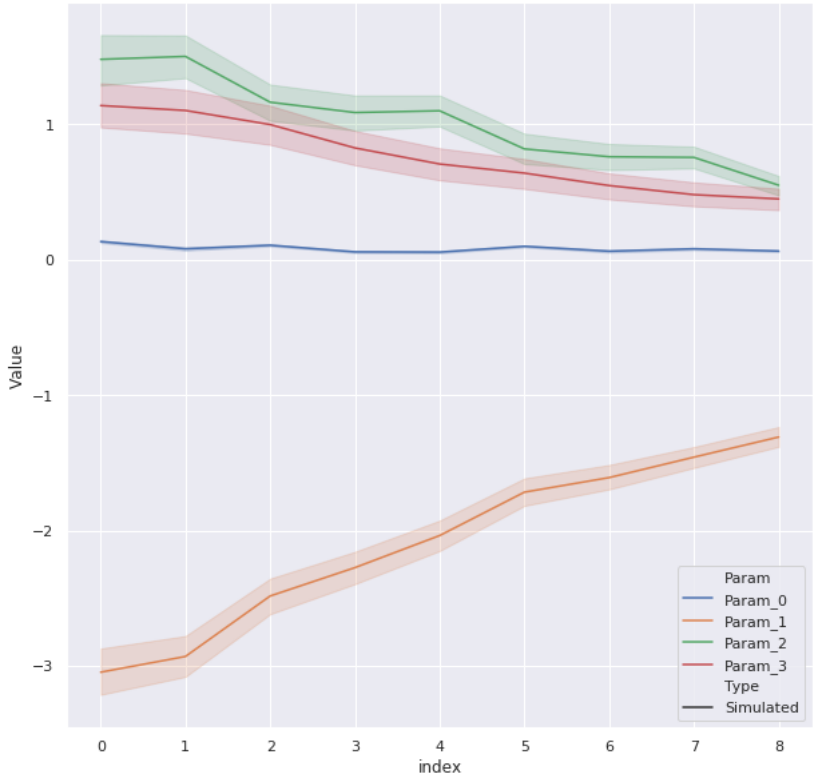

Figure 3: Progression of the β parameter

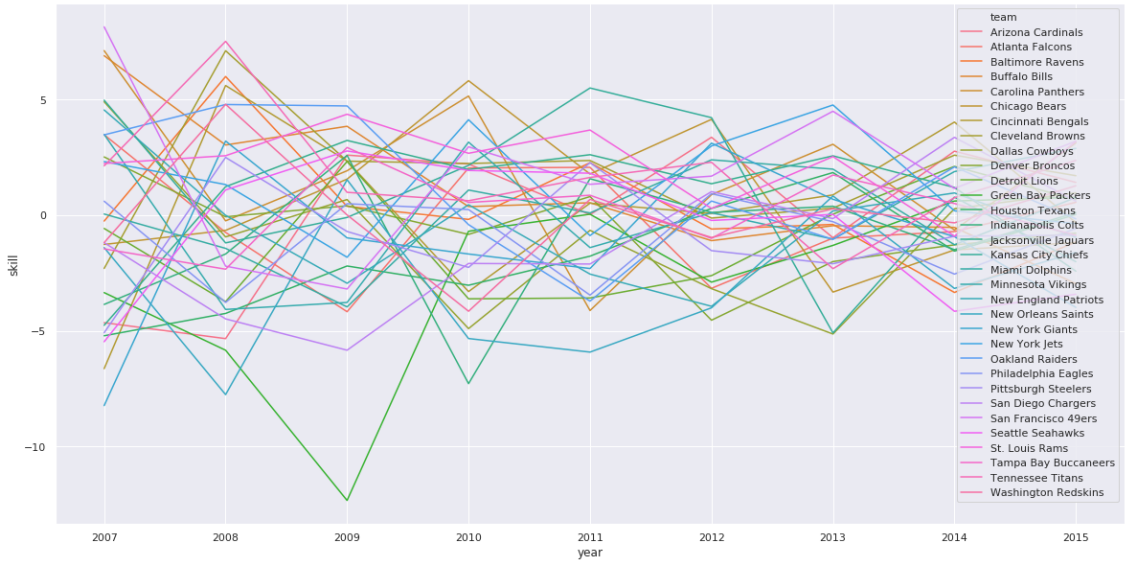

Figure 4: The ability score λ for 32 teams in NFL league 2007 - 2015

6 Conclusion and Remarks

This report details the Dynamic structured Bradley-Terry model and showed how it can be used, both on simulated and real-world data for explainable models and for dynamic ranking. Although a powerful, flexible model, there are many possible extensions.

In our implementation of the structured model, we introduced a linear relationship between the covariates and the ability factor. This is a strong assumption, especially without an error term on this mapping. Introducing an error term and investigating non-linear relationship between covariates and the ability factor is a natural next step. One possible way of doing so is to search for such a function in the Reproducing Kernel Hilbert Space (RKHS) and appeal to representer theorems for a tractable functional form of such non-linear functions. A simplification to the RKHS approach is to apply basis functions on the covariates to extract informative features.

An additional complexity of interest would be to set the evolution of state β_t as unknown, with a prior, and then learn this as part of the model fit.

As with many time-series applications, we may wish to learn the parameters in an online fashion. This would be an interesting extension of the simulation smoother approach.

It may be of interest to use this model for prediction. In the prediction use-case, a cross-validated error on test data may be used as a metric to assess performance. This performance can then be assessed alongside other models in a larger study.

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

James Durbin and Siem Jan Koopman. A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616, 2002.

Marek Jarociński. A note on implementing the durbin and koopman simulation smoother. *Computational Statistics & Data Analysis*, 91:1–3, 2015.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Jesse Windle, Carlos M Carvalho, James G Scott, and Liang Sun. Efficient data augmentation in dynamic models for binary and count data. *arXiv preprint arXiv:1308.0774*, 2013.