Hierarchical Bradley-Terry Models

Coen Jones
Supervisor: Dr. Alan Huang

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

A thesis submitted for the degree of Bachelor of Mathematics (Honours) at The University of Queensland in 2021

Abstract

Despite being one of the cornerstone models for head-to-head comparisons for almost 70 years, there is little research regarding how one can model homeground advantage using the Bradley-Terry model. At present, modern software only provides the ability to model a single level of homeground advantage, shared by all teams and in all situations. It is our belief that this is insufficient. In particular, we wish to consider how the hierarchical structure that exists within some sporting competitions may impact on modelling this effect. This paper first examines the current state of the model, from inception and initial recognition, through to modern developments and additions to the model. We then provide the mathematical detail behind our extensions, describing the series of nested models that can be used to consider differing levels of depth for modelling homeground advantages. Information on the implementation of these models in R then follows, before a series of examples of its usage, as well as a discussion on the validity of these results due to the effect time may have on the strengths of teams.

Contents

Chapter 1

Literature Review

1.1 Introduction

The model initially proposed by Zermelo (1929), and made famous by Bradley & Terry (1952), has become a commonly used model in the analysis of paired comparisons in a variety of fields, whilst also becoming a popular topic of research in its own right. This review of the literature shines light on the development of this statistical model, from its entrance into the mainstream in 1952, through to the current state-of-the-art implementations of the model in R, highlighting important extensions to the model which have been made along the way.

The model was originally proposed by Zermelo (1929) to solve the problem of ordering players at chess tournaments where players may have had unequal fixtures, whether that be differing in number of games or strength of opponents. However, it was not until the model was presented, with more details, by Bradley & Terry (1952) that it became popular, and in fact it is for this paper that the model is named. The model has proven both popular and useful in ranking the strengths of items in a wide variety of fields where pairwise comparisons can be made. Some examples include strengths of prominent journals based off their citations found in Agresti (2013), the effects of environmental factors on the fighting abilities of

African lizards as documented in Whiting et al. (2006), rankings of universities by Dittrich et al. (1998) and many sporting competitions, including the study of 'Home Ground Advantage' in the MLB, an example also pulled from Agresti (2013).

A number of natural extensions to the model have been proposed and accepted, some of which are directly relevant to our course of study. As discussed below, natural elements of sporting competitions such as ties, home-ground advantage and covariates such as weather have been factored into more modern versions of the model, and are discussed as a part of this review.

1.2 Bradley and Terry

Bradley & Terry (1952) is considered, for the most part, to be the first point of interest for the model. The paper proposes the following: Consider t treatments in an experiment involving paired comparisons. Consider that these treatments have true 'ratings', $\pi_1, \ldots, \pi_t$, which are considered only in relation to each other within the experiment. We can add the restrictions that each $\pi_i \geq 0$, and that $\Sigma\pi_i = 1$, to specify the model.

We propose that the probability some item i 'wins' against some item j is given by:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \pi_j}$$

As a result, Bradley and Terry consider a variable r_{ijk} which denotes the ranking of item i when it comes against item j in the k-th repetition of such a matchup, such that when item i wins, we have $r_{ijk} = 1$, and $r_{jik} = 3 - r_{ijk} = 2$. This then leads to the likelihood function proposed in the original paper:

$$\begin{aligned} L(\pi_i, \pi_j | r_{ijk}) &= \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{2-r_{ijk}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{2-r_{jik}} \\ &= \frac{\pi_i^{(2-r_{ijk})}\pi_j^{(2-r_{jik})}}{(\pi_i + \pi_j)} \end{aligned}$$

We can then generalise this for all possible comparisons, and for all n repetitions, ending up at the following likelihood function:

$$L(\pi|\mathbf{r}) = \prod_i \pi_i^{2n(t-1)-\sum_{j\neq i}\sum_k r_{ijk}} \prod_{i<j}(\pi_i + \pi_j)^{-n}$$

Bradley and Terry go on to propose a number of hypothesis tests possible using this model, leaning on likelihood ratio tests and a number of values for what they refer to as B statistics, a brief number of which are published in this paper, with an extended table available in the follow-up paper Bradley (1954). The third installment in this series of papers, Bradley (1955), then goes on to prove a number of results regarding the statistics proposed in the initial paper, as well discussing some other statistical extensions such as the variance and covariances of estimators, as well as the measuring of the power of a test under the model.

Whilst extensions on the model are thoroughly discussed below, it should also be noted that the usage of r_{ijk} to denote the outcome of an event appeared to have disappeared quite quickly from the literature, with many papers instead using a variable w_{ij} to indicate the number of times i had been preferred over j for the sake of simplicity, leading to a likelihood function of:

$$L(\pi|\mathbf{W}) = \prod_i \prod_{j<i} \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{w_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{w_{ji}}$$

In this case, the lazy implementation of ties would be to treat a tie between i and j as $w_{ij} = w_{ji} = \frac{1}{2}$, allowing the model to incorporate the information into the estimation of the π_is, but not able to predict a tie as a potential outcome - something which the extension discussed below aims to do.

1.3 Handling Ties

Whilst ties may be uncommon in some of the potential uses for this model, they are very common in a number of sports - especially in Chess, for which Zermelo first proposed the model. However, the model in its 1955 form notably lacks a way to properly estimate and predict this common outcome. Thus, handling ties would appear as perhaps the most natural way to extend such a model, and a number of methods to manage these outcomes were proposed throughout the 1960s - including those by Glenn & David (1960), Rao & Kupper (1967), and Davidson (1970)

The first method worth considering is that proposed by Rao & Kupper (1967), in what would seem to be a loose adaptation of the method proposed for extending the Thurstone-Mosteller model by Glenn & David (1960). In their paper, Rao and Kupper propose the use of a threshold parameter η. In the case they describe, when the difference of the logs of the 'skills' of two teams is less than η, the judges will declare a tie, allowing for differing levels of sensitivities. As a result, Rao and Kupper propose that if we then take $\theta = \exp\{\eta\}$,[1] that we can have a modified model of:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \theta\pi_j}$$

$$\mathbb{P}(i < j) = \frac{\pi_j}{\theta\pi_i + \pi_j}$$

$$\mathbb{P}(i = j) = \frac{\pi_i\pi_j(\theta^2 - 1)}{(\pi_i + \theta\pi_j)(\theta\pi_i + \pi_j)}$$

which we can clearly see reduces back to a classic Bradley-Terry model for $\theta = 1$, as this implies that $\eta = 0$ and that the absolute difference in log-skills must be negative to declare a tie. Their paper goes on to discuss a likelihood function for this model, and thus ways to estimate θ and each π_i under the extension.

A follow-up proposal by Davidson (1970) is also worth discussing. Closely linked to Rao and Kupper's proposal, it considers a model where the probability of a tie is proportional to the geometric mean of the two items winning - i.e.:

$$\mathbb{P}(i = j) = \nu\sqrt{\mathbb{P}(i > j)\mathbb{P}(i < j)}$$

for some $\nu \geq 0$ which does not depend on i or j. We can then determine

[1] Note that this usage of θ varies from its more common usage throughout this paper - perhaps in part due to the fact that this proposal came before the widespread adoption of the log-linear parametrisation.

that the extended model looks like:

$$\mathbb{P}(i > j) = \frac{\pi_i}{\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}}$$

$$\mathbb{P}(i = j) = \frac{\nu\sqrt{\pi_i \pi_j}}{\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}}$$

Similar to the Rao and Kupper model, this simply yields the original Bradley-Terry model when we set $\nu = 0$.

It is worth noting that the Davidson model appears to have become the standard method for handling ties in the literature. It is also worth noting that not all implementations of the model aim to predict and model ties in this manner, and it is also quite common to simply include draws as results in the data by assigning half a win to each team. Whilst this means that the model will not predict draws in the future, the model is still able to use the information provided when a draw occurs in any given matchup.

1.4 Order Effects

In many competitions, the order in which two items are presented may have some bearing on their probability of winning. One common example of this effect is the 'Home Ground Advantage' that teams may hold in many sports. A similar advantage exists in Chess too, as the player handling the white pieces is given the first move. Being able to extend the model to incorporate this sort of effect is something which can help the model to better explain the true strengths of teams, in particular when teams may experience uneven draws - such as in the AFL, where teams often only play each other once, making it impossible for each side to have the opportunity to play each opponent both at home and away.

These order effects appear to be first considered by Beaver & Gokhale (1975) by simply adding on some linear factor to the item presented first. They consider some δ_{ij} with restriction that $|\delta_{ij}| \leq \min\{\pi_i, \pi_j\}$ to represent the 'bias' attached to the situation that item i is presented before item j.

In this additive model, we have the following when i is presented before j:

$$\mathbb{P}(i > j) = \frac{\pi_i + \delta_{ij}}{\pi_i + \pi_j}$$

$$\mathbb{P}(i < j) = \frac{\pi_j - \delta_{ij}}{\pi_i + \pi_j}$$

Following this, Davidson & Beaver (1977) proposed that a multiplicative order effect could be used instead. In their paper, they make note of the fact that the Bradley-Terry model possesses the property that the values $\ln \pi_i$ can be used to represent the merits of the objects on a linear scale. They conclude that due to this property, it would seem sensible to use a multiplicative factor to value the home strength, as this then becomes an additive factor on the logarithmic scale. For their model, they consider a $\gamma_{ij} \geq 0$, and make the assumption that $\gamma_{ij} = \gamma_{ji}$, implying that the benefit gained by item i being presented first is equivalent to the benefit item j may gain by being presented first instead. The model then treats the probabilities as being:

$$\mathbb{P}(i > j)_i = \frac{\pi_i \gamma_{ij}}{\pi_i \gamma_{ij} + \pi_j}$$

$$\mathbb{P}(i < j)_i = \frac{\pi_j}{\pi_i \gamma_{ij} + \pi_j}$$

We can see that $\gamma_{ij} = 1$ yields the vanilla Bradley-Terry model, and that $\gamma_{ij} > 1$ is of benefit to item i, with $\gamma_{ij} < 1$ benefiting item j. It should also be noted that we do not have to make the same assumption that Davidson and Beaver have that $\gamma_{ij} = \gamma_{ji}$, and that there may a number of cases where it would not make sense for these two to equal - for instance, in the comparison of aromatics, a more overpowering scent could have a strong advantage if placed before judges first, however the advantage gained by a weaker scent would not necessarily be equal should it be presented first instead.

The paper also discusses the case where $\gamma_{ij} = \gamma$ for all $\{i, j\}$, and that one can obtain some $\alpha_{ij} = \ln \gamma_{ij}$ to represent the preference probabilities given by the model. The paper also discusses how to include these order effects into both the Rao-Kupper and the Davidson tie models.

Again, it should be noted that the multiplicative model has become the standard model for measuring order effects in the literature, and is the model that we plan to adapt and extend upon.

1.5 Log-Linear Parametrisation

Perhaps the most important development in the field of Bradley-Terry models is that of the shift to a log-linear parametrisation of the model. This change provides many benefits, but one key point is that this shift has allowed us to take the results we know hold true for logistic regression, and apply them to this model, allowing us to use a more general and better developed framework.

Bradley & Terry (1952) does give mention to the fact that the $\log \pi_i$s are actually comparable on a linear scale, and even suggests that this is the scale upon which all comparisons of difference should be made. However, despite this, Bradley and Terry do not propose or give consideration to a full log-linear parametrisation of their model. The first record of formulating the model as the log odds of two team's strengths appears in Cox (1970), where it is shown that

$$\log\left(\frac{\pi_i}{\pi_j}\right) = \theta_i - \theta_j$$

can be used. Throughout the 1970s, the log-linear parametrisation gained some traction, with papers such as Atkinson (1972), but still appears to be considered as almost an alternative by the bulk of the literature - perhaps due to a lack of understanding by those applying the model, as noted by Maxwell (1974). However by the 1990s, the log-linear parametrisation had become the preferred presentation of the Bradley-Terry model, as it remains today.

It is of note that many of the nice things previously incorporated into the model, such as ties and external covariates, may now simply be considered as linear terms of the log-linear model. For instance, when considering some home ground advantage α, we now have that

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i - \theta_j + \alpha$$

using the previously defined $\theta_i = \ln \pi_i$ and $\alpha = \ln \gamma$, and where p_{ij} denotes the probability that team i beats j at home. This model also allows us to include factors such as external covariates, as discussed below.

1.6 Modelling External Covariates

Another interesting extension upon the model is the ability to incorporate external covariates into the strengths of items - for instance, in Dittrich et al. (1998), which concerns modelling external covariates with an application to University Rankings. In said example, covariates are added to consider factors specific to relationships between a university and particular students, such as the ability of a student to speak the university's local language. To relate back to our sporting examples, elements of a contest such as the weather - for example, the strength of a team when it is raining compared to when it is not - could be modelled in this way.

The modelling of external covariates appears to have been first touched upon by Critchlow & Fligner (1991), who make use of the aforementioned log-linear parametrisation of the Bradley-Terry model where $\theta_i = \ln \pi_i$. They consider that we can have

$$\theta_i = \sum_{j=1}^{p} c_{ij}\beta_j, \quad \text{for } i = 1, \ldots, t,$$

with unknown parameters $\beta_1, \ldots, \beta_p, p \leq t$ and c_{ij} representing the known value of the j-th covariate for item i. They note that c_{ij} may take on the form on an indicator variable. The paper by Critchlow and Fligner discusses an example of these covariates for comparing taste samples with covariates c_{i1} for "Flavour concentration", and c_{i2} for "Gel concentration". However, I would like to consider another example perhaps more pertinent to our interests in applying the Bradley-Terry model to sporting pursuits.

One interesting debate in Australian Rules Football is whether teams should play with two dedicated ruckmen, as is traditional, allowing one to be replaced by a specialist ruckman while resting, or whether to rotate through a tall forward instead (Cleary (2020)). This could potentially be investigated using an external covariate extension to the Bradley-Terry model with an indicator variable c_{i1} representing whether team i is playing 2 specialist ruckmen or not, thus leading to a model with

$$\theta_i = \beta_0 + c_{i1}\beta_1$$

and a null hypothesis of $\beta_1 = 0$, where here the β_0 represents a teams underlying strength without the additional covariate. This model could also be used to help explain external phenomena in contests such as the effects of weather, accounting for teams abilities in the wet or when playing at night.

Critchlow & Fligner (1991) propose that one can use the GLIM package to fit these models - a process which became popular in the literature of the time - given that in the log-linear parametrisation it is akin to fitting a logistic regression. The modern R package, discussed below, uses the penalised quasi-likelihood algorithm of Breslow & Clayton (1993) to fit the model in this manner.

1.7 Packages in R

Given that part of our interest in studying this field is to be able to produce an R Package suitable to our goals, it is important to survey what work has already been done in this domain. In the literature, there exist only 2 popular R packages for formulating the Bradley-Terry model - **BradleyTerry** by Firth (2005), and its follow-up **BradleyTerry2** by Turner & Firth (2012).

1.7.1 BradleyTerry

The first package was published in 2005 by David Firth, and was capable of handling a number of the previously mentioned extensions to the model, albeit not all. The package's capabilities include:

- Formulating a Bradley-Terry model with $\lambda_1 = 0$, and by default using the MLE to estimate the other λ_is
- Ability to model linear predictors (as mentioned in Section 6 above)
- A single model specific order effect α (as per Section 4)
- Produce statistics regarding the model, such as residuals of pairings, as well as standard errors, Z-scores and p-values for team strengths
- Ability to fit the model using a bias-reduced Maximum Likelihood proposed in Firth (1993) as an alternative

1.7.2 BradleyTerry2

An update to the package was published in 2012, with a few minor additions. These additions include:

- The ability to use a probit or a cauchit formulation of the model, rather than the logit formulation
- The inclusion of modelling prediction errors U_i, such that $\lambda_i = \sum_r \beta_r x_{ir} + U_i$ and $U_i \sim N(0, \sigma^2)$. It would seem that σ^2 represents an estimate of the variance in an item's strength, and is common across all items
- Incorporating more contest specific predictors, which could implement any factor that affects a specific contest

1.8 Our Research

Our main interest regarding this problem is to extend the model's ability to measure varying order effects, which we will refer to from here on in as ‘homeground advantages’ given our particular interest in sporting competitions such as the AFL and the NBA. As discussed above, we do not believe in the assumption proposed by Davidson & Beaver (1977) that $\gamma_{ij} = \gamma_{ji}$ should necessarily hold. As such, we consider a model where they are unequal, and with a unique γ_{ij} for every possible combination of teams to be our largest model. Also, when considering that current packages only allow for a single model specific order effect α, we think that it is important to be able to handle more complex interpretations of homeground advantage.

Of particular interest to us is to consider ‘Hierarchies’ within competitions. These sorts of structures are typical of American sporting competitions, which are often quite large - for instance, the NFL and NHL consist of 32 teams, while the NBA and MLB each consist of 30 teams. To handle this, teams are often split into conferences - sometimes using the geographical ‘East-West’ system such as in the NBA and the NHL, and sometimes using historical groupings such as the ‘American’ and ‘National’ leagues in both the MLB and the NHL. Within a conference, each of these four leagues are split again into ‘Divisions’, grouping teams which share a similar geographical area. For example, the New York Knicks belong to the Atlantic *Division*, which is one of the divisions that makes up the Eastern *Conference* of the NBA.

Considering these structures, we propose that it is both possible and valuable to be able to model differing homeground advantages based off the relationship between two teams - whether they are in the same division, the same conference but not the same division, or whether they are in opposing conferences. It is also worth noting that this sort of hierarchical model need not be restricted only to the ‘Conference + Division’ style setup that is found in America, but could also be used in any other setting where groupings of items can be made. For instance, an analysis of European football may involve considering a number of editions of the UEFA Champions League and Europa Leagues, with a simple overlay of which country teams belong to, as it would be expected that a team may have a differing level of advantage if it is playing a club from the same nation compared to when it plays a foreign club.

We would also like to include some other simple extensions on the homeground advantage that are not currently implemented in R, allowing us to consider a number of non-hierarchical models also, and to be able to perform hypothesis tests as to which models may provide a significantly better fit. These include the pairwise-unique model discussed above, with an α_{ij} for each pairing of home team i with away team j, but also the smaller model of giving each team a unique homeground advantage α_i which is common against all opponents. This will also allow us to test hypotheses such as whether the hierarchical structure of a competition does have a significant impact on the advantage a team carries when playing at home.

We do not intend to incorporate a way of predicting ties, nor a way of modelling external covariates in our research. For the former, we know that in the sports we are primarily aiming to model that draws are either non-existent (i.e. in Basketball[2]) or incredibly rare (i.e. in Australian Football[3]), and as such it is not of great interest to us to be able to predict such outcomes. For the latter, it is simply a case where we believe that a substantial amount of research has already been done in this area, to the extent that we believe there are much better results to be found in considering the homeground advantages teams may have, which until now appear to have been largely ignored. However, we are of the opinion that they are still both interesting and important, and including these in a future extension to our R package is certainly in our interests.

[2]In professional Basketball, extra time is played until a result transpires

[3]Only 162 draws have been played since the advent of the VFL in 1897, and only 9 draws from 2012 to 2019, the only completed seasons since the introduction of the GWS Giants which lead to the current season length of 205 games including finals. As such, 9 draws over 8 seasons represents only 0.5% of all AFL matches. At time of publishing, a list of all draws played out in VFL/AFL history could be found at `https://afltables.com/afl/stats/biglists/bg8.txt`

Chapter 2

Model

Having now considered the current state of the literature and what we would like to contribute to it, we need to introduce the models that we intend on making use of. This chapter first provides a summary of important definitions to be used throughout the paper. Each of the models we have chosen to consider are then presented, considering important information such as how each individual comparison is modelled, the likelihood function of said model, as well as more technical aspects such as formulas for each element of the Fisher-Information matrices. This chapter then concludes with a handful of short proofs which provide some useful properties for these models.

2.1 Common Notation

Whilst some of these may have been introduced throughout Chapter 1, they are restated here so as to provide a complete list of all symbols used throughout the paper.

Data

- $W -$ The matrix of home wins for a team, defined such that W_{ij} is the number of wins that team i has achieved at home over team j.
- $L -$ The matrix of home losses for a team, defined such that L_{ij} is the number of losses that team i has sustained at home against team j.
- $M -$ The matrix of home matches played between each pair, defined such that M_{ij} is the number of times team i has hosted team j, with the restriction that $M = W + L$
- $T -$ The number of teams in a given model. As such, each of the W, L and M matrices are to be $T \times T$ matrices.
- $\mathcal{H} -$ The set of hierarchies in a given model, with components $H_1 \subset H_2 \subset \ldots \subset H_N$
- $N -$ the number of levels in any hierarchical system $\mathcal{H}$
- $R -$ the relationship matrix, such that R_{ij} represents the highest level of the hierarchy $\mathcal{H}$ that the relationship between two teams i and j belongs to, characterised by $\min_{k<N} \{(i, j) \in H_k\}$

Events and Parameters

- $(i > j)_i -$ The event that team i beats team j at home.
- $\lambda_i -$ The 'strength' of team i
- $\theta_i -$ The 'log-strength' of team i, given by $\theta_i = \ln \lambda_i$
- $\gamma_m -$ The 'strength' of homeground advantage m. This subscript could pertain to any number of things, as discussed throughout the paper, and indeed may have multiple subscripts at times.
- $\alpha_m -$ The 'log-strength' of homeground advantage m, given by $\alpha_m = \ln \gamma_m$

2.2 Implemented Models

A number of variations on the Bradley-Terry model are considered for this paper, and indeed are all implemented in the R package which complements it. As can be seen, they form a series of nested models considering different treatments of the homeground advantages - essentially considering different possible ways to subscript γ and α to represent different levels of specifying these advantages.

For each of these models, we present the following attributes:

- Number of Parameters
- Characterisation of $\mathbb{P}(i > j)$
- Likelihood Function
- Log-Likelihood Function
- Gradient Functions
- Second-Order derivatives, for forming $\mathcal{I}$

2.2.1 Model 0 - Vanilla Bradley-Terry

The base model to be implemented is that proposed by Bradley and Terry in their original paper.
This model has T parameters - a λ_i for each team, such that

$$\mathbb{P}(i > j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

This model has likelihood function

$$L(\lambda|W, L) = \prod_{i=1}^{T} \prod_{j=1}^{T} \binom{M_{ij}}{W_{ij}} \left(\frac{\lambda_i}{\lambda_i + \lambda_j}\right)^{W_{ij}} \left(\frac{\lambda_j}{\lambda_i + \lambda_j}\right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\lambda|W,L) = \sum_{i=1}^{T}\sum_{j=1}^{T} W_{ij}\theta_i + L_{ij}\theta_j - M_{ij}\log\left(e^{\theta_i} + e^{\theta_j}\right) + \text{constant}$$

Thus, we need to consider all partial derivatives with respect to each θ_i, all of which are of the form

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - (M_{ij} + M_{ji})\left(\frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}\right)$$

From, this we have both

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\sum_{j=1}^{T}(M_{ij} + M_{ji})\left(\frac{e^{\theta_i+\theta_j}}{(e^{\theta_i} + e^{\theta_j})^2}\right)$$

and for $i \neq j$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = (M_{ij} + M_{ji})\left(\frac{e^{\theta_i+\theta_j}}{(e^{\theta_i} + e^{\theta_j})^2}\right)$$

2.2.2 Model 1 - Common Homeground

This model has $T+1$ parameters - T many λs and a single common γ. This is the homeground model included in **BradleyTerry2**. The model is characterised by

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma}{\lambda_i \gamma + \lambda_j}$$

and

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma + \lambda_j}$$

This yields the likelihood function

$$L(\lambda, \gamma|W,L) = \prod_{i=1}^{T}\prod_{j=1}^{T}\binom{M_{ij}}{W_{ij}}\left(\frac{\lambda_i \gamma}{\lambda_i \gamma + \lambda_j}\right)^{W_{ij}}\left(\frac{\lambda_j}{\lambda_i \gamma + \lambda_j}\right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\theta, \alpha | W, L) = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij}(\theta_i + \alpha) + L_{ij}\theta_j - M_{ij} \log \left(e^{\theta_i + \alpha} + e^{\theta_j} \right) + \text{constant}$$

Thus, considering the partials with respect to each θ_i, we have

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - \left(\frac{M_{ij} e^{\theta_i + \alpha}}{e^{\theta_j} + e^{\theta_i + \alpha}} + \frac{M_{ji} e^{\theta_i}}{e^{\theta_i} + e^{\theta_j + \alpha}} \right)$$

and for α

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij} - \frac{M_{ij} e^{\theta_i + \alpha}}{e^{\theta_i + \alpha} + e^{\theta_j}}$$

From this, we have second order derivatives of:

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = - \sum_{j=1}^{T} e^{\theta_i + \theta_j + \alpha} \left(\frac{M_{ij}}{(e^{\theta_i + \alpha} + e^{\theta_j})^2} + \frac{M_{ji}}{(e^{\theta_j + \alpha} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = e^{\theta_i + \theta_j + \alpha} \left(\frac{M_{ij}}{(e^{\theta_i + \alpha} + e^{\theta_j})^2} + \frac{M_{ji}}{(e^{\theta_j + \alpha} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha} = \sum_{j=1}^{T} e^{\theta_i + \theta_j + \alpha} \left(\frac{M_{ji}}{(e^{\theta_j + \alpha} + e^{\theta_i})^2} - \frac{M_{ij}}{(e^{\theta_i + \alpha} + e^{\theta_j})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \alpha^2} = - \sum_{i=1}^{T} \sum_{j=1}^{T} \frac{M_{ij} e^{\theta_i + \theta_j + \alpha}}{(e^{\theta_i + \alpha} + e^{\theta_j})^2}$$

where $i \neq j$

2.2.3 Model 2 - Common Hierarchical Homeground

This model has $T + N$ parameters - T many λs and N many γs. We refer to these γs as being γ_n, where $R_{ij} = n$ represents the level of the hierarchy to which the relationship between the two teams being considered belongs. As such, when team i competes against team j, we know that the element

R_{ij} contains the level of the hierarchy to which the relationship between two teams belongs, and thus the homeground advantage that the home team would possess in these matchups is $\gamma_{R_{ij}}$. Thus, the model can be characterised by:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_{R_{ij}}}{\lambda_i \gamma_{R_{ij}} + \lambda_j}$$

and

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{R_{ij}} + \lambda_j}$$

This yields the likelihood function

$$L(\lambda, \gamma | W, L) = \prod_{i=1}^{T} \prod_{j=1}^{T} \binom{M_{ij}}{W_{ij}} \left(\frac{\lambda_i \gamma_{R_{ij}}}{\lambda_i \gamma_{R_{ij}} + \lambda_j} \right)^{W_{ij}} \left(\frac{\lambda_j}{\lambda_i \gamma_{R_{ij}} + \lambda_j} \right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\theta, \alpha | W, L) = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij}(\theta_i + \alpha_{R_{ij}}) + L_{ij}\theta_j - M_{ij} \log \left(e^{\theta_i + \alpha_{R_{ij}}} + e^{\theta_j} \right) + \text{constant}$$

Thus, considering the partials with respect to each θ_i, we have

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - \left(\frac{M_{ij} e^{\theta_i + \alpha_{R_{ij}}}}{e^{\theta_j} + e^{\theta_i + \alpha_{R_{ij}}}} + \frac{M_{ji} e^{\theta_i}}{e^{\theta_i} + e^{\theta_j + \alpha_{R_{ij}}}} \right)$$

and for each $n \in \{1, \dots, N\}$

$$\frac{\partial \ell}{\partial \alpha_n} = \sum_{i=1}^{T} \sum_{j=1}^{T} \left(W_{ij} - \frac{M_{ij} e^{\theta_i + \alpha_{R_{ij}}}}{e^{\theta_i + \alpha_{R_{ij}}} + e^{\theta_j}} \right) \mathbb{I}(R_{ij} = n)$$

From this, we have second order derivatives of:

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\sum_{j=1}^{T} e^{\theta_i+\theta_j+\alpha_{R_{ij}}} \left(\frac{M_{ij}}{(e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j})^2} + \frac{M_{ji}}{(e^{\theta_j+\alpha_{R_{ij}}} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = e^{\theta_i+\theta_j+\alpha_{R_{ij}}} \left(\frac{M_{ij}}{(e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j})^2} + \frac{M_{ji}}{(e^{\theta_j+\alpha_{R_{ij}}} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha_n} = \sum_{j=1}^{T} e^{\theta_i+\theta_j+\alpha_{R_{ij}}} \left(\frac{M_{ji}}{(e^{\theta_j+\alpha_{R_{ij}}} + e^{\theta_i})^2} - \frac{M_{ij}}{(e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j})^2} \right) \mathbb{I}(R_{ij} = n)$$

$$\frac{\partial^2 \ell}{\partial \alpha_n \partial \alpha_m} = 0$$

$$\frac{\partial^2 \ell}{\partial \alpha_n^2} = -\sum_{i=1}^{T} \sum_{j=1}^{T} \frac{M_{ij} e^{\theta_i+\theta_j+\alpha_{R_{ij}}}}{(e^{\theta_i+\alpha_{R_{ij}}} + e^{\theta_j})^2} \mathbb{I}(R_{ij} = n)$$

where $i \neq j$, and $n \neq m$

2.2.4 Model 3 - Team-Specific Homeground

This model has $2T$ parameters - a λ and a γ for each possible T. The model is characterised by

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_i}{\lambda_i \gamma_i + \lambda_j}$$

and

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_i + \lambda_j}$$

Thus, we have the likelihood function

$$L(\lambda, \gamma | W, L) = \prod_{i=1}^{T} \prod_{j=1}^{T} \binom{M_{ij}}{W_{ij}} \left(\frac{\lambda_i \gamma_i}{\lambda_i \gamma_i + \lambda_j} \right)^{W_{ij}} \left(\frac{\lambda_j}{\lambda_i \gamma_i + \lambda_j} \right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\theta, \alpha | W, L) = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij}(\theta_i + \alpha_i) + L_{ij}\theta_j - M_{ij} \log \left(e^{\theta_i+\alpha_i} + e^{\theta_j} \right) + \text{constant}$$

Thus, considering the partials with respect to each θ_i we have

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - \left(\frac{M_{ij} e^{\theta_i + \alpha_i}}{e^{\theta_i + \alpha_i} + e^{\theta_j}} + \frac{M_{ji} e^{\theta_i}}{e^{\theta_j + \alpha_j} + e^{\theta_i}} \right)$$

and for each α_i

$$\frac{\partial \ell}{\partial \alpha_i} = \sum_{j=1}^{T} W_{ij} - \frac{M_{ij} e^{\theta_i + \alpha_i}}{e^{\theta_i + \alpha_i} + e^{\theta_j}}$$

From this, we have second order derivatives of:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta_i^2} &= - \sum_{j=1}^{T} e^{\theta_i + \theta_j} \left(\frac{M_{ij} e^{\alpha_i}}{(e^{\theta_i + \alpha_i} + e^{\theta_j})^2} + \frac{M_{ji} e^{\alpha_j}}{(e^{\theta_j + \alpha_j} + e^{\theta_i})^2} \right) \\
\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} &= e^{\theta_i + \theta_j} \left(\frac{M_{ij} e^{\alpha_i}}{(e^{\theta_i + \alpha_i} + e^{\theta_j})^2} + \frac{M_{ji} e^{\alpha_j}}{(e^{\theta_j + \alpha_j} + e^{\theta_i})^2} \right) \\
\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha_i} &= - \sum_{j=1}^{T} \left(\frac{M_{ij} e^{\theta_i + \theta_j + \alpha_i}}{(e^{\theta_i + \alpha_i} + e^{\theta_j})^2} \right) \\
\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha_j} &= \frac{M_{ji} e^{\theta_i + \theta_j + \alpha_j}}{(e^{\theta_j + \alpha_j} + e^{\theta_i})^2} \\
\frac{\partial^2 \ell}{\partial \alpha_i \partial \alpha_j} &= 0 \\
\frac{\partial^2 \ell}{\partial \alpha_i^2} &= - \sum_{j=1}^{T} \frac{M_{ij} e^{\theta_i + \theta_j + \alpha_i}}{(e^{\theta_i + \alpha_i} + e^{\theta_j})^2}
\end{aligned}$$

where $i \neq j$

2.2.5 Model 4 - Hierarchical Model

This model has $T(N+1)$ parameters - a λ for each team, and a γ for each team and each level of the hierarchy. We index the γs first by team, and then by hierarchy, such that we have $\gamma_{i,n}$ representing the homeground advantage afforded to team i when competing against teams with which their relationship belongs to level n of the hierarchy. As a result, this model is characterised by:

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_{i,R_{ij}}}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j}$$

and

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j}$$

This yields the likelihood function

$$L(\lambda, \gamma | W, L) = \prod_{i=1}^{T} \prod_{j=1}^{T} \binom{M_{ij}}{W_{ij}} \left(\frac{\lambda_i \gamma_{i,R_{ij}}}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j} \right)^{W_{ij}} \left(\frac{\lambda_j}{\lambda_i \gamma_{i,R_{ij}} + \lambda_j} \right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\theta, \alpha | W, L) = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij}(\theta_i + \alpha_{i,R_{ij}}) + L_{ij}\theta_j - M_{ij} \log \left(e^{\theta_i + \alpha_{i,R_{ij}}} + e^{\theta_j} \right) + \text{constant}$$

Thus, considering the partials with respect to each θ_i, we have

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - \left(\frac{M_{ij} e^{\theta_i + \alpha_{i,R_{ij}}}}{e^{\theta_j} + e^{\theta_i + \alpha_{i,R_{ij}}}} + \frac{M_{ji} e^{\theta_i}}{e^{\theta_i} + e^{\theta_j + \alpha_{j,h(j,i)}}} \right)$$

and for each $n \in \{1, \ldots, N\}$

$$\frac{\partial \ell}{\partial \alpha_{i,n}} = \sum_{j=1}^{T} \left(W_{ij} - \frac{M_{ij} e^{\theta_i + \alpha_{i,R_{ij}}}}{e^{\theta_i + \alpha_{i,R_{ij}}} + e^{\theta_j}} \right) \mathbb{I}(R_{ij} = n)$$

From this, we have second order derivatives of:

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\sum_{j=1}^{T} e^{\theta_i+\theta_j} \left(\frac{M_{ij} e^{\alpha_{i,R_{ij}}}}{(e^{\theta_i+\alpha_{i,R_{ij}}} + e^{\theta_j})^2} + \frac{M_{ji} e^{\alpha_{j,R_{ij}}}}{(e^{\theta_j+\alpha_{j,R_{ij}}} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = e^{\theta_i+\theta_j} \left(\frac{M_{ij} e^{\alpha_{i,R_{ij}}}}{(e^{\theta_i+\alpha_{i,R_{ij}}} + e^{\theta_j})^2} + \frac{M_{ji} e^{\alpha_{j,R_{ij}}}}{(e^{\theta_j+\alpha_{j,R_{ij}}} + e^{\theta_i})^2} \right)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha_{i,n}} = -\sum_{j=1}^{T} \left(\frac{M_{ij} e^{\theta_i+\theta_j+\alpha_{i,R_{ij}}}}{(e^{\theta_i+\alpha_{i,R_{ij}}} + e^{\theta_j})^2} \right) \mathbb{I}(R_{ij} = n)$$

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \alpha_{j,n}} = \frac{M_{ji} e^{\theta_i+\theta_j+\alpha_{j,R_{ij}}}}{(e^{\theta_j+\alpha_{j,R_{ij}}} + e^{\theta_i})^2} \mathbb{I}(R_{ij} = n)$$

$$\frac{\partial^2 \ell}{\partial \alpha_{i,n} \partial \alpha_{k,m}} = 0$$

$$\frac{\partial^2 \ell}{\partial \alpha_{i,n}^2} = -\sum_{j=1}^{T} \frac{M_{ij} e^{\theta_i+\theta_j+\alpha_{i,R_{ij}}}}{(e^{\theta_i+\alpha_{i,R_{ij}}} + e^{\theta_j})^2} \mathbb{I}(R_{ij} = n)$$

where $i \neq j$, and that at least one of $m \neq n$ or $i \neq k$ holds.

2.2.6 Model 5 - Pairwise Unique Homeground

This model has $T(T+1)$ parameters - T many λs, and then a γ for each possible pairing of teams - T^2 of them. The model is characterised by

$$\mathbb{P}(i > j)_i = \frac{\lambda_i \gamma_{ij}}{\lambda_i \gamma_{ij} + \lambda_j}$$

and

$$\mathbb{P}(i < j)_i = \frac{\lambda_j}{\lambda_i \gamma_{ij} + \lambda_j}$$

Thus, we have the likelihood function

$$L(\lambda, \gamma | W, L) = \prod_{i=1}^{T} \prod_{j=1}^{T} \binom{M_{ij}}{W_{ij}} \left(\frac{\lambda_i \gamma_{ij}}{\lambda_i \gamma_{ij} + \lambda_j} \right)^{W_{ij}} \left(\frac{\lambda_j}{\lambda_i \gamma_{ij} + \lambda_j} \right)^{L_{ij}}$$

and thus log-likelihood function

$$\ell(\theta, \alpha | W, L) = \sum_{i=1}^{T} \sum_{j=1}^{T} W_{ij}(\theta_i + \alpha_{ij}) + L_{ij}\theta_j - M_{ij} \log \left(e^{\theta_i + \alpha_{ij}} + e^{\theta_j} \right) + \text{constant}$$

and so we have the following sets of partial derivatives:

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{j=1}^{T} W_{ij} + L_{ji} - \left(\frac{M_{ij} e^{\theta_i + \alpha_{ij}}}{e^{\theta_j} + e^{\theta_i + \alpha_{ij}}} + \frac{M_{ji} e^{\theta_i}}{e^{\theta_i} + e^{\theta_j + \alpha_{ji}}} \right)$$

and

$$\frac{\partial \ell}{\partial \alpha_{ij}} = W_{ij} - \frac{M_{ij} e^{\theta_i + \alpha_{ij}}}{e^{\theta_j} + e^{\theta_i + \alpha_{ij}}}$$

Given how large this model is, and how unlikely it is to provide a meaningful interpretation, we have decided not to incorporate a method for calculating standard errors for this model, and so have no need for finding the second order partial derivatives.

2.3 Proofs

2.3.1 Win and Loss Matrices as Sufficient Statistics

Considering the above formulations for each model, we intend to make use of the factorisation theorem to show that the W and L matrices can combine to be a sufficient statistic for each model. Indeed, if we form an augmented matrix $\mathbf{x} = [W, L]$, we can indeed use a single matrix made up of these two as the sufficient statistic for the model. Considering this case, we should note that $L_{ij} = \mathbf{x}_{i,\ j+T}$.

Another point worth contemplating is that showing this property for the largest model, the pairwise unique homeground model, is sufficient to show that it is true for each smaller model. For instance, we can see that we also attain the unique homeground model by placing a restriction on the values of γ such that each $\gamma_{i1} = \gamma_{i2} = \ldots = \gamma_{iT}$ for each team i. Similar restrictions can be imposed for each of the other models proposed. As such, it remains to show that the likelihood of the pairwise model can be represented in the form

$$L(\mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x});\theta)$$

Indeed, if we consider

$$\begin{aligned}
\mathbf{x} &= [W, L] \\
h(\mathbf{x}) &= \prod_{i=1}^{T}\prod_{j=1}^{T}\binom{\mathbf{x}_{i,\ j} + \mathbf{x}_{i,\ j+T}}{\mathbf{x}_{i,\ j}} \\
T(x) &= \mathbf{x} \\
g(T(x), \lambda, \gamma) &= \prod_{i=1}^{T}\prod_{j=1}^{T}\left(\frac{\lambda_i\gamma_{ij}}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{T(x)_{i,j}}\left(\frac{\lambda_j}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{T(x)_{i,\ j+T}}
\end{aligned}$$

we can quite clearly see that:

$$\begin{aligned}
h(\mathbf{x})g(T(\mathbf{x});\theta) &= \left(\prod_{i=1}^{T}\prod_{j=1}^{T}\binom{\mathbf{x}_{i,j} + \mathbf{x}_{i,j+T}}{\mathbf{x}_{i,j}}\right) \times \\
&\quad \left(\prod_{i=1}^{T}\prod_{j=1}^{T}\left(\frac{\lambda_i\gamma_{ij}}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{T(x)_{i,j}}\left(\frac{\lambda_j}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{T(x)_{i,\ j+T}}\right) \\
&= \prod_{i=1}^{T}\prod_{j=1}^{T}\binom{M_{ij}}{W_{ij}}\left(\frac{\lambda_i\gamma_{ij}}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{W_{ij}}\left(\frac{\lambda_j}{\lambda_i\gamma_{ij} + \lambda_j}\right)^{L_{ij}} \\
&= f(W, L|\lambda, \gamma)
\end{aligned}$$

2.3.2 Asymptotic Normality of the MLE

So as to be able to provide meaningful confidence intervals for our estimates, we wish to show that the errors of our maximum likelihood estimates should converge towards a normal distribution as we increase the number of matches in our data, i.e.

$$\sqrt{N}\left((\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})^\top - (\boldsymbol{\theta}, \boldsymbol{\alpha})^\top\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\boldsymbol{\theta}, \boldsymbol{\alpha})\right)$$

where N represents the number of matches played in the set of data.

Indeed, as this is a common property of many MLEs, and that many textbooks include a proof of this property, we shall not include a full proof but instead will simply show that the preconditions of this statement hold. This can, for instance, be found as Theorem 5.3.3 in Bickel & Doksum (2001). Here, they state that this property holds given the following conditions (with some notation changed to suit our above notation):

1. $\mathcal{P}$ is a canonical exponential family
2. The parameter space $\mathcal{E}$ is open
3. $X_1, \ldots, X_N$ are an independent sample from $\mathcal{P}_{\boldsymbol{\theta}, \boldsymbol{\alpha}} \in \mathcal{P}$
4. $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})$ is defined as the MLE

These all hold trivially for our model. The first condition holds, as we've seen previously that the likelihood function for each of the models is a product of many binomial distributions. As the binomial distribution is an exponential family, so too are each of our models. We also can see that as our parameter space for $(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is $\mathbb{R}^d$, where d is the number of parameters being estimated in any given model, that indeed we have an open parameter space. The final condition obviously holds too, as the idea behind this entire principle is that is in fact the MLEs we are considering.

Condition 3, perhaps surprisingly the most complex condition to examine, simply considers that our data is an independent sample from the model, and that the model is an exponential family with (fixed) parameters $(\boldsymbol{\theta}, \boldsymbol{\alpha})$ - this is perhaps the most difficult assumption to consider, but it is primarily a qualitative argument regarding what period of time the set of parameters may stay fixed for. As such, a discussion of this assumption can be found in Chapter 5. It does also state that the sample needs to be independent, and whilst we do not provide a full discussion of this, this a very popular topic of debate in statistical circles. Whilst the general consensus does appear to lean towards the idea that form and 'hot streaks' are nothing more than a fallacy, it is certainly not a unanimous opinion. Some papers worth considering for debate on this topic include the works of Steeger et al. (2021), Bar-Eli et al. (2006), Dorsey-Palmateer & Smith (2004), Hales (1999) and Gilovich et al. (1985). It could also be discussed whether the situation of a team within the season is of note - does a team stop playing at full strength once they've reached however many wins they need to make playoffs, or perhaps if they experience a poor start to the season would they write-off any hope at a championship and opt to provide experience to weaker players. Whilst not discussed above, it is another avenue of investigation that could potentially give us some insight on this assumption.

2.3.3 Hypothesis Testing of Nested Models

Another important property that we would like to have is the ability to perform hypothesis tests between these models, so as to determine whether there is significant evidence that using a more complex model provides a better fit. In particular, we want to take advantage of Wilks' Theorem, which gives us that for a null model M_0, the parameters of which are a subset of the parameters of the alternative model M_1, that

$$-2\left(\ell_{M_1}(\mathbf{X}|\hat{\boldsymbol{\theta}}_1) - \ell_{M_0}(\mathbf{X}|\hat{\boldsymbol{\theta}}_0)\right) \xrightarrow{d} \chi^2_{n-m}$$

where n and m are the dimension of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ respectively. Indeed, Bickel & Doksum (2001) state that if the assumptions for the previous theorem hold, so too do the assumptions for this theorem. As such, we are able to use such a technique to compare our nested models.

Chapter 3

R Implementation

With the models defined, and a number of nice properties provided thanks to the proofs given, we now wish to discuss the package which has been created to complement this paper. First, we provide a short discussion on how the models are fitted in the package. This is followed by outlining some of the extra functions which are included as a part of the package, and finally a short section benchmarking this new package against the current state-of-the-art **BradleyTerry2**. Access to the package in it's form at the time of this submission can be found on Github.[1]

3.1 Overview

The framework for this module is contained within two files - `functions.r` and `module.r`. The former of these consists mainly of functions for performing grunt work, which are then called by functions in `module.r`, which contains the functions which are intended to be operated by the user. The linked repository includes a number of other useful tools and resources, including data and code for all the examples considered throughout this paper.

[1] `https://github.com/coenjones/Honours`

3.2 Fitting the Model

At present, the function `BTC()`, found in `module.r`, fits a Bradley-Terry model of a given type using a `list` consisting of (at least) the win and loss matrices. This model then returns estimates of the parameters, along with their associated standard errors. This fitting is done by feeding the Log-Likelihood function and its associated gradients into R's inbuilt optimiser `optim()`, while the standard errors for most models are evaluated by taking the inverse Fisher information matrix. For the sakes of stepping through how this process works, we will take a nice small example, that being the 1916 VFL Season as recorded by Rogers & Brown (1998). Due to the First World War, only 4 of the 10 VFL teams of the day contested the season - Carlton, Collingwood, Fitzroy and Richmond - playing each other twice at home and twice away for a grand total of 24 matches.[2]

3.2.1 Inputs

The `BTC()` function takes 4 inputs

- 'data' – the list of matrices, with data$w representing the home-wins matrix, and data$l representing the home-loss matrix.
- 'mod' – an integer representing which model type, as specified in Chapter 2, that the user wishes to utilise.
- 'rel' – the R matrix for fitting a hierarchical model.
- 'ref' – the reference team whose value for θ will be set to 0 so as to specify the model. This defaults to the weakest team in the data, so that all other teams are positive and easily compared to the lowest value.

[2]A fun fact regarding the 1916 season is that all 4 teams contested the finals series so that the VFL could continue using the Top 4 finals system which had been in place since the league's 1897 season. As a result, Fitzroy, who finished bottom of the ladder with only 2 wins, finished the season as premiers. To date, they are the only club to claim both the 'wooden spoon' and the premiership in the same season - and should forever remain so.

It is acknowledged that sporting data is unlikely to come in the form of the matrices required. Indeed, the data used throughout this paper comes in a form similar to below, where there are columns specifying the home and away teams, as well as their scores.

	home.team	away.team	home_scores	away_scores
1	Carlton	Fitzroy	60	64
2	Richmond	Collingwood	69	77
3	Fitzroy	Richmond	66	37
4	Collingwood	Carlton	45	71
5	Fitzroy	Collingwood	50	50
6	Richmond	Carlton	61	66
7	Collingwood	Richmond	78	53
8	Fitzroy	Carlton	51	84
9	Carlton	Collingwood	75	76
10	Richmond	Fitzroy	127	68
11	Collingwood	Fitzroy	79	65
12	Carlton	Richmond	92	31
13	Collingwood	Richmond	74	75
14	Fitzroy	Carlton	50	79
15	Carlton	Collingwood	65	53
16	Richmond	Fitzroy	65	55
17	Collingwood	Fitzroy	59	58
18	Carlton	Richmond	91	41
19	Carlton	Fitzroy	81	62
20	Richmond	Collingwood	93	67
21	Fitzroy	Richmond	65	75
22	Collingwood	Carlton	70	72
23	Fitzroy	Collingwood	57	75
24	Richmond	Carlton	65	82

Figure 3.1: Raw 1916 VFL results for processing

The function `Data2Mat()` takes a series of results in the above form, and returns the matrices required. For instance, when applied to this dataset, we are given the W matrix shown in Figure 3.2, as well as its respective L matrix in Figure 3.3.

As we can see, draws are indeed represented by a value of 0.5 in the data as previously discussed. Fitzroy's record against Collingwood at home - a draw in game 5, and a loss in game 23 - is reflected by the fact that $W_{\text{Fitzroy, Collingwood}} = 0.5$.

```
> mat = Data2Mat(data)
> mat$w
            Carlton Richmond Fitzroy Collingwood
Carlton           0        2       1         1.0
Richmond          0        0       2         1.0
Fitzroy           0        1       0         0.5
Collingwood       0        1       2         0.0
```

Figure 3.2: Home Wins Matrix from 1916 VFL Data

```
> mat$l
            Carlton Richmond Fitzroy Collingwood
Carlton           0        0       1         1.0
Richmond          2        0       0         1.0
Fitzroy           2        1       0         1.5
Collingwood       2        1       0         0.0
```

Figure 3.3: Home Losses Matrix from 1916 VFL Data

3.2.2 Process

As stated previously, we use the log-likelihood, the score vector and the Fisher matrices to fit our model. In particular, one can see that the bulk of the `functions.r` file consists of these functions for each of the models. Indeed, they bear the names `LLx`, `LLxdash` and `Hessx`, where 'x' is replaced by the integer representing the particular model.[3]

To fit the model, we take an initial guess using the `runif` function. Whilst this does add some level of randomness to the process, it can be considered quite negligble - for instance, the difference in Theta's over 100 fits of the 1916 VFL data with a vanilla model and Fitzroy as reference yields a difference between the maximum and minimum estimates for each θ on the scale of 10^{-5}, and standard deviations on the order of 10^{-6}, suggesting that these estimates should be consistent to at least 4 decimal places.

Once the parameters have been estimated and scaled to the reference, `BTC()` evaluates the negaitve Hessian with respect to the estimates and

[3] i.e. when fitting a Team-Specific Homeground model, `BTC()` will call `LL3()`, `LL3dash()` and `Hess3()`

provided data, removes the column and row relating to the reference parameter (as indeed this parameter is fixed and does not have error) and inverts the matrix, returning the standard errors by taking the square root of the diagonal. For most of the models, this process works fine. However, for the unique hierarchical model, the Fisher Information matrix is often numerically singular.

To see why, we need to consider the second derivatives as discussed in Section 2.2.5. We know that the Hessian should be in the form

$$\nabla^2 \ell = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta^2} & \frac{\partial^2 \ell}{\partial \theta \partial \alpha} \\ \frac{\partial^2 \ell}{\partial \theta \partial \alpha} & \frac{\partial^2 \ell}{\partial \alpha^2} \end{bmatrix}$$

As shown, we know that the top left of this matrix should be entirely filled, whilst the bottom-right shall be a diagonal matrix, as the second derivative on the same α yields a value, but deriving over any two different α's yields 0. However, when we assess the behaviour of the $\frac{\partial^2 \ell}{\partial \theta \partial \alpha}$ sections of the matrix, the hierarchical nature of the model causes problems that are not seen in the other models. Indeed, each element of this section includes an indicator term for whether or not the two teams being considered belong to the hierarchy being considered at this particular element when the teams are unique.

In essence, for some team i and level of hierarchy n, with $k = \sum_j \mathbb{I}(H_{ij} = n)$ representing the number of teams that team i shares a relationship with belonging to hierarchy n, and when considering a model with T teams and N levels to the hierarchy, we know that the column corresponding to the derivatives regarding $\frac{\partial \ell}{\partial \alpha_{i,n}}$ will have

- A non-zero element for $\frac{\partial^2 \ell}{\partial \alpha_{i,n}^2}$
- $TN - 1$ zero-elements, corresponding to all the zero-valued $\frac{\partial \ell}{\partial \alpha_{i,n} \alpha_{j,m}}$ elements, where at least one of $i \neq j$ and $n \neq m$ holds
- A non-zero element for $\frac{\partial \ell}{\partial \alpha_{i,n} \theta_i}$, assuming that $k > 0$

- k many non-zero elements, for each of the $\frac{\partial \ell}{\partial \alpha_{i,n} \theta_j}$ terms where $\mathbb{I}(H_{ij} = n) = 1$, and
- $T - k - 1$ zero elements, for the teams that do not have $\mathbb{I}(H_{ij} = n) = 1$

Indeed, to extend upon this, we know that the term commonly found in the second derivatives throughout Chapter 2

$$\frac{M_{ij} e^{\theta_i + \theta_j + \alpha}}{(e^{\theta_i + \alpha} + e^{\theta_j})^2}$$

is of the order $M_{ij} p(1-p)$, as the exponential term simply consists of $\mathbb{P}(i > j)_i \times \mathbb{P}(i < j)_i$. As such, we know that most of the terms without sums - those that involve two unique teams - should be of a similar magnitude when operating under the assumption that within a given set of data, teams within the same hierarchy should be facing off against each other a similar number of times, leading to similar values for M_{ij}.

Thus, considering the NBA example, for the column regarding a divisional alpha, meaning that the team only has 4 other rivals on the same level of the hierarchy, we see that we should have

- 2 larger terms, pertaining to $\frac{\partial \ell}{\partial \theta_i \alpha_{i,n}}$ and $\frac{\partial^2 \ell}{\partial \alpha_{i,n}^2}$
- 4 terms of similar order, pertaining to $\frac{\partial \ell}{\partial \theta_j \alpha_{i,n}}$
- 114 zero terms

We could also see that, when comparing to another rival in the same division, 3 of the same terms will still be of similar order - the $\frac{\partial \ell}{\partial \theta_k \alpha_{i,n}}$ expressions for the other 3 teams in the division. Thus, as only 114 of the 120 elements should bear significant difference - the 2 pairs of 0s swapping with the large terms, as well as the pair of theta-alpha cross terms which change - we can quite easily see how a computer program may see linear dependence, especially when these type of columns represent $100\frac{N}{N+1}\%$ of the matrices columns - or 75% for the NBA case. Thus, to avoid this computational issue, we opt to use a parametric bootstrap when estimating the standard errors for the hierarchical model, as discussed below in Section 3.3.

3.2.3 Output

```
> BTC(mat ,0)
$Theta
               Theta
Carlton    2.3650373
Richmond   0.7876102
Fitzroy    0.0000000
Collingwood 1.2285269

$Std.Errors
              errors
Carlton    0.9901086
Richmond   0.8256171
Collingwood 0.8506180
```

```
> BTC(mat ,1)
$Theta
               Theta
Carlton    2.3722619
Richmond   0.7900827
Fitzroy    0.0000000
Collingwood 1.2323912

$Alpha
[1] -0.12131

$Std.Errors
              errors
Carlton    0.9930236
Richmond   0.8270449
Collingwood 0.8523406
Alpha      0.4936103
```

Figure 3.4: Outputs for Model Types 0 and 1 on 1916 VFL Data

Figures 3.4 and 3.5 show what the output of a `BTC()` call looks like for the given set of data. As we can see, Fitzroy is automatically selected as the reference team in each of these models, thanks to their place as the weakest team. A number of other interesting results can be seen in these models,[4] however they are not pertinent to this section.

[4] In particular, Carlton won all 6 of their away games in 1916, skewing their Alpha to quite an unrealistic figure. This also yields large errors, and raises an important point in aiming to make sure that the data includes at least 1 of each possible outcome (Home/Away $\times$ Win/Loss)

```
> BTC(mat ,3)
$Theta
                 Theta
Carlton     14.6237117
Richmond     0.3735001
Fitzroy      0.0000000
Collingwood  1.7726275

$Alpha
                  Alpha
Carlton     -13.1324886
Richmond      1.8098951
Fitzroy       0.4989973
Collingwood  -0.4786770

$Theta.Errs
            theta.errors
Carlton       281.381167
Richmond        1.523324
Collingwood     1.550280

$Alpha.Errs
            alpha.errors
Carlton       281.380973
Richmond        1.771546
Fitzroy         1.748295
Collingwood     1.721365
```

Figure 3.5: Output for Model Type 3 on 1916 VFL Data

The return of a `BTC()` call can then be fed into another function, `BT_plot()`, to produce a lollipop chart of the estimates from the given model. Examples of these can be seen throughout Chapter 4. These models can also be used for a number of other purposes, as discussed further below.

3.3 Other Functions

Below, I list a number of the functions that currently exist within the project that are intended for public use. These include

- `SimpleHierarchy()` - a function that you can pass a number of row vectors too, each consisting of a list of teams, and return a relationship matrix for the simple hierarchy where $H_{ij} = 2$ for teams that are in the same group, and $H_{ij} = 1$ for teams that do not share the same group. It is possible to combine calls of `SimpleHierarchy()` to create more complex hierarchies, such as in the case of the NBA.

- `SeasonSim()` - takes a fitted model and a fixture, and returns a simulated season by allocating either 1-0 or 0-1 scores to the winning side. A similar function, `MatchSim()` performs the same operation for a single matchup between two given teams based off a model.

- `BT_plot()` - creates the plots seen in Chapter 4

- `WLTable()` - takes a set of matrices used for the fitting of a model, and returns the Win-Loss table according to said matrices.

- `BT_predict()` - returns the probability of `home_team` defeating `away_team` for a given model. `MatchSim()` and `SeasonSim()` are built upon this.

- `BT_test()` - performs the hypothesis test described in Section 2.3.3 for two models on a given set of data.

- `BT_bootstrap()` - returns the errors for some model on a given fixture using a parametric bootstrap. Each run consists of using `SeasonSim()` to simulate the given fixture, before fitting a model to it, and adding the given estimates to the current set of estimates. The standard deviation of each set of estimates is then returned.

- `Data2Mat()` - converts a set of results, consisting of 4 columns: home.team, away.team, home_scores and away_scores, to the matrices required by the model.

- `ResultSplitter()` - takes a set of results where a 'result' column exists rather than 2 indivudal score columns, and turns them into two columns as required by `Data2Mat()`.

3.4 Benchmarking

As stated in Chapter 1, some of the models that we have included in our R Package have already been implemented in the package **BradleyTerry2** - namely the vanilla and the common homeground models. As such, it is important that we compare our estimates for both the parameters and their associated standard errors for the purposes of quality assurance. It is worth noting that the methods of estimation are different for the two packages - where **BradleyTerry2** has opted to use the `glm` function to fit their model, our package directly uses the likelihood function and it's first-order derivatives with the `optim` function to find the maximum likelihood estimates for our parameters. Indeed, it is natural that we should expect minor differences in estimates as a result, however these should be of such a small magnitude that they can easily be waived away as computational error, or computational difference of methods. For the above models, along with the common hierarchical and unique homeground models, we will also benchmark the analytic standard errors found using the Fisher Information matrix against those found using a parametric bootstrap.

When benchmarking these estimates, we will consider both the errors themselves, as well as a handful of metrics regarding the bias of these. We will consider the p-value for the two-sided alternative hypothesis against a null hypothesis that the proportion of biased results for each parameter should be distributed by $\mathrm{Bin}(n, 0.5)$, where n is the number of parameters being estimated.

All of these results consider the NBA data used in Chapter 4.2, with all benchmarking performed over 500 runs.

Metric	BT2 Estimates	BT2 Errors	Bootstrap Errors
Mean Error	2.697×10^{-5}	2.089×10^{-7}	0.00754
Mean % Error	0.01206%	0.00012%	4.427%
Max % Error	0.26218%	0.00105%	12.101%
Min % Error	2.949×10^{-6}%	2.454×10^{-7}%	0.704%
Mean Bias	3.143×10^{-7}	-8.756×10^{-8}	-0.0068
% of overestimates	41.38%	41.38%	6.897%
p-value	0.45826	0.45826	< 0.0001

Table 3.1: Benchmarking Results for the Vanilla Model

Metric	BT2 Estimates	BT2 Errors	Bootstrap Errors
Mean Error	3.629×10^{-5}	4.523×10^{-7}	0.00231
Mean % Error	0.00939%	0.00028%	1.377%
Max % Error	0.1372%	0.00069%	3.694%
Min % Error	0.0008%	0.000032%	0.038%
Mean Bias	-1.62×10^{-5}	4.162×10^{-7}	0.00082
% of overestimates	36.67%	90%	56.67%
p-value	0.2649	< 0.0001	0.4583

Table 3.2: Benchmarking Results for the Common Model

Metric	Common Hierarchical	Unique
Mean Error	0.00395	0.0092
Mean % Error	2.344%	3.592%
Max % Error	5.399%	10.496%
Min % Error	0.389%	0.00175%
Mean Bias	0.0017	0.0063
% of overestimates	59.375%	76.27%
p-value	0.2962	< 0.0001

Table 3.3: Benchmarking of Errors of Models 2 and 4 against the Bootstrap

First of all, it's import to note that both the estimates and the errors provided by our package line up very closely with those provided by **BradleyTerry2**. Whilst there are some minor differences, they appear small enough that they can be ignored - indeed, we cannot guarantee that the estimates and errors provided by **BradleyTerry2** are perfect either.

It's noteworthy that the errors are quite different to the bootstrapped errors. In particular, whilst the bootstrap appears to overestimate the errors for most of the models, there is significant evidence to suggest that it underestimates the errors in the vanilla model. Indeed, these estimates are not out by much on average, but it does suggest that where possible using the Fisher Information matrix to estimate errors is not only optimal due to algorithm run time - for instance, 500 bootstraps takes around 15 to 20 minutes on my computer[5] - but indeed provides a better fit for the standard errors.

On the note of timing, we can also compare the time it takes to run our Likelihood optimisation algorithm against the `glm()` framework favoured by `BradleyTerry2`. Indeed, over 100 runs on the NBA 2016-2020 dataset, we see the following run times (all recorded in seconds)

Function	Mean	SD
My Vanilla	0.3590459	0.04205163
BT2 Vanilla	0.009923978	0.002749742
My Common	0.4359825	0.06213588
BT2 Common	0.01789508	0.004612482
My Team-Specific	1.100939	0.1230922
My Hierarchical	1.445455	0.5577743

Whilst `BradleyTerry2` does appear to display a vastly quicker run-time, neither of these could be considered significantly prohibitive. Indeed, the data consists of 4904 observations across 30 teams - it is unlikely that many larger datasets would be considered at least in the sporting world, other than perhaps College Football and Basketball competitions in the United States, or MLB data, where each MLB season is only twice the size of an NBA season. Indeed, in Chapter 5, we discuss that even in the NBA, perhaps this dataset of 4 seasons is too large for our stationarity assumption.

[5] Using 16GB RAM and an Intel i5-6500 CPU @ 3.20GHz

With regards to how this could perhaps be improved, it would be interesting to implement a Fisher-Scoring algorithm, or to try and use the `glm()` package as a part of our project instead and compare the difference in speed.

Chapter 4

Applications

4.1 AFL Data

The Australian Football League, or **AFL** as it is most commonly referred to, is a professional Australian Football competition contested in the Australian winter by 18 teams from across the country. The league has a number of peculiarities that lead to inequalities across the league, and suggest that a hierarchical modelling may be valuable.

First of all, each team plays 22 matches throughout the course of a regular AFL season, ensuring that they are able to face every other team at least once. However, this does leave 5 teams which a club will play twice in a year. Unlike many other leagues, there are no rules set in stone to guarantee any level of fairness as to how these extra games are allocated. Instead, they are usually allocated to ensure the maximum number of 'blockbuster' matches in a season, so as to boost the revenue of the league. Unsurprisingly, what this means is that between 2010 and 2019, both the South Australian and Western Australian derbies (Adelaide vs Port Adelaide, and West Coast vs Fremantle respectively) have been contested twice in each and every season. A similar result holds true for the Queensland and Sydney derbies from 2011 and 2012 respectively, when a second team entered the league from these markets. Other match-ups are

# of Matches	Team 1	Team 2
20	Port Adelaide	Adelaide
	West Coast	Fremantle
19	Collingwood	Essendon
	Geelong	Hawthorn
18	Collingwood	Carlton
	Sydney	Hawthorn
	Brisbane	Gold Coast

Table 4.1: The most common matchups in the AFL Home and Away season from 2010 to 2019

also over-represented, including some traditional rivalries such as Collingwood and Essendon, who play off in the ANZAC match each year, or Collingwood and Carlton, who have played each other 260 times since the league began, occuring more often than any other matchup,[1] as well as some match-ups which have been favoured due to recent history, such as between Geelong and Hawthorn, who amazingly won 7 of the 9 Grand Finals played between 2007 and 2015, including the 2008 Grand Final where these two teams indeed played each other, or Sydney and Hawthorn, who played each other in both the 2012 and the 2014 Grand Finals.[2]

On the other hand, some matchups have occurred only the minimum number of times, or close to, over the last 10 seasons. It should be noted that Gold Coast entered the competition in 2011, followed by GWS in 2012, and so these sides have only played in 9 and 8 of the included seasons respectively. However, there are still some matchups which they have avoided as much as possible - for instance, Collingwood have played both GWS and Gold Coast the least possible number of times, with Gold Coast vs Richmond also having been played only once in each and every possible season. Some reasons for this could be the vast difference in ability between the sides - as can be seen by the model estimates later - and the idea that it

[1] These two teams have also played each other in 22 Finals and 6 Grand Finals. At time of writing, the match-up is curiously even - each side has beaten the other 128 times, whilst there have been 4 draws.

[2] Indeed the only two seasons in our dataset where these two teams did not play each other twice were in 2011, before either side's successes, and in 2019, long after these Grand Finals.

# of Matches	Team 1	Team 2
8	GWS Giants	Collingwood
9	Gold Coast	Richmond
	Gold Coast	Collingwood
	GWS Giants	North Melbourne
	GWS Giants	Hawthorn
	GWS Giants	Fremantle
	GWS Giants	Brisbane
	GWS Giants	Carlton

Table 4.2: The least common matchups in the AFL Home and Away season from 2010 to 2019

would be a waste of a match to have the most well attended clubs playing such small market teams more often than necessary.[3] The other type of match-up which occurs the least often is those between teams which are both quite unpopular, such as GWS vs North Melbourne, Fremantle or Brisbane,[4] as these are clearly the least profitable games for the league.

Regardless of the reason behind these lopsided fixtures, it is quite clear that over a few seasons, teams cannot be compared solely off their Win-Loss record, as some teams will have had a comparatively easier fixture than others. This makes it a perfect choice for using a Vanilla Bradley-Terry Model to compare teams. A plot showing the estimates of θ for each team is shown in Figure 4.1

[3] In 2019, Richmond and Collingwood were the top two teams for average attendance, while Gold Coast and GWS were the two least attended teams

[4] Brisbane and North Melbourne were in the Bottom 4 for average attendance in 2019, whilst Fremantle had the worst average attendance away from home in 2019.

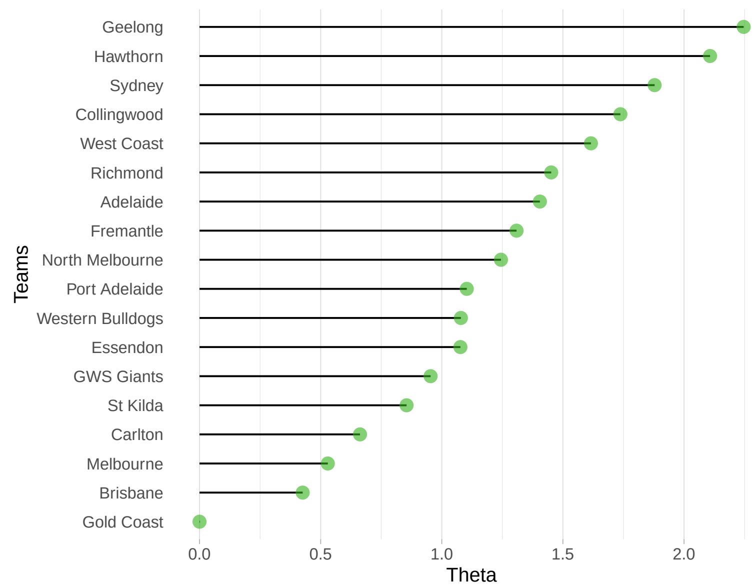

Figure 4.1: Vanilla Bradley-Terry Model of AFL 2010-2019 dataset

Whilst not unique to the AFL, the advantage that a team may gain by playing at home in this league may be exacerbated in comparison to competitions in some foreign countries due to the vast expanse of Australia's landmass. As such, it seems entirely reasonable that there exists some level of advantage for playing at home. Performing our likelihood ratio test, we indeed see that there is very strong evidence ($p < 0.001, \chi^2 = 39.67$) to suggest that the model is better explained by including a common α term than by ignoring it. The resulting plot from our package is shown in Figure 4.2.

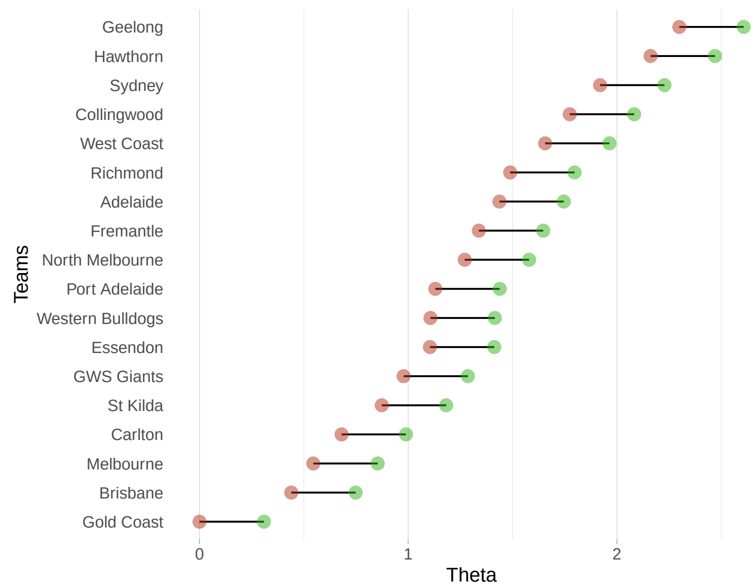

Figure 4.2: Common Homeground Bradley-Terry Model of AFL 2010-2019 dataset (*Red*: θ_i, *Green*: $\theta_i + \alpha$)

One peculiarity of note is the manner in which the winner of the competition is decided in the AFL. In most Association Football leagues around the globe, the team which sits atop the ladder at the end of the Home & Away season is considered the champion, whilst in American competition, a series of playoffs culminate in a match, or series of matches, at either a neutral venue or spread across the venues belonging to the two teams. However, owing to the competitions history as the Victorian state league, the AFL Grand Final is played at the MCG in Melbourne, Victoria each and every season. This is a source of discontent for many fans when we consider that the MCG is the nominated home ground for 5 clubs, and that 9 of the 18 teams across the league call Melbourne home (as well as a 10th Victorian team, Geelong, representing a city which is a short drive west of Melbourne). As such, it is perhaps useful for us to consider whether there is evidence that Victorian teams have a differing level of advantage when playing teams from interstate compared to against their fellow Victorians. Indeed, when we consider the Common Hierarchical model against the Common Homeground model, we see quite strong evidence to suggest that this is the case ($p = 0.00336, \chi^2 = 8.601$). Our estimates for these advantages are $\alpha_1 = 0.4533$ when playing between the hierarchies (i.e. when a Victorian team hosts a non-Victorian team), whilst $\alpha_2 = 0.1631$ is clearly much smaller, representing the advantage when two teams are in the same group (i.e. Victorian vs Victorian). This is in comparison to the common estimate of $\alpha = 0.3087$ under the previous model. Plots of this model for each level of the hierarchy can be found in Figure 4.3.

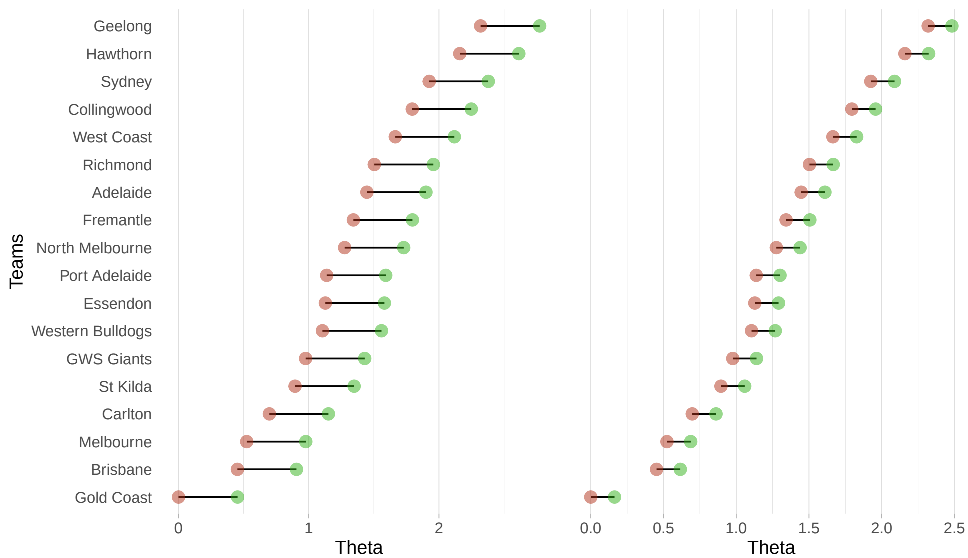

Figure 4.3: Common Hierarchical Bradley-Terry Model of AFL 2010-2019 dataset - across group on left, within group on right (*Red*: θ_i, *Green*: $\theta_i + \alpha$)

Another strange quirk pertaining to the AFL is that the size of fields upon which a match is played are not standardised. Whilst the dimensions must exist within some reasonable set of bounds, they can vary quite drastically - for instance, Kardinia Park where Geelong play measures 170m by 116m, a little longer than most AFL grounds, but one of the narrowest, especially when compared to cricket grounds such as the Gabba in Brisbane (156 by 138) and the Melbourne Cricket Ground (161 by 138) which are much rounder, owing to their original purpose as cricket grounds rather than football fields. A full list of the venues at which each AFL team plays at, and where they are located within Australia, can be found in the appendix.

It also happens in the AFL that many sides share stadia - both the West and South Australian teams share the same field in their state's capital, whilst the 9 Melbourne teams share 2 main football grounds. As such, it would make sense that there are varying levels of homeground advantage between sides, and so the unique homeground model may be a better choice. However, this would not appear to be the case, as when compared against the previously mentioned Common Hierarchical model we fail to see any significant evidence for this hypothesis ($p = 0.3713, \chi^2 = 17.22$). Regardless, a plot of this model is shown in Figure 4.4

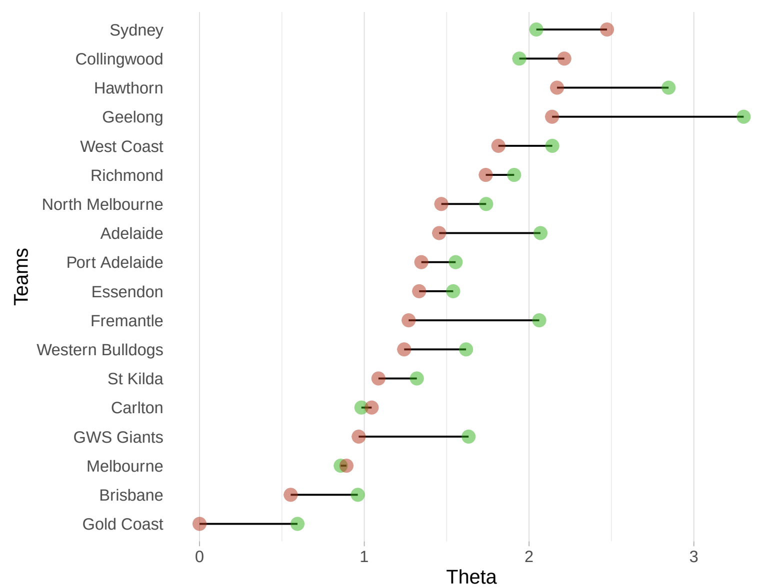

Figure 4.4: Team-Specific Homeground Bradley-Terry Model of AFL 2010-2019 dataset (*Red*: θ_i, *Green*: $\theta_i + \alpha_i$)

4.2 NBA Data

The National Basketball Association, or **NBA**, is universally considered to be the strongest professional basketball league on the planet. The league boasts 30 teams, divided evenly and geographically into 2 conferences of 15 teams, which again is divided geographically into 3 divisions of 5 teams each. The exact specifics of this hierarchy can be found in Table A2.

Given that the NBA has this hierarchical structure, it is a natural choice for applying our model. On the one hand, both levels of the hierarchy are made on geographical factors - as such, the travel to visit an intra-division rival is most likely much shorter than the travel to visit an intra-conference rival, and shorter again than visiting an inter-conference rival. These divisions have also stayed static for quite some time - the structure has not changed since the league last grew in 2004-05[5], leading to strong rivalries between some teams, or exacerbating pre-existing rivalries between neighbouring cities. Finally, we also see that the NBA uses this hierarchy to craft an uneven fixture. As of the 2019-20 season, the NBA fixture for any given team is set using the following rules:

- Four matches against all other teams within the division
- Four matches against 6 of the 10 teams within the conference, but in other divisions
- Three matches against the other 4 intra-conference rivals
- Two matches against each of the 15 teams in the opposing conference

The teams who one will play three or four matches against rotate over a period of 5 seasons, such that at the end of a 5 year period, any team will have played 20 matches against division rivals, 18 against conference rivals, and 10 against inter-conference rivals, with Home and Away splits also made even, within season where possible (i.e. two at home and two away when playing a team 4 times in a season), or over the life of the cycle (for when only 3 meetings occur in a season).

Qualification for the playoffs is secured by finishing in the top 8 of a conference, based off the number of matches that a team wins. Once the playoffs commence, they are conducted first as a knockout competition

[5]Indeed, some elements of the stucture have stayed the same since the early days of the NBA when there was only an Eastern and a Western Division - for instance, the Boston Celtics and the New York Knicks have been in the same division for every single NBA season since it's advent in 1946.

within each conference. The NBA champion is then decided by a best-of-7 series between the two conference champions. This is the first time during the playoffs that two teams from opposing conferences are pitted against each other. As such, should one conference be deemed to be of a lower standard than the other, a team in this conference could be presented with a much easier run at the finals than a team in the tougher conference.

These inequities in fixture make our NBA data an example ripe for treating with the Bradley-Terry model. The hierarchical nature of these inequities would also suggest to use that applying our hierarchical extension to the problem would also be a sensible step. As such, we consider the following results, procured by applying our R package to the provided NBA dataset, consisting of all official matches - both regular season and playoff - played from the 2016-17 season through to the final match prior to the Covid-19 break in March 2020.

Plots for the Vanilla, Common, Team-Specific and Hierarchical models are found in Figures 4.5, 4.6, 4.7 and 4.8 respectively. Table 4.2 compares the strengths of teams across these 3 models, ranked by their θ_i estimate within the model, and compared to their Win-Loss record across the given time period.

Whilst there is strong evidence to suggest that there is a common homeground advantage ($p < 0.001, \chi^2 = 137.46$), there is weak evidence at best to suggest that there are further hierarchical effects. Comparisons to the common hierarchical ($p = 0.384, \chi^2 = 1.914$) and the full hierarchical model ($p = 0.218, \chi^2 = 99.106$) are inconclusive, whilst there is weak evidence to suggest that the team-specific homeground advantage does a better job of modelling the competition than the common homeground ($p = 0.064, \chi^2 = 41.327$).

It is certainly interesting, but not entirely unsurprising, to see that these effects are negligible. For one, basketball courts are of a regulation size, and so the court alone has no real impact on play. The large number of games and short breaks between matches that occur in the NBA may also play some part - each match is followed quite quickly by the next, sometimes even on back to back days. NBA players, along with the large teams of health and performance professionals they're surrounded by, would certainly be able to ensure that even when on the road, these players are able to still perform at their best. Teams are also guaranteed to visit every opposition venue at least once a season, so the intimidation and unfamiliarity factors of playing in front of these hostile crowds may be lesser.

Another interesting piece of research conducted in 2021 by Home Team Sports, alongside Wakefield Research Partners and YouGov may also suggest that the passion of fans for specific teams is lower in the NBA than other leagues. The study asks fans of the top four American leagues[6] a number of questions regarding their opinions on the league and their relationship with it. It is of note that the proportion of NBA fans who agreed with statements such as “My favourite team of any sport is in this league”, “My favourite team in this league is in/near my hometown” and “Being a fan has been a long time family tradition” were the lowest of the four leagues. This could suggest that NBA fans are perhaps less passionate, or at the very least less heavily invested in their own team. Given that teams play 41 games at home over 5 months would suggest that only the most die-hard of fans would even consider attending the majority of matches that their team plays.

[6] NFL, NBA, NHL and MLB

Teams	Vanilla	Team-Specific	Hierarchical	Win/Loss [1]
Houston Rockets	1	2	3	1
Golden State Warriors	2	1	1	3
Toronto Raptors	3	3	2	2
Milwaukee Bucks	4	5	5	5
Boston Celtics	5	4	4	4
Utah Jazz	6	7	7	8
Los Angeles Clippers	7	6	6	7
San Antonio Spurs	8	10	11	9
Denver Nuggets	9	9	9	6
Oklahoma City Thunder	10	8	8	10
Portland Trail Blazers	11	13	13	13
Philadelphia 76ers	12	16	15	12
Indiana Pacers	13	14	12	11
Miami Heat	14	11	10	14
Los Angeles Lakers	15	15	16	16
New Orleans Pelicans	16	12	14	18
Cleveland Cavaliers	17	17	17	17
Washington Wizards	18	20	21	15
Minnesota Timberwolves	19	25	24	21
Detroit Pistons	20	26	26	19
Memphis Grizzlies	21	24	25	22
Charlotte Hornets	22	19	19	20
Dallas Mavericks	23	21	22	23
Sacramento Kings	24	18	18	24
Orlando Magic	25	22	23	25
Brooklyn Nets	26	23	20	26
Atlanta Hawks	27	28	28	27
Chicago Bulls	28	27	27	28
New York Knicks	29	30	29	29
Phoenix Suns	30	29	30	30

Table 4.3: Comparisons of Rankings by Log-Skill across Models on the 2016-2020 NBA Dataset

[1] Win-Loss records for the dataset can be found in Table A3

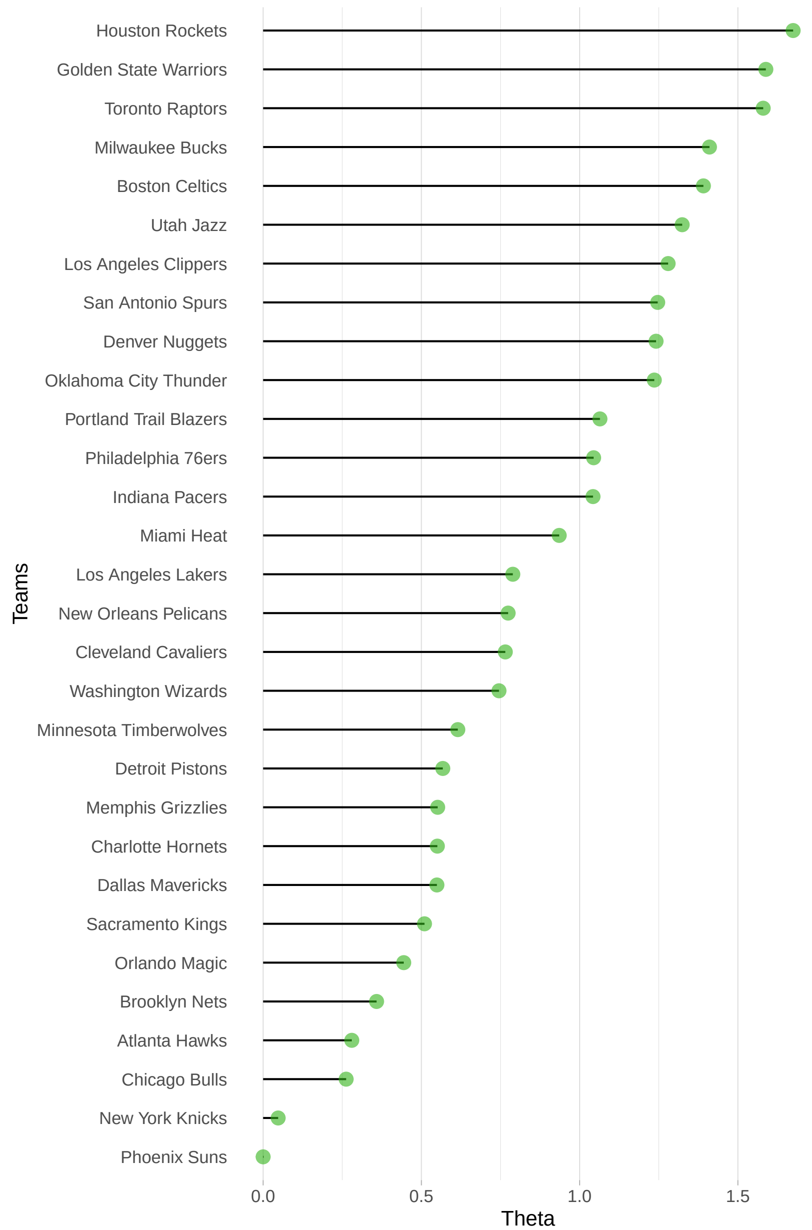

Figure 4.5: Vanilla Model of NBA 2016-2020 dataset

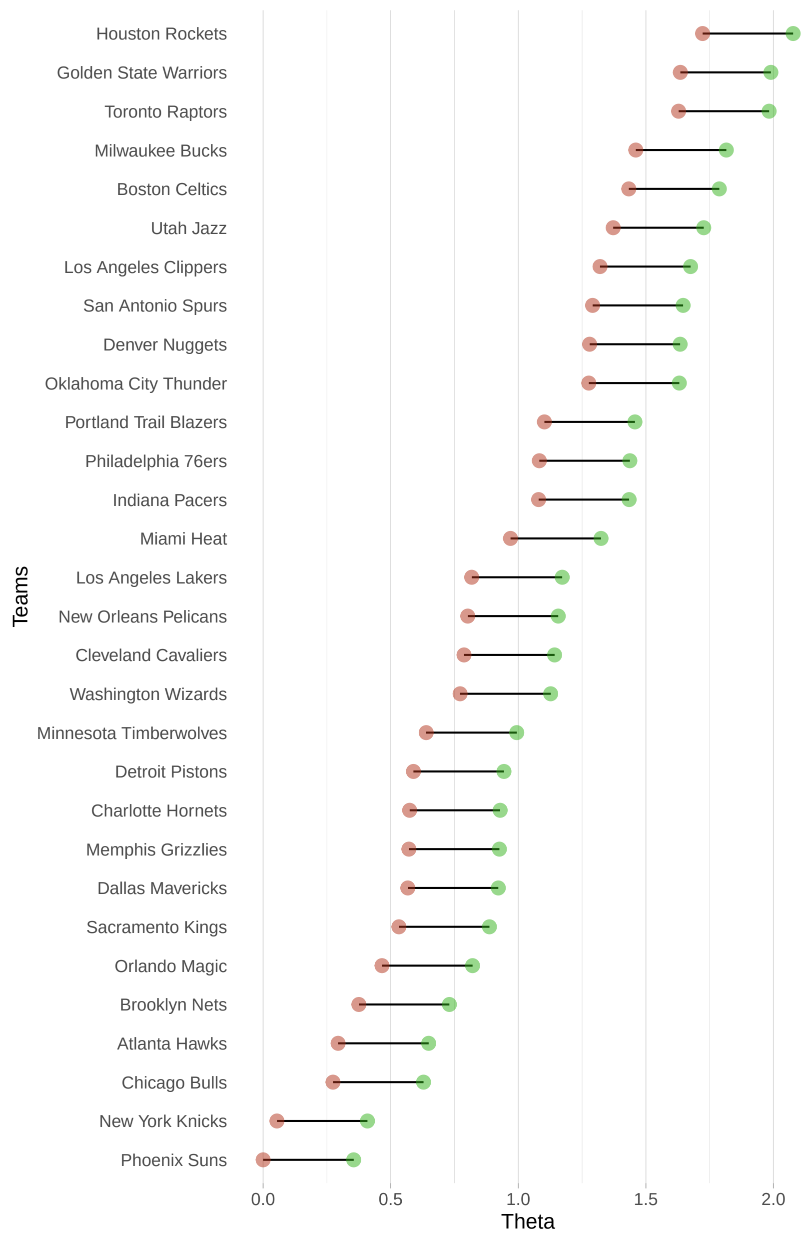

Figure 4.6: Common Model of NBA 2016-2020 dataset

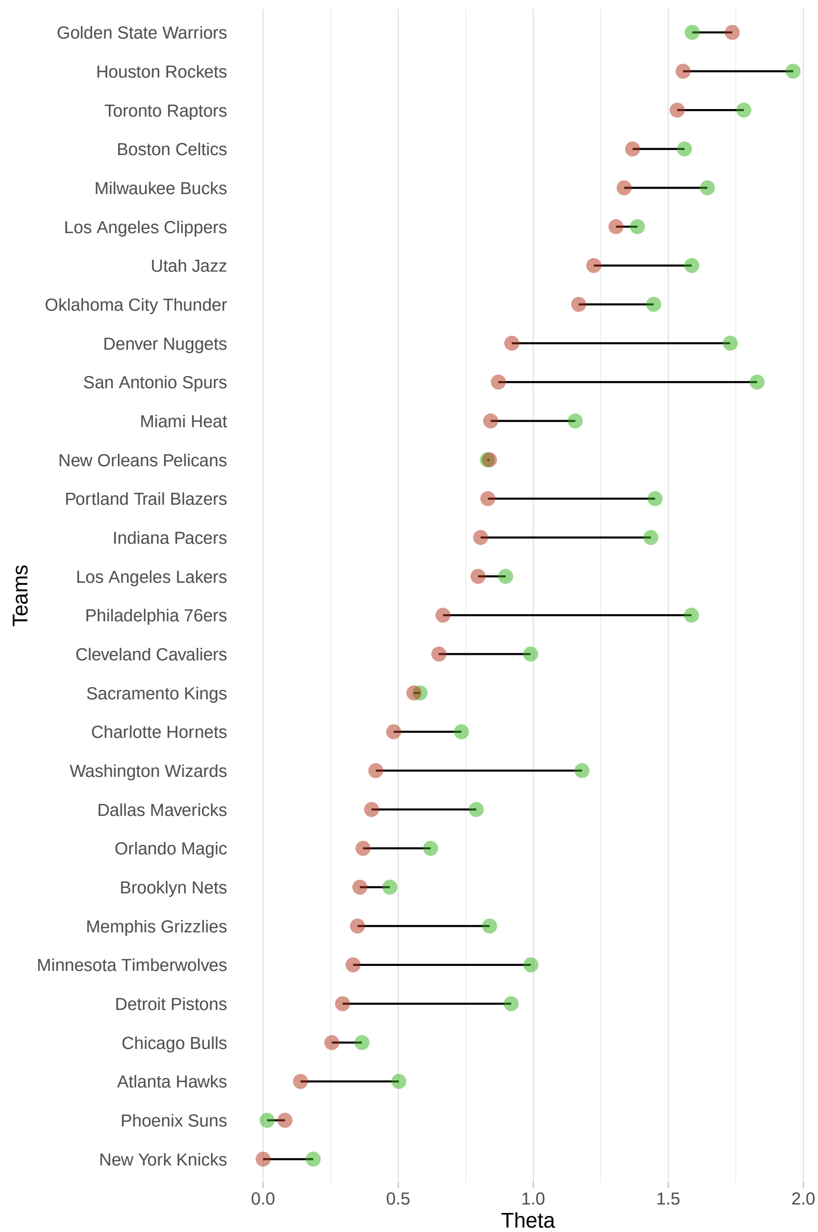

Figure 4.7: Team-Specific Homeground Model of NBA 2016-2020 dataset (Red: θ_i, $Green$: $\theta_i + \alpha_i$)

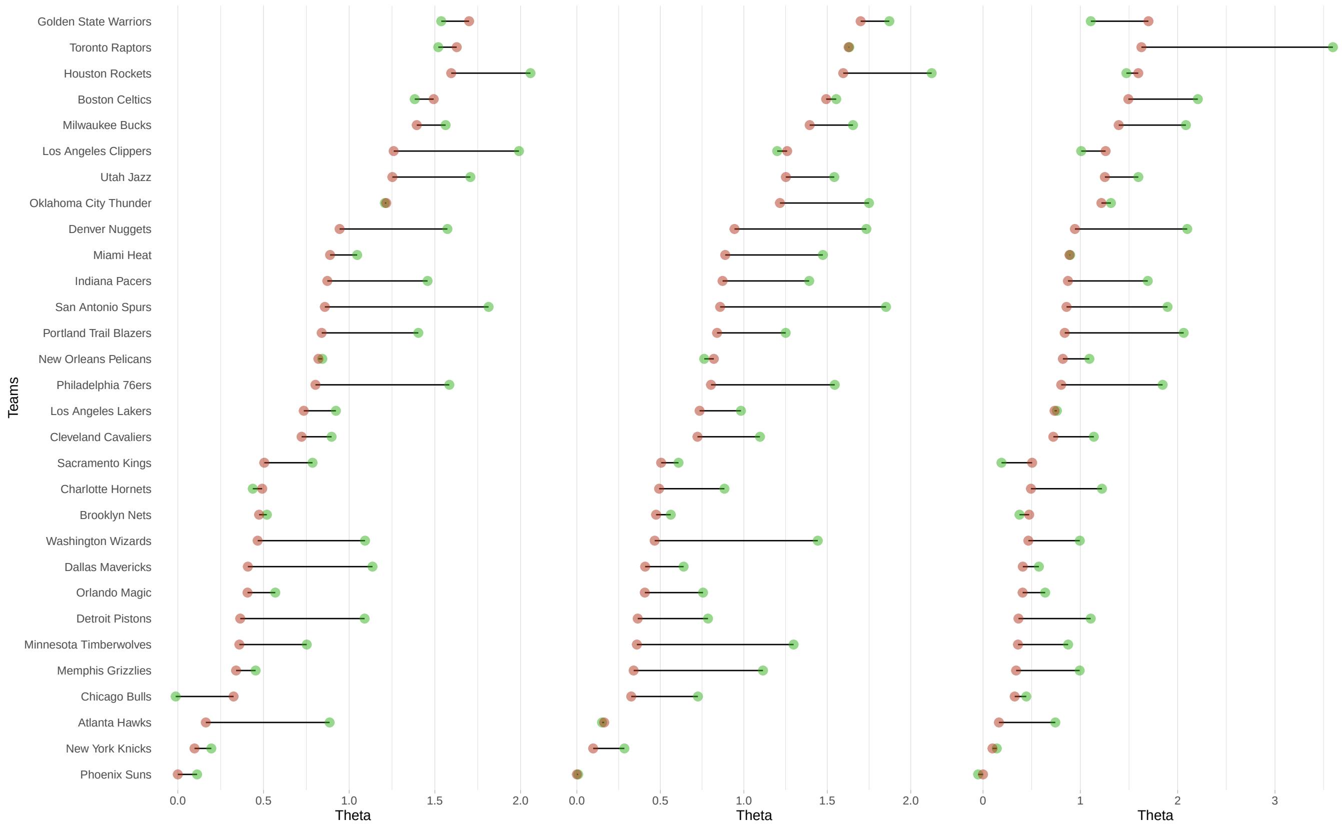

Figure 4.8: Hierachical Model on the NBA 2016-2020 dataset - from left to right, these are Inter-Conference, Intra-Conference and Intra-Division (*Red*: θ_i, *Green*: $\theta_i + \alpha_{i,1}$)

Chapter 5

Discussion on the Stationarity of Sporting Competitions

Throughout a student's undergraduate studies, the assumption that the observations from a set of data are all independent and identically distributed is something that one may often simply smile and nod at, comforted by the fact that in any reasonable experiment such an assumption should hold true. However, in the case of the data we have considered, it should be questioned thoroughly. Indeed, it has the potential to cause problems, as our asymptotic results from Section 2.3 require the ability to take the size of an iid sample to infinity. As such, any long term differences or time dependency in our parameters would contradict this assumption, and bring these results into disrepute. In spite of this, try as we might, many real world constraints prevent us from having near-infinite samples in the sporting sphere. As a result, we have included some qualitative discussion on these assumptions - what real world effects underpin changes in strength and how quickly do these changes in strength appear to occur - as well as a brief overview on how modern literature models these issues, with the hope that perhaps somebody may take an interest in expanding this discussion further in future.

5.1 Challenges to Stationarity

It comes as no shock that sporting competitions do not stay stationary. Indeed, it is the hope that "next year will be our year" that keeps fans returning to their club season after season. Yet it is also the nature of life that a professional athlete can not keep operating at peak performance forever - for instance, the last time a player won the Brownlow Medal[1]aged 30 or older was Nathan Buckley in 2003, whilst all but one[2] of the last 13 Brownlow Medallists have been between 25 and 29 years old at time of winning. Many leagues, including both the AFL and the NBA, also implement measures such as salary caps and drafts for young talent as a mechanism for preventing clubs from having long periods of success.

Comparing teams many seasons removed from each other is somewhat like comparing apples with oranges. For instance, if we compare the 97-98 Chicago Bulls - a team fresh off winning 6 of the last 8 championships, boasting 5 future Hall of Fame players and a Hall of Fame coach at its peak, and achieving a 74.7% win rate across the past 8 years of regular season play - to the same franchise 20 years later, finishing the 2018-19 season with only 22 wins from the 82 matches, we can clearly see that they are nothing alike. Similar examples can be found in any sport - in the AFL, Brisbane went from playing 4 straight Grand Finals from 2001 to 2004, to finishing in the bottom 2 positions on the ladder in 2015, 2016 and 2017. Of course, this can happen in the opposite direction too - recent NRL Premiers Penrith missed finals all but once from 2005 to 2013, and recent AFL heavyweights Richmond had not won a final since 2001 before their drought-breaking 2017 premiership - now they've won three of the last five flags.

But sometimes, for reasons known or unknown, the fate of a team can change overnight too. Indeed, the same 97-98 Chicago Bulls discussed above dropped off the face of the earth the very next year, finishing 98-99 with a 26% win record, with 20.7%, 18.3% and 25.6% records in the seasons to follow. Conversely, the Sydney Roosters came within 40 minutes of

[1]the AFL's annual award for the Fairest and Best player during the Home and Away season

[2]Indeed, the one player to win one while younger than 25 during this period was Nat Fyfe in 2015, who went on to win a second Brownlow in 2019, then aged 27

winning the 2010 NRL Premiership a mere 12 months after collecting the wooden spoon. The former of these can easily be explained - the team fell apart. Michael Jordan retires, Scottie Pippen moves to Houston, Dennis Rodman joins the LA Lakers, and coach Phil Jackson leaves the organisation too. The latter is perhaps a little more difficult - whilst the Roosters had a change of coach, and picked up the problematic Todd Carney after serving a 12-month ban from the game, the team kept Braith Anasta as captain, with key players such as Shaun Kenny-Dowall, the team's leading try-scorer in both seasons, Anthony Minichiello and Mitchell Pearce already on the books in 2009.

Thus, considering the variety of ways that the fortunes of individual teams may change, it becomes clear that we cannot just take some sample of data, throw it into the model and hope for the best. Below, we look at a couple of cases from the examples discussed in Chapter 4 to consider how many seasons is the optimal number for sample size. In particular, we know that the asymptotic results discussed in Chapter 2 rely on an assumption of the sample size approaching infinity. However we need to remain equally cautious about taking a window of seasons too large, potentially misconstruing the data and violating assumptions regarding independence and identicalness of samples.

5.2 Case Study

5.2.1 AFL

As discussed in Section 4.1, the data provided consists of all regular season matches played from 2010 to 2019 inclusive. We have clipped this dataset to only include matches from 2012 onwards for this section, as this was the first season of the current 18 team league. Throughout this section, we are going to consider the estimates of log-skill for two teams with vastly different fortunes over the given time period - the Hawthorn Hawks and the Greater Western Sydney Giants. The former of these teams went from being considered one of the greatest teams of the modern era, to a team that was better placed in the middle of the pack. The latter only joined the league in 2012, starting with the youngest and most inexperienced team list in the competition, but finished the period as Grand Finalists in 2019. Table 5.1 details the season-by-season results of these two teams, including their Home and Away record, as well as their stage of elimination from the finals (or DNQ when the team did not qualify). Details of the AFL Finals system can be found in Table A4.

Season	Hawthorn Hawks			GWS Giants		
	Record	Ladder	Finals	Record	Ladder	Finals
2012	17-5	1st	Runner-Up	2-20	18th	DNQ
2013	19-3	1st	**Premiers**	1-21	18th	DNQ
2014	17-5	2nd	**Premiers**	6-16	16th	DNQ
2015	16-6	3rd	**Premiers**	11-11	11th	DNQ
2016	17-5	3rd	Semi Final	16-6	4th	Preliminary Final
2017	10-11-1	12th	DNQ	14-6-2	4th	Preliminary Final
2018	15-7	4th	Semi Final	13-8-1	7th	Semi Final
2019	11-11	9th	DNQ	13-9	6th	Runner-Up

Table 5.1: Results for Hawthorn and GWS by Season

Having considered the fates of these two teams, we should now turn our heads to how the model interprets their strengths. Figures 5.1 and 5.2 show the estimates for log-skill by season using one, three and five season windows, with the 90% error bars for each estimate also. In addition, the estimate for log-skill when fitting the entire 8 seasons of data is given as a line in red, with the pink shaded area around it representing the 90% error for this estimate. The windows estimated are rolling windows, meaning that, for instance, the five year windows shown are 2012-2016, 2013-2017, and so on.

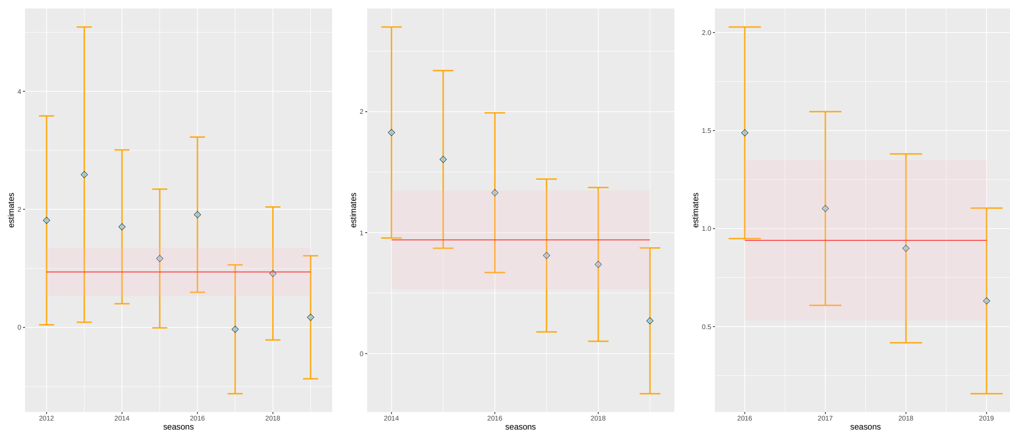

Figure 5.1: Estimates for Hawthorn under Common Homeground Model

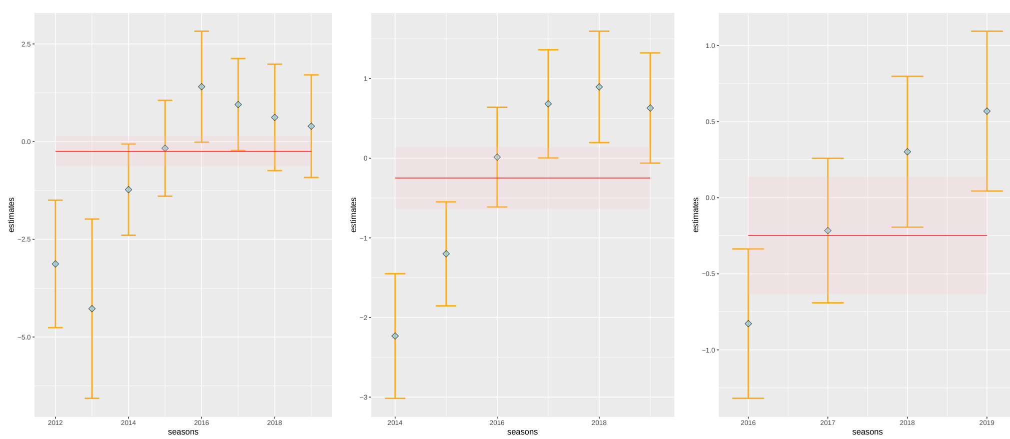

Figure 5.2: Estimates for GWS under Common Homeground Model

The first point of note is that when considering the season-by-season record of teams alongside the one season window graphs, the estimates for log-skill line up almost perfectly. Indeed, Hawthorn's three 17-5 seasons are all roughly given the same estimate, and a season where a team wins more games than another always results in a higher log-skill on the given dataset - perhaps an indication that the slightly uneven fixture discussed in Section 4.1 is not as much of a problem as one may think.

The more pertinent point is the difference in relationships that we can see on these graphs. Whilst there is clearly a negative trend on Hawthorn's one season graph, it is not a monotonic trend. Indeed, the longest run of consistent change is 2 decreases from 2013 to 2015. The error bars for all 8 single season estimates also overlap with the estimate for log-skill from the full dataset. However, the three season graph displays a monotonic trend, whilst the five season graph displays a relationship that appears to be almost perfectly linear. Similar results can be seen on the GWS graph, and although there are larger gaps on the single-season window estimates, they're hardly unexpected when considering the club went from being non-existent in 2010, to being packed to the brim with the very best youth at their first few drafts[3], followed by developing these players to full strength.

It should be noted that these relationships are the two most extreme examples found in the dataset. For example, Figure 5.3 shows the same series of plots for the Geelong Cats, a team who made finals in all but one of the eight seasons, but not a single grand final. Their worst record was 11-9-1[4] in 2015, whilst their best was 2018's 18-4 season.

[3]GWS were given extremely generous concessions upon entrance to the competition. These included a larger salary cap than other clubs, a large 'recruiting zone' for which they had exclusive rights to pick players from, alongside 9 of the first 15 picks at the 2011 draft, and 5 of the first 14 at the 2012 draft - these include getting to take picks 1, 2 and 3 at both drafts.

[4]Their Round 14 match against Adelaide was cancelled due to the murder of Adelaide coach Phil Walsh

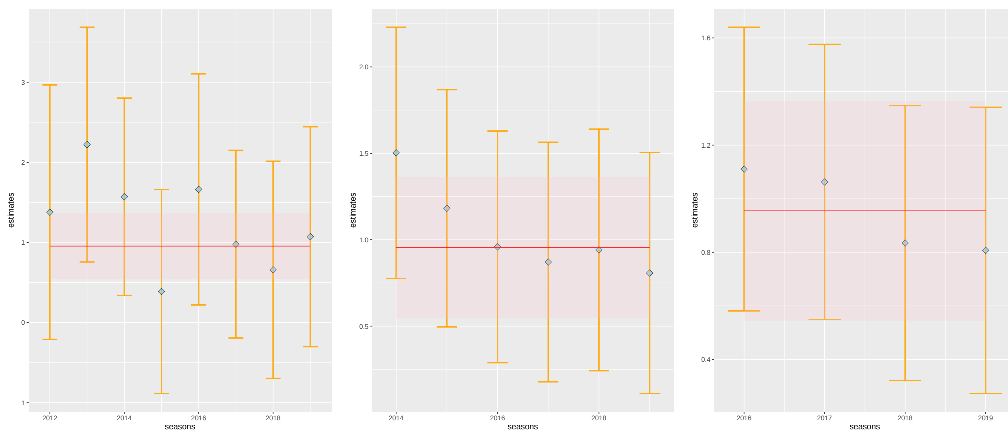

Figure 5.3: Estimates for Geelong under Common Homeground Model

Considering these, we can see that whilst their may be a slight trend with the five season windows, it is not anywhere near as pronounced as the previous two examples. The one season window plot shows that all of the error bars comfortably include the full data estimate, and indeed appear to bob above and below it almost randomly. The three season window also shows similar behaviour, with perhaps a decrease from the highs of the first two windows - perhaps either a sign of Geelong decreasing in ability themselves, but also perhaps due to the increasing competitiveness of the competition as a whole due to GWS beginning to catch up with the more established clubs.

Based off this qualitative discussion alone, it remains to find a more perfect solution to this problem. However, it would seem as though using a three season window may be best for modelling the AFL. Indeed, it seems clear that using the full 8 seasons is too large to be accurate, considering the number of significant differences between the overall estimate and some of the three season estimates. There could potentially be some argument made for the five season window also - a number of teams, including the Cats as shown, do not exhibit much of a relationship over three season windows either, and only begin to exhibit this in the five season window scenario. However, we do require that the entire competition remains relatively stationary to ensure accuracy, not just a small number of clubs. Whilst anecdotally, individual clubs appear to maintain similar levels of strength for at most five years - for instance, most of the modern day AFL 'dynasties' lasted for approximately this long,[5] as have the stints that clubs have spent in the bottom 4.[6] On the one hand, this does suggest that perhaps the AFL's equalisation measures are working well, but on the other, it does create a level of difficulty in establishing a window for analysis.

One side note, and a point worth potentially investigating further, is to try and explain how these changes may happen. In particular, this could give a simple way to check when the assumptions may have been violated by checking the underlying values of certain attributes. It's been noted above many times that GWS' improvement could most likely be explained by the maturing of their list. Equivalently, we also mention that players drop off as they get older, so it is of interest to consider whether potentially it was the aging of Hawthorn's list that saw their results slide. Figure 5.4 shows the one season estimates from above plotted against the average age of the club's playing squad for the season, as recorded by DraftGuru (2021). The points alone appear to give evidence to both of these hypotheses, alongside potentially indicating that the peak performance of an AFL team occurs when the average squad age is a touch above 24.

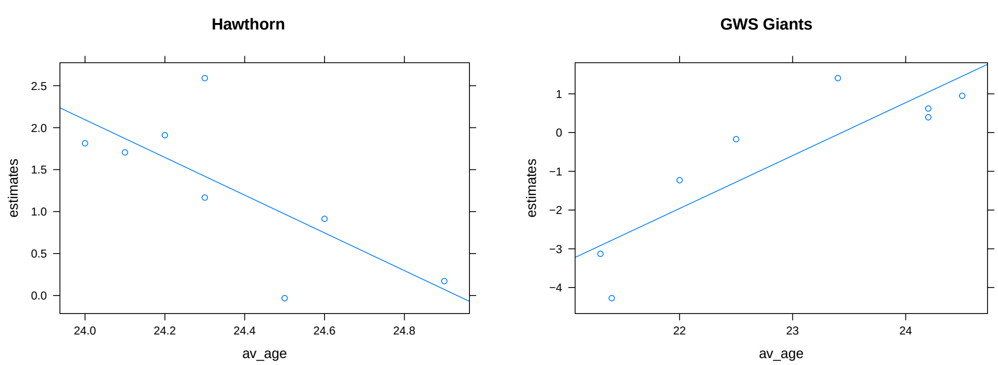

Figure 5.4: Log-Skill against average Squad Age

[5]Some examples would include Brisbane (2001-2004), Geelong (2007-2011), Hawthorn (2012-2015) and Richmond (2017-2020)

[6]Brisbane (2014-2018) and Gold Coast (2015-2019) are the only examples of teams that spend more than 3 consecutive seasons in the bottom 4 since 2000. Carlton (2005-2007 & 2017-2019), GWS (2012-2014), Melbourne (2012-2014), Richmond (2002-2004) and St Kilda (2000-2002) are the cases of three consecutive seasons.

Figure 5.5 shows the same relationship for all 18 teams, again lending some weight to this theory. It is of course worth noting that this is by no means a perfect analysis, but simply a starting point for further exploration - indeed, Collingwood, Carlton and Hawthorn entered 2021 with average ages of 24.2, 24.3 and 24.4, but all finished amongst the bottom 6 teams on the ladder.

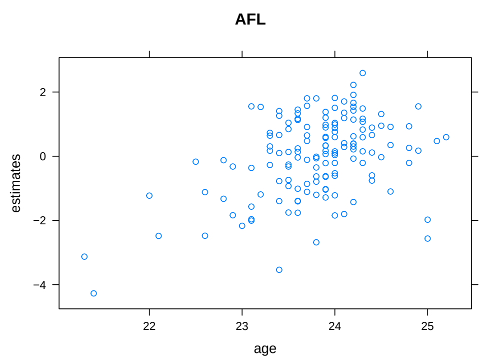

Figure 5.5: Log-Skill against average Squad Age

5.2.2 NBA

The other key hypothesis suggested above was that the strength of a team can change overnight with the right (or wrong) transfer policy - such as the aforementioned 98-99 Chicago Bulls. This type of change is more typical of professional basketball too, a sport where only 5 players take the court at any one time, and are expected to offer contributions to all phases of play. As such, one superstar player can have a significant impact on the fortunes of a team. Indeed, this spurns a common phrase in NBA media - the *Franchise Player* - the superstar that a franchise can build themselves around, both on and off the court. Think Michael Jordan at the Bulls, Steph Curry at the Warriors, LeBron James at Cleveland, or even slightly less famous players such as Tim Duncan at the Spurs or Dirk Nowitzki at the Mavericks.

As such, we turn to the dataset used in Section 4.2, covering all NBA matches from the start of the 2016-17 season through to the COVID-19 Break in 2020. In particular, we want to consider the sudden change in the fortunes of the Philadelphia 76ers, as shown in Figure 5.6

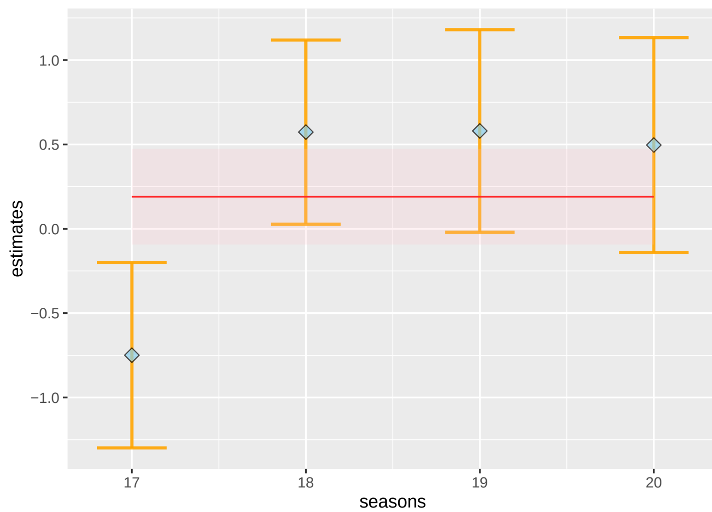

Figure 5.6: Estimates for the 76ers under Model 1 with one season windows

In 2014, Philadelphia selected Joel Embiid with Pick 3 of the NBA draft. They followed this up in 2016 by selecting Ben Simmons at Pick 1. Due to injury, Ben Simmons was forced to miss the 2016-17 season, whilst injury kept Embiid to only 31 matches. However, both players were able to play the majority of matches - 81 for Simmons, 63 for Embiid - in the 2017-18 season. Embiid also picked up an All-Star nod in 17-18, before both Simmons and Embiid were then selected as All-Stars for the 3 consecutive seasons thereafter. Whilst these are the fortunes of only 2 players, they can be quite easily seen in the results that the club achieved, and the estimates for log-skill that our model provides. As such, it may be the case that either such events could act as a trigger for closing a window, or perhaps even that in a competition such as the NBA, that star players be modelled by their own coefficients, rather than being lumped in with the rest of the team.

A similar case of 'overnight success' can be seen in the Milwaukee Bucks. Though perhaps not travelling as poorly as the 76ers had been, they too enjoyed a sudden leap in success between the 17-18 and the 18-19 NBA Seasons, as can be seen in Figure 5.7. A change in coach, and an improvement by franchise player Giannis Antetokounmpo to be named the

NBA's Most Valuable Player in both the 18-19 and 19-20 seasons are the most likely catalysts for such a rise - one which extends past this dataset and into having won the most recent NBA Championship in the 2020-21 season.

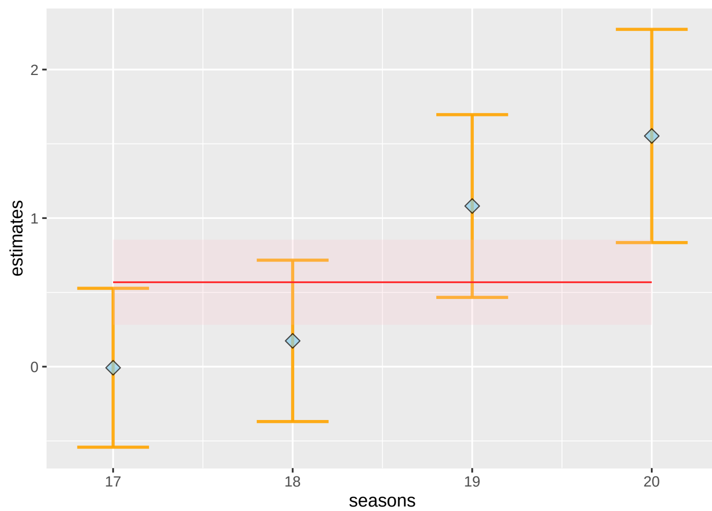

Figure 5.7: Estimates for the Bucks under Model 1 with one season windows

5.3 Alternative Approaches

Indeed, this problem in particular - the question of how large is large enough for n, without sacrificing the identicalness of the distribution - is one that does not appear to have a body of literature attached to it. Some interesting work regarding the Bradley-Terry model's asymptotics has been discussed, such as the work by Simons & Yao (1999) which proves that normality holds, under certain conditions, when the number of parameters approaches infinity if the number of matches between each pair of teams is fixed to some value. This is then extended upon by Han et al. (2020) in discussing how these asymptotics still hold for a sparse network, noting that as the number of parameters t approaches infinity, the number of matches between any specific pair would approach 0. Their assumptions can give us some information - normality should still hold so long as the probability that any two teams have faced each other in the data, p_t, is bounded from below by $\frac{\log(t)^{1/5}}{t^{1/10}}$ for $t \to \infty$. This assumption does not cause us a great deal of issue, as both the NBA and the AFL ensure that each team play each other at least once per season, as do the MLB, NHL and NRL. The only major sporting league where this may cause trouble is the NFL, where due to the season consisting of only 17 matches, 6 of which are against the same 3 teams each and every year, a given team will only be guaranteed to play each team in the opposing conference once every four years. As the above bound is equal to approximately 0.9066592 for 32 teams, this suggests that we would be forced to consider a minimum of a 4 year window when analysing NFL data. However, if we extend this model to some other interesting problems, such as ATP Tennis, where hundreds of players may compete each season without the guarantee of many of the possible pairings happening, or into the sphere of online gaming, where there may even be millions of possible players simply being matched on a first-come-first-served basis - for instance, Chess.com sees over 10 millions games played each and every day[7]. It could, however, be argued that such large systems would be computationally beyond the scope of our approach to the Bradley-Terry model anyway.

Despite the lack of work on this particular question, a great deal of literature has developed to answer the adjacent problem of "How do these models evolve?". In particular, we choose to highlight three main avenues of research on this question. First, we consider the well-known Elo ranking system, as used by many organisations such as the *Fédération Internationale des Échecs* (FIDE) in providing ratings for all players who compete in chess tournaments globally, as well as by *Fédération Internationale de Football Association* (FIFA), and it's future adaptations including the Glicko rating systems, as implemented in particular by many online chess websites, such as chess.com and Lichess. This is then followed by two more experimental methods, regarding the usage of exponentially weighted moving averages to provide 'decay' for old results over time, as well as the usage of interpolation between estimates provided at different points in time to paint a picture of change in skill over time.

[7]As stated in a recent Instagram post by the website, see `https://www.instagram.com/p/CU7qsWBBor7/`

5.3.1 Elo Ratings

Whilst the ratings proposed by Elo originally used the normal curve to decide what the probability of some player i defeating player j should be, it has subsequently been established that the usage of a logistic curve provides a more accurate fit - in particular, the logistic curve does not underestimate lower rated players as much as the normal curve will. As such, when using the logistic curve, we establish that

$$\mathbb{P}(i > j) = \frac{1}{1 + e^{R_j - R_i}}$$

where R_i and R_j are the established rankings for two competitors. In particular, we can see that this is equivalent to the Bradley-Terry model, by multiplying by e^{R_i}/e^{R_i}, and noting that $R_i = \theta_i$ for the established usage of θ throughout this paper. In particular, to yield values that were found to fit more with the tradition of chess, we could use the formula

$$\mathbb{P}(i > j) = \frac{1}{1 + 10^{(R_j - R_i)/400}}$$

as noted by Glickman (1995). To update a rating, the following formula is used. Take our current rating, R_0, and our score S in a game against our opponent, where 1 represents a win, 0 represents a loss and 1/2 shall represent a draw. Then, we have that our new rating is given by

$$R_n = R_0 + K(S - \mathbb{E}[S])$$

for some factor K as decided by the rating authority - indeed, it is this K that determines how quickly a rating should change based off a number of results, and itself provides the basis of Glickman's research into improving the Elo system. It can also be noted that whilst for a single match we simply have $\mathbb{E}(S) = \mathbb{P}(i > j)$, we can extend the formula further to cover a series of games, which is what is used in practice for calculating the ratings change between rating periods. The paper referenced above, *A*

Comprehensive Guide To Chess Ratings provides quite an in-depth discussion of this system, it's usage in practice and some of the results that stem from it. It also considers the order effects we have discussed throughout this paper, alongside how to estimate the probability of a draw as we discuss in Section 1.3.

Glickman has gone on to develop the modern Glicko rating system, the technical details of which he describes in Glickman (1999). We can summarize the model as using the Bradley-Terry model to yield the probability of player i defeating player j, but using a Bayesian updating algorithm with a Gaussian prior to select the K which governs how quickly a player's rating will change.

5.3.2 Dynamic Bradley-Terry

Cattelan et al. (2013) propose a method of updating a Bradley-Terry model dynamically using an exponentially weighted moving average (EWMA) type process. In particular, we consider a slight adjustment of the model which yields

$$\mathbb{P}(Y_i = 1 | Y_{i-1} = y_{i-1}, \ldots, Y_1 = y_1) = \frac{\exp\{a_{h_i}(t_i) - a_{v_i}(t_i)\}}{1 + \exp\{a_{h_i}(t_i) - a_{v_i}(t_i)\}}$$

where Y_i denotes the outcome of the i-th match, with $a_{h_i}(t_i)$ and $a_{v_i}(t_i)$ representing the abilities of the home and visiting team of the i-th match respectively, as estimated at time t_i. The model considers an evolution of 'home ability' and 'away ability' respectively, considering only matches played at home for the former, and away for the latter. This eliminates the need for a homeground advantage parameter, and could be considered similar to Model 3 described in Section 2.2.4. The updating equations are then given by

$$a_{h_i}(t_i) = \lambda_1 \beta_1 r_{h_i}(t_{i-1}) + (1 - \lambda_1) a_{h_i}(t_{i-1})$$
$$a_{v_i}(t_i) = \lambda_2 \beta_2 r_{v_i}(t_{i-1}) + (1 - \lambda_2) a_{v_i}(t_{i-1})$$

where λ_1 and λ_2 are smoothing parameters which govern the speed of updating, r_{h_i} and r_{v_i} are the outcome of the i-th match for the home and visiting team respectively, whilst β_1 and β_2 are described simply as 'home-specific' and 'visitor-specific' parameters. These models also need an initial condition for the beginning of the season. Cattelan et al. suggest that a suitable condition is to take equal abilities for every team, valued at $\beta_1 \bar{r}_{h_i}$ and $\beta_2 \bar{r}_{v_i}$, where $\bar{r}_{h_i}$ represents the average result for a home team across the *previous* season. Thus, if in the previous season 60% of NBA matches were won by the home team, we would take every teams initial ability to be equal to $0.6\beta_1$.

Indeed, in running such a model on the 2009-2010 NBA season, Cattelan et al. find an MLE for the two lambdas as being $\hat{\lambda}_1 = 0.043$, and $\hat{\lambda}_2 = 0.025$. Under this model, they found a log-likelihood of -752.22 compared to a log-likelihood of -830.56 for the model with $\lambda_1 = \lambda_2 = 0$, lending evidence to the hypothesis that these systems are dynamic - perhaps implying that form may exist within the NBA.

Another potential course of investigation along this line could be applying an EWMA type process directly to the data, refitting after each match. For instance, perhaps the number of wins that a win d days ago is worth in the algorithm could be given by $\exp\{-\lambda d\}$, or some other arbitrary way of decaying the value of matches played some time into the past. Given the generalisation of the choose function to taking continuous n and k by using gamma functions, this should not cause any issue for the current models discussed in this paper, so long as the equality $W + L = M$ still holds.

5.3.3 Interpolation of Estimates

Another novel approach to the problem is that of Krese & Štrumbelj (2021). Due to the greater complexity of these methods compared to the previous two, the details of their framework will be spared, but can be found in the referenced paper. At a high level, the paper describes a method where a set of times are selected across the dataset, with estimates at these time points being calculated, followed by creating a curve to interpolate these points using one of a number of methods - Vanilla

Barycentric Rational Interpolation (BRI), fixing the λ and β parameters used in the process), a Bayesian BRI where these parameters are treated as random variables, as well as modelling the series with a Gaussian Process, where the mean at time t is the estimate for θ at time t.

The paper considers an empirical evaluation of these methods on two datasets - NBA data from 2013 to 2018, as well as the matches played by 20 ATP[8] players from 2015 to 2019. A discussion of how many nodes should be used is had, as well as comparisons between feeding the entire dataset into a Bradley-Terry model against the interpolated models considered. Of particular interest to us is the fact that, on the ATP data, there is no real difference in the model performances. This perhaps suggests that, on a time scale of 5 years, the top 20 Tennis players kept reasonably stationary skill levels. This is not unrealistic - indeed, the issues regarding the trading of players could not possibly create difficulties here, as the 'teams' being discussed are individual players. 5 years is also perhaps not long enough to see enough of these players' fortunes change dramatically. In particular, over the given 5 years, all but one of the 20 Grand Slam titles available were won by either Djokovic, Nadal, Federer or Murray - the 2015 French Open was won by Stan Wawrinka. The number one ranked player was also one of these four during the entire time period.

This does not hold for the NBA data, where they to find that the interpolated estimates provide a better fit, but we have already established that a period of five seasons would seem much too long for the NBA to remain stationary, so this is not really of great concern. The paper also performs an interpolation of estimates for strengths based off of the odds published by bookmakers for each match - this outperforms any and all of the interpolated models by some margin, showing that the strength of these models is perhaps not as great as what one may desire.

[8]Association of Tennis Professionals

Bibliography

AFL Tables (2019). Crowds 2019. Data retrieved from the AFL Tables website on the 27th of July 2021, found at `https://afltables.com/afl/crowds/2019.html`.

AFL Tables (2021). All drawn games in chronological order. Data retrieved from the AFL Tables website on the 21st of July 2021, found at `https://afltables.com/afl/stats/biglists/bg8.txt`.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley-Interscience, 3rd ed.

Atkinson, A. C. (1972). A test of the linear logistic and Bradley-Terry models. *Biometrika* **59**, 37–42.

Austadiums (2021). All data regarding the stadiums used in the AFL was retrieved from the Austadiums website on the 27th of July 2021, found at `https://www.austadiums.com/stadiums/`.

Bar-Eli, M., Avugos, S. & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise* **7**, 525–553.

Beaver, R. J. & Gokhale, D. V. (1975). A model to incorporate within-pair order effects in paired comparisons. *Communications in Statistics* **4**, 923–939.

Bickel, P. J. & Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, 2nd ed.

Bradley, R. A. (1954). Rank analysis of incomplete block designs: Ii. additional tables for the method of paired comparisons. *Biometrika* **41**, 502–537.

BRADLEY, R. A. (1955). Rank analysis of incomplete block designs: Iii some large-sample results on estimation and power for a method of paired comparisons. *Biometrika* **42**, 450–470.

BRADLEY, R. A. & TERRY, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**, 324–345.

BRESLOW, N. E. & CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

CATTELAN, M., VARIN, C. & FIRTH, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society* **62**, 135–150.

CLEARY, M. (2020). Death of a ruckman? why clubs may abandon two big men in 2020.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen, 1st ed.

CRITCHLOW, D. E. & FLIGNER, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on glim. *Psychometrika* **56**, 517–533.

DAVIDSON, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65**, 317–328.

DAVIDSON, R. R. & BEAVER, R. J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics* **33**, 693–702.

DITTRICH, R., HATZINGER, R. & W, K. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society* **47**, 511–525.

DORSEY-PALMATEER, R. & SMITH, G. (2004). Bowlers' hot hands. *The American Statistician* **58**, 38–45.

DRAFTGURU (2021). All data regarding the average age of squads in the AFL was retrieved from the DraftGuru website on the 21st of October 2021, found at `https://www.draftguru.com.au/lists/2012`.

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

FIRTH, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software, Articles* **12**, 1–12.

GILOVICH, T., VALLONE, R. & TVERSKY, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* **17**, 295–314.

GLENN, W. A. & DAVID, H. A. (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics* **16**, 86–109.

GLICKMAN, M. (1995). A comprehensive guide to chess ratings. *American Chess Journal* **3**, 59–102.

GLICKMAN, M. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society* **48**, 377–394.

HALES, S. D. (1999). An epistemologist looks at the hot hand in sports. *Journal of the Philosophy of Sport* **26**, 79–87.

HAN, R., YE, R., TAN, C. & CHEN, K. (2020). Asymptotic theory of sparse Bradley–Terry model. *The Annals of Applied Probability* **30**, 2491–2515.

KRESE, B. & ŠTRUMBELJ, E. (2021). A Bayesian approach to time-varying latent strengths in pairwise comparisons. *PLoS ONE* **16**.

MAXWELL, A. E. (1974). The logistic transformation in the analysis of paired-comparison data. *British Journal of Mathematical and Statistical Psychology* **27**, 62–71.

NATIONAL BASKETBALL ASSOCIATION (2020). Official rules 2020-21.

Rao, P. V. & Kupper, L. L. (1967). Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *Journal of the American Statistical Association* **62**, 194–204.

Rogers, S. & Brown, A. (1998). *Every Game Ever Played: VFL/AFL Results 1897-1997*. Viking Books, sixth ed.

Simons, G. & Yao, Y.-C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics* **27**, 1041–1060.

Steeger, G. M., Dulin, J. L. & Gonzalez, G. O. (2021). Winning and losing streaks in the national hockey league: are teams experiencing momentum or are games a sequence of random events? *Journal of Quantitative Analysis in Sports* **17**, 155–170.

Turner, H. & Firth, D. (2012). Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software, Articles* **48**, 1–21.

Wakefield Research Partners, YouGov and Home Team Sports (2021). State of american sports fandom `https://public.tableau.com/views/StateofAmericanSportsFandom/Story1?:showVizHome=no`.

Whiting, M. J., Stuart-Fox, D. M., O'Connor, D., Firth, D., Bennett, N. C. & Blomberg, S. P. (2006). Ultraviolet signals ultra-aggression in a lizard. *Animal Behaviour* **72**, 353–363.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **29**, 436–460.

Appendix A

Appendix

Stadium	Length	Width	Location	Teams
York Park [8]	175m	145m	Launceston, Tas	HAW[1]
Marrara Oval[9]	175m	135m	Darwin, NT	N/A
Kardinia Park[10]	170m	116m	Geelong, Vic	GEE
Traeger Park[7]	168m	132m	Alice Springs, NT	N/A
Adelaide Oval	167m	123m	Adelaide, SA	ADE, POR
Cazaly's Stadium[7]	165m	135m	Cairns, Qld	N/A
Perth Stadium[11]	165m	130m	Perth, WA	FRE, WCE
Sydney Showground Stadium[12]	164m	128m	Sydney, NSW	GWS[2]
Manuka Oval	162.5m	138m	Canberra, ACT	GWS[2]
MCG	161m	138m	Melbourne, Vic	COL, HAW[1], MEL, RIC, CAR[3], ESS[4]
Jiangwan Stadium[7]	160m	136m	Shanghai, China	N/A
Docklands Stadium[13]	160m	129m	Melbourne, Vic	BUL[5], NOR[6], STK CAR[3], ESS[4]
Eureka Stadium[14]	160m	129m	Ballarat, Vic	BUL[5]
Bellerive Oval[15]	160m	124m	Hobart, Tas	NOR[6]
Metricon Stadium	158m	134m	Carrara, Qld	GCS
Gabba	156m	138m	Brisbane, Qld	BRI
SCG	155m	136m	Sydney, NSW	SYD

Table A1: Field Dimensions of AFL fields, sorted by Length

[1] Hawthorn play 7 of their home games at the MCG, and their other 4 home games in Launceston

[2] The GWS Giants play 8 of their home games at the Showgrounds, and their other 3 home games in Canberra.

[3] Carlton play 6 of their home games at Docklands Stadium, and their other 5 home games at the MCG.

[4] Eseendon play 7 of their home games at Docklands Stadium, and their other 4 home games at the MCG

[5] The Western Bulldogs play 9 of their home games at Docklands Stadium, and their other 2 home games in Ballarat.

[6] North Melbourne play 8 of their home games at Docklands Stadium, and their other 3 homes games in Hobart.

[7] Exhibition matches only.

[8] Known as 'UTAS Stadium' for sponsorship purposes.

[9] Known as 'TIO Stadium' for sponsorship purposes. Exhibition matches only.

[10] Known as 'GMHBA Stadium' for sponsorship purposes.

[11] Known as 'Optus Stadium' for sponsorship purposes.

[12] Known as 'GIANTS Stadium' for sponsorship purposes.

[13] Known as 'Marvel Stadium' for sponsorship purposes.

[14] Known as 'Mars Stadium' for sponsorship purposes.

[15] Known as 'Blundstone Arena' for sponsorship purposes.

Division	Team	Location
Eastern Conference		
Atlantic	Boston Celtics	Boston, MA
	Brooklyn Nets	New York City, NY
	New York Knicks	New York City, NY
	Philadelphia 76ers	Philadelphia, PA
	Toronto Raptors	Toronto, Ontario
Central	Chicago Bulls	Chicago, IL
	Cleveland Cavaliers	Cleveland, OH
	Detroit Pistons	Detroit, MI
	Indiana Pacers	Indiana, IN
	Milwaukee Bucks	Milwaukee, WI
Southeast	Atlanta Hawks	Atlanta, GA
	Charlotte Hornets	Charlotte, NC
	Miami Heat	Miami, FL
	Orlando Magic	Orlando, FL
	Washington Wizards	Washington, D.C.
Western Conference		
Northwest	Denver Nuggets	Denver, CO
	Minnesota Timberwolves	Minneapolis, MN
	Oklahoma City Thunder	Oklahoma City, OK
	Portland Trail Blazers	Portland, OR
	Utah Jazz	Salt Lake City, UT
Pacific	Golden State Warriors	San Francisco, CA
	Los Angeles Clippers[1]	Los Angeles, CA
	Los Angeles Lakers[1]	Los Angeles, CA
	Phoenix Suns	Phoenix, AZ
	Sacramento Kings	Sacramento, CA
Southwest	Dallas Mavericks	Dallas, TX
	Houston Rockets	Houston, TX
	Memphis Grizzlies	Memphis, TN
	New Orleans Pelicans	New Orleans, LA
	San Antonio Spurs	San Antonio, TX

Table A2: NBA Teams by Division and Conference

[1] The Clippers and the Lakers are the only two teams in the NBA to share an arena - both teams play out of what is currently known as the Staples Center in downtown Los Angeles

Team	Wins	Losses	Win Rate
Houston Rockets	280	216	0.5645161
Toronto Raptors	285	223	0.5610236
Golden State Warriors	303	243	0.5549451
Boston Celtics	280	231	0.5479452
Milwaukee Bucks	263	220	0.5445135
Denver Nuggets	237	208	0.5325843
Los Angeles Clippers	249	220	0.5309168
Utah Jazz	254	225	0.5302714
San Antonio Spurs	241	216	0.5273523
Oklahoma City Thunder	245	220	0.5268817
Indiana Pacers	234	218	0.5176991
Philadelphia 76ers	233	219	0.5154867
Portland Trail Blazers	239	228	0.5117773
Miami Heat	228	221	0.5077951
Washington Wizards	223	230	0.4922737
Los Angeles Lakers	219	227	0.4910314
Cleveland Cavaliers	244	253	0.4909457
New Orleans Pelicans	227	239	0.4871245
Detroit Pistons	209	230	0.4760820
Charlotte Hornets	210	233	0.4740406
Minnesota Timberwolves	207	231	0.4726027
Memphis Grizzlies	210	236	0.4708520
Dallas Mavericks	208	234	0.4705882
Sacramento Kings	211	241	0.4668142
Orlando Magic	209	242	0.4634146
Brooklyn Nets	209	246	0.4593407
Atlanta Hawks	205	247	0.4535398
Chicago Bulls	209	252	0.4533623
New York Knicks	196	255	0.4345898
Phoenix Suns	198	261	0.4313725

Table A3: Win-Loss Record of NBA teams from 2016 to 2020

AFL Finals System

The top 8 teams in the Home and Away season compete in the AFL Finals series. In all matches, bar the two qualifying finals, the losing team is eliminated whilst the winning team advances to the next round. The winner of a qualifying final receives a bye during the second week before hosting a preliminary final, whilst the loser hosts a semi final in the second week. The format of the series is as such:

Final Name	Home Team	Away Team
Week 1		
1st Qualifying Final (QF1)	1st	4th
2nd Qualifying Final (QF2)	2nd	3rd
1st Elimination Final (EF1)	5th	8th
2nd Elimination Final (EF2)	6th	7th
Week 2		
1st Semi Final (SF1)	QF1 Loser	EF1 Winner
2nd Semi Final (SF2)	QF2 Loser	EF2 Winner
Week 3		
1st Preliminary Final (PF1)	QF1 Winner	SF2 Winner
2nd Preliminary Final (PF2)	QF2 Winner	SF1 Winner
Week 4		
Grand Final	PF1 Winner	PF2 Winner

Table A4: The AFL Final Eight System, used since 2000

The Grand Final is almost always played at the Melbourne Cricket Ground - considered a ‘neutral’ ground despite the fact that 6 teams call it home, as per Table A1. Indeed, since 1902 the MCG has hosted all but 7 Grand Finals, for 3 unique reasons. The Grand Final was held at Princes Park, Carlton in 1942, 1943 and 1945, and Junction Oval, St Kilda in 1944 due to the usage of the MCG for military purposes during the Second World War. In 1991, Waverley Park in Melbourne’s east hosted the game whilst the southern stand of the MCG underwent renovations, whilst the COVID-19 Pandemic forced the Grand Final away from Melbourne for the first time in 2020, and again in 2021. These matches were played at the Gabba, Brisbane, and Perth Stadium, Perth.