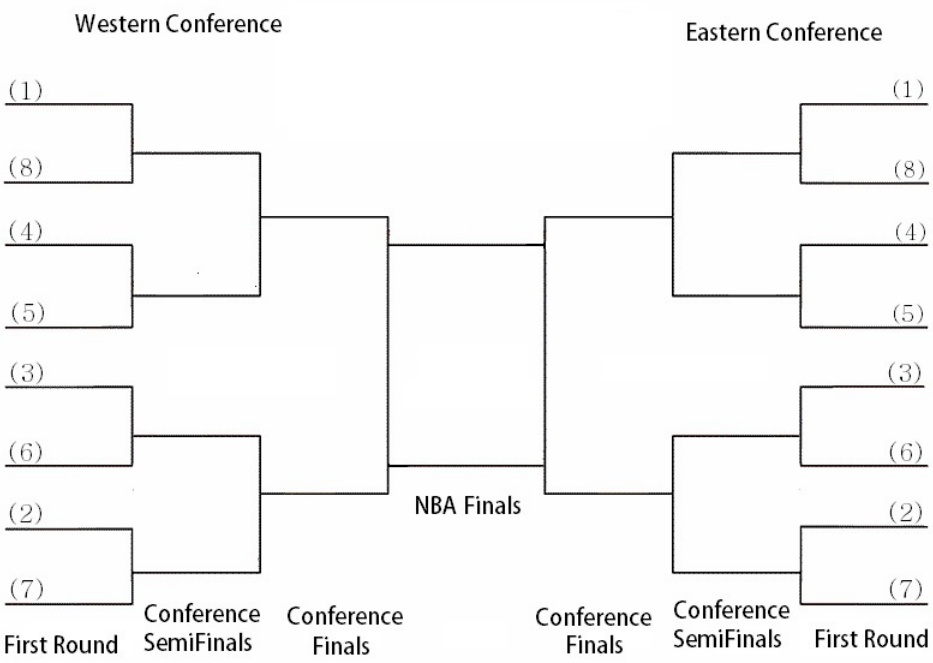
THE UNIVERSITY OF WESTERN AUSTRALIA

Bayesian Forecasting of NBA Playoff Game Results



Author:
Yasith KARIYAWASAM
School of Mathematics and Statistics

Supervisor:
Berwin TURLACH
School of Mathematics and Statistics

This thesis is presented for the partial requirements of the degree of
Bachelor of Science with Honours
of the University of Western Australia
October 20, 2017

Abstract

The NBA has become one of the most watched leagues in the world and the popularity of the sport has opened wider opportunities for punters to speculate on games. The accessibility to a large volume and variety of data has expanded the opportunities for statisticians to use predictive models to speculate results. A large proportion of the research conducted about forecasting NBA outcomes is based on the regular season. Compared to the regular season, the task of predicting playoffs is a much tougher challenge. Research into modelling the outcomes of the playoffs is limited but the results have been moderately successful.

This thesis takes on the challenge of predicting score differences of playoff games using a continuous distribution. Initially, numerous models are fitted to the regular season, the home and away strength estimates from the regular season models are then used as prior information in the predictive process of the playoff games; both models are expressed in Bayesian hierarchical frameworks. Research conducted up to thus far has considered the home effect as a constant term but this thesis investigates the effect as an unique effect for each team. A reasonable level of accuracy is evident throughout the predictions, especially during the first series and Conference Finals of the playoffs.

Acknowledgements

This year would not have been possible if it was not for a number of special people. First of all, I would like to give a massive thank you to my supervisor Berwin Turlach. The guidance, advice and inspiration you have provided me in this project, as well as providing my first taste of Bayesian statistics have all been invaluable experiences.

I would also like to thank my family for their constant support and dealing with my roller-coaster of emotions throughout this challenging year.

Finally a special thank you to my friends, especially my fellow Honours colleagues. I'm grateful for your motivation that kept me sane throughout the year. A special mention to my good friend Kenyon Ng, for assisting and supporting me with countless challenges. Thanks to you, I can proudly write a for-loop!

Contents

Abstract iii

Acknowledgements v

List of Figures ix

List of Tables xi

1 Introduction 1

2 Data Acquisition and Exploratory Data Analysis 7
- 2.1 Web-scraping . . . 7
- 2.2 Exploratory Data Analysis . . . 9
  - 2.2.1 Number of Rest Days . . . 12

3 Methodology 17
- 3.1 Model Details . . . 17
- 3.2 Estimating Team Strengths from the Regular Season . . . 19
- 3.3 Model Overview . . . 23
- 3.4 Playoffs Predictive Model . . . 24

4 Parameter Estimation and Predictive Performance 27
- 4.1 Estimating Strengths from Regular Season . . . 27
- 4.2 Predictive Performance . . . 32

5 Concluding Thoughts and Extensions 39
- 5.1 Possible Extensions . . . 39
- 5.2 Concluding Thoughts . . . 40

Appendices 43

A Appendix 45

List of Figures

1.1 Outline of the NBA Playoffs Bracket 2

2.1 Distribution score difference in the playoffs per series. 10
2.2 Distribution of score differences in the regular season and playoffs . . 11

3.1 Comparing the density function of Student's t-distribution and the standard normal distribution. 18
3.2 The directed acyclic graph (DAG) representation of the regular season hierarchical model. 23

4.1 Mean Home and Away Strengths of Teams Qualified for Playoffs . . 28
4.2 Estimated player strengths top 10 players in the 2015/16 regular season. 29
4.3 Posterior distributions of beta coefficients with 99% intervals 30
4.4 A trace plot, density estimate and autocorrelation of the samples from the posterior distribution of replicated score difference for Round 1 of Hawks vs Celtics. 31
4.5 50% Posterior Prediction intervals for all rounds in Series 1 33
4.6 Prior and posterior distributions of home and away strengths of the Cavaliers and Warriors. 36
4.7 Prior distributions for game 7 of the NBA finals between Cavaliers and Warriors. 37
4.8 Prior and posterior distributions of home effects 38

A.1 Trace plots of games 1-4 in round 1 series 1. 46
A.2 Trace plots of games 5-8 in round 1 series 1. 47

List of Tables

2.1 Days of rest between games and proportion for home teams. 13
2.2 Days of rest between games and proportion for away teams. 14
2.3 Observed Regular Season Results 15

3.1 Data format for dynamic team model. 20

4.1 Prediction accuracy per Series 33
4.2 Predicted results of the first series. 34

A.1 Strength estimates from static player level model. 45

Introduction

The flawless game we now refer to as basketball started with a group of rattled students at the Young Men's Christian Association (YMCA) in Springfield, Massachusetts. As they restlessly awaited for instructions from their physical education teacher, James Naismith. The snow from last night had frozen the school yard into an ice rink. Harsh winter weather denied Naismith's option of an outdoor physical activity. Pressure grew on the inexperienced teacher to settle his students. Naismith had no other choice but to invent a new indoor game involving two peached style baskets and a soccer ball; now known as basketball.

James Naismith's clutch idea to keep his class occupied in 1981 was the origin of the game that is now beloved worldwide. A simple game, innovated in a worn down Massachusetts' gym has now evolved into a culture and reached billions of fans all around the world.

Consisting of 30 teams, the highest and largest level of competition is played in North America known as the National Basketball Association (NBA). The 30 teams are split evenly into a western and eastern conference according to their geographical location.

The Structure of the Regular Season

The NBA regular season consists of 82 games in between the months of October and June. The regular season provides teams a platform to prepare for the playoffs by developing their strengths and team chemistry, improving weaknesses and making team adjustments.

Compared to leagues like the Australian Football League or English Premier League, the structure of the season can be quite unique. Perhaps, the largest discrepancy is the volume of games a team may play during a week. A team can be scheduled to play between one to four games in a week anywhere in the country. This could mean finishing a game on the eastern coast and playing another game the very next day after a five hour flight journey. With an amplitude amount of travelling and inadequate rest, the regular season is physically and mentally demanding on athletes.

Each team will compete against their opponents that are in their own conference four times with equal home court opportunity but only play teams in the other conference twice. For example, Boston Celtics (eastern conference) plays the New York Knicks (eastern conference) four times in the regular season with the home court advantage split evenly. However, Boston will only face a team from the Western Conference twice in the regular season.

As the season progresses, teams are ranked within their conference according to

their win-loss records. These records become vital at the end of the regular season when determining the top eight teams from each conference that will progress to the playoffs.[1]

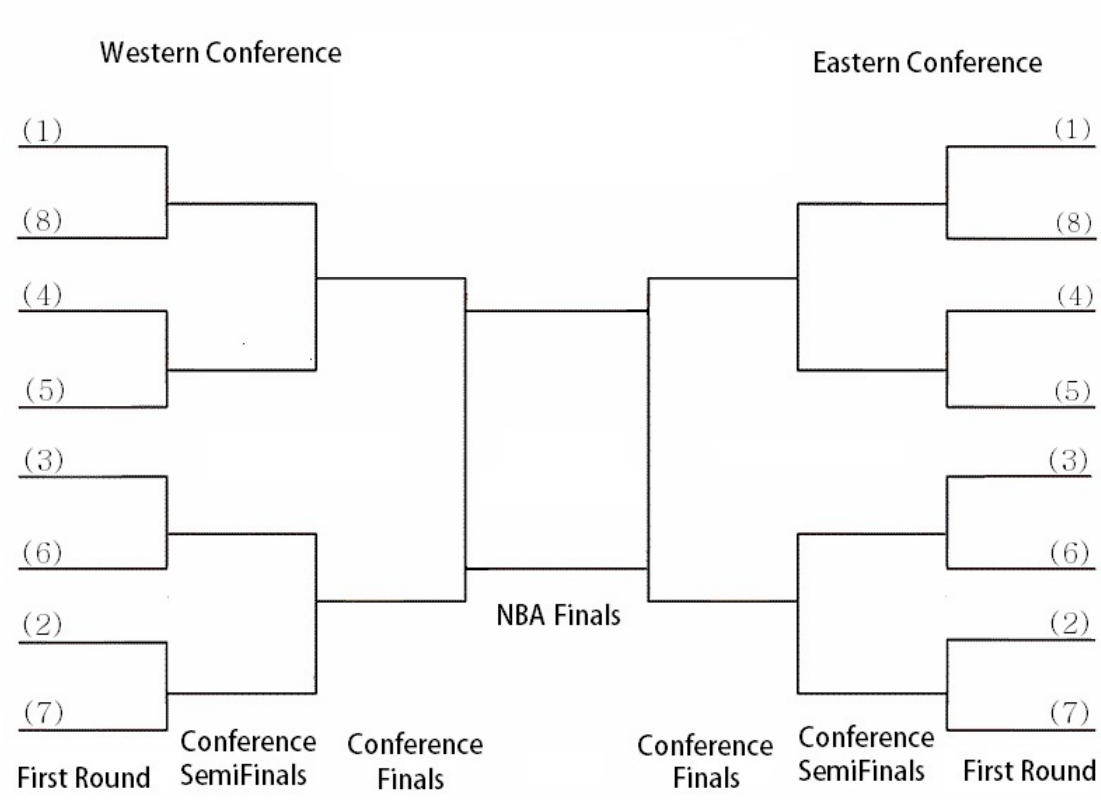


Figure 1.1: Outline of the NBA Playoffs Bracket

The Structure of the Playoffs

Each playoff series is a best of seven games played in a 2-2-1-1-1 format. Games 1, 2, 5 and 7 are played at the home venue of the higher seeded team whereas the lower seeded team will only host games 3,4 and 6. Games 5, 6 and 7 are only played if a winner cannot be determined prior to these games. The final win-loss records from the regular season are used to position teams in the playoffs bracket. An outline of the playoffs bracket is provided in Figure A.2 to help visualise the journey to a championship.

Take the example of the series between the 1st seed and 8th seed. The first two games will be played at the home venue of the first seed while the 8th seed will host the next two. If a team is not able to beat their opposition four times, the series will be extended to game 5, 6 or 7 if necessary. The four winners of the first round

[1]If two teams end the season with the same record, the head-to-head results will determine which team will be ranked the higher seed.

in each conference advance to the conference semifinals, with a match-up between the 1st or 8th seed against the 4th or 5th seed and a match-up between the 2nd or 7th against the 3rd or 6th.

The NBA playoffs structure does not allow teams in opposing conferences to play against each other until the NBA Finals. The final of the playoffs is a series between the best two teams in the western and eastern conference (Figure A.2). Home court advantage is awarded to the team with the higher seed from the regular season. In the event where two teams have the same seed, the home-court advantage will be determined by the head-to-head results in the regular season.

Prior research

The NBA playoffs can be a very exciting period basketball; a period of basketball where legends are born, losers are criticised and champions are crowned. Elimination styled rounds forces the best out of teams; evoking fans' highest attention to the game. The fans are not the only beneficiaries of the playoffs. Media, businesses and betting organisations find the playoffs to be kind to them. The American Gaming Association estimates at least $150 billion dollars invested in sports betting each year by Americans (Purdum; 2017).

NBA attracts a global fan base. Fans from China, India, Australia, Europe pay attention to the outcome of games and their favourite player. With a global fan base, interests for predicting games via betting markets arise. Bookies can offer many various markets for fans; outright winners of the regular season or playoffs, outright winners on individual awards and betting odds for individual games.

Most participants often place their odds subjectively based on their personal opinion and preference without any scientific reasoning. It is of interest of the fans to find or develop systems, algorithms or models to help make informed bets. Statistical models are the ideal tool for this. They have the ability to incorporate various covariates, account for variance with in the season and at least provide the punter with a guide for their bets.

There have been numerous and various attempts of building predictive machine learning models for NBA games. Much of the research is concentrated on outcomes of the regular season. Due to the unpredictability and high variety of performances in playoff games a reliable model for the playoffs is yet to be published. Past research have playoff models based on regular season positions, regular season match ups, home court advantage and win-loss streaks. Some research has extra emphasis on home court advantage (Swartz and Arce; 2014)?, and some believe experienced teams don't need home court advantage so they look to rest their players towards the end of the regular season. Cheng et al. (2016) identifies the problem of predicting games as a classification problem and apply the principle of maximum entropy to construct a NBA maximum entropy model that fits to the discrete statistics of NBA games,

and then predict the outcome of playoffs game with the model.

Teramoto and Cross (2010) discusses the relative importance of performance factors in winning basketball games in the past 10 years of the NBA. Specifically, the contributions of overall efficiency in defence and offense to winning games in the regular season and the playoffs. Using a multiple linear regression and a logistic regression Teramoto and Cross (2010) found the importance of defence in winning games may be greater in the playoffs than in the regular season. Shooting efficiency on both ends of the floor (offensive and defensive effective field goal percentages) seems to be the most critical aspect of the game in the regular season as well as the playoffs. Like Teramoto and Cross (2010) there were numerous models that used individual player statistics, team ratings in offense and defence to build team strengths. We didn't want to head down a path similar to this, as it would have required more time to understand the efficiency ratings and extract the relevant data for them.

Perhaps the most recognised prediction model in the basketball community is of Nate Silver's. Nate Silver is a well-recognised statistician who has published several analytical articles in regards to sports and politics. In 2015 Nate Silver and his team at FiveThirtyEight released their first of many predictive model for NBA regular season games. The model incorporates an Elo Rating system similar to that of chess. Each team is assigned a rating based on the final score and whether it is played at home or away. Ratings are dynamic as they are updated after each game. This method is simple and requires minimal data as it only requires the final score and location of games. However, the Elo rating system lacks the capability to account for mid-season trades and player movements during the season. This dissertation accounts for this weakness. If a player is traded mid-season, they are introduced as a new player on the understanding that their output in the new team will be different to their previous team.

We were intrigued by the volume of research conducted for football data. Baio and Blangiardo (2010) developed a Bayesian hierarchical model for the number of goals scored by the two team in each match. Their method was implemented on predicting winner of the 2007-2008 Italian Serie A championship. Baio and Blangiardo (2010) used linear relationship between home field advantage (for home strengths) and attack and defence strengths to estimate team strengths for each week. The estimate of the strength of a team from the linear regression would be used as the parameter in the Bivariate-Poisson. Due to the high scoring nature of basketball games, it would be difficult to model the scoring intensity of basketball using a Poisson model.

However, Kharratzadeh (2017) showed it could be possible to use a continuous distribution to model the score differences. His research incorporated a t-distribution in a hierarchical Bayesian model to forecast score differences in English Premier League games.

Outline of Thesis

The aim of this dissertation is to build a statistical predictive model to forecast NBA playoffs series games and rounds. The remaining chapters of this thesis are as follows. Chapter 2 contains a brief explanation of how data was acquired, describes the process of web scraping, and the exploratory analysis of the data. Chapter 3 discusses various model implemented, including regular season models used to train model parameters. Chapter 4 contains a discussion on model fitting and results. Chapter 5 is a collection of possible extensions and concluding thoughts.

Data Acquisition and Exploratory Data Analysis

Given its popularity, one may think obtaining historical data on NBA results to be a simple task. Multiple web pages provide historical data only for the games played in the ongoing 2016/17 season. But for the scope of this thesis, we require data on game results and player statistics for all games in the 2015/16 season [1] However, basketball-reference is an exception to this. The web page is a thorough database with historical data dating back to 1947 presented in hyper text mark-up language (HTML) format. Thus, problems arise in storing the information locally from the web page. Entering a season of data with 1230 games is an extremely time consuming task. Which brings us to one of the earliest challenges of this thesis, acquisition of data through web-scraping.

2.1 Web-scraping

Web-scrapping is a method used to convert information presented in a unstructured format from web pages to a structured format which can be stored locally and used. With the appropriate tools, web scrapping can provide an efficient method of acquiring data. Hadley Wickham's rvest package (Wickham; 2016) is used to scrape basic box-score information of the 2015/16 on each game from basketball-reference. For each of the games in the 2015/16 season we need to obtain data on game results, location of the game, date and as well as information about the players and their time played on court.

The rvest package has logical functions that simplifies the data collecting process. For example, by simply providing the URL to the `html()` function we can store the HTML codes and data of that web page locally. If we need to select only parts of the web page, in our case just the basic box-score statistics we can use the `html_nodes()` function to extract parts of the web page. As an assignment to the `html_nodes()` function we need to enter the cascading style sheet (CSS) node that allows the function to identify the data we require. We can use the open source software named Selector Gadget tool to assist us in finding the CSS nodes. The Selector Gadget tool is a Google chrome extension that can generate the simplest CSS node for an element by simply clicking on it. For more information visit `www.selectorgadget.com`.

The iterative process for web-scraping can become quite complicated if the URLs are not created in a logical manner. Fortunately basketball-reference has used the date of the game and an abbreviation of the home team for its URLs. All URLs contain a common hyper link at the start, followed by a unique date and home team abbreviation. For example the first game of the season that took place in Atlanta on October 27, 2015 has the URL `www.basketball-reference.com/boxscores/201510270ATL.`

[1] Analyses on 2016/17 season was not possible as the season was yet to completed at the start of this thesis.

`html`. As one can see the end of the URL is constructed using the date of the game followed by an abbreviation of the home team. Using this logic, we create an iterative loop to produce URLs of each game and save it.

Once we have each URL we create an iterative process to extract the data we need from each web page. A general outline of the iterative process is as follows:

1. Load the required R packages.
2. Create the unique URL for each game and store it locally.
3. Extract data from each web page.
   (a) Use `regexpr()` and `strsplit()` to extract the date of the game from the URL.
   (b) Extract the team names using `html_nodes()` , use selector gadget to obtain CSS node.
   (c) Extract the box score table using `read_table()`, use selector gadget to obtain CSS node.
   (d) Extract the location of the game, use `strsplit()` to get rid of unessacary information.
   (e) Extract the final score of the game for the home and away team.
4. Save data onto a data frame.

An example of this iteration for the first game of the season is provided in Listing 2.1.

```
#Use the URL from the first game
#www.basketball-reference.com/boxscores/201510270ATL.html
url1 <- "www.basketball-reference.com/boxscores/201510270ATL.html"

#extract the date of the game
start <- regexpr("[[:digit:]]", url1)
date <- substr(url1, start, start + 7)

#read the url as an html
xxx <- read_html(url1)
#extract the team names played
teams <- html_text(html_nodes(xxx, css="strong a"))

# extracting the box score table using html_table()
yyy <- html_nodes(xxx, ".stats_table[id*='_basic']")
zzz<-  html_table(yyy)

#extract the date of game
www <-  html_text(html_nodes(xxx, css=".scorebox_meta div"))
Location_1 <- www[2]
Location_2 <- strsplit(Location_1, ",")
Location <- as.character(Location_2[[1]][2])
trim.leading <- function(t) sub("^\\s+", "", t)
Location <- trim.leading(Location)

#extract final score of the game
qqq <-  html_text(html_nodes(xxx, css=".score"))
HS <- as.numeric(qqq[1])
AS <- as.numeric(qqq[2])
```

Listing 2.1: An example of the web-scraping process

2.2 Exploratory Data Analysis

Once the data is web-scraped and stored locally, we conduct an exploratory data analysis to investigate patterns, the shape, spread and frequency of our data. This is an important step as it may assist with modelling and variable selection of the predictive process. Exploratory data analysis will be conducted on the score differences for regular season and playoffs, home effect and days of rest between home and away teams.

Score Differences

The model in this thesis is built to predict games through the score difference of each game, with the assumption that the differences are symmetrically distributed around zero. The histograms represented in Figure 2.2 display the distribution of score differences in the 2015/16 regular season and playoffs respectively. The histogram of the regular season score differences confirms they are reasonably symmetric and centred around five, which could possibly reflect the home court advantage. We

are also interested to observe if both histograms reflect similar shape and spread as it will allow us to approach modelling both in a similar fashion. Although we expect playoff games to be more tightly contested compared to the regular season, the histogram of score differences reflect a similar range of that to the regular season.

In the predictive process explained in Chapter 3 we assume the score differences are distributed similarly and each series of the playoffs. The boxplots in Figure 2.1 display the distribution of score differences for each playoff series, where 1 denotes the First Round and 5 denotes the NBA Finals. The spread of the score differences upholds the assumption as box plots are evident to be approximately similar in each series.

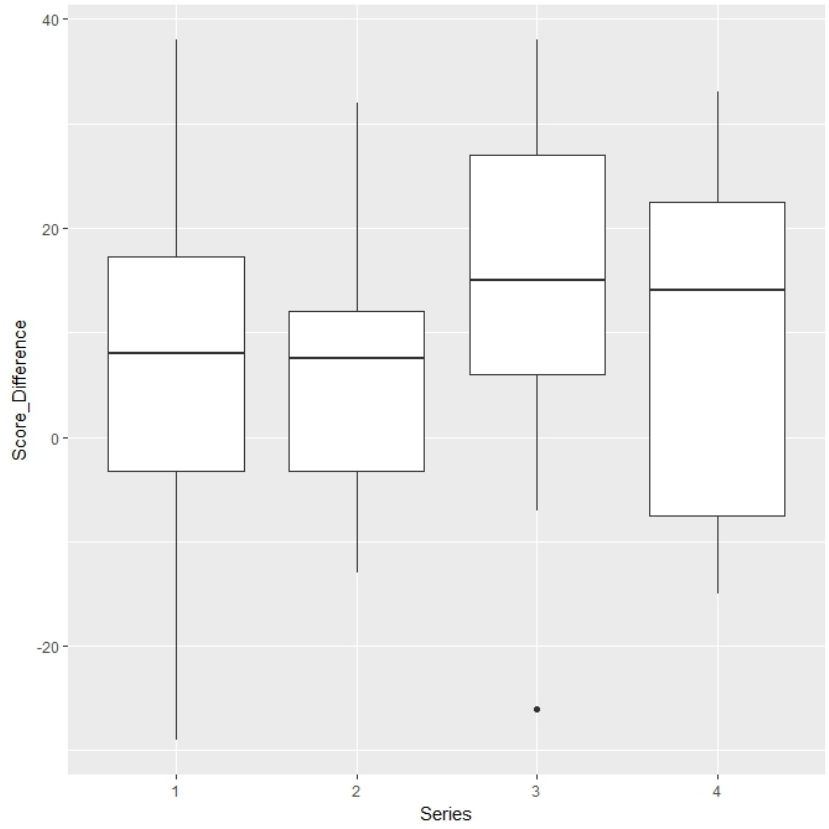
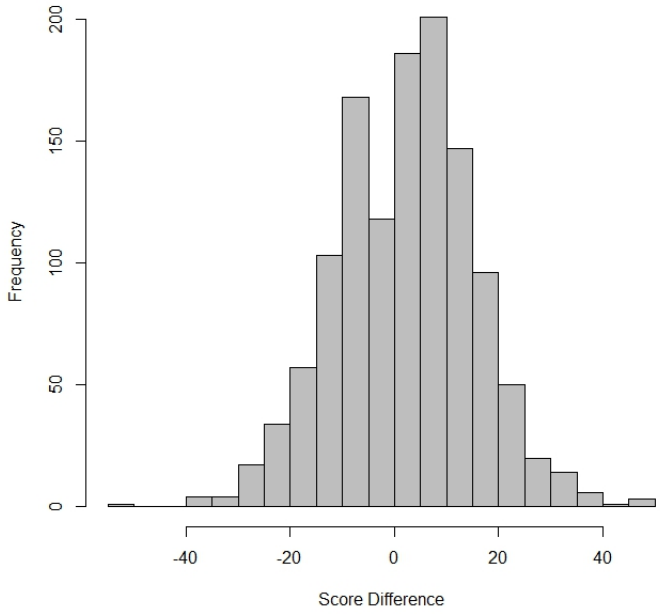
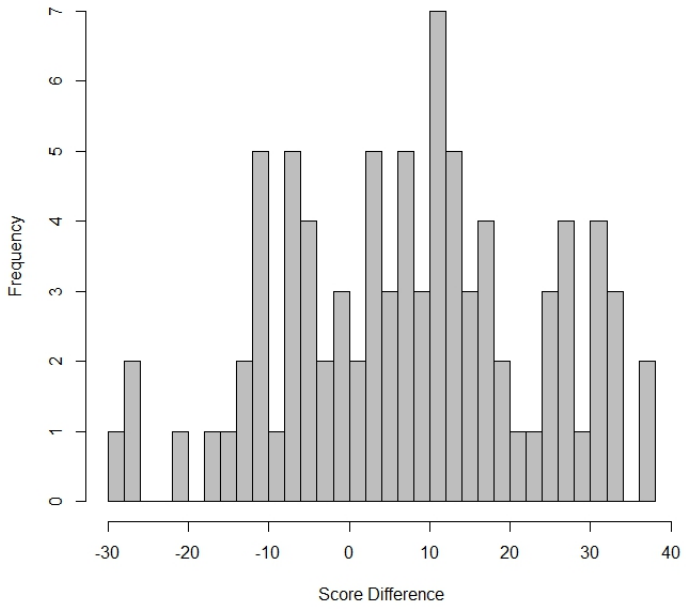


Figure 2.1: Distribution score difference in the playoffs per series.



(a) Regular season.



(b) Playoffs.

Figure 2.2: Distribution of score differences in the regular season and playoffs

Home Effect

In most sports there is a common assumption that a team's performance likely to vary whether they are playing at home or away. Before including any home effect in models, we need to explore whether there is any effect at all and it's consistency between teams. A summary of teams' performance at home and away are provided in Table 2.3. From the evidence of average difference in points at home compared to away, it is evident that teams tend to perform better at home compared to away games and the effect also does not seem to be consistent across all teams. For example, the average difference in points of home and away for the Miami Heat is 6.12. Whereas the Houston Rockets has a slightly positive advantage for away games, with -0.56 point difference. This can't be used as conclusive evidence, but the information gained in the exploratory data analysis can certainly be of assistance during the modelling process. We will explore a model with a home effect parameter that is unique for each team and investigate whether there is statistical evidence of a unique effect.

2.2.1 Number of Rest Days

The irregular structure of the NBA may cause advantages and disadvantages for teams and these effects may differ whether they play at home or away. Table 2.1 and 2.2 presents a summary of the number of games and winning proportion when played with various days of rest. The first column presents the scenario where teams have to play a game a day after their previous game with no rest days in between. It is evident not all teams play the same amount of back to back games. Comparing the proportion of games won with no days of rest for home and away, the away teams are at disadvantage compared to home. Team's are very rarely given the opportunity of four of more days of rest and the most two or three days of rest are the most common occurrence. The number of days of rest for home and away teams will be investigated in the modelling of the regular season.

Table 2.1: Days of rest between games and proportion for home teams.

,

ID	Team	0 Days of Rest	Proportion Won	2 or 3 Days of Rest	Proportion Won	4 or more Days of Rest	Proportion Won
1	Hawks	11	0.82	26	0.62	2	1.0
2	Celtics	8	0.75	31	0.65	1	1.0
3	Nets	2	0.50	37	0.32	0	
4	Hornets	9	0.67	32	0.75	0	
5	Bulls	5	0.40	33	0.64	1	1.0
6	Cavaliers	7	0.86	31	0.77	2	1
7	Mavericks	4	0.50	37	0.57	0	
8	Nuggets	2	1.0	39	0.46	0	
9	Pistons	9	0.78	30	0.63	0	
10	Warriors	5	1.0	33	0.94	2	1.0
11	Rockets	8	0.38	30	0.67	0	
12	Pacers	6	0.67	35	0.63	0	
13	Clippers	6	0.83	33	0.7	0	
14	Lakers	5	1.0	30	0.33	2	0.5
15	Grizzlies	6	0.67	32	0.62	1	1.0
16	Heat	2	1.0	37	0.68	0	
17	Bucks	6	0.67	31	0.55	2	1.0
18	Timberwolves	5	0.20	35	0.37	0	00
19	Pelicans	6	0.50	32	0.5	1	0.5
20	Knicks	7	0.57	33	0.39	1	1.0
21	Thunder	2	1.0	37	0.78	0	
22	Magic	4	0.75	33	0.55	1	0.5
23	76ers	7	0.14	31	0.19	0	
24	Suns	1	1.0	37	0.35	1	1.0
25	Blazers	7	0.43	30	0.7	2	1.0
26	Kings	4	0.25	32	0.47	1	0.33
27	Spurs	4	1.0	36	0.97	1	1.0
28	Raptors	4	0.50	35	0.8	1	1.0
29	Jazz	6	0.67	34	0.59	0	
30	Wizards	7	0.29	30	0.57	2	0.67

Table 2.2: Days of rest between games and proportion for away teams.

,

ID	Team	0 Days of Rest	Proportion Won	2 or 3 Days of Rest	Proportion Won	4 or more Days of Rest	Proportion Won
1	Hawks	8	0.38	30	0.57	3	0.66
2	Celtics	11	0.55	27	0.48	2	1.0
3	Nets	13	0.08	27	0.19	1	0
4	Hornets	7	0.86	31	0.35	1	1.0
5	Bulls	11	0.36	27	0.37	2	1.0
6	Cavaliers	12	0.42	27	0.67	1	1.0
7	Mavericks	13	0.62	25	0.4	1	1.0
8	Nuggets	14	0.36	25	0.36	0	
9	Pistons	11	0.45	27	0.41	1	0
10	Warriors	15	0.87	25	0.84	0	
11	Rockets	12	0.42	27	0.44	1	0
12	Pacers	11	0.27	27	0.52	1	1.0
13	Clippers	14	0.50	24	0.58	2	0.50
14	Lakers	13	0.08	27	0.15	1	0
15	Grizzlies	12	0.42	27	0.37	2	0.50
16	Heat	15	0.40	25	0.52	0	
17	Bucks	14	0.29	27	0.22	0	
18	Timberwolves	9	0.11	30	0.43	0	
19	Pelicans	11	0.09	28	0.29	1	0
20	Knicks	10	0.40	29	0.31	0	
21	Thunder	14	0.29	27	0.7	0	
22	Magic	15	0.20	25	0.36	1	0
23	76ers	12	1.0	27	0.11	0	
24	Suns	13	0.31	27	0.19	1	0
25	Blazers	12	0.25	29	0.45	0	
26	Kings	15	0.20	24	0.42	2	0.50
27	Spurs	12	0.83	25	0.6	1	1.0
28	Raptors	13	0.69	24	0.5	3	0.33
29	Jazz	12	0.5	27	0.37	0	
30	Wizards	13	0.38	26	0.46	1	1.0

Table 2.3: Observed Regular Season Results

	Team	Home	Home Lose	Average Home Score	Away Win	Away Lose	Average Away Score
1	Atlanta Hawks	27	14	103.59	21	20	102.10
2	Boston Celtics	28	13	106.05	20	21	105.39
3	Brooklyn Nets	14	27	98.93	7	34	98.37
4	Charlotte Hornets	30	11	105.46	18	23	101.34
5	Chicago Bulls	26	15	101.83	16	25	101.46
6	Cleveland Cavaliers	33	8	106.63	24	17	102.02
7	Dallas Mavericks	23	18	104.32	19	22	100.27
8	Denver Nuggets	18	23	103.66	15	26	100.12
9	Detroit Pistons	26	15	105.05	18	23	98.88
10	Golden State Warriors	39	2.00	116.27	34	7	113.51
11	Houston Rockets	23	18	106.27	18	23	106.83
12	Indiana Pacers	26	15	102.90	19	22	101.41
13	Los Angeles Clippers	29	12	104.80	24	17	104.20
14	Los Angeles Lakers	12	29	96.80	5	36	97.88
15	Memphis Grizzlies	26	15	101.85	16	25	96.34
16	Miami Heat	28	13	103.22	20	21	96.88
17	Milwaukee Bucks	23	18	100.85	10	31	97.24
18	Minnesota Timberwolves	14	27	101.51	15	26	103.32
19	New Orleans Pelicans	21	20	105.93	9	32	99.51
20	New York Knicks	18	23	98.83	14	27	97.88
21	Oklahoma City Thunder	32	9	109.54	23	18	110.90
22	Orlando Magic	23	18	104.51	12	29	99.61
23	Philadelphia 76ers	7	34	98.07	3	38	96.76
24	Phoenix Suns	14	27	103.10	9	32	98.63
25	Portland Trail Blazers	28	13	108.05	16	25	102.24
26	Sacramento Kings	18	23	107.56	15	26	105.61
27	San Antonio Spurs	40	1	105.15	27	14	101.93
28	Toronto Raptors	32	9	104.95	24	17	100.46
29	Utah Jazz	24	17	98.56	16	25	96.80
30	Washington Wizards	22	19	106.02	19	22	102.12

Methodology

As we mentioned earlier, the aim of this dissertation is to be able to predict NBA playoff game rather than regular season results. However, regular season information is vital for the model to understand and learn about team strengths and weaknesses. We first investigated models for regular season data at the team level with covariates that contributes directly to team strengths. Models are also fitted at the player level, taking into account individual player strengths and their proportion of time on court. The first part of this chapter will briefly explain details of the statistical models used for regular season followed by the model specifics used for the playoffs.

3.1 Model Details

Derivation of the t-distribution

Suppose Z and U are two independent random variables with Z as the standard normal distribution and U a chi-square distribution with ν degrees of freedom. The random variable

$$T = \frac{Z}{\sqrt{\frac{U}{\nu}}}$$

follows a Student's t-distribution when $\nu > 0$, ν is the degrees of freedom, not nesscarily the integer. The probability density function is given by

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

where Γ is the gamma function. Traditionally Student's t-distribution occurs in classical statistics as test statistics for parameters whose magnitude of variability has to be estimated from the data. Whereas in Bayesian statistics, it is often used as an error distribution in robust regression methods to identify for outliers.

The shape of the probability density function is similar to the bell shape nature of the standard normal distribution but with heavier tails (Figure 3.1). It will approach a standard normal distribution as the degrees of freedom tends to infinity (Figure 3.1). This becomes an appealing attribute of the distribution in the case where weakly informative priors are desirable.

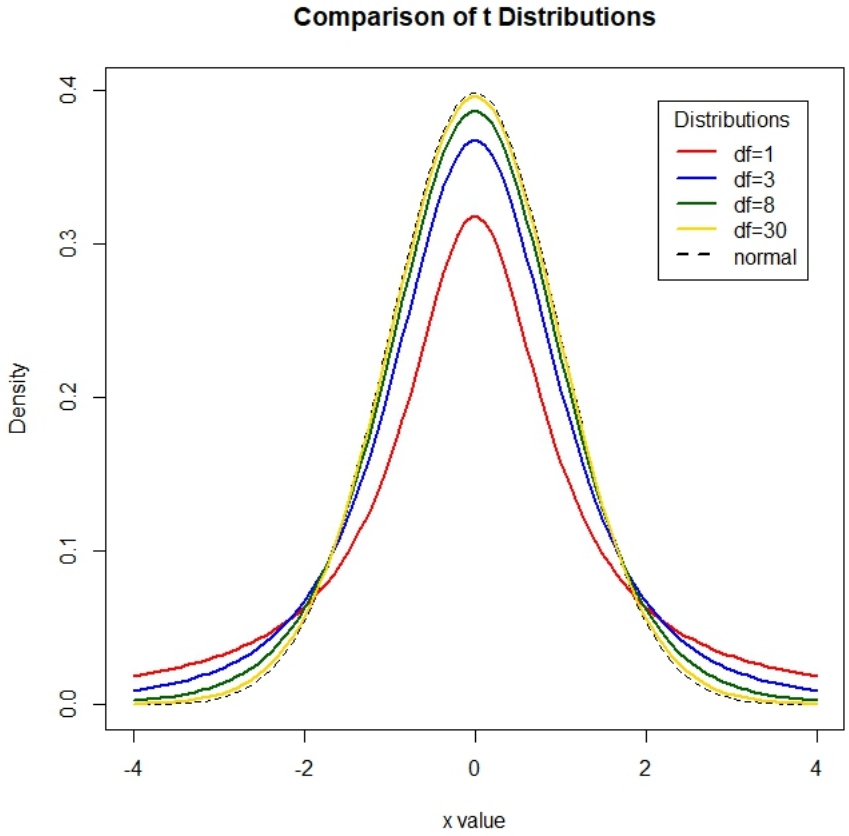


Figure 3.1: Comparing the density function of Student's t-distribution and the standard normal distribution.

Cauchy Distribution

In this thesis, the Cauchy distribution is used often as prior distributions for variance parameters. For large values of the scale parameter in the Cauchy distribution, the gentle slope in the tails allows to be used as a weakly informative prior distribution (Polson et al.; 2012). Gelman et al. (2006) discusses the issues of using the inverse-gamma family of non-informative prior distributions and recommends the Half-Cauchy prior distribution when a weakly informative prior is desired.

The Cauchy distribution is a special case of the Student's t-distribution with 1 degree of freedom. Let X_1 and X_2 be independent random variables with a standard normal distributions

$$X_1 \sim N(0,1), \quad X_2 \sim N(0,1),$$

the ratio $\frac{X_1}{X_2}$ will follow a Cauchy distribution which has the following probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

3.2 Estimating Team Strengths from the Regular Season

Dynamic Team Level Model

Our initial approach for estimating strength parameters in the regular season follows a similar to approach Kharratzadeh (2017) hierarchical model for English Premier League outcomes. The score difference of a game, denoted as y_g is modelled using a Student's t-distribution

$$y_g \sim t_\nu(\mu_g, \sigma_y), \tag{3.1}$$

where

$$\mu_g = \alpha_g + \pi. \tag{3.2}$$

The difference in team strengths for each game is denoted by α_g. We denote the strength of the home team and away team at week j as $\gamma_{j,1}$ and $\gamma_{j,2}$ respectively. Thus, the α_j is computed as

$$\alpha_j = \gamma_{j,1} - \gamma_{j,2}, \quad j = 1, 2, ..., 82. \tag{3.3}$$

We need to denote the home and away week in Equation 3.3 due to the uneven nature of the NBA structure. When two teams play against each other, the number of games played by the home and away team may not be identical.

Harville and Smith (1994) study on the macroscopic home court advantages in NBA observed teams playing at home to win around 60% of the games. Within our study of the 2015-16 season we observed the home team to win 52% of the games. Thus, the minor effect for playing at the home venue is introduced as π in Equation 3.4. The home effect is given the same prior for all teams and used as an additive term to α_i. A standard normal prior is placed on the home advantage:

$$\pi \sim N(0, 1). \tag{3.4}$$

The variation of the score differences denoted as σ_y in Equation 3.1, is given an uninformative prior:

$$\sigma_y \sim N(0, 5), \qquad \text{truncated to} \quad [0, \infty). \tag{3.5}$$

During the long and tenuous NBA season, we expect team abilities to change due to injuries, trades, player acquisitions and improvements in team chemistry. The evolution of team abilities is modelled as an AR[1]. Specifically, the ability evolution of a home and away team from week $j-1$ to week j is modelled as

$$\gamma_{j,1} \sim (\gamma_{j-1,1}, \sigma_a), \qquad \gamma_{j,2} \sim (\gamma_{j-1,2}, \sigma_a) \qquad j = 1, 2, ..., 82 \tag{3.6}$$

In the hierarchical nature of this model, the variance is sampled from

$$\sigma_{a,j} \sim N(0, \tau_a), \qquad \text{truncated to} \quad [0, \infty). \tag{3.7}$$

where τ_a is the hyper-parameter for game-to-game variation and is sampled from a Cauchy distribution

$$\tau_a \sim Cauchy(0, 1). \tag{3.8}$$

Data Structure

Extracting our own data through web-scraping has its disadvantages. For instance, a large proportion of time was spent on cleaning and formatting the scraped dataset. The data structure required to implement the model is presented in Table 3.1. Multiple data cleaning processes are required to clean and format the data structure from the web-scrapped structure to Table 3.1.

To save tedious time and effort from more data cleaning, all the models fitted have a common data structure. The data is structured in a manner where each game g contains the name and identification key of teams playing, and the score difference. The indexes $h(g)$ and $a(g)$ are uniquely associated with one of the 32 teams. For example, in Table 3.1 the Atlanta Falcons are always associated with the identification key 1 regardless of home or away.

Table 3.1: Data format for dynamic team model.

g	Home	Away	h(g)	a(g)	score_diff(g)
1	Atlanta	Detroit	1	9	-12
2	Chicago	Cleveland	5	6	2
3	New Orleans	Golden State	10	19	16
4	Boston	Philadelphia	2	23	17
...	...	...	...	...	...
1229	Portland	Denver	25	8	8
1230	Washington	Atlanta	30	1	11

Static Player Level Model

A static model on the team abilities with an emphasis on player performance is also investigated. The length of rest days between games is considered in the process of estimating home and away strengths.

Basketball is commonly perceived as a team sport, however with only 5 players per team on the court extremely talented players can turn the game into their individual show. Historically, we have observed dominant players take over games and in some cases dominate the league over multiple seasons. Take Michael Jordan, arguably the greatest player of all time. His abilities were far superior than his opponents such that if we knew Jordan was playing a large proportion of game, the Chicago Bulls

were considered the favourite to win. We incorporate this simple and effective logic to estimate the home and away team strengths.

The score difference for each game is still modelled using a Student's t-distribution

$$y_g \sim t_\nu(\mu_g, \sigma_y), \tag{3.9}$$

where

$$\mu_g = \theta_g - \phi_g + \pi_{home\_team_{(g)}} + \beta_{XH} R_X + \beta_{XA} R_X. \tag{3.10}$$

For each game, the home and away ability is estimated by a weighted sum of individual player abilities

$$\theta_g = \sum_{n_h=1}^{n_h} W_{(i,g,1)} * \rho_{player(i)}, \quad \phi_g = \sum_{n_a=1}^{n_a} W_{(i,g,2)} * \rho_{player(i)}. \tag{3.11}$$

Let θ_g be the estimated home ability for a game g and $W_{(i,g,1)}$ the proportion of time played by player i in game g for the home. The parameter $\rho_{player(i)}$ is the estimated player strength. Similarly, the away team ability denoted as ϕ_g is the weighted sum of the player strengths and proportion of time played by the away team players.

The weights denoted as $W_{(i,g,1)}$ and $W_{(i,g,2)}$, are the proportions of minutes played by player i in game g for the home and away team respectively. The number of players played in an NBA game can vary between five and twelve; thus, n_h and n_a are introduced as the number of players played by the home and away team in game g respectively, and m denotes the time in minutes played. The weights are computed by the following equation

$$W_{(i,g,1)} = \frac{m_{(i)}}{\sum_{n_h=1}^{n_h} m_{(i)}}, \qquad W_{(i,g,2)} = \frac{m_{(i)}}{\sum_{n_a=1}^{n_a} m_{(i)}} \tag{3.12}$$

Player strengths ρ_i are treated as a fixed effects over the season[1]. The standard deviation of a player's ability at home will differ for each player, so we sample from a normal distribution of unknown variance:

$$\rho_i \sim N(0, \sigma_a), \quad \psi_i \sim N(0, \sigma_a), \tag{3.13}$$

Where σ_a is sampled from a half Cauchy prior of

$$\sigma_a \sim Cauchy(0, 5), \qquad \text{truncated to} \quad [0, \infty). \tag{3.14}$$

[1]it is unlikely a player's ability will be constant over a season, but due to time and complexity reasons we will concentrate on a static version of this model.

Length of Rest - The NBA Scheduling Problem

Every year, the league office has a tough challenge in creating a schedule satisfying various constraints while minimising the teams' total travel distance. To our knowledge, there has been at least three studies on the complexity of building a efficient NBA schedule: see Bean and Birge (1980), Bao (2009) and Ashman et al. (2010).

Ashman et al. (2010) research focused on the role of fatigue in NBA wagering markets. Nineteen years of historical data showed that betting markets tend to repeatedly misprice betting spreads in situations when the home team was playing a game without any days of rest after their previous games whereas the away team had 1 or 2 days of rest. Ashman et al. (2010) also found evidence for betting lines to be most inefficient when the home teams played with no rest but away teams had 1 or 2 days of rest and the away team were considered as the underdog. Whereas, Entine and Small (2008) study related to the role of rest on NBA home court advantage found evidence to days of rest playing an important role in the margins of victory in NBA games.

We will consider the cases which a home or away team had 1,2,3 or more days of rest between games. In Equation 3.10 X represents the number of rest days between games, and $XR_(g)$ is an indicator variable to represent whether a team had rested X days. In Equation 3.10, π is treated as the intercept variable. Thus, we are unable to use all indicator variables as this will lead to identifiable problems. For a categorical variable with k levels, $k-1$ dummy variables are used with each dummy variable coded as 0 or 1 (Anderson et al.; 2014).

For each coefficient of the indicator variables we use the following uninformative normal priors

$$\beta_{XH} \sim N(0, 10), \qquad \beta_{XA} \sim N(0, 10)$$

Unique home advantage

Many research papers considered the home effect as a fixed effect since they assumed it to be constant for all teams. For example, Baio and Blangiardo (2010) study on the outcomes of soccer results considered home advantage to be constant for team in the Italian Serie A and in Andrew (2015) study of forecasting Australian Football League results the home effect is considered constant.

The reasons behind home field advantage may be mysterious, in this thesis the home effect is used as an unique but static estimate for each team in the league. In Equation 3.10 the home effect is denoted by π_g, and modelled as

$$\pi_g \sim N(0, \sigma_b). \tag{3.15}$$

We expect the variance of the home effect to behave differently for teams, thus sampling its standard deviation from a half Cauchy distribution with the prior:

$$\sigma_b \sim Cauchy(0, 5), \qquad \text{truncated to} \quad [0, \infty). \tag{3.16}$$

3.3 Model Overview

The directed acyclic graph (DAG) represented in 3.2 is similar to Baio and Blangiardo (2010) and Andrew (2015). The observed information of the proportion of minutes played at home and away are represented as *Home Weights* and *Away Weights* respectively. All other nodes are parameters that sampled from prior distributions. Unlike Baio and Blangiardo (2010), the hyper prior distribution for the initial nodes are not represented in this figure.

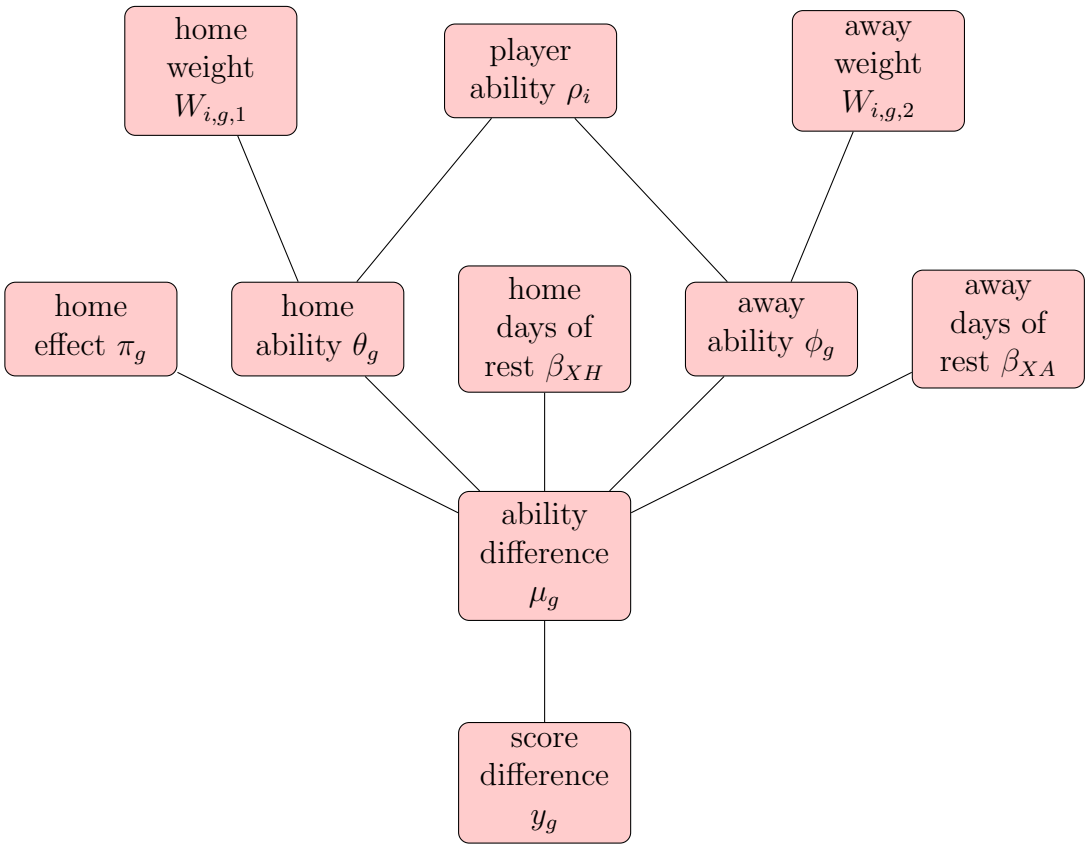


Figure 3.2: The directed acyclic graph (DAG) representation of the regular season hierarchical model.

3.4 Playoffs Predictive Model

Playoffs games are modelled similar to that of the regular season with a few minor adjustments. The score difference for each round[2], denoted as y_g is modelled using a Student's t-distribution

$$y_g \sim t_\nu(\mu_g, \sigma_y), \tag{3.17}$$

where

$$\mu_g = \theta_{g,h[g]} - \theta_{g,a[g]} + \pi_g. \tag{3.18}$$

Each playoffs series is played in a 2-2-1-1-1 format (see Chapter 2). Meaning, the away strengths of the home in team and home strengths of the away teams in the first two games will not be updated by the posterior distributions of the first two games. To incorporate this, we use a multivariate normal distribution using the priors for team i based on the regular season previous games during the playoffs. The correlations of home and away strengths are calculated using a hypothetical parameter $\phi_{g,i}$, if the team was to play the same game away. The prior distribution of the home and away strength is as follows

$$\begin{pmatrix} \theta_{g,i} \\ \phi_{g,i} \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_i \\ \theta_i \end{pmatrix}, \begin{pmatrix} \sigma^2_{i,1} & \rho_i \sigma_{i,1} \sigma_{i,2} \\ \rho_i \sigma_{i,1} \sigma_{i,2} & \sigma^2_{i,2} \end{pmatrix} \right]$$

The variation of score differences is denoted as σ_y with a half Cauchy prior

$$\sigma_y \sim Cauchy(0, 5), \qquad \text{truncated to} \quad [0, \infty).$$

The degrees of freedom noted as ν has a prior with a Gamma distribution, specifically,

$$\nu \sim Gamma(2, 0.1).$$

The model used for the playoffs incorporates a unique but static home court advantage for each team. Research on home court advantage in playoffs and regular season show that its harder to win in away games in the playoffs games than regular seasons (Paine; 2017). Historical data from 1998-2008 showed homed teams in regular season won had a winning percentage of 60.6% in the regular season When it came to the playoffs, their winning percentage increased to 65% (Paine; 2017).

It can be debated that some of this disparity is simply due to the nature of playoffs seeding, since the teams finishing with better regular season records are given home-court advantage in the playoffs. In general, home court advantage is an area that requires and deserves more research due to the absence of credible explanation for this phenomenon. Either way, each team is assigned an individual home court

[2]During our modelling process we refer to a round as a game played in a particular series.

advantage effect as it would be unfair to give the same home effect of a higher seeded team to a lower seeded.

The home advantages for each teams in the playoffs are estimated in a hierarchical nature. The estimation of the home effect (π_g) is

$$\pi_g \sim N(\Lambda, \sigma_b^2),$$

where the mean is sampled from a uninformative normal distribution

$$\Lambda \sim N(0, 10)$$

and the variance is given a half Cauchy prior

$$\sigma_b^2 \sim Cauchy(0, 5), \qquad \text{truncated to} \quad [0, \infty).$$

Parameter Estimation and Predictive Performance

This chapter will discuss the parameter estimates from the regular season and the predictive performance of the playoffs model. The regular season is used as a source of information to understand home and away strengths of playoff teams. These strengths are then used as prior distributions for the playoffs and predictions are then made per playoff round. Estimates from both models are presented in this section and can be cross checked with observed results in the regular season (win/loss records and seeds). The model that resembles the observed results should be considered to use as prior distributions for the playoffs.

4.1 Estimating Strengths from Regular Season

Markov Chain Monte Carlo Estimation

Markov Chain Monte Carlo (MCMC) simulations are common technique in Bayesian statistics; used to solve the problem of sampling from a complicated posterior distribution. MCMC simulations can be simplified to follow a general concept. As the name indicates, the method is composed of two components, the *Markov Chain* and *Monte Carlo* integration. The Markov Chain component is to solve the common problem of sampling from the distribution of a complexed parameter. Markov Chains are constructed as sequential models that transition from one state to another, in a probabilistic fashion; the next state the chain samples from is conditioned on the previous. The process is continued iteratively until samples have converged. Please see Robert (2004) for detail on Markov chain Monte Carlo (MCMC) estimation.

The free and open-source probabilistic programming language Stan is used to estimate parameters. Stan has a excellent interface with the R, allowing the convenience of building specific models directly from command lines in R. Stan uses a variant of the Hamiltonian Monte Carlo called the no-U-turn sampler (Hoffman and Gelman; 2014) as its fitting procedure. Performing multiple steps per iteration allows it to move more efficiently through the parameter space and to sample from the posterior distribution. For more details on Stan please see `http://mc-stan.org/` and Hoffman and Gelman (2014) for information on sampler details.

Estimates from Static Player Level Model

To asses the accuracy of the estimates we can compare the estimated top teams with observed final standings in the 2015-16 regular season. Observed records and results in the 2015/16 regular season are shown in Table 2.3.

Figure 4.1 displays the average home and away strengths of the 16 teams that qualified for the playoffs. The strengths of the Warriors, Spurs and Thunder are clearly

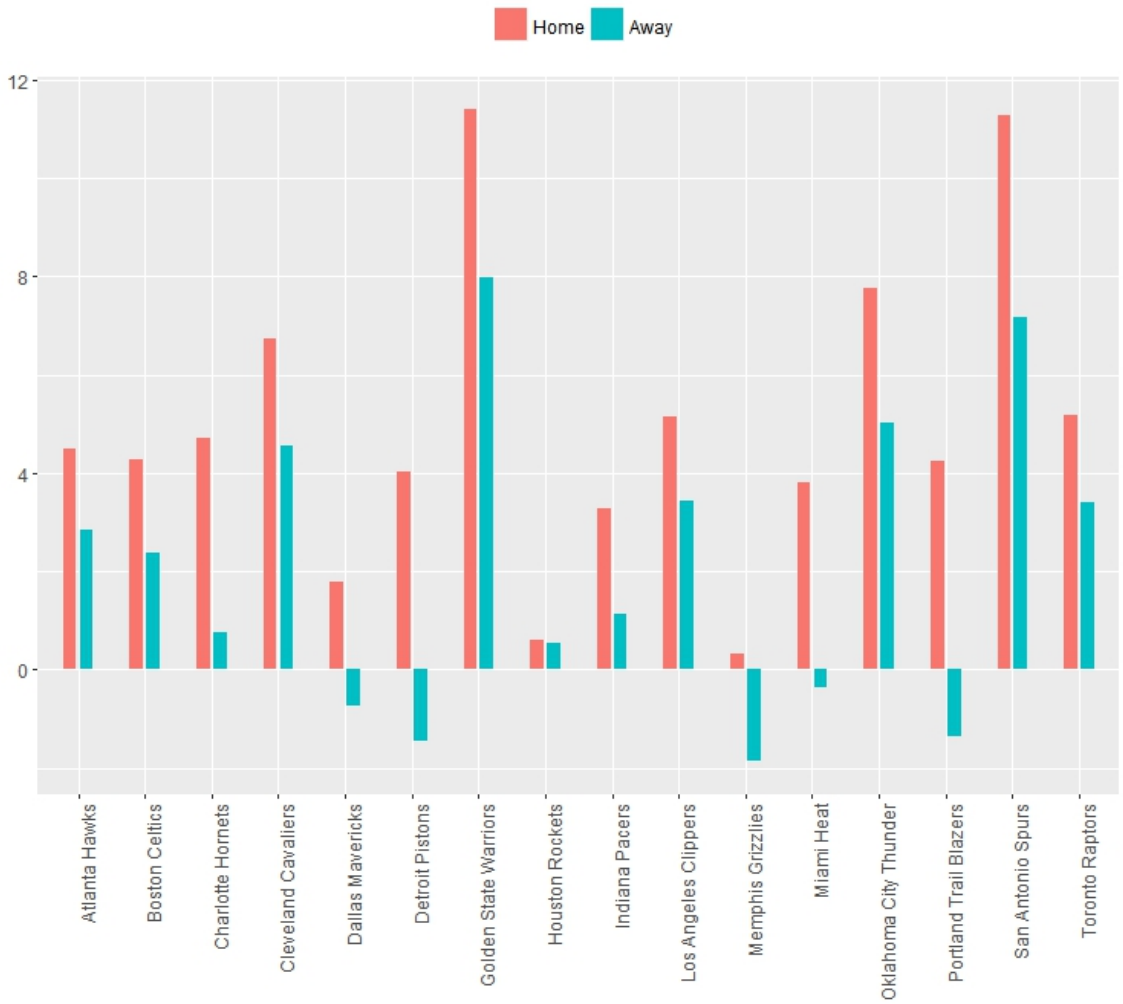


Figure 4.1: Mean Home and Away Strengths of Teams Qualified for Playoffs

the league's best. Comparing the estimated strengths with the observed win/loss records of the regular season, Figure 2.3 reflects a reasonable level of accuracy within our estimations. For example, the top average home strength estimates were the Warriors (11.44) and Spurs (11.40). During the 2015/16 regular season, the Warriors were almost unbeatable on their home court with only two losses and averaging 117 points per game (the league's highest) and the Spurs were also one of the best teams at home with only one loss.

Using a model at the player level provides another possible approach of model validation. Using Figure 4.2 we can check whether the estimated player strengths in our model matches the current top players in the league. As the home and away estimates are a direct reflection of players' strengths, if there are notably wrong estimates in player strengths it presents an indication of imprecise estimates of team strengths.

Compared to the top 10 players in the NBA (Golliver and Mahoney; 2015) the

player estimates are accurate with six players; Kawhi Leonard, Lebron James, Kevin Durant, Stephen Curry and Draymond Green. There is a common theme behind the remaining four players; they are all reserve players from strong teams. For example Andre Iguodala from the Warriors and Cameron Payne from the Thunder are reserve players from two of the top three teams in the league. It is common for these teams to beat opponents by a large margin with considerable time remaining in a game. When games are out of reach and the reserve players manage to hold onto the lead, the model overestimates their player abilities. Further research could conducted on ways to resolve this issue.

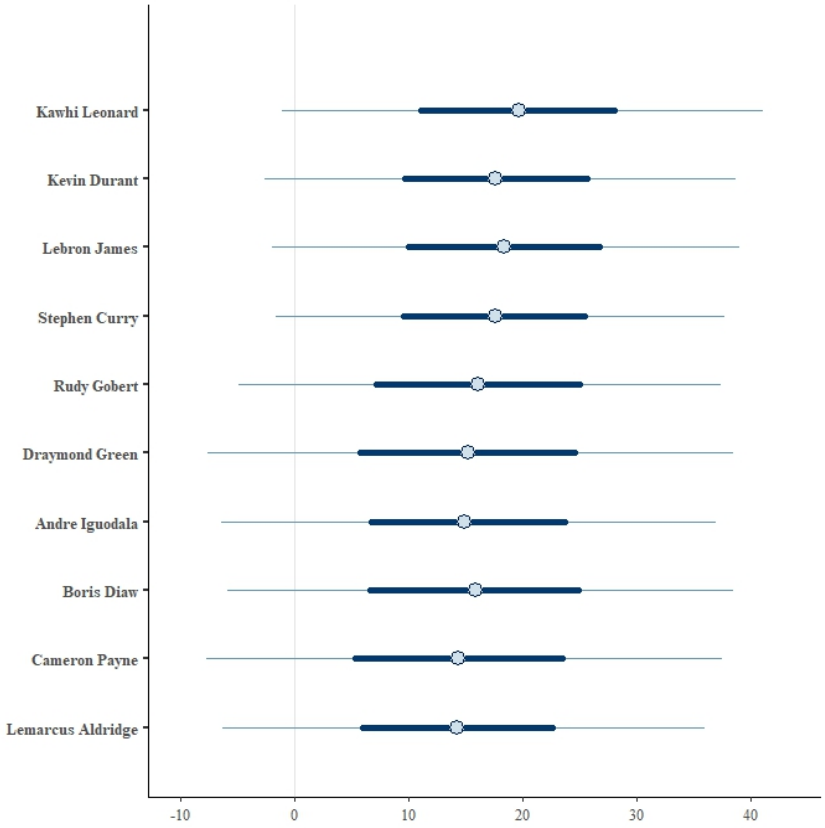


Figure 4.2: Estimated player strengths top 10 players in the 2015/16 regular season.

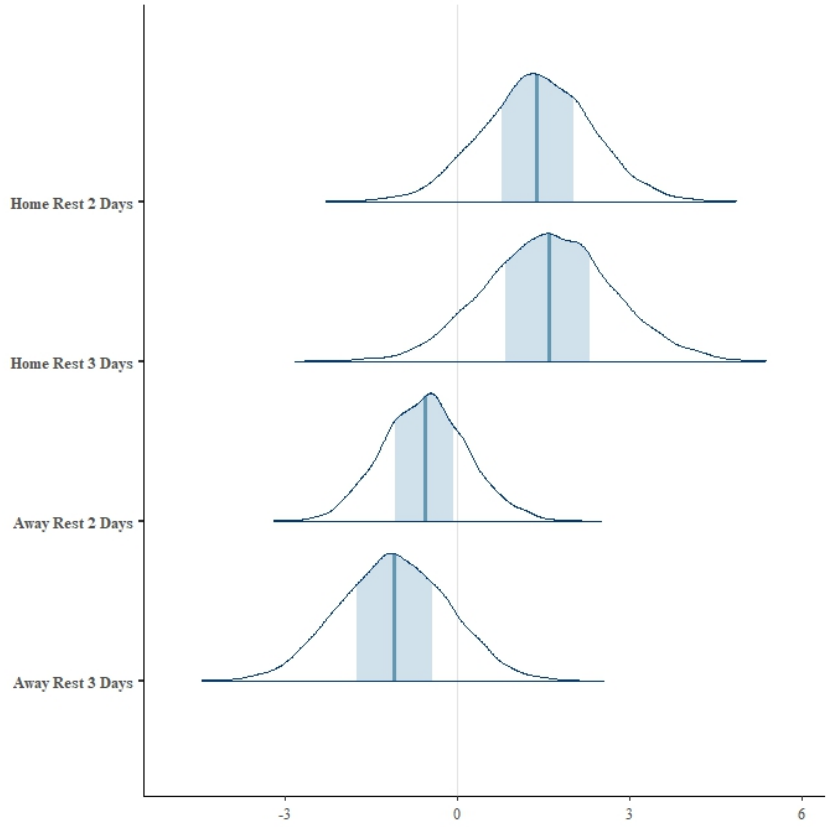


Figure 4.3: Posterior distributions of beta coefficients with 99% intervals

Figure 4.3 displays the posterior distributions of the rest covariates used in the model. From a classical statistical point of view the coefficients do not reflect statistical significance for predicting score differences; we do not concern ourselves with this as the intent is not to find a parsimonious model. Comparing the strength estimates with and without the rest covariates reflected the rest covariates help in predicting game outcomes. The posterior distributions indicate that the length of rest affects home teams strength more than away teams. Entine and Small (2008) study on the role of rest in NBA team supports the above results.

MCMC Diagnostics

To asses convergence in Markov Chains there are a number of diagnostics available to use. A trace plot shows the history of a parameter value across iterations of the chain. If a chain has converged, it should not be showing any long-term trends as the average value for the chain should approximately be flat. Evidence of an apparent trend could correspond to MCMC chains failing to converge.

The first diagnostic plot in Figure 4.4 displays a trace plot for the replicated score difference between Hawks and Celtics in Round 1. The trace plot does not show any obvious trend in long-term average providing support to assume convergence

in the first replicated score differences of round 1. The symmetrical nature of the estimated density plot (second plot in 4.4) gives further evidence of a converged chain.

Another important measurement to check for efficient sampling is autocorrelation. A number between negative 1 and positive 1 measures the linear dependency between the current value of the chain to past values called lags. Autocorrelation is an important metric because it's an indication of how much information is available in the Markov Chains. The autocorrelation plot in Figure 4.4 provides evidence for efficient sampling.

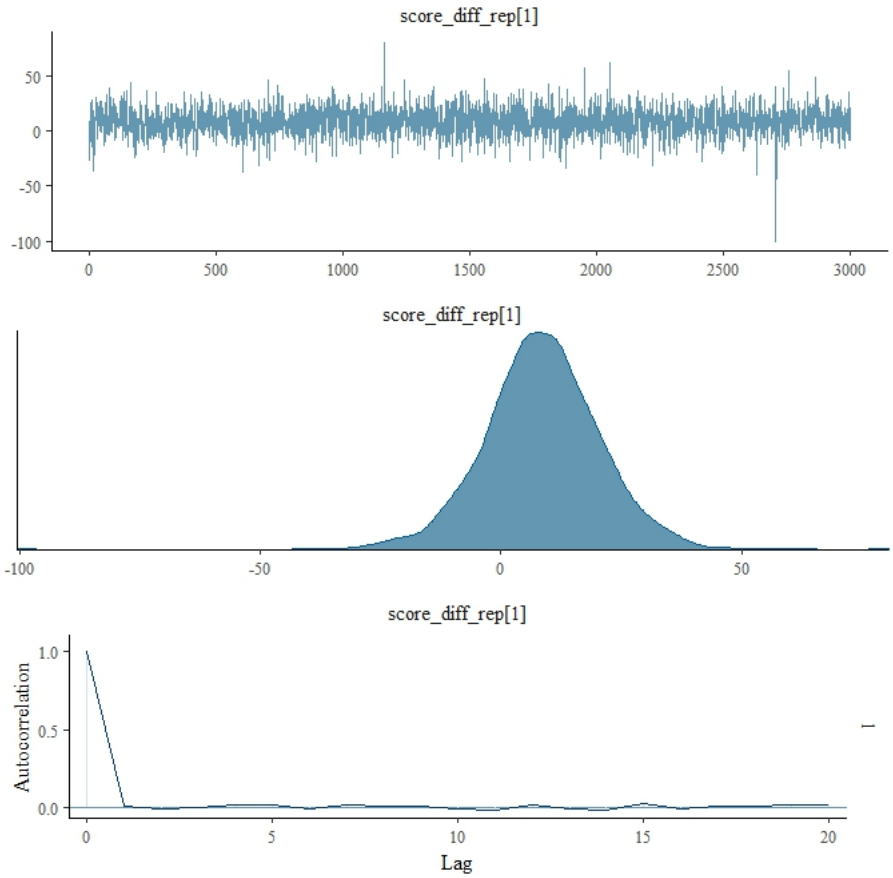


Figure 4.4: A trace plot, density estimate and autocorrelation of the samples from the posterior distribution of replicated score difference for Round 1 of Hawks vs Celtics.

4.2 Predictive Performance

Simulating the regular season provided information about team strengths and weaknesses. The mean home and away strengths for each team and their variances from the regular season is used as prior distributions for in the firs round of the first playoff series. The posterior distribution of $\theta_{g,1}$, $\theta_{g,2}$, σ_1 and σ_2 are then used as prior distribution for the next round.

Predictions are computed on a per round basis. Meaning after each round we fit the observed score difference as observed information, allowing the model to adapt to any injuries or other effects in a team that may outcome games. If a star player is a team is injured the outcome of the series is likely to change. The team is with the injured player may lose in a large margin. If this occurs, the team strength posterior distribution is expected to shift in favour of the opposition thus decreasing the likelihood of the injured player's team from winning.

Predicting the outcome of a round g in a series is an iterative process and is done as follows:

1. Update estimates for $\theta_{g,1}$, $\theta_{g,2}$, σ_1, σ_2 using posterior distributions from round $g-1$.

2. Select from the database for the complete playoff series only those games prior to round g, and using these games as observed data in Stan for predicting the outcomes in round g.

3. Check which teams are at home and away and start the sampling process.

4. If $g \geq 4$ check if a team has been eliminated and updated match ups accordingly.

Prior to updating estimates in Step 1, the convergence of the Markov Chains in the posterior distributions of round $g-1$ should be investigated. Once there is sufficient evidence for the convergence of the Markov Chains, the sample mean and sample standard deviation of home and away strengths can be computed. Table 4.2 displays the match-up, observed winners and predicted outcomes of the first series. The first series of the playoffs consists of 8 teams in total, 4 teams from each conference competing in a best of seven game series (Figure A.2).As mentioned previously, the primary focus of this dissertation is predicting outcomes of playoff rounds but for model validation predicted score differences are also computed, displayed in the predicted score difference column of Table 4.2.

A weakness needs to addressed with the model is evident in the predicted mean score of round 2 between Raptors and Pacers. The continuous distribution predicts an overall score difference of zero suggesting a tie game. In context of an NBA game, this is impossible as overtime periods are played until a winner is determined.

Table 4.1: Prediction accuracy per Series

	First Series	Conf. Semi-Finals	Conf. Finals	NBA Finals
Number of rounds	36	17	11	6
Predicted Correctly	25	9	7	3
Percentage	69%	53%	64%	50%

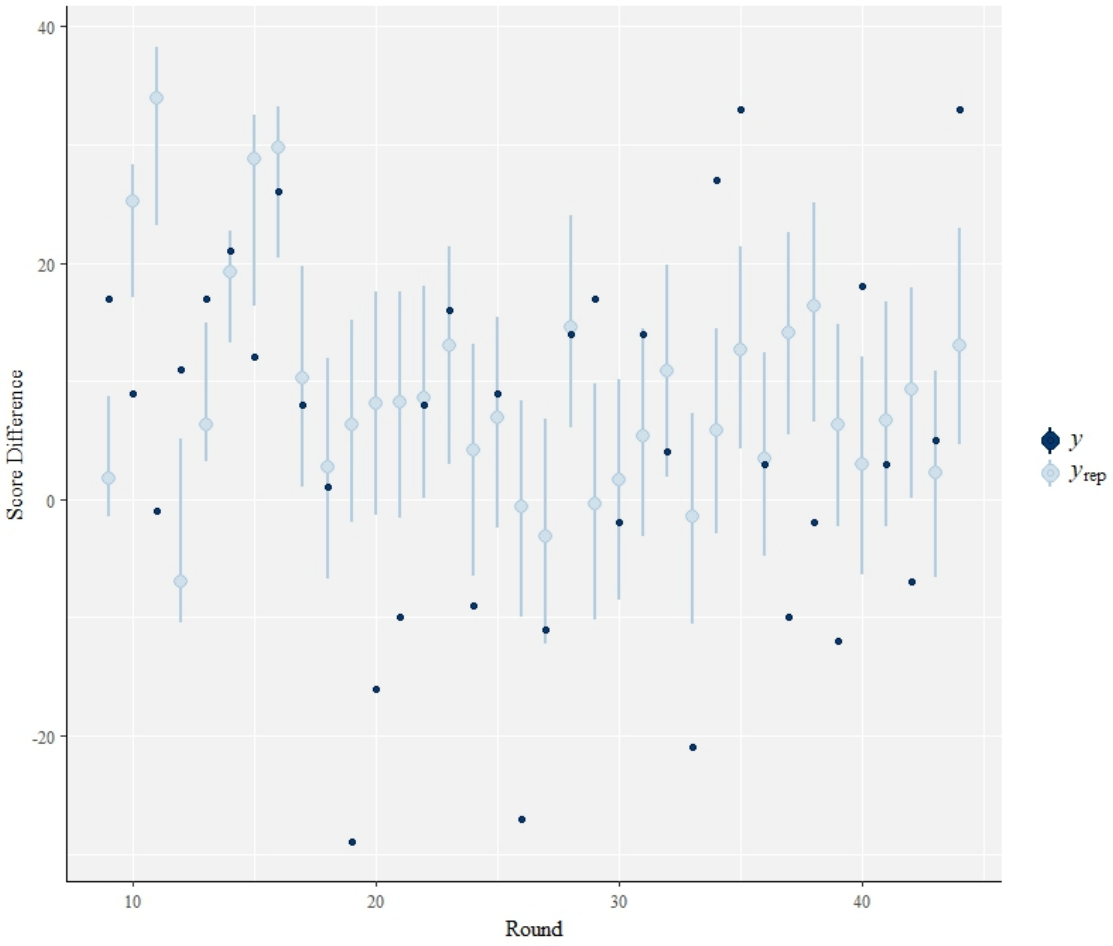


Figure 4.5: 50% Posterior Prediction intervals for all rounds in Series 1

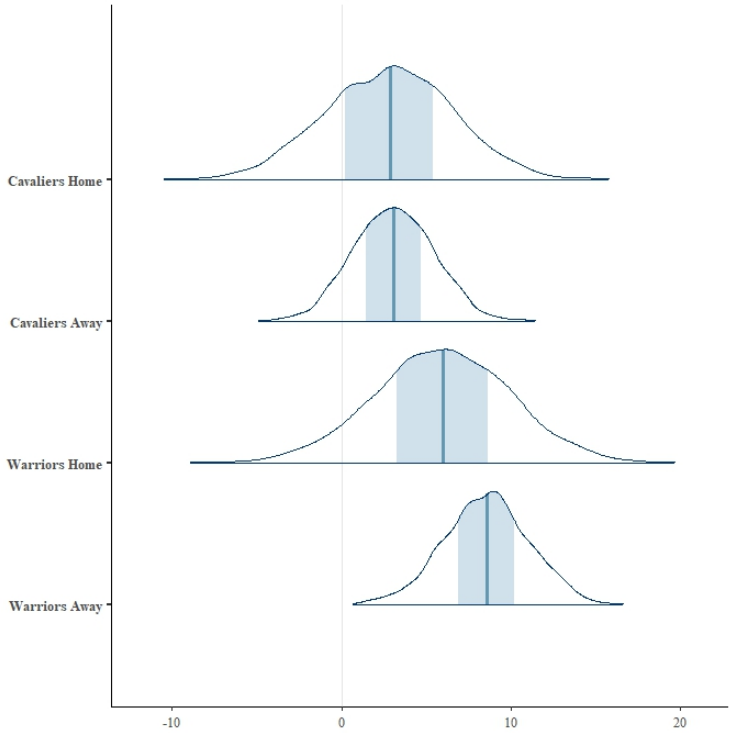
Table 4.2: Predicted results of the first series.

,

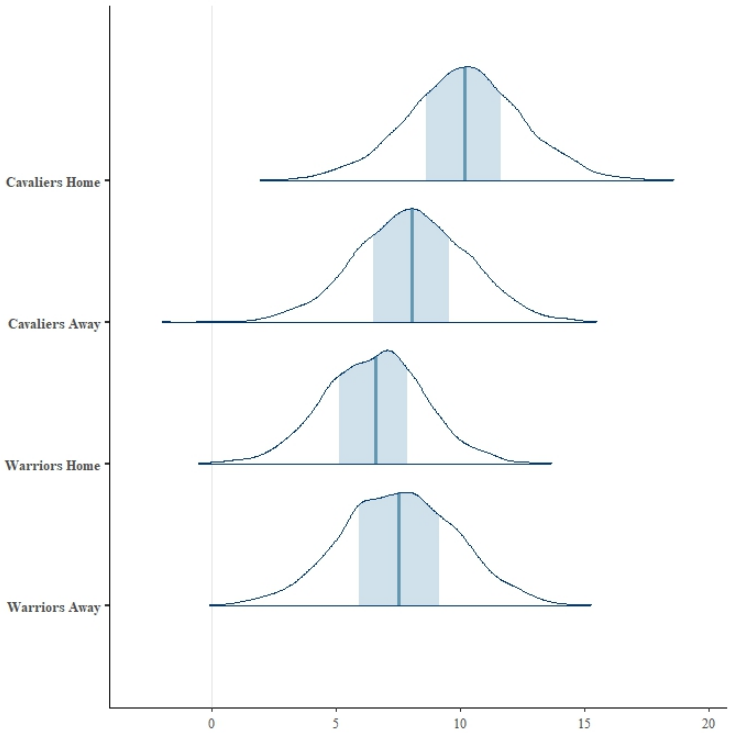
Round	Home	Away	Obs. Winner	Obs.d Score Diff	Pred. Winner	Pred. Score Diff	Pred. Correct?
2	Hawks	Celtics	Hawks	17	Hawks	5	TRUE
2	Warriors	Rockets	Warriors	9	Warriors	21	TRUE
2	Thunder	Mavericks	Mavericks	-1	Thunder	28	FALSE
2	Raptors	Pacers	Raptors	11		0	FALSE
2	Cavaliers	Pistons	Cavaliers	17	Cavaliers	10	TRUE
2	Clippers	Trail Blazers	Clippers	21	Clippers	17	TRUE
2	Heat	Hornets	Heat	12	Heat	22	TRUE
2	Spurs	Grizzlies	Spurs	26	Spurs	25	TRUE
3	Celtics	Hawks	Celtics	8	Celtics	11	TRUE
3	Rockets	Warriors	Rockets	1	Rockets	3	TRUE
3	Mavericks	Thunder	Thunder	-29	Mavericks	7	FALSE
3	Pacers	Raptors	Raptors	-16	Pacers	8	FALSE
3	Pistons	Cavaliers	Cavaliers	-10	Pistons	8	FALSE
3	Trail Blazers	Clippers	Trail Blazers	8	Trail Blazers	9	TRUE
3	Hornets	Heat	Hornets	16	Hornets	13	TRUE
3	Grizzlies	Spurs	Spurs	-9	Grizzlies	4	FALSE
4	Celtics	Hawks	Celtics	9	Celtics	6	TRUE
4	Rockets	Warriors	Warriors	-27	Warriors	-1	TRUE
4	Mavericks	Thunder	Thunder	-11	Thunder	-3	TRUE
4	Thunder	Mavericks	Thunder	14	Thunder	15	TRUE
4	Pacers	Raptors	Pacers	17	Raptors	-1	FALSE
4	Pistons	Cavaliers	Cavaliers	-2	Pistons	1	FALSE
4	Trail Blazers	Clippers	Trail Blazers	14	Blazers	5	TRUE
4	Hornets	Heat	Hornets	4	Hornets	11	TRUE
5	Grizzlies	Spurs	Spurs	-21	Spurs	-3	TRUE
5	Hawks	Celtics	Hawks	27	Hawks	6	TRUE
5	Warriors	Rockets	Warriors	33	Warriors	14	TRUE
5	Raptors	Pacers	Raptors	3	Raptors	3	TRUE
5	Clippers	Trail Blazers	Trail Blazers	-10	Clippers	14	FALSE
5	Heat	Hornets	Hornets	-2	Heat	15	FALSE
6	Celtics	Hawks	Hawks	-12	Celtics	6	FALSE
6	Pacers	Raptors	Pacers	18	Pacers	3	TRUE
6	Trail Blazers	Clippers	Trail Blazers	3	Trail Blazers	7	TRUE
6	Hornets	Heat	Heat	-7	Hornets	9	FALSE
7	Toronto	Indiana	Toronto	5	Toronto	16	TRUE
7	Miami	Charlotte	Miami	33	Miami	7	TRUE

The advantage of predicting per round is the ability for us to cross check forecasts with historical data. Table 4.1 depicts a model with a 60% accuracy across the 2015-16 playoffs. Performing its best in the first series and weakest in the NBA Finals. It should be noted, the 2015-16 NBA Finals provided one of the most unexpected and greatest NBA finals comeback played between the Cavaliers and Warriors. The Cavaliers were trailing 1-3 in the series with the last game at Warrior's home court. Given the dominance Warriors displayed on their home court the predictions showed the Warriors would finish the series and win the championship. However, the Cavaliers managed to miraculously rally back and win the series in seven rounds.

Due to the round-by-round prediction nature of the model, the prior distributions of home and away strengths is expected to shift towards the posteriors as each series progresses. Meaning, when the Cavaliers were climbing back in the series the posterior distributions of Cavaliers' strengths should have updated sharply and predicted Game 7 accurately. Figure 4.6a displays the prior distributions used for the Cavaliers and Warriors before the start of the playoffs. The two teams are equally strong but the Warriors seem to be the stronger team by a slight margin which is expected as they had the best regular season record in NBA history. The prior distributions before round 7 of the finals (Figure 4.7) still depicts the Warriors to be the stronger team even though the series was even with Cavaliers winning the last two rounds. The strength parameters were not updated fast enough; most likely due to the weak estimates of home and away correlation.



(a) Prior distributions of strengths for Cavaliers and Warriors before playoffs commence.



(b) Posterior distributions of strengths for Cavaliers and Warriors at the end of the Conference Finals.

Figure 4.6: Prior and posterior distributions of home and away strengths of the Cavaliers and Warriors.

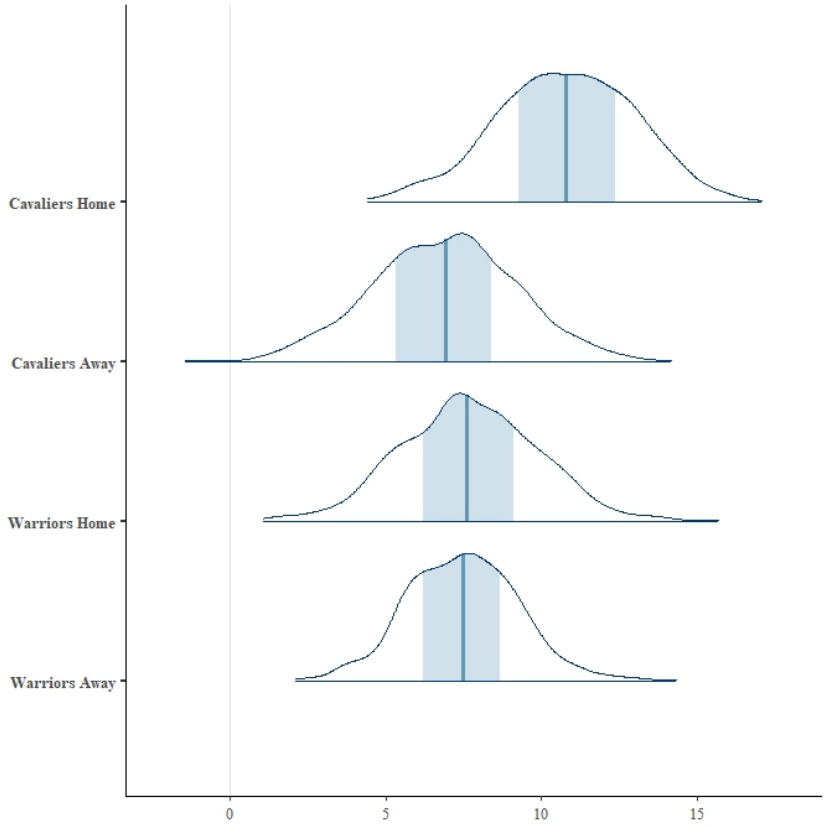
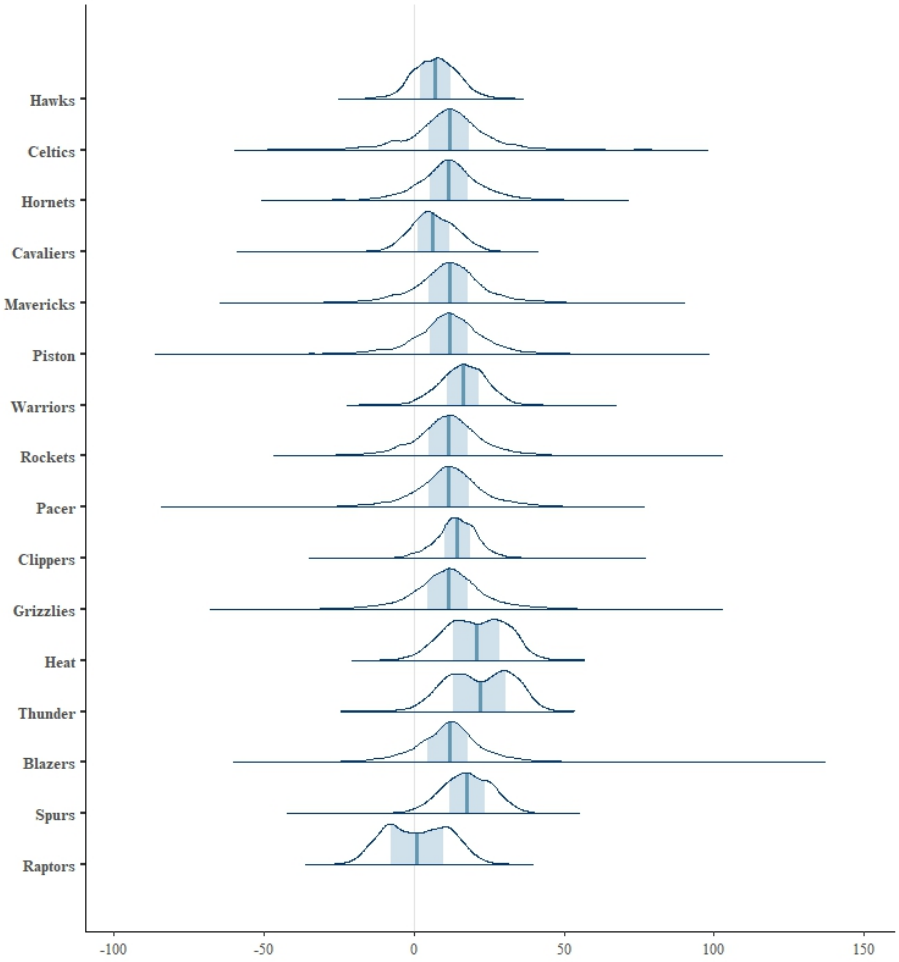


Figure 4.7: Prior distributions for game 7 of the NBA finals between Cavaliers and Warriors.

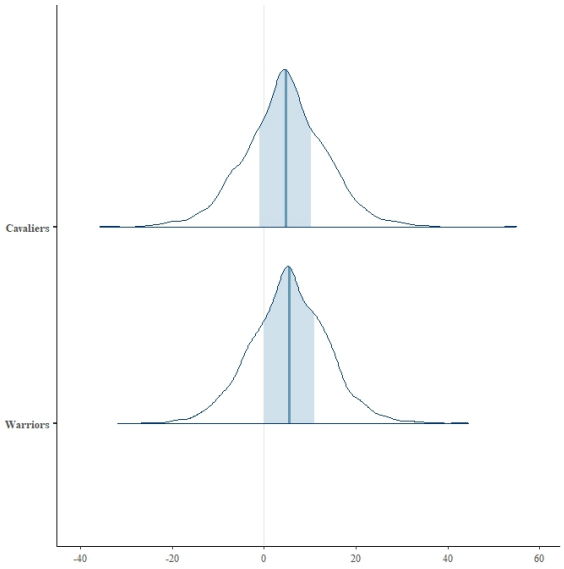
Unique Home Advantage Estimates

The home effect in our model was fitted as a unique parameter for each team. Figure 4.8a displays the prior distributions of the home effect for each team in the playoffs. These distributions are attained from the posterior distributions of the regular season. Each team's home effect has around 5 point affect to their overall score but it is hard to find any clear discrepancies between the teams. Warriors, Thunder and Heat have has their posteriors shifted slightly to the left.

Figure 4.8b displays the posterior distributions of the home effect of the Cavaliers and Warriors after the last game of the Finals.It is evident the posterior has not shifted much from the prior in 4.8a. This is expected as the prior is unlikely to shift within a small volume of games in the playoffs.



(a) Prior distributions of home effects at the start of the playoffs



(b) Posterior distributions of home effects of the Cavaliers and Warriors after NBA Finals

Figure 4.8: Prior and posterior distributions of home effects

Concluding Thoughts and Extensions

5.1 Possible Extensions

The aim of this thesis was to use build a Bayesian Hierarchical model and forecast the outcomes of the games during the NBA playoffs. Prior to any forecasting. a web-scraper was developed to attain data. With the intent of learning about teams' strengths and their variability a dynamic and static model were fitted to the regular season. The static model also included the length of rest between games as a explanatory variable and home court advantage as a unique parameter for each team. In this thesis, the player level model was chosen over the dynamic because the covariates used provides estimates for teams that are capable of playing in any situation. A possible extension could be to include explanatory variables in the dynamic model and observe weather it provides better estimates.

Using more Data

Throughout the thesis, a number of issues have been mentioned. Using a model at the player level made it infeasible to use data from multiple seasons at once. Since players move around due to trades, free agent signings and drafts we cannot fit a multiple seasons of data for a single playoff season. If we wanted to fit multiple seasons of data, we would need to do reintroduce players and fit the model separately. Unfortunately due to time constraints the models were only tested for the 2015/16 regular season and playoffs. With additional time the regular season model could be fitted for more seasons of data and we can expect it to provide improved estimates of team strengths; using the current estimates as priors for upcoming seasons.

Score Difference of Zero

NBA games can never end in a draw, yet our continuous distribution can predict a score difference of zero. Although in the 2015/16 playoffs there was not a game where a the mean score difference between two teams was predicted to be zero, it is a possibility. To avert this problem, the distribution could be adjusted in a manner where a constraint on the outcome of zero needs to be introduced. A possible solution could be to model the direction of the score difference seperately as a binary outcome and model the score difference separately between 1 to ∞. This particular problem can be looked further into as an extension of this thesis.

Improving Estimates of Player Strengths

The static player level model used in the regular showed some reserve players in winning teams to have a larger than deserved estimated player strengths. This may have occurred due to reserve players playing in situations where games were already out of hand and both teams have elected to play their reserves. Thus the model

estimated their role to be important, even though they played against the opponents' weaker players as the game was already out of hand. This becomes a problem when these players play again and their teams' strength are overestimated by the model. To improve the estimation of player strengths, further research could be conducted on methods to incorporate avoid this problem. Incorporating individual player statistics (points, rebounds and assists) as explanatory variables to assist with player strength estimation may avoid overestimation of player strengths.

5.2 Concluding Thoughts

With the increase of volume and variety of data in basketball, we can only expect the motivation to predict outcomes using statistical models to grow. The static Bayesian hierarchical model built in this thesis is only a small step towards a viable method for forecasting playoff games. Predicting playoff outcomes is a much tougher challenge than regular season. The situation teams face in the playoffs compared to the regular season are far more consequential. The emotional and psychological aspect of playoffs can arise a larger variability in outcomes and strengths; as we saw the Cavaliers climb back from a 3-1 deficit. The model built in this thesis forecast to 2015/16 playoff games, was accurate on 60% of the games.

An achievement in this thesis that cannot be overlooked is the development a a efficient web-scraper. The web-scraper described in Chapter 2 consumed much of time to develop but it can be considered as important as the fitted model. It has the capability of attaining information for free in a timely fashion. With minor adjustments, the scraper can be adapted to other sports and purposes.

To clearly determine a winner between two equally good teams (e.g. Cavaliers vs Warriors), priors of the strengths must shift sharply after observing results. In the model fitted, the strength priors were not able to shift quick enough after each round and therefore kept predicting the similar outcomes. The major disadvantage of this lies when star players are injured and the series shifts in the other teams way. During the first series, Los Angeles Clippers all star guard Chris Paul was sidelined by a knee injury and this allowed the Portland Trail Blazers to climb back. However the model's posterior distributions did not incorporate this change and therefore still estimated the Los Angeles Clippers as the favourite.

As part of the research, this thesis also looked into fitting the home effect parameter as a unique but static for each team. The posterior distributions of the home effects for the playoffs team show that each team has approximately the same effect. However, when compared against teams that did not qualify for the playoffs (Appendix A.1) it's evident the home advantage is unique for each venue.

Bibliography

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D. and Cochran, J. J. (2014). Statistics for Business & Economics, revised, Cengage Learning.

Andrew, M. . (2015). Dynamic Bayesian forecasting of AFL match results using the Skellam distribution, Master's thesis.

Ashman, T., Bowman, R. A. and Lambrinos, J. (2010). The role of fatigue in NBA wagering markets: The surprising "Home Disadvantage Situation", Journal of Sports Economics **11**(6): 602–613.

Baio, G. and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results, Journal of Applied Statistics **37**(2): 253–264.

Bao, R. (2009). Time relaxed round robin tournament and the NBA scheduling problem, PhD thesis, Cleveland State University.

Bean, J. C. and Birge, J. R. (1980). Reducing travelling costs and player fatigue in the National Basketball Association, Interfaces **10**(3): 98–102.

Cheng, G., Zhang, Z., Kyebambe, M. N. and Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle, Entropy **18**(12): 450.

Entine, O. A. and Small, D. S. (2008). The role of rest in the NBA home-court advantage, Journal of Quantitative Analysis in Sports **4**(2).

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), Bayesian analysis **1**(3): 515–534.

Golliver, B. and Mahoney, R. (2015). Si.com's top 100 nba players of 2016.
URL: *https://www.si.com/nba/top-100-nba-players-2016?page=5devicetype=default*

Harville, D. A. and Smith, M. H. (1994). The home-court advantage: How large is it, and does it vary from team to team?, The American Statistician **48**(1): 22–28.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., Journal of Machine Learning Research **15**(1): 1593–1623.

Kharratzadeh, M. (2017). Hierarchical Bayesian modeling of the English Premier League.
URL: *http://andrewgelman.com/2017/05/17/using-stan-week-week-updating-estimated-soccer-team-abilites/*

Paine, N. (2017). A home playoff game is a big advantage — unless you play hockey, Online Article.

URL: *https://fivethirtyeight.com/features/a-home-playoff-game-is-a-big-advantage-unless-you-play-hockey/*

Polson, N. G., Scott, J. G. et al. (2012). On the half-cauchy prior for a global scale parameter, Bayesian Analysis **7**(4): 887–902.

Purdum, D. (2017). Sports betting legalization: Where do we stand right now.
URL: *http://www.espn.com.au/chalk/story/_/id/*20704273*/gambling − where − does − sports − betting − legalization − us − stand − right − now*

Robert, C. P. (2004). Monte Carlo Methods, Wiley Online Library.

Swartz, T. B. and Arce, A. (2014). New insights involving the home team advantage, International Journal of Sports Science & Coaching **9**(4): 681–692.

Teramoto, M. and Cross, C. L. (2010). Relative importance of performance factors in winning NBA games in regular season versus playoffs, Journal of Quantitative Analysis in Sports **6**(3).

Wickham, H. (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2.
URL: *http://CRAN.R-project.org/package=rvest*

Appendices

Appendix

Table A.1: Strength estimates from static player level model.

,

	Team	ID	Mean Home	Mean Away	Std. Home	Std. Away	Home effect
1	Golden State Warriors	10	11.44	8.10	2.58	2.82	6.69
2	San Antonio Spurs	27	11.40	7.17	2.65	2.86	6.52
3	Oklahoma City Thunder	21	7.81	5.04	2.80	2.66	4.53
4	Cleveland Cavaliers	6	6.74	4.60	3.03	2.84	3.34
5	Toronto Raptors	28	5.29	3.41	2.69	2.86	3.02
6	Los Angeles Clippers	13	5.13	3.45	2.95	2.87	3.19
7	Charlotte Hornets	4	4.83	0.78	2.89	3.09	3.18
8	Utah Jazz	29	4.66	-0.24	3.14	3.64	2.77
9	Atlanta Hawks	1	4.53	2.95	2.56	2.56	3.36
10	Boston Celtics	2	4.25	2.50	2.53	2.58	1.98
11	Portland Trail Blazers	25	4.19	-1.33	2.47	2.56	2.98
12	Detroit Pistons	9	4.05	-1.43	2.77	2.69	2.20
13	Miami Heat	16	3.79	-0.28	3.23	3.23	2.73
14	Indiana Pacers	12	3.37	1.16	2.77	2.71	2.12
15	Dallas Mavericks	7	1.84	-0.76	2.72	3.19	1.11
16	Washington Wizards	30	1.67	-1.19	3.25	2.94	1.31
17	Orlando Magic	22	1.59	-3.29	3.07	3.19	1.18
18	Chicago Bulls	5	1.59	-1.75	3.07	3.33	0.50
19	Houston Rockets	11	0.64	0.66	3.23	2.84	0.96
20	Memphis Grizzlies	15	0.38	-1.76	3.09	4.24	0.13
21	New Orleans Pelicans	19	-0.59	-3.67	3.79	3.46	-1.25
22	Milwaukee Bucks	17	-0.82	-3.98	3.21	2.89	-1.69
23	Sacramento Kings	26	-1.29	-1.36	2.92	3.41	-1.15
24	New York Knicks	20	-1.55	-2.13	2.76	2.85	-1.29
25	Denver Nuggets	8	-2.54	-1.11	2.92	3.13	-1.62
26	Phoenix Suns	24	-2.77	-7.24	3.76	3.70	-1.93
27	Minnesota Timberwolves	18	-3.84	-1.13	2.57	2.65	-2.46
28	Brooklyn Nets	3	-4.67	-6.11	3.03	3.73	-3.42
29	Los Angeles Lakers	14	-5.43	-8.80	2.95	3.10	-4.02
30	Philadelphia 76ers	23	-8.42	-7.54	3.07	3.09	-5.25

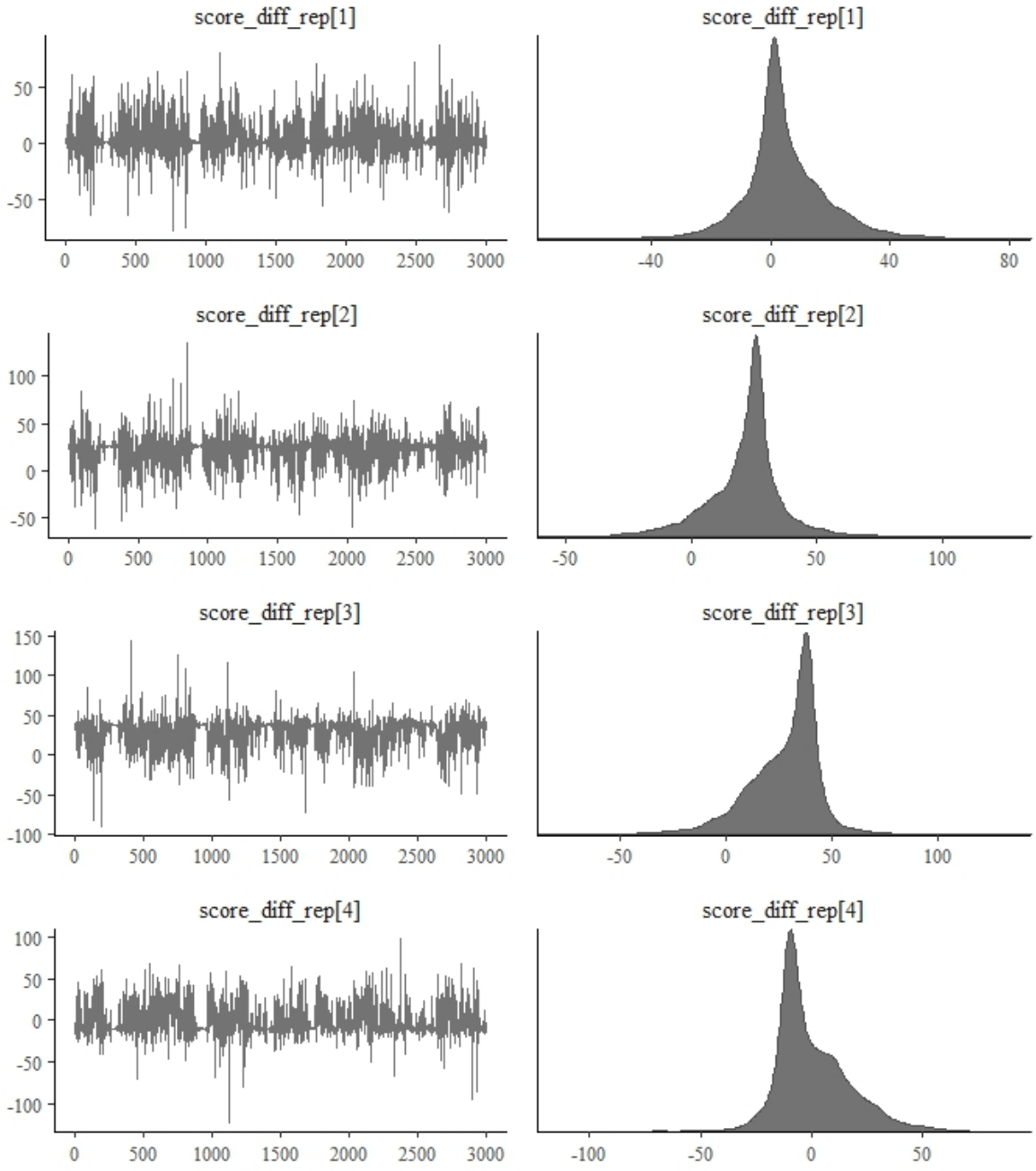


Figure A.1: Trace plots of games 1-4 in round 1 series 1.

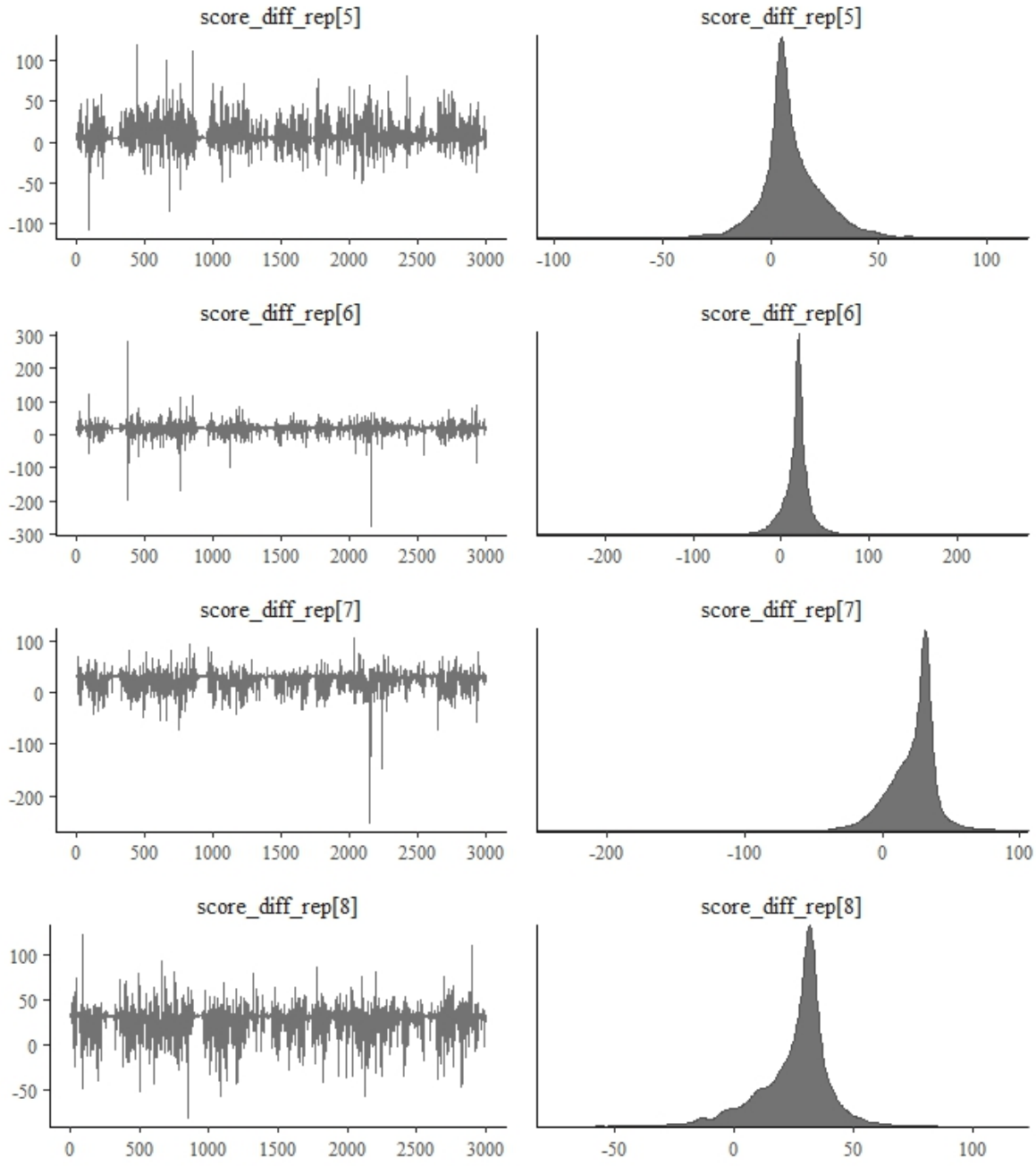


Figure A.2: Trace plots of games 5-8 in round 1 series 1.