Tennis Players: who is the greatest player of all time?

Geoffrey Liu

Research School of Finance, Actuarial Studies and Statistics,
Australian National University

Abstract

Two models are used to analyse the tennis players since 1998. A Bradley Terry model is fit using a Gibbs sampler on match-level data and a Bayesian hierarchical model is used to analyse point-level data of players' service and return games on different court surfaces. According to the Bradley Terry model, the skill difference between the top three players, Roger Federer, Rafael Nadal and Novak Djokovic is not significantly different. However, all three players are significantly higher skilled than the fouth-placed player, Andy Murray. The hierarchical model gives evidence that Federer is the greatest in terms of overall skill level on all surfaces except for clay where Nadal is indisputably the strongest player.

1. Introduction

There is a large debate within the tennis community about the greatest modern-era ATP player of all time. Some analysts propose it's Roger Federer while others nominate Rafael Nadal or Novak Djokovic. The current way of determining the best tennis player is through the number of competition wins, head-to-head matchups or through a tournament based point system i.e. ATP ranking system. In this paper, I aim to bring some statistical evidence to this debate by fitting the skill level of each player to a model using a Bayesian approach. After obtaining a distribution of each players' skill level, players can be compared and the probability of each player having a higher skill compared to another can be calculated. The challenge in this project is to incorporate the data of all tennis matches. I show that it is hard to incorporate all match data in the Bradley Terry model if a frequentist approach is used since direct matches between players are sparse i.e. the probability that there is match data between two players directly is very low. To get around this problem, priors were placed on the skill levels of players. The second model presented is a Bayesian hierarchical model that adjusts for the shortcomings of the Bradley Terry model.

The dataset used throughout the paper is provided by Jeff Sackmann. He is the author of Tennis Abstract (tennisabstract.com) and provides match-level data and detailed summary statistics of players and their matches. The data can be found on his GitHub page.[1] ATP Tour match data starting from 1998 was used since this marks the start of Roger Federer's professional career. Each row of the dataset contains a match between two players and there are 49 variables in the raw data. There are 58,786 matches and 1638 unique players. Only 9 variables in the data are used: winner name, loser name, match surface, match date, match score, total number of service points (winner), service points won (winner), total number of service points (loser), service points won (loser).

2. Bradley Terry Model

The Bradley-Terry model is a model for pairwise comparisons (similar to Elo) that is used quite commonly in modelling team sports. The model estimates the probability of player i beating player j as,

$$p_{ij} = Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

where $i, j \in K$ indicate player i and player j, K is the set of players and λ_i can be interpreted as the skill level of player i.

Usually these models are quite easy to fit through optimisation of the maximum likelihood estimator (MLE), however, for a unique finite MLE there must be a way to compare ever pair of players. Graph theory is used to explain a necessary condition for the existence of a finite MLE. Let each vertex represent a player and let a directed edge from player i to j represent the occasion that player i has beaten player j. Ford (1957) provides a necessary condition on the uniqueness of the MLE. [2] [3]

Theorem 1 (Ford 1957)**.** *For a unique finite MLE to exist, the comparison graph must be fully connected (i.e. there is a directed path from i to j, $\forall i, j \in K$).*

This condition breaks down in the case of tennis data. A simple reason is that not all players attend each tournament. The high level players attend more exclusive tournaments compared to lower level players, however, both high and low level players do meet up occasionally at Grand Slams and Masters 1000 tournaments. Moreover, regardless of what time-frame analysed, there will always be newer players who don't yet have a path connecting them to all other players. For example, Roger Federer defeated Yoshihito Nishioka (ranked 177 during the match) this year (2018). For Nishioka to be connected to all the other players, Nishioka must have defeated a player who has indirectly defeated all other players. There are 655 players in the dataset who aren't fully connected (roughly a third of all players). Other sports (e.g. ice hockey, soccer, cricket, rugby etc.) have a set amount of teams that must all play each other every season so this issue does not arise. A network graph of the data is attached in the appendix.

To solve this problem, players who are not fully connected must be removed or a prior distribution imposed on the skill (λ) of each player. The first option is a poor choice since it fails to take into account the data of players who are not fully connected (e.g. Federer's win against Nishioka) when estimating the player's skill. Therefore, it is preferable to use the Bayesian approach since it allows us to use all the data. Caron (2012) provides details for the Gibbs sampler to draw from the posterior distribution of λ given the data and a prior distribution of λ. [4]The essential equations of the Gibbs sampler are outlined below.

2.1. Gibbs Sampler

2.1.1. Latent variables, z

First, introduce latent variable, z, in order to find the complete likelihood.

$$p(z|y, \lambda) = \prod_{1 \le i \le j \le K | n_{ij} > 0} \text{Gamma}_{z_{ij}}(n_{ij}, \lambda_i + \lambda_j)$$

where K is the number of players.

2.1.2. Complete likelihood

$$\begin{aligned} L_c(\lambda, z) &\propto \prod_{i=1}^{K} \left[\lambda_i^{w_i} \exp\left\{ - \left(\sum_{i<j|n_{ij}>0} z_{ij} + \sum_{i>j|n_{ij}>0} z_{ji} \right) \lambda_i \right\} \right] \\ &\propto \prod_{i=1}^{K} \text{Gamma} \left(w_i + 1, \sum_{i<j|n_{ij}>0} z_{ij} + \sum_{i>j|n_{ij}>0} z_{ji} \right) \end{aligned}$$

2.1.3. Priors

The prior distribution of λ_i and $\lambda_i \overset{i.i.d}{\sim} \text{Gamma}(a, b)$

$$p(\lambda) = \prod_{i=1}^{K} \text{Gamma}_{\lambda_i}(a, b)$$

2.1.4. Gibbs updates

The posterior distribution is $p(\lambda, z|y) \propto p(\lambda)L_c(\lambda, z)$. Notice that $p(\lambda)$ is a conjugate prior to $L_c(\lambda, z)$. To draw from the posterior distribution of λ, first draw z by (2.1.1), then draw from $p(\lambda|z, y)$ by

$$p(\lambda|z, y) \propto L_c(\lambda)p(\lambda) = \prod_{i=1}^{K} \text{Gamma}\left(a + w_i,\ b + \sum_{i<j|n_{ij}>0} z_{ij} + \sum_{i>j|n_{ij}>0} z_{ji}\right)$$

The complete details of the derivation can be found in the Caron's 2012 paper. [4]

2.2. Methodology

First, the dataset was transformed into a square matrix of wins, W. Each entry W_{ij} is the number of times that i beats j. For our data, this matrix is sparse – only 1.43% of the matrix entries are non-zero. The shape parameter, a, of the prior distribution was set to 2.5. The rate parameter b in the Gamma distribution is not likelihood identifiable, therefore the prior distribution solely relies on the specification of the shape parameter a. For a given a, $b := aK - 1$. The prior distributions graphs for different values of a is attached in the appendix. I implement the Gibbs sampling by writing our code in R with reference to the example MATLAB code and equations provided in the paper. The Gibbs sampler ran for 5000 iterations, and the first 100 iterations treated as burn-in. From the trace plots, the posterior draws did not require thinning. The trace plots are attached in the appendix. The code is also attached to the paper.

2.3. Results

The posterior distribution of λ_i for the top 6 players can be seen in Figure 1. The top 4 players in the model are not surprising since these 4 players are known as the 'Big Four' players. [5]

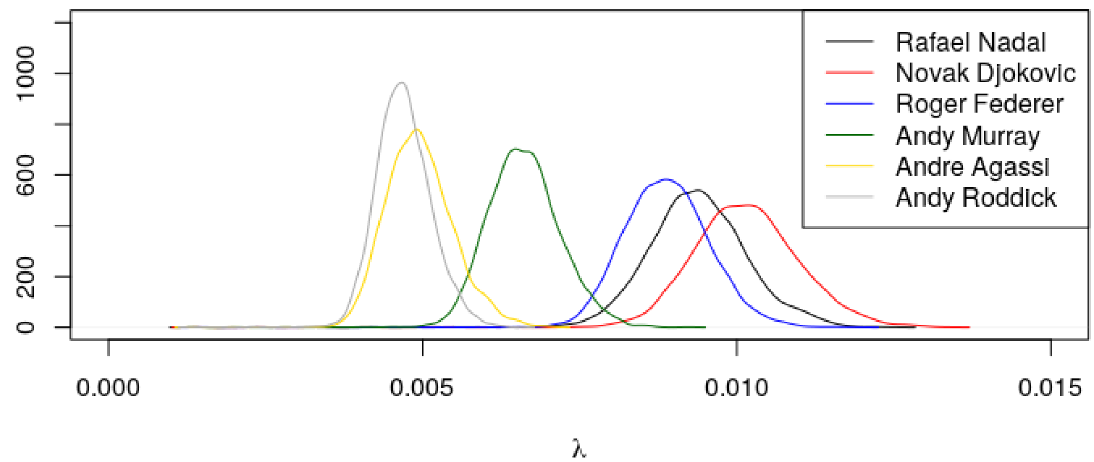

Figure 1: Posterior distribution of λ for top 6 players

Tables of results are provided for the top 4 players with the highest skill level (λ).

Table 1: Table 1: $\Pr(\lambda_i > \lambda_j)$

	Nadal	Djokovic	Federer	Murray
Nadal	–	0.24	0.69	1
Djokovic	0.76	–	0.89	1
Federer	0.31	0.11	–	1
Murray	0	0	0	–

Table 2: $\Pr(Player_{row}$ beats $Player_{col})$

	Nadal	Djokovic	Federer	Murray
Nadal	–	0.48	0.51	0.59
Djokovic	0.52	–	0.53	0.61
Federer	0.49	0.47	–	0.57
Murray	0.41	0.39	0.43	–

Table 3: No. times Player$_{row}$ beats Player$_{col}$ historically

	Nadal	Djokovic	Federer	Murray
Nadal	–	25	23	17
Djokovic	27	–	25	25
Federer	15	22	–	14
Murray	7	11	11	–

Table 4: Pr(Player$_{row}$ beats Player$_{col}$) historically

	Nadal	Djokovic	Federer	Murray
Nadal	–	0.48	0.61	0.71
Djokovic	0.52	–	0.53	0.69
Federer	0.39	0.47	–	0.56
Murray	0.29	0.31	0.44	–

Figure 1 shows the distribution of the skill levels of the top 6 players. Djokovic has the highest skill, followed by Nadal, Federer, Murray, Agassi then Roddick. The probability of $\lambda_i > \lambda_j$ for the top four players are presented in Table 1. The skill difference between Rafael Nadal, Roger Federer and Novak Djokovic is not significant at the 95% significance level. However, all three players have significantly higher skill than Andy Murray.

The mean of the posterior distribution of λ was used to calculate $Pr(i$ beats $j)$. Compared to the data, the model is fairly accurate when calculating the probability of winning for Nadal vs Djokovic, Djokovic vs Federer, and Federer vs Murray matches. The model underpredicts Nadal and Djokovic's win rate against Murray and overpredicts Federer's win rate against Nadal. These results are presented in Tables 2,3, and 4.

Figure 2 is the distributional equivalent of Table 2 where each plot is the distribution behind the corresponding cell in Table 2. The red lines mark the 95% confidence interval, and the black vertical line represents the data. In half of these plots, the data does not lie within the 95% confidence interval. This indicates that our model may be overconfident in estimating match outcomes. Our model may also be overly simplistic since it does not consider other variables apart from match outcomes. Variables such as player age, match surface, serve ability, return ability all have predictive ability in modelling tennis outcomes. I will introduce some of these predictors in our next model.

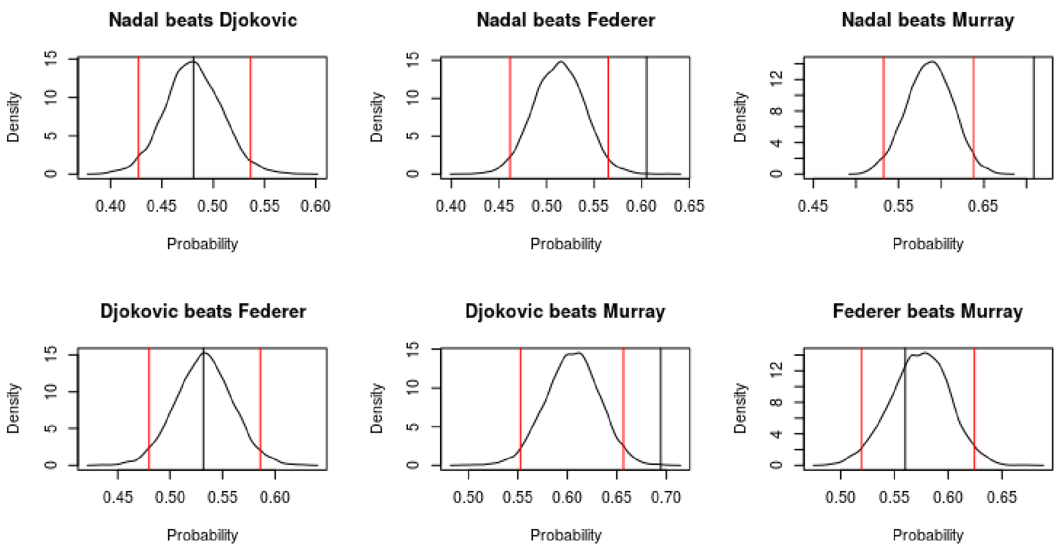

Figure 2: Posterior distributions of $Pr(i \text{ beats } j)$

3. A Bayesian Hierarchical Model

The second model presented is a Bayesian hierarchical model. To fit this model, each match was broken into two service matches. The rationale is that tennis can be broken down into two distinct games, one where the player is serving and the other where the player is receiving. In most cases, the serving player has a higher chance of winning his game compared to the receiving player. The situation where the serving player fails to win his service game is called a 'break' or 'breaking' of serve. Therefore the service game is modelled separately from the returning game. The returner's skill level must also taken account when facing his opponent's serve. A player who is good at returning serves has a higher probability of winning a point off his opponents serve.

These types of tennis models have been suggested before by Newton (2004), O'Malley (2008), Barnett (2006), Ingram (2017). [6][7][8][9]

3.1. Data

Each match in the original data is broken down into two distinct services matches; the service match of the winner and that of the loser. Breaking down the data in this way doubles the number of data points in the original data. The odd rows of data has the server (who was previously the winner), the returner (who was previously the loser), match surface, the number of points won by the server (winner) and the total points played by the server (winner). The even rows is similar except the server and returner are reversed. The corresponding row of the data has the server (who was previously the loser), the returner (who was previously the winner), match surface, the number of points won by the server (loser) and the total points played by the server (loser). Each observation contains only points won by the server however it still characterises the points won by the returner since the total number of points won by the returner equals the total number of service points played by the server minus the number of points won by the server.

(There are now $N = 117572$ rows of data and 5 columns described above.)

3.2. Model

The model used is a partial pooling Bayesian hierarchical model characterised below.

$$Y_m \sim \text{Binomial}(n_m, \theta_m), \qquad m \in N$$

Y_m is the number of service points won by the server in match m
n_m is the total number of service points played by the server in match m
θ_m is the probability of the server winning one service point in match m

θ_m is specified as

$$\theta_m = \text{logistic}(a_{m_i} - b_{m_j} + \delta_{m_i}(s) - \delta_{m_j}(s) + c)$$

where $\text{logistic}(\alpha) = \frac{\exp\alpha}{1+\exp\alpha}$ and m_i is the server i and m_j is the returner j in match m.

The parameters can be interpreted as follows
a_i is the serving skill of player $i \in I$
b_j is the returning skill of player $j \in I$
$\delta_i(s)$ is the surface-specific skill of player $i \in I$ on surface $s \in S$
c is the intercept, i.e. baseline skill level common to each player
$I = \{1, \ldots, 1638\}$ the set of unique players in the dataset.
$S = \{\text{Hard, Clay, Carpet, Grass}\}$ are the unique surfaces in the dataset.

3.2.1. Priors

The parameters a, b, and δ follow a normal distribution with hierarchical priors and c has a non-hierarchical prior.

$$\begin{aligned} a &\sim \text{Normal}(0, \sigma_a) \\ b &\sim \text{Normal}(0, \sigma_b) \\ \delta(s) &\sim \text{Normal}(0, \sigma_s) \quad \forall s \in S \\ c &\sim \text{Normal}(0, 100) \end{aligned}$$

where $a = [a_1, \ldots, a_{1638}]$ is a vector of serve skills for each player. The skill level of all players is normally distributed with mean 0 and standard deviation σ_a. σ_a determines the variability of the players serve skill around 0. The mean of the distribution is 0 since a, b and δ augment the baseline skill level c. Technically, a is interpreted as a skill multiplier or bonus based on the persons service ability.

An assumption used here is that a_i, b_i, c_i are independent. This is a reasonable assumption since the serve/return/surface ability of one player is independent to the serve/return/surface ability of other players. c is common to all players and represents the common baseline skill of all professional tennis players in the dataset.

3.2.2. Hyperpriors

The hyperpriors are weakly informative

$$\begin{aligned} \sigma_a &\sim \text{Gamma}(0.1, 0.1) \\ \sigma_b &\sim \text{Gamma}(0.1, 0.1) \\ \sigma_s &\sim \text{Gamma}(0.1, 0.1) \quad \forall s \in S \end{aligned}$$

3.3. Methodology

3.3.1. Stan

Stan is a probabilistic programming language with R and Python interfaces that allows users to perform full Bayesian statistical inference with MCMC sampling. Stan implements the No U-Turn Sampler (NUTS) and Hamiltonian Monte Carlo (HMC) to generate draws from the posterior distribution. [10] HMC is similar to the Metropolis Hastings algorithm except in the way it treats the proposal distribution. [11] The HMC also performs better in high-dimension situations. The model was implemented in Stan with help from the online Stan community and available tutorials and demonstrations. [12] [13] [9]

The model was fit using the RStan package in R with 10 chains and 1000 iterations per chain. By default, 500 of these iterations are warm-up, and the other 500 are samples. This gives 5000 posterior samples for each of the specified parameters. The model took roughly 150 minutes to fit. The 'feather' package in R was used to save the RStan results into a pickle file and read using Python. Python was used to plot the results since it is more capable at handling and visualising a large amount of data.

After fitting the model, each player has unique parameters a_i, b_i, and $\delta_s(s),\ s \in S$. There are 7 group parameters that are common across all players, c, σ_a, σ_b, σ_s for $s \in S$.

3.4. Results

3.4.1. Posterior Distributions

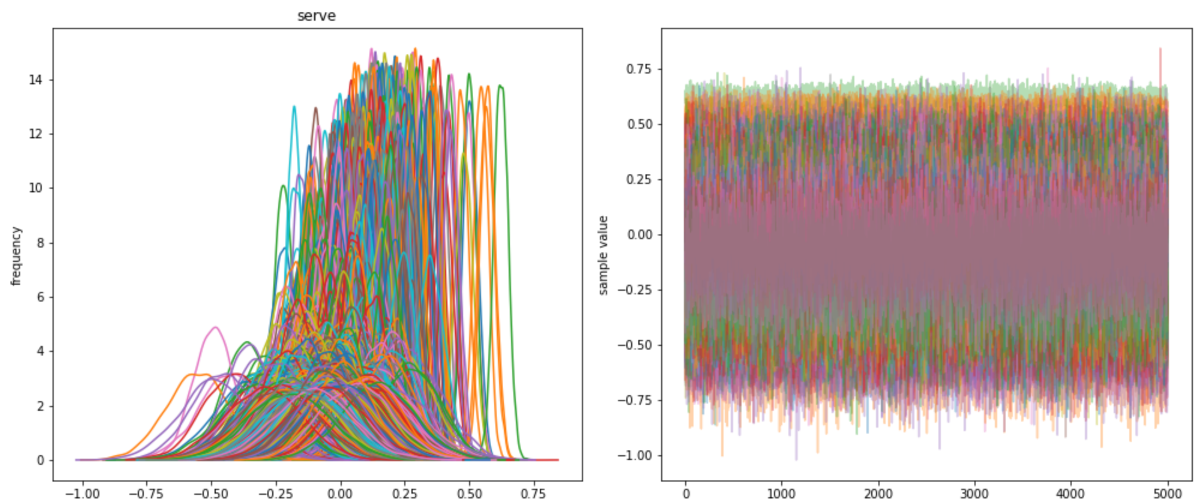

Figure 3: Posterior distribution of a

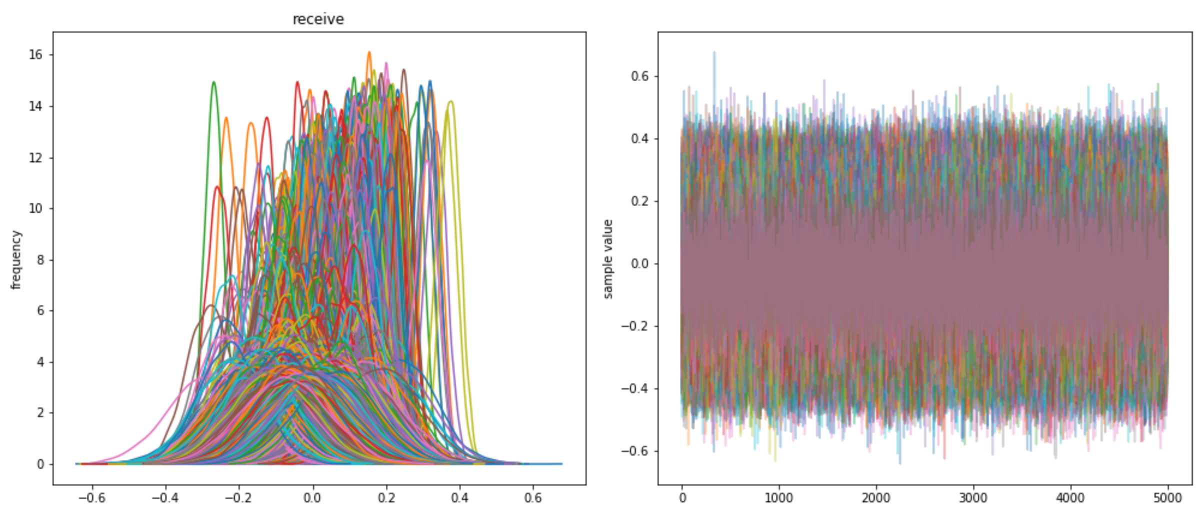

Figure 4: Posterior distribution of b

Figure 3 and 4 show the posterior distribution of a and b. Each density plot represents the distribution of posterior draws for an individual player's serve skill. There are 1638 distributions in each of the plots. The trace plots for all 1638 posteriors can be found on the right in the figures.

Although hard to see in the above graph, trace plots for a random sample of parameters was checked and the posterior samples have converged to a stationary distribution.Furthermore over 99% of the fitted parameters have an effective size of over 1000 and most have effective sizes in the 4000-5000 range. A histogram of the effective sizes is in the appendix. The ACF plots of the posterior draws converge very fast and are also in the appendix. The posterior distributions for the other parameters are similar and placed in the appendix.

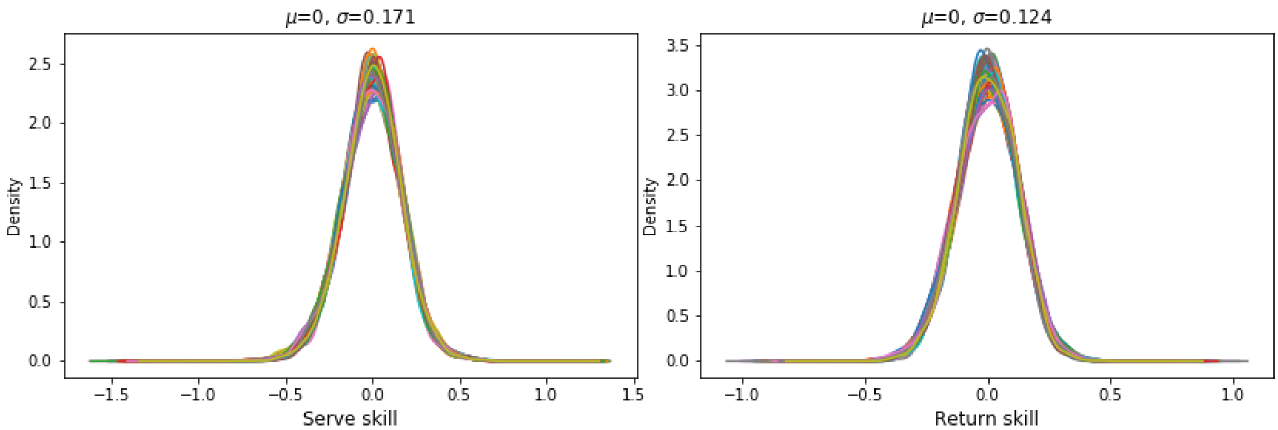

Figure 5: Posterior draws of a and b

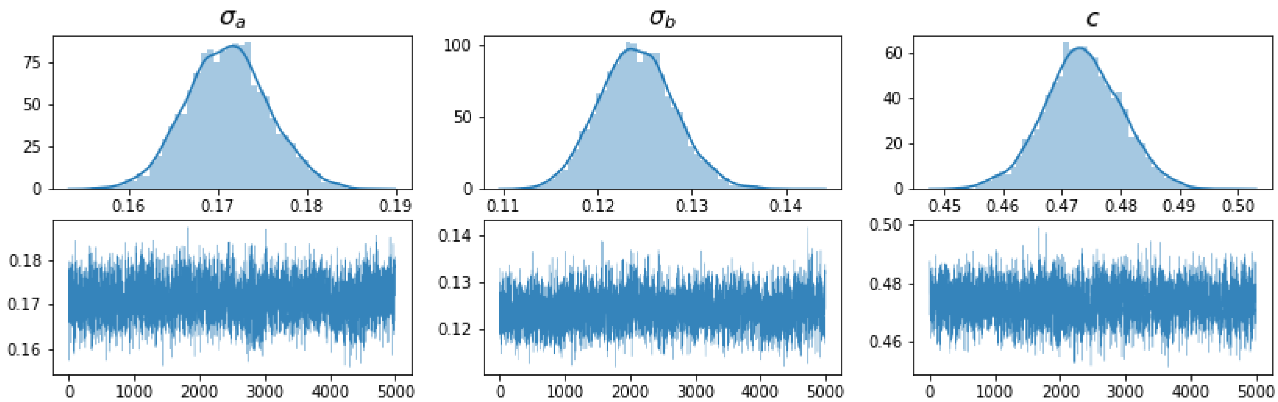

Figure 6: Posterior draws of σ_a, σ_b, c

Figure 5 shows 100 posterior distributions of a and b. These are normally distributed which matches our model specification. Moreover, the average standard deviation of a and b clearly lie within the posterior distribution of σ_s. This result is coherent with our model specification.

3.4.2. Posterior Predictive

Since the full posterior predictive distribution is too hard to check (as it requires 5000 draws from a binomial distribution for each row of data), a partial posterior predictive distribution was generated. 50% of the total data was simulated with 1000 samples from the posterior distribution of the model parameters.

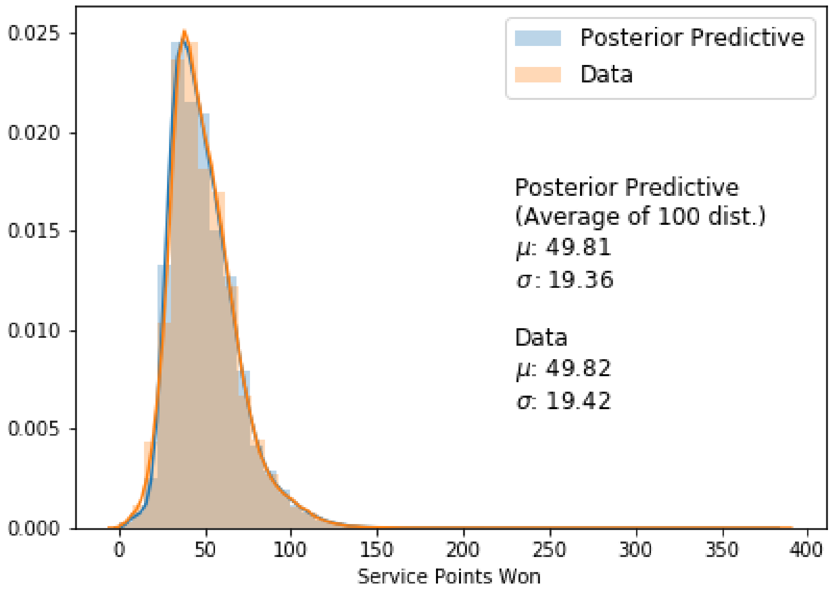

Figure 7: Posterior predictive

Figure 7 shows the posterior predictive distribution. The 1000 simulations of the posterior predictive was flattened into a single vector in order to plot one distribution instead of 1000. The posterior predictive distribution fits the data fairly well. The average mean of the 1000 posterior predictive distributions is equal the mean of the data; however, the average standard deviation of the 1000 posterior predictive distributions is significantly lower compared to the data. This may indicate that our model is slightly overconfident.

Posterior predictive distributions was also generated for a random subset of the data and 95% equal tail confidence intervals were calculated. 92.1% of the data lies within the 95% confidence interval of the posterior predictive. These plots can be found in the appendix.

3.4.3. Analysis

The model is used to analyse our question of interest. The same 4 players analysed in the Bradley Terry model will be analysed again with the Bayesian hierarchical model. θ is calculated the top 4 player matchups and plot the distributions.

Table 5: No. times Player$_{row}$ beats Player$_{col}$ on hard court historically

	Nadal	Djokovic	Federer	Murray
Nadal	–	7	9	7
Djokovic	18	–	18	20
Federer	11	16	–	12
Murray	5	8	10	–

Comparing Figure 8 to Table 5, the the model lines up well compared to the data. The probability of winning a service point on hard court as predicted by the model lines up with historic match outcomes. A player who has beaten another player more times historically also has a higher probability of winning a service point in the model. The only ambiguous matchup is Federer vs Djokovic. Since the posterior means of the distributions for both players in this matchup is nearly identical, each player can be assigned a 50% chance of winning the match since their serve and return probabilities are symmetric. [8]. 18 wins for Djokovic compared to

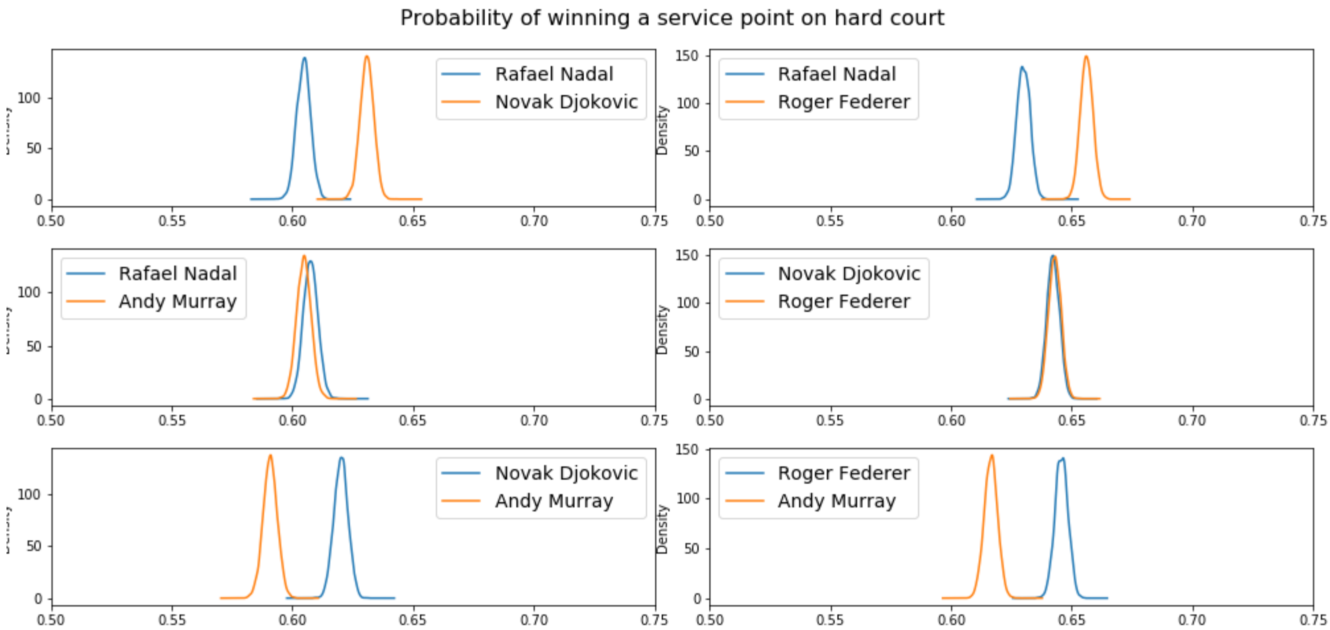

Figure 8: Big four player matches on hard surface

16 wins for Federer historically is a feasible prediction from the model if historic matches are considered to be draws from Binomial$(34, 0.5)$. The probability of drawing less or equal to 16 from Binomial$(34, 0.5)$ is roughly 43%.

The skill level of players changes dramatically when clay court matches are analysed.

Rafael Nadal has a clear advantage over all other players when the match is played on clay and the advantage is also quite large compared to Federer/Djokovic's advantage on hard court. This result is intuitive since he has won 11 French Open titles and a large number of Master 1000 clay tournaments. He also has a career average match win rate of 91% on clay. The model also fits the historic data table (Table 6) quite nicely.

The model also lets us compare Federer and Murray who have never played a clay court match with each other. The model tells us that Roger Federer would have a higher chance of winning the match. The actual probability of a player winning the match given the probability of winning a service point requires challenging combinatorics and is left out of this report. Details of the calculation can be found in Newton (2004) and the appendix. The model could be extended to predict match outcome probabilities given more time. [8]

The model can also be used to identify the players with the highest serve (a), return $(b) and surface (\delta(s)$ skills. Ivo Karlovic, John Isner, Milos Raonic, and Andy Roddick have the top 4 highest service skill (in that order). Looking at the statistics of these players, they all hold top 5 fastest serve speed records on the ATP website and known for having a good serve due to their height. Andy Murray has the highest return skill (with Djokovic coming in as close second). Filippo Volandri has the highest clay skill (Nadal coming in third) and Novak Djokovic

Table 6: No. times Player$_{row}$ beats Player$_{col}$ on clay court historically

	Nadal	Djokovic	Federer	Murray
Nadal	–	15	13	7
Djokovic	7	–	4	5
Federer	2	4	–	0
Murray	2	1	0	–

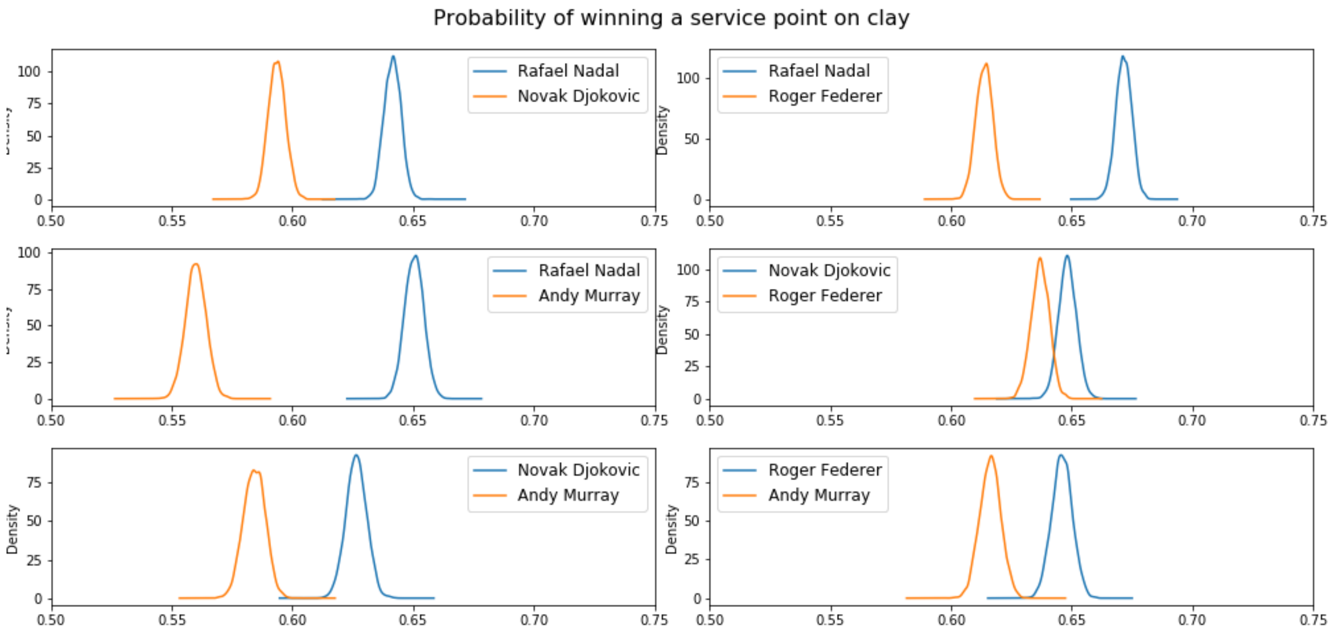

Figure 9: Big four player matches on clay surface

has the highest hard court skill. It is important to realise that surface skill is not indicative of the total skill of a player on that surface. A players surface skill is relative to his skill on other surfaces. A player who has equal ability on all surfaces will have a surface skill of 0 on all surfaces and can be characterised solely by his serve and return skill. Although Volandri has the highest clay skill, the model predicts that Nadal has a much higher probability of winning a point compared to Filippo Volandri on clay.

In terms of overall skill defined by $a_i + b_i + \delta_i(s) + c$, the model concludes that Roger Federer has the highest overall skill on grass. Djokovic and Federer are very close on hard courts. Nadal has the highest aggregate skill on clay by a sizeable margin, however, he has lower skill than Federer and Djokovic on hard and grass courts. At the 95% significance level, Nadal is significantly better than Djokovic, Federer and Murray on clay and Federer is significantly better than Djokovic, Nadal and Murray on grass.

From these results, I conclude that Federer is the best player on grass and, Nadal is the undisputed 'King of Clay'. Djokovic is hard to classify since he is no better than Federer on hard, worse than Federer on grass and worse than Nadal on clay. I call him the 'Jack of all surfaces' since he falls in between Nadal and Federer on grass and clay. I did not take into account carpet surface in this analysis since modern players rarely, if ever, play on carpet courts.

4. Conclusions

In this paper I looked at two Bayesian models to analyse the skill level of ATP tour tennis players since 1998. The Bradley Terry model gives Djokovic the highest skill parameter, Nadal comes in second, then Federer, then Murray. However at the 95% confidence, the Bradley Terry shows no significant difference in skill level between the top 3 players. The Bayesian hierarchical model gives us a more detailed ranking of the players. The hierarchical model gives Nadal the highest ranking on clay courts, and Federer the highest ranking on grass courts. Federer also comes first on hard courts and Djokovic comes in second, however there is no significant difference between them. Andrew agassi (a top player from the older generation) ranks third, however both Djokovic and Federer are significantly higher skilled than Andrew on hard court. Nadal's ranking drops significantly on all surfaces apart from clay, and Federer's ranking drops significantly on clay. Djokovic maintains a similar ranking on all surfaces. From this, there is enough significant evidence to label Nadal as the 'King of Clay', Federer as the 'King of Grass' and Djokovic as the 'Jack of all surfaces'.

4.1. A model extension

The same Bayesian hierarchical model was fit with one addition; the year that the match was played was introduced as $\zeta(year)$. This parameter adjusts a players ability based on the year (similar to the surface parameter).

I fitted the same Bayesian hierarchical model with one change. The year of the match was introduced as a parameter in the extended model as $\zeta(year)$. This parameter adjusts a players ability based on the year (similar to the surface parameter). The model remains the same otherwise and this paramater is introduced exactly the same as the surface parameter in the base model. The year a match was played is an important predictor since player skills vary greatly by year for many reasons. 2015-2018 data was used to reduce time taken to fit the model and the overall results from the extended model are similar to the base model. The model gives Andy Murray the highest ζ in 2016, Rafael Nadal in 2017 and Marco Cecchinato in 2018.

For context, Andy Murray had his best season in 2016 where he reached the final of 2 Grand Slams, won Wimbledon, ATP Tour Finals and the Olympic Gold and reached world number 1 for the first time. He also won 90% of his matches that season. [14] Nadal also had his best (post 2015) season in 2017 where he won 2 Grand Slams (his first Grand Slam win since 2014) and only lost 1 match in the clay season. Marco Cecchinato also had one of the best 2018 seasons as he climbed from rank 109 to rank 19 and achieved a career high win rate. Player skills can also be compared by year; these graphs are attached in the appendix.

Another use of the model is to predict matches between two different players in different years. For example, a common question of debate in the tennis community is "Was 2010 Nadal better than 2011 Djokovic?" since they both won 3 Grand Slams in these years. The extended model makes this analysis possible.

4.2. Limitations and Self Criticism

There are quite a few limitations to the model. First, the simplicity of the specified model for θ doesn't provide adequate detail for tennis betting. There are a few parameters missing that could improve the model such as a time dependent skill parameter. This is because player performance changes drastically throughout the years due to injuries and physical or mental well-being. Ideally, if the purpose of the model is prediction, it would be preferable to fit the model only on more recent data since model fitting would be much quicker. It would also be good to add a parameter that discounts previous matches compared to recent matches since only relatively recent performance is important placing bets and determining current skill level. Barnett (2006) goes into more detail and proposes to incorporate a tournament level parameter for each player since players might play in specific geographic locations and weather conditions.

If there were more time on this project, or it had to be redone, I would have skipped the Bradley Terry model and focused on fine-tuning the hierarchical model. It would be interesting to optimise the Stan code using some of Gelman's tricks. For example, Gelman proposes that setting $\theta_i = \mu \times \eta_i$ allows for noticeably faster computation; however, the reason behind this is unknown to me. Granted more time, it would be interesting to rent server time to give the HMC more time to run and allow it to fit a bigger set of parameters by introducing more variables into the proposed model. A question raised by a colleague is whether the model takes into account that Federer had no real rivals from 2001-2005 since Djokovic and Nadal are both 5 years younger than Federer. That is, it is possible that Federer's statistics in the current model are boosted by the lack of strong players early on in his career. A short answer to this is that the model partially takes this into account through partial pooling of the parameters and including the returner's parameters in the model. If it were the case that Federer faced weaker players, his opponents skill level would go down as well as his own increasing. Therefore, even if it were the case that Federer's statistics boosted due to weaker players, the model limits this bias. However, a more in-depth analysis would be needed to confirm this answer.

5. Appendix

5.1. A note on match probabilities

The model also does not calculate the probability that one player beats another in terms of the overall match, it only calculates the probability of a player winning a service point against the other. This doesn't matter too much because the probability of winning the match is monotonic in the probability of winning a service point. The calculation is much easier if there was a single probability for a player winning a service game and return game however this is not the case. Given that both these probabilities need to be incorporated and the complex point structure of the game, the probability of a player winning a match is a complex combinatoric calculation due to the deuce-advantage structure, the set tie breaker and the many different ways of reaching different match outcomes. For example, there are over 20 different ways of getting to 6-3. If the probability winning a service point and return point is equal then each way has the same probability, however if they aren't then each way may have a different probability. Moreover, the player who serves first becomes important to the calculation since some set outcomes have different probabilities depending on who serves first. The complete analytic solution to this problem is a large nested summation which can be found in Newton Keller (2004). One advantage of the point-based model explored in this paper is the calculation of intra-match probabilities. The probability that a match end in 3,4 or 5 sets and the probability most common set outcomes (e.g. 6-3 or 6-4) in the match can be calculated. [8]

5.2. Bradley Terry model

Figure 10: Network visualisation of tennis matches

The network visualisation shows that not all players are connected for example the outside ring of players only have arrows away from them and none towards them.

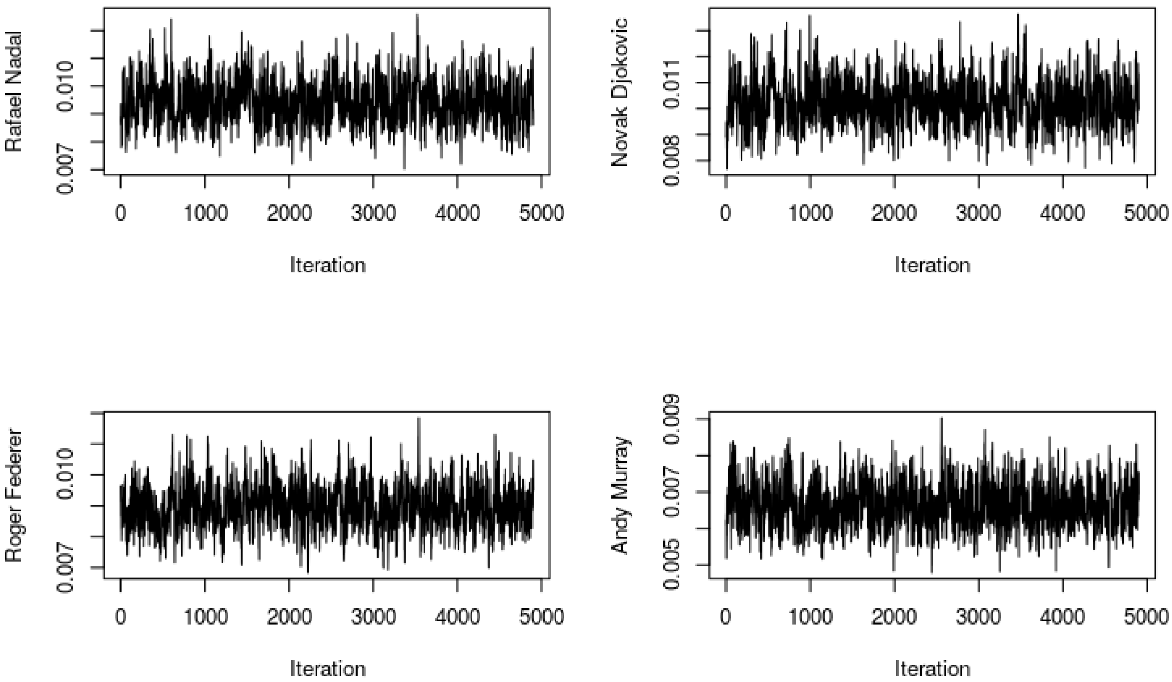

Figure 11: Traceplots of λ for Bradley Terry

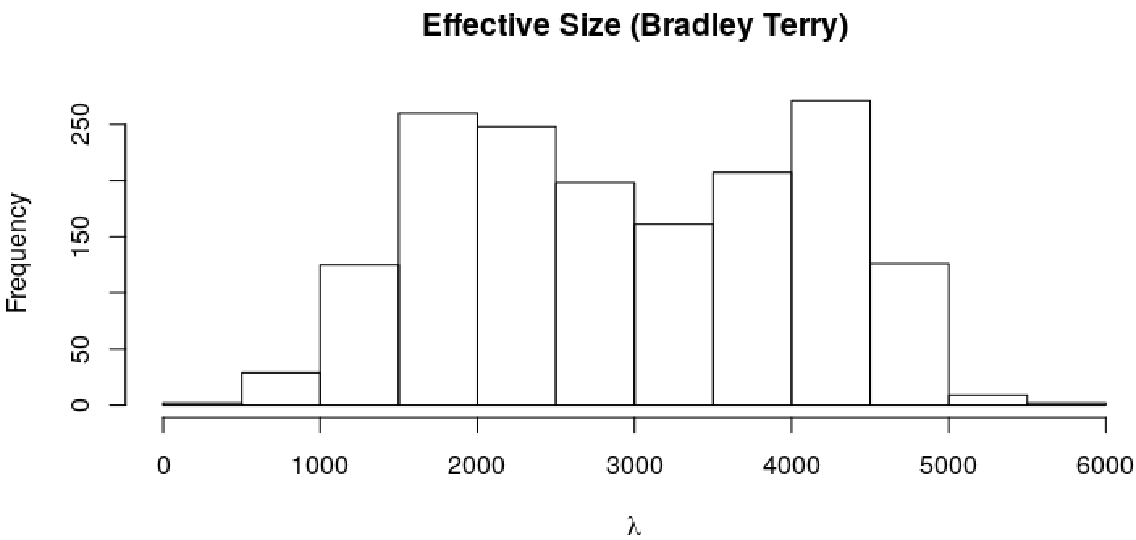

Figure 12: Effective size for λ_i

5.3. Bayesian hierarchical model

5.3.1. Posterior distributions

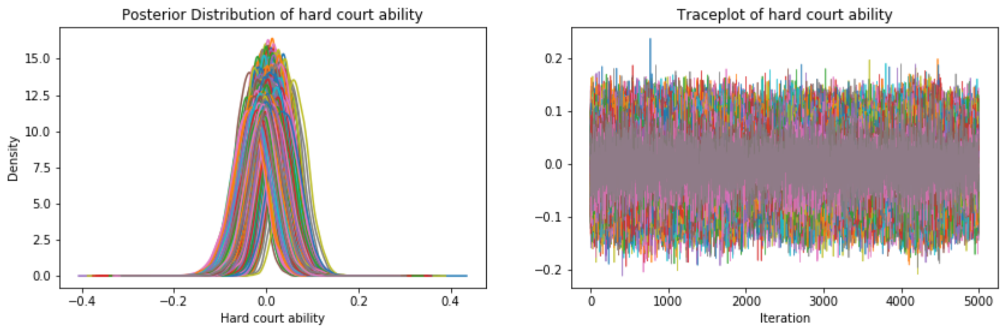

Figure 13: Posterior distribution for $\delta(hard)$

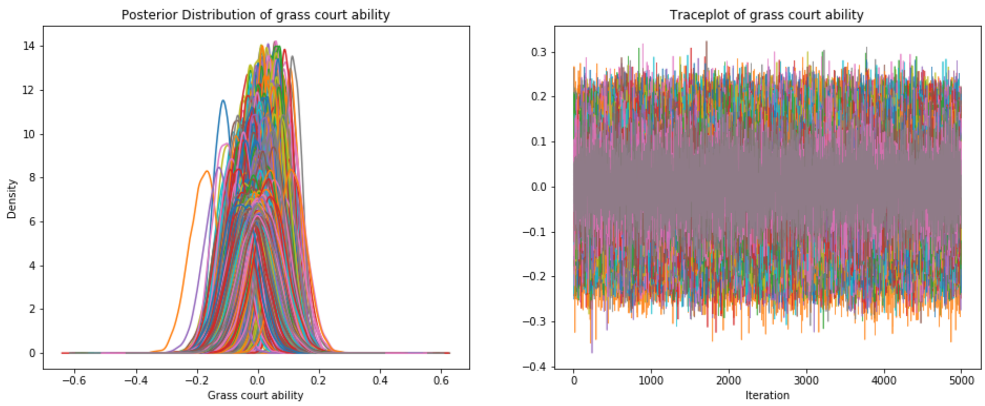

Figure 14: Posterior distribution for $\delta(grass)$

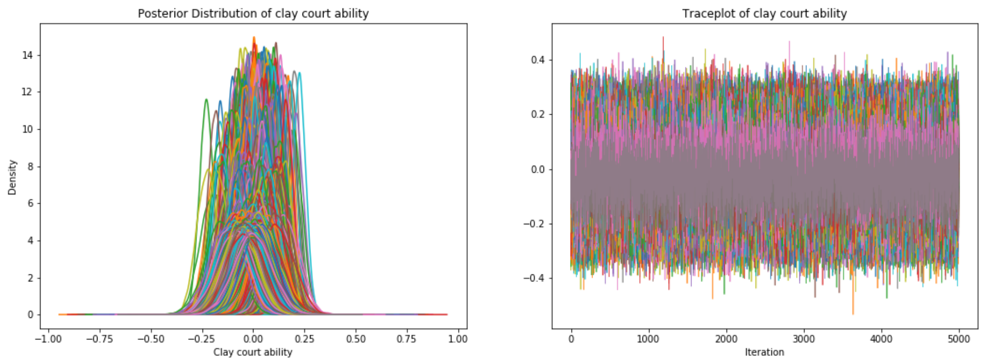

Figure 15: Posterior distribution for $\delta(clay)$

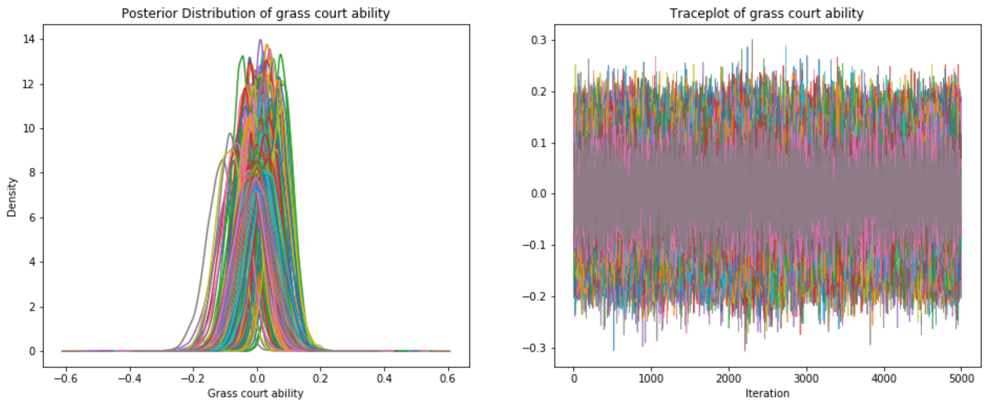

Figure 16: Posterior distribution for $\delta(carpet)$

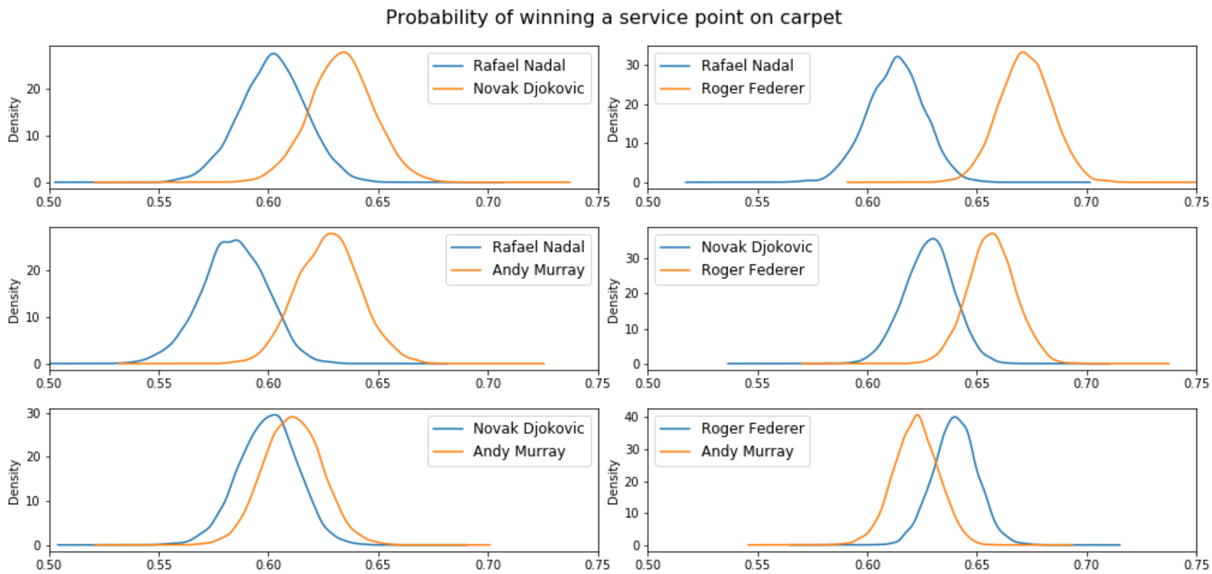

Figure 18: Top 4 player matchups on carpet

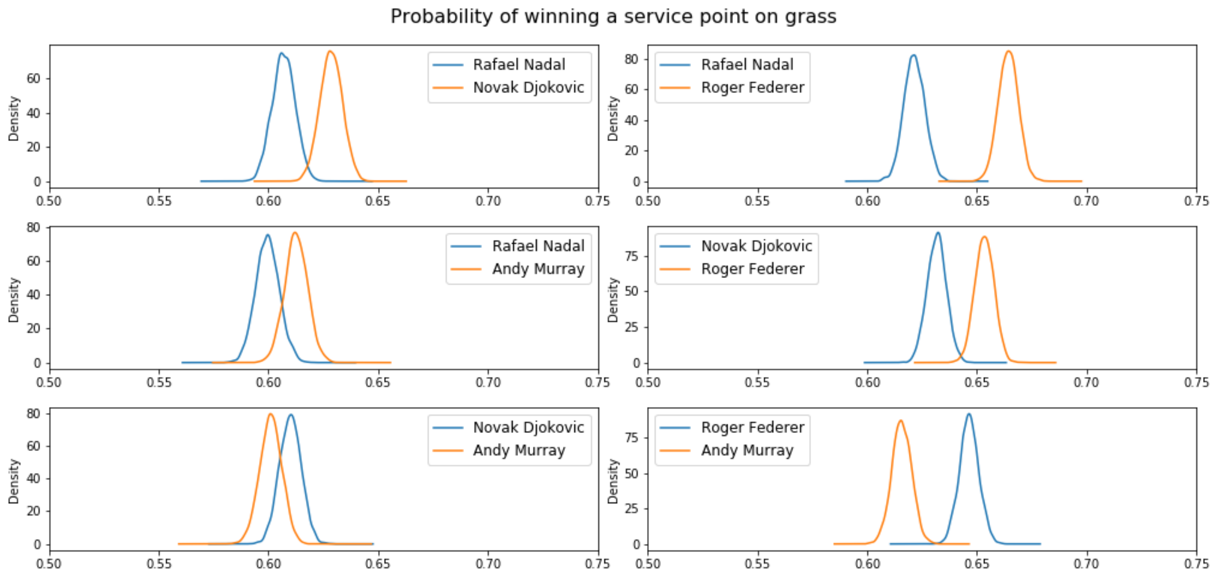

Figure 19: Top 4 player matchups on grass

5.3.2. Posterior checks

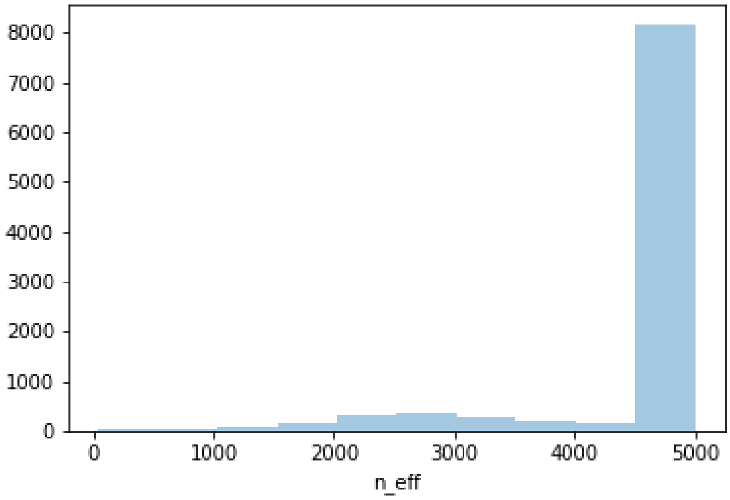

Figure 20: Effective sizes of parameters

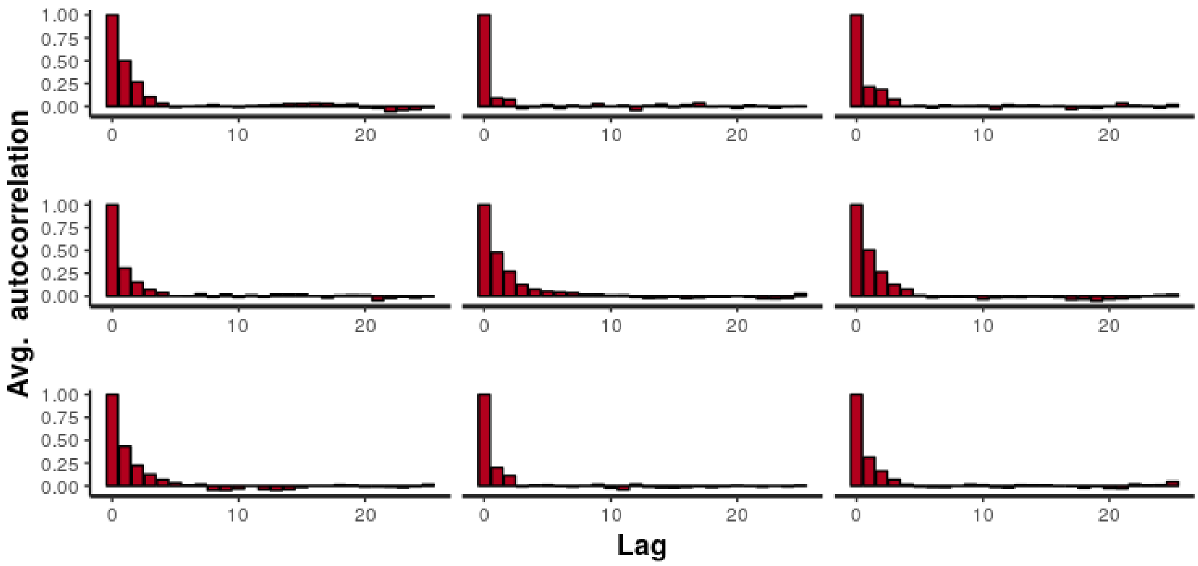

Figure 21: ACF plots of select parameters

5.3.3. Posterior predictive

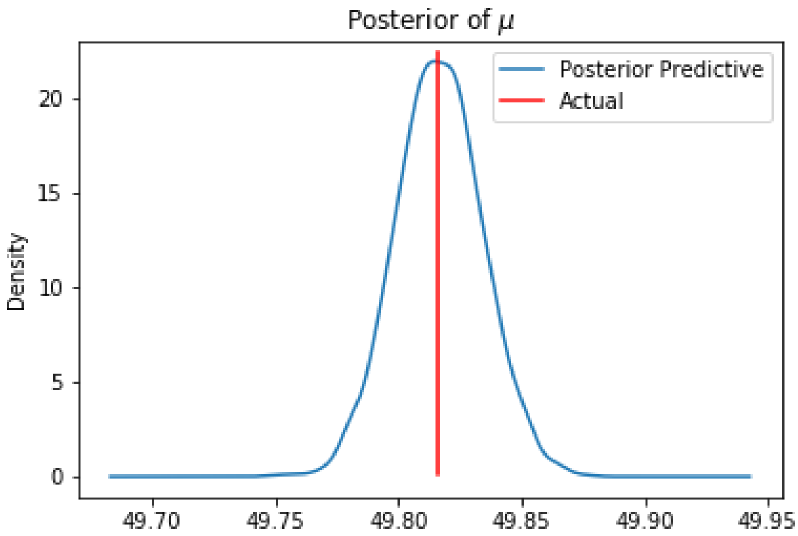

Figure 22: Average μ of 1000 simulated posterior predictive distributions

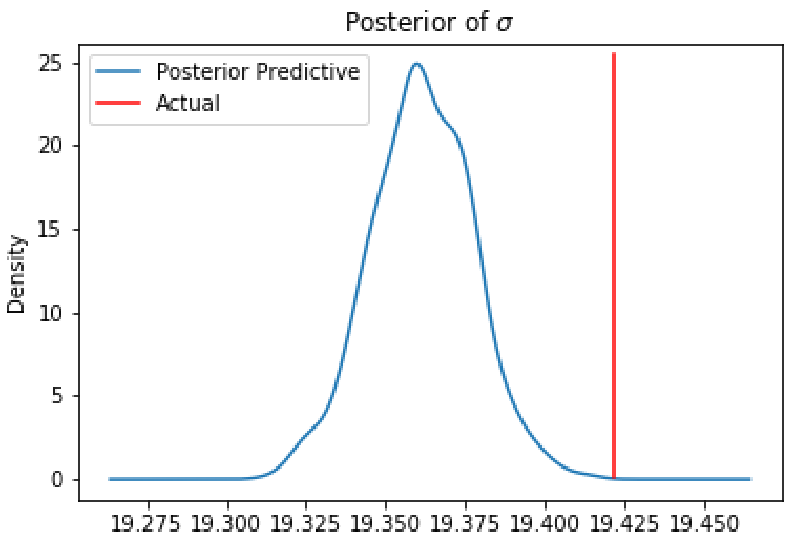

Figure 23: Average σ of 1000 simulated posterior predictive distributions

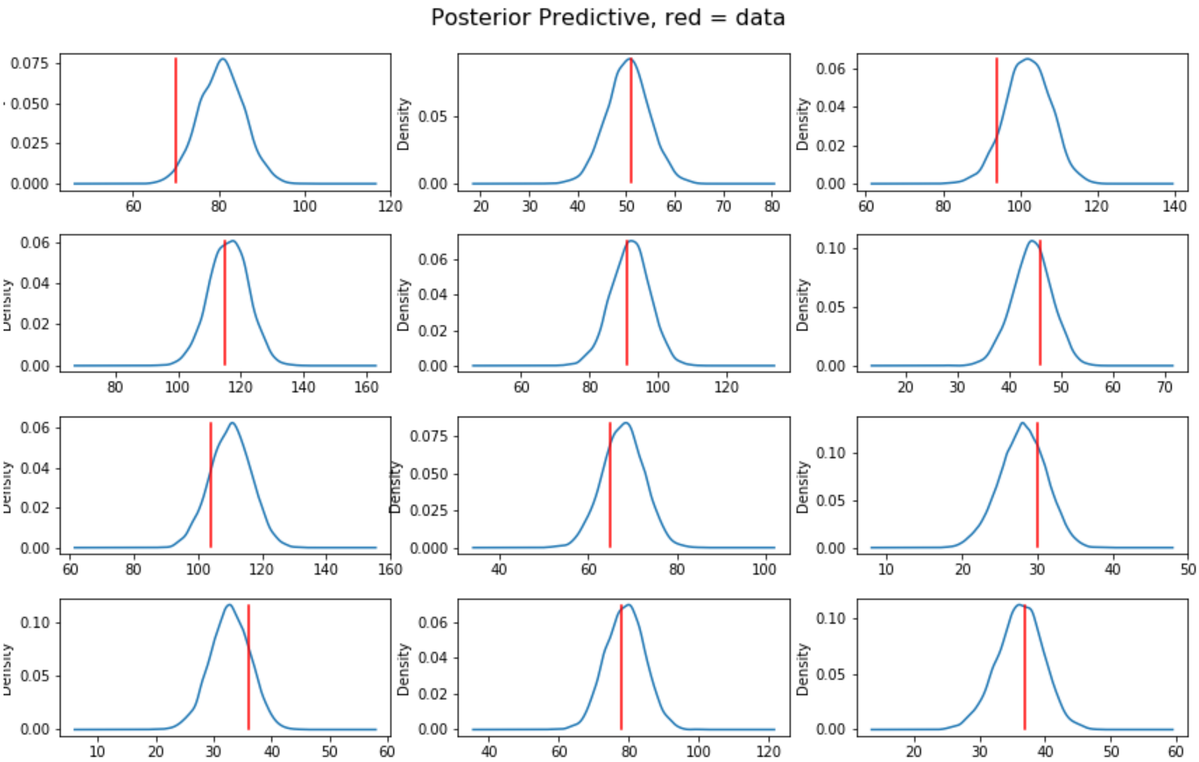

Figure 24: Posterior predictive distributions for 12 rows in the data

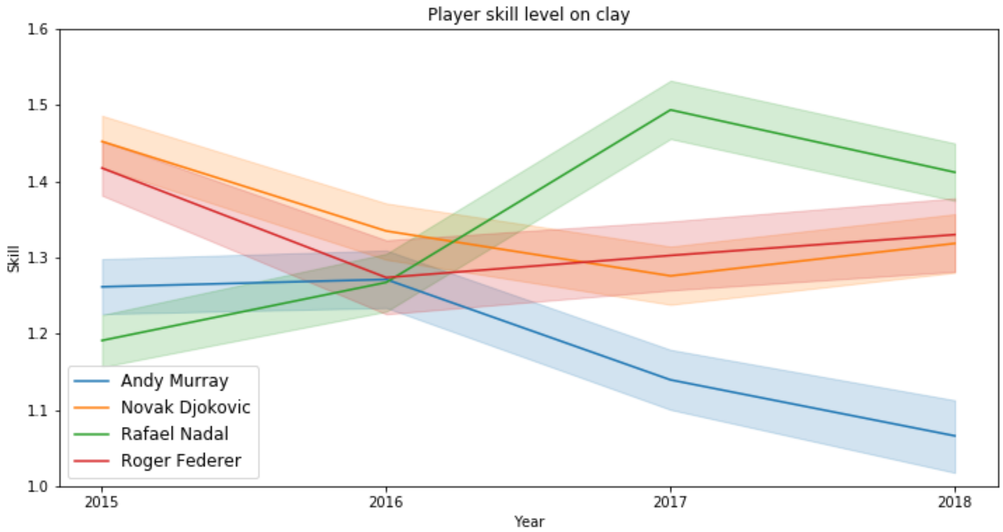

Figure 25: Big four players skills on clay court by year (50% CI)

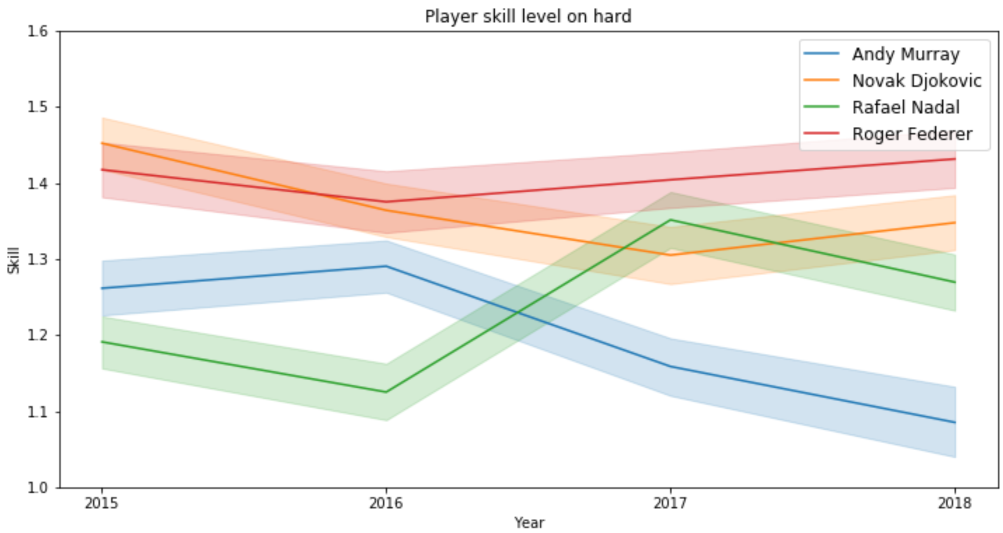

Figure 26: Big four players skills on hard court by year (50% CI)

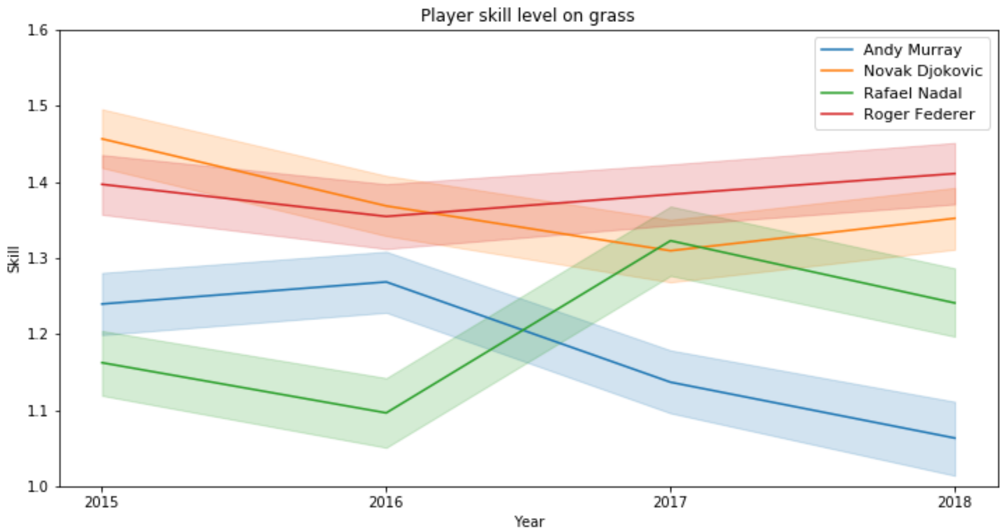

Figure 27: Big four players skills on grass court by year (50% CI)

[1] J. Sackmann, Assocation of tennis professionals match data, 2018.

[2] L. R. Ford, Solution of a ranking problem from binary comparisons, The American Mathematical Monthly 64 (1957) 28–33.

[3] E. Kaye, D. Firth, Fitting the bradley-terry model to large and potentially sparse datasets, https://cran.r-project.org/web/packages/BradleyTerryScalable/vignettes/BradleyTerryScalable.html, 2017.

[4] F. Caron, A. Doucet, Efficient bayesian inference for generalized bradleyterry models, Journal of Computational and Graphical Statistics 21 (2012) 174–196.

[5] Wikipedia, Big Four (tennis) — Wikipedia, the free encyclopedia, http://en.wikipedia.org/w/index.php?title=Big%20Four%20(tennis)&oldid=867446509, 2018. [Online; accessed 09-November-2018].

[6] T. J. Barnett, Mathematical modelling in hierarchical games with specific reference to tennis (2006).

[7] O. A. James, Probability Formulas and Statistical Analysis in Tennis, Journal of Quantitative Analysis in Sports 4 (2008) 1–23.

[8] P. K. Newton, J. B. Keller, Probability of winning at tennis: Theory and data (2004).

[9] M. Ingram, A point-based bayesian hierarchical model to predict the outcome of tennis matches, http://nessis.org/nessis17/Ingram.pdf, 2017.

[10] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language, Journal of Statistical Software, Articles 76 (2017) 1–32.

[11] M. Betancourt, A Conceptual Introduction to Hamiltonian Monte Carlo, ArXiv e-prints (2017).

[12] C. Fonnesbeck, A primer on bayesian multilevel modeling using pystan, http://mc-stan.org/users/documentation/case-studies/radon.html, ????

[13] S. D. Team, Bayesian statistics using stan, http://www.stat.columbia.edu/~gelman/bda.course/_book/, ????

[14] Wikipedia, 2016 Andy Murray tennis season — Wikipedia, the free encyclopedia, http://en.wikipedia.org/w/index.php?title=2016%20Andy%20Murray%20tennis%20season&oldid=867118103, 2018.