A Story of Comeback in ATP Tennis Tour

Yinan Fang

May 8, 2015

1 Introduction

For the past decade, the men's professional tennis world has been dominated by four greatest players we have ever seen: Roger Federer, Rafael Nadal, Novak Djokovic and Andy Murray. Media and fans often refer them as "Big Four". Since Australian Open 2005, almost all the ATP tour trophies have been shared by Big Four. They have won 39 out of the last 43 men's Grand Slam Single titles; they have also won 10 out of previous 12 World Tour Finals. Also,they have been consistently occupying the top four places in ATP single's ranking since 2008.

During their whole professional careers so far, what makes Big Four so successful on tennis court is not just those sweeping dominating wins, but also the tough ones when they managed to battle back after dropping one set. In this study, I will focus on their ability to manage a victory after dropping the first set. I gathered the data which consist of the scores of all matches which Big Four played throughout their careers from the time they turned pro to the end of 2014 season. An autoregressive model descirbed in section 3 will be utilized to study the probability of comeback win for each player over their careers. Bayesian methods will be used to fit the model. Through the posterior samples from the MCMC sampler, we will see how the comeback probability has been evolving over years for Big Four and how these comeback performance patterns match with their overall career paths.

2 Data

The original data are from the website http://tennis-data.co.uk/alldata.php. The data files contain records of all matches played on the ATP tour each year, including the information of tournament, surface, players, scores, result, etc. I extract all matches involving "Big Four" from the time they turned pro to the end of 2014. So, the data used in this project contain records of Murray from 2005, Djokovic from 2003, Nadal from 2002 and Federer from 1998. Among these records, I picked out the matches in which they lost the first set. After summarizing the number of comeback victories by each player at each year, Table 1 shows the datasets which are analyzed in this project. Figure 1 contains four plots, each of which shows a player's number of matches in which he fell behind after the first set and number of matches in which he eventally won after dropping the first set.

3 Notations and Model

3.1 Notations

In this study, we have Murray's data for 10 years (2005-2014), Djokovic's data for 12 years (2003-2014), Nadal's data for 13 years (2002-2014) and Federer's data for 17 years (1998-2014). The key notations used in the model are listed below:

- Use p as the index for players: $p = 1$ represents Murray; $p = 2$ represents Djokovic; $p = 3$ represents Nadal; $p = 4$ represents Federer.
- Use t as the index for year: for $p = 1$, $t \in \{1, ..., 10\}$; for $p = 2$, $t \in \{1, ..., 12\}$; for $p = 3$, $t \in \{1, ..., 13\}$; for $p = 4$, $t \in \{1, ..., 17\}$.
- N_{pt}: the number of matches in which player p dropped the first set in year t.
- y_{pt} : the number of comeback matches won by player p after dropping the first set in year t.
- θ_{pt} : the comeback probability for player p in year t.

3.2 Model

Here, we assume the independence among four players. Figure 2 shows the plots of "logit(comeback rate)" against "year" for Big Four. From the patterns of points in the plots, it is appropriate to assume an autoregressive model on $logit\left(\theta_{pt}\right)$ for each player. Consider the following model:

- given θ_{pt}, binomial model for y_{pt}: $y_{pt} \overset{iid}{\sim} Bin(N_{pt}, \theta_{pt})$;
- if $t = 1$, $logit\left(\theta_{p1}\right) = \mu_p + \epsilon_{p1}$, where $\epsilon_{p1} \overset{iid}{\sim} N\left(0,\ \sigma_p^2\right)$;
- if $t > 1$,$logit(\theta_{pt}) = \mu_p + \rho_p\left(logit\left(\theta_{p,t-1}\right) - \mu_p\right) + \epsilon_{pt}$, where $\epsilon_{pt} \overset{iid}{\sim} N\left(0,\ \sigma_p^2\right)$.

We need to assign prior distirbutions on parameters μ_p, ρ_p, σ_p. Given limited prior information for these parameters, noninformative priors are applied in the bayesian analysis: the uniform priors are used for μ_p and ρ_p, and half cauchy prior is assigned for σ_p. Thus, for each player,

- $p\left(\mu_p\right) \propto 1$, $p\left(\rho_p\right) = I\left(-1 < \rho_p < 1\right)$, $p\left(\sigma_p\right) \sim Cauchy^{+}(0, 1)$;
- Priors for μ_p, ρ_p and σ_p are mutually independent.

Thus

$$p\left(\mu_p, \rho_p, \sigma_p\right) = p\left(\mu_p\right) p\left(\rho_p\right) p\left(\sigma_p\right) \propto \frac{1}{1+\sigma_p^2} I\left(\sigma_p > 0\right) I\left(-1 < \rho_p < 1\right).$$

Fit the model described above in stan, run four separate Markov Chains with 2000 iterations (first 1000 iterations as burn-in), and we obtain 4000 MCMC samples of all parameters: θ_{pt},μ_p, ρ_p, σ_p. For all parameters, the potential scale reduction factors are less than 1.1. By checking the trace plots and autocorrelation function plots for all parameters, there is no indication of lack of convergence for the Markov Chains.

We assess the model fit using posterior replicates of data. To obtain a replicate of the data, do the following:

1. obtain a set of posterior samples for θ_{pt} from Markov chain as $\theta_{pt}^{(j)}$;
2. generate a replicate of data: $y_{pt}^{(j)} \sim Bin(N_{pt}, \theta_{pt}^{(j)})$, for every p, t.

Figure 3 gives the histograms of 4000 posterior replicates of data. The dashed black line represents the observed data y_{pt} in each histograms. According to Figures 3, except two extreme low observations, Nadal in 2004 and Federer in 1998, there is no sign of significant lack of fit.

4 Result

Now, let us see how each player's comeback victory rate changes over years. In Figure 4, the 95% credible intervals, as well as posterior median, are presented for each parameter θ_{pt}. The pattern of how comeback rate evolves for each player can be summarized from these plots.

Murray: Overall, Murray's comeback ability has been gradually improving since he turned pro, except the 2010 season. The posterior median of comeback probability increases from 0.363 in 2005 to 0.444 in 2014.

Djokovic: Unlike Murray, Djokovic's comeback probability did not show much increase at the early stage of his career until 2011. While, in 2011, Djokovic's comeback probability dramatically increased, and stayed at relatively high level until 2014. This pattern is consistent with Djokovic's career path. He had a breakout year in 2011, winning three Grand Slams and finishing the year with a 70-6 record. Since 2011, he has been one of the most dominant players in the tour.

Nadal: In his first year, Nadal had a very high starting point (posterior median 0.458) of comeback probability. It fell back for the next two years. However, his comeback ability improved significantly in 2005 season. This is the year when he won his first grand slam and started to rise as a top player. From 2005 to 2011, Nadal's comeback ability seems to be relatively consistent, except for a drop in 2009, the same year when he lost the only French Open title in the past ten years. Since 2011, his comeback probability has a increasing trend and the poesterior median is above 0.5 in the past two seasons.

Federer: Despite the low starting point in his first few years, from 2000 to 2006, the posterior median of Federer's comeback probability remarkably increased from 0.269 to 0.619. However, in 2007, it suffered a sudden drop to 0.462. During 2007 and 2013, the posterior median of Federer's comeback probability wondered between 0.4 and 0.5. In 2014, it seems to have an impressive increase again to 0.522. Most of these shifts match well with Federer's career path. From 2001 to 2006, he gradually rose to be the best player in the world. Nonetheless, since 2007, the maturity of Nadal, along

with the emergence of Djokovic and Murray, challenged Federer's dominance over men's tennis.

In Figure 5, we show the posterior means of θ_{pt} changing over years for Big Four. This plot demonstrates how they compares with each other in comeback ability over years. Examine the comparison starting from 2005, when all players' careers began to overlap. Between 2005 and 2008, it shows sthat Federer and Nadal had better chance to achieve comeback than Djokovic and Murray. However, starting from 2009 season, the difference diminished and four players were almost at the same level until 2014. This also validates the widely accepted opinion that Djokovic and Murray joined Big Four around 2009 to end the duopoly of Federer and Nadal at the summit of tennis.

5 Discussion

In the analysis above, we fit a model for the number of comeback victories for Big Four. For the comeback probability parameters (θ_{pt}), an autoregressive model of $logit\,(\theta_{pt})$ is assigned for each player. Assuming the noninformative priors, the posterior samples are obtained by running MCMC in stan. From the summaries above, we can see that the patterns of posterior medians and credible intervals for θ_{pt} over years reflect Big Four's individual career paths very well. And, the trend of posterior means of θ_{pt} in Figure 5 indicates the time when Big Four emerged in the ATP tour. In this project, we use the independent models among different players. An hierarchical model can also be considered here by adding mixing distributions to ρ_p, μ_p and σ_p.

In this project, we mainly focus on how the overall comeback performance of Big Four develops during their careers. Other interesting questions can also be studied. For example, the surface (hard, clay, grass) information for each match is available in original data. A Bayesian analysis can be conducted to study their comeback rates on different surfaces as well.

Tables and Figures:

year	player	Y	N
2005	Andy Murray	5	20
2006	Andy Murray	7	28
2007	Andy Murray	9	20
2008	Andy Murray	11	26
2009	Andy Murray	6	13
2010	Andy Murray	5	20
2011	Andy Murray	7	18
2012	Andy Murray	11	21
2013	Andy Murray	9	15
2014	Andy Murray	13	27

(a) Murray

year	player	Y	N
2003	Novak Djokovic	1	4
2004	Novak Djokovic	6	13
2005	Novak Djokovic	3	13
2006	Novak Djokovic	7	23
2007	Novak Djokovic	7	22
2008	Novak Djokovic	9	24
2009	Novak Djokovic	11	27
2010	Novak Djokovic	7	24
2011	Novak Djokovic	9	13
2012	Novak Djokovic	10	19
2013	Novak Djokovic	10	17
2014	Novak Djokovic	7	12

(b) Djokovic

year	player	Y	N
2002	Rafael Nadal	5	9
2003	Rafael Nadal	5	18
2004	Rafael Nadal	2	17
2005	Rafael Nadal	9	15
2006	Rafael Nadal	13	22
2007	Rafael Nadal	9	21
2008	Rafael Nadal	9	18
2009	Rafael Nadal	7	21
2010	Rafael Nadal	7	13
2011	Rafael Nadal	5	15
2012	Rafael Nadal	3	7
2013	Rafael Nadal	10	16
2014	Rafael Nadal	13	21

(c) Nadal

year	player	Y	N
1998	Roger Federer	0	4
1999	Roger Federer	5	22
2000	Roger Federer	5	28
2001	Roger Federer	11	29
2002	Roger Federer	5	20
2003	Roger Federer	9	22
2004	Roger Federer	7	10
2005	Roger Federer	7	8
2006	Roger Federer	8	10
2007	Roger Federer	4	11
2008	Roger Federer	8	18
2009	Roger Federer	9	16
2010	Roger Federer	6	14
2011	Roger Federer	3	11
2012	Roger Federer	11	20
2013	Roger Federer	7	20
2014	Roger Federer	9	14

(d) Federer

Table 1: Comeback records of Big Four: N represents the number of matches in which the player dropped the first set; Y represents the number of comeback victories after dropping the first set

(a) Murray (b) Djokovic

(c) Nadal (d) Federer

Figure 1: Plots of each player's comeback record over years: black points represent the number of matches in which each player lost the first set; red points represent the number of matches in which each player achieved a comeback win after dropping the first set.

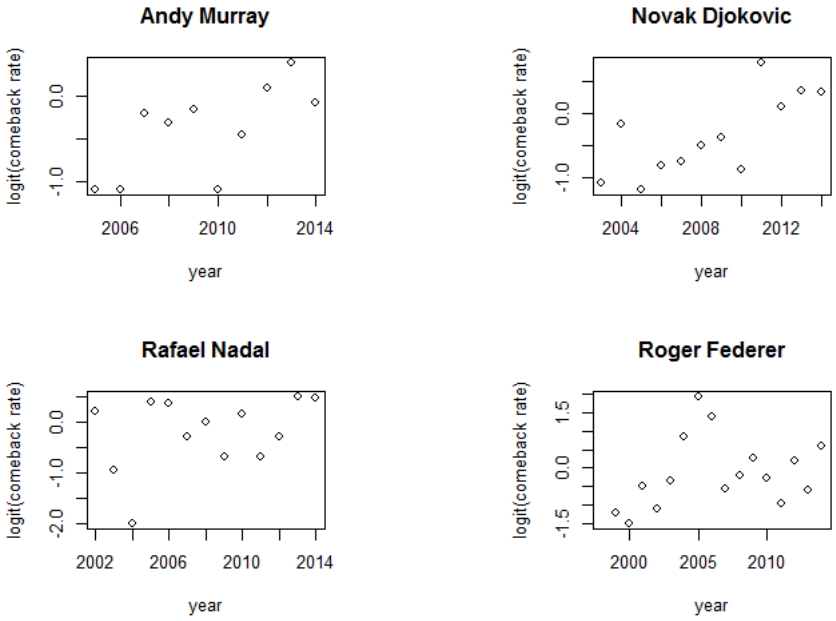
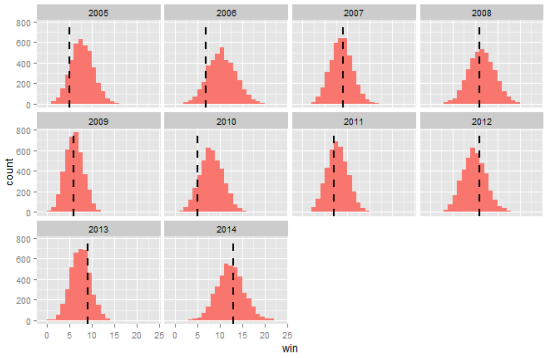
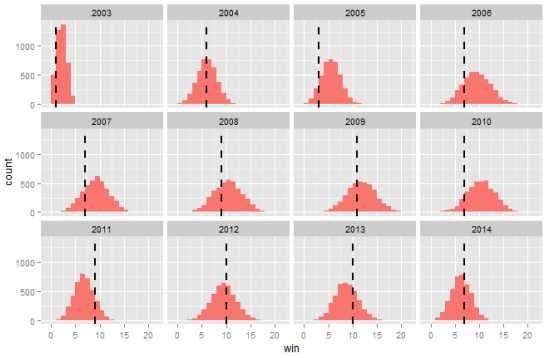


Figure 2: plots of observed"logit(comeback rate)" vs. "year" for Big Four

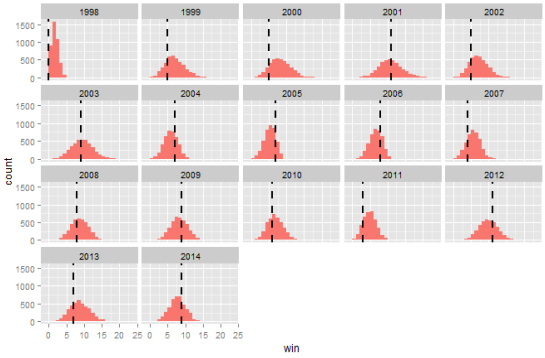


(a) Murray



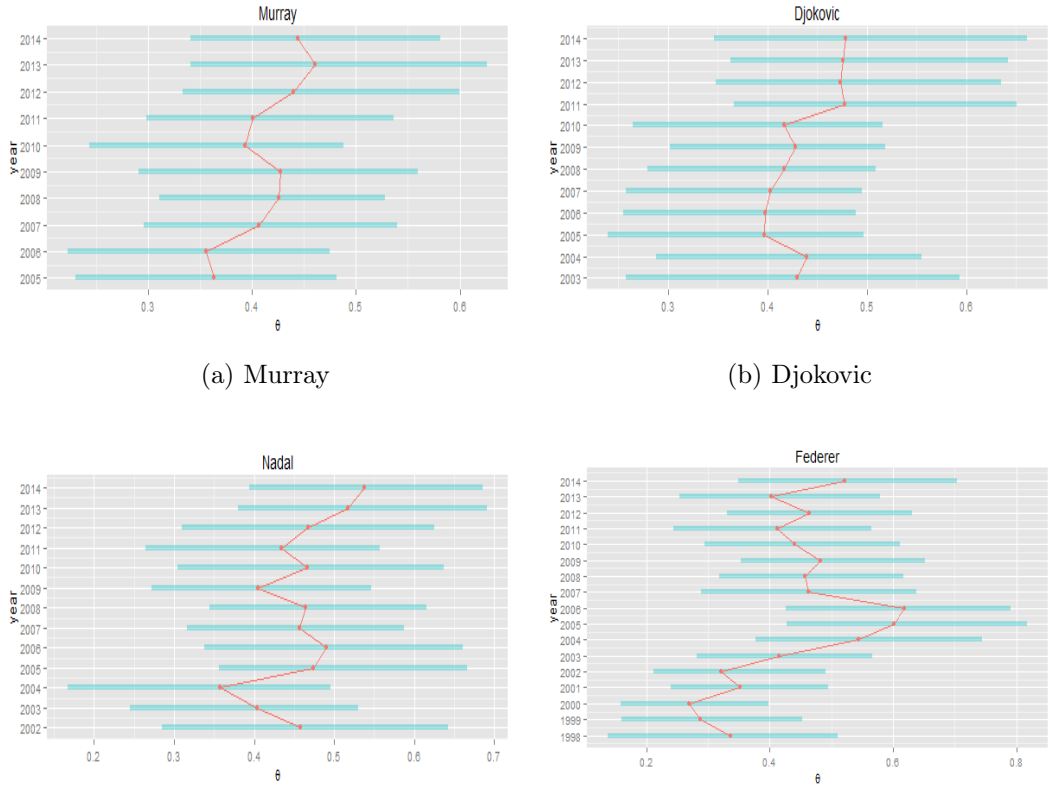
(b) Djokovic

(c) Nadal



(d) Federer

Figure 3: Histograms of posterior replications $\widetilde{y_{pt}}$; the dashed black line represents the observed data y_{pt}



(a) Murray (b) Djokovic

(c) Nadal (d) Federer

Figure 4: 95% credible intervals of θ_{pt}(comback probability for each player over years): the blue segement at each year represents the 95% credible interval; the red dot represents the median of MCMC samples

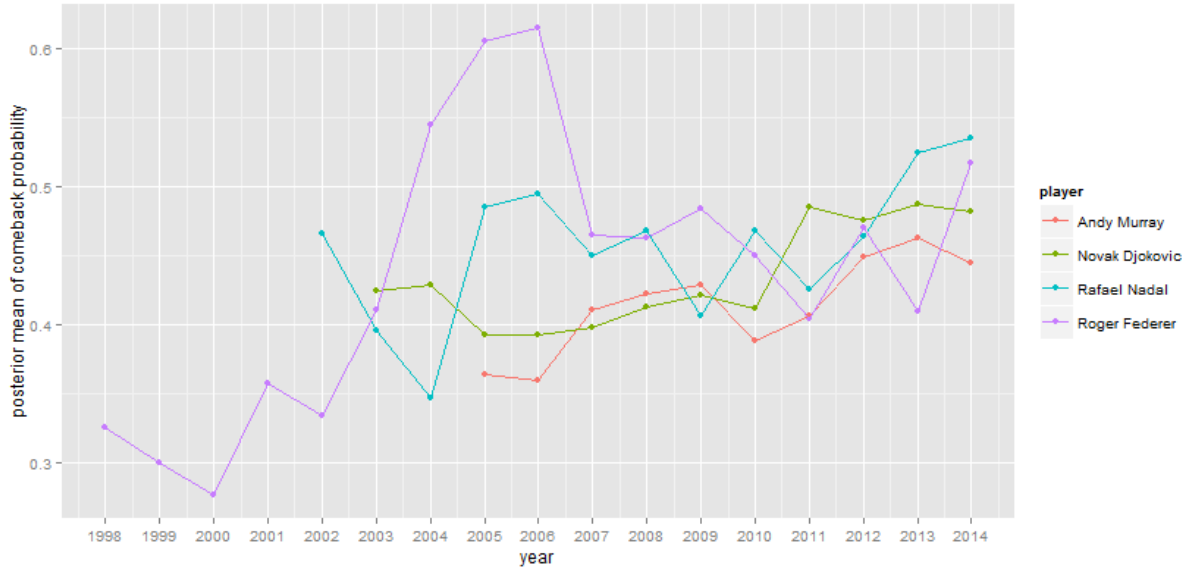


Figure 5: posterior mean of comeback probabilityθ_{pt} changing over years for Big Four

Key Code in Stan:

```
model1=" data{
  int<lower=0> M;
  int<lower=1> len[4];
  int<lower=0> y[M];
  int<lower=0> N[M];
 }
parameters{
   vector<lower=0,upper=1>[M] theta;
   vector[4] mu;
   vector<lower=-1,upper=1>[4] rho;
   vector<lower=0>[4] sigma;
 }
transformed parameters{
   vector[M] q;
   q <- log(theta)-log(1-theta);
 }
model{
   int ind;
   ind <- 0;  ###pointer in the loop to get specific player
   sigma ~ cauchy(0, 1);  ###prior for sigma
   rho ~ uniform(-1, 1);   ####prior for rho
   y ~ binomial(N, theta); ###data model
   for(k in 1:4){  ###autoregressive model for logit(theta)
      q[ind+1] ~ normal(mu[k], sigma[k]);
      for(i in 2:len[k]){
        q[ind+i] ~ normal(mu[k]+rho[k]*(q[ind+i-1]-mu[k]), sigma[k]);
      }
      ind <- ind + len[k];
   }
}
"
m1=stan_model(model_code="model1")
d <- list(M=nrow(dat8), y=dat8$win, N=dat8$lose1, len=len)
s1 <- sampling(m1,d,c("theta","q", "sigma","mu", "rho"))
result1 <- extract(s1)
```