Comparing yearly batting average for Sachin Tendulkar using Bayesian Hierarchical Model

Aditya Ranade

April 30, 2021

1 Introduction

Cricket is the most popular game in India and is played and watched in every corner of the country. Consistency is an important and desired quality for any sportsperson. If we look closely at the most well known and best players in any sports, more often then not, they are known for their consistant performances. In this project we analyse the cricketing career of Sachin Tendulkar, considered to be one of the best players to have ever played the game of Cricket who broke into the Indian Cricket team in 1989 as a 16 year old and went on to play for 25 years which is remarkable for any sportsperson at the International level.

We will compare the average number of runs scored per innings in each of the 25 years of the said player's career. During the career of the said player, Cricket was majorly played in 2 formats. First is Test matches which are played over 5 days and each team gets to bat and bowl for 2 innings each. Second is One Day International matches (ODI) where each teams gets to bat and bowl for one inning each of 50 overs. More details of the format are given in section 4.

2 Data

The data has been obtained from kaggle. For this report, we have considered the number of runs scored in a match, the year/season of the match, the format of the match and the location of the match (if the match is played in home country or away from home). Since some games get abandoned due to various reasons most likely due to bad weather, those were not considered. Similarly the games where the player did not get a chance to bat have also not been considered.

A note on the interpretation of average according to official cricket record keeping is the number of runs scored by the player between two times the player gets out. However, the average we are considering in this project is in the usual sense where we divide the total runs scored by the number of innings taken to score the runs. To illustrate, if a player bats and scores 300 runs cumulatively in 10 matches in a season and gets out only once then the average will be 300 according to the official records but for the project it will be 30.

Here, season is the calender year since international cricket is played throughout the year. The way the game is played, we can expect the variation to be extremely high. For example, runs scored by any player who gets an opportunity to bat is a non-negative integer. So in a match a player can get out on 0 and in the next match the player can score 100 runs. Similarly if a player plays 20 matches in a season, high variation can be expected. In fact the way the formats are designed, variation in One Day Internationals (ODI) can be expected to be much higher than Test matches.

3 Methods

3.1 Model

We are comparing the averages for the said player through seasons using Hierarchical Bayesian model.

The model can be described as follows.

$$Y_i \overset{i.i.d}{\sim} N(\mu_{year[i]}, \sigma^2)$$
$$\mu_{year} \overset{ind}{\sim} N(\theta, \tau^2)$$
$$\theta \sim N(\theta_0 = 50, \sigma_0 = 16)$$
$$\sigma \sim Ca^+(0, 20) \text{ for Test format and } Ca^+(0, 40) \text{ for ODI format}$$
$$\tau \sim Uniform(10, 50)$$

Y is runs scored in a particular inning. μ_{year} indicates the average for a particular calender year / season with common variance for the seasons. We use the same model for Test format and One Day international format, except for σ, where the prior values are different.

3.2 Tools

The data was analysed using JAGS through the rjags package in R to generate the samples from the posterior distribution of the parameters. We run 3 chains in model for each format with 10,000 iterations.

4 Results

4.1 Test Match Career

Test match is played over 5 days and each team bats alternately twice and the team that scores the more number of runs cumulatively wins the test match. The match is considered a draw if all the innings do not get complete in 5 days where an innings is complete when 10 out of 11 batsmen get out. A batsmen can bat till the innings is complete or till the particular batsmen gets out. Test cricket is considered the toughest format of cricket.

Sachin Tendulkar is the only player in the history to play 200 test matches with the next best player having played 168 test matches. In figure 1, we look at the posterior 95% credible intervals (values can be found in table 1 in Appendix A) for the average in each seasons (years 1989 to 2013) of test career. Since the number of matches played in a season is different, the variance is slightly on the higher side in general.

We can see the credible intervals for the average score in every season are more or less constant throughout the career with a couple of seasons at the end have considerably lower interval. The red vertical line is the Test career average (47.92). We can note intervals for all the years contain the career batting average and hence we can say the player has been a consistent player throughout his career. We can see from the intervals that year 2010 was probably the best season of the player's career. Overall we can say the player has been consistent throughout the career.

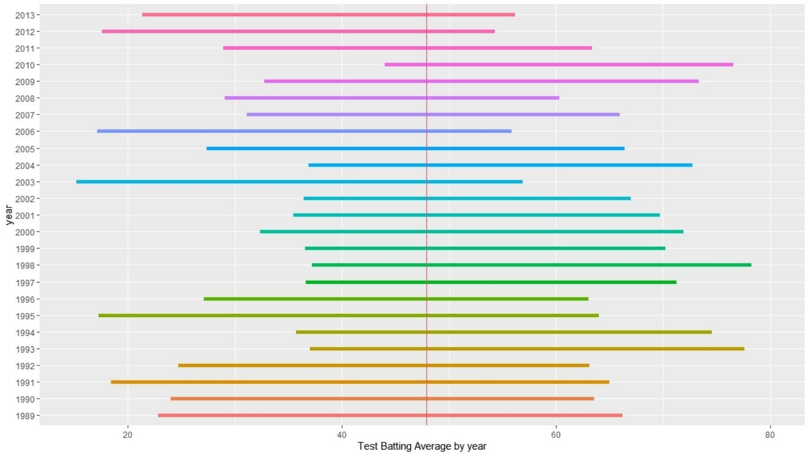

Figure 1: Credible Intervals of averages by year in Test Career

The posterior plot for the model parameters θ, σ and τ can be seen in figure 2 below (Posterior 95% credible interval values can be found in table 1 in Appendix A) and they seem to be on expected lines. Trace plots indicates proper mixing and convergence for each of the parameters listed. For the reasons discussed earlier, the variance is considerably high in the data which is evident in the plot for σ and τ.

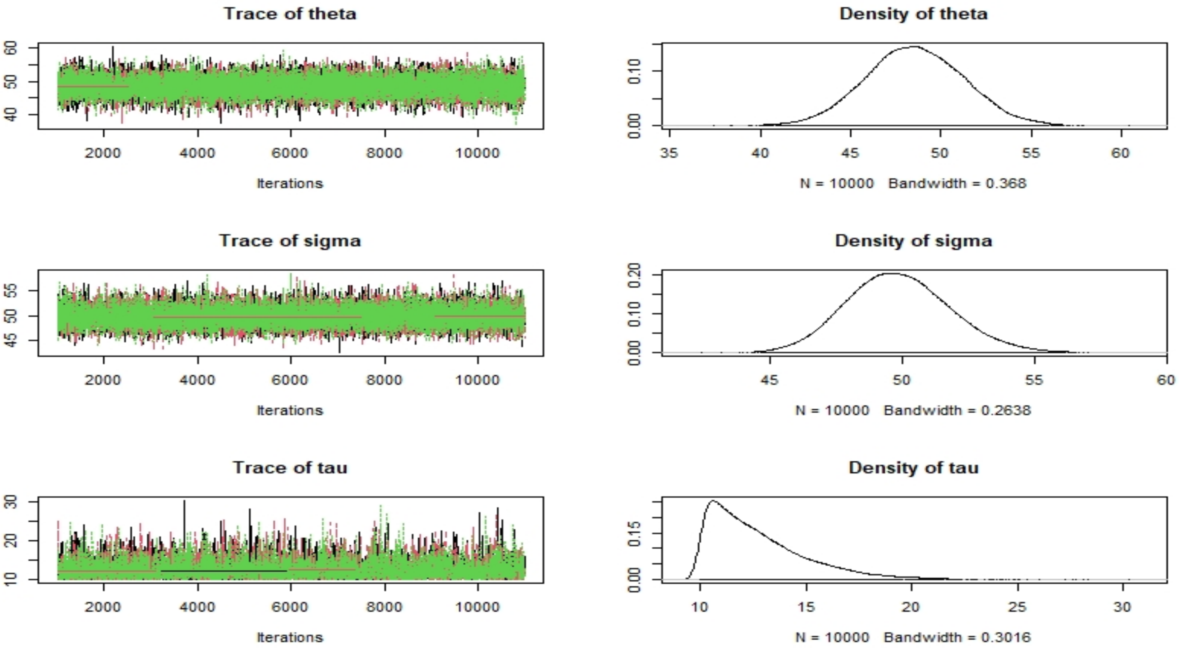

Figure 2: Posterior for model parameters (Test format)

Since there are mean parameter for 25 years, their posterior density plots can be found in Appendix A in figure 7. We do not include trace plots for the mean of each year. However trace plot for 3 randomly chosen years (1991, 2000 and 2010) can be found in appendix A in figure 8 which indicates proper mixing and convergence.

4.2 One Day International Match Career

One Day International match is more like an instant result match although the game lasts an entire day. It is played in a single day and each team bats once for a maximum of 50 overs. The team that scores the more number of runs wins the One day International match. If the team batting second scores one run more than the team batting first within their quota of 50 overs, then the team batting second wins the match and the match is considered complete and stopped at that point. So unless the match cannot be completed due to weather conditions, the match is guaranteed to produce a result unlike test match format.

Again Sachin Tendulkar has played the maximum number of One Day International matches ever at 463 ODIs. The next best player has played 448 ODIs. Since this format has limited number of overs in a match, each team tries to score maximum number of runs and usually every batsmen tries to score ‘fast’ and play more risky shots than required in order to squeeze in as much runs as possible and in turn the chance of the player getting out is always high. So naturally the variation in ODI is considered higher than Test matches for any player.

In figure 3, we look at the posterior 95% credible intervals (values can be found in table 1 in Appendix A) for the average in each seasons (years 1989 to 2012) of ODI career. We can see the credible intervals for the average score in every season are more or less constant throughout the career with a couple of seasons overall have considerably lower average.

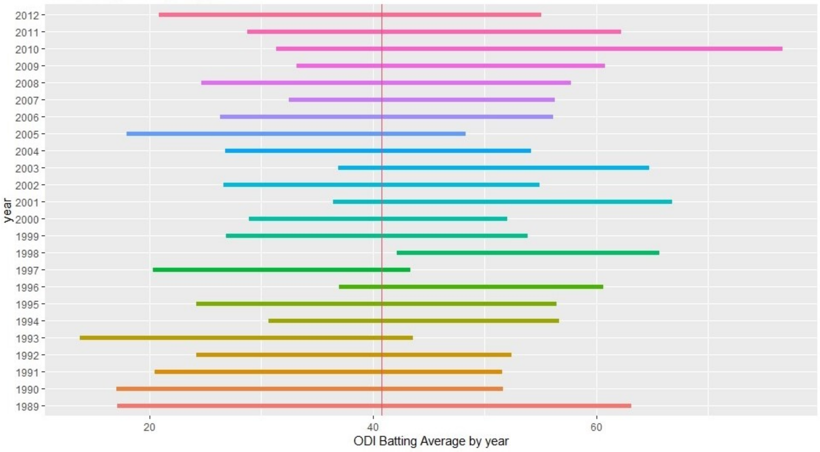

Figure 3: Credible Intervals of averages by year in ODI Career

The red vertical line is the player's career ODI average (40.77) and we can say the intervals in all but one year (1998) contain the career average. Hence we can say the player has been consistent player in One Day International format as well. Year 1998 was extraordinary for the said player as the interval contains all the values higher than the career average. As in test cricket, year 2010 can be considered good as well but the variation is considerably high. We can see the variation is high between the years as compared to test career.

The posterior plot for the model parameters θ, σ and τ can be seen in figure 4 below (Posterior 95% credible interval values can be found in table 1 in Appendix A) and they seem to be on expected lines. Trace plots indicates proper mixing and convergence for each of the parameters listed. For the reasons discussed earlier, the variance is considerably high in the data which is evident in the plot for σ and τ.

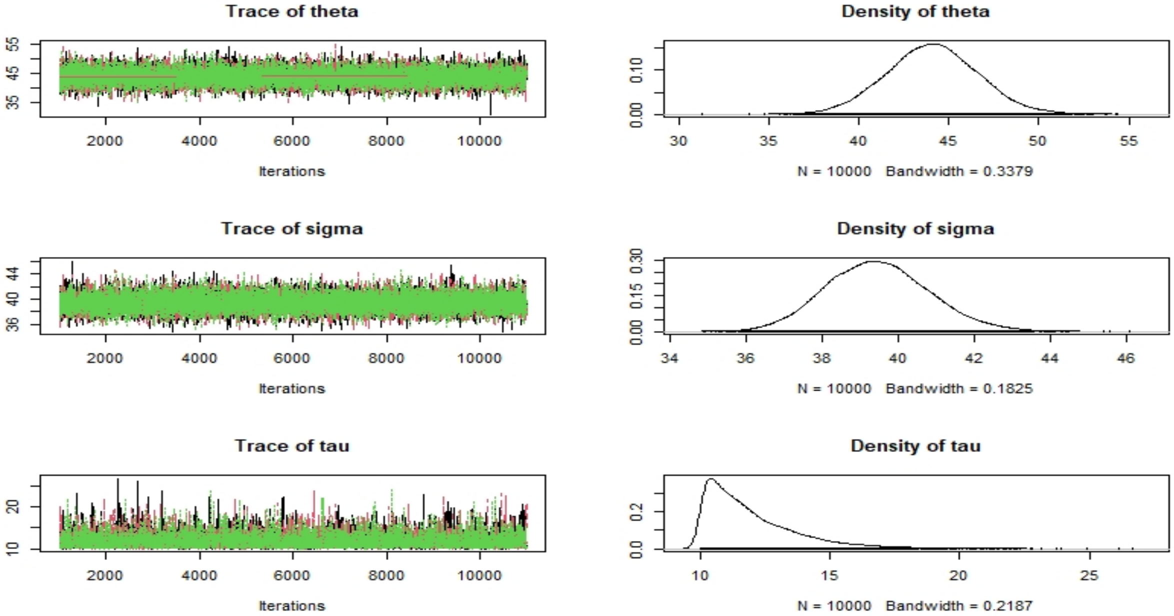

Figure 4: Posterior for model parameters (ODI format)

Since there are mean parameter for 24 years, their posterior density plots can be found in Appendix A figure 9. We do not include trace plots for the batting average of each year. However trace plot for 3 years (1998, 2005 and 2012) can be found in appendix A figure 10 which indicates proper mixing and convergence.

5 Discussion

As far as the analysis is concerned, the posterior for μ and σ remain similar even if we change the prior or hyperprior parameters. Whereas the posterior for τ (the standard deviation of the mean in each of the season) is sensitive to the prior choice depending on the format and the number of observations for 1 season. Hence a high variation is expected and more so since the sample points (in each year) are on the lower side as compared to career sample points in our data.

To check if the model used behaves well, a posterior predictive p-value is calculated using data replicates from the posterior distribution. We calculate the difference in runs scored on home soil (matches played in India) and away soil (matches played outside India) using a standard t-test using our observed data and replicates from posterior $p(y^{rep}|y)$. Comparison is made between the test statistic for observed data and predicted data. The same technique is used for Test and ODI formats and the plots for the respective formats can be seen in figure 5 and 6 respectively. Since the observed test statistic is around the middle of the replicated statistic, it is safe to say the model works well for both test and ODI formats.

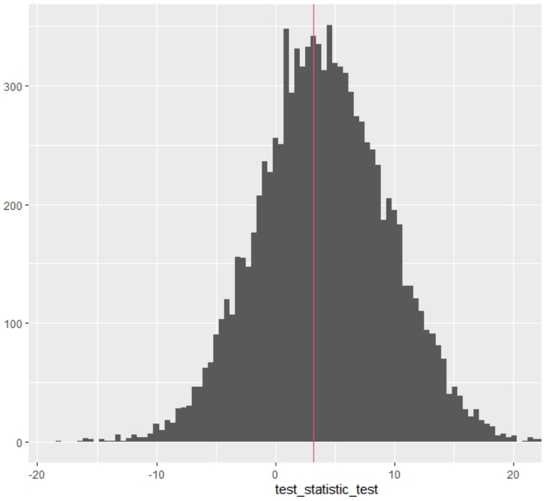

Figure 5: Posterior p-value (Test Format)

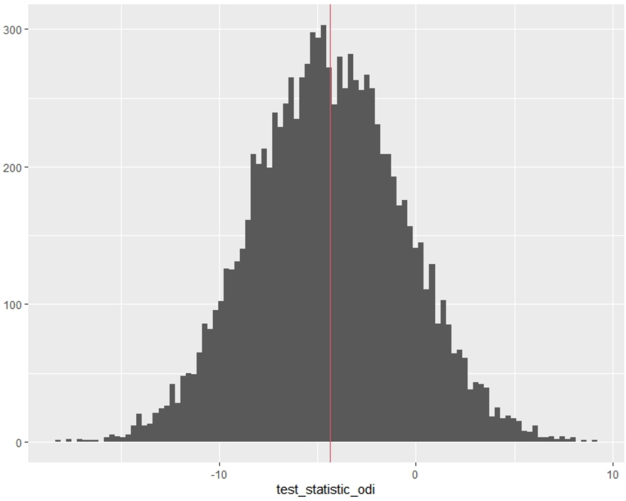

Figure 6: Posterior p-value (ODI Format)

Considering the two formats, we can say the player has been consistent player throughout his career. Playing the game for 25 years at the international level is certainly commendable but to be consistently performing well at that level speaks volumes about the skill and ability of Sachin Tendulkar. In cricketing circles, Sachin Tendulkar is highly regarded and is considered one of the best player to step on the cricketing field.

6 Future Work

In this age of data analytics and big data, there are specific tools developed which are specifically used in Sports as there are so many factors which affect the performance of a player in any sports to be specific. In terms of cricket some of the factors which can affect the performance is the opposition team, the location where the match is played as the pitch on which the game is played impacts the performance, the weather conditions to name a few. If we incorporate the opposition and the location where the match is played in the model, the variation would surely be better explained and in turn, the intervals can be expected to be on the shorter side. A model with linear regression for the mean parameter with the specified variables we just discussed can be considered for future work.

7 References

Internet
Dataset : https://www.kaggle.com/anirbna/sachin-tendulkar-batting-stats
Bayesian Data Analysis (Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin)

Appendix A

The posterior density plots for Test format can be seen in figure 7 below

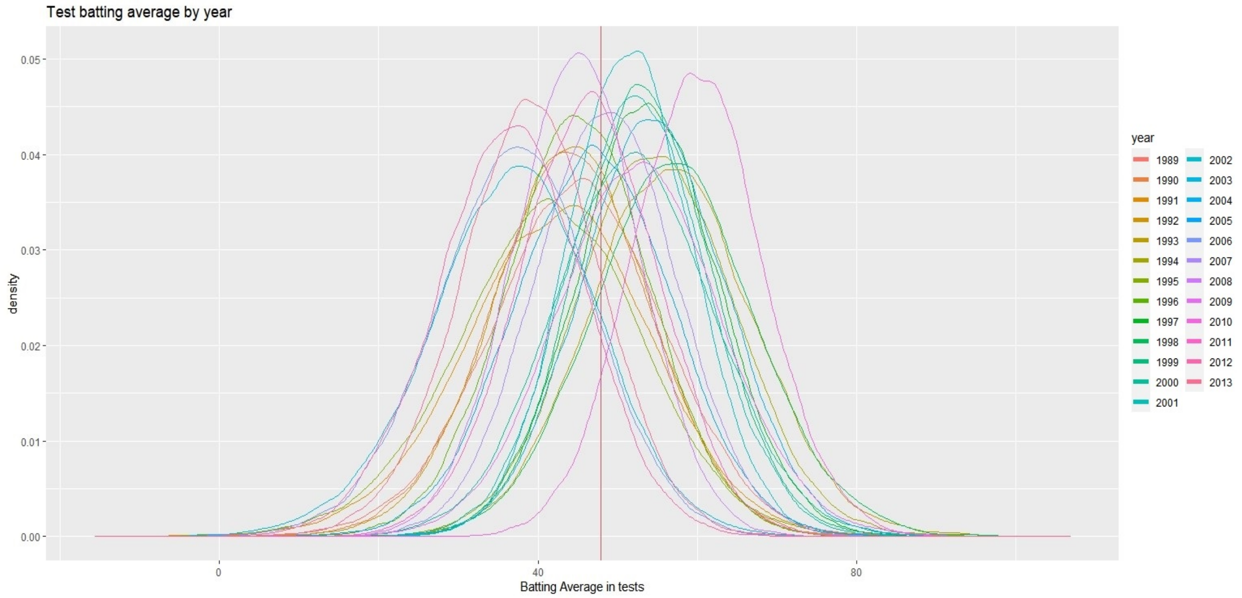

Figure 7: Posterior density plots of averages by year in Test Career

The trace plots for the mean of years 1991, 2000 and 2010 (μ_{1991}, μ_{2000} and μ_{2010}) from the Test career are in figure 8 below. It indicates proper mixing and convergence.

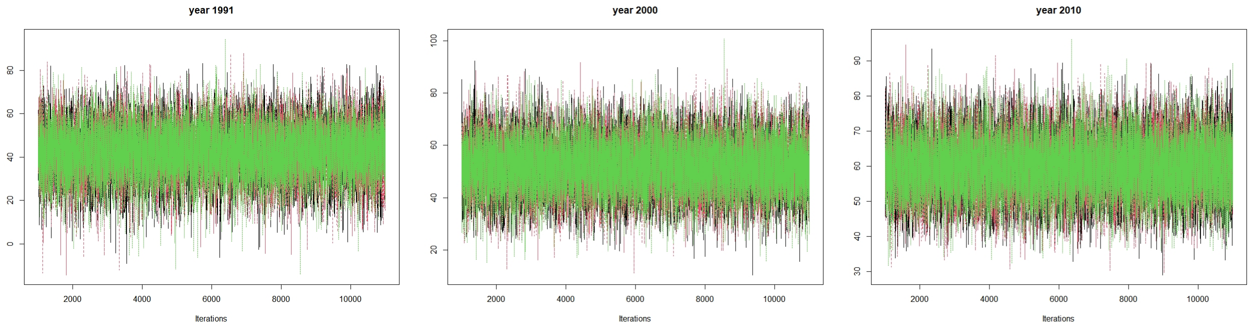

Figure 8: trace plot for μ_{1991}, μ_{2000} and μ_{2010} of test career

The posterior density plots for One Day Internationals (ODI) format can be seen in figure 9 below

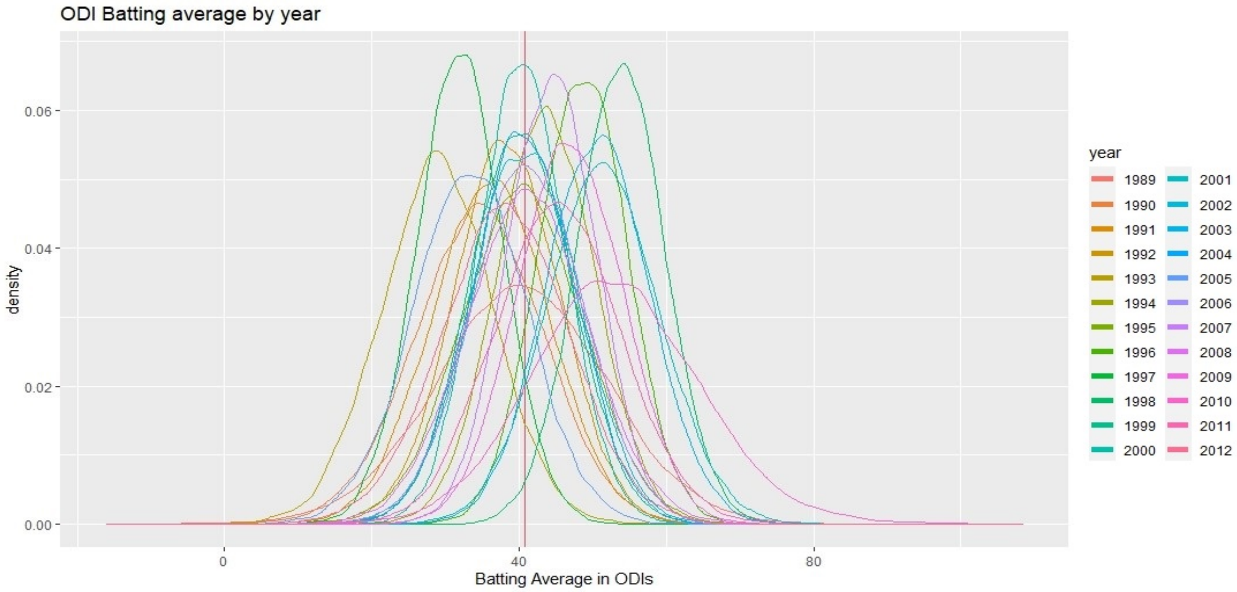

Figure 9: Posterior density plots of averages by year in ODI Career

The trace plots for the mean of years 1998, 2005 and 2012 (μ_{1998}, μ_{2005} and μ_{2012}) from the ODI career are in figure 10 below. It indicates proper mixing and convergence.

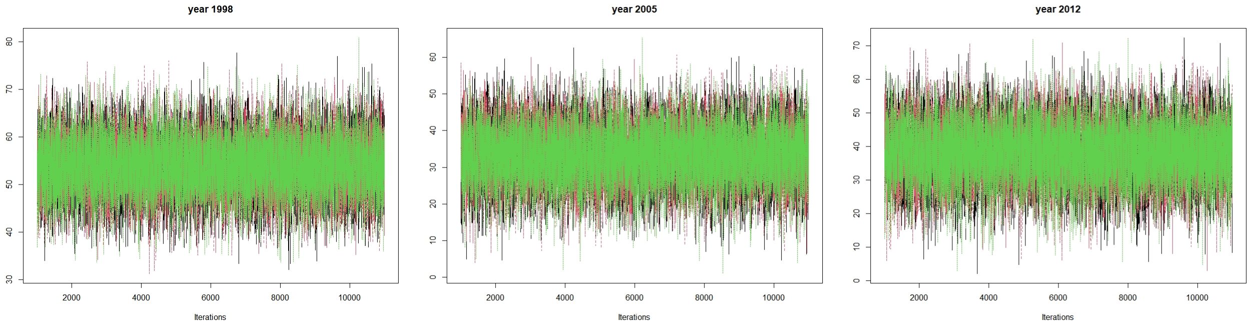

Figure 10: trace plot for μ_{1998}, μ_{2005} and μ_{2012} of ODI career

Table 1 below contains the 95% credible intervals for each of the parameter in the model for Test and ODI career

parameter	Mean(Test)	Credible Interval (Test format)	Mean(ODI)	Credible Interval (ODI format)
μ_{1989}	44.93	(22.48, 66.47)	40.36	(17.08, 63.24)
μ_{1990}	44.12	(23.87, 63.74)	34.54	(17.34, 51.63)
μ_{1991}	42.37	(18.64, 64.78)	36.16	(20.33, 51.86)
μ_{1992}	44.29	(24.55, 63.26)	38.44	(24.52, 52.26)
μ_{1993}	56.91	(36.8, 78)	28.79	(13.67, 43.67)
μ_{1994}	54.85	(35.58, 74.56)	43.71	(30.72, 56.61)
μ_{1995}	41.29	(17.05, 63.92)	40.46	(24.3, 56.96)
μ_{1996}	45.14	(27.16, 62.98)	48.74	(36.98, 60.56)
μ_{1997}	53.9	(36.6, 71.51)	31.92	(20.6, 43.12)
μ_{1998}	57.25	(37.39, 78.3)	53.9	(42.01, 65.73)
μ_{1999}	53.41	(36.6, 70.31)	40.22	(26.59, 53.72)
μ_{2000}	52.17	(32.55, 72.27)	40.41	(28.75, 51.91)
μ_{2001}	52.39	(35.1, 69.7)	51.31	(36.33, 66.44)
μ_{2002}	51.68	(36.49, 66.89)	40.81	(26.59, 55.12)
μ_{2003}	36.88	(15.03, 57.01)	50.72	(37.02, 64.68)
μ_{2004}	54.73	(36.76, 73.12)	40.51	(26.9, 54.23)
μ_{2005}	47.03	(27.14, 67.02)	33.39	(18.32, 48.44)
μ_{2006}	37.01	(17.17, 56.02)	41.28	(26.33, 56.26)
μ_{2007}	48.6	(31.04, 66.33)	44.45	(32.6, 56.38)
μ_{2008}	44.88	(29.42, 60.31)	41.08	(24.83, 57.38)
μ_{2009}	52.86	(32.93, 73.52)	46.87	(32.9, 60.82)
μ_{2010}	60.08	(44.33, 76.47)	53.14	(30.99, 76.58)
μ_{2011}	46.31	(28.98, 63.72)	45.39	(28.83, 62.04)
μ_{2012}	36.4	(17.4, 54.27)	38.08	(20.98, 55.05)
μ_{2013}	38.98	(21.37, 55.95)	NA	(NA, NA)
σ	49.86	(46.25, 53.86)	39.46	(36.96, 42.18)
τ	12.86	(10.09, 19.38)	11.92	(10.05, 16.52)
θ	48.56	(43.11, 54.02)	44.07	(39.2, 49.07)

Table 1: Parameter posterior 95% credible intervals for Test and ODI career

Please note Sachin Tendulkar had retired from One Day International (ODI) format in 2012 and hence no parameter estimates for μ_{2013} for ODI format.