Modelling career trajectories of cricket players using Gaussian processes

Oliver Stevenson

COMPASS Seminar Series 2018

University of Auckland

Statistics in cricket

- Many previous statistical studies in cricket — few on measuring and improving performance
- Our focus is on measuring player batting ability
- Batting ability primarily recognised using a single number
- **Batting average** $= \frac{\text{Total \# runs scored}}{\text{Total \# dismissals}}$

‘Getting your eye-in’

Batting is initially difficult due to external factors such as:

- The local pitch and weather conditions

Pitch conditions

Credit: *http://www.abc.net.au/news/image/6941478-3x2-340x227.jpg*

Credit: *http://sportbox.co.nz/wp-content/uploads/2013/12/WACA-pitch.jpg*

‘Getting your eye-in’

Batting is initially difficult due to external factors such as:

- The local pitch and weather conditions
- The specific match scenario

The process of batsmen familiarising themselves with the match conditions is nicknamed ‘getting your eye-in’.

Predicting the hazard

- **Hazard** = probability of a batsmen being dismissed on their current score
- Due to the 'eye-in' process, a constant hazard model is no good for predicting when a batsman will get out
  - Will under predict dismissal probability for low scores
  - Will over predict dismissal probability for high scores (i.e. when a player has their 'eye-in')

Predicting the hazard

Therefore it would be of practical use to develop models which quantify:

1. How well a player bats when they first arrive at the crease
2. How much better a player bats when they have their ‘eye-in’
3. How long it takes them to get their ‘eye-in’

Kane Williamson's career record

Kane Williamson
New Zealand

Full name Kane Stuart Williamson

Born August 8, 1990, Tauranga

Current age 27 years 120 days

Major teams New Zealand, Barbados Tridents, Gloucestershire, Gloucestershire 2nd XI, New Zealand Under-19s, Northern Districts, Sunrisers Hyderabad, Yorkshire

Playing role Top-order batsman

Batting style Right-hand bat

Bowling style Right-arm offbreak

Relation Cousin - D Cleaver

Explore Kane Williamson's performance

Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	62	111	10	5117	242*	$\boxed{50.66}$	10161	50.35	17	25	560	12	54	0
ODIs	117	111	10	4678	145*	46.31	5575	83.91	9	32	440	39	48	0
T20Is	42	40	6	1173	73*	34.50	959	122.31	0	7	133	16	20	0
First-class	125	215	17	9600	284*	48.48	18715	51.29	27	48	1124	29	115	0
List A	178	168	18	6799	145*	45.32	8218	82.73	13	44	606	60	75	0
T20s	127	119	12	2930	101*	27.38	2475	118.38	1	16	293	52	54	0

Credit: *www.cricinfo.com*

Psychological factors

- Statistical milestones play a large role in cricket and can impact a player's performance

Kane Williamson’s career record

Kane Williamson

New Zealand

Full name Kane Stuart Williamson

Born August 8, 1990, Tauranga

Current age 27 years 120 days

Major teams New Zealand, Barbados Tridents, Gloucestershire, Gloucestershire 2nd XI, New Zealand Under-19s, Northern Districts, Sunrisers Hyderabad, Yorkshire

Playing role Top-order batsman

Batting style Right-hand bat

Bowling style Right-arm offbreak

Relation Cousin - D Cleaver

insights Explore Kane Williamson's performance

Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	62	111	10	5117	242*	50.66	10161	50.35	17	25	560	12	54	0
ODIs	117	111	10	4678	145*	46.31	5575	83.91	9	32	440	39	48	0
T20Is	42	40	6	1173	73*	34.50	959	122.31	0	7	133	16	20	0
First-class	125	215	17	9600	284*	48.48	18715	51.29	27	48	1124	29	115	0
List A	178	168	18	6799	145*	45.32	8218	82.73	13	44	606	60	75	0
T20s	127	119	12	2930	101*	27.38	2475	118.38	1	16	293	52	54	0

Credit: *www.cricinfo.com*

Psychological factors

- Statistical milestones play a large role in cricket and can impact a player's performance
- Not uncommon to see players bat more cautiously near milestones
- Psychological studies have indicated that player mood can have a significant impact on a cricket player's performance (Totterdell, 1993)

Nervous 90s

Credit: *www.stuff.co.nz*

Aim

- The main aim was to develop models which quantify a player's batting ability at any stage of their innings
  - Should provide a better measure than batting average of how well a player is batting *during* an innings
- Models fitted within a Bayesian framework:
  - Nested sampling
  - C++, Julia & R

The exponential varying-hazard model

Deriving the model likelihood

If $X \in \{0, 1, 2, 3, ...\}$ is the number of runs scored by a batsman:

$$\text{Hazard function} = H(x)$$

$H(x) =$ The probability of getting out on score x, given you made it to score x

Deriving the model likelihood

If $X \in \{0, 1, 2, 3, ...\}$ is the number of runs scored by a batsman:

$$\begin{aligned} \text{Hazard function} &= H(x) \\ &= P(X = x | X \geq x) \end{aligned}$$

$H(x) =$ The probability of getting out on score x, given you made it to score x

Deriving the model likelihood

If $X \in \{0, 1, 2, 3, ...\}$ is the number of runs scored by a batsman:

$$\begin{aligned} \text{Hazard function} &= H(x) \\ &= P(X = x | X \geq x) \\ &= \frac{P(X = x)}{P(X \geq x)} \end{aligned}$$

$H(x) =$ The probability of getting out on score x, given you made it to score x

Deriving the model likelihood

Assuming a functional form for $H(x)$, conditional on some parameters θ, the model likelihood is:

$$L(\theta) = L_O(\theta) \times L_{NO}(\theta)$$

$$L_O(\theta) = \prod_{i=1}^{I-N} \left(H(x_i) \prod_{a=0}^{x_i-1} [1 - H(a)] \right)$$

$$L_{NO}(\theta) = \prod_{i=1}^{N} \left(\prod_{a=0}^{y_i-1} [1 - H(a)] \right)$$

$\{x_i\}$ = set of out scores

$\{y_i\}$ = set of not out scores

I = Total number of innings

N = Total number of not out innings

Parameterising the hazard function

- To reflect our cricketing knowledge of the ‘getting your eye-in’ process, $H(x)$ should be higher for low scores, and lower for high scores
- From a cricketing perspective we often refer to a player’s ability in terms of a batting average

The effective average function

- Instead, we can model the hazard function in terms of an 'effective batting average' or 'effective average function', $\mu(x)$.
- This is the batsman's batting ability on score x, in terms of a batting average and evolves with score as batsmen 'get their eye-in'
- This allows us to think in terms of batting averages, rather than dismissal probabilities
- Relationship between the hazard function and effective average function:

$$H(x) = \frac{1}{\mu(x) + 1}$$

The effective average function

- Therefore, our model and the hazard function depend on the parameterisation of the effective average function, $\mu(x)$
- Reasonable to believe that batsmen begin an innings playing with some initial batting ability, μ_1
- Batting ability increases with number of runs scored, until some peak batting ability, μ_2, is reached
- The speed of the transition between μ_1 and μ_2 can be represented by a parameter, L

$$\mu(x; \mu_1, \mu_2, L) = \mu_2 + (\mu_1 - \mu_2) \exp\left(-\frac{x}{L}\right)$$

The effective average function

Constraints:

- $\mu_1 \leq \mu_2$
- $L \leq \mu_2$

To implement these constraints, we re-parameterise the effective average function, $\mu(x)$:

- $\mu_1 = C\mu_2$
- $L = D\mu_2$

Where C and D are restricted to the interval [0, 1].

The effective average function

$$\mu(x; C, \mu_2, D) = \mu_2 + \mu_2(C-1)\exp\left(-\frac{x}{D\mu_2}\right)$$

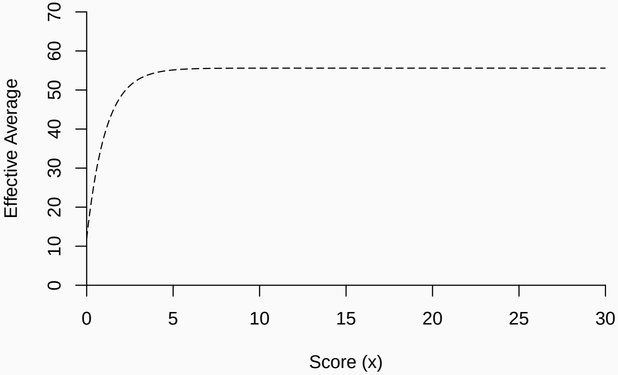


Figure 1: Examples of various plausible effective average functions, $\mu(x)$.

The effective average function

$$\mu(x; C, \mu_2, D) = \mu_2 + \mu_2(C-1)\exp\left(-\frac{x}{D\mu_2}\right)$$

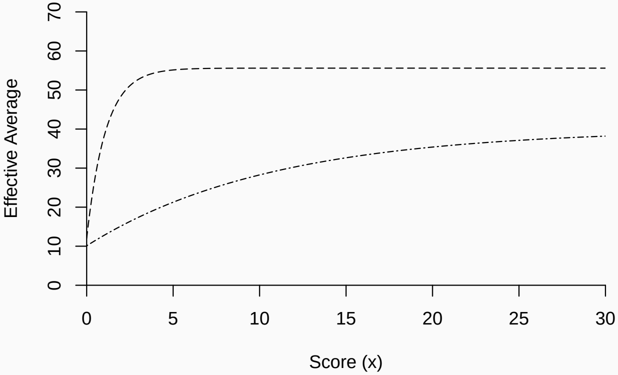


Figure 2: Examples of plausible effective average functions, $\mu(x)$.

The effective average function

$$\mu(x; C, \mu_2, D) = \mu_2 + \mu_2(C-1)\exp\left(-\frac{x}{D\mu_2}\right)$$

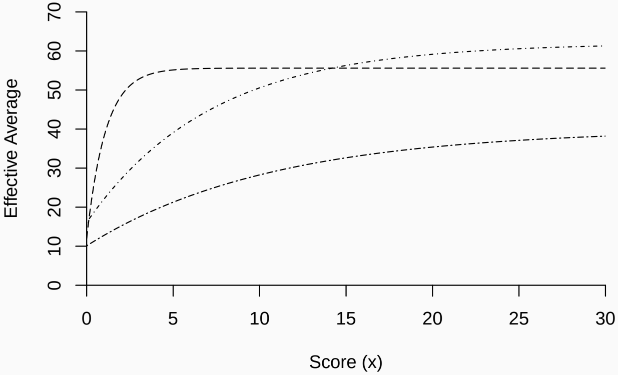


Figure 3: Examples of plausible effective average functions, $\mu(x)$.

The effective average function

$$\mu(x; C, \mu_2, D) = \mu_2 + \mu_2(C-1)\exp\left(-\frac{x}{D\mu_2}\right)$$

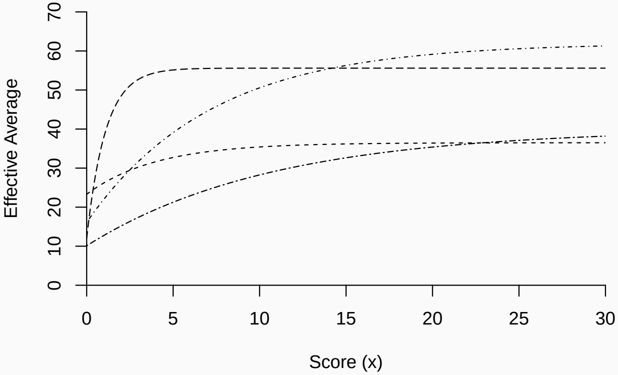


Figure 4: Examples of plausible effective average functions, $\mu(x)$.

The effective average function

$$\mu(x; C, \mu_2, D) = \mu_2 + \mu_2(C-1)\exp\left(-\frac{x}{D\mu_2}\right)$$

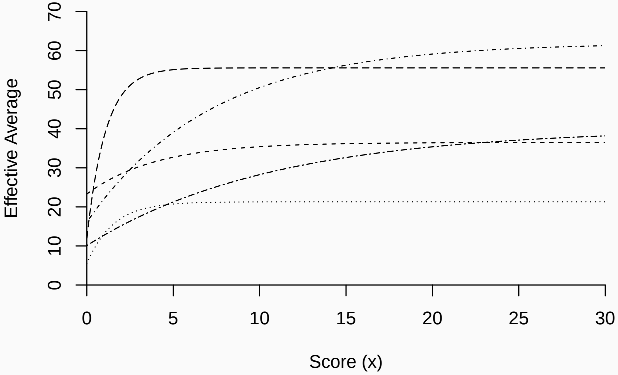


Figure 5: Examples of plausible effective average functions, $\mu(x)$.

Data

Fit the model to player career data:

Runs	Out/not out
13	0
42	0
53	0
104	1
2	0
130	0
2	0
1	0
176	0

- $0 =$ out, $1 =$ not out

Prior specification

Bayesian model specification:

$$\mu_2 \sim \text{Lognormal}(25, 0.75^2)$$
$$C \sim \text{Beta}(1, 2)$$
$$D \sim \text{Beta}(1, 5)$$

- Implemented in C++, using a nested sampling algorithm using Metropolis-Hastings updates

Results: the exponential varying-hazard model

Posterior summaries

Table 1: Parameter estimates and uncertainties for each analysed player using the exponential varying-hazard model. 'Prior' indicates the prior point estimates and uncertainties.

Player	μ_1	μ_2	L	Average
V.Kohli (IND)	$23.6^{+8.5}_{-6.9}$	$62.9^{+10.7}_{-8.1}$	$11.0^{+12.1}_{-7.0}$	53.4
J.Root (ENG)	$24.8^{+8.8}_{-6.9}$	$58.9^{+7.8}_{-6.5}$	$7.2^{+6.3}_{-3.5}$	52.6
K.Williamson (NZL)	$17.6^{+7.3}_{-4.9}$	$59.1^{+8.2}_{-6.9}$	$7.4^{+6.2}_{-3.8}$	50.4
AB de Villiers (SAF)	$25.7^{+8.5}_{-7.0}$	$54.6^{+5.3}_{-4.5}$	$4.7^{+5.2}_{-2.8}$	50.7
S.Al-Hasan (BAN)	$25.9^{+7.1}_{-6.5}$	$44.0^{+6.6}_{-5.1}$	$7.0^{+9.0}_{-4.9}$	40.4
Prior	$6.6^{+12.8}_{-5.0}$	$25.0^{+27.7}_{-13.1}$	$3.0^{+6.7}_{-2.3}$	N/A

Posterior summaries

Table 2: Parameter estimates and uncertainties for each analysed player using the exponential varying-hazard model. 'Prior' indicates the prior point estimates and uncertainties.

Player	μ_1	μ_2	L	Average
V.Kohli (IND)	$23.6^{+8.5}_{-6.9}$	$62.9^{+10.7}_{-8.1}$	$11.0^{+12.1}_{-7.0}$	53.4
J.Root (ENG)	$24.8^{+8.8}_{-6.9}$	$58.9^{+7.8}_{-6.5}$	$7.2^{+6.3}_{-3.5}$	52.6
K.Williamson (NZL)	$17.6^{+7.3}_{-4.9}$	$59.1^{+8.2}_{-6.9}$	$7.4^{+6.2}_{-3.8}$	50.4
AB de Villiers (SAF)	$25.7^{+8.5}_{-7.0}$	$54.6^{+5.3}_{-4.5}$	$4.7^{+5.2}_{-2.8}$	50.7
S.Al-Hasan (BAN)	$25.9^{+7.1}_{-6.5}$	$44.0^{+6.6}_{-5.1}$	$7.0^{+9.0}_{-4.9}$	40.4
Prior	$6.6^{+12.8}_{-5.0}$	$25.0^{+27.7}_{-13.1}$	$3.0^{+6.7}_{-2.3}$	N/A

Predictive hazard functions

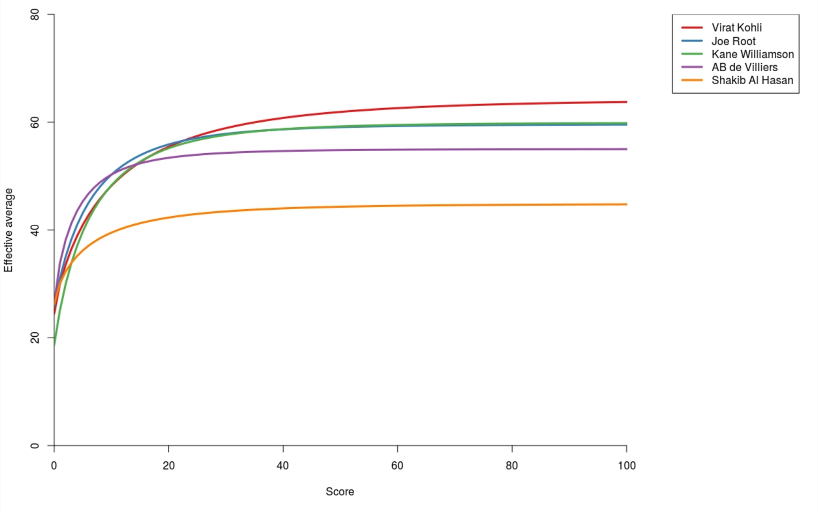


Figure 6: Predictive hazard functions in terms of effective average, $\mu(x)$.

Predictive hazard functions

An interesting comparison can be made between Kane Williamson and AB de Villiers, two top order batsmen with similar career Test batting averages (50.35 vs. 50.66).

De Villiers appears to arrive at the crease batting with greater ability and gets his 'eye-in' quicker, however Williamson appears to be the superior player once familiar with match conditions.

Predictive hazard functions

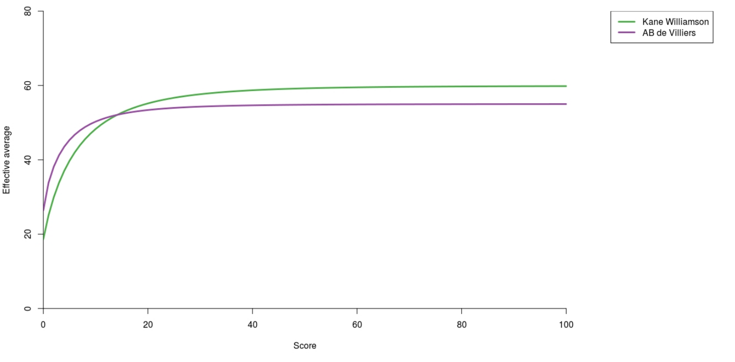


Figure 7: Predictive hazard functions in terms of effective average, $\mu(x)$, for Williamson and de Villiers.

Developing more flexible models

Developing more flexible models

- The exponential varying-hazard model does a reasonable job at identifying batsmen who are particularly capable or vulnerable early in their innings
- Limited to monotonically increasing effective average functions
- Cannot account for scored-based fluctuations in ability (due to nerves/pressure)

Gaussian hazard model

Flexibility is introduced by multipling the effective average function, $\mu(x; C, \mu_2, D)$ from the exponential varying-hazard model, by the exponential of a Gaussian function.

$$g(x; k, \phi, m) = -k \exp\left(\frac{-1}{2\phi^2}(x-m)^2\right)$$

Where $k =$ strength, $\phi =$ width and $m =$ midpoint, of the Gaussian function.

The effective average function

The only change required to implement the Gaussian hazard function is to the effective average function:

$$\mu(x; C, \mu_2, D, k, \phi, m) \quad = \mu(x; C, \mu_2, D) \times \exp(g(x; k, \phi, m))$$

The exponential varying-hazard model

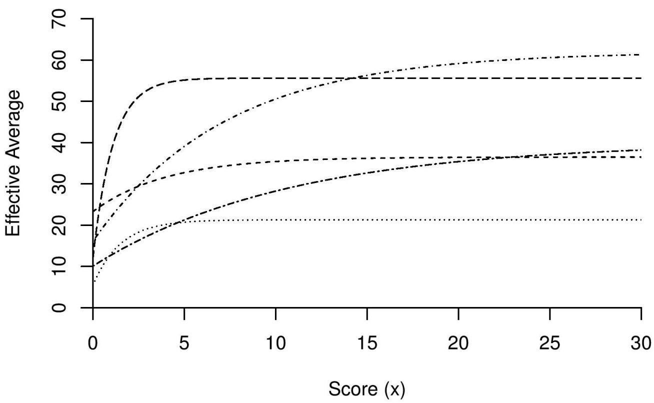


Figure 8: Examples of various plausible effective average functions $\mu(x)$, ranging from small to large differences between the initial and equilibrium effective averages μ_1 and μ_2, with both fast and slow transition timescales L.

Gaussian hazard model

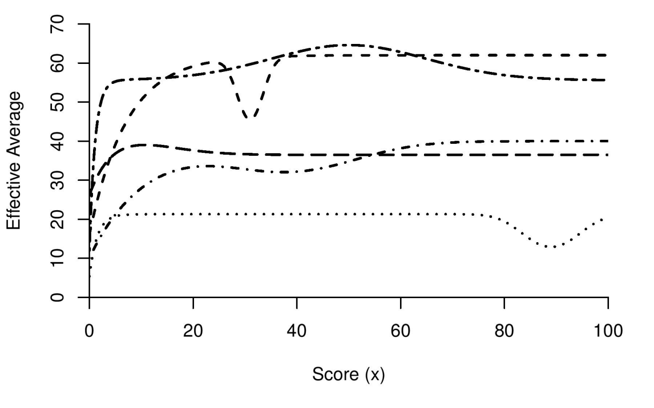


Figure 9: Examples of effective average functions, $\mu(x; C, \mu_2, D, k, \phi, m)$ allowed under the Gaussian hazard model, with varying levels and timings of temporal deviation in batting ability.

Gaussian hazard model

Bayesian model specification:

$$
\begin{aligned}
\mu_2 &\sim \text{Lognormal}(25, 0.75^2) \\
C &\sim \text{Beta}(1, 2) \\
D &\sim \text{Beta}(1, 5) \\
k &\sim \text{Uniform}(-1, 1) \\
\phi &\sim \text{Uniform}(0, 20) \\
m &\sim \text{Uniform}(0, 400)
\end{aligned}
$$

- Implemented in C++, using a nested sampling algorithm using Metropolis-Hastings updates

Predictive hazard functions

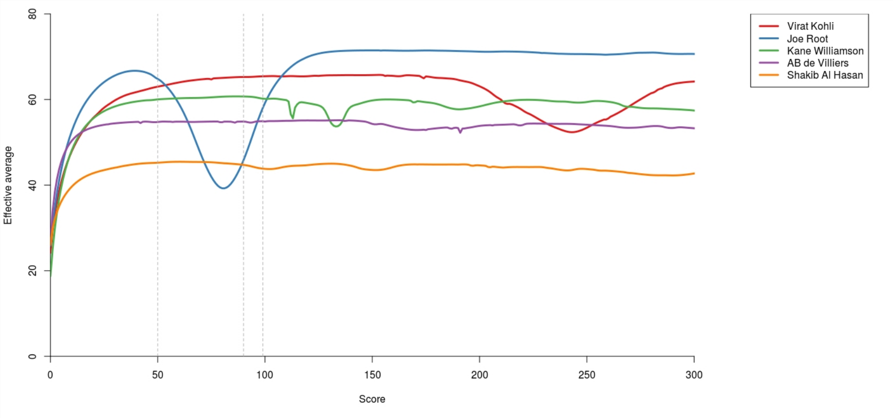


Figure 10: Predictive hazard functions for the Gaussian hazard model in terms of effective average, $\mu(x)$.

Gaussian hazard model

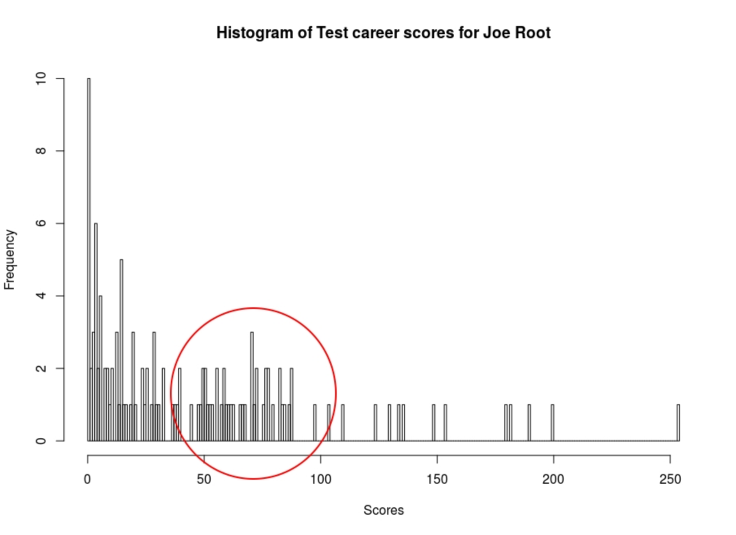


Figure 11: Histogram of Test match career scores for Joe Root.

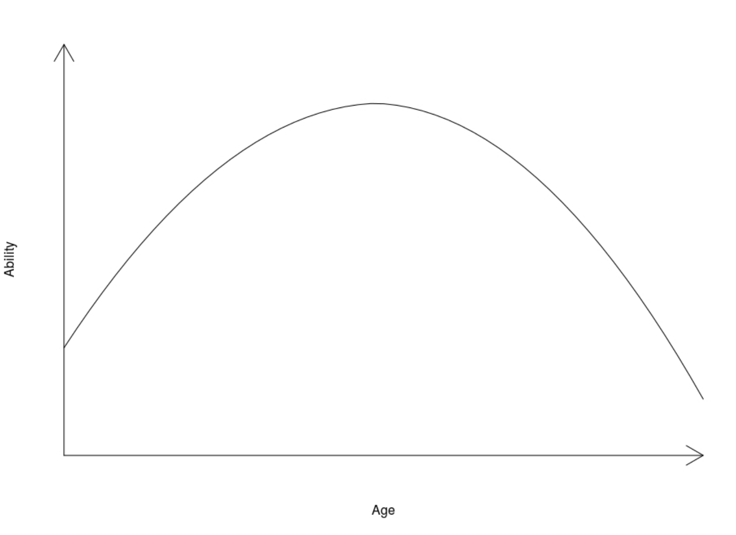
Looking at the bigger picture

Looking at the bigger picture

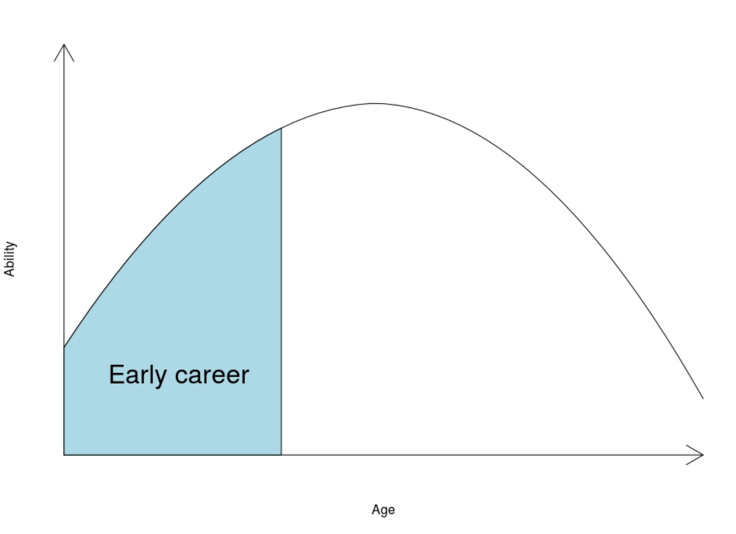
So far the effective average allows us to quantify how the batting abilities of players change *within* an innings, in terms of a batting average.

What about how batting ability changes across a player's career?

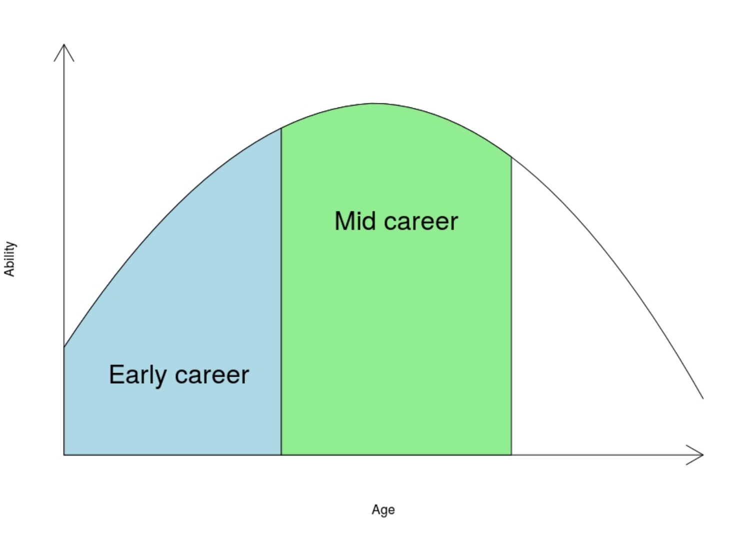
Looking at the bigger picture



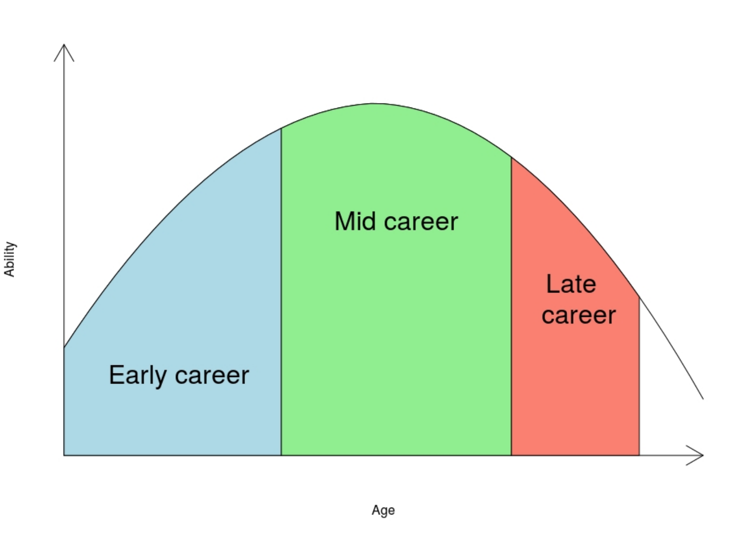
Looking at the bigger picture



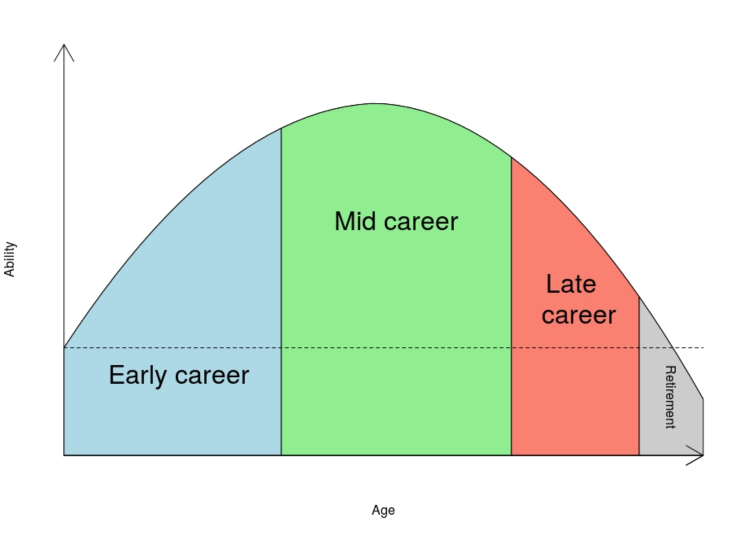
Looking at the bigger picture



Looking at the bigger picture



Looking at the bigger picture



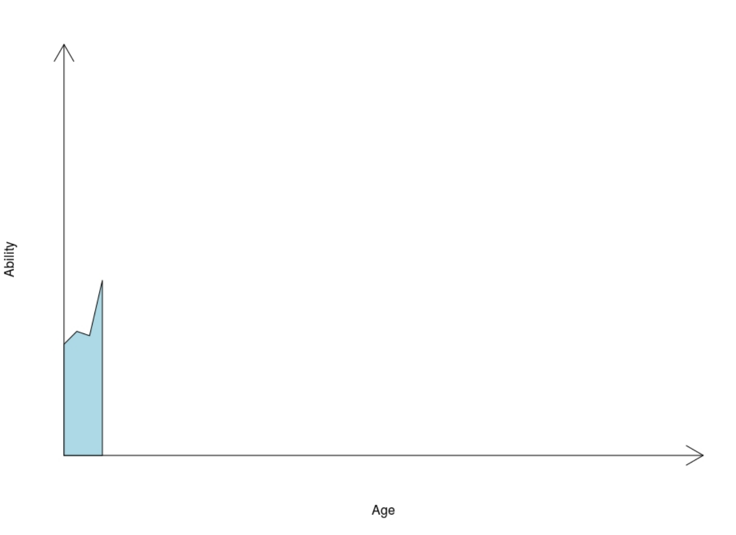
Modelling career trajectories of batsmen in cricket

Modelling batting career trajectories

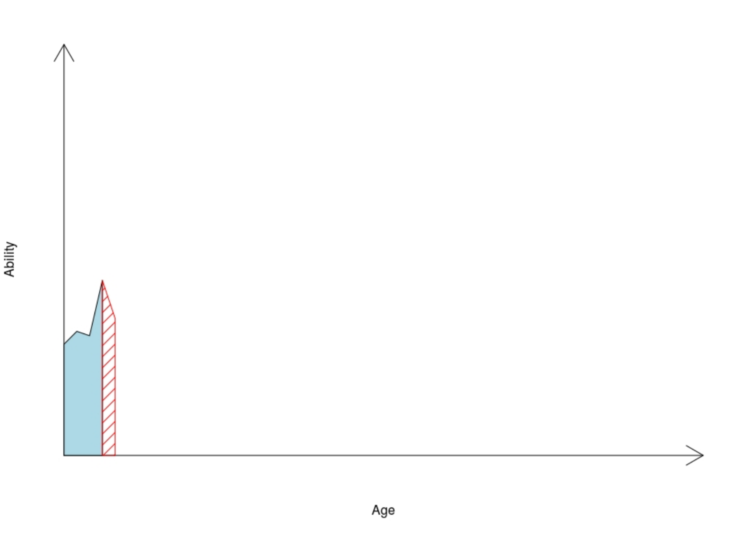
- Due to the nature of the sport, batsmen fail more than they succeed
- Not uncommon to see players get stuck in a rut of poor form over a long period of time
- Coaches more likely to tolerate numerous poor performances in a row than in other sports
- Interestingly, players frequently string numerous strong performances together
- Suggests external factors such as a player's current form and fitness levels are important variables to consider
- Due to the 'random' element of these external factors, players may exhibit multiple peaks in ability during a long career

Typical sporting career trajectories

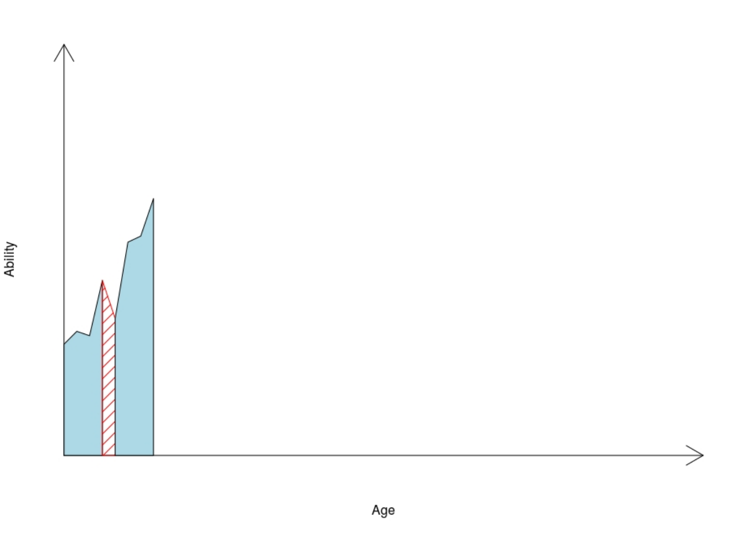
Typical sporting career trajectories



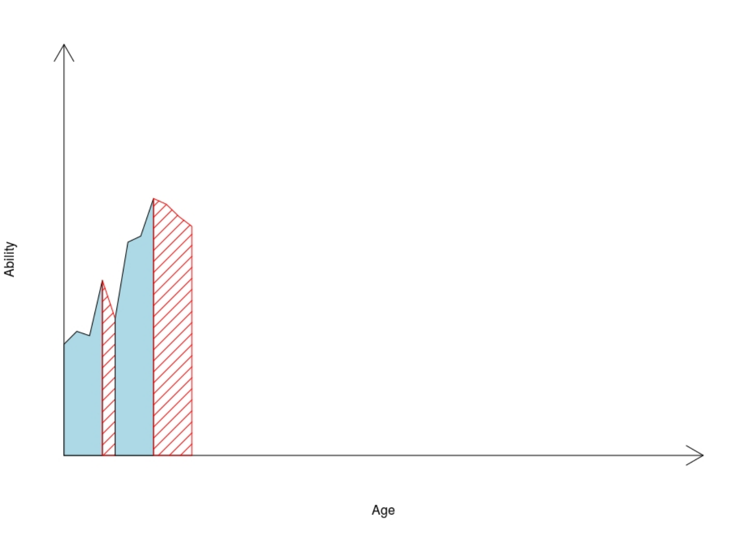
Typical sporting career trajectories



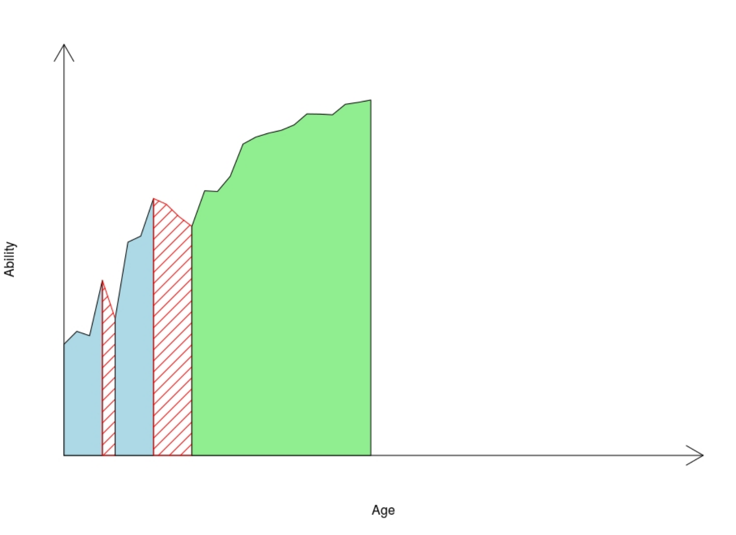
Typical sporting career trajectories



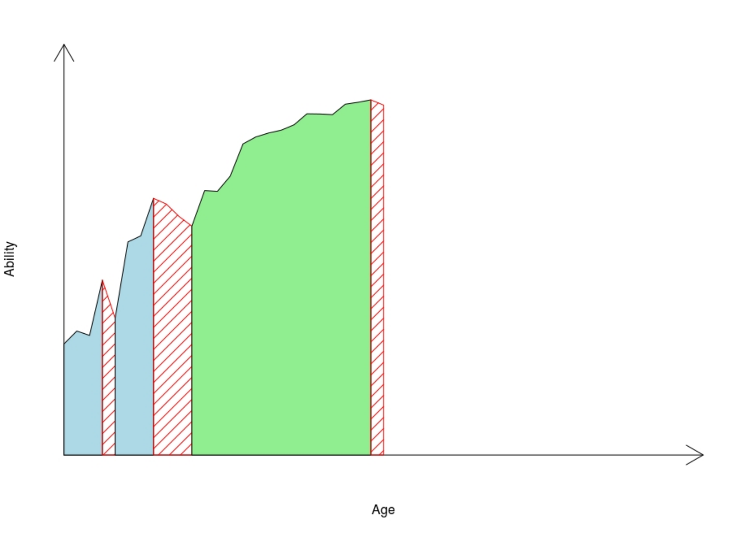
Typical sporting career trajectories



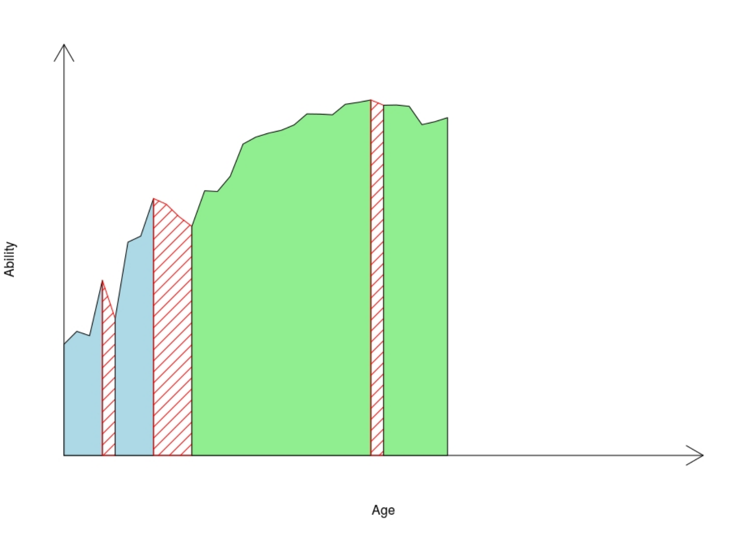
Typical sporting career trajectories



Typical sporting career trajectories

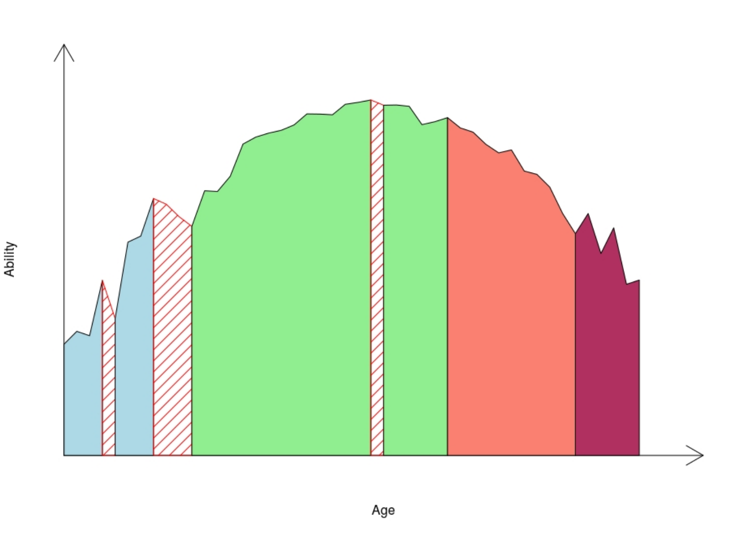


Typical sporting career trajectories

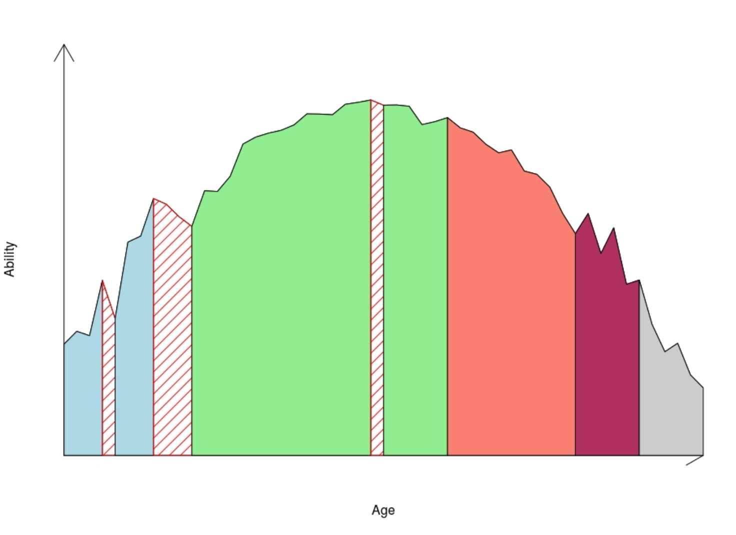


Typical sporting career trajectories

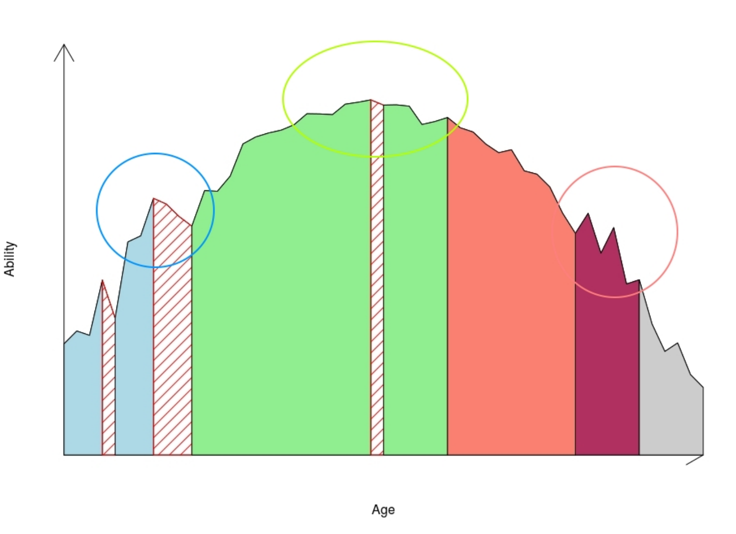
Typical sporting career trajectories



Typical sporting career trajectories



Typical sporting career trajectories



Modelling batting career trajectories

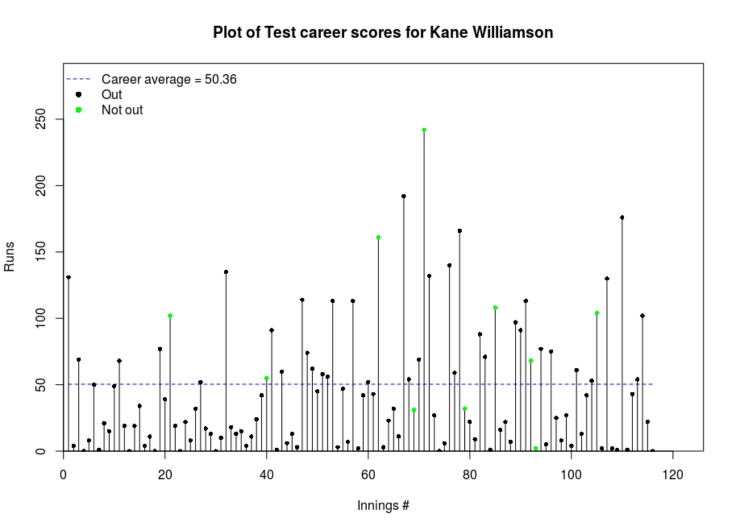


Figure 12: Plot of Test career scores for Kane Williamson.

Modelling batting career trajectories

Our aim is to build a model which can measure and predict player batting ability at any given stage of career.

Needs to be able to handle random fluctuations in performance due factors such as:

- Player form
- Player fitness (both mental and physical)
- Random chance!

Gaussian processes

Gaussian processes are a class of schotastic process, made up of a collection of random variables, such that every finite collection of those random variables has a multivariate normal distribution (Rasmussen & Williams, 2006).

A Gaussian process is completely specified by its:

- Mean function, $m(x)$
- Covariance function, $K(x, x)$

Covariance functions

There are a number of covarince functions available to choose from. A common choice is the *squared exponential covariance function*.

$$K(X_i, X_j) = \sigma^2 \text{exp}(\tfrac{-(X_i - X_j)^2}{2l^2}) + n_{ij}$$

$\sigma =$ 'signal variance', determines how much a function value can deviate from the mean
$l =$ 'length-scale', roughly the distance required to move in the input space before the function value can change significantly
$n =$ 'noise variance', used by the Gaussian process model to allow for any noise present in the i observations. This term is only included when $i = j$

Example: Gaussian processes

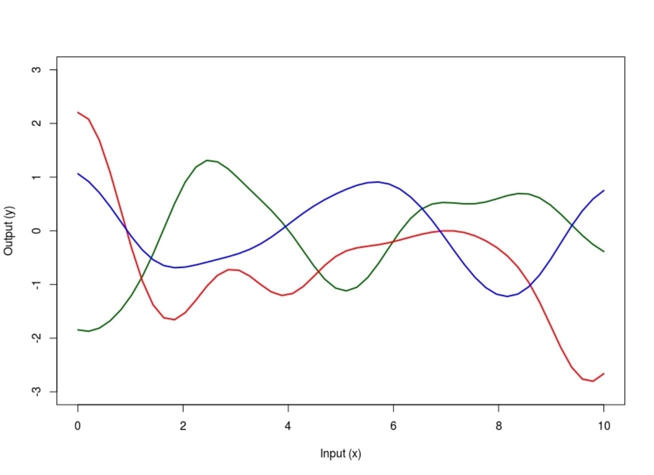


Figure 13: Gaussian processes drawn from a null distribution (i.e. uninformed by any data), with a mean value of 0, and varying values for σ and l.

Example: Gaussian processes

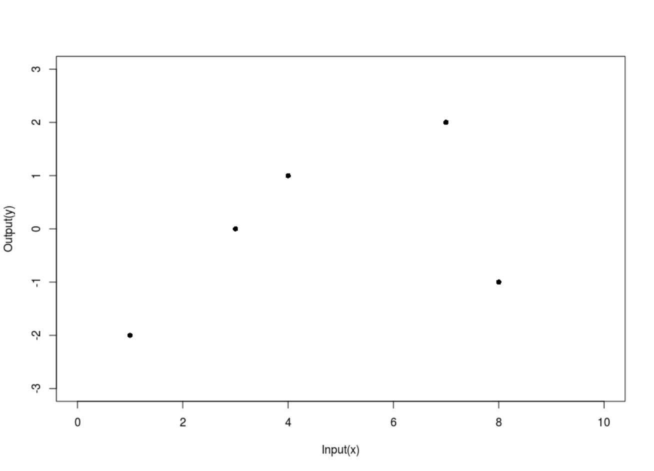


Figure 14: Some observed data in the input/output space.

Example: Gaussian processes

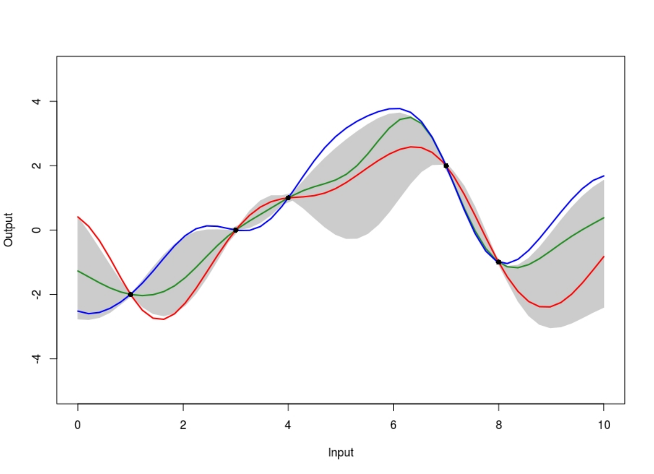


Figure 15: Example Gaussian processes fitted to some noiseless data. Shaded area represents a 95% confidence interval.

Example: Gaussian processes

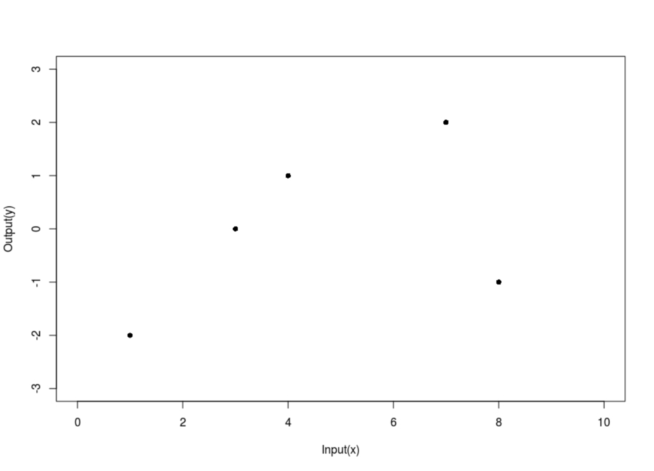


Figure 16: Some observed data in the input/output space.

Example: Gaussian processes

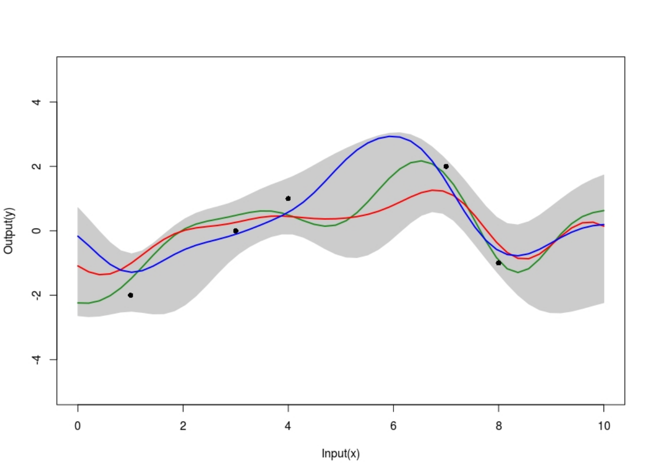


Figure 17: Example Gaussian processes fitted to some noisy data. Shaded area represents a 95% confidence interval.

Modelling batting career trajectories

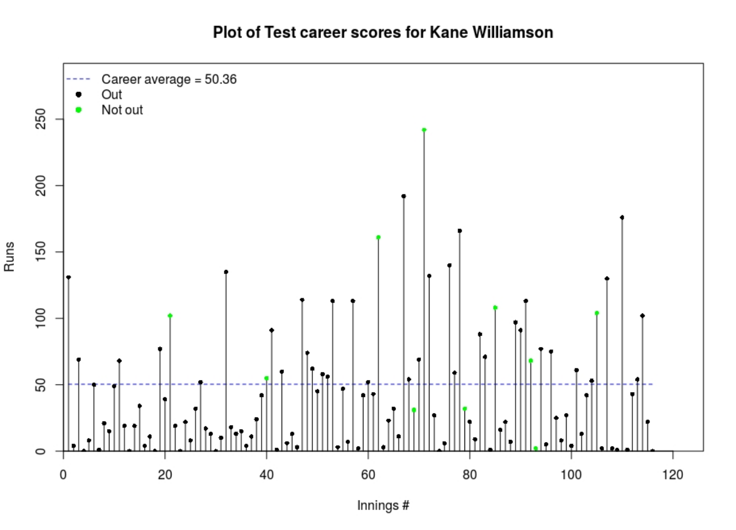


Figure 18: Plot of Test career scores for Kane Williamson.

Modelling batting career trajectories

Recall the effective average function, $\mu(x)$:

$$\mu(x; C, \mu_2, D) = \text{Player batting ability on score } x$$

- $\mu_2 =$ 'peak' batting ability *within* an innings

If we re-define $\mu(x)$, to $\mu(x, i)$:

$$\mu(x, i; C, \mu_2, D) = \text{Player batting ability on score } x\text{, in } i^{th} \text{ career innings}$$

- $\mu_{2_i} =$ 'peak' batting ability within batsman's i^{th} career innings

Constructing the Gaussian process

Now, instead of estimating the posterior distribution for μ_2, we must estimate posterior distributions for each of the μ_{2_i} terms, one for each innings the player has batted in.

This is achieved by introducing a set of noise terms, $\{n_i\}$ in the model, which are used to construct the Gaussian process for μ_2.

To ensure positivity in our estimates for μ_{2_i}, we model $\log(\mu_2)$ as a Gaussian process and back-transform accordingly.

Prior specficiation

Bayesian model specification:

$$\log(\mu_{2_i}) \sim \text{GP}(m,\ K(X_i, X_j))$$
$$\{n_i\} \sim \text{Normal}(0, 1)$$
$$C \sim \text{Beta}(1, 2)$$
$$D \sim \text{Beta}(1, 5)$$
$$m \sim \text{Lognormal}(25, 0.75^2)$$
$$\sigma \sim \text{Exponential}(\mathrm{mean} = 0.1)$$
$$l \sim \text{Uniform}(0, 100)$$

Calculating the predictive hazard function

The model output provides us with posterior distributions for the set of $\{n_i\}$, noise terms. Some clever matrix algebra (Rasmussen & Williams, 2006), allows us to use these terms to construct posterior predictive functions for μ_2 across a career.

However, we aren't interested in μ_2, at each innings, rather the innings-specific effective average, $\mu(i)$:

$$\begin{aligned}\mu(i) &= \text{expected number of runs scored in } i^{th}\text{innings} \\ &= \text{expected } \underline{\text{batting average}} \text{ in } i^{th}\text{innings}\end{aligned}$$

Which we can compute analytically.

Predictive hazard function

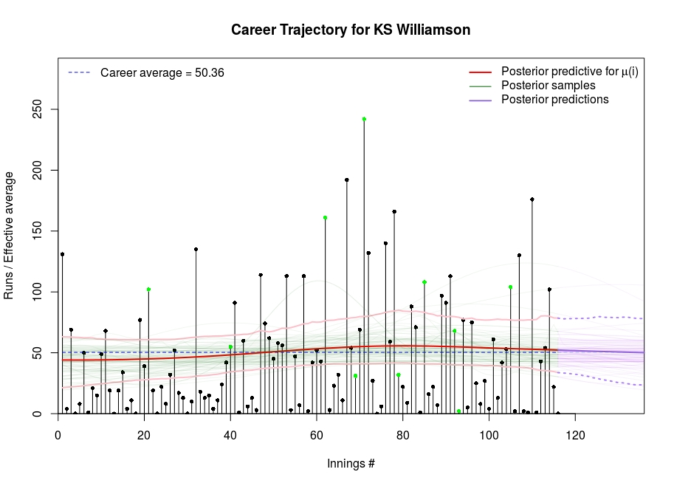


Figure 19: Predictive hazard function for $\mu(i)$, in terms of effective average, with 95% credible intervals.

Predictive hazard function

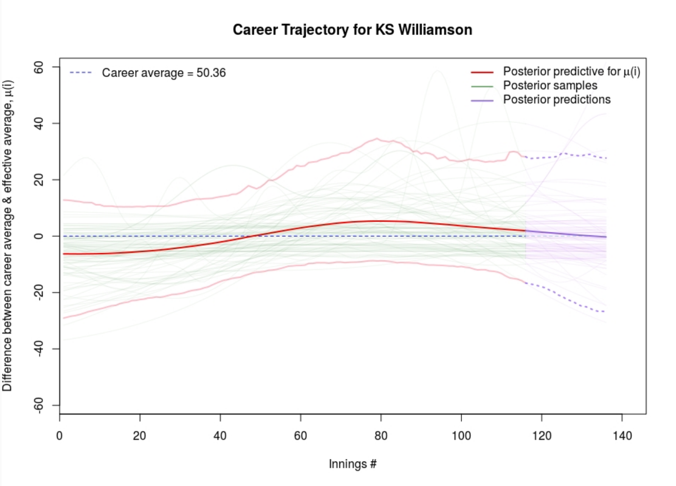


Figure 20: Difference between career average and predictive hazard function for $\mu(i)$, in terms of effective average.

Predictive hazard functions

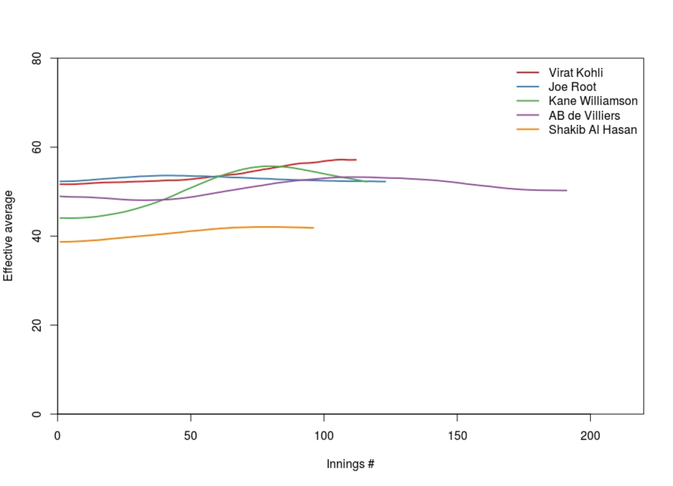


Figure 21: Predictive hazard functions for $\mu(i)$, in terms of effective average.

Predictive hazard functions

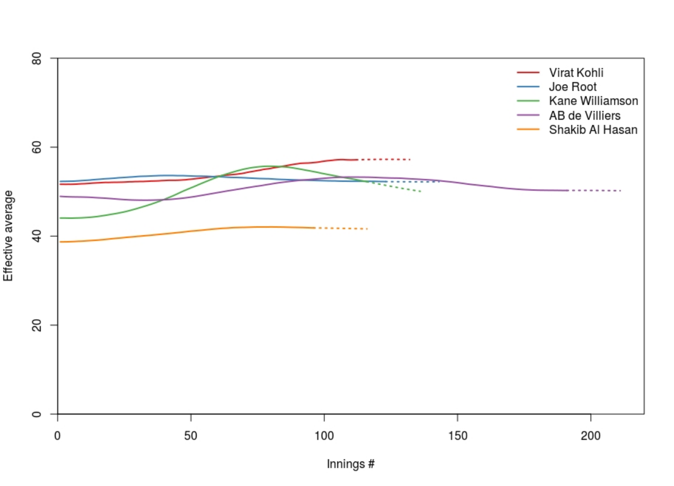


Figure 22: Predictive hazard functions for $\mu(i)$, in terms of effective average. Dotted lines are predictions for the next 20 innings.

Predictive hazard functions

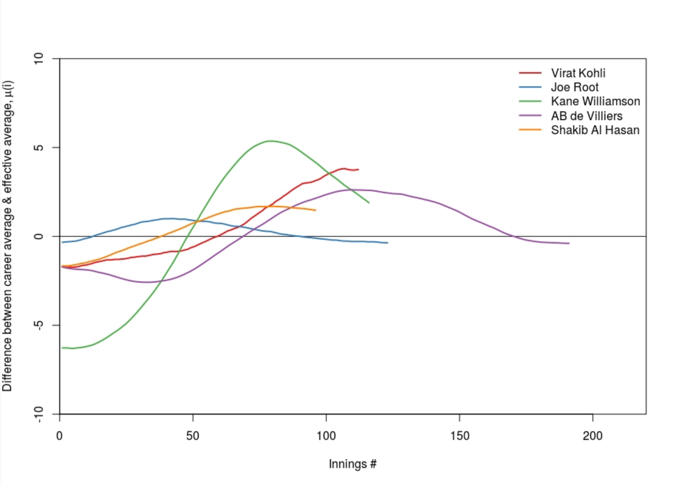


Figure 23: Difference between career averages and predictive hazard functions for $\mu(i)$, in terms of effective average.

Predictive hazard functions

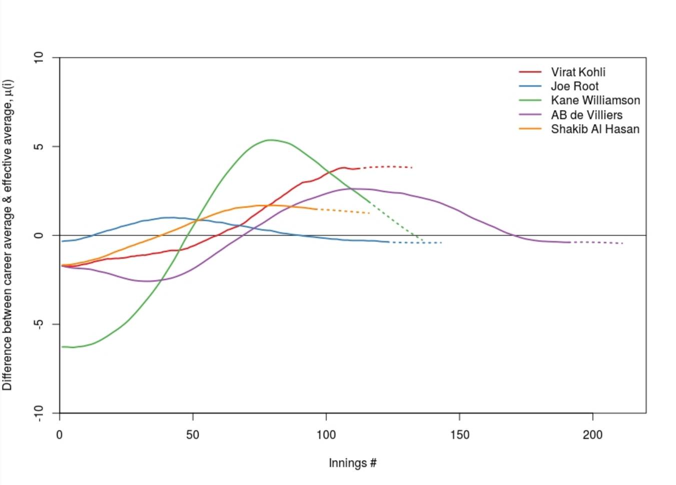


Figure 24: Difference between career averages and predictive hazard functions for $\mu(i)$, in terms of effective average. Dotted lines are predictions for the next 20 innings.

Concluding statements, limitations and further work

Limitations and future work

- Models ignore variables such as balls faced and minutes batted
- Historic data such as pitch and weather conditions difficult to obtain
- Haven't accounted for the likes of opposition bowler ability
- Models assume player ability isn't influenced by the match scenario
    - Limits usage to Test/First Class matches, possibly One Dayers

Concluding statements

- There has been a recent boom in statistical analysis in cricket, particularly around T20 cricket
- However, many analyses stray away from maintaining an easy to understand, cricketing interpretation
- We have developed tools which allow us to quantify player batting ability both within *and* between innings
  - ~~Batting average~~
  - Effective average ✓

Effective average visualisations

Stevenson & Brewer (2017)

www.oliverstevenson.co.nz

Cricket Visualisations | Documentation

Cricket Visualisations

Select countries:

- ☑ Australia
- ☑ Bangladesh
- ☑ England
- ☑ India
- ☑ New Zealand
- ☑ Pakistan
- ☑ South Africa
- ☑ Sri Lanka
- ☑ West Indies
- ☑ Zimbabwe

Select player(s):

- Ricky Ponting (AUS)
- Kevin Pietersen (ENG)
- Sachin Tendulkar (IND)
- Stephen Fleming (NZL)
- Mark Richardson (NZL)
- Brian Lara (WI)

Plotting options:

Select range of scores:

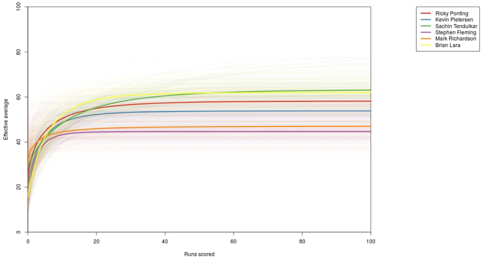
0 100 500

0 50 100 150 200 250 300 350 400 450 500

Number of samples to plot (max 1000):

100

Clear plot



Player	μ_1	μ_2	L
Ricky Ponting	22.22	57.48	5.77
Kevin Pietersen	19.90	53.41	4.65
Sachin Tendulkar	28.74	62.63	13.00
Stephen Fleming	11.87	44.44	2.39
Mark Richardson	30.75	46.36	3.63
Brian Lara	15.10	61.45	6.09

Thanks

THE UNIVERSITY OF AUCKLAND NEW ZEALAND

NZC

References

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Stevenson, O. G., & Brewer, B. J. (2017). Bayesian survival analysis of batsmen in test cricket. *Journal of Quantitative Analysis in Sports*, *13*(1), 25–36.

Questions?