Getting Your Eye In: A Bayesian Analysis of Early Dismissals in Cricket

Brendon James Brewer
School of Mathematics and Statistics
The University of New South Wales

brendon.brewer@unsw.edu.au

February 17, 2013

Abstract

A Bayesian Survival Analysis method is motivated and developed for analysing sequences of scores made by a batsman in test or first class cricket. In particular, we expect the presence of an effect whereby the distribution of scores has more probability near zero than a geometric distribution, due to the fact that batting is more difficult when the batsman is new at the crease. A Metropolis-Hastings algorithm is found to be efficient at estimating the proposed parameters, allowing us to quantify exactly how large this early-innings effect is, and how long a batsman needs to be at the crease in order to "get their eye in". Applying this model to several modern players shows that a batsman is typically only playing at about half of their potential ability when they first arrive at the crease, and gets their eye in surprisingly quickly. Additionally, some players are more "robust" (have a smaller early-innings effect) than others, which may have implications for selection policy.

1 Introduction

It is well known to cricketers of all skill levels that the longer a batsman is in for, the easier batting tends to become. This is probably due to a large number of psychological and technique-related effects: for example, it is generally agreed that it takes a while for a batsman's footwork to "warm up" and for them to adapt to the subtleties of the prevailing conditions and the bowling attack. Consequently, it is frequently observed that players are far more likely to be dismissed early in their innings than is predicted by a constant-hazard model, where the probability of getting out on your current score (called the *Hazard*) is exactly the same regardless of your current score. Note that a constant hazard model leads to an exponential probability distribution over the non-negative integers (i.e. the geometric distribution) as describing our prediction of a batsman's score. The aim of this paper is to develop a Bayesian method (O'Hagan & Forster, 2004) for inferring how a player's Hazard varies throughout an innings, thus giving quantitative answers to the questions "how long do we have to wait until batsmen get their eye in, and how much better do they really become?". This question has been addressed previously using nonparametric frequentist survival analysis (Kimber and Hansford, 1993; Cai, Hyndman and Wand, 2002). However, using a nonparametric approach in a Bayesian setting would give the hazard function far too much freedom and would lead to very poorly constrained inferences of the hazard function if applied to individual players. To simplify matters, this paper uses a parametric model, which is effectively a single change-point model with a smooth transition rather than a sudden jump.

2 Sampling Distribution

Consider predicting the score $X \in \{0, 1, 2, 3, ...\}$ that a batsman will make in a single innings. We will now assign a probability distribution for X, conditional on some parameters. Define a hazard function $H(x) \in [0, 1]$ as the probability of being dismissed on score x (i.e. $P(X = x)$) given that the batsman is currently on score x (i.e. given $X \geq x$):

$$H(x) = P(X = x|X \geq x) = \frac{P(X = x, X \geq x)}{P(X \geq x)} = \frac{P(X = x)}{P(X \geq x)} \tag{1}$$

Define a backwards cumulative distribution function by:

$$G(x) = P(X \geq x) \tag{2}$$

Using $G(x)$ rather than the conventional cumulative distribution $F(x) = P(X \leq x)$ simplifies some expressions in this case, and also helps because $G(x)$ will also serve as the likelihood function for the "not-outs", or uncompleted innings. With this definition, Equation 1 becomes, after some rearrangement, a difference equation for G:

$$G(x + 1) = [1 - H(x)]\, G(x) \tag{3}$$

With the initial condition $G(0) = 1$, this can be solved, giving:

$$G(x) = \prod_{a=0}^{x-1} [1 - H(a)] \tag{4}$$

This is the product of the probabilities of surviving to score 1 run, times the probability of reaching a score of 2 runs given that you scored 1, etc, up to the probability of surviving to score x given that you scored $x - 1$. Thus, the probability distribution for X is given by the probability of surviving up to a score x and then being dismissed:

$$P(X = x) = H(x) \prod_{a=0}^{x-1} [1 - H(a)] \tag{5}$$

This is all conditional on a choice of the hazard function H, which we will parameterise by parameters θ. Assuming independence (this is not a physical assertion, rather, an acknowledgement that we are not interested in any time dependent effects for now), the probability distribution for a set of scores $\{x_i\}_{i=1}^{I-N}$ (I and N are the number of innings and not-outs respectively) and a set of not-out scores $\{y_i\}_{i=1}^{N}$ is:

$$p(\mathbf{x}, \mathbf{y}|\theta) = \prod_{i=1}^{I-N} \left(H(x_i; \theta) \prod_{a=0}^{x_i-1} [1 - H(a; \theta)] \right) \times \prod_{i=1}^{N} \left(\prod_{a=0}^{y_i-1} [1 - H(a; \theta)] \right) \tag{6}$$

When the data $\{\mathbf{x}, \mathbf{y}\}$ are fixed and known, Equation 6 gives the likelihood for any proposed model of the Hazard function - that is, for any value of θ. The log likelihood is:

$$\log p(\mathbf{x}, \mathbf{y}|\theta) = \sum_{i=1}^{I-N} \log H(x_i; \theta) + \sum_{i=1}^{I-N} \sum_{a=0}^{x_i-1} \log [1 - H(a; \theta)] + \sum_{i=1}^{N} \sum_{a=0}^{y_i-1} \log [1 - H(a; \theta)] \tag{7}$$

3 Parameterisation of the Hazard Function

Rather than seek clever parameterisations of $H(x; \theta)$ and priors over θ that are conjugate to the likelihood, we will take the simpler approach of simply defining a model and prior, and doing the inference with a Metropolis-Hastings sampler (C++ source code and data files for carrying this out will be provided by the author on request). To capture the phenomenon of "getting your eye in", the Hazard function will need to be high for low x and decrease to a constant value as x increases and the batsman becomes more comfortable. Note that if $H(x) = h$, a constant value,

the sampling distribution becomes a geometric distribution with expectation $\mu = 1/h - 1$. This suggests modelling the Hazard function in terms of an "effective batting average" that varies with time, which is helpful because it is easier to think of playing ability in terms of batting averages than dismissal probabilities. $H(x)$ is obtained from $\mu(x)$ as follows:

$$H(x) = \frac{1}{\mu(x) + 1} \tag{8}$$

A simple change-point model for $\mu(x)$ would be to have $\mu(x) = \mu_1 + (\mu_2 - \mu_1)\text{Heaviside}(x - \tau)$, where τ is the change-point. However, a more realistic model would have μ changing smoothly from one value to the other. Replacing the Heaviside step function with a logistic sigmoid function of the form $1/(1 + e^{-t})$ gives the following model, which will be adopted throughout this paper:

$$\mu(x) = \mu_1 + \frac{\mu_2 - \mu_1}{1 + \exp\left(-(x - \tau)/L\right)} \tag{9}$$

Hence

$$H(x) = \left[1 + \mu_1 + \frac{\mu_2 - \mu_1}{1 + \exp\left(-(x - \tau)/L\right)}\right]^{-1} \tag{10}$$

This has four parameters: μ_1 and μ_2, the two effective abilities of the player, τ, the midpoint of the transition between them, and L, which describes how abrupt the transition is. As $L \to 0$ this model resembles the simpler change-point model. A few examples of the kind of hazard models that can be described by varying these parameters are shown in Figure 1. It is possible (although we don't really expect it) for the risk of being dismissed to increase as your score increases; more commonly it will decrease. Slow or abrupt transitions are possible and correspond to different values of L.

4 Prior Distribution

Now we must assign a prior probability distribution of the space of possible values for the parameters (μ_1, μ_2, τ, L). All of these parameters are non-negative and can take any positive real value. It is possible to take into account prior correlations between μ_1 and μ_2 (describing the expection that a player who is excellent when set is also more likely to be good just after arriving at the crease, and that μ_2 is probably greater than μ_1). However, this will almost certainly be supported by the data anyway. Hence, for simplicity we assigned the more conservative independent Normal$(30, 20^2)$ priors[1], truncated to disallow negative values:

$$p(\mu_1, \mu_2) \propto \exp\left[-\frac{1}{2}\left(\frac{\mu_1 - 30}{20}\right)^2 - \frac{1}{2}\left(\frac{\mu_2 - 30}{20}\right)^2\right] \qquad \mu_1,\, \mu_2 > 0 \tag{11}$$

The joint prior for L and τ is chosen to be independent of the μ's and also independent of each other. A typical player can expect to become accustomed to the batting conditions after ~ 20 runs. An exponential prior for τ with mean 20 and an exponential prior with mean 3 to L were found to produce a range of plausible hazard functions:

$$p(\tau, L) \propto \exp\left(-\frac{\tau}{20} - \frac{L}{3}\right) \qquad \tau,\, L > 0 \tag{12}$$

Some hazard functions sampled from the prior are displayed in Figure 1. The posterior distribution for μ_1, μ_2, τ and L is proportional to `prior` $\times$ `likelihood`, i.e. the product of the right hand sides of Equations 6, 11 and 12. Qualitatively, the effect of Bayes' theorem is to take the set of possible hazard functions and their probabilities (Figure 1) and reweight the probabilities according to how well each hazard function fits the observed data.

[1] A reasonable first-order description of the range of expected variation in batting abilities and hence our state of knowledge about a player whose identity is unspecified - we intend the algorithm to apply to any player. It is possible to parameterise this prior with unknown hyperparameters and infer them from the career data of many players, yielding information about the cricket population as a whole. However, such a calculation is beyond the scope of this paper.

5 The Data

Data were obtained from the StatsGuru utility on the Cricinfo website (`http://www.cricinfo.com/`) for the following players: Brian Lara, Chris Cairns, Nasser Hussain, Gary Kirsten, Justin Langer, Shaun Pollock, Steve Waugh and Shane Warne. These players were chosen arbitrarily but subject to the condition of having recently completed long careers. A selection of batsmen, quality all-rounders and bowlers was chosen. The MCMC was run for a large number of steps - mixing is quite rapid because the likelihood evaluation is fast and the parameter space is only 4-dimensional. For brevity, we will display posterior distributions for Brian Lara only. For the other players, summaries such as the posterior means and standard deviations will be displayed instead.

6 Results

6.1 Marginal Posterior Distributions

In this section we will focus on the posterior distributions of the parameters (μ_1, μ_2, τ, L) for Brian Lara. The marginal distributions (approximated by samples from the MCMC simulation) are plotted in Figure 2. These results imply that when Lara is new to the crease, he bats like a player with an average of ~ 15, until he has scored ~ 5 runs. After a transition period with a scale length of ~ 2 runs, (although the form of the logistic functions shows that a transition is more gradual than indicated by L), he then begins to bat as though he has an average of ~ 60. In this case, the analysis has confirmed the folklore about Brian Lara - that if you don't get him out early, you can never really tell when he might get a huge score. "Form" doesn't really come into it. The only surprise to emerge from this analysis is the low value of the change-point τ - Lara is halfway through the process of getting his eye in after scoring only about 5 runs. However, there is still a reasonable amount of uncertainty about the parameters, even though Brian Lara's long test career consisted of 232 innings. The posterior distribution for Brian Lara's parameters does not contain strong correlations (the maximum absolute value in the correlations matrix is 0.4). This is also true of the posterior distributions for the other players. Hence, in the next section, summaries of the marginal distributions for the four parameters will be presented for each player.

6.2 Summaries

The estimates and uncertainties (posterior mean $\pm$ standard deviation) for the four parameters are presented in Table 1. Figure 3 also shows graphically where each of the eight players is estimated to lie on the μ_1-μ_2 plane. One interesting result that is evident from this analysis is that it is not just the gritty specialist batsmen that are robust in the sense that μ_1 is quite high compared to μ_2. The two aggressive allrounders Shaun Pollock and Chris Cairns also show this trait, and are even more robust than, for example, Justin Langer and Gary Kirsten. It is possible that the technique or mindset shown by these players is one that does not require much warming up, or that it is more difficult to get your eye in at the top of the order than in the middle/lower order - although this ought to be a very tentative conclusion given that it is based on only two examples.

The estimated value of μ_1 for Steve Waugh is lower than all other players in the sample apart from Shane Warne. Even Shaun Pollock appears to be a better batsman than Steve Waugh at the beginning of his innings. The plausibility of this statement can be measured by asking the question "what is the posterior probability that $H_{Pollock}(0) < H_{Waugh}(0)$?". From the MCMC output, this probability was found to be 0.92.

The marginal likelihood or "evidence" for this entire model and choice of priors can be measured effectively using annealed importance sampling, or AIS (Neal, 1998; Jarzynski, 1997a,b). AIS is a very generally applicable MCMC-based algorithm that produces an unbiased estimate of $Z = \int \texttt{prior}(\theta) \times \texttt{likelihood}(\theta)\, d\theta$. Z is the probability of the data that were actually observed, under the model, averaged over all possible parameter values (weighted according to the prior). It is the crucial quantity for updating a state of knowledge about which of two distinct models is correct (MacKay, 2003; O'Hagan & Forster, 2004). To test our model for the hazard function, we computed the evidence value for each player for the varying-hazard model and also for a constant hazard model, with a truncated $N(30, 20^2)$ prior on the constant effective average. The logarithm

Table 1: Parameter estimates for the players studied in this paper. The right hand column, the logarithm of the Bayes Factor, shows that the data support the varying hazard model over a constant hazard model by a large factor in all cases. The smallest Bayes Factor is still over 2500 to 1 in favour of the varying Hazard Model. Thus, the varying hazard model would likely still be significantly favoured even if the priors for the parameters were slightly modified.

Player	μ_1	μ_2	τ	L	$\log_e(Z)$	$\log_e(Z/Z_0)$
Cairns	26.9 ± 9.2	36.7 ± 5.5	14.5 ± 17.7	3.1 ± 3.0	-444.11	8.82
Hussain	15.6 ± 9.1	42.1 ± 4.4	5.2 ± 7.1	2.2 ± 1.0	-707.16	12.28
Kirsten	16.6 ± 9.3	54.1 ± 5.7	7.3 ± 5.5	2.9 ± 2.4	-757.16	16.94
Langer	24.3 ± 11.5	49.6 ± 4.9	8.9 ± 14.3	2.8 ± 2.9	-810.34	11.66
Lara	14.5 ± 8.3	60.2 ± 4.7	5.1 ± 2.9	2.8 ± 1.8	-1105.65	21.62
Pollock	22.1 ± 7.7	38.9 ± 5.4	9.7 ± 9.3	3.1 ± 2.9	-519.39	7.87
Warne	3.5 ± 2.0	21.1 ± 2.0	1.1 ± 0.6	0.5 ± 0.4	-686.59	22.54
Waugh	10.5 ± 5.5	57.3 ± 4.4	1.8 ± 1.6	0.8 ± 1.2	-1030.69	25.29
Prior	32.8 ± 17.6	32.8 ± 17.6	20.0 ± 20.0	3.0 ± 3.0	N/A	N/A

of the Bayes Factor (evidence ratio) describing how well the data support the varying-hazard model over constant hazard is shown in the right-hand column of Table 1. Since these were computed using a Monte-Carlo procedure, they are not exact, but the AIS simulations were run for long enough so that the standard error in the Bayes Factor for each player was less than 5% of its value. The data decisively favour the varying-hazard model in all cases, and this would be expected to persist under slight changes to the hazard function parameterisation or the prior distributions.

6.3 Predictive Hazard Function

In the usual way (O'Hagan & Forster, 2004), a predictive distribution for the next data point (score in the next innings) can be found by averaging the sampling distribution (Equation 6) over all possible values of the parameters that are allowed by the posterior. Of course, all of these players have retired, so this prediction is simply a conceptual device to get a single distribution over scores, and hence a single estimated hazard function via Equation 1.

This predictive hazard function is plotted (in terms of the effective average) for three players (Brian Lara, Justin Langer and Steve Waugh) in Figure 4. The latter two are noted for their grit, whilst Brian Lara is considered an aggressive batsman. These different styles may translate to noticable differences in their predictive hazard functions. It is clear from Figure 4 that Justin Langer is more "robust" than Brian Lara, as the difference between his abilities when fresh and when set is smaller ($P\left(\left(\frac{\mu_2}{\mu_1}\right)_{Lara} > \left(\frac{\mu_2}{\mu_1}\right)_{Langer}\right) = 0.80$). This is probably a good trait for an opening batsman. On the other hand, Steve Waugh is actually significantly worse when he is new to the crease than Brian Lara ($P(H_{Waugh}(0) > H_{Lara}(0)) = 0.85$). This surprising result shows that popular perceptions are not necessarily accurate, given that many people regard Steve Waugh as the player they would choose to play "for their life"[2]. However, Steve Waugh's predictive hazard function has a transition to its high equilibrium value that is sooner and faster than both Lara and Langer ($P(\tau_{Waugh} < \tau_{Langer}$ and $\tau_{Waugh} < \tau_{Lara}$ and $L_{Waugh} < L_{Langer}$ and $L_{Waugh} < L_{Lara}) = 0.53$, the prior probability of this is 0.0625). Therefore, perhaps his reputation is upheld, except at the very beginning of an innings.

Note that the general tendency of the effective average to drift upwards as a function of score does not imply that all batsman get better the longer their innings goes on - since once the transition has occurred, our model says that the hazard rate should stay basically constant. Instead, the hazard function of the predictive distribution describes a gradual change in our state of knowledge:

[2]Technically, the correct choice would be to choose a player that minimised the expected loss, where loss is defined as the amount of injury inflicted on the spectator as a function of the batsman's score.

the longer a batsman stays in, the more convinced we are that our estimate of their overall ability μ_2 should be higher than our prior estimate. This is why there is an upwards tendency in the predictive ability function for all players, even well after the change-point transition is completed.

7 Conclusions

This paper has presented a simple model for the hazard function of a batsman in test or first class cricket. Applying the model to data from several cricketers, we found the expected conclusion: that batsmen are more vulnerable towards the beginning of the innings. However, this analysis now provides a quantitative measurement of this effect, showing how significant it is, and the fact that there is substantial variation in the size of the effect for different cricketers. Surprisingly, we found that Steve Waugh was the second most vulnerable player in the sample at the beginning of an innings - only Shane Warne, a bowler, was was more vulnerable. Even Shaun Pollock is better at the beginning of his innings. This surprising result would have been very hard to anticipate.

From this starting point, there are several possible avenues for further research. One interesting study would involve much larger samples of players so we can identify any trends. For instance, is it true that all-rounders are more robust batsman in general, or are Chris Cairns and Shaun Pollock atypical? Also, it should be possible to create a more rigorous definition of the notion of robustness discussed above. Once this is done, we could characterise the population as a whole, and search for possible correlations between batting average, strike-rate (runs scored per 100 balls faced) and robustness. Modelling the entire cricket population would also allow for a more objective choice for the parameterisation of the hazard function, and the prior distribution over its parameter space. Depending on the results, these kinds of analyses could have implications for selection policy, especially for opening batsmen where consistency is a highly desirable trait.

References

Cai, T., Hyndman, R.J. and Wand, M.P. (2002). *Mixed model-based hazard estimation*, Journal of Computational and Graphical Statistics, 11, 784-798.

Jarzynski, C. 1997a, *Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach*, Physical Review E, 56, 5018.

Jarzynski, C. 1997b, *Nonequilibrium Equality for Free Energy Differences*, Physical Review Letters, 78, 2690.

Kimber, A. C., Hansford, A. R., *A Statistical Anaylsis of Batting in Cricket*, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 156, No. 3. (1993), pp. 443-455.

MacKay, D. J. C. 2003, *Information Theory, Inference and Learning Algorithms*, ISBN 0521642981. Cambridge, UK: Cambridge University Press.

Neal, R. M. 1998, *Annealed Importance Sampling*, arXiv:physics/9803008, Technical Report No. 9805 (revised), Dept. of Statistics, University of Toronto. Available online at `http://www.cs.toronto.edu/~radford/papers-online.html`

O'Hagan, A., & Forster, J. 2004, *Kendall's advanced theory of statistics. Vol.2B: Bayesian inference*, 2nd ed., by A. O'Hagan and J. Forster. 3 volumes. London: Hodder Arnold, 2004.

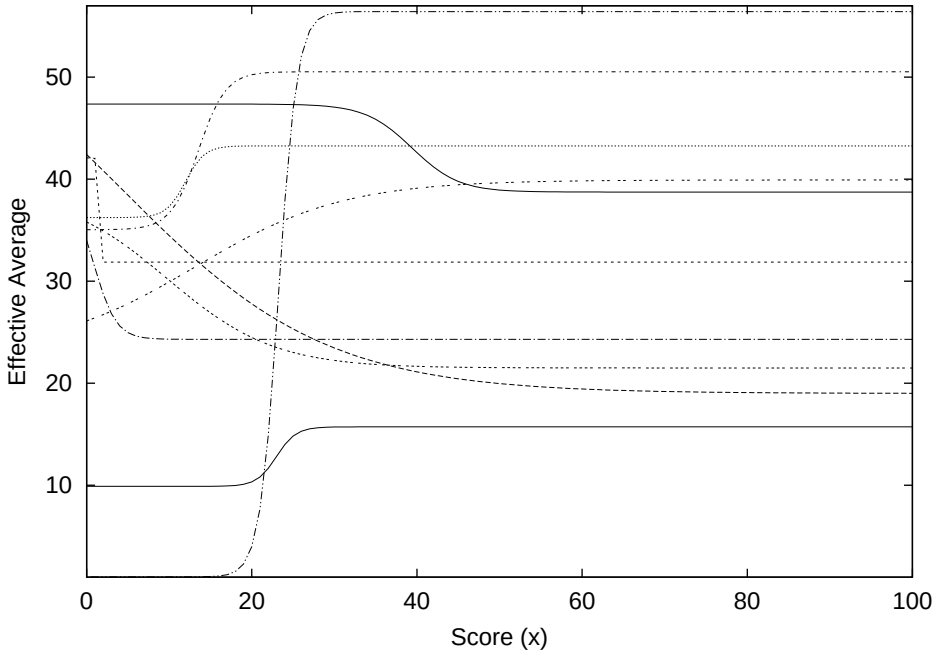


Figure 1: Illustrative examples of the kind of functions produced by Equation 9, with a range of typical values (chosen from the prior, see Section 4) of the four parameters μ_1, μ_2, τ and L.

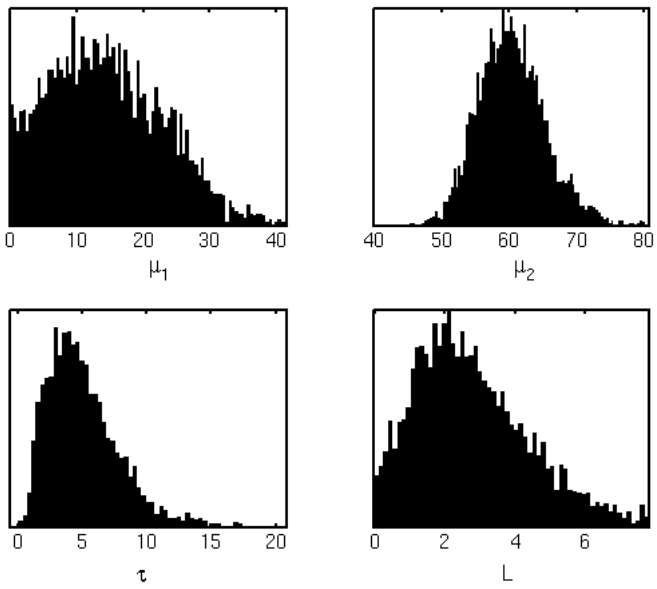
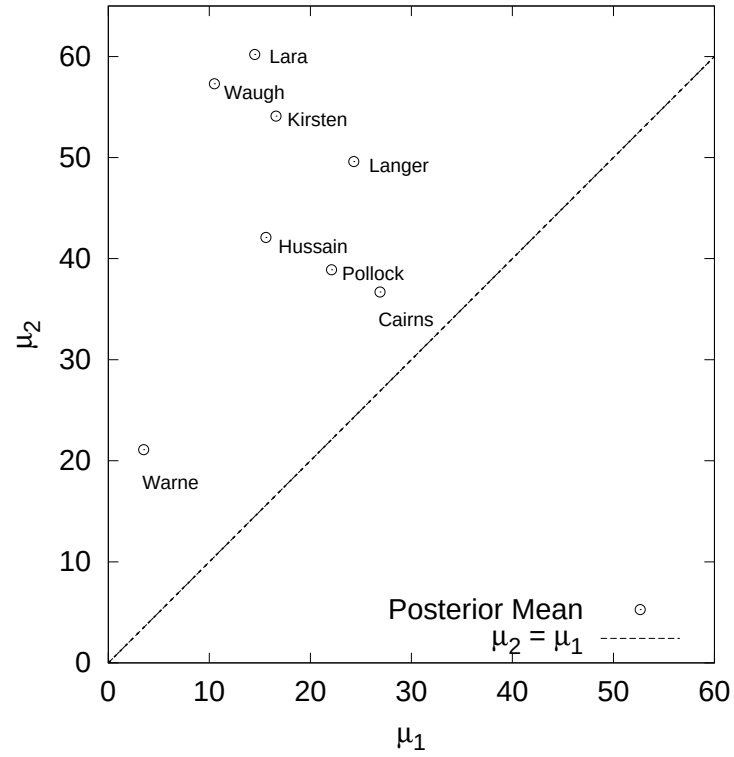


Figure 2: Results for the four parameters for Brian Lara. The top two panels show the posterior distributions for his two abilities (effective batting averages) μ_1 and μ_2, while the lower panels show the distributions for the change-point τ and change-timescale L. See the text for interpretation.

Figure 3: Estimated location of each of the players on the μ_1-μ_2 plane. The line corresponds to $\mu_2 = \mu_1$, and the closer a player is to the line, the more robust the player. Surprisingly, Shaun Pollock and Chris Cairns lie closest to the line. Although each player has been represented by a point on this plot, these are only estimates (posterior means), and each point is actually just the centre of a large zone of uncertainty.



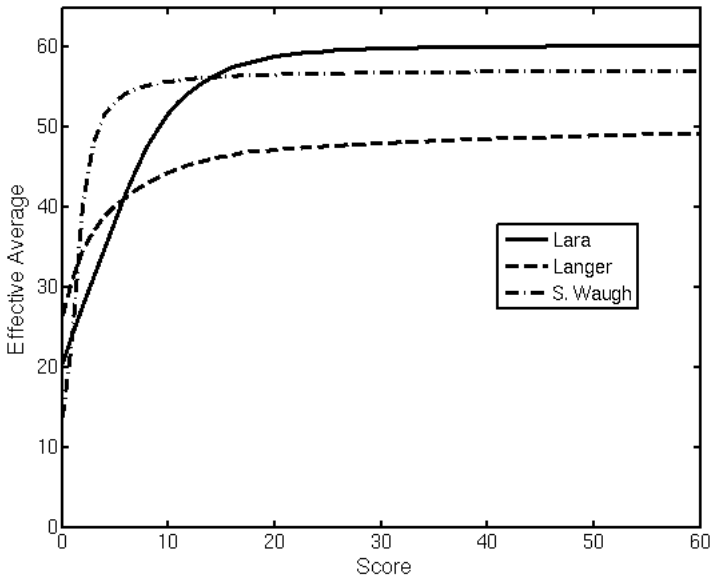
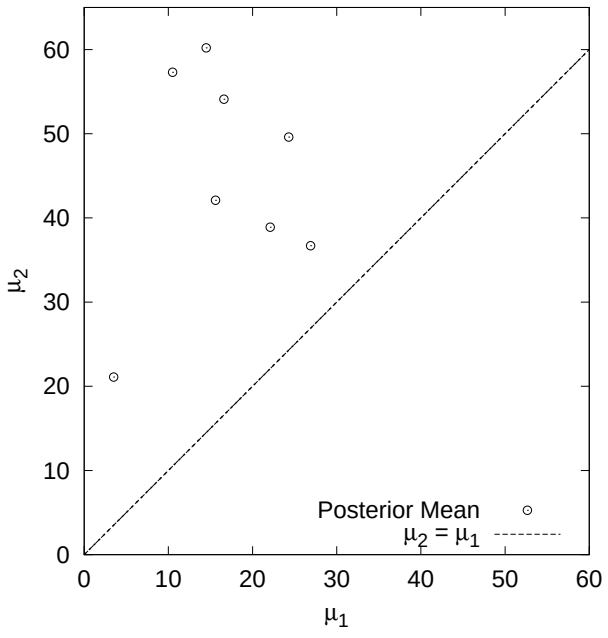


Figure 4: Predictive Hazard Functions for Brian Lara, Justin Langer and Steve Waugh. As expected, Justin Langer proves to be less vulnerable than Brian Lara at the beginning of his innings. However, surprisingly, Steve Waugh is more vulnerable than Brian Lara, although he gets his eye in sooner.


60
50
40
30
20
10
0
0
10
20
30
40
50
60
μ_2
μ_1
Posterior Mean
$\mu_2 = \mu_1$