A SIMULATOR FOR TWENTY20 CRICKET

JACK DAVIS, HARSHA PERERA AND TIM B. SWARTZ[1]

Simon Fraser University

Summary

This paper develops a Twenty20 cricket simulator for matches between sides belonging to the International Cricket Council. As input, the simulator requires the probabilities of batting outcomes which are dependent on the batsman, the bowler, the number of overs consumed and the number of wickets lost. The determination of batting probabilities is based on an amalgam of standard classical estimation techniques and a hierarchical empirical Bayes approach where the probabilities of batting outcomes borrow information from related scenarios. Initially, the probabilities of batting outcomes are obtained for the first innings. In the second innings, the target score obtained from the first innings affects the aggressiveness of batting during the second innings. We use the target score to modify batting probabilities in the second innings simulation. This gives rise to the suggestion that teams may not be adjusting their second innings batting aggressiveness in an optimal way. The adequacy of the simulator is addressed through various goodness-of-fit diagnostics.

Keywords: Empirical Bayes, Markov chain Monte Carlo.

[1]Author to whom correspondence should be addressed.
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC, Canada V5A1S6
The authors wish to thank two anonymous reviewers for helpful comments that resulted in an improvement to the manuscript. Swartz has been supported by the Natural Sciences and Engineering Research Council of Canada.

1 INTRODUCTION

The game of cricket has a long history dating back to the 16th century. The most recent form of cricket, known as Twenty20 cricket (or T20 cricket), began in 2003 involving matches between English and Welsh domestic sides. Since 2003, Twenty20 cricket has exploded in popularity with five World Cups having been contested (2007, 2009, 2010, 2012 and 2014). The Indian Premier League (IPL) which had its inaugural season in 2008 is known as the showcase for T20 cricket. The IPL continues to grow in popularity with respect to the number of teams, television contracts, salaries, etc.

Except for some subtle differences (e.g. fielding restrictions, limits on the number of overs for bowlers, etc.), Twenty20 cricket shares many of the features of one-day cricket. One-day cricket was introduced in the 1960s, and like T20 cricket, it is a version of cricket based on limited overs. The main difference between T20 cricket and one-day cricket is that each batting side in T20 is allotted 20 overs compared to 50 overs in one-day cricket. This difference allows Twenty20 matches to finish in roughly three hours, a length of time comparable to the duration of matches in many other professional sports.

Simulation methodologies have been developed and proven useful for many types of complex systems. For example, the simulation of weather systems using mathematical models has a long history in both short-term weather forecasts and in the prediction of climate change (Lynch 2008). A match simulator for T20 cricket would likewise be useful. For example, the prediction of match outcomes is obviously of interest to cricket enthusiasts. A match simulator would also facilitate the investigation of various match characteristics for which there does not exist a sufficient number of actual matches. For example, suppose that a T20 team is considering a new batting lineup. They may be interested in the distribution of runs scored by the hypothetical lineup. Naturally, a good simulator for T20 cricket is one which is realistic and captures the complexity of the game. To our knowledge, there have not been any realistic simulators developed for Twenty20 cricket. A difficulty in the development of a realistic T20 match simulator involves gaining a detailed understanding how various interacting factors (e.g. overs, wickets, batsmen, bowlers, the target, the powerplay[2], etc.) affect run progression.

Simulators have been investigated for other forms of cricket. The earliest "simulators" were proposed by Elderton (1945) and Wood (1945) who fitted simple geometric distributions for the number of runs scored in test cricket. Dyte (1998) also considered the simulation of test cricket matches where the only inputs were career batting and bowling averages. In one-day cricket, Bailey & Clarke (2006) introduced covariates related to run scoring and used the normal distribution for the generation of runs. In test cricket, Scarf et al. (2011) model the number of runs by fitting a zero-inflated negative binomial distribution to each of the 10 partnerships. More closely related to this paper, Swartz, Gill & Muthukumarana (2009) developed a Bayesian latent variable model which provided batting outcome probabilities in one-day cricket. A criticism of Swartz, Gill & Muthukumarana (2009) is that they use a coarse discretization of wickets lost and overs consumed based on 9 overall categories. In particular, their structure does not account for powerplays.

Section 2 is concerned with preliminaries related to the T20 simulator. We first introduce the extensive dataset which is used throughout the paper. Exploratory data analyses are carried out to motivate the subsequent modeling. The T20 simulator is then described in simple terms for first innings batting. Section 3 discusses the inputs to the simulator. Specifically, batting outcomes are enumerated and the corresponding probabilities are derived from multinomial distributions. Our model is highly parametrized and we use an amalgam of classical estimation techniques and a hierarchical Bayesian model to estimate the multinomial parameters. One of the key features of the approach is that the estimators from a given scenario borrow information from related scenarios to improve reliability. Another noteworthy aspect of the approach concerns the detail which is provided in ball-by-ball scoring. Simulators which simply generate the total number of runs for each over do not address the manner in which runs are scored. In section 4, the simulator is extended in various ways. We consider the case of specific batsman/bowler matchups, the home team advantage and second innings simulation where the target score is taken into account. Clearly, higher target scores force the second innings batting team to be more aggressive. When they are more aggressive, they score more runs but are more likely to be dismissed. In section

[2]the powerplay is defined later

5, we demonstrate the realism of the simulator via some goodness-of-fit diagnostics. A notable consequence of the validation exercise is the suggestion that teams may not be batting optimally during the second innings. Specifically, teams that are falling behind in the second innings may not be increasing their aggressiveness in an incremental fashion. We also illustrate the utility of the simulator by addressing some problems of prediction. We conclude with a short discussion in section 6.

2 PRELIMINARIES

For the analysis, we consider all T20 matches that took place from 2005 until the end of 2013 which involved full member nations of the International Cricket Council (ICC). Currently, the 10 full members of the ICC are Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. Details from these matches can be found in the Archive section of the CricInfo website (www.espncricinfo.com). A proprietary R-script was used to parse and extract ball-by-ball information from the Match Commentaries. In total, we obtained data from 250 matches. In Table 1, we provide summary statistics for the matches where we observe that Bangladesh and Zimbabwe are clearly the weakest T20 teams. Amongst the other 8 teams, the winning percentages do not vary greatly. When looking at the differences between runs scored versus runs allowed for individual teams, it appears that Sri Lanka's win percentage is lower than what might be expected.

We now study various features related to batting. We temporarily ignore extras (sundries) that arise via wide-balls and no-balls, and note that there are only 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

$$\begin{array}{lcl} \text{outcome } j=0 & \equiv & 0 \text{ runs scored} \\ \text{outcome } j=1 & \equiv & 1 \text{ runs scored} \\ \text{outcome } j=2 & \equiv & 2 \text{ runs scored} \\ \text{outcome } j=3 & \equiv & 3 \text{ runs scored} \\ \text{outcome } j=4 & \equiv & 4 \text{ runs scored} \\ \text{outcome } j=5 & \equiv & 5 \text{ runs scored} \\ \text{outcome } j=6 & \equiv & 6 \text{ runs scored} \\ \text{outcome } j=7 & \equiv & \text{dismissal} \end{array} \tag{1}$$

Team	Matches	Win %	$\bar{R}^{(S)}$	$\bar{R}^{(A)}$
Australia	63	52%	161.7 (33)	157.8 (30)
Bangladesh	22	18%	134.0 (07)	169.5 (15)
England	58	55%	159.7 (23)	158.7 (36)
India	42	50%	159.4 (23)	164.7 (19)
New Zealand	64	44%	153.6 (32)	157.8 (32)
Pakistan	67	55%	152.3 (36)	144.9 (31)
South Africa	58	62%	147.6 (29)	148.9 (29)
Sri Lanka	52	48%	159.7 (26)	139.3 (26)
West Indies	49	41%	159.7 (28)	147.5 (21)
Zimbabwe	25	20%	129.9 (13)	171.8 (12)

Table 1: Summary statistics for the T20 dataset corresponding to matches from February 17, 2005 through November 13, 2013. The variables $\bar{R}^{(S)}$ and $\bar{R}^{(A)}$ denote the average number of first innings runs scored and runs allowed, respectively, with the number of matches in parentheses.

In the list (1) of possible batting outcomes, we include byes, leg byes and no balls where the resultant number of runs determines one of the outcomes $j = 0, \ldots, 7$. We note that the outcome $j = 5$ is rare but is retained to facilitate straightforward notation.

We first calculate the proportions $\hat{p}_0, \ldots, \hat{p}_7$ corresponding to the first innings batting outcomes. Table 2 provides a comparison of these proportions based on the T20 dataset compared with the proportions for fourth innings batting in test cricket as reported by Perera, Gill & Swartz (2013). We observe that T20 batting is much more aggressive than batting in test cricket. For example, 6's occur with a much greater frequency (by a factor of 14) in T20 cricket than in test cricket. Consequently, the modeling of runs is dependent on the particular form of cricket under consideration.

	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$
T20 (1st innings)	0.301	0.411	0.078	0.006	0.103	0.002	0.042	0.057
Test (4th innings)	0.743	0.128	0.035	0.010	0.065	0.000	0.003	0.016

Table 2: Sample proportions corresponding to batting outcomes in two forms of cricket.

In one-day cricket, it well-known that opening batsmen begin matches cautiously, attempting to avoid wickets and hoping to develop a batting rhythm. As the match proceeds, batting tends to become more aggressive. In Twenty20 cricket, there are only 20 overs, and it is conceivable that batsmen behave uniformly throughout a match. In other words, T20 batsmen exhibit aggressiveness at the beginning of a match, and display the same level of aggressiveness as the match proceeds. The intuition is that it is beneficial to always be aggressive since the 10 allocated wickets are likely to suffice for 20 overs. For example, in our dataset, teams were made "all out" during the first innings in only 11% of the matches. If it were true that batsmen displayed constant aggressiveness in T20, this would facilitate modeling since batting characteristics would not change with respect to wickets lost nor overs consumed.

In Figure 1, we provide plots of the proportions of batting outcomes in the first innings of the T20 dataset stratified by over. We observe that the above intuition about constant aggressive in T20 batting is clearly false. For example, we observe that batsmen have very few 4's during the first two overs which is a period of adjustment to the bowler, to the ball, to the pitch, to the weather, etc. This initial period is followed by a spike in 4's which corresponds to the powerplay (the first 6 overs of the match). Once the powerplay terminates, the proportion of 4's plummets in the 7th over, and then there is a gradual (nearly monotonic) increase in 4's until the completion of the innings. These observations are important for our subsequent modeling assumptions.

Note that it is also possible to produce a plot of the proportion of batting outcomes in the first innings stratified by wickets lost. Such a plot suggests that batting characteristics are also dependent on wickets lost. Furthermore, with respect to batting characteristics, there is clearly an interaction between the number of overs consumed and the number of wickets lost. For example, a batsman is more aggressive in the 19th over with two wickets lost than in the 19th over with 9 wickets lost. The existence of the interaction is one of the guiding principles in the development of the Duckworth-Lewis resource table (Duckworth & Lewis 2004) used to reset targets in rain-interrupted matches.

According to the enumeration of the batting outcomes in (1), the preceding discussion suggests a statistical model for the number of runs scored by the ith batsman:

$$(X_{iow0}, \ldots, X_{iow7}) \sim \text{multinomial}(m_{iow};\ p_{iow0}, \ldots, p_{iow7}) \tag{2}$$

where X_{iowj} is the number of occurrences of outcome j by the ith batsman during the oth over when w wickets have been taken. In (2), m_{iow} is the number of balls that batsman i has faced in the dataset corresponding to the oth over when w wickets have been taken. The multinomial

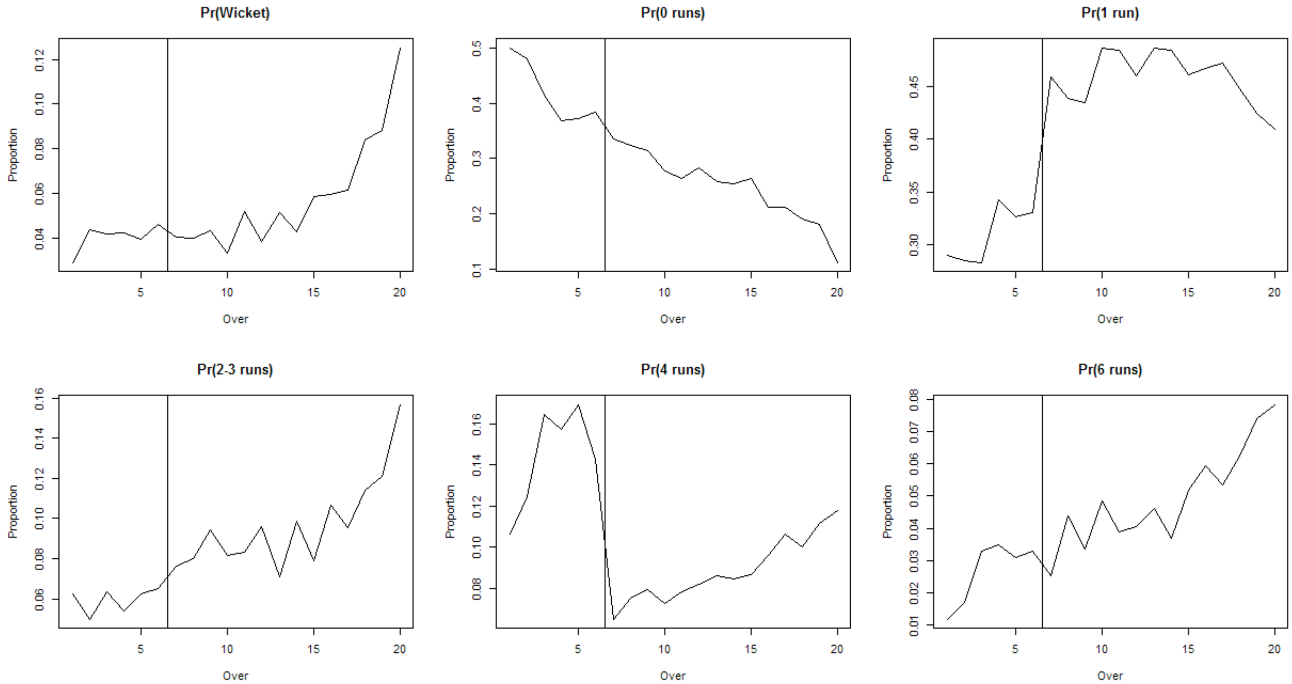

Figure 1: Proportion of batting outcomes stratified by over. The vertical line denotes the termination of the powerplay.

distributions (2) define the likelihood which is used to estimate the characteristics p_{iowj}. In section 3, we address the difficulty of parameter estimation in a highly parametrized setting with sparse data; i.e. $m_{iow} \approx 0$ for many of the situations (i, o, w). In section 4, we address the problem where batsmen face bowlers of varying quality.

Assume temporarily that we are able to obtain (i.e. estimate) the multinomial parameters in (2). We would then be able to generate (from a multinomial distribution with $m = 1$) the outcome of a single ball. A straightforward algorithm for simulating first innings runs against an average bowler would proceed as follows: The batting team begins with a fixed batting lineup where $k = 1$, $w = 0$ and the 11 batsmen are described by their batting characteristics p_{iowj}. **Start**: Suppose that ball $k \leq 120$ of the match is being delivered which determines the corresponding over o. Provided that $w < 10$ wickets have been taken, and provided that $o \leq 20$, a variate $u_1 \sim \text{Uniform}(0, 1)$ is generated. Otherwise, the innings are complete. If $u_1 \leq q_1$, then an extra has occurred, a single run is added to the run counter and we return to **Start**. Based on our extensive T20 dataset, we have determined that extras occur with probablity $q_1 = 0.033$.

If $u_1 > q_1$, then a second variate $u_2 \sim$ Uniform$(0, 1)$ is generated. Then, according to the specific probabilities $p_{iow0}, \ldots, p_{iow7}$ with the ith batsman batting, one of the outcomes in (1) is determined. The run counter, the ball counter k and the wicket counter w are updated accordingly. If the batsman scores 1 run, 3 runs or 5 runs (essentially impossible), then his batting partner faces the next ball. If the batsman is dismissed, the batsman is replaced by the next batsman in the batting lineup. We then return to **Start**.

Therefore, given the multinomial parameters p_{iowj} in (2), it is a very simple coding exercise to develop a first innings simulator for T20 cricket. In section 4.3, we introduce modifications for second innings simulation. With a match simulator, we are then able to investigate various situations of interest with respect to Twenty20 cricket. In the following section, we discuss the fundamental problem of parameter estimation.

3 PARAMETER ESTIMATION

Under model (2), we are concerned with the estimation of the multinomial parameters p_{iowj} subject to the constraints $\sum_{j=0}^{7} p_{iowj} = 1$, $\forall i, o, w$. Whereas maximum likelihood estimation of the p_{iowj} is "easy", it does lead to reliable estimation due to the sparsity of the data in many of the situations $(iowj)$. For example, a bowler would never batting in the early overs of an innings. With so many parameters, we make some simplifying assumptions based on our observations in section 2. Specifically we let

$$p_{iowj} = \frac{\tau_{owj}\ p_{i70j}}{\sum_j \tau_{owj}\ p_{i70j}} \ . \tag{3}$$

In (3), the parameter p_{i70j} represents the baseline characteristic for batsman i with respect to batting outcome j. The characteristic p_{i70j} is the probability of outcome j associated with the ith batsman at the juncture of the match immediately following the powerplay (i.e. the 7th over) when no wickets have been taken. The multiplicative parameter τ_{owj} scales the baseline performance characteristic p_{i70j} to the stage of the match corresponding to the oth over with w wickets taken. The denominator in (3) ensures that the relevant probabilities sum to unity.

Therefore, there is an implicit assumption in (3) that the stage of the first innings (overs completed and wickets lost) affects all batsmen in the same fashion. We argue that the stage of the match determines the aggressiveness of batting in general. Although batsmen are unique, their batting characteristics change by the same multiplicative factor which is essentially an indicator of aggression. For example, when aggressiveness increases relative to the baseline state, one would expect $\tau_{ow4} > 1$ and $\tau_{ow6} > 1$ since bolder batting leads to more 4's and 6's. In section 5, we demonstrate that the batting characteristics modeled according to (3) lead to a realistic match simulator.

Given (3), we need to estimate the multiplicative parameters τ_{owj} and the individual baseline probabilities p_{i70j}. With $N = 384$ batsmen in the dataset, there are $(19)(9)(7) + (384)(7) = 3885$ unknown parameters. And with 30458 batting outcomes, the ratio of data to parameters is roughly 8:1. Ideally, we would like to estimate the parameters simultaneously. However, we were unable to do so because of our inability to obtain a convergent Markov chain; the corresponding posterior distribution is complex and is specified below. Instead, we opted for a hybrid method of estimation. We first estimate the multiplicative parameters τ_{owj}, and then given the τ_{owj}, we estimate the baseline probabilities p_{i70j} for individual batsmen. The spirit of the two-step estimation procedure is reminiscent of profile likelihood methodology (Davison 2003).

In the Appendix, we discuss the estimation of the multiplicative factors τ_{owj}. In the remainder of this section, we describe the hierarchical model and the methodology used to estimate the baseline probabilities p_{i70j} given the τ_{owj}.

Model (2) describes the sampling distribution of the data. In a Bayesian framework, we also require a prior distribution on the parameters p_{i70j}. Since the p_{i70j} are probabilities defined on simplices, we make the prior assumption $(p_{i700}, \ldots, p_{i707}) \sim \text{Dirichlet}(a_0, \ldots, a_7)$. Letting p denote the vector of all baseline parameters and letting $[x \mid y]$ denote generic notation for the conditional density of x given y, we obtain the posterior density

$$\begin{aligned} [p \mid X] &\propto [X \mid p]\,[p] \\ &\propto \left(\prod_{i,o,w} \left(\frac{\tau_{ow0}\ p_{i700}}{\sum_j \tau_{owj}\ p_{i70j}} \right)^{X_{iow0}} \cdots \left(\frac{\tau_{ow7}\ p_{i707}}{\sum_j \tau_{owj}\ p_{i70j}} \right)^{X_{iow7}} \right) \left(\prod_i p_{i700}^{a_0-1} \cdots p_{i707}^{a_7-1} \right) \end{aligned} \quad (4)$$

where independence is assumed across the batsmen, overs and wickets.

In a Bayesian setting, it is standard practice to use posterior means as estimators. Given the complexity of the posterior density (4), we propose a sampling based approach based on Markov chain Monte Carlo (MCMC) methods to estimate the parameters p_{i70j}. With the τ's specified, we note that the posterior (4) is amenable in the sense that it factors according to each batsman i. Specifically, we first consider a Gibbs sampling algorithm (Gilks, Richardson & Spiegelhalter 1996) where the full conditional densities take the form

$$[p_{i700}, \ldots, p_{i707} \mid \cdot\,] \;\propto\; \frac{p_{i700}^{(\sum_{o,w} X_{iow0})+a_0-1} \cdots p_{i707}^{(\sum_{o,w} X_{iow7})+a_7-1}}{\prod_{o,w}(\sum_j \tau_{owj} p_{i7oj})^{m_{iow}}} \,. \tag{5}$$

Unfortunately, the full conditional densities (5) are nonstandard in the sense that there does not appear to be a simple way to generate variates directly from the corresponding distributions. We therefore consider a Metropolis within Gibbs step where the proposal distributions are Dirichlet with parameters corresponding to the exponents in the numerator of (5).

The described procedure is fully Bayesian and only requires the subjective setting of the hyperparameters $a_0, \ldots, a_7$. Note that the default setting $a_0 = \cdots = a_7 = 0$ is an obvious choice although it does not take prior knowledge into account. As the a_i's get larger, there is greater shrinkage of the individual estimates towards a common characteristic. In this application, we take an empirical Bayes approach where we use the data to specify the hyperparameters.

We begin by setting $a_j = c \sum_{i,o,w}(X_{iowj}/\tau_{owj}) / \sum_{i,o,w,k}(X_{iowk}/\tau_{owk})$ for some $c > 0$ and $j = 0, \ldots, 7$. As $c \to \infty$, the posterior density (4) approaches a Dirichlet density dominated by the a_j terms and where the posterior mean of p_{i70j} is

$$\hat{p}_{i70j} \to \frac{\sum_{i,o,w}(X_{iowj}/\tau_{owj})}{\sum_{i,o,w,k}(X_{iowk}/\tau_{owk})}$$

which may be thought of as the common mean (over all batsmen) for batting in over 7 with zero wickets after transforming all situations to the case $o = 7$ and $w = 0$.

Therefore, our problem is the selection of $c > 0$ such that it is not too large (where all batsmen have identical characteristics) but is also not too small (where small sample sizes may give rise to unrealistic characteristics). We note that the standard deviations of the prior parameters p_{i70j} are proportional to $1/\sqrt{c+1}$. Therefore, the tuning parameter $c > 0$ may be thought of as the prior sample size. We have tinkered with various choices and have found that $c = 60.0$ provides realistic characteristics. Also, we have observed that the simulation results do not differ practically based on the selection of $c \in (50, 80)$.

Again, we emphasize that there is a need to choose c substantially greater than zero since there are batsmen with limited batting histories. We do not want the sparsity of their batting attempts to result in unrealistic batting characteristics. Choosing c substantially larger than zero shrinks their observed proportions closer to the mean characteristics for all batsmen. In Figure 2, we provide a histogram of the number of batting attempts by the batsmen in our dataset. We observe that more than 200 batsmen have faced fewer than 40 balls. Only a handful of batsmen have faced more than 500 balls.

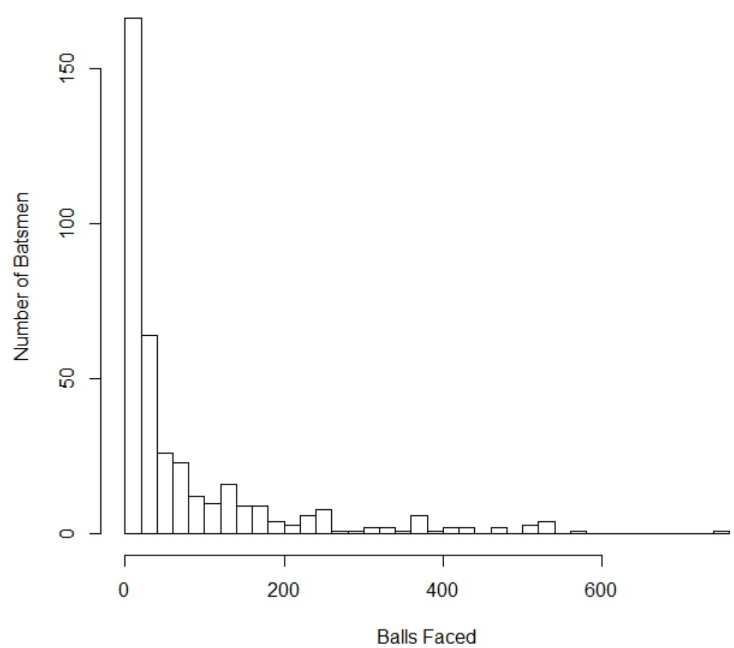

Figure 2: Histogram of the number of first innings batting attempts by the 384 batsmen in the dataset.

We note that the player characteristics p_{iowj} estimated in this section may be viewed as average characteristics taken over a player's career. In some applications, it might be more meaningful to carry out simulations based on current form. The natural way to do this is to weight the data with more weight given to recent observations (i.e. matches). Operationally, this was achieved by replacing X_{iowj} in (4) with $\sum_g r^g X_{iowjg}$ where r is a decay ratio and the subscript g is the match index that denotes the number of games before the last game was played. We determined $r = 0.88$ via maximization of the posterior density.

4 EXTENDING THE SIMULATOR

Although the proposed simulator is detailed and captures many of the essential features of T20 cricket, it may be extended in various ways to enhance realism.

4.1 The Impact of the Bowler

In model (1), the data X_{iowj} correspond to batting outcomes. Noting the symmetry between batting and bowling, one can also specify a model from the point of view of bowlers. Specifically,

$$(Y_{iow0}, \ldots, Y_{iow7}) \quad \sim \quad \text{multinomial}(m_{iow};\ q_{iow0}, \ldots, q_{iow7}) \tag{6}$$

where Y_{iowj} is the number of occurrences of outcome j as defined in (1) experienced by the ith bowler during the oth over when w wickets have been taken. The parameters q_{iowj} therefore describe bowling characteristics with respect to an average batsman. The parameters q_{iowj} can be estimated using the same empirical Bayes approach described in section 3.

In a given match simulation, rather than having batsmen face average bowlers, it is more realistic to have batsmen face specified bowlers. It is suggested that a modification can be made in the case of batsman i_1 facing bowler i_2 by using the characteristic

$$p_{i_1owj} + q_{i_2owj} - \bar{p}_{owj} \tag{7}$$

for outcome j where $\bar{p}_{owj}$ is the average batting characteristic of outcome j taken over all batsmen. Note that $\bar{p}_{owj}$ is also the average bowling characteristic of outcome j taken over all bowlers.

The batting characteristics given by (7) are sensible in the sense that if batsmen i_1 is average, then (7) reduces to q_{i_2owj}, and if bowler i_2 is average, then (7) reduces to p_{i_1owj}. The quantity (7) provides a synthesis of the individual characteristics of the batsman and the bowler.

4.2 The Home Team Advantage

Although the underlying causes of the home team advantage are difficult to pinpoint precisely, the effect of the home team advantage is real and the magnitude of the advantage varies according to the sport (Swartz & Arce 2014). One of the interesting findings is that there is not a strong case for differential home team advantages amongst teams that compete in the same league (hockey and basketball).

We note that de Silva, Pond & Swartz (2001) considered the effect of the home team advantage in one-day international cricket. They defined the home team advantage as the number of runs one would expect a home team to defeat the road team when both teams are of equal strength. The advantage was estimated to be worth approximately 16 runs (their Model D).

In our T20 simulator, we take a simple approach for implementing the effect of the home team advantage. From the data set, the average number of first innings runs scored by the home team was 158.4 and the average number of first innings runs scored by the away team was 149.4. Therefore, in the simulator, we scale the number of runs scored by the home team by the factor $2(158.4)/(158.4+149.4) = 1.03$, and we scale the number of runs scored by the away team by the factor $2(149.4)/(158.4+149.4) = 0.97$. This is done at the individual batsmen level which can lead to a non-integer number of runs scored. At the end, we round the total number of runs scored. In a match played at a neutral site, no adjustment is made for the home team advantage.

4.3 Second Innings Simulation

Up until now, we have focused on first innings simulation. We would like to extend the simulator to the second innings so that we can address questions such as "What is the probability that the team batting second wins the match given a specific target score obtained during the first innings?"

As we have seen, batting characteristics vary according to the aggressiveness of the batsmen, and aggressiveness is determined by the state of the match. And in the second innings, the target score forms a component of the state of the match. Naturally, batsmen need to be more aggressive when the target score is greater.

The idea for second innings simulation (borrowed from Swartz, Gill & Muthukumarana 2009) is that batting characteristics are tweaked according to changes in aggressiveness. For this, let r_1 denote the number of runs scored during the first innings (the target) and let r_2 be the number of runs scored in the second innings up to the current stage of the match (determined by the number of overs completed and the number of wickets taken). Then the team batting second requires $r_1 - r_2 + 1$ additional runs to win the match.

Next, we let d denote the number of *resources* remaining in the match from the given stage of the match. The value d is obtained from the Duckworth-Lewis table (Duckworth & Lewis 2004) modified for T20. Resources are a synthesis of the number of overs completed and the number of wickets taken. As the number of overs increases and the number of wickets increases, a team's batting resources decrease. Therefore, in order for the team batting second to win the match, they need to bat with a "runs to resources ratio" of at least

$$\frac{r_1 - r_2 + 1}{d} \ . \tag{8}$$

The quantity (8) is a measure of the required level of aggressiveness where larger values indicate increasing aggressiveness. In terms of aggressiveness, (8) is essentially a ratio of what is needed to what is typically available.

But according to the batting characteristics p_{iowj} of the ith batsman, how aggressive is the batsman actually batting? At the given stage (o, w) of the match, the expected number of runs scored on the next ball bowled is

$$E^{(1)}_{iow} = p_{iow1} + 2p_{iow2} + 3p_{iow3} + 4p_{iow4} + 6p_{iow6} \ .$$

Similarly, the expected number of resources consumed on the next ball bowled is

$$E^{(2)}_{iow} = xp_{iow0} + xp_{iow1} + xp_{iow2} + xp_{iow3} + xp_{iow4} + xp_{iow6} + (x + y)p_{iow7}$$

where x are the resources lost due to the current ball and y are the resources lost due to a wicket. The values x and y are obtained from the Duckworth-Lewis table.

Putting this all together, the logic is that if the current level of batting aggressiveness $E^{(1)}_{iow}/E^{(2)}_{iow}$ is not sufficiently large to win the match, then batting aggressiveness should be increased. Using the above notation, batting aggressiveness should be increased if

$$\frac{E_{iow}^{(1)}}{E_{iow}^{(2)}} < \frac{r_1 - r_2 + 1}{d}. \tag{9}$$

Since batting aggressiveness increases as the overs increase for a fixed number of wickets lost, our approach is to find the minimum value $o^* > o$ such that

$$\frac{E_{io^*w}^{(1)}}{E_{io^*w}^{(2)}} \geq \frac{r_1 - r_2 + 1}{d}\ . \tag{10}$$

When we have solved for o^* in (10), then even though batting is taking place at stage (o, w) of the match, we will use the batsman's characteristics p_{io^*wj}. The numerical determination of o^* is straightforward. We begin with $o^* = o$, and then increment o^* until (10) is satisfied. If o^* cannot be determined, then the simulation proceeds with the batsman's personal maximum aggressiveness as given by the characteristics p_{iowj} where $o = 20$.

The modeling of the second innings was one of the more challenging aspects of the paper. Whereas everyone agrees that batting tendencies change as the match progresses, it is not clear exactly how batting characteristics are modified. As a player, you may know that your team needs more runs, but is it possible to increase your run output without sacrificing additional wickets? If a batsman is able to do this, then the obvious question is why don't they enact this batting behaviour all of the time? A feature of our second innings modeling is that batsmen modify their batting characteristics in a manner that is within their capabilities. When they need to be more aggressive, they simply behave as though they were batting later in the innings, and for this we have good estimation procedures.

We also emphasize that the above description of second innings batting is how we believe batsmen should modify their batting when trailing in a match. It is an optimal behaviour which we are suggesting. In section 5, we observe that batsmen do not quite behave in this optimal fashion.

5 ADEQUACY OF THE SIMULATOR

Our general approach in the development of the simulator has been to simulate matches and then critically examine the output. When the output is in conflict with empirical results or our intuition, we revisit the underlying theory and iterate towards a realistic simulator.

Our resulting simulator has been coded in the R programming language. The time required to simulate a single innings is virtually instantaneous. To simulate 1000 innings requires approximately 1.5 minutes of computation on an older laptop computer.

We first investigate individual player characteristics obtained through our hybrid estimation scheme based on the multinomial model (2) and the simplifying assumptions given by (3). In Table 3, we provide batting characteristics for two prominent batsmen, Shane Watson of Australia and AB de Villiers of South Africa. Watson is an all-rounder who is typically an opening batsman and has faced 673 balls in T20 cricket. AB de Villiers is a pure batsman who usually bats in the third or fourth position in the lineup and has faced 666 balls. In comparing batsmen, we see that de Villiers resembles an average batsman. In fact, his characteristics are a little on the cautious side, scoring slightly fewer sixes and getting out less often than an average batsman. On the other hand, Watson is a power hitter who scores sixes at a high rate but is also dismissed frequently. It is interesting to compare batting performances at different stages of a match. As expected, run rates increase in the 10th over with zero wickets compared to the period immediately after the powerplay (7th over with zero wickets). Run rates continue to increase at the stage of the 15th over with two wickets lost. This is sensible as there are many wickets remaining and batsmen can take chances by increasing their aggressiveness. Note that the probabilitity of a wicket increases dramatically at this stage. We then observe a decrease in productivity in the 15th over with 8 wickets. This also corresponds to our intuition as batsmen need to protect their wickets as only two replacement batsmen remain.

We also investigate the adequacy of the simulator by looking at the larger picture in terms of team performance. For each of the 10 ICC teams, we recorded their first innings runs for all T20 matches from the past two years. We then determined a team batting lineup that was representative during that period, and simulated 1000 matches against average opposition. The

Player	Over	Wicket	0 runs	1 run	2/3 runs	4 runs	6 runs	Out	E(RR)
Average	7	0	0.3421	0.3700	0.0956	0.1212	0.0341	0.0370	7.5
Watson	7	0	0.2804	0.3794	0.0868	0.1046	0.0935	0.0553	9.2
de Villiers	7	0	0.3087	0.4035	0.1153	0.1116	0.0254	0.0356	7.4
Average	10	0	0.2905	0.4000	0.0998	0.1074	0.0432	0.0591	7.7
Watson	10	0	0.2294	0.3950	0.0873	0.0893	0.1139	0.0851	9.7
de Villiers	10	0	0.2604	0.4333	0.1197	0.0982	0.0319	0.0565	7.5
Average	15	2	0.2056	0.4556	0.0944	0.0838	0.0533	0.1072	7.8
Watson	15	2	0.1532	0.4246	0.0779	0.0657	0.1329	0.1456	9.8
de Villiers	15	2	0.1826	0.4889	0.1120	0.0759	0.0390	0.1014	7.5
Average	15	8	0.2200	0.4921	0.0846	0.0836	0.0455	0.0742	7.6
Watson	15	8	0.1687	0.4718	0.0718	0.0675	0.1165	0.1037	9.5
de Villiers	15	8	0.1948	0.5264	0.1002	0.0755	0.0332	0.0700	7.4

Table 3: Batting probabilities (characteristics) for an average batsman, Shane Watson and AB de Villiers at different stages of a match. The quantity E(RR) is the expected run rate for the over where 2/3's are treated as 2's. Note that 3's occur very rarely ($< 1\%$ of the time).

actual runs and the corresponding simulated quantiles are given in Figure 3 for Australia and Zimbabwe. According to Table 1, Australia and Zimbabwe are the highest and the lowest scoring teams respectively. The Q-Q plots suggest that the simulator produces first inning runs that are in line with the actual number of runs scored. Similar plots were obtained for the other ICC teams.

To investigate wicket estimation, Figure 4 provides a plot of the average number of wickets lost versus the number of overs completed for first innngs batting. Figure 4 contains two lines; one based on average wickets lost from actual matches and the other based on averages wickets lost from simulated matches involving randomly chosen batsmen. We observe that the wicket rate increases as the match progresses. There appears to be reasonable agreement between the two lines. This is important because the occurrence of wickets greatly affects run scoring.

To investigate the second innings batting formulation, we considered the lineups used in the 2014 World Cup final between Sri Lanka and India held on April 6. Our simulations give strikingly different probabilities of winning depending on which team bats first. We obtained Prob(SL wins | SL bats first) = 0.46 and Prob(SL wins | India bats first) = 0.61. In contrast, various studies including de Silva & Swartz (1997) and Saikia & Bhattacharjee (2010) have suggested that batting

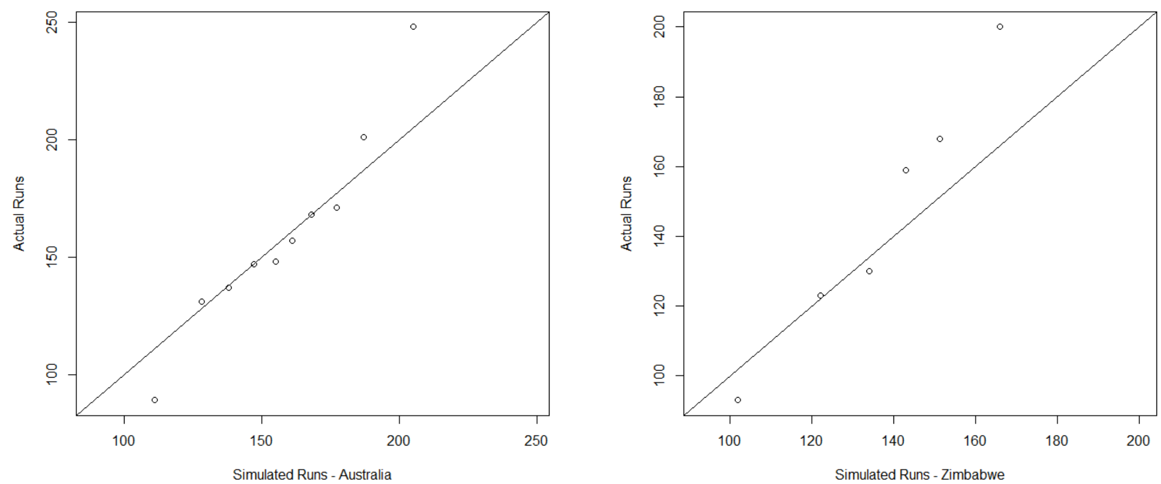

Figure 3: Q-Q plots for Australia and Zimbabwe for first innings runs where the fits appear reasonable.

second confers at most a minor advantage.

How do we reconcile these observations? It seems to us intuitive that batting second should provide a competitive advantage as the team batting second has knowledge of the target and can adjust their batting strategy accordingly. This appears to be the case in Major League Baseball where home teams (which bat in the bottom half of innings) win roughly 54% of their games (Stefani 2008). In second innings simulation, we emphasize that our modified batting characteristics are not unattainable batting characteristics. In fact, they are the characteristics that batsmen display at various stages of a match. It is within their capabilities to modify their characteristics in the manner which we have prescribed. What we posit is that batsmen do not behave in this "optimal" manner. Instead, we believe that batsmen delay increasing their aggressiveness when their team begins falling behind in the second innings. To investigate this, we modify the condition (9) which stipulates an increase in aggressiveness. We adjust the condition for increased aggressiveness by multiplying the right hand side of (9) by the factor 0.8. This states that the team batting second must fall behind an additional 20% before they begin altering their style. In a match with 150 runs, this is essentially saying that a team increases its aggressiveness when it perceives that it is on track to lose by 30 runs. When we introduce the factor 0.8, we

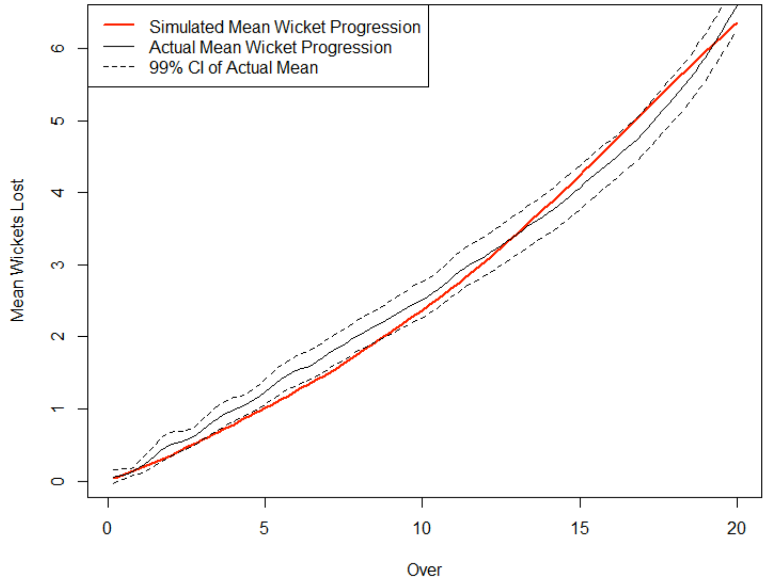

Figure 4: Average number of wickets lost versus overs completed for actual and simulated matches.

obtain Prob(SL wins | SL bats first) = 0.51 and Prob(SL wins | India bats first) = 0.55, and now, the benefit of batting second is much reduced.

The preceding discussion has implications for batting strategy in the second innings. We believe that teams would be better served by increasing their aggressiveness incrementally when they begin falling behind rather than panic at some later stage when it becomes obvious that they are on the verge of losing.

5.1 An Example Concerning the Practical Use of the Simulator

The 2014 World Cup that took place in Bangladesh from March 16 through April 6 provided an interesting application for our methodology.

We considered matches beyond the qualification stage that involved the teams from our dataset. We excluded matches involving Bangladesh since the data collected on Bangladesh (see Table 1) was not as comprehensive. Bangladesh had several "new" players for whom we had little/no data and we did not want to introduce a home team effect for Bangladesh. We note that the Netherlands were the "surprise" team of the tournament as they advanced to the qualification stage at the expense of Zimbabwe. We also did not consider matches involving the Netherlands since we had no data on their past performances.

For a match between Team A and Team B, we simulated 10,000 first innings for each team and calculated the proportion of time that Team A had more runs than Team B. We used this as a proxy for the probability that Team A defeats Team B. Note that sportsbook odds do not take into account which team bats first since this is determined by the coin flip at the beginning of a match. The batting and bowling lineups that we selected in the simulations were the lineups used in the actual matches.

In Table 4, we present the win probabilities from the simulations and the win probabilities implied by sportsbook odds. We see fairly strong agreement between the two sets of probabilities. This is a further endorsement of the realism of the simulator since sportbooks are thought to be "efficient markets" in the sense that sportsbook odds capture all of the available information. One of our observations from the exercise is that the inclusion/exclusion of key players in the lineup can have a meaningful impact on the probabilities. We also note that relative to the sportsbook, our winning probabilities for Pakistan were considerably higher. We believe that this was partly due to the inclusion of Zulifiqar Babar and Biliawal Bhatti into the lineups as relatively new bowlers. Whereas the sportbook discounted their abilities, our model provided them with performance characteristics that were in line with average performance. Pakistan also did badly in some of their T20 matches leading up to the World Cup, matches for which we did not collect data. We also note that sportsbook odds are dynamic and sometimes the odds can change by several percent in the hours leading up to a match.

To investigate the simulator further, and possibly assess whether its output is superior to the sportsbook odds, we wagered a hypothetical \$100 on each of the 15 matches from Table 4. The team that we wagered on was the team whose simulated probabilities exceeded the implied sportsbook probabilities. The \$100 was wagered at the odds corresponding to the sportsbook. The net result of this exercise was a hypothetical profit of \$399 where 9 of the 15 winning teams were chosen correctly. Of course, this is too small a sample of matches to guarantee long run profitability.

				Prob(Team A wins)	
Date	Team A	Team B	Winner	Simulator	Sportsbook
March 21	India	Pakistan	India	0.54	0.59
March 22	Sri Lanka	South Africa	Sri Lanka	0.44	0.48
March 22	England	New Zealand	New Zealand	0.47	0.44
March 23	Pakistan	Australia	Pakistan	0.51	0.35
March 23	West Indies	India	India	0.39	0.45
March 24	New Zealand	South Africa	South Africa	0.41	0.42
March 27	England	Sri Lanka	England	0.36	0.38
March 28	West Indies	Australia	West Indies	0.40	0.37
March 29	England	South Africa	South Africa	0.42	0.44
March 30	India	Australia	India	0.52	0.48
March 31	Sri Lanka	New Zealand	Sri Lanka	0.69	0.59
April 1	West Indies	Pakistan	West Indies	0.33	0.49
April 3 (Semifinal)	Sri Lanka	West Indies	Sri Lanka	0.64	0.53
April 4 (Semifinal)	India	South Africa	India	0.55	0.57
April 6 (Final)	Sri Lanka	India	Sri Lanka	0.53	0.42

Table 4: Win probabilities for specified 2014 World Cup matches beyond the qualification stage.

6 DISCUSSION

In the development of our simulator, batting outcome probabilities are dependent on the batsman, the bowler, the number of overs consumed, the number of wickets lost, the home team advantage and the target score (in the case of the second innings). Whereas the proposed model is complex and captures the essential features of T20 cricket, there is no doubt that there are other variables that may influence batting performance. For example, the fielding quality of the opposing team affects run scoring. Also, if various players are in particularly good or poor form, one may consider tinkering with their characteristics. As discussed at the end of section 3, one way to accomplish this may involve a weighted estimation scheme where more weight is given to recent performances. The implementation of these sorts of ideas is something that may be considered in future research.

One of the interesting by-products of our work is that we have posited that teams are not batting optimally in the second innings. We suggest that teams are not incrementally increasing their aggressiveness when they begin falling behind. Instead, we believe that they wait until the situation becomes dire, and only then, increase their aggressiveness. Although it may be difficult to train batsmen to increase their aggressiveness incrementally in the prescribed fashion, we see an opportunity where players can move somewhat in this direction. This change of strategy could provide a significant benefit to teams.

We believe that the modeling of batting behaviour and the subsequent development of the simulator are important steps in gaining a deeper understanding of strategic aspects related to T20 cricket. For example, with a realistic simulator, it may be possible to determine player worth and to investigate optimal team selection and optimal batting orders. These are topics which we plan to pursue in future work. We also understand that in-game cricket forecasting is a difficult problem which has applications to wagering. The methodology of section 4.3 may be useful in this regard. We therefore see this paper as seminal work in the advancement of T20 cricket analytics.

7 REFERENCES

de Silva, B.M. & Swartz, T.B. (1997). Winning the coin flip and the home team advantage in one-day international cricket matches. *New Zealand Statistician*, 32, 16-22.

de Silva, B.M., Pond, G.R. & Swartz, T.B. (2001). Estimating the magnitude of victory in one-day cricket. *The Australian and New Zealand Journal of Statistics*, 43, 259-268.

Elderton, W.E. (1945). Cricket scores and some skew correlation distributions. *Journal of the Royal Statistical Society, Series A*, 108, 1-11.

Bailey, M.J. & Clarke, S.R. (2006). Predicting the match outcome in one day international cricket matches while the match is in progress. *Journal of Science and Sports Medicine*, 5, 480-487.

Davison, A.C. (2003). *Statistical Models*, Cambridge: Cambridge University Press.

Duckworth, F.C. & Lewis, A.J. (2004). A successful operational research intervention in one-day cricket. *Journal of the Operational Research Society*, 55, 749-759.

Dyte, D. (1998). Constructing a plausible test cricket simulation using available real world data. In *Mathematics and Computers in Sport*, N. de Mestre and K. Kumar, editors, Bond University, Queensland, Australia, 153-159.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (editors) (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227, 3431-3444.

Perera, H., Gill, P.S. & Swartz, T.B. (2014). Declaration guidelines in test cricket. *Journal of Quantitative Analysis in Sports*, 10, To appear.

Saikia, H. & Bhattacharjee, D. (2010). On the effect of home team advantage and winning the toss in the outcome of T20 international cricket matches. *Assam University Journal of Science and Technology*, 6, 88-93.

Scarf, P., Shi, X. & Akhtar, S. (2011). On the distribution of runs scored and batting strategy in test cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174: 471-497.

Stefani, R. (2008). Measurement and interpretation of home advantage. In *Statistical Thinking in Sports*, J. Albert and R.H. Koning, editors, Chapmana & Hall/CRC: Boca Raton.

Swartz, T.B. & Arce, A. (2014). New insights involving the home team advantage. *International Journal of Sports Science and Coaching*, 9, To appear.

Swartz, T.B., Gill, P.S. & Muthukumarana, S. (2009). Modelling and simulation for one-day cricket. *The Canadian Journal of Statistics*, 37, 143-160.

Wood, G.H. (1945). Cricket scores and geometrical progression. *Journal of the Royal Statistical Society, Series A*, 108, 12-22.

8 APPENDIX

Recall that the multinomial model (2) is highly parametrized where the data are sparse and even nonexistent over regions of the parameter space. The simplifying assumption (3) leads to a more tractable model where the parameters p_{i70j} and τ_{owj} are estimated in two steps. In section 3, we described a hierarchical model where a Bayesian approach was taken to estimate the p_{i70j}. A key component of the approach was the recognition of similar batting characterstics amongst players. Here, in the Appendix, we describe the estimation of τ_{owj}; the parameters used to describe the modification of batting characteristics with respect to the stage of the match (i.e. overs consumed and wickets taken).

Let x_{iowj} denote the number of occurrences of outcome j by batsman i for all batting attempts in the oth over with w wickets taken. The corresponding empirical probability is $\hat{p}_{iowj} = x_{iowj}/n_{iow}$ where $n_{iow} = \sum_j x_{iowj}$.

Next, we define the transition factor $\tilde{\alpha}_{iowj} = \hat{p}_{io'wj}/\hat{p}_{iowj}$ which represents the change in empirical probabilities for batsman i when going from the stage of the match (o, w) to the adjacent stage $(o', w) = (o + 1, w)$ corresponding to the next over. We then average the transition factors over all batsmen giving

$$\hat{\alpha}_{owj} = \frac{\sum_i v_{iowj}^{-1/2} \ \tilde{\alpha}_{iowj}}{\sum_i v_{iowj}^{-1/2}} \tag{11}$$

where the Delta Theorem is used to obtain the variance expressions for ratios

$$v_{iowj} = \tilde{\alpha}_{iowj}^2 \left(\frac{1 - \hat{p}_{io'wj}}{n_{io'w}\hat{p}_{io'wj}} + \frac{1 - \hat{p}_{iowj}}{n_{iow}\hat{p}_{iowj}} \right) \ .$$

We can therefore view the estimates $\hat{\alpha}_{owj}$ as forming a matrix with the rows corresponding to overs ($o = 1, \ldots, 20$) and the columns correponding to wickets ($w = 0, \ldots, 9$). For any stage (o, w) of a match, the matrix entry $\hat{\alpha}_{owj}$ is the transition factor for changing the probability p_{iowj} to the probability $p_{io'wj}$ for any batsman i. With respect to the matrix, the movement corresponds to going down column w from row o to row $o' = o + 1$. We smooth the matrix to improve the estimates.

Analogous to (11), transition factors $\hat{\beta}_{owj}$ can be defined when going from the stage of the match (o, w) to the adjacent stage $(o, w') = (o, w + 1)$ corresponding to the next wicket. We then have a second matrix where $\hat{\beta}_{owj}$ describes the movement along row o from column w to column $w' = w + 1$.

Finally, to obtain the parameter τ_{owj}, we recall that τ_{owj} is the multiplier that is used to modify the baseline probability p_{i70j} in (3) to the probability p_{iowj}. We obtain τ_{owj} by taking the straight line from the matrix position from the start of the innings $(o = 1, w = 0)$ to (o, w) and use the nearest transition factors $\hat{\alpha}$ and $\hat{\beta}$ as multipliers.

We remark that the proposed estimation procedure for τ_{owj} is based on incremental changes to overs and wickets. It is not possible to estimate directly from the baseline state $(o = 7, w = 0)$ to a distant stage (o, w) since there are very few (if any) batsmen who have batted in both stages. However, by approaching the estimation incrementally, we have common batsmen who bat in adjacent stages.