FORECASTING MLB PERFORMANCE UTILIZING A BAYESIAN APPROACH IN ORDER TO OPTIMIZE A FANTASY BASEBALL DRAFT

A Dissertation
Presented to the Faculty of
Claremont Graduate University
and
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Computational Science - Statistics

by
Daniel Luke Herrlin
Fall 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

APPROVAL OF THE REVIEW COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Daniel Luke Herrlin as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy.

Richard Levine, Chair
Department of Mathematics and Statistics, San Diego State University
Department Chair

Joey Lin
Department of Mathematics and Statistics, San Diego State University
Professor

Barbara Bailey
Department of Mathematics and Statistics, San Diego State University
Associate Professor

John Angus
Institute of Mathematical Sciences, Claremont Graduate University
Professor

Allon Percus
Institute of Mathematical Sciences, Claremont Graduate University
Associate Professor

ABSTRACT OF THE THESIS

FORECASTING MLB PERFORMANCE UTILIZING A BAYESIAN APPROACH IN ORDER TO OPTIMIZE A FANTASY BASEBALL DRAFT
by
Daniel Luke Herrlin
Doctor of Philosophy in Computational Science - Statistics
Claremont Graduate University and San Diego State University, 2015

Fantasy baseball has been increasing in popularity dramatically over the past decade. The game begins with participants selecting a team through a fantasy draft at the beginning of the season and then tracking their players' statistics, compared to those of their competitors, throughout the season. Selecting the players who will perform the best in the upcoming season is the goal at the start of the year, and there are many different rankings and algorithms devoted to assisting a participant in creating their team. This dissertation will focus on predicting the outcomes of the upcoming season and propose a selection algorithm based on the evaluations in order to optimize a participant's fantasy draft.

While the vast majority of fantasy baseball rankings do not disclose any analytical rigor behind them, this approach will focus on the methodology utilized to forecast player statistics throughout the upcoming season. A Bayesian approach will be utilized in conjunction with nonlinear growth curves and nonparametric regression tree approaches in order to predict future outcomes.

DEDICATION

Dedicated to my beautiful and loving wife Trista and sons Tealson and Wakelon. Trista has been an incredible support and inspiration both to pursue my degree, to keep up the pursuit, and to bring it to completion. Without her support and encouragement this would not have been possible. Tealson and Wakelon have been an incredible inspiration and motivation to me to be the best man that I can be.

ACKNOWLEDGEMENTS

First, I would like to thank my committee chair Dr. Rich Levine. He has been both a help and inspiration since day 1. He gave me direction and encouragement that I could turn this topic into a dissertation, and has been a great help all along the way.

Additionally, I would like to thank my committee members Dr. Barbara Bailey, Dr. John Angus, Dr. Joey Lin and Dr. Allon Percus for their assistance and insights over the years. Each has proven to be a valuable resource both in the courses taken and in research assistance.

Finally, I would like to thank Dr. Jose Castillo, Parissa Plant, and the Computational Science Research Center in providing the resources: faculty, administrative, and funding, associated with the joint SDSU/CGU Ph.D. program.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

CHAPTER 1
INTRODUCTION

Baseball has been a field of increased visibility in statistical analysis over the past decade. Major League Baseball teams are also beginning to focus on more analytical approaches in their scouting and player development as well as other areas. This revolution began in the 1990's when Billy Beane came to the fore as the Oakland Athletics general manager who used an advanced analytical approach to field a competitive team despite the teams disadvantages in revenue [37], and has spread across many other major league teams. While it is spreading and growing in popularity there is hesitancy to use it in certain aspects of the game. For example, the San Diego Padres were one of the early adopters of this increasingly analytical approach, yet they hired only one analyst, who is no longer with the team, and are not interested in using analytics in day to day team operations such as lineup selection or optimization.

Fantasy sports, a game in which players construct their own virtual teams and utilize their player's live statistics to score the game, has increased dramatically in popularity over the past decade. The premise of the game relies upon a players ability to:

- Predict future performance of players
- Strategize team composition based on league rules
- Select players that will produce better statistics than their opponents

There are many sites available that rank players and provide forecasts for each player for the upcoming season. Most of the rankings are static and assume a rigid set of rules for the fantasy games that do not apply to most leagues outside of the website that is providing them. In fact many of the leagues within the websites (espn.com, yahoo.com, and cbssportsline.com are among the most popular) are based on custom rules which means that the rankings are not necessarily a good metric even if the underlying assumptions about player performance are accurate.

This dissertation will seek to address those issues as well as the more fundamental one of accurately predicting player performance, or a reasonable baseline of expectation. In order to assist the reader in following the methodology in this dissertation a flowchart outlines the methodology in Figure 1.1. Since the location of the batter in the order is also relevant, we will discuss batting order, and how it can be optimized first in Chapter 2. Chapter 3 will begin by describing the batter production and the statistics that will be utilized. Following will be the methodology by which production will be evaluated and modeled to produce a reasonable set of expectations along with parameters to assess the volatility of the estimates. Chapter 4 will be analogous to Chapter 3 but will evaluate pitcher production. Chapter A will outline the simulation methodology which will utilize the production probabilities outlined in the previous two chapters. Chapter 5 will evaluate the results of these models and compare them to the best industry results that are available. These results will then be used to develop an algorithm to provide the best set of players possible in a fantasy draft. Chapter 6 will discuss the contributions that this research provides as well as possible extensions and future research.

Fantasy baseball is broken up into two main categories: batter performance and pitcher performance, and this dissertation will be broken up in the same manner.

1.1 Summary Statistics Definitions

- run: When the offensive player crosses home plate and scores a run.
- home run: When the batter hits the ball and scores on the play, without an error.
- run batted in: When a run scores due to the batters hit, including himself when a home run is hit.
- stolen base: When a runner advances a base without the batter hitting the ball, or an error being credited to one of the fielders.
- caught stealing: When the runner attempts to advance a base without the batter hitting the ball, but is tagged out by a fielder.

- plate appearance: When a batter comes up to bat and an outcome for the batter is recorded (either a hit, out, walk, or other event).

- at bat: When a plate appearance results in either a hit or out.

- batting average: Proportion of at bats where a hit is recorded.

- on base percentage: Proportion of plate appearances where the batter reaches base.

- slugging percentage: Ratio of bases reached based on batters hits to at bats.

- isolated power: Slugging percentage less batting average.

- inning pitched: Number of outs recorded while the pitcher is in the game, divided by 3.

- strikeout: Attributed to both batters and hitters, occurs when the batter is out on strikes.

- walk: Attributed to both batters and hitters, occurs when the batter reaches base on balls.

- earned run: A run that scores, and is not attributed to an error in the field. This is attributed to the pitcher in the game when the runner reached base.

- ERA: Earned Run Average is the number of earned runs attributed to a pitcher divided by the number of innings pitched and multiplied by 9.

- WHIP: Walks plus Hits per Inning Pitched adds walks and hits allowed by a pitcher divided by the number of innings pitched.

- quality start: Occurs when the starting pitcher pitches at least six innings in a game and allows three or fewer earned runs.

- win: Attributed to the pitcher who is in the game for the winning team at the last lead change of the game. If the win would be credited to the starting pitcher but they did not pitch five full innings, then the first relief pitcher is credited with the win.

- loss: Attributed to the pitcher who is in the game for the losing team at the last lead change of the game.

- save: Attributed to the last pitcher who pitches for the winning team if they do not get credit for the win and: the team is winning by three or fewer runs when the pitcher comes into the game or the pitcher pitches at least three innings.

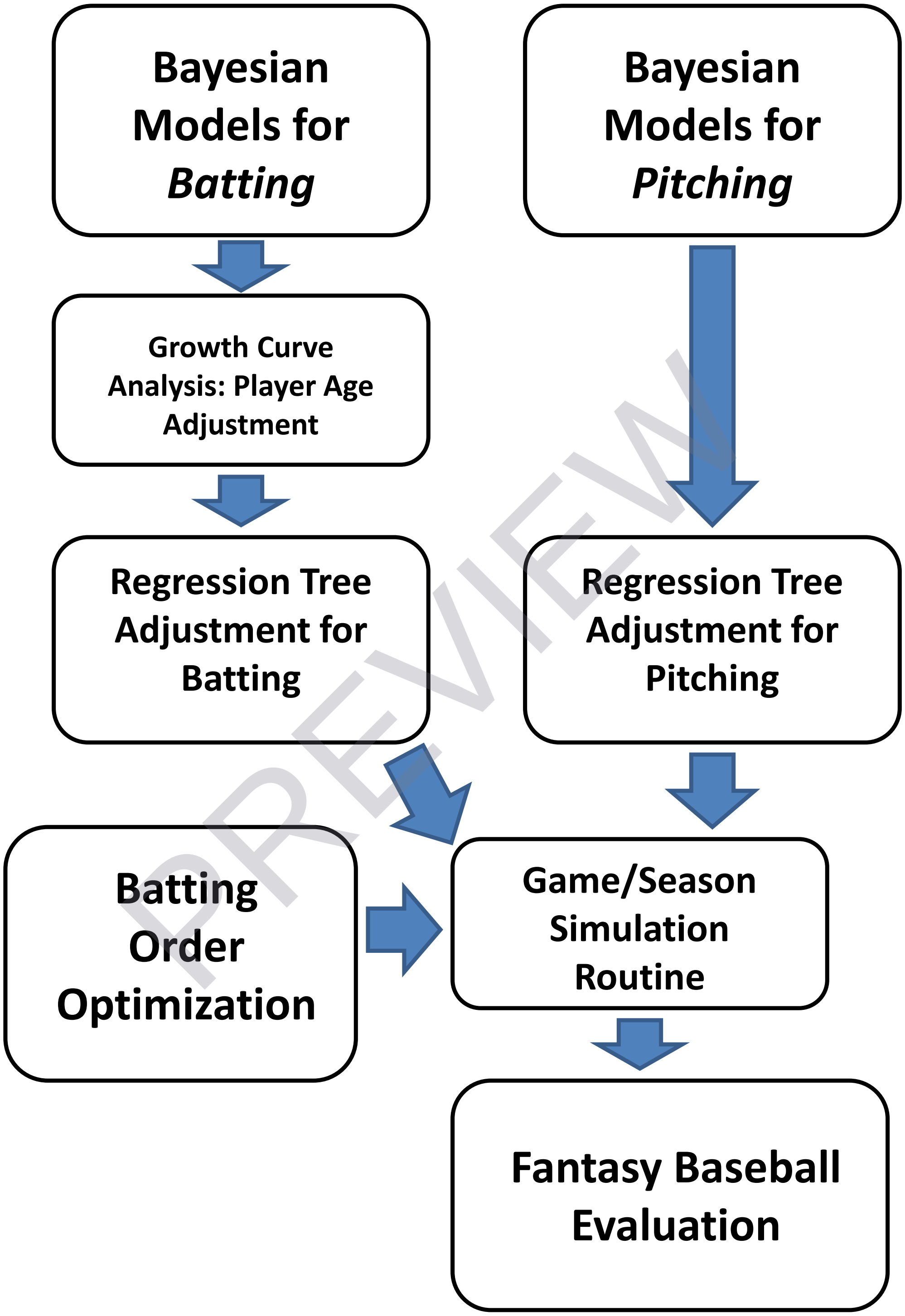
Bayesian Models for *Batting*
Bayesian Models for *Pitching*
Growth Curve Analysis: Player Age Adjustment
Regression Tree Adjustment for Batting
Regression Tree Adjustment for Pitching
Batting Order Optimization
Game/Season Simulation Routine
Fantasy Baseball Evaluation

Figure 1.1. Flowchart of dissertation methods

CHAPTER 2

BATTING ORDER OPTIMIZATION

2.1 INTRODUCTION

Baseball teams are faced with a substantial question before each game, what order should their batters hit in the lineup? Traditionally baseball teams have left this question up to their managers to decide, who typically use one of two approaches: the lineup that they always use so as not to throw their hitters out of rhythm, or their “gut feeling” comprising of which players they believe will hit better in which lineup positions on a given evening (or afternoon). This dissertation will provide a more analytical approach to the concept of where players should bat in a lineup (and to a lesser extent, who should be included) via a Markovian method as shown in Figure 2.1.

Markov chain methods are a natural approach for run production in baseball as a given sequence of events (namely the batters batting) occurs in repetition. The method requires that probabilities be used to facilitate the transition between different states in the game. These probabilities are easily obtained from readily available player batting data. Each of these probabilities are unique to individual players, but since the batting order cycles through and does not change (with the exception of in-game lineup changes) a Markov method is a natural fit.

There have been a number of papers attempting to answer this same question using a Markov chain method: Howard [32], Cook [19], Thorn and Palmer [58], Pankin [46], Stern [55], Bukiet et al. [16], Takei et al. [57], Sokol [54], and Nobuyoshi [42]. These papers have all used the same Markov chain approach which will be extended in this dissertation. We will briefly discuss the primary contributions.

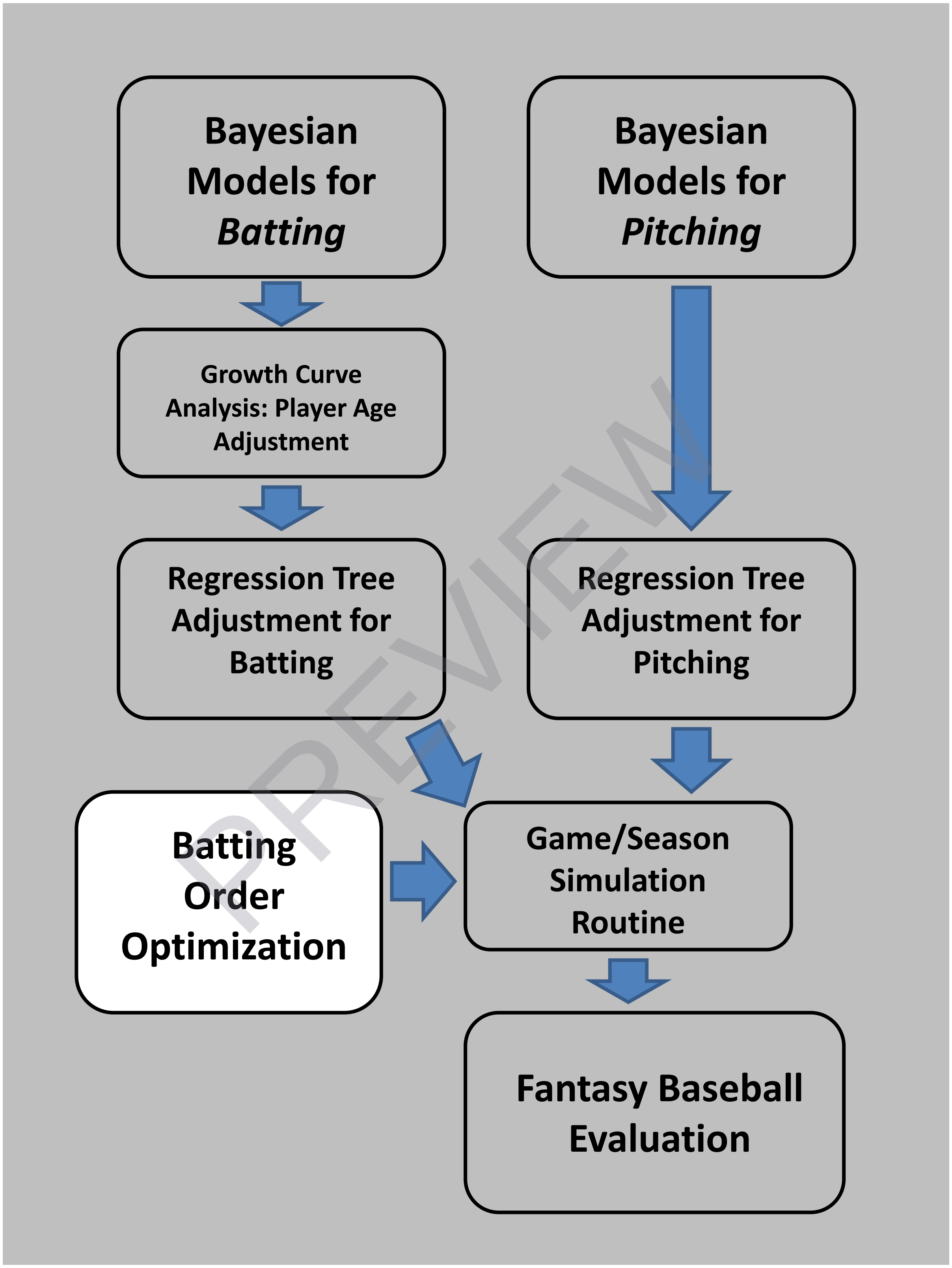
Bayesian Models for *Batting*
Bayesian Models for *Pitching*
Growth Curve Analysis: Player Age Adjustment
Regression Tree Adjustment for Batting
Regression Tree Adjustment for Pitching
Batting Order Optimization
Game/Season Simulation Routine
Fantasy Baseball Evaluation

Figure 2.1. Flowchart of dissertation methods

Pankin [46] was the first to study many possible orders, testing 200 possible rotations and evaluating the 9 lineups (with each batter in the rotation as the leadoff batter) associated with each rotation. While he could make no claim as to the absolute accuracy of his lineups, he was able to compare them with the lineups that were currently being used and show that his best lineups were an improvement. Much of the data that we use today was not readily available at the time of Pankin [46], such as double play rates and runner advancement probabilities. The paper also looked at different statistical metrics and associated their importance with different positions in the order. He found that on base percentage was most important for the leadoff batter, while both slugging percentage and on base percentage were very important for the batters hitting second, third, and fourth. Many of these conclusions will be re-affirmed in this dissertation.

Bukiet et al. [16] was able to calculate optimal orders using an exhaustive search (which took 5.5 days per team in 1997) and establish batter positioning rules so that a near optimal order could be produced while testing fewer than 1000 different lineups. The exhaustive search employed today still takes many hours to perform even on very powerful machines, thereby rendering that methodology impractical in practice since the employment of this process by Major League Baseball (MLB) would mean testing different potential sets of players for each game. There are significant limitations with the paper's results, however, since they did not consider many of the things that MLB managers consider very important when determining their lineup, e.g. stolen bases, base running ability, runner advancement, and double plays. These factors will need to be considered for the result to be applicable to and accepted by MLB teams.

Takei et al. [57] extended the methodology of Bukiet et al. [16] to Japanese League baseball. They also included double plays, stolen bases, and batting average with runners in scoring position. The issues with baserunning remain a concern in this approach, however, in that Takei et al. [57] assumed that all base runners will advance two bases on a single and all baserunners will score on a double, when some of these actions happen less than fifty percent of the time in the American game (first to third on a single, for example). This approach also assumes that all ground balls result in double plays and automatically places the catcher in the eighth spot and the pitcher in the ninth spot of the batting order. As we will see later, batting the pitcher ninth is sub-optimal in nearly all situations in Major League Baseball, and catcher production often suggests a higher placement in the lineup.

Sokol [54] produced the most actionable results as he attempted to consider all of the factors that are relevant to manager decision-making. He used a runner advancement approach suggested by Pankin [46] which groups the runners into three categories in order to determine their probability of advancement. This approach is moving in the right direction, but using play-by-play data, the specific likelihood for each player to advance from first to third on a single, or score from first on a double, is far more precise. Another drawback of the Sokol [54] approach is that the players were categorized and then an order was estimated, creating a similar looking order for all MLB teams. Nonetheless, Sokol [54] demonstrated that all batting orders display some level of robustness under uncertainty: while the optimal order is far from robust, near optimal orders are robust to variations in ability. Baumer [13] estimated the maximum impact of baserunning to be 70 runs over the course of a 162 game season. In this dissertation we will examine the differences in optimal or near optimal orders based on team composition and evaluate what types of players are more valuable to different teams. Additionally, we will account for each individual runner's ability to advance the extra base, as well as the batter's ability to advance runners when they themselves do not reach base. These two factors are prevalent in Major League Baseball and prove very important in demonstrating that traditional orders are not near optimal.

As part of our methods development for order optimization, we propose a greedy algorithm for identifying new near optimal lineups under our Markovian methodology. This approach generally finds a near optimal order, and when it is implemented across a number of randomly chosen starting lineups, it identified the optimal lineup for all teams tested.

2.2 METHODOLOGY

This dissertation will use methodology along the lines of Bukiet et al. [16], but will involve a much more robust approach and unique optimization algorithms. The first step of the Markov process is to establish transition matrices.

2.2.1 Transition Matrix

The Markov process begins by assessing the different states of the game. This process is effective in baseball because there are a relatively small number of unique states in the game, as opposed to football where every down, distance, and yard line would be a different state. In baseball we can break the game down into half-inning increments. During each inning both teams have an opportunity to bat and score runs, a half-inning is one teams' opportunity up to bat. Baseball can be limited to 25 states: since there are three bases that can be occupied at any time, there are 2^3 or 8 possible states for runners on base. When that is combined with the three possible states for the number of outs (0, 1 or 2), 24 possible states are generated. The final state is the one that results in the end of the half-inning, or the three out state. At the end of any plate appearance the game will be in one of these 25 states.

The matrix T (Table 2.1) is the block transition matrix which stores the probabilities of a players' at-bat resulting in the transition from one state to another similar to the tables displayed in Bukiet et al. [16]. The rows and columns that represent 0, 1, and 2 outs are 8×8 block matrices (upper left 3×3 block in Table 2.1), and the row representing three outs has three 1×8 row vectors of zeros followed by a scalar value of 1. The column representing three outs (column 4) has three 8×1 column vectors. DP represents the players' probability of hitting into a double play in each of the 8 one out states. Pout is the probability that the player will record an out in each of the 8 two out states (no runners will score). Note that this matrix is upper triangular since it is not possible to reduce the number of outs at any point during a half-inning. This block matrix assumes independence between plate appearances and the number of outs in the inning. It is easy to change this assumption, although there is no evidence that a significant relationship exists between number of outs and a players' ability to reach base safely or perform other specific actions such as hitting a home run and the number of outs in an inning. This notion was dispelled by Barra and Schwartz [11]. While not addressed in this dissertation, one set of outcomes that may be out dependent is runner advancement. It is likely that baserunners would be more likely to advance an extra base with two outs on a fly ball as they are not required to ensure that the ball is not caught before advancing.

States	No outs	1 out	2 outs	3 outs
No outs	$A_{8\times 8}$	$B1_{8\times 8}$	$B2_{8\times 8}$	$0_{8\times 1}$
1 out	$0_{8\times 8}$	$A_{8\times 8}$	$B1_{8\times 8}$	$DP_{8\times 1}$
2 outs	$0_{8\times 8}$	$0_{8\times 8}$	$A_{8\times 8}$	$Pout_{8\times 1}$
3 outs	$0_{1\times 8}$	$0_{1\times 8}$	$0_{1\times 8}$	$1_{1\times 1}$

Table 2.1. Full transition matrix, T. Key: 0 is a zero-vector, DP = double play, Pout = probability out recorded; A, $B1$, and $B2$ defined in Tables 2.2-2.4.

The 8×8 block A (Table 2.1) represents the player reaching base safely without recording an out, see Table 2.2 for the transitions. This basic transition matrix operates under the assumption that each runner on base will advance the same number of bases as the batter (with the notable exception of a walk where the runners advance only if they are forced). The B1 (Table 2.1) blocks represent the player generating an out, broken down in Table 2.3. In the basic transition matrix this would be a diagonal matrix with no runners advancing when an out is recorded. In our robust approach, the batter's ability to advance the runners an extra base, while not reaching base themselves, is accounted for as well as the batter's probability of hitting into double plays. Table 2.4 represents the players' propensity for generating two outs in a single plate appearance (double plays) with no outs in the inning. Moving from the zero out state to the three out state (or players hitting into a triple play) is not differentiable from zero, thus zero is the probability assumed for all batters.

This Markov chain approach does not account for who is on base. Since the baserunner is never specified they are probabilistically determined by the bases that are