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

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: Thank you for everything throughout this process. This degree is really a joint degree that we earned together, as I could never have come close to completing this without you. I appreciated your playful nudges as you would ask me how things were coming along. The graduation gift of skydiving with my brother was truly a testament to the support and encouragement that you have been. It took planning and scheming over the course of a year, and of course due to my continual unbridled optimism as it relates to my completion of this degree, was planned and executed four years prior to the actual completion. I know how self sacrificial that was for you, as it was something that you knew I really wanted to do, but it put far more butterflies in your stomach than mine. I'm sure there was a great deal of relief in your heart when the parachute finally opened. Who else can say that their wife gave them a graduation present that was purely sacrificial in nature, and unintentionally more than four years early!

As life stops for no one, we have had our ups and downs over the past ten years as I pursued this degree. The birth of our two amazing boys were certainly highlights, and the length of the process was a struggle for both of us, though you rarely ever showed it. There may have been times when you had had enough, secretly wishing this journey to be over, but you never communicated that to me. While there were certainly times where you would have preferred me to be doing something other than working on this project and completing this degree, you have always valued this accomplishment for me and graciously endured the long hours and early mornings that this project required. You have been an incredible support and

encouragement through it all, and I am a lucky man to be able to call you my wife.

To my sons Tealson and Wakelon: When I look at you boys I am reminded of the weight of my actions. You are both constantly looking up to me and watching me, I hope that I am able to be an example of a man of integrity, worthy to be looked up to. There are many things in my life that I hope you do not emulate, but I look forward to continuing to lead you and our family, helping to mold you into the men that God wants you to be.

One example of this emulation came when both of you requested to play fantasy football this year. Not many seven year old children request to play fantasy sports, and for that I am proud. I look forward to seeing you both leverage this work, and potentially build upon it, to assist your fantasy teams for years to come.

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

PAGE

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.

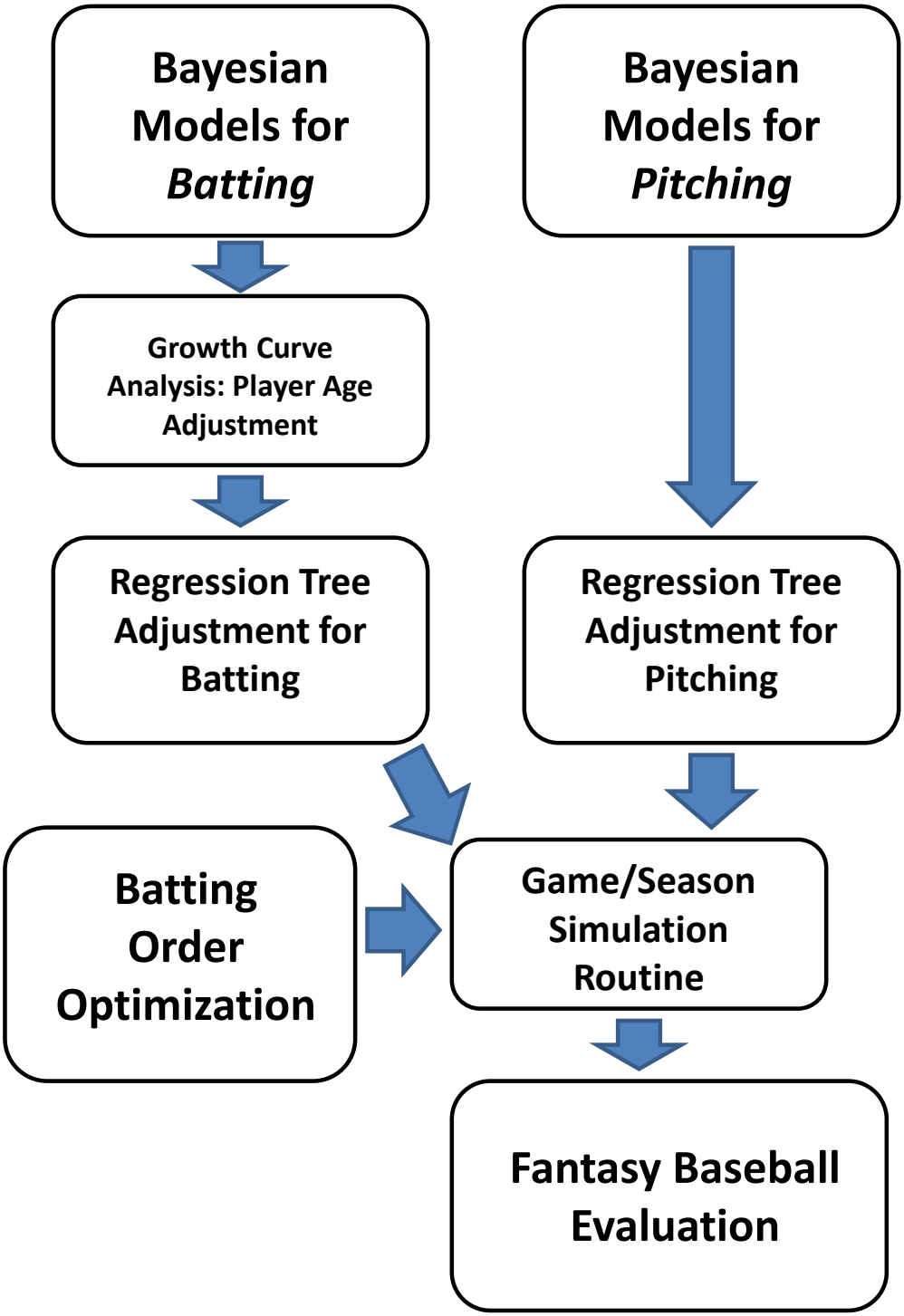

Figure 1.1. Flowchart of dissertation methods

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.

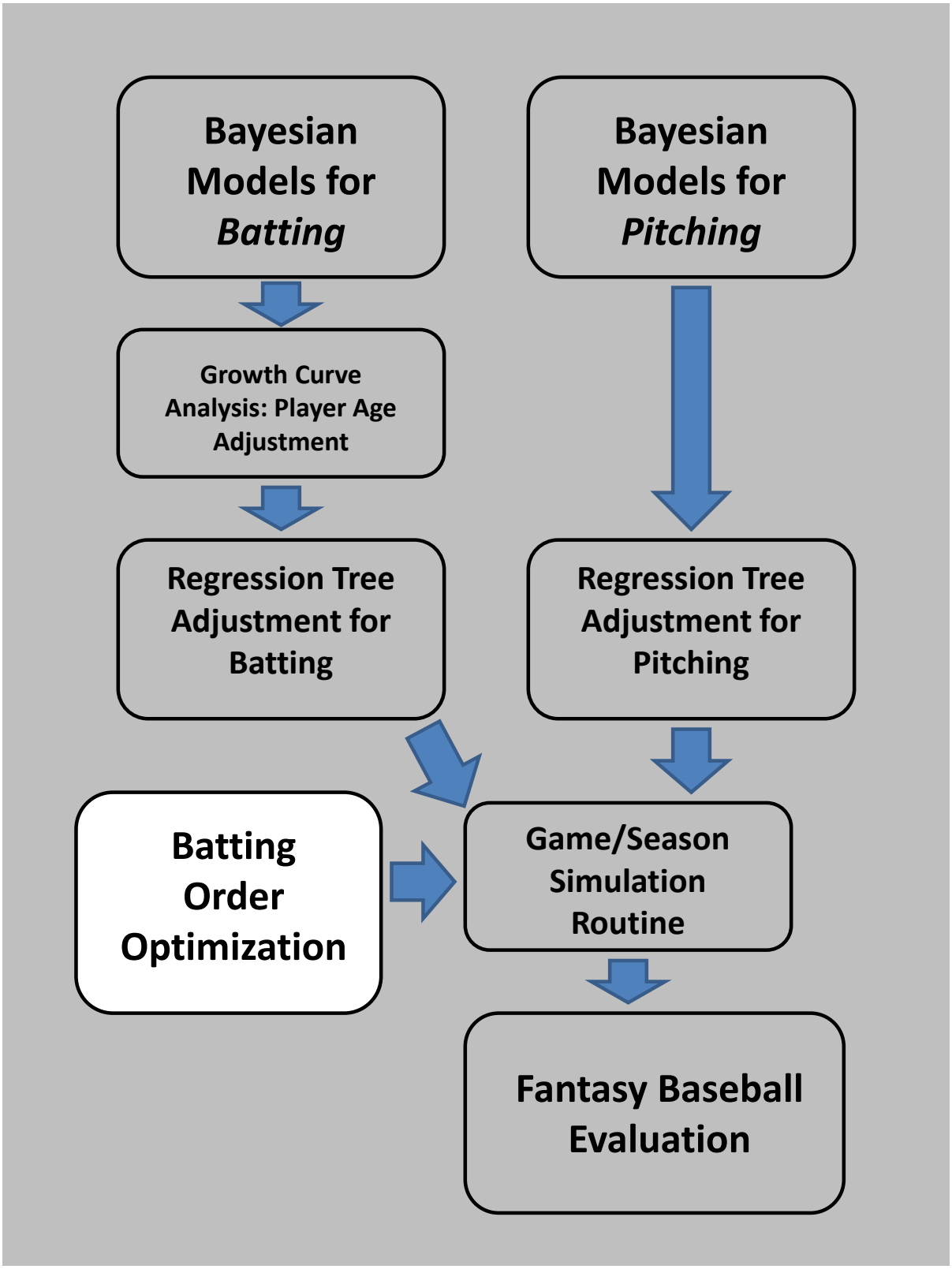

Figure 2.1. Flowchart of dissertation methods

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	$A_{8\times8}$	$B1_{8\times8}$	$B2_{8\times8}$	$0_{8\times1}$
1 out	$0_{8\times8}$	$A_{8\times8}$	$B1_{8\times8}$	$DP_{8\times1}$
2 outs	$0_{8\times8}$	$0_{8\times8}$	$A_{8\times8}$	$Pout_{8\times1}$
3 outs	$0_{1\times8}$	$0_{1\times8}$	$0_{1\times8}$	$1_{1\times1}$

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

A	Empty	1B	2B	3B	1B, 2B	1B, 3B	2B, 3B	Loaded
Empty	HR	BB+1B	2B	3B	0	0	0	0
1B	HR	0	0	3B	BB+1B	0	2B	0
2B	HR	0	2B	3B	BB	1B	0	0
3B	HR	1B	2B	3B	0	BB	0	0
1B, 2B	HR	0	0	3B	0	0	2B	BB+1B
1B, 3B	HR	0	0	3B	1B	0	2B	BB
2B, 3B	HR	0	2B	3B	0	1B	0	BB
Loaded	HR	0	0	3B	0	0	2B	BB+1B

Table 2.2. No outs generated by the plate appearance. Key: BB = Walk, 1B = Single, 2B = Double, 3B = Triple, HR = Home Run.

B1	Empty	1B	2B	3B	1B, 2B	1B, 3B	2B, 3B	Loaded
Empty	OUT	0	0	0	0	0	0	0
1B	0	OUT(1-DP-S1)	S1*OUT	0	0	0	0	0
2B	0	0	OUT(1-S2)	S2*OUT	0	0	0	0
3B	0	0	0	OUT	0	0	0	0
1B, 2B	0	0	0	0	OUT(1-DP-SA)	0	OUT*SA	0
1B, 3B	0	OUT*SF	0	0	0	OUT(1-DP-SF)	0	0
2B, 3B	0	0	OUT*SF	0	0	0	OUT(1-SF)	0
Loaded	0	0	0	0	OUT*SF	0	0	OUT(1-SF)

Table 2.3. One out generated by the plate appearance. Key: DP = Double Play, S1 = Sacrifice a player from first base, S2 = Sacrifice a player from second base, SF = Sacrifice which scores a player from third base, SA = Sacrifice advancing two baserunners, * = Multiplication.

occupied, who is up to bat, and how many outs have been made in that inning. For example, if there is a runner on second base with no outs and the third batter is up to bat, the second batter must be on second base. In fact if there are no outs, since the base paths represent a queue, we know exactly who is on which base at all times. If there were one out in the prior situation then the runner on second base could have been either the first or second batters. The first batter would be on first base if the second batter recorded an out, and the second batter would be on first base if either the first batter had recorded an out, or the first batter reached base safely, and the second batter hit into a fielder's choice, where the first batter was put out advancing to second base, but the second batter reached safely. Any other batters that had come to bat in that inning must have scored based on the nature of the queue. With two outs the runner could have been the first, second, or ninth batters based on the prior logic where

B2	Empty	1B	2B	3B	1B, 2B	1B, 3B	2B, 3B	Loaded
Empty	0	0	0	0	0	0	0	0
1B	OUT*DP	0	0	0	0	0	0	0
2B	0	0	0	0	0	0	0	0
3B	0	0	0	0	0	0	0	0
1B, 2B	0	0	0	OUT*DP	0	0	0	0
1B, 3B	OUT*DP	0	0	0	0	0	0	0
2B, 3B	0	0	0	0	0	0	0	0
Loaded	0	0	0	OUT*DP	0	0	0	0

Table 2.4. Two outs generated by the plate appearance.

two of those three batters have been put out while all other batters that had batted prior in that inning must have scored. This development is important when considering whether a runner will score if the batter hits a single. To account for this the initial transition matrices for each player must be changed based on who is hitting before them. Conditional probabilities must be determined, e.g., P(first batter is on first|one out and second batter batting with a runner on first). This creates 216 conditional probabilities, only 54 of which are non-zero based on the prior criteria, for every base runner that is on base. These probabilities are distinct for a lone runner on base, two runners on base, and with the bases loaded. Which base the runner is occupying is of no direct consequence here, only their position in the base path queue. The ultimate result is 324 distinct conditional probabilities. In this application it will be assumed that each runner with a non-zero probability to be any given baserunner has an equal probability of occupying that base. One fairly simple adjustment could be to utilize players on base percentages as their comparative probability of occupying any given base. This was tested and had no significant differences either in the runs per game or in the optimal orders. Further research is required to establish accurate distinct probabilities for each baserunner since the number of outs and base that the runner is on clearly has a significant impact on which runner is more likely to occupy the base.

The ability of a player on second to score on a single has to do not only with the runner's ability but also with where, and how hard, the ball was hit. In this model the batter

will not factor into the ability of the runner to advance an extra base. We will assume that the distribution of strength and location of balls hit for a single will remain constant. While this is not the case in reality, the runner on base typically plays a much larger role in his ability to score on a single than the batter does.

In Table 2.2 it is assumed that a player advances from base to base. Incorporating this baserunning ability will necessitate changing the batter's transition matrix depending on who is on base ahead of him. For example, if there are no outs and a runner on second base, in the initial matrix the runner would stop at third. Now the runner will score in accordance with the baserunner's ability to score from second on a single. In this particular instance the baserunner is the player preceding the current batter. However, if there was one out, the runner could either be the player immediately preceding the current batter or the player before him. This same process is done for runners scoring from first on a single as well as scoring from third base on an out.

The last runner-impacted effect on the transition matrix is the runner's propensity and ability to steal a base. Runners' probabilities of stealing both second and third, as well as their success rates are incorporated into the transition matrices as well. A batter hitting a single and then stealing second will appear as though the batter hit a double in the transition matrix with the notable exception that any other runners on base will react as if the batter has hit a single.

Double steals are not factored into the model but players getting caught stealing is accounted for. In this model, the only time that a runner may steal a base is immediately following his at-bat. It is assumed that if the runner attempts to steal he will do so during the following batter's at-bat, before another batting outcome takes place. While this is not always the case, it is typically the case because often the batter and the runner work together to facilitate a stolen base (the batter will swing at the pitch with no intent to make contact in some instances, desiring only to distract the catcher from making the throw). The bias that is created based on this assumption is a positive one where the runner advances an extra base on the steal before he would have in game play in some instances. Since there are generally less

than two steals per team per game, and the timing of the steal is accurate in general, this assumption will not significantly affect the distribution of runs scored. No other baserunning outs are included such as being thrown out at home plate, introducing an additional small positive bias.

As a final note this process is technically a "Markov-like" process as we need to know not only the base/out state, but also require knowledge of which batter is on base, no longer being memoryless in the true Markovian sense. Of course we may include such information in the state space, but resulting in a large number of additional states and significantly increased complexity in chain accounting within the scheme. As we see such generality more for mathematical elegance than practical significance, we keep with the transition matrix proposed herein and remain with the spirit of previously termed Markovian methods for batting order optimization referring to our routine as a Markov process.

2.2.2 Transition Matrix Pseudo code

1. Begin with the transition matrices displayed in Tables 2.1 through 2.4.
2. Establish an order to test run production.
3. Perform the runner advancement step starting with Table 2.2.
    - Identify the matrix elements that result in a batter hitting a single with a runner on second (3,6), (5,8), (7,6), and (8,8).
    - Determine the baserunner on third base after the batter hits a single with no outs in these matrix elements for each batter, and who the two and three potential baserunners on third base could be with one and two outs respectively.
    - For each of the matrix elements determine the potential runner(s) probability(s) of scoring from second base on a single.
    - If there is more than one player who could be on second base (i.e., the one or two out state), average the players' probabilities of scoring from second. (Note: with more research a weighted average may be more appropriate.)
    - For the elements that are a result of both a single and a walk ((5,8) and (8,8)) use the batter's ratio of singles to walks to identify the proportion of the probability in the element that is a result of a single to distribute in the next step.

- Use the resulting probability of scoring from second to distribute the probability in the given matrix element between the initial matrix element and the element corresponding to the state indicating the baserunner had scored (for (3,6) the other element would be (3,2)).

- Perform the same steps for the matrix elements that are a result of a runner on first moving from first to second on a single with no runner advancing to third ((2,5) and (6,5)).

- Perform the same steps for scoring from first on a double with the elements that identify baserunners on second and third as a result of a double (2,7), (5,7), (6,7), (8,7).

4. Perform the stolen bases step

    - Assess the probability of attempting and having success stealing second base for all of the non-zero elements of the Table 2.2 result from step three in the second and sixth columns.
      Example: Begin with the element (1,2) and leave only the probability of landing in that position and not attempting to steal second there. The probability of successfully stealing second would go to (1,3) and the probability of being caught stealing goes to Table 2.3, element (1,1).

    - Assess the probability of attempting and having success stealing third base for all non-zero elements in the third column to the respective elements in the fourth column (success), and the first column of Table 2.3 (caught stealing).

    - Assess the probability of attempting and having success stealing third base for all non-zero elements in the third column to the respective elements in the sixth column (success), and the second column of Table 2.3 (caught stealing).

2.2.3 Markovian Baseball

Initial transition matrices must be defined for all of the batters in the lineup. The player's base running probabilities outlined in Section 2.2.1 must also be obtained and appropriate modifications made to the transition matrices based on the lineup that is being evaluated. In this analysis it is assumed that the lineup will remain static throughout the game. In this algorithm it is straightforward to insert substitutions, or pinch hitters, with a given probability after a set number of times through the order. It is also straightforward to base the substitutions on the current inning rather than the number of trips through the order.

To calculate the number of runs scored in a game we need to begin with the initial state of the game (no outs and no runners on base) and then multiply the transition matrices in a specific and unique manner that will be described later. We will begin with a 21×25 matrix with each row representing the number of runs scored in the game, from 0 to 20. Each of the 25 columns represents a state of the game, with the 25th column representing 3 outs, or the end of the half-inning. The game will begin with probability one in the upper left corner of the matrix, and all other probabilities are zero.

The 21×25 matrix can now be used to calculate the distribution of runs scored in the first inning. In order to calculate the runs scored for a nine inning game stack nine of these matrices vertically creating one 189×25 matrix (R_0) with the first 21 rows representing zero to 20 runs scored in the first inning, rows 22 to 42 then represents the cumulative total of runs scored in the first and second innings from zero to 20 runs and so on. The cumulative sum of this matrix' entries must maintain unity as each of the entries represent the probability that the current game be in the given state (where the current inning and number of runs scored are now also incorporated into the state of the game) after any number of batters have come up to bat.

R_n is now a 189×25 matrix representing the probabilities of: the current inning, the number of runs that have scored, the amount and location of runners on base, and the number of out recorded in the given inning, before the n^{th} batter bats for the given team. The matrix does not account for the opponents at bats, runs, etc. The rows of the matrix represent the inning and the number of runs that have scored, the first 21 rows representing the first inning, rows 22-42 the second inning, and so on. Rows 1,22,43 etc. represent no runs scored, rows 2,23,44 etc. 1 run scored and so on. The first eight columns of the matrix represent no outs in the current inning, the next eight 1 out, the third eight two outs, and the last column represents three outs. The runners on base mirror the columns in matrix A where columns 1,9, and 17 represent the bases being empty, columns 2,10, and 18 represent a runner on first base only, and so on.

Since each row in R_n represents the number of runs scored in the game before the n^{th} batter comes up to bat standard matrix multiplication will not work. The transition matrices T generated after the baserunning steps must be decomposed into $T0$, $T1$, $T2$, $T3$, and $T4$, where $T0$ represents the matrix probabilities of T where no runs will score, $T1$ represents the probabilities that result in one run scoring, and so on. To advance to the next batter the decomposition of the batter's T matrix is combined with the R matrix according to equation 1, where the first entry in the parenthesis represents the row and the second the column, in typical Matlab notation (the colon, ':', indicates all columns). Multiplication in this way affects a maximum of five rows in R_n when multiplied by the 25×25 transition matrix T, and tracks the number of runs scored with the row of R_n.

$$R_{n+1}(i,:) = R_n(i,:)T0 + R_n(i-1,:)T1 + R_n(i-2,:)T2 + R_n(i-3,:)T3 + R_n(i-4,:)T4. \tag{2.1}$$

Take note that for the first three rows of R_{n+1} this formula will be abridged to include only positive rows of R_n. R_0 is combined with the transition matrix for the leadoff batter ($T1$) using equation 1, and that result is combined with the transition matrix for the second batter ($T2$) and so on. Once the last batter in the lineup ($T9$) is reached, we cycle back to the leadoff batter. To account for the changes in innings (once three outs are reached) the probabilities in $R_{n,25}$ is moved to $R_{n+21,1}$. This will preserve the number of runs scored up to the given point (denoted by the row of the matrix) while moving to the next inning and resetting the outs to zero.

The batters will be cycled through in this manner until the probability that the game is over (last 25 rows of the last column) is sufficiently close to one. In general seven trips through the lineup will accomplish this. (Note: The code must ensure that when runs are scored that the matrix row does not move past a multiple of 21, from row 21 to row 22 for example. This would happen if the team scores 21 or more runs. If we do not prevent the

algorithm from making this jump then scoring the 21st run would cause a bookkeeping effect of losing all runs scored and advancing one inning.) Once the algorithm has stabilized, and the probability of the game being over is sufficiently close to 1, we compute the average number of runs scored by summing the product of the probability that the lineup generated the given number of runs by the number of runs generated, from 0 to 20.

An additional advantage of this approach is that it is also possible (though out of the scope of this dissertation) to differentiate between lineups with similar average runs scored per game since what is produced is a distribution of runs scored. For example based on the individual team's pitching and the distribution of runs given up per game, it is possible to directly estimate the probability of winning a particular game by simulating both teams' distributions of runs scored given the pitching match-ups on any particular day. As mentioned before it is also relatively straightforward to adjust the statistics when the starting pitcher leaves the game by using the distribution of batters faced or a probability associated with any given inning. Of course some reconciliation would have to be done to account for the correlation between runs allowed by the starting pitcher and batters faced or innings pitched.

2.3 OPTIMIZATION

Using this algorithm it is computationally intensive to perform an exhaustive search of all 362,880 possible lineups for any given set of players. The exhaustive search takes 23.5 hours on an Intel Dual Core 3.2 Ghz PC with 2 GB of RAM using Matlab R2007a. Once additional player possibilities are inserted into the lineup (e.g. platoon situations), the computational time could easily expand to weeks of CPU time. Three different routines were suggested by Bukiet et al. [16] and we will begin with a modified version of one of those methodologies: the single placement approach.

The batters are inserted into the lineup one at a time starting with the player that was furthest from the average in terms of team ability, and once that player is placed in the batting order the second most distinctive player is inserted and so on. OPS (on base percentage plus slugging percentage) was used to evaluate player ability in terms of order insertion. In order to

determine the first batter's position in the lineup, the statistics for the remaining eight player's in the lineup are assumed to be the average of the other eight players' statistics. This is done under the assumption that each remaining player has equal probability of being in every unoccupied position in the lineup. Therefore since they each could occupy every unoccupied lineup position with the same probability, the average performance of the players not inserted is the performance of each unoccupied position. The first batter is then tested in each of the nine possible batting order positions and his final position is determined by the order which yields the highest expected number of runs. When inserting the second batter, the first batter is left in his position and the remaining seven players' statistics are averaged for the vacant lineup slots, and so on. This generally produces a near optimal order, but can miss by a handful of runs over the course of the season and is very dependent on the order in which the players are inserted. We thus propose applying a greedy algorithm on this sub-optimal order.

2.3.1 Greedy Algorithm

Using the lineup obtained by the above algorithm (the order will be referred to as 123456789), begin with the leadoff batter, and try inserting him into the second position with the player in the second position cycling up to the first position (213456789). Continue moving the original leadoff batter into the other seven possible positions until a better order is found (testing 231456789, 234156789 etc.). When a better lineup is found, use that lineup and look at the new leadoff man in each of the other eight possible lineup positions. If the leadoff man's best position is to bat leadoff, then step down to the batter hitting second and test him batting in all of the other possible positions in the batting order. If second is his optimum position move on to the third batter and do the same. Once a better lineup is obtained start back with the leadoff man and cycle through the routine again. The algorithm ends once a complete cycle is made and no better lineup is found. If the lineups that are tested are not tracked, then the same lineup will be evaluated multiple times so a log was kept to track the lineups tested and if the given lineup had already been examined it was skipped.

To assess the robustness of the starting point a random insertion algorithm was also used. In the random insertion algorithm players are selected at random for each of the lineup positions. The random insertion algorithm proved to be more useful when running the greedy algorithm multiple times as the starting point plays a large role in how the algorithm traverses the space and hence whether or not it gets stuck in a local maxima. A randomization step was also introduced to help avoid getting stuck in local extrema. With the randomization step 84% of runs resulted in a solution that was within one run per season of optimal, without the randomization step 66% were within one run per season. The randomization step was utilized with the first 400 lineup comparisons, after which the probability of accepting a lineup that was not an improvement was set to zero.

2.3.2 Greedy Algorithm Pseudo code

1. First batter is the test batter, test run production of the lineup with the test batter moved down to the second position and the batter in the second position sliding up to leadoff.
2. If runs increases use new lineup and return to step 1.
3. Randomization step: If fewer than 400 lineups have been tested generate a random number between 0 and 1. If twice the inverse of that number is greater than the number of lineups tested then use the new lineup and return to step 1.
4. If the test batter is not in the ninth position in the lineup, swap the lineup position of the test batter with the position of the batter one spot lower in the lineup, evaluate the run production of the new lineup and go back to step 2.
5. Return the test batter to his original lineup position and designate the batter following him as the test batter. Move the new test batter to first in the lineup, evaluate the new order and go back to step 2.
6. If fewer than 400 lineups have been tested go back to step 1.
7. Upon completion all batters will have been tested in all other positions in the order. This resulting order is is at least a local optima, and was found to be the global optimum in many of the cases tested.

2.3.3 Bootstrapping

To test the robustness of the optimal order a bootstrap algorithm was employed on the individual player production matrices. Batting average (AVE) and isolated power (ISO) were simulated using asymptotically normal distributions as follows

$$AVE \sim AN\left(ave, \sqrt{(ave * (1 - ave)/600)}\right)$$

$$ISO \sim AN\left(iso, \sqrt{(iso * (1 - iso)/600)}\right).$$

Isolated power is similar to slugging percentage: (2B+2*3B+3*HR)/At-bats, and was used because it is less correlated with batting average than slugging percentage. Isolated power was bootstrapped first and the probabilities of hitting a double, triple, or home run were scaled maintaining the batters original ratio of doubles to triples to home runs. The probability of hitting a single was then adjusted to yield the batting average result from the bootstrap.

2.4 RESULTS

There are many standard practices in batting order selection that have been called into question by analysts in recent years. Where the pitcher should bat in the order is a question that has been raised in the past few years. Pitchers typically bat last in the order, but recently the Milwaukee Brewers and St. Louis Cardinals have begun experimenting with positioning the pitcher eighth rather than ninth in the order. Also historically the first batter (or leadoff batter) has been a fast runner who can steal bases, and a secondary consideration has been the batter's ability to get on base, with the second position in the lineup filled by a batter who rarely strikes out and is able to direct the ball to one side of the field or the other (which assists in allowing runners on base to advance a base even though the batter himself may not reach base). The last question that we will address is where should the power batter who is also the leader in on-base percentage (OBP) hit. The traditional position for this batter is

either third or fourth. To perform this analysis we will examine four teams from the 2008 MLB season. The play-by-play data for this analysis was collected from retrosheet.org [40].

The results from the analysis of these teams led us to an additional test using generic lineups. The MLB teams' results indicated that the players' on-base percentage was the most significant factor in determining their lineup position. The generic lineups were made up of fictional batters that had identical on-base percentages but the attributes of speed, power, and contact (the ability to advance base runners without reaching base) were varied. There were also additional player profiles inserted to represent the pitcher, a better batter, a weaker batter, and a batter with power but a low on base percentage who tends to strike out.

In general when the lineup was selected via the Bukiet et al. [16] one at a time insertion routine it produced a result that was both highly predicated on the order of player insertion and almost 0.5 run per game (or approximately 8 wins) off from the optimal lineup. However once the greedy move method was employed the result was rather stable both in the number of runs per game and the actual order. Three different player orders were used for the Bukiet et al. routine, as well as twenty random orders for each team, and the greedy algorithm was employed. For each of the teams the optimal order was generated in at least twenty percent of the orders. For three of the teams there was a near optimal order that was generated with more frequency than the optimal order. In each of the tables (2.5,2.7,2.8) this order is labeled greedy and the optimum is labeled as such.

Bukiet et al. [16] also recommended a criteria based set of 987 orders, which were generated using the results of National League teams from 1989 and tested against National League teams from 1969. While their criteria based code identified the optimal order under their assumed conditions in five of twelve lineups, it did not find the optimal lineup in any of the four 2009 lineups, and in fact was as much as 6.3 runs per season off from optimal in the Phillies lineup (which was attained in fifty five percent of the greedy algorithm runs). Sokol's [54] heuristic fared much worse ranging from 6.3 - 10.5 runs per season below optimal. In all lineups the greedy algorithm proposed fared better than both Sokol and Bukiet et al. Bukiet et

al.'s algorithm fared within 2 runs per season of the greedy algorithm in three of the four teams.

The baserunning element has a big impact on the lineup that is chosen, and in selecting the best possible lineup. To test this the 987 orders identified by Bukiet et al. [16] were also run without including the baserunning algorithm. The algorithm was just over 2.5 runs per season below optimal when the baserunning algorithm was incorporated, but almost 10 runs per season below optimal when the baserunning was not utilized. In some cases the result was not a significant improvement over the most common lineup utilized by the team.

When the player performance metrics were bootstrapped, the resulting run productions followed an approximately normal distribution that had a slight positive skew. While the standard deviations of runs per game was between .2579 in table 2.7 and .3184 in table 2.5 there was consistency in the optimal order. In all four of the teams tested the optimal order out performed the most commonly used order in over 99.5% of the simulations. In three of the four lineups there was a lineup that was identified more frequently than the optimal, but was always within one run per season of optimal.

The bootstrapped production was also utilized to evaluate the orders selected by Bukiet et al.'s [16] criterion and Sokol's [54] heuristic as well as the traditional order in addition to the optimal and greedy algorithm orders. Each order was tested with 200 different bootstrap iterations and the mean and standard deviations of the results are included on the team results tables. The variance of production based on the bootstrap was not significantly different for the different orders. It is interesting to note that the rank of the orders remained consistent with all of the teams across the bootstrapped results. This result suggests that prior assertions that the near optimal orders proposed by [16] and [54] were sufficient, may not be correct. The greedy algorithm outperforms the other routines across player variability. Additionally, the benefit that the greedy algorithm has of obtaining the true optimum with high probability in a short amount of time (has been successful in all of the lineups tested in fewer than 10

iterations, or 15 minutes) is underscored. With any one bootstrapped order the optimal order has a relatively low probability of remaining the optimal, but it remains optimal in probability.

On-base percentage (OBP) seemed to be the most significant driver in determining the lineup. In all four teams, the Optimal lineup placed the teams' top three players in OBP in the 2nd to 5th positions in the lineup. In addition, all of the top tier batters in OBP on a team batted together in the lineup. The 1st position in the lineup was relegated to the lowest position player in OPS (OBP plus SLG) in three of the four lineups (all three NL lineups with the pitcher batting last), with the fourth team batting a player in the bottom third in OPS. In all four lineups tested the lowest batter in OPS batted 9th.

The pitcher hit ninth in the optimal order for all three of the National League (NL) teams and remained robust for over 80% of the top 25 orders overall for each team. It appeared to be the case that batting the pitcher last made it desirable to bat a less capable hitter in the 1st position. In two of the three NL lineups the player with the biggest difference between on-base percentage and slugging was placed either 7th or 8th. In all NL lineups the player with the biggest difference between OBP and SLG was the last batter before the bottom tier of batters. This way there is a lower probability of the pitcher batting with runners on base. While this works in theory, it may not in practice since if a player like Ryan Howard or Carlos Delgado were batting right in front of the pitcher, the opposing team may be more likely to allow the batter to draw a walk and face the pitcher instead.

The biggest difference between the optimal orders found and the traditional approach was clearly in the first position of the batting order. While some teams have been experimenting with batting the pitcher eighth in the order rather than ninth, none have taken the approach of intentionally batting a weak hitter first. This is in clear opposition to Bukiet et al.'s [16] assertion that the two worst batters must be placed in the bottom third of the order, however even their algorithm finds the seventh best batter either bats first, or in the bottom of the order. Sokol [54] also has a less productive hitter bat in the first position since the worst of the table setters, or players with above team average ability to get on base but below team

average ability to drive in runs, bats first. Baserunner advancement does play a significant role in this as well. When testing Bukiet et al.'s 987 orders both with and without the runner advancement, a more capable batter consistently hit first in the order when the runner advancement metrics were not employed.

The best batter hit in the 2nd to 4th positions. There were two teams that had a player that was clearly the best in terms of OBP (Milton Bradley of the Rangers and Albert Pujols of the Cardinals). Those players both hit before the other top tier batters on their team (2nd and 3rd respectively). On the other two teams where there was not one batter that was clearly the best in terms of OBP, the best batter in terms of OPS hit in the 4th position.

The "speed" batters, traditionally batting leadoff, were not part of the top of the order. It was not obvious that the speed metrics played a significant role in the lineup optimization. These prototypical leadoff men were placed below the sluggers that batted 3rd, 4th, and 5th traditionally (because the sluggers had higher OBP). The notable exception was Cesar Izturis of the Cardinals who hit 1st in the optimal order, all others were placed in the 4th to 7th positions in the lineup. This was probably due to the fact that while Izturis is a "speed" batter, he is also the least capable batter in the Cardinals lineup in terms of OPS.

2.4.1 Texas Rangers

The Rangers were used because they have a fairly diverse lineup with three very good hitters (2008 statistics); Milton Bradley is the best hitter by the OPS metric (0.939) followed by Ian Kinsler (0.933) and Josh Hamilton (0.892). They also have a consistent leadoff hitter in Ian Kinsler, who fits the traditional expectation of a leadoff batter with one exception, he can hit for power. Michael Young is also a good test for the second batter, since he has hit second for the Rangers over 70% of the time from 2005 thru 2008 and his on-base percentage is significantly lower than five other players on the team (2008).

The best player (Bradley) hit second in the optimal order as well as in nearly all of the top 25 overall orders. The next four batters in on-base percentage (Ian Kinsler 0.369, Josh Hamilton 0.368, Ramon Vazquez 0.367, Marlon Byrd 0.365) hit in the next four spots in the

order with Kinsler (highest Slugging Percentage 0.564), hitting fifth (Table 2.5). Ramon Vazquez, who fits the prototype for a second batter, hit third behind Bradley. Michael Young hit in the leadoff position as the worst batter in terms of OPS hitting before 8th in the order.

Rangers	OBP	SLG	SB	Most Common	Bukiet	Sokol	Greedy	Optimal	Bootstrapped Greedy
Ian Kinsler	.369	.564	26	1	5	3	5	5	6
Michael Young	.340	.410	10	2	7	1	6	1	8
Josh Hamilton	.368	.524	9	3	3	6	3	4	4
Milton Bradley	.425	.514	5	4	2	4	2	2	2
David Murphy	.321	.482	7	5	6	7	7	7	5
Marlon Byrd	.365	.461	7	6	4	5	4	6	9
Brandon Boggs	.324	.371	3	7	8	8	9	9	1
Gerald Laird	.310	.396	2	8	9	9	8	8	7
Ramon Vazquez	.367	.396	0	9	1	2	1	3	3
Runs Per Game				5.764	5.848	5.812	5.849	5.851	5.819
Order Rank				93,697			8	1	2,178
Percent Obtained				–			45	20	20
Bootstrapped Avg				5.781	5.868	5.831	5.868	5.870	
Standard Dev				.3189	.3211	.3187	.3204	.3184	

Table 2.5. Rangers position players for 2009 and summary statistics from 2008.

Cardinals	OBP	SLG	SB	Most Common	Bukiet	Sokol	Greedy	Optimal	Bootstrapped Greedy
Skip Schumaker	.354	.400	8	1	6	4	2	2	6
Rick Ankiel	.330	.486	2	2	5	7	6	6	5
Albert Pujols	.459	.593	7	3	3	3	3	3	3
Ryan Ludwick	.365	.572	4	4	4	6	5	5	4
Troy Glaus	.369	.482	0	5	2	5	4	4	2
Yadier Molina	.347	.376	0	6	1	2	8	8	8
Felipe Lopez	.326	.376	8	7	7	9	7	7	7
Pitcher	.197	.216	0	8	9	8	9	9	9
Cesar Izturis	.300	.349	24	9	8	1	1	1	1
Runs Per Game				4.989	5.0639	5.009	5.071	5.071	5.070
Order Rank				14,641			1	1	2
Percent Obtained				–	–		35	35	40
Bootstrapped Avg				4.997	5.071	5.017	5.077	5.077	
Standard Dev				.2751	.2758	.2754	.2753	.2753	

Table 2.6. Cardinals position players for 2009 and summary statistics from 2008

2.4.2 St. Louis Cardinals

The Cardinals were selected primarily because they had arguably the best hitter in the game in Albert Pujols (1.052 OPS). We wish to examine how the algorithm would position

him with their two other best hitters (Ryan Ludwick 0.937, Troy Glaus 0.851 OPS). They also had a batter with a low OBP (0.330) and higher SLG (0.486) in Rick Ankiel who hits a lot of home runs but does not get on base much otherwise. A number of industry analysts suggested that Pujols should hit first. Also the Cardinals were one of the few teams in Major League Baseball who had begun to bat the pitcher in the 8th position rather than hitting him last. We wish to study these ideas with our algorithm.

The algorithm placed Pujols, Glaus, and Ludwick third, fourth, and fifth respectively in the order (Table 2.6). One of the primary reasons why the algorithm selected Pujols to hit as low as third in the order, was because of the pitcher. (One criterion that Bukiet et al., [16] suggested was that there were at least three lineup slots between the best and worst hitters). If Pujols were to hit earlier in the order then the pitcher would have to either bat earlier in the order, or there would be no more than two batters in between him and Pujols, which would lead to fewer opportunities to produce runs from Pujols' at-bats. The Cardinals were the only team that had a batter who actually batted leadoff for them at some point in the season batting leadoff in the optimal order (Izturis). After Cesar Izturis, was Skip Schumaker, the fourth best batter in terms of on-base percentage, with a very small difference between on-base percentage and slugging (indicating that the batter has little power).

The Cardinals are the only team that had near optimal lineups that resembled a traditional lineup. In fact this is the only lineup where each of the players has appeared in his optimal lineup position at least once. This appears to be mostly by chance, as the only reason that Cesar Izturis appeared as the leadoff batter is because he can steal bases, not because he was the worst batter on the team in terms of both on-base percentage and slugging.

2.4.3 Philadelphia Phillies

The Phillies team was selected for a number of reasons. First, we wanted to select a number of National League teams to vet the location of the pitcher in the order. Second, the Phillies generally bat Jimmy Rollins first in the order (a batter who steals a lot of bases but does not have one of the better on-base percentages on the team). The Phillies lineup is also

Phillies	OBP	SLG	SB	Most Common	Bukiet	Sokol	Greedy	Optimal	Bootstrapped Greedy
Jimmy Rollins	.349	.437	47	1	5	2	5	6	6
Shane Victorino	.352	.447	36	2	6	3	6	5	5
Chase Utley	.380	.535	14	3	4	4	4	4	4
Ryan Howard	.339	.543	1	4	1	7	8	7	7
Pat Burrell	.367	.507	0	5	2	5	2	2	2
Jayson Werth	.363	.498	20	6	3	6	3	3	3
Pedro Feliz	.302	.402	0	7	7	9	7	8	8
Carlos Ruiz	.320	.300	1	8	9	1	1	1	1
Pitcher	.197	.216	0	9	8	8	9	9	9
Runs Per Game				4.751	4.853	4.827	4.891	4.892	4.892
Order Rank				58,992			2	1	1
Percent Obtained				–	–		30	55	60
Bootstrapped Avg				4.778	4.878	4.851	4.916	4.918	
Standard Dev				.2434	.2407	.2419	.2429	.2435	

Table 2.7. Phillies position players for 2009 and summary statistics from 2008.

relatively diverse, having three batters who stole at least twenty bases (Jimmy Rollins, Shane Victorino, and Jayson Werth), a fairly diverse range of on-base percentages (Pedro Feliz 0.300 to Chase Utley 0.380), and one of the most dynamic power hitters in the game (Ryan Howard).

The optimal order placed the top three batters in on-base percentage in the 2nd to 4th positions in the order (Table 2.7). The best hitter both in on-base percentage and on-base plus slugging percentage (Chase Utley) was positioned fourth in the order. It is interesting to note the positions of Pat Burrell (2nd) and Jayson Werth (3rd). These are two similar players in on-base percentage. Their most notable differences were that Burrell hit more home runs and Werth stole more bases. Conventional baseball wisdom would result in batting Werth before Burrell. In fact Burrell has never hit 1st or 2nd in MLB play. What is most likely the cause of this is that Burrell walks more and Werth gets on base more often as a result of a hit. The Phillies fastest players (Rollins and Victorino), who most frequently occupied the top two positions in the order, were placed 6th and 5th in the order respectively. Ryan Howard was placed 7th right in front of the weakest portion of the order, which does point out a current weakness in the model. If Howard did hit in front of weak batters, he would likely walk more frequently (especially when first base was un-occupied) with teams electing to face the weakest portion of the order with an additional runner on base rather than Howard. This would change his transition matrix, but exactly how it would change is another topic

altogether. Also if this analysis were done using a prior year's statistics the result would likely be quite different as 2008 was Ryan Howard's worst year up to that point both in terms of batting average and on-base percentage.

2.4.4 New York Mets

The Mets were selected as another National league team with a traditional leadoff batter (Jose Reyes). Much like the Phillies they also have a diverse lineup, with some runners who steal more bases, and some who hit more home runs. Their lineup is also pretty stationary at the top of the order as Reyes, Castillo, Wright, Beltran, and Delgado almost always bat in that order when they are all in the lineup.

The optimal order included the top five batters in on-base percentage batting in the 2nd through 6th spots in the order (Table 2.8). This was the only lineup tested that had either of the first or second batters in the most common lineup batting second (Luis Castillo). Castillo hit second both in the Mets lineup as well as in the optimal one. As with each of the other National League lineups, the Mets had a batter with a low on-base percentage and high slugging batting before the least capable batters (Carlos Delgado 8th). The Mets had only one position player and the pitcher who performed poorly both in on-base percentage and slugging, and they were placed first and ninth in the order.

Mets	OBP	SLG	SB	Most Common	Bukiet	Sokol	Greedy	Optimal	Bootstrapped Greedy
Jose Reyes	.358	.475	56	1	5	2	6	6	8
Luis Castillo	.355	.305	17	2	9	1	2	2	4
David Wright	.390	.534	15	3	3	4	4	4	5
Carlos Beltran	.376	.500	25	4	2	3	3	3	3
Carlos Delgado	.353	.518	1	5	7	6	8	8	2
Fernando Tatis	.369	.484	3	6	4	5	5	5	6
Ryan Church	.346	.439	2	7	6	7	7	7	7
Brian Schneider	.339	.367	0	8	1	9	1	1	1
Pitcher	.197	.216	0	9	8	8	9	9	9
Runs Per Game				5.052	5.120	5.092	5.135	5.135	5.102
Order Rank				19,313			1	1	147
Percent Obtained				–	–		0	97.5	10
Bootstrapped Avg				5.069	5.128	5.099	5.143	5.143	
Standard Dev				.2776	.2848	.2809	.2819	.2819	

Table 2.8. Mets position players for 2009 and summary statistics from 2008.

2.4.5 Additional Lineup Tests

Since on-base percentage played such a large role in determining the optimal order for MLB teams, the question arises: what if OBP remained constant, then where would different types of players hit; and where would the best and worst batters bat in this type of lineup? Additionally, at what point will other factors outweigh on base percentage? In order to answer these questions different player profiles were created. Fictional batters designed to answer the above questions.

Three different players were created all having identical on-base percentages (.360):

Speed Batter: The prototypical leadoff batter that steals a lot of bases and is very likely to advance the extra base when on the base paths. Player attempted a steal with no runners on base 50% of the time with an 85% success rate. Runner advancement was based on the average advancement probabilities of the top 10 players in stolen bases from 2008. Hits into a double play 5% of the time. Slugging percentage of .405 includes slightly more doubles and triples than the contact batter due to the runners speed.

Contact Batter: Prototypical batter to hit second. This player steals some bases and is more likely than average to advance the extra base. He rarely strikes out and does the "little things" to advance runners that are on base without necessarily reaching base himself. Attempts stealing second with no runners on base 10% of the time with 75% accuracy. Sixty percent of the time when the batter makes an out, he advances other runners already on base. Hits into a double play 5% of the time. Slugging percentage of .395. Runner advancement was based on overall league average.

Power Batter: Prototypical cleanup batter (fourth in the lineup). This batter does not usually advance the extra base and has a tendency to hit into double plays. However, they also hit a lot of doubles and home runs. Player ability was based on the average of the sluggers who do not steal bases from 2008 (players like Alex Rodriguez and Albert Pujols were omitted because they have more speed than the traditional power batter.) They hit

approximately 35 doubles and 30 home runs over the course of the season, but also hit into double plays 30% of the time when a runner was on first base.

Additional Players: In addition to these three player types, other players, with different on-base percentages were also included. A typical pitcher, as well as another batter with lower OBP and SLG was also used. In addition, a best player, with higher OBP and SLG numbers was included as well. The last type was a player with very low OBP and high SLG. This player was modeled after Mark Reynolds, a prototypical cleanup hitter but without the ability to consistently reach base safely.

The speed batter was evaluated to determine the effect of stolen bases on the lineup production. It was found that when the runner attempted to steal regardless of whether or not there was already a runner on base, if he were not successful at least 93% of the time attempting to steal, it was a detriment to overall run production. When the player only attempted to steal with the bases empty the threshold was reduced to 82%. This is why the speed batters were assumed to only attempt to steal with no runners on base. When the runner attempts a stolen base is a management issue, not a player ability issue. Further research could be done to analyze threshholds for productive stealing across each of the twelve states of the game (four possibile combinations for baserunners and three different possibilities for number of outs) where steal attempts are possible and with varying lineups around the base stealer(s).

The first lineup that was tested utilized three speed batters, three contact batters, and three power batters. In these orders all speed batters are exactly identical, all contact batters are identical, and all power batters are identical initially and then variations were included to test sensitivity. The optimal order was then obtained using the random insertion algorithm coupled with the greedy algorithm as described in previous sections. Other orders were also tested utilizing one to three of the additional players in place of some of the players from the first lineup. In all of these additional lineups there were at least one of each of the speed, contact, and power batters, and in most lineups tested there were two of each.

Here there is less variability between the players, when we speak of the worst player it is the player that generates the fewest runs per game when the entire lineup is filled with that player. These orders showed consistency with the team lineups tested in that the best batters hit together, the worst batter hit last, and the second worst hit first. Generally the power batters hit in the fourth and fifth positions. The interesting thing here is that since there is less disparity between the players, the best lineups tend to bat the third power batter in either the eighth or ninth positions, effectively building two mini lineups within the lineup. This is true regardless of the quality of the power batters included in the lineup. This actually is consistent with the MLB lineups ran as well since in many of those optimal lineups there was a power batter hitting at the bottom of the lineup as well.

Player Type	OBP - .340	OBP - .350	OBP - .360	OBP - .370	OBP - .380
Contact batter	4.401	4.721	5.060	5.418	5.796
Speed batter	4.886	5.224	5.579	5.925	6.341
Power batter	5.406	5.733	6.082	6.454	6.851

Table 2.9. Runs per game for lineups made exclusively of one type of batter at varying On Base Percentages

Table 2.9 shows the relative value of the three different types of batters. Of course the specific value to any individual team is predicated on the remainder of the lineup, however this table shows the relative value of the players overall, and the number of runs per game that would be generated if a team had nine identical players as indicated by the descriptions of the batter types, with the only fluctuation being the players on base percentage. This table shows that for a contact batter to be as productive as a speed batter he requires an improved on base of about fifteen points. Likewise the difference between the speed and the power batters is also about fifteen points in terms of on base percentage. These generalities show some trade-offs although the exact trade offs will vary from lineup to lineup.

Overall this work demonstrates the opportunity that teams have to capitalize on lineup optimization in a way that no paper before it has. While all of the nuances of the game are not

captured fully in this method, enough of the factors are incorporated to make the approach actionable in a manner that has not been done before. An MLB manager with an analytical mindset reading this dissertation would be forced to at least consider the idea that there are opportunities on his team to win more games by utilizing lineup optimization.

CHAPTER 3
BATTERS AND OFFENSIVE PRODUCTION

3.1 OVERVIEW

This chapter will go over the offensive analysis from models to results and discussion. Since the focus is on predicting performance for fantasy baseball, it is important to understand what statistics are most important in this context. There are five core offensive categories in fantasy baseball: runs, home runs, runs batted in, stolen bases, and batting average, while on base percentage and slugging percentage are two additional categories that are used in a substantial number of fantasy leagues. The goal of this section is to be able to accurately predict these seven summary statistics for each player over the course of a season.

Traditional fantasy leagues have a draft before the season starts and that draft determines the players for each team throughout the course of the season. There are trades and free agent acquisitions in most leagues but the biggest part of player acquisition occurs before the first pitch of the regular season is thrown. Thus projecting performance will be based on prior seasons data only and projections will be made for the subsequent season.

The biggest piece of fantasy performance is predicated on the outcomes of the given players at bats. Home runs, batting average, on base percentage, and slugging percentage are all directly determined by the player's batting outcomes. Runs batted in are jointly determined by a player's batting outcomes as well as the batters on base when the given player bats. Runs scored is jointly determined by the player's batting outcomes, his baserunning, and the players who come up to bat after that player reaches base. Stolen bases can only happen after a player reaches base, and is also predicated on the next base being vacant when the player attempts to steal and the managers discretion.

Section 3.2 presents a review of the literature on modeling offensive production in baseball and where this dissertation lies therein. Section 3.3 describes the goals of the model and what will be addressed in Sections 3.4 - 3.8. Section 3.4 details which distributions are used and how they are used in the models. Section 3.5 describes the growth curves used and how they are layered into the models described in Section 3.4. Section 3.8.2 details how model variances are analyzed and addressed through the inclusion of decision trees. Section 3.7 details the algorithm employed in the simulations, once the ability models described in Sections 3.4 - 3.8.2 are implemented. Section 3.8 describes and assesses the models accuracy, as well as comparing the model results to two best in class fantasy projections.

3.2 LITERATURE REVIEW

Baseball is a sport that has been analyzed statistically for over six decades now, becoming more popular in the 1990's and brought to the mainstream with Bill James and his book Moneyball. Many individual events and outcomes have been analyzed as well as a various number of total game and season simulations including: Howard [32], Cook [19], Freeze [26], Jensen [34], [36], [35], Thorn [58], Pankin [46], Stern [55], Bukiet [16], Takei [57], Sokol [54], Stevens [56], Null [44], and Nobuyoshi [42]. All of these prior methodologies utilize a Markov chain approach which is effective for tracking the number of runs scored, however it does not track individual player performances. There have been no publications to our knowledge that iteratively simulate games and seasons so that individual player performances are trackable.

The Markov chain approaches that began with Howard [32] and gained traction with Bukiet [16] are very useful and quick in order to determine game and team based simulations, but less so when individual player performances are sought out. Bukiet [16] utilized the Markov chain approach to determine optimal batting orders, an approach extended in Chapter 2. Null [44] utilized the Markov chain approach to predict individual game winners as well as teams win loss record and divisional winners over the course of a season. Null also compared the results of the Markov chain method to other simpler methods of simulation that make

assumptions about scoring based on players on base percentages and slugging percentages. These simpler approaches were shown to be less accurate than the Markov chain approach in terms of total runs scored.

Our approach is a play by play approach but it is unique in that it is not Markov chain based and hence both more flexible and more informative at the individual player level. This research is unique in that it is the first known player forecasting system that simulates seasons play by play.

3.2.1 Game Simulations

There has been much published on player batting performance as well as a fair amount on baserunning. It is largely segmented as each paper will evaluate one aspect of the game whether it be hitting for power (Albert [3]), hitting over time (Albert [6], Fair [25], Schall [51]), evaluating the impact of particular outcomes (Albert [7], Baumer [12], Cover [20]), scoring runs (D'Esopo [22]), hitting streaks and clutch hitting (Albert [4], Albright [10], Cramer [21]), or more external factors such as park factors and pinch hitting (Acharya [2] and Hirotsu [29]).

While many have used Markov Chains to evaluate team based batting performances, the first to develop comprehensive models to forecast performance were Jensen [34] and Null [44]. They utilized a Bayesian framework with Dirichlet-Multinomial conjugate distributions as will be utilized in this dissertation. Null utilized a Nested Dirichlet to account for covariances between the outcomes. This was especially important for Null's model as it was built across players. Across players there is a significant positive covariance between strikeouts, walks, and homeruns, while the covariance is negative between homeruns and other hits such as singles, doubles, and triples. This dissertation will also utilize Dirichlet-Multinomial conjugate distributions in a Bayesian framework, but will incorporate more within player data (the player's entire career) rather than utilizing data between players.

Null [44] also incorporates a quadratic age factor to account for the age bias found in the model before age is accounted for. Null's age factor is universal across players. This will

extend the quadratic age effects utilizing both within and between player data and incorporating an iterative Bayesian method in order to find unique age curves for each player.

3.2.2 Growth Curves

Albert [3] modeled home run rates and total home runs utilizing Poisson mixture distributions as well as quadratic log-linear models. He concluded that that the Bayesian quadratic model was better because accounting for age was important when evaluating power. His quadratic model was player specific and utilized Bayesian weights to account for outliers in the data. Our model will take this idea a step further and utilize the data available to create a posterior view of the player's underlying ability at each season rather than reducing the weight of a season when the evidence indicates that the player outperformed the expectation of his ability. A further limitation of Albert's research is that he only evaluated back in time and did not use his models for forecasting.

Null [44] had the advantage of being usable for forecasting regardless of the amount of data that was available. The advantage of Albert [3] is the use of distinct player curves more accurately accounting for individual age effects. Null's strength was Albert's weakness and vice versa. Our research will create a hybrid of the two approaches, while extending Albert's outlier impact so that the growth curves are as individualized as possible given the available data, and are built on posterior ability expectations of each individual player.

Albert [6] also evaluated overall performance utilizing quadratic growth curves. The overall performance metric utilized was one proposed by Thorn [58] which provided weights for each outcome as an alternative metric to OPS (on base percentage plus slugging percentage) before that metric was widely utilized. He used a Bayesian exchangable model in order to smooth career trajectories and account for the knowledge of how time effects other players of a similar era. The era piece is an important one to consider as we are now in an era commonly referred to as the "Post Steroid Era" and growth rates have a much more precipitous decline than that of the "Steroid Era". Some work has been done on bridging the gap between eras by Berry et al. [15], [14]. They utilized the continuous overlap between

players in order to assess the impact of the era on both batting average and home runs. Albert [6] found that the model did not do a great job of forecasting what the player's peak season would be. This is not surprising since players spend a number of seasons near, or at, their peak ability and production variability will outweigh underlying ability most of the time in these situations. Albert [6] also evaluated the relationship between the length of the player's career and the slope of their curves and found that shorter careers did indeed have steeper slopes. Again Albert [6] was not forecasting but evaluating post career. This finding suports the idea that a player who comes into the league at a more advanced age will likely have a more precipitous decline. The reason being that in order for a player to have a lengthy career, they must begin their career earlier (players in general are not going to have a twenty year career that starts at the age of 26). Identifying the relationship between slope and longevity is important for us because this will help us to identify players who are likely on the decline and should be avoided for fantasy purposes.

Fair [25] also modeled age effects, using on base plus slugging percentage (OPS) as his independent variable. Fair [25] utilized fixed effects nonlinear regression to fit a piecewise quadratic function. There were two quadratic functions, one for growth and the other for decline, pieced together at the modeled peak age of 27. This peak was confirmed by the prior research of Schulz [52]. This is a similar model to the one employed in this chapter for players with fewer than five years of experience or with convex independent curves. When more data is available then our model allows each players data to influence the slope of the curve as well as the intercept. The chief improvement of the quadratic models in this chapter is that a Bayesian approach is utilized to iteratively evaluate the underlying ability of the player at each season, using all other seasons data to create the prior distribution, thus reducing the effect of outliers in the data.

3.2.3 Regressing toward the Mean and Mean Reversion

What is being addressed in this section is the idea that batters who significantly outperform either their own historical averages or the average of the league at large significantly, are not likely to repeat the same feat in the subsequent season. Null [44] and Schall [51] have both addressed the issue of regressing toward the mean or mean reversion. Schall evaluated batting averages utilizing linear models to predict batting average based on the z - scores calculated by year of prior years' batting averages for the same player. He found mean reversion to be evident, and observed that this would be expected given that the evaluation is based on a sample where many of the higher observations are likely outliers rather than true underlying ability. Our model addresses this phenomenon in the methodologies described in sections 3.5 and 3.8.2.

3.2.4 Evaluating Outcome Metrics

Albright [10] evaluated the hypothesis of streaky hitters. Streakiness is the idea that a player's most recent plate appearances impact the probability of success for the next plate appearance. He utilized a logistic regression model concluding that there was not significant evidence of streakiness for at bat success in baseball. Albert [4] wrote a response to the article emphasizing the idea that the modeling strategies employed was not able to show evidence of streakiness, but that did not mean that streakiness does not exist. Albert [9] extended this work to include strikeouts and home runs as well as utilizing a Bayesian approach with beta-binomial conjugate distributions. He did find evidence of streakiness within one season for each of the statistics, but not for players across the statistics. This suggested that while there is some evidence of streakiness within a season, that does not necessarily support the idea that the players themselves are streaky (i.e., that the same players will continue to show evidence of streakiness in other seasons).

Clutch hitting is the idea that some players are better hitters when it is most important for the team (i.e., when there are runners on base or later in a close game). In order to analyze

this idea Mills [39] created a statistic called a player win average PWA to quantify the impact of each at bat on the team's probability of winning. Cramer [21] utilized this statistic in order to evaluate whether clutch hitters exist. He did not find evidence to suggest that, among the players that he evaluated (all batters in the 1969 and 1970 seasons), clutch hitting was present. More recently Albert [5] and James [33] came to similar conclusions.

3.2.5 Distributional Background

The most important distributional family for this work is the Dirichlet distributions which are conjugate prior to the multinomial data that we see most frequently in baseball. The kernel of the Dirichlet distribution is the beta-binomial compound distribution which was introduced by Skellam [53] in 1948. This distribution was then generalized to the Dirichlet-multinomial by Mosimann [41]. A further generalization was proposed by Connor and Mosimann [18] which is now known as the generalized Dirichlet distribution. The nested Dirichlet is a natural extension of the generalized Dirichlet. While the generalized Dirichlet requires a specific nesting structure, the nested Dirichlet is a family of distributions with an unspecified nesting structure.

The Dirichlet family of distributions is most commonly used in baseball research when multiple outcomes of an at bat are of interest. Null [43], [44], Jensen [34], Paisley [45], and Quintana [49] have all utilized Dirichlet distributions for baseball modeling. Null [44] has done the most work on the distribution in a baseball context utilizing maximum likelihood estimation to fit a nested Dirichlet model on batting performance. He required a nested Dirichlet because the between batters covariance of a number of the parameters are positive. In his maximum likelihood approach he was able to fit positive covariances between between walks and strikeouts, but not the other positively correlated variables that he identified. In fact the covariance between the highest correlated pair of variables (walks and fly ball homeruns) was identical to the covariance of the Dirichlet. Jensen also utilizes a Dirichlet distribution in his hierarchical model for evaluating home run hitters. This dissertation will utilize Dirichlet distribution (and not the nested Dirichlet) as our research indicates that between batters the

covariance of a number of parameters are positive, but the modeling that will be outlined in Sections 2.3 - 2.5 is done within a given batter, where the covariances are not substantively different from those assumed in the Dirichlet distribution.

3.3 Background: Metrics for the Batting Model

In order to project future performance one must first evaluate the past. There are many different potential outcomes of an at bat that could be considered. The focus of this dissertation will be on the outcomes that (a) occur with enough frequency to be predictable with some accuracy, and (b) are relevant to fantasy baseball, since that is the focus of this dissertation. The direct outcomes of an at bat that will be the modeling focus are: outs, walks, singles, doubles, triples, home runs, and stolen bases. Outcomes such as catchers interference, reaching on an error and being hit by a pitch are not relevant as individual outcomes for fantasy baseball and as such are not included. Being hit by a pitch occurs with enough frequency, and is not a totally batter independent random phenomenon, so they are included with walks. They are not relevant for fantasy independently but do contribute to a player's on base percentage. Indirect outcomes are important as well, such as runs scored and runs batted in (RBI) but these metrics will be accounted for in the simulation section of the paper and will not be discussed here.

The outcomes will be fit with a Bayesian model utilizing a Dirichlet prior updated with multinomial data. A quadratic age curve will also be utilized to account for a players change in home run rate and batting average over time. The age curve model will also be built utilizing a Bayesian framework which makes use of more of the existing knowledge of the player's underlying ability each season. Once the posterior distribution of the player's abilities for the upcoming season are modeled, seasons are simulated following the methodology outlined in section 3.7.

The data that has the most importance for batters in fantasy baseball is what happens when they are up to bat. Two of the five standard fantasy categories (home runs and batting

average) are directly attributed to the event of their at bat while two others are indirectly attributed to the event (runs and runs batted in). The fifth most common fantasy category is stolen bases, which is dependent on the batting outcomes as well.

3.4 Bayesian Modeling and Distributions

The result of an at bat can be any one of a number of outcomes, how the at bat events are categorized varies slightly but the main events which are always present are: outs, walks, singles, doubles, triples, and home runs. Some separate out the hits into whether they are generated by a ground ball or a fly ball [43], or separate the outs as a result of strikeouts from those as a result of a batted ball. Since the outcomes are discrete, the multinomial distribution is the logical choice. This section is higlighted in Figure 3.1 The multinomial distribution is a generalized form of the binomial distribution. The multinomial also has the family of Dirichlet distributions as conjugate priors which makes the distribution desirable from a Bayesian perspective. Characteristics of the multinomial distribution are shown as follows:

$$pmf \qquad \frac{n!}{x_1!...x_k!} p_1^{x_1}...p_k^{x_k} \tag{3.1}$$

$$E(X_i) \;=\; np_i \tag{3.2}$$

$$\tag{3.3}$$

where $p_1, ..., p_k$ are the event probabilities ($\sum p_i = 1$), n is the number of trials, and X_i denotes the number of observable outcomes in category i.

3.4.1 Dirichlet Distributions

Both the Dirichlet and the nested Dirichlet distributions are conjugate prior to the multinomial distribution. The benefits and limitations of both will be described and a conclusion drawn as to the best distribution for the purposes of modeling batter production

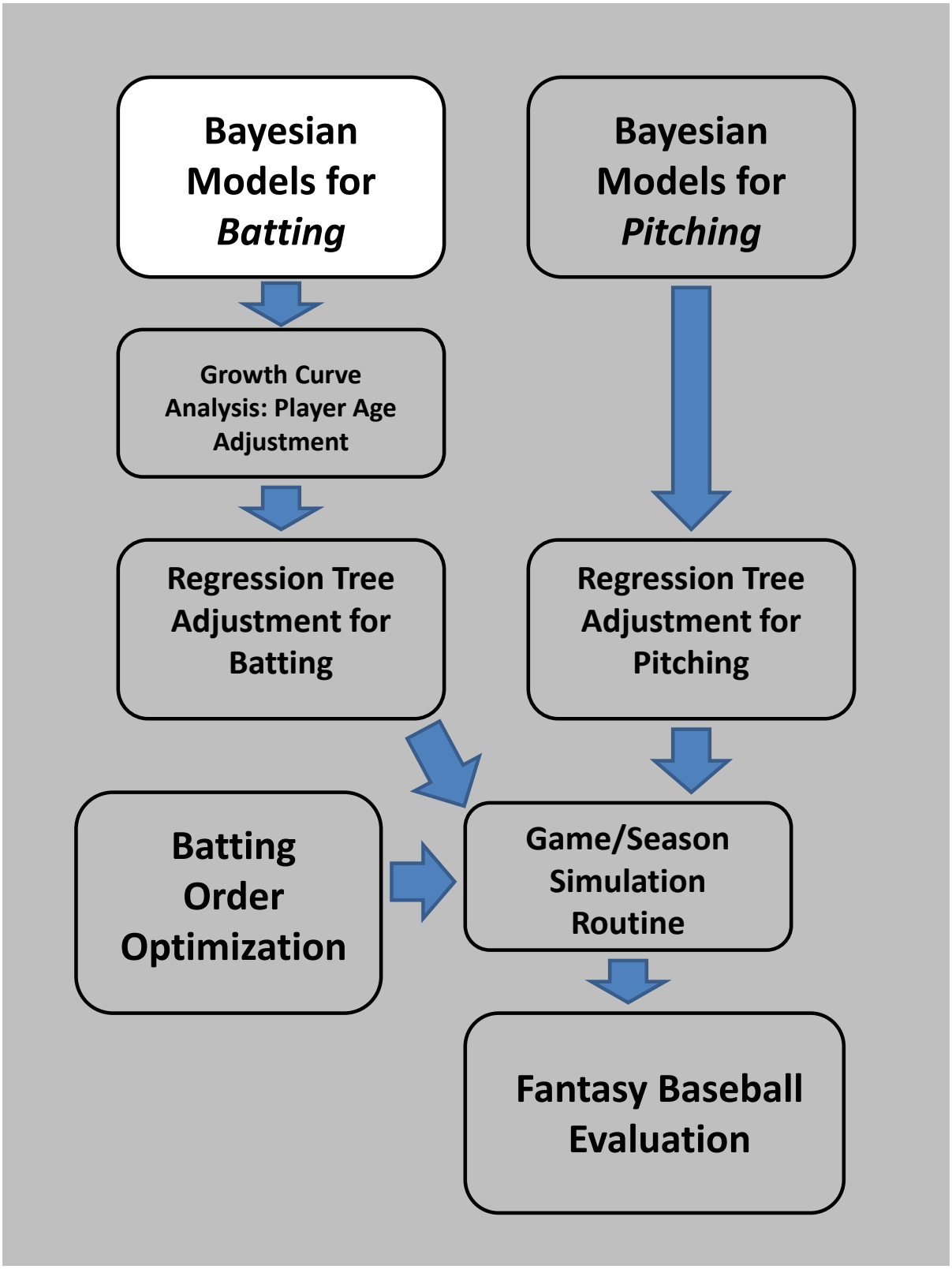

Figure 3.1. Flowchart of dissertation methods highlighting the bayesian batting methods

with fantasy baseball performance in mind. We will begin with the Dirichlet distribution whose joint probability density function is defined as follows:

$$D(\mathbf{x}, \alpha) = \frac{1}{B(\alpha)} \prod_{j \in x_i} x_i^{\alpha_i - 1} \tag{3.4}$$

where $\mathbf{x}$ represents a series of random variables where $x_i \geq 0 \; \forall \; i$ and $\sum_i = 1$. The α vector represents the parameters where $B(\alpha)$ is the multinomial beta function as defined below:

$$B(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(A)} \tag{3.5}$$

$$A = \Sigma_i \alpha_i. \tag{3.6}$$

The benefits of the Dirichlet distribution is that it is easy to use and has a defined structure, to go along with a defined expected value and covariance matrix. The limitation is that the covariance matrix is inflexible, and uniformly negative. The nested Dirichlet distribution is defined by a series of hierarchical Dirichlet distributions, as defined by the user, in order to provide the desired covariance matrix. The joint probability density function of the nested Dirichlet is defined by Null [44] as follows:

$$f(x_1, ..., x_n) = \frac{\prod_{i=1}^{n} x_i^{\alpha_i - 1} \prod_{j=1}^{k} x_{n+j}^{\alpha_{n+j} - \bar{A}_j}}{\prod_{j=0}^{k} B(A_j)} \tag{3.7}$$

where $x_n, ..., x_{n+k}$ are degenerate.

The nested Dirichlet was utilized by Null [43],[44] in order to fit a positive covariance matrix. The positive covariance matrix shown by Null compares the covariance of players marginal home run, walk, and strikeout rates over a four year span from 2003-2006. These covariances are significantly positive indicating on the surface that a nested Dirichlet is the best approach for this type of data. However these results are marginal rates between players

of differing abilities. The overall ability sets do have positive correlations and hence the marginal covariances will be positive. The methodology that is used here to model player ability is independent of other player's data, and hence this positive covariance is irrelevant. Data is utilized across players when evaluating the age curve in Section 3.5, however the age curves are done grouping together outcomes into a beta-binomial distribution.

What is relevant to the analysis of the optimal modeling structure for this approach is whether or not there is a positive covariance of production, given that abilities remain constant. This is a much more difficult question to answer as for an individual player, his ability matrix is constantly changing. An analysis was performed of the correlations within batters between home run, walk, and strikeout rates. Player data from 1990-2011 was utilized and only players with at least ten seasons of 300 or more plate appearances were utilized.

Since we are trying to understand the covariance between walks and home runs within players, once ability step changes have been factored in, the data was further paired down. The first two seasons of data was removed as those seasons are where the players ability fluctuates the most. The number of seasons used per player was also capped at ten in an effort to remove the tail end of a players career when age effects are most pronounced. This also removed outliers from PEDs such as the latter part of Barry Bonds' career.

The players that will be most impacted by a change in methodology are the power hitters, since we are more interested in a player who may go from 30 to 40 home runs in a season than from six to eight. For this reason, the results are displayed for players with a career home run rate of at least .04, or players who averaged at least 25 home runs in a season where they received 625 plate appearances. The means were within one standard error when all players were included regardless of home run rate.

The tables below show the mean and standard errors of the covariances taken from the sixty six player observations.

The correlation between walks and home runs within batters is significantly positive with a one tailed p-value of .0001. The correlations associated with strikeouts are not

Table 3.1. Within Batters Correlation Matrix: Mean

Stat	Walk	Home Run	Strikeout
Walk	1	.176	.023
Home Run	.176	1	.026
Strikeout	.023	.026	1

Table 3.2. Within Batters Correlation Matrix: Standard Error

Stat	Walk	Home Run	Strikeout
Walk	0	.043	.044
Home Run	.043	0	.042
Strikeout	.044	.042	0

Table 3.3. Between Batters Correlation Matrix: Mean

Stat	Strikeout	Walk	Single	Double	Triple	Home Run
Strikeout	1	.006	-.519	-.237	-.122	.163
Walk	.006	1	-.326	-.137	-.181	.222
Single	-.519	-.326	1	.138	.167	-.296
Double	-.237	-.137	.138	1	.100	-.008
Triple	-.122	-.181	.167	.100	1	-.152
Homerun	.163	.222	-.296	-.008	-.152	1

Table 3.4. Between Batters Correlation Matrix: Standard Error

Stat	Walk	Home Run	Strikeout
Walk	0	.021	.018
Home Run	.021	0	.028
Strikeout	.018	.028	0

significantly positive (one tailed p-value for strikeouts and walks is .3014, strikeouts and home runs is .2690), and while they are suggestive of a covariance that differs from that of the Dirichlet distribution it does not significantly so. We must also realize that while the within batters correlations are our best opportunity to sample from a constant distribution, the distribution is not really constant. The slight positive correlations that we see here are likely due to the fact that as a player's ability changes in one area (home run rate for example), the player's ability in other areas changes as well. This can be explained by the fact that the between batters correlation matrix is significantly positive between strikeouts and home runs as well as between strikeouts and walks.

It is also relevant to consider how much of an impact the nested Dirichlet may have on the covariance matrix of the distribution. Null [43] fit the nested Dirichlet using the method of maximum likelihood and identified three positive covariances that he was attempting to account for with this approach (covariances between walks, strikeouts, and home runs). Using the maximum likelihood approach he was only able to account for one of the three positive covariances (between strikeouts and walks) the other covariances were identical to that of the Dirichlet distribution.

As mentioned above given the minimal impact of the nested Dirichlet from prior research [43] coupled with the likelihood that the positive covariances are due to a change in ability rather than a positive covariance in the distribution, it makes the most sense to utilize the Dirichlet in this context. The expected values along with the variance and covariance matrices are defined as:

$$E(X_i) = \frac{\alpha_i}{\Sigma\alpha_k} \tag{3.8}$$

$$Var(X_i) = \frac{\alpha_i(\Sigma\alpha_k - \alpha_i)}{(\Sigma\alpha_k)^2(\Sigma\alpha_k + 1)} \tag{3.9}$$

$$Cov(X_i, X_j) = -\frac{\alpha_i\alpha_j}{(\Sigma\alpha_k)^2(\Sigma\alpha_k + 1)}. \tag{3.10}$$

A batter's outcome matrix fits nicely into a Bayesian model where the prior is a dirichlet distribution updated with multinomial data. All of the batting outcomes are dependent on the batter's environment, with the pitcher being the most influential factor. Over the course of a season, however, we assume that these dependencies are averaged out so that each player is on relatively equal footing. Over the course of the season each team in the league plays every other team in the league, and plays each team at most 19 times. Some of the outcomes that are a direct result of another player's actions, errors for example, are not accounted for in this model. Since the Dirichlet distribution is a conjugate prior to the multinomial has a closed form:

$$Prior \quad - \quad \beta \sim Dir(\alpha_{BB}, \alpha_{1B}, \alpha_{2B}, \alpha_{3B}, \alpha_{4B}, \alpha_{Out}) \tag{3.11}$$

$$Data \quad - \quad X|\beta \sim Mult(X_{BB}, X_{1B}, X_{2B}, X_{3B}, X_{4B}, X_{Out}) \tag{3.12}$$

$$Posterior - \beta|X \sim Dir((\alpha + X)_{BB}, (\alpha + X)_{1B}, (\alpha + X)_{2B}, (\alpha + X)_{3B}, (\alpha + X)_{4B}, (\alpha + X)_{Out}) \tag{3.13}$$

where subscripts $BB, 1B, 2B, 3B, 4B, Out$ represent the outcomes, α_i is the rate of outcome i in the prior distribution, and X_i is the count of outcome i observed in the data.

The nested Dirichlet is also a conjugate prior to multinomial data and is sometimes used to fit this type of data [44]. The benefit of the nested Dirichlet is its flexibility in specifying the covariance matrix of the data. The parameters of a Dirichlet are all negatively correlated with one another, while a nested Dirichlet has the ability to create positively correlated parameters. The reason that some prefer the nested Dirichlet is due to the assertion that walks, strikeouts, and home runs are all positively correlated [44]. While it is fairly obvious that these parameters are all positively correlated between batters it is not obviously the case that they are positively correlated within batters, and within batters is the modeling focus of this dissertation. In fact the correlation tends to be negative within players across

years, especially when a player's developmental years are discounted. The one effect that can create a positive correlation between home run rate and walk rate over the years of a batter's career can be development. It is often the case that a players walk rate increases during the same season that his power, or home run rate increases. This suggests a step change in the underlying ability, and not positive correlation of outcomes within a particular year. As such this positive correlation can be discounted and there is no need for the added complexity of the nested Dirichlet. Decision trees were utilized to assess the step change in ability mentioned above, which will be discussed in Section 3.8.2.

3.4.2 Baserunning and Final Comments

Baserunning is also built into the model. While baserunning is dependent on other players on the field, some of these dependencies will be accounted for in the simulation routine section, and others will average out over the course of the season. A nested Bayesian model will be used for stealing bases. Distributions (3.14,3.15,3.16) below show the binomial distribution of the number of steals given the number of attempts with a beta conjugate prior resulting in the closed form posterior distribution in (3.16). The data utilized to update this model are the number of attempts $N_{attempts}$ and probability of success p_{sucess}, that the player has achieved over the past three seasons.

$$Prior \quad - \quad \beta \sim Beta(1,1) \tag{3.14}$$

$$Data \quad - \quad X|\beta \sim Bin(N_{attempts}, p_{success}) \tag{3.15}$$

$$Posterior \quad - \quad \beta|X \sim Beta(\alpha, \beta) \tag{3.16}$$

The probability of attempting a stolen base utilizes the same conjugate-prior distributions where the number of opportunities to steal second base are the total number of times a player is on first with second base empty over the course of the past three seasons. This model will only account for the stealing of second base. Attempting to steal third base

and home are both possible, but are rare events that are too dependent on other players on the field (most notably the pitcher) to model with accuracy. The rare nature of the events also makes them less relevant for fantasy purposes. In 2013, 85% of the attempted steals were players attempting to steal second base. There were 3404 attempts to steal second base in Major League Baseball, compared with 522 attempts to steal third and only 59 attempts to steal home. Of the attempts 2386 attempts to steal second base were successful, compared with 384 successful attempts to steal third base and 37 successful attempts to steal home.

The number of games played over the course of a year is also modeled with a beta-binomial conjugate prior distribution. Modeling this is more complicated since there are a number of distinct reasons why a player who generally starts may not start on a particular day. These classifiers will be grouped into two main categories: short term and long term. The short term will account for everyday reasons like a routine day off, sitting a player versus a particular pitcher, platoon situations, and minor injuries. These factors are more predictable and modeled well in a beta-binomial conjugate family as described in Equations 3.17

$$\begin{aligned} Prior &- \beta \sim Beta(1,1) \\ Data &- X|\beta \sim Bin(N,p) \\ Posterior &- \beta|X \sim Beta(\alpha,\beta). \end{aligned} \tag{3.17}$$

Here N represents the number of games over the past three years for the given player that were not impacted by major injuries and p represents the probability that the player will start in any given game defined by N.

The long term focuses on longer term injuries of a month or more. These are more random events than the short term and the question with the long term injuries is how likely it is to happen, rather than how many days will a player miss over the course of the year as a result. Long term injuries will be introduced as a probability of occurring, and when they do occur will significantly impact the posterior distribution. Short term injuries are modeled

utilizing the past three years of data, while long term injuries were modeled with decision trees to project average number of games missed. Long term injuries were also tested by aggregating the results from a collection of unpruned trees, and while the average number of games missed is slightly more accurate than the single decision tree, aggregating the trees significantly under estimates the variance in games missed, while the single decision tree was able to be created to replicate our level of uncertainty.

The distributions outlined above assume a static underlying distribution which is not the case. Each player's abilities are constantly changing, but the most significant changes happen from year to year. In the context of this dissertation we will not attempt to identify the changes in a players abilities over the course of a year, although the Bayesian framework is set up to easily adjust a player's underlying ability distribution in the light of new data.

The two significant ways that a player's batting ability changes are the rate at which they are able to hit safely (batting average) and the power with which they are able to hit the ball. There are a number of power metrics one could use, and we will use both home run rate and isolated power (slugging percentage - batting average) to model changes in ability. The following two subsections will outline the methods used to assess changes in each player's underlying abilities. Nonlinear growth curves are sometimes used [43] to address the change in player abilities over time. In the next section, we discuss a quadratic curve which was found to fit the data reasonably well.

3.5 Quadratic Age Curve

One method to assess changes over time would be to use a time series regression, but that is less effective when information is known about the expected trajectory of the distribution. A more common method to address changes in player ability over time are non-linear growth curves [43]. A quadratic curve fits the data reasonably well and will be modeled for both batting average and home run rate using the algorithm outlined below as shown in Figure 3.2.

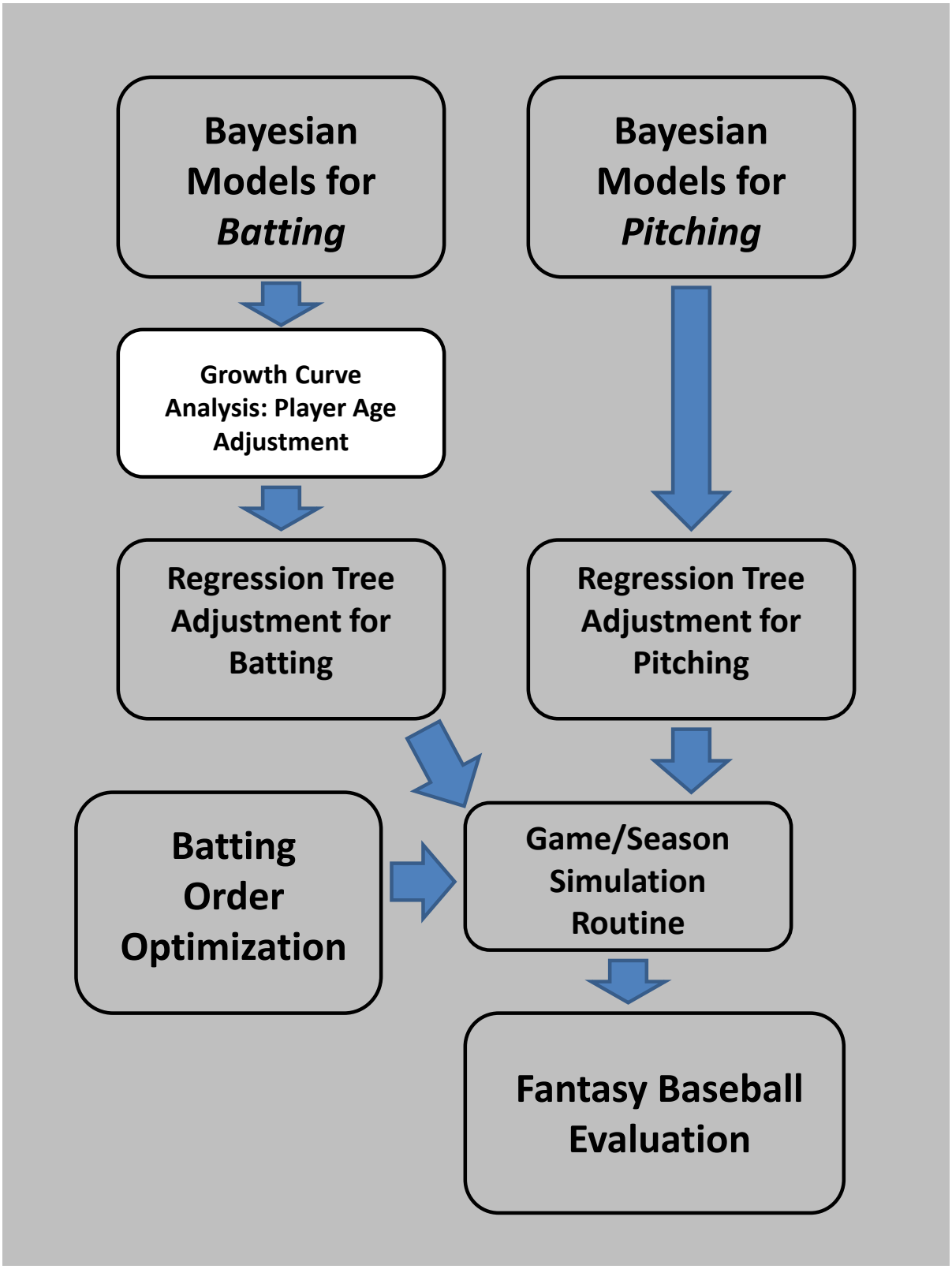

Figure 3.2. Flowchart of dissertation methods highlighting the growth curve analysis

All players who entered Major League Baseball after 1986 and were at least 32 years of age prior to the 2007 baseball season were taken and the overall batting average and home run percentages by age were calculated. Players with high evidence of using performance enhancing drugs were removed (i.e., confession or citation in the Mitchell report). Interestingly enough of the players with high evidence of performance enhancing drugs did not follow the same distribution as many of them appeared to be impervious to the deteriorating effects of age on their abilities (i.e., Barry Bonds). The resulting proportions closely followed a quadratic distribution with a peak at the age of 28.

Each player was then fit with their own quadratic age curves. Due to the random variability of the data and its limited nature (generally 5-15 seasons of data available per player) the results were more stable if there were limitations to the quadratic. The peak of the MLB population occurred at 28, and was the most stable parameter across players, so that was fixed for all players. It was also found that where there were fewer than five seasons of data available the slope of the quadratic was also unstable. The parameters for slope were fixed to the population average in this case as shown in equations (3.19, 3.21). Furthermore when a player had at least five seasons of data the quadratic was tested to ensure that it was concave, indicating a maximum at 28 rather than a minimum, and if it was not, the slope was fixed for those players as well. For the rest of the players the quadratic was fit as shown in equations (3.18, 3.20)

$$HomeRun \quad rate \quad = -\beta_1 \times Age^2 + 56 \times \beta_1 \times Age + \beta_2 \tag{3.18}$$

$$HomeRun \quad rate \quad = -.00015 \times Age^2 + .00961 \times Age + \beta_2 \tag{3.19}$$

$$Batting \quad average \quad = -\beta_1 \times Age^2 + 56 \times \beta_1 \times Age + \beta_2 \tag{3.20}$$

$$Batting \quad average \quad = -.00015 \times Age^2 + .00961 \times Age + \beta_2. \tag{3.21}$$

The quadratic models were fit utilizing the Gauss-Newton algorithm for least squares. As such it does require the model coefficients to be initialized. The initial conditions were set at .1 for β_1 and zero for β_2. The model is robust to initial conditions and was fit within two iterations in almost all cases. However the models are very sensitive to outliers and the resulting variance is not reflective of the actual uncertainty surrounding the estimate. To account for this, an iterative Bayesian process was utilized. The quadratic models were first fit leaving out the player's first season of data, the model forecast for the first season was the prior and it was updated using the data of that year's observations. Since there was more data for the prior than the observation the prior was given more weight by a three to one ratio. This created a posterior estimate for the player's ability (either home run rate or batting average) for their first season. The posterior was then used to create the prior for the second season of data. The second season was updated in the same manner and the posterior for the second season was then utilized for the third season's prior, and so on. Upon completion of one full iteration a posterior estimate had been created for the players underlying ability level utilizing all of the players available data. This process provided a clearer understanding of players underlying ability at each season and was then utilized to fit the final quadratic curve to project the player's ability for the upcoming season.

These results from the quadratic approach were then fit into the Dirichlet distribution obtained from the past three seasons of data. The new home run percentage directly replaces the original value from the Dirichlet, while the batting average value must be inserted more creatively since there is no direct value for batting average in the Dirichlet. The probabilities of: single, double, triple, and out are scaled from the posterior distribution so that the posterior batting average fits the quadratic model result as follows:

$$\bar{x}_{1bq} = \bar{x}_{1b}\left[\frac{\bar{q}_{ave} - \bar{q}_{hr}}{\bar{x}_{1b} + \bar{x}_{2b} + \bar{x}_{3b}}\right]$$
$$\bar{x}_{2bq} = \bar{x}_{2b}\left[\frac{\bar{q}_{ave} - \bar{q}_{hr}}{\bar{x}_{1b} + \bar{x}_{2b} + \bar{x}_{3b}}\right]$$
$$\bar{x}_{3bq} = \bar{x}_{3b}\left[\frac{\bar{q}_{ave} - \bar{q}_{hr}}{\bar{x}_{1b} + \bar{x}_{2b} + \bar{x}_{3b}}\right] \tag{3.22}$$

where $\bar{x}_{1b}$ represents the Dirichlet mean posterior probability of a single, $\bar{x}_{2b}$ that of a double, and $\bar{x}_{3b}$ represents that of a triple. The final, quadratic adjusted, posterior probabilities are represented by: $\bar{x}_{1bq}$, $\bar{x}_{2bq}$, $\bar{x}_{3bq}$ respectively. $\bar{q}_{ave}$ represents the quadratic point estimate of the batting average and $\bar{q}_{hr}$ represents the quadratic point estimate for home run rate.

3.6 Predicting Breakout and Regression

In order to project breakout and regression candidates we need to evaluate the model error rates and see if we can predict the errors and correct the model as shown in Figure 3.3. To do this the model was run forecasting 2011 performance and the model error rates of batting average and isolated power were analyzed. Isolated power is an alternative approach for evaluating power to home run rate. Both isolated power and home run rate were tested and the errors were found to be more predictive when isolated power was used. This is done both to validate the quadratic assumption and to test for regression to the mean. If there is regression toward the mean then the higher predicted values for batting average and isolated power will be decreased by the regression trees.

The trees were built using player age, experience, predicted values for batting average and isolated power, as well as performance metrics for the four prior years. The model was built using data prior to 2012 in order to predict player performance for 2012, and then tested predicting 2013 performance. Only players projected to come to bat (plate appearance) at least 250 times were used in the training set. The predictions, rather than actual plate appearances in the training set, were used as the cutoff because we are very interested to see how the model may be biased on young players with limited Major League experience. Bootstrapped,

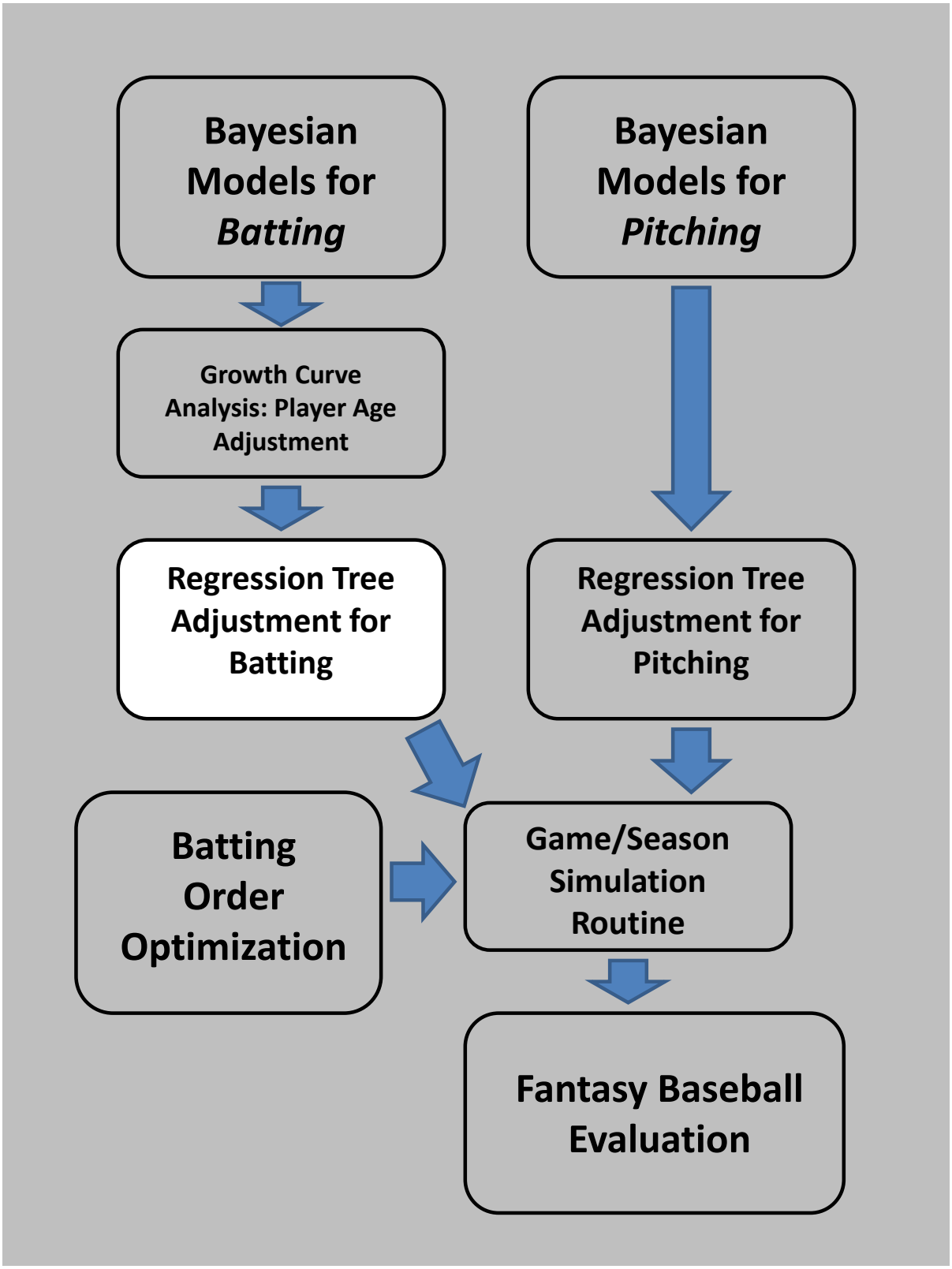

Figure 3.3. Flowchart of dissertation methods highlighting the regression tree methods

unpruned trees, were built and used to validate the errors. The trees improved the model accuracy as shown in Figure 3.4. The next item that we vetted was the model variance. We found when utilizing too many trees the parameter variances were underestimated. Parameter variances tended to be higher when there was less data available, and less so when more data data was available (as mentioned in Section 3.4.1 a limitation of the Dirichlet distribution is that the variances are not flexible). Utilizing the trees to adjust the posterior distribution has the benefit of providing variance flexibility in the final posterior distribution. We found that six hundred plate appearances (approximately one full season) was the point at which the observed variances reduced significantly. Utilizing one tree for players with fewer than six hundred career plate appearances and calculating the average of four trees for players with at least six hundred plate appearances matched the observed variability in the outcomes.

Figure 3.4 compares the error rates pre and post tree implementation. The negative values are over estimates and the positive values are under estimates. The distribution of the error rates are approximately normal for all of the statistics, although they do all have shorter negative tails and longer positive ones, the most pronounced of which is for stolen bases. When looking at the runs scored and runs batted in (RBI) in Figure 3.4 we can see that they both had a positive bias (over estimates) before the tree implementation (an average of 12.6 runs and 10.0 runs batted in) and were brought down to within 2.5 runs and runs batted in after the implementation. The error rates did not change significantly for home runs, stolen bases, batting average, and slugging percentage.

3.7 ALGORITHM

We run the algorithm with data prior to 2013 to project 2013 player performances. The performances are run iteratively team by team in conjunction with the pitcher performances which will be outlined in detail in the next chapter as shown in Figure 3.5.

1. Begin with an non-informative prior
2. Aggregate the last three years of data

Figure 3.4. Histograms of the errors and how they have decreased as a result of utilizing decision trees to predict breakout performances and regression tendencies both as a result of age and a declining skill set or as regression to the mean.

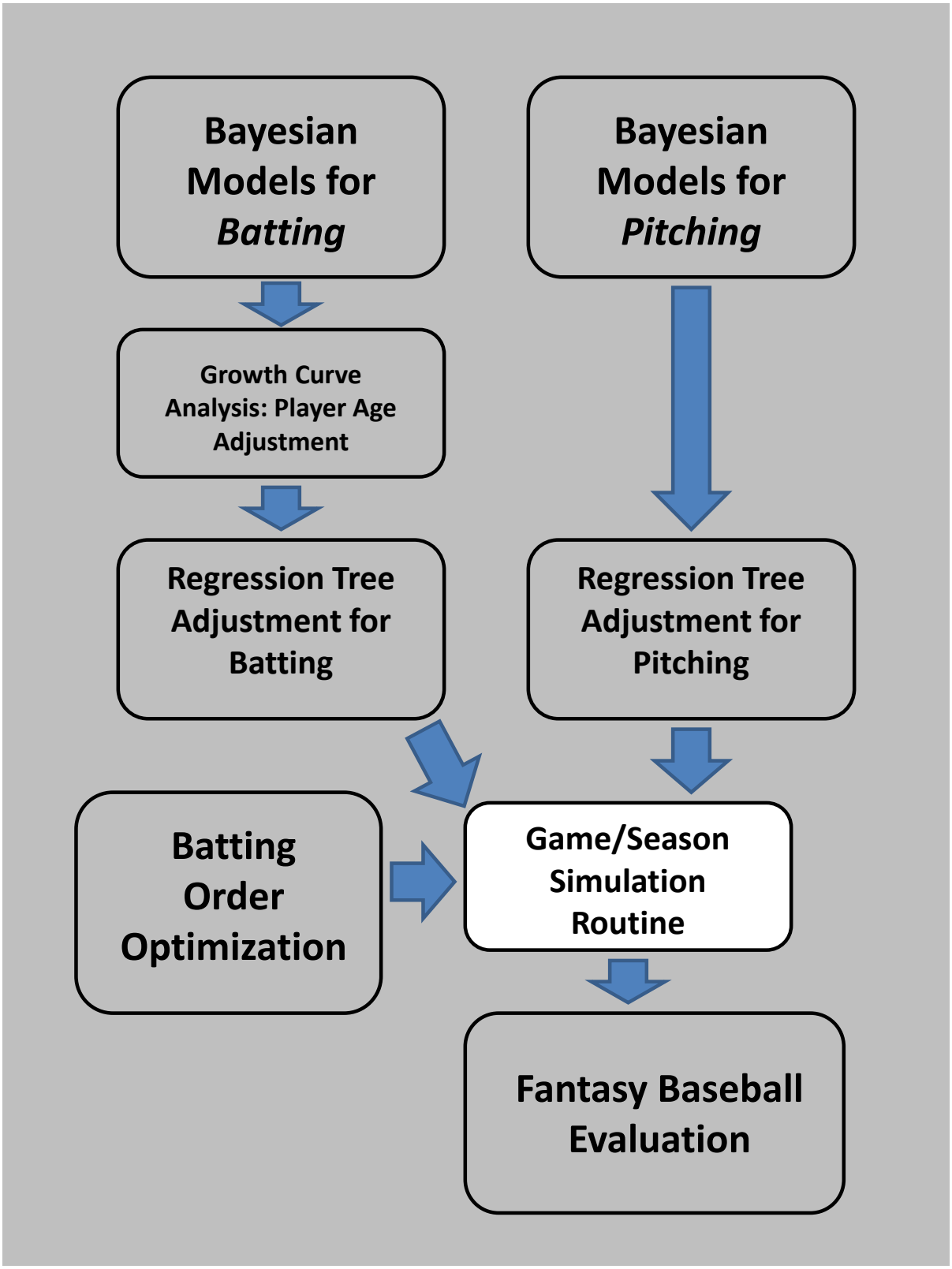

Figure 3.5. Flowchart of dissertation methods highlighting the simulation routine

3. Utilize aggregated data to update the prior
4. Then utilize batting average and home run rates for each season to build a quadratic curve for each
5. Beginning with the players rookie season build a quadratic function utilizing all the other available years of data to forecast the batters home run and batting average ability during their rookie season
6. Utilizing the players data as the prior, update with the quadratic result weighting the quadratic update with a ratio of three to one
7. Perform the prior two steps for each other season of data, until all seasons have been updated in the same manner
8. Build a final quadratic curve utilizing the updated point estimates for the players underlying ability in each season
9. Forecast batting average and home run rate for the upcoming season
10. Directly insert the forecasted home run rate into the Dirichlet posterior
11. Scale singles, doubles, and triples in Dirichlet so the posterior Dirichlet batting average matches the quadratic average
12. Batting order is determined by actual batting orders for the prior season as identified through retrosheet.org [40]. Changes are made based on the new roster and projected changes in the players role (as well as to account for injuries over the prior season)
13. Begin Seasonal Simulation
14. Run the players data through the regression trees to update average and isolated power in the posterior Dirichlet, forecasting breakout and regression performances
15. Adjust isolated power for the players that breakout or regress by scaling doubles, triples, and home runs
16. Adjust batting average to the final average as a result of breakout or regression by scaling singles to obtain batting average
17. Simulate production for the leadoff batter and record event(s) that take place (out, hit, baserunners, runs scored, rbi)
18. When runners are on base, players advance with the same probability that they have advanced in their prior two years (move from first to second or third on a single etc.)

19. If a runner is on first base and there is no runner on second simulate the probability of attempting a stolen base, and if a stolen base is attempted simulate the probability that it is successful and record the outcome

20. Repeat items 17 to 19 until there are three outs

21. Once three outs are reached clear the on base vector

22. Simulate a pitching inning

23. Begin next inning with the next batter in the batting order. After the ninth batter has batted, cycle back to the first batter

24. Repeat items 17 to 22 until nine innings have been played

25. Repeat game simulations until all starting pitchers have pitched their allotted number of starts

26. Record all player productions and begin a new season. Repeat 10 seasons and record the averages

3.8 RESULTS

The model was run to predict performance for the 2013 season. The goal of the model is to be able to rank players from best performing to worst, as well as to understand what categories will be the player's strengths and weaknesses, making it easier to build a balanced team in each of the categories. This is important in most fantasy baseball formats, as comparisons are made category by category and either all the teams are ranked for each category over the course of the season, or teams are matched up head to head with the winner of each category getting a point. There is also a simpler, but less popular methodology where each result is given a point value and the points are totaled to determine the winner. This dissertation will focus on the categorical style of fantasy baseball scoring, but it is straight forward to sum up the total number of points based on the player's production and rank the players that way as well. The fantasy baseball evaluation step is highlighted in Figure 3.6

There are still a number of other challenges that must be addressed in order to evaluate the results. First we expect that there will be a fairly wide variance in production as the season by season variability is quite high. To compare the model's results to a baseline we will look

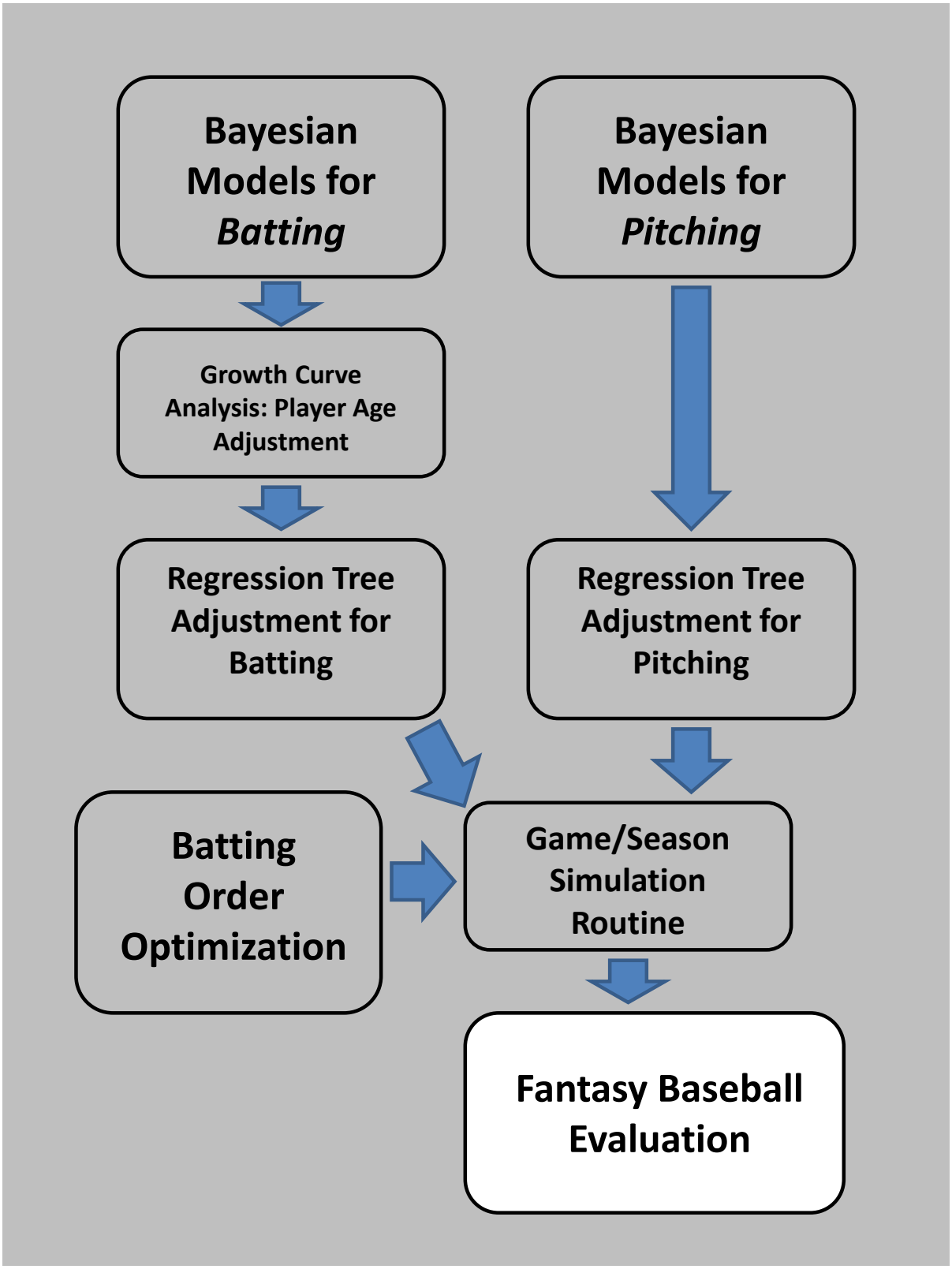

Figure 3.6. Flowchart of dissertation methods highlighting the fantasy baseball evaluation

at two popular magazine rankings as the best in class comparisons. The two magazines are Rotoworld, which is affiliated with the biggest fantasy baseball site on the web, Yahoo sports, and Athlon sports, which is a premier sports publisher that is also well regarded for their fantasy magazines.

The next challenge is how to compare the three results. The way that most fantasy baseball publications publish their results is through rankings. Each publication will rank players at each position and also provide an overall composite ranking blending all players together. We will provide the same, based on the model results, as well as ranking the overall production of the players. This methodology will also be utilized when looking at the pitching model results in the next chapter.

Blending the statistics together first necessitates comparing each category on equal footing. The categories that we will look at for these results will be the five basic fantasy baseball categories on which the comparison rankings are also based: runs scored, runs batted in, home runs, batting average, and stolen bases. Of these, all of these categories but stolen bases can be approximated with a normal distribution. As shown in Figure 3.7 the counting statistics (runs, home runs, runs batted in) are positively skewed but the normal distribution still does a good job as an approximation, so creating z-scores for each of categories will work well. Stolen bases is the trouble spot, more closely approximating a Poisson distribution, albeit with a few outliers. The biggest challenge that this poses is that stolen bases are a more rare commodity, and hence more valuable. Integrating base stealing poses a significant problem, it is important not to weight stolen bases too much or we may severely over-value premier base stealers. It is also important not to under value them, or it will be difficult to obtain a well rounded team. Since home runs, runs batted in, and runs scored all have a significant positive skew, it is possible to extrapolate stolen bases so that the distribution more closely resembles those three, by creating additional negative data to create a more normal looking distribution, as shown on the bottom of Figure 3.7. The adjusted distribution reduces the standard deviation from 10.0 to 8.2 making stolen bases effectively equivalent in value to

home runs (standard deviation of 8.5). The standard deviation was robust to changes in the extrapolated portion of the distribution. This is not a large change but it does have a maximum impact of 1.3 standard deviations on the premier base stealer of 2013. This methodology does a good job of giving the players that contribute in all categories including stolen bases, a significant edge over those who do not steal bases. A great example of this is the question of who should be drafted first overall, Miguel Cabrera or Mike Trout? Miguel Cabrera is better in most categories but he does not steal bases while Mike Trout is a significant asset on the basepaths.

3.8.1 Model Accuracy

Once we have an overall metric for assessing the value of players, there are still a number of different ways to look at accuracy. At first glance we could look at the top two rows of Table 3.5 and claim to have done a great job at evaluating the best overall performers, which is true, but it is also not exceptional as most evaluators had Miguel Cabrera and Mike Trout ranked first and second, or at least in the top five overall. It is important to be able to classify those whom most are evaluating as the best overall, but what can give the fantasy baseball drafter the edge when constructing their team is being able to predict those who will outperform the experts ranks and those who are ranked too highly for where their actual performance will merit at the end of the season.

Table 3.5 on the following pages provides the player ranks for all batters who were ranked in the top 25 in any of the four metrics. They are sorted by actual performance so it is easier to compare across rankings quickly. The last two columns are assessments of the model rank. The column titled Relative Rank compares the model rank to the two best in class rankings. For example the top three rows (Miguel Cabrera, Mike Trout, and Chris Davis) are all classified as Average because the model ranks are right in line with the best in class rankings. Paul Goldschmidt is identified as High because the model rank of 17 is significantly higher than the best in class ranks of 39 and 46. The last column titled Value compares the median rank to the actual rank. Again the first two rows are Average because the median rank

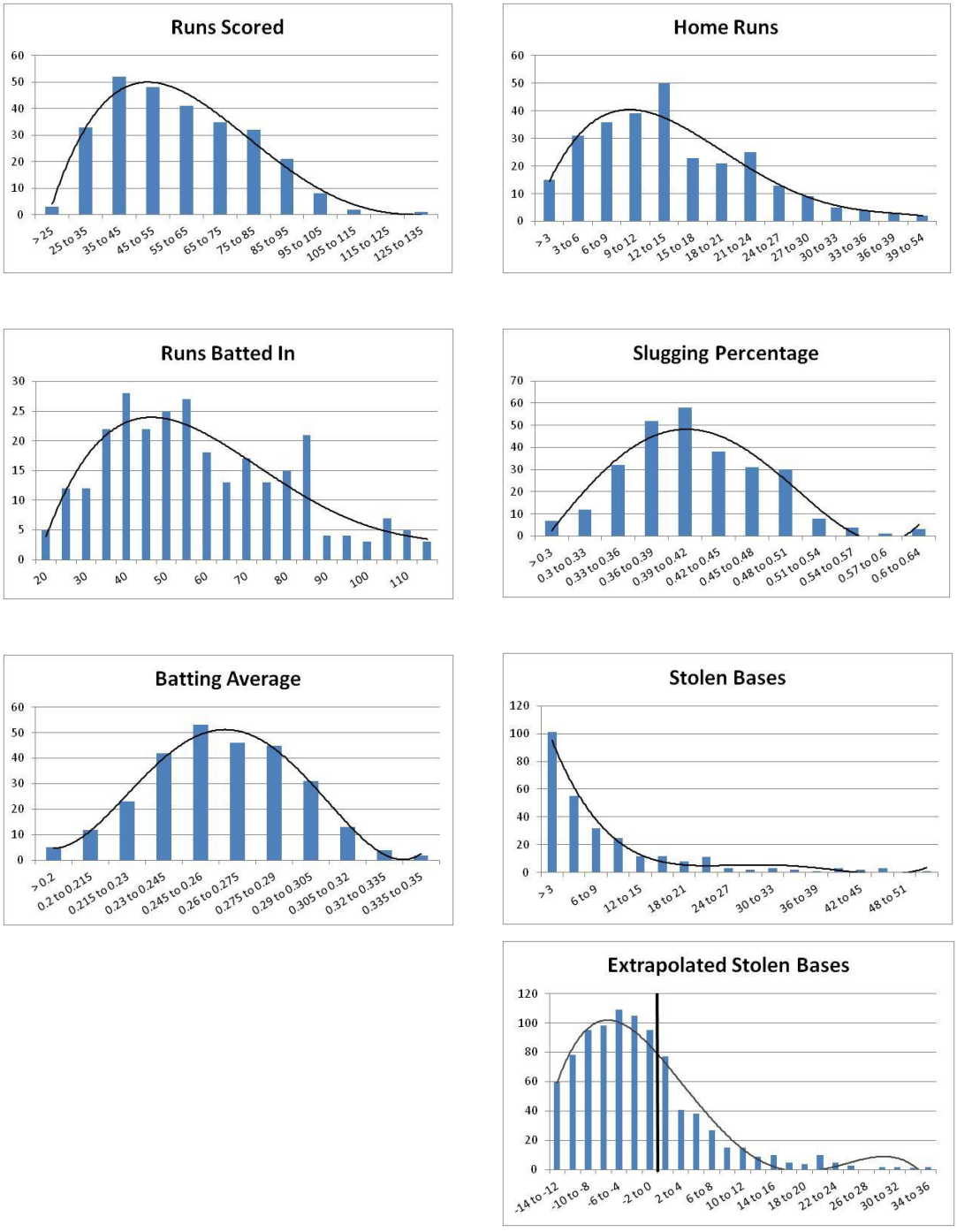

Figure 3.7. Histograms of observed statistics for players with at least 300 plate appearances in the 2013 Major League Baseball season.

(2 and 1 respectively) was right in line with the model rank. The next two rows were identified as very high and high respectively because the actual ranks of 3 and 4 were substantially higher than the median ranks of 103 and 39. When the last two columns are in alignment that indicates that the model did a better job on the player than the best in class models. When they are not in alignment the model did not do as good a job as the best in class models (e.g., Alex Rios actual rank was 7 but the model classified him as the 102 best player. The best in class models ranked him 59 and 48 indicating that all of the rankings missed substantially on Alex Rios but the model ranking was worse than the best in class rankings).

Using the last two columns as a guide, the model did a good job projecting players compared with the best in class. Of the 52 rankings on Table 3.5 the model result agreed with the best in class rankings on only 14 and disagreed on 38 of them. When evaluating the discrepancies between the model rank and best in class, the model projected the players more accurately than the best in class models for 26 of the 38 discrepancies for a success rate of over 68%. If the same analysis is performed comparing the Rotoworld ranking to the model rankings and Athlon Sports they achieve a 37% success rate. Comparing Athlon Sports rankings to the others yields a similar result of 36% accuracy.

Another way to assess model accuracy is to measure the distance between the forecasted rank and the actual rank for each of the rankings. Since accuracy is most important at the top of the rankings, a weighted scheme is appropriate. We measured the distance between the forecasted and actual ranks for the top 25 batters using weights of the square root of the reverse of the weighting (sqrt(25), sqrt(24) ...). A cap of 99 was used on the rankings both to mitigate the impact of outliers and to account for the missing values from some of the rankings. The model had an average miss of 29.3 positions, while Roto World missed by 36.2 and Athlon by 34.9.

Table 3.5. Ranking of Results

		Model	Actual	Roto	Athlon	Relative	

Player	Team	Rank	Rank	World	Sports	Rank	Value
Miguel Cabrera	Det	2	1	2	1	Average	Average
Mike Trout	Ana	1	2	1	3	Average	Average
Chris Davis	Bal	103	3	96	123	Average	Very High
Paul Goldschmidt	Ari	17	4	39	46	High	High
Andrew McCutchen	Pit	5	5	29	10	High	High
Jacoby Ellsbury	Bos	23	6	17	26	Average	High
Alex Rios	CHW	102	7	59	48	Low	Very High
Adam Jones	Bal	10	8	38	45	High	High
Carlos Gomez	Mil	30	9	103	74	High	High
Hunter Pence	SF	13	10	71	69	Very High	High
Jason Kipnis	Cle	12	11	37	27	High	High
Alfonso Soriano	CHC	185	12	94	140	Low	Very High
Jean Segura	Mil	151	13	107	83	Low	Very High
Edwin Encarnacion	Tor	37	14	34	56	Average	High
Robinson Cano	NYY	11	15	4	5	Low	Below Ave
David Ortiz	Bos	36	16	65	50	High	High
Elvis Andrus	Tex	25	17	36	39	High	High
Freddie Freeman	Atl	83	18	42	49	Low	High
Matt Holliday	StL	63	19	41	53	Low	High
Jayson Werth	Was	129	20	81	135	Low	Very High
Carlos Gonzalez	Col	9	21	5	6	Low	Low
Adrian Beltre	Tex	48	22	21	15	Low	High
Jay Bruce	Cin	15	23	45	40	High	Average
Shin-Soo Choo	Cin	49	24	57	41	Average	High
Matt Carpenter	StL	116	25	123	133	Above Ave	Very High

Joey Votto	Cin	3	29	9	11	High	Low
Dustin Pedroia	Bos	22	33	33	25	Average	Below Ave
Evan Longoria	TB	4	36	11	23	High	Low
Prince Fielder	Det	19	39	24	13	Average	Low
Justin Upton	Atl	14	40	8	19	Average	Low
Ryan Zimmerman	Was	38	43	28	14	Low	Low
Hanley Ramirez	LAD	16	44	13	34	Average	Low
Troy Tulowitzki	Col	46	45	6	8	Low	Low
Adrian Gonzalez	LAD	18	47	10	29	Average	Low
David Wright	NYM	41	49	16	28	Low	Low
Ian Kinsler	Det	21	50	23	35	Average	Low
Jose Bautista	Tor	6	54	15	21	High	Low
Yoenis Cespedes	Oak	52	70	20	20	Low	Very Low
Bryce Harper	Was	108	72	12	24	Very Low	Very Low
Austin Jackson	Det	20	76	54	33	High	Low
Carlos Santana	Cle	24	82	35	42	High	Low
Josh Hamilton	LAA	39	86	52	12	Low	Very Low
Buster Posey	SF	40	94	14	17	Low	Very Low
Billy Butler	KC	26	96	31	22	Average	Low
Jose Reyes	Tor	28	102	19	7	Low	Very Low
Joe Mauer	Min	105	115	30	16	Very Low	Very Low
Giancarlo Stanton	Mia	32	117	18	32	Low	Very Low
Pablo Sandoval	SF	74	122	25	71	Low	Very Low
Albert Pujols	LAA	8	137	7	4	Low	Very Low
Starlin Castro	CHC	44	146	22	18	Low	Very Low
Jason Heyward	Atl	50	149	32	9	Low	Very Low

Matt Kemp	LAD	7	166	3	2	Low	Very Low

3.8.2 Breakout vs Outlier

While the model did miss on a number of the top performers in 2013, a very reasonable question is: Are these results model misses, or outliers? One way to assess this is to see how the players performed in the following season. If the model missed, then the players should sustain their performance, to some degree at least, while if they are performance outliers one would expect them to regress back in line with previous expectations.

Of the top 20 performers in the 2013 season, seven of them were ranked by the model outside the top 40. These could easily be seen as significant model misses. However if the players were ranked consistently in our ranking when compared to the best in class models, we would still potentially obtain those players. Furthermore, if our ranking was significantly higher than the best in class we would be likely to obtain them. Unfortunately our ranking was only in line with the best in class on two of them, and below the other rankings on the remaining five, leaving us little chance of drafting most of those players.

Now lets examine the following season of the seven players and examine their production in the 2014 baseball season. Of the seven players only Jayson Werth and Freddie Freeman finished in the top 40 in fantasy scoring for the 2014 season (neither was in the top 20). For Freddie Freeman 2013 marked a breakout campaign and he will likely remain a top 40 fantasy batter for years to come (he was 23 years old in 2013). Jayson Werth has been among the top 50 fantasy batters since 2008, however he was 34 years old in 2013 and while he has maintained a high batting average causing the model to underestimate his production, his home run rate did drop significantly in 2014 as he begins to show signs of his age. The remaining five players (Alex Rios, Chris Davis, Alfonso Soriano, Jean Segura, Matt Holliday) have all regressed significantly the worst of whom, Alfonso Soriano, was actually released by

his MLB team. There has also been discussions sending another, Jean Segura, back down to the minor leagues. These two players were ranked the lowest in our 2013 rankings.

These results are helpful, but since the sample size is very small it will be more enlightening if we can examine the consistency of these results in different seasons. For this reason we evaluated 2012 and 2014 under the same criteria. Overall for the three seasons there were fifteen players that finished among the top twenty fantasy producers and yet the model ranked them outside the top forty. Of these players the model was significantly better than best in class rankings for four of them, and at least in line with the best in class rankings for eight players leaving only seven of the sixty top twenty performers that the model significantly missed on and we would not likely have the opportunity to draft, giving us an 88% success rate by this metric. Only three of the players that we would not have drafted, and six overall, maintained top forty performance in the following season.

CHAPTER 4
PITCHING MODEL

This chapter will go over the pitching analysis from the literature review to models, results, and discussion. Since the focus is on predicting performance for fantasy baseball, it is important to understand which statistics are most important in this context as well. There are five core pitching categories in fantasy baseball: wins, strikeouts, earned run average (ERA), walks plus hits per inning pitched (WHIP), and saves; with other categories such as quality starts, losses, and innings are frequently used as categories as well. This chapter will target accurately predicting these eight summary statistics for each player over the course of the season.

This chapter will address the season in the same manner as the batting model chapter, projecting performance for a given season based on prior seasons of data in anticipation of a fantasy draft which occurs before the season starts. Projecting a starting pitcher's performance can be broken down into three interrelated areas: number of games started, how many batters a pitcher faces during a given start, and the outcomes of the batters that the pitcher faces. Each of these areas will be discussed separately in the subsequent sections, as well as the algorithm to bring the methodologies together. Projecting relief pitchers performance is slightly different, the most sought after asset from a fantasy perspective is saves, so special attention will be given to that category. Saves are of primary consideration for relief pitchers because no other pitchers can generate that statistic and there are at most 30 pitchers generating that statistic on a regular basis. Saves are an elusive category because they are predicated on opportunity. The manager decides who will have the opportunity to accrue saves for his team, so predicting who the manager will choose both to start the season and whether that player will maintain that role or another player will take over is very important. There will be

opportunity to evaluate pitchers who will not accrue saves to start the season but to assess their likelihood to step into that role as the season unfolds.

4.1 Literature Review

As mentioned in the batting literature review of Section 3.2, baseball has been analyzed statistically for decades. However, the bulk of the analysis has been focused on the batting outcomes, the work on pitcher evaluation has been relatively sparse. That being said there has been some work done evaluating the use of pitchers [60], [30], [31] assessing the predictability of their statistics [8], [38], [47] and the impacts of age on pitching [25], [52], [24]. Since this dissertation is more interested in evaluating pitchers how they are used currently and not optimizing their use, we will focus our literature review on the latter two topics.

4.1.1 Pitcher Performance Predictability

Not surprisingly, luck plays a major role in identifying top performers [8]. What about the general case? As with hitting, defense impacts a pitchers statistics. There are some statistics that are not (noticeably anyway) impacted by the defense. These are strikeouts, walks, and home runs. Most of the categories that relevant to fantasy however, rely heavily on the defense. Wins, Losses, ERA, WHIP, quality starts, and innings pitched are all highly predicated on the defense.

When looking at the defense independent statistics of strikeouts, walks, and home runs McCracken found that past performance is a good indicator of future behavior [38]. Strikeouts and walks of the prior season were found to be highly correlated with the strikeouts and walks of the subsequent season. Home runs allowed were also correlated, but not as strongly as strikeouts and walks. This is not surprising as home runs are a much more rare occurrence and vary from batter to batter.

The results found for the defense dependent statistics were much more surprising. McCracken focused on hits allowed (less home runs) and ERA evaluating their predictability

[38]. He found that hits allowed and ERA were not very predictable year over year. He further studied the hit rates allowed of the entire team, rather than the individual player, and the resulting ERA was more predictable than the actual ERA from the prior season. This result suggested that the defense behind the pitcher was a better predictor for hits on balls in play than the pitchers ability. This result will be analyzed later.

4.1.2 Age Impacts of Pitchers

There has been a fair amount of work done in this area, as those analyzing the age impacts of batters tend to do the same for pitchers. Both Fair [25] and Schulz [52] looked at the impacts of age on a number of summary statistics over time. Dun et al. [24] analyzed pitcher kinematics for a small group of pitchers of various ages. They concluded that the pitcher's delivery motion changes over time, based both on experience and physiological changes.

While all of the analysis reviewed suggested changes over time to a pitcher's performance, and hence fantasy value, the type and rate of change was not consistent across the results. Dun et al. [24] focused on the kinematics suggesting changes were happening, but did not attempt to quantify the impacts of those changes. Fair and Schulz both quantified the changes, but the results were significantly different. Both Fair [25] and Schulz [52] assessed the impact of age on ERA, with the expected peak age differing by more than 2.5 years (26.54 in Fair's analysis, compared with 29.11 in Schulz). Fair presented a standard error of 1.4 years, while Schulz presented a standard error of 4.56. Both samples were similar in size (144 to 153). The methods for collecting the sample were significantly different, as Fair required ten seasons of at least 150 innings pitched between 1921 and 2004, while Schulz took his sample from players active in 1965 with ten seasons of experience (no minimum innings threshold). The methodology was significantly different as well, Fair utilized a piecewise quadratic model while the approach in Schulz was nonparametric. It was shown in Section 3.5 of this dissertation that the quadratic approach tended to under estimate the standard error. This is likely the case with Fair's results as well.

There are a few possible reasons for the discrepancy. Since Fair's constraints are focused on ten relatively injury free years as a starting pitcher, his sample may be biased towards more productive pitchers. In the modern era it is more likely for players to start and/or end their careers in the bullpen, coupled with the fact that having ten years of injury free service is no small feat (Sandy Koufax had only 9 seasons of 150 innings). ERA also is not the best measure of pitcher ability as shown in the previous section. The noise of random chance and team defense is substantial on ERA.

Fair's age curve analysis was limited to ERA, while Schulz also looked at measures that are more stable (to varying degrees) such as strikeouts, walks, and innings pitched. He found that strikeouts and innings pitched peaked at a younger age (27 to 28) while walks did not peak until 31.

4.2 Background: On Pitchers and Batters

The starting pitching model has many similarities to the batting model. In essence, the outcomes being assessed are virtually identical, it is the point of view that is being changed. As such the same distributions and methodologies will tend to hold true for the pitching model as it does for the batting model. One issue that must be addressed, however is the relative impacts. McCracken [38] asserted that on balls hit into the field of play, the pitcher had little effect on what happened, that was the batter's and fielders domain.

This hypothesis will be assessed later on in this dissertation, where the methodologies will assume that the pitcher's impact on balls hit into the field of play is negligible. This assumes that the batter and the defense are the largest contributors in these areas. Home runs may be more impacted by the batter than anyone else, but the pitcher also plays a significant role. If it is assumed that over the course of a season pitchers face the same quality of batters than they have previously, it is straightforward to assume that the batter's impact is negligible over the course of a season. Adjustments may need to be made when a pitcher changes teams, especially when that change also spans leagues, as the number of runs per game is consistently higher in the American League than the National League. (This discrepancy is

not surprising to anyone familiar with the rules, as the DH rule in the American League permits the team to allow one player to hit and not field, while another (virtually always the pitcher) may field and not hit.)

4.3 Bayesian Modeling and Distributions

The pitching model utilizes the same Bayesian principles and conjugate priors as the batting model, just from a different point of view. The same results of an at bat will be modeled as the batting model, but strikeouts will be considered independently, resulting in seven categories: strikeouts, walks, home runs, singles, doubles, triples, and outs (exclusive of strikeouts). Models exist in the literature that integrate batting and pitching effects on a single plate appearance [44]. Such an approach is relevant for individual games, but less relevant over the course of the season. As with the batting model the discrete outcomes make the multinomial distribution the natural choice. The characteristics of the multinomial distribution can be found in the batting chapter, equations (3.1, 3.2) as shown in Figure 4.1.

As with the batting model, the Dirichlet distribution (3.8, 3.9, 3.10) is conjugate prior to the multinomial distribution. This yields the closed form Bayesian family shown in the batting chapter equation (3.11, 3.12, 3.13). It is also important to ensure that the data is not violating the covariance assumptions of the Dirichlet distribution, or we will need to test the nested Dirichlet to see if we are able to fit the covariance matrix to the data. The most pitcher dependent statistics are strikeouts, walks, and home runs, so we will use those metrics to assess the covariance. For this test starting pitchers will be used since they pitch a larger number of innings and hence have a more substantive sample size year over year.

All pitchers with more than five starts in at least five seasons between 2008 and 2013 were used in the sample. The resulting sample was 420 player-seasons of data and 75 different pitchers. The standard error is assessed by taking the standard deviation of 20 bootstrapped samples. There is a small positive correlation between pitchers when assessing strikeouts and walks, and a negative correlation that is greater than the Dirichlet covariance between home run and strikeout rates. These results suggest that pitchers who strike more

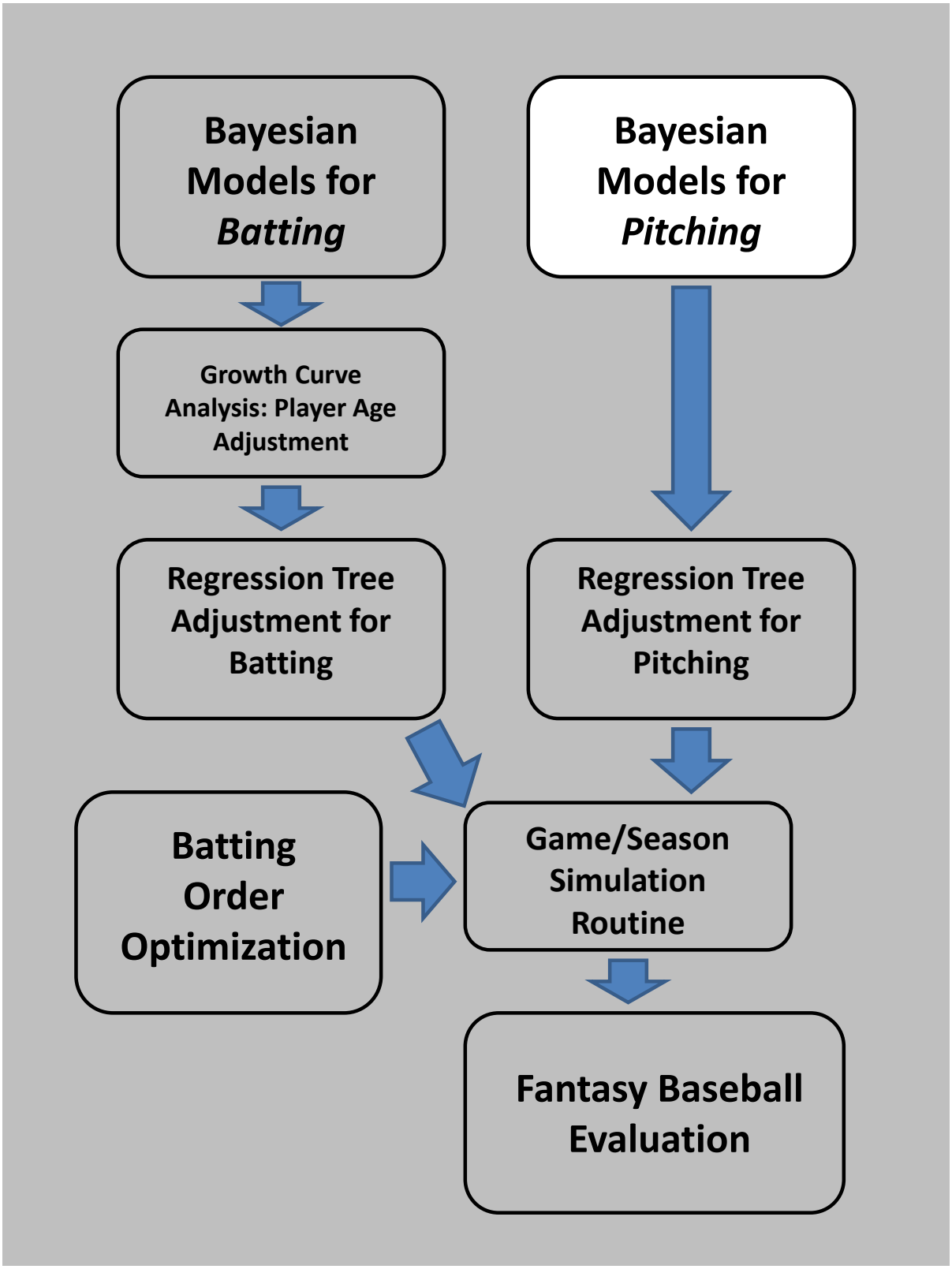

Figure 4.1. Flowchart of dissertation methods highlighting the bayesian pitching methods

batters out also tend to walk more batters and give up fewer home runs. The real test for utilizing the Dirichlet distribution is with the within pitchers assessment, however.

Table 4.1. Between Pitchers Correlation Matrix: Mean

Stat	Walk	Home Run	Strikeout
Walk	0	-.047	.051
Home Run	-.047	0	-.201
Strikeout	.051	-.201	0

Table 4.2. Between Pitchers Correlation Matrix: Standard Error

Stat	Walk	Home Run	Strikeout
Walk	0	.050	.053
Home Run	.050	0	.044
Strikeout	.053	.044	0

When looking at the data set as a whole there is some evidence suggesting that the Dirichlet distribution may not be the best model. However we are more interested in the covariance within players. For our analysis the correlation between higher strikeout rates and lower home runs allowed between pitchers does not impact our models. More importantly, when a given pitcher strikes more batters out, will he also tend to give up home runs or walks in a rate disproportionate to the assumed Dirichlet distribution model. Tables 4.1 and 4.2 present the within pitchers mean and standard error of the correlation matrix between walks, strikeouts, and home runs.

Table 4.3. Within Pitchers Correlation Matrix: Mean

Stat	Walk	Home Run	Strikeout
Walk	0	.084	-.120
Home Run	.084	0	-.129
Strikeout	-.120	-.129	0

Table 4.4. within Pitchers Correlation Matrix: Standard Error

Stat	Walk	Home Run	Strikeout
Walk	0	.060	.063
Home Run	.060	0	.063
Strikeout	.063	.063	0

These results suggest that walks and home runs are more negatively correlated with strikeouts than the Dirichlet covariance, however the covariance is within a 95% confidence interval. The positive correlation between home runs and walks in the sample, similarly is suggestive of a positive correlation but also falls within a 95% confidence interval. As a results the Dirichlet distribution will be utilized for pitchers.

4.3.1 Defense Impacts

The theory provided by McCracken [38] that the defense had a larger impact on balls hit into the field of play (BIP) than the pitcher is one that was tested in our methodology. We began by using the pitcher's prior three seasons of data as observations for balls hit into the field of play. We then tested utilizing random forests to correct the errors, where the data provided in the trees was the pitcher's prior data, age, handedness, etc. This was compared against utilizing the entire team's defensive results on balls hit into the field of play and the results were compared for predictive power.

In order to assess this impact a comparison was made utilizing the data for the model outlined in Section 4.3, the data used was a hybrid of the individual pitcher and the team defense. The pitcher data was used for strikeouts, walks, and home runs, while the team data was utilized for outs (non-strikeouts), singles, doubles, and triples. The pitchers data utilized was production from 2010 thru 2012 while the team data was the 2012 season. The results were compared predicting 2013 performance. The results for starting pitchers shown in Table 4.5 demonstrate that utilizing the team's data for the prior year has greater predictive power than just the pitcher's prior data, and the prior data with non-parametric correction applied. This suggests that the defense does have more influence on balls hit into the field of play than

the pitcher. Making this correction had the biggest impact on ERA, Wins, and Losses. The Table below shows the bias that was corrected when making this change. There is still a positive bias on innings pitched, which results in increased strikeouts and wins, but the improvement in earned run average is substantial with this change.

Table 4.5. Average Modeling Results Comparison

	ERA	WHIP	Ks	Wins	Loss
Actual	3.85	1.30	134	9.9	8.6
Team BIP	3.73	1.29	148	12.7	8.7
Pitcher BIP	2.90	1.29	147	13.2	7.9

4.3.2 Starts and Batters Faced

In addition to determining the productivity of the starting pitcher when they are in the game, it is also important to determine how many games they will start and how far into the game they will pitch in each start. Games started fits well in the beta-binomial conjugate family of distributions. The prior three years of data are used where a maximum number of 36 starts is assumed. The league leader in games started has been at 34 for most of the past decade, with Justin Verlander and CC Sabathia both eclipsing that number (to 35) once. The last player to actually start 36 games was Barry Zito back in 2002, when a number of starters routinely started 35 games each year. 36 is an appropriate number in today's climate, although if one wished to perform this analysis historically, the cap in number of starts would have to be much larger as pitchers would routinely start over 40 games as recently as the 1970's.

The games started for each of the prior three years are the number of successes, and the number of failures is delta between that number and 36. To account for injuries and players who were not starting pitchers for the entire season, the number of games not started is capped at 15% of the games started. This was done so there were not drastic under-estimates that needed to be corrected in the decision tree phase. It is still the case that players with a high number of missed starts were adjusted in the decision tree phase outlined in the next section.

How to address how far into a game a starting pitcher pitches is an interesting question. Marginal pitchers will be pulled from the game when they have given up a number of runs, even if they are not yet fatigued, leading to a correlation between productivity and how much a pitcher will pitch in a game. Better pitchers will pitch further into a game even when they are not pitching particularly well, leading to the number of batters faced by a particular pitcher per game to have a low correlation with productivity. This analysis utilizes batters faced as the metric by which the length of the pitchers stay in the game is determined.

The benefit of using batters faced is that it is not highly correlated with production so one can assume independence without dramatically affecting the results. The downside of using batters faced is that the pitcher will be removed in the middle of an inning most often, underrepresenting the proportion of times a pitcher is relieved after an inning is complete. These factors, while not completely irrelevant, are also not of primary concern. The significant impacts of these factors will be on games won or lost, as well as quality starts. Only one of these metrics (Wins) is a standard fantasy category, and both wins and losses are highly predicated on external factors and hence very hard to predict.

4.3.3 Saves

As identified at the beginning of chapter 4 the statistic that generates the most fantasy attention (and real life attention for that matter) among relievers is saves. Since a save is earned by a reliever who finishes a game when entering with a lead of three or fewer runs, it is a statistic predominantly predicated on opportunity. The manager chooses which players will have the opportunity to finish games, and hence accrue saves. This makes predicting saves a combination of predicting the success of a relief pitcher, as well as the manager's inclinations.

Predicting which relief pitcher will pitch in the last inning of close games, and hence accrue saves, is done utilizing random forests. The data included in the predictions are the player's statistics over the past three years, including games in which the player pitched the eighth or ninth innings of close games, as well as who is expected to be the closer going in to the season. Of course it is not difficult to predict a player to finish games and hence accrue

saves when they have done so for the past several years and have been named the closer by the manager prior to the season. While these players do suffer injuries and lose their jobs, it is much more difficult to predict players who are not named closers at the season outset to accrue saves accurately.

There are at most thirty players who can accrue saves on a regular basis at a time (one per team) which means that if we only use one season for evaluation the sample size is going to be problematic. The past three seasons of data were utilized to evaluate the results of the trees and hence predictability of saves. The model was only able to correctly predict closers, as defined by accruing twenty or more saves over the course of the season, twenty eight out of sixty times generating a success rate of just under fifty percent. While this may seem like failure, we will discuss in Section 4.6 how even these meager results prove useful.

4.4 Age Effects and Mean Reversion

There has been some work on the impacts of age on the productivity of pitchers. However the results have not been consistent, and there is more variability in pitching than there is in hitting overall. Some of that is likely due to the fact that so much of the pitcher's stat line is out of the pitcher's control. Due to the variability, however, a non-parametric approach to age effects and mean reversion was chosen as shown in Figure 4.2.

To assess the changes in strikeout, walk, and home run rates decision trees were utilized. Other methods such as quadratic age curves and random forests were tested, but decision trees proved to have the most predictive power. The section is titled age effects and mean reversion because the decision trees account for those variations. Age and years of experience were generally not important variables in the model, however variables identifying trends in the statistic of interest were influential. Mean reversion is identified because those projected to perform in the extrema are almost unilaterally moved into a more moderate position by the decision trees.

The decision tree results in Figure 4.3 show mixed results from the decision tree implementation. The one variable that noticeably improved as a result of the implementation

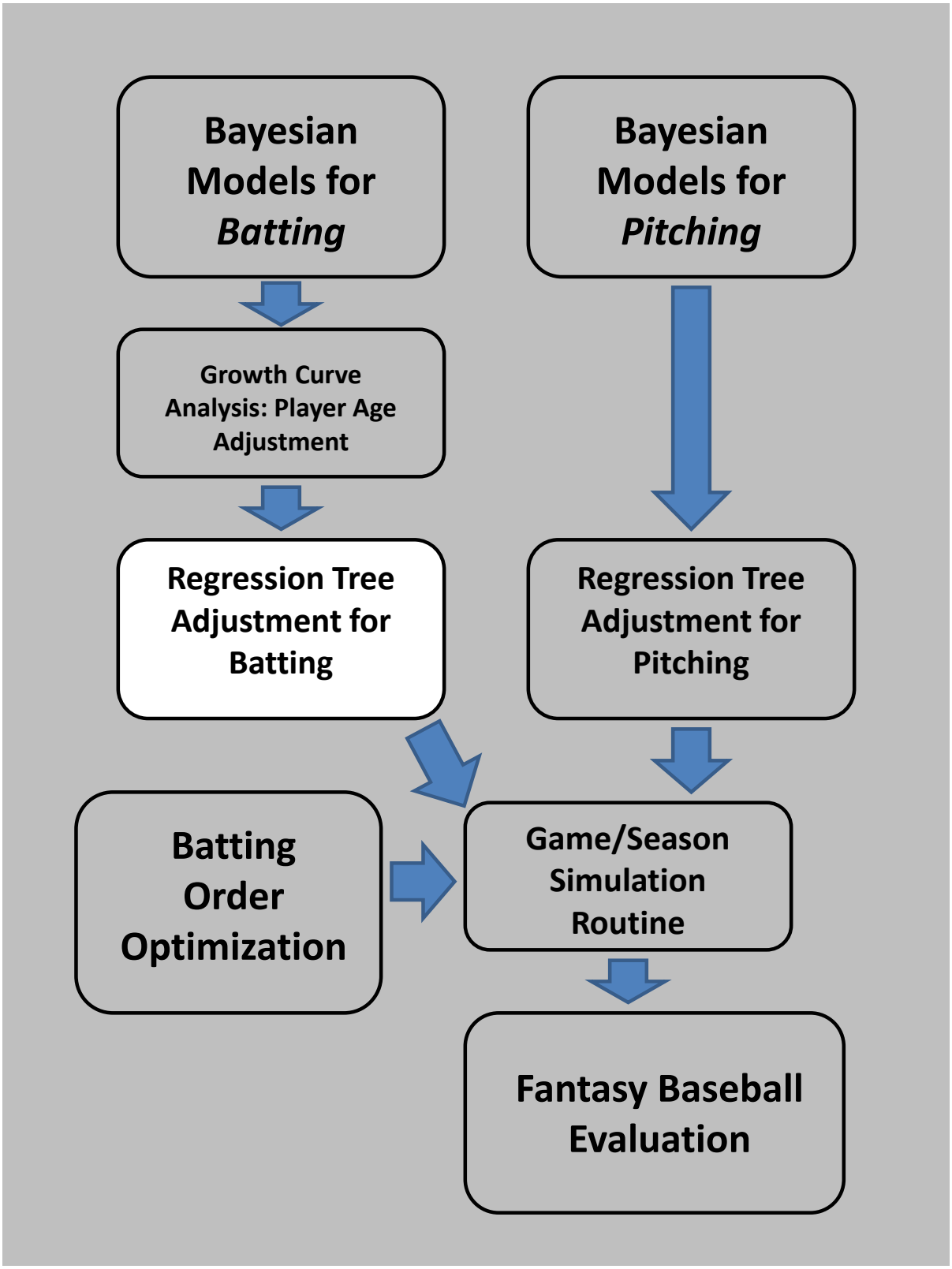

Figure 4.2. Flowchart of dissertation methods highlighting the regression tree pitching methods

was earned run average. Before the decision trees the results were positively biased by .8 earned runs. Figure 4.3 shows a clear positive bias for earned run average as the peak of the distribution is between zero and two, while the tree adjusted errors are centered on zero. The other biases were reduced, at least slightly, for strikeouts, wins, and losses, while there was not a discernable difference in innings, strikeouts, or WHIP. The variance increased slightly for innings and strikeouts, but overall the results were improved.

Batting average against and isolated power were tested though the metrics that proved to be significant were: strikeout rate, walk rate, home run rate, batters faced per game, and games started. Both strikeout rate and walk rate tend to decrease over time, which is consistent with the findings of Schulz [52] on age effects. Pitchers tended to peak in their strikeout rates very early on in their careers, and minimized their walk rates much later on. The impacts of batters faced per game and games started were largely mean reverting. Players with lower projected games started or batters faced were increased (generally younger players), while players with the highest projected games started or batters faced were projected to regress. The regression was minor while the increases particularly in games started were substantial at times.

4.5 ALGORITHM

We will run the algorithm with data prior to 2013 to project 2013 player performances. The performances are run iteratively team by team in conjunction with the batter performance algorithm outlined in the previous chapter as shown in Figure 4.4. The batter performances are integrated with the pitching performance as they impact wins and losses.

1. Begin with a non-informative prior
2. Aggregate the last three years of data
3. Utilize aggregated data to update the prior
4. Utilize the individual pitcher rates for strikeouts, walks, and home runs while imputing the team rates for balls hit into the field of play

Figure 4.3. Histograms of model error rates before and after the decision tree implementation for starting pitchers in the 2013 Major League Baseball season.

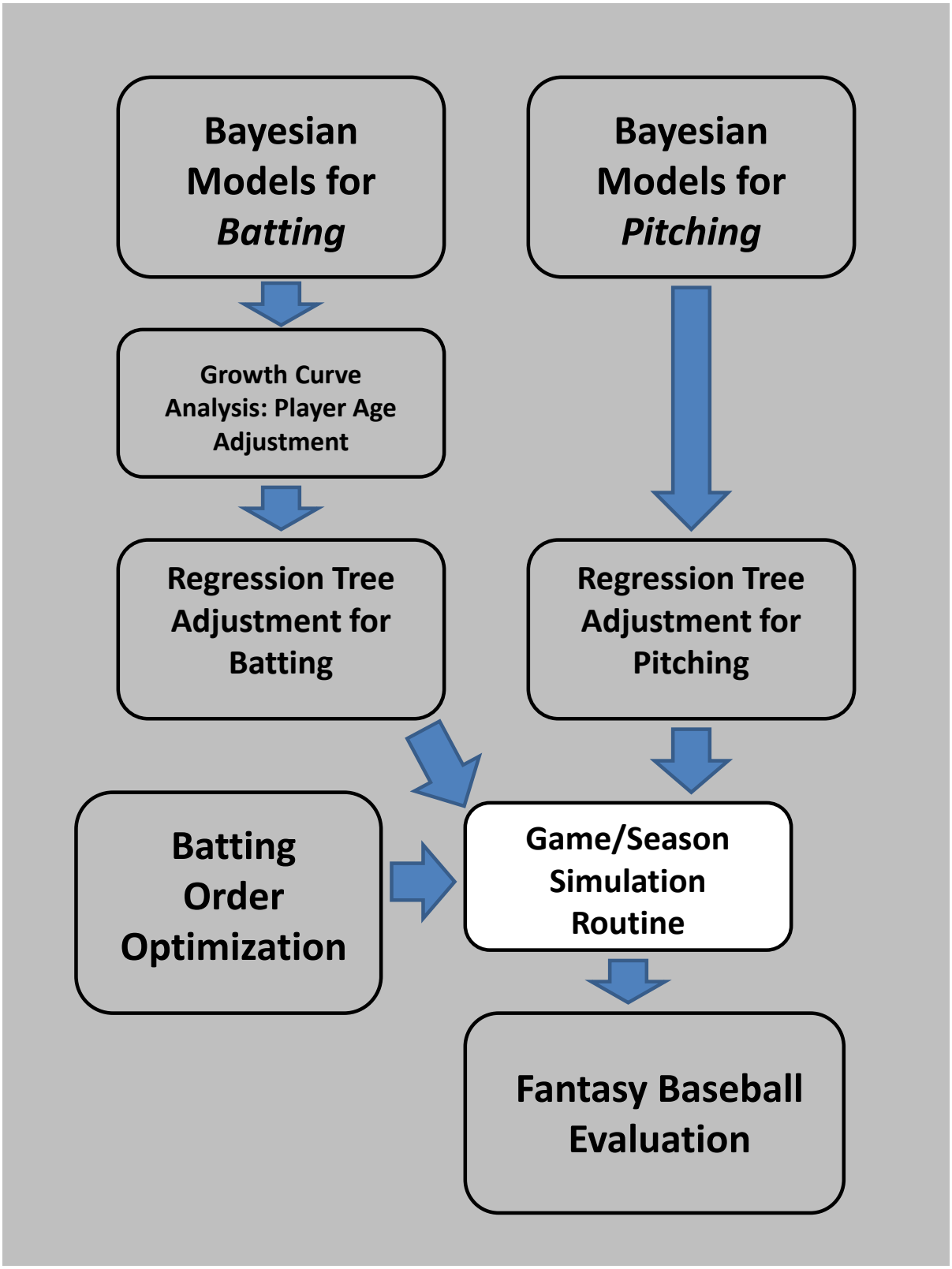

Figure 4.4. Flowchart of dissertation methods highlighting the simulation routine

5. The rates for balls hit into the field of play are scaled to unity for each players strikeout, walk, and home run, rates

6. Run the players data through the breakout and regression trees to forecast likelihood of a player to either outperform or underperform the final posterior distributions

7. Run relief pitchers through closer trees to forecast likelihood of a player to be named the closer

8. Begin Seasonal Simulation

9. Adjust strikeout rate, walk rate, and home run rate in the Dirichlet distribution in accordance with the tree results

10. Adjust games started and batters faced per game in accordance with the tree results

11. Simulate the number of starts for the season

12. Simulate the closer for the game

13. Simulate the number of batters that the starter will face in the given game

14. Simulate the result of the first batter batting

15. If there are runners on base and the batter reaches base, determine the ending position of each runner. Runners only advance if forced with a walk, while league average baserunner advancement is applied to runners on first and second when a single is hit, and runners on first with a double. All other baserunners score.

16. Record event(s) that took place (strikeout, walk, hit, earned run(s))

17. Repeat items 13 to 15 until there are three outs or the number of batters faced has been reached

18. Once three outs are reached clear the on base vector

19. Simulate one batting inning for the same team

20. Repeat items 13 to 18 until either nine innings have been played or the number of batters has reached the total number of batters faced for the pitcher

21. If the pitchers team is ahead in the game and the pitcher has pitched five innings or more, identify the pitcher as in line for a win

22. If the pitchers team is behind in the game, identify the pitcher as in line for a loss

23. If the pitcher has pitched at least six innings and given up fewer than four runs, record a quality start

24. If nine innings have not yet been played repeat items 13 to 18 until nine innings have been played (extra innings could be simulated if desired, we have chosen to end games at nine innings)

25. After each half inning identify if there has been a lead change (or tie), if there has then identify the pitcher as no longer in line for a win or a loss

26. If the starting pitcher is no longer in the game, and it is not a save situation, randomly select a relief pitcher to pitch that inning

27. If it is a save situation, insert the closer to pitch that half inning, and record a save if the team is still winning the game at the end of that half inning

28. At the end of the game record the win or loss, if the pitcher finished the game in line for a win or a loss

29. Repeat game simulations until 162 games have been simulated

30. Begin a new season. Repeat 10 seasons and record the averages

4.6 RESULTS

As with Chapter 3, and as shown in Figure 4.5, the model was run to predict pitcher performances for the 2013 baseball season based on their statistics in 2012 and prior. The methodology for comparison is similar as well. We are again comparing the model results to that of Rotoworld and Athlon Sports, two industry leaders in their own right. These publications projected rankings for starting pitchers, and we must again create a ranking system from the raw statistics that are the model output. The four basic starting pitcher categories that the best in class ranks are based on is wins, strikeouts, ERA, and WHIP. As with the batting ranks we create z-scores to aggregate the statistics. Each of these statistics can be approximated by a normal distribution as shown in Figure 4.6, so we do not run into the issue of assessing the scarcity of a statistic as we did with stolen bases in the batting model. These statistics all fit more closely to the normal distribution than the batting statistics. Wins and innings pitched have a slight negative skew due to injuries, a significant number of pitchers each year have their seasons cut short or dramatically impacted by injuries, much more so than batters.

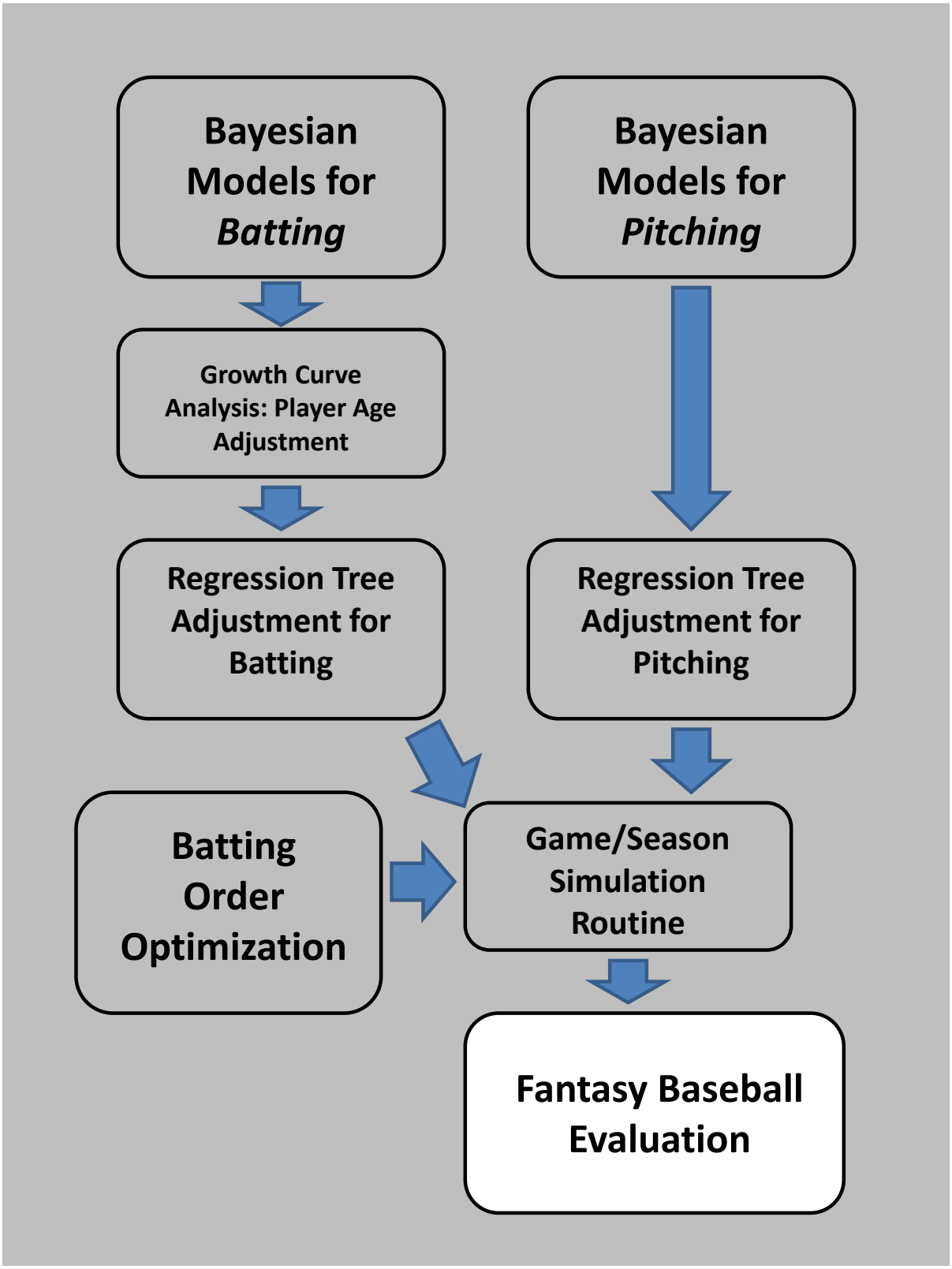

Figure 4.5. Flowchart of dissertation methods highlighting the fantasy evaluation

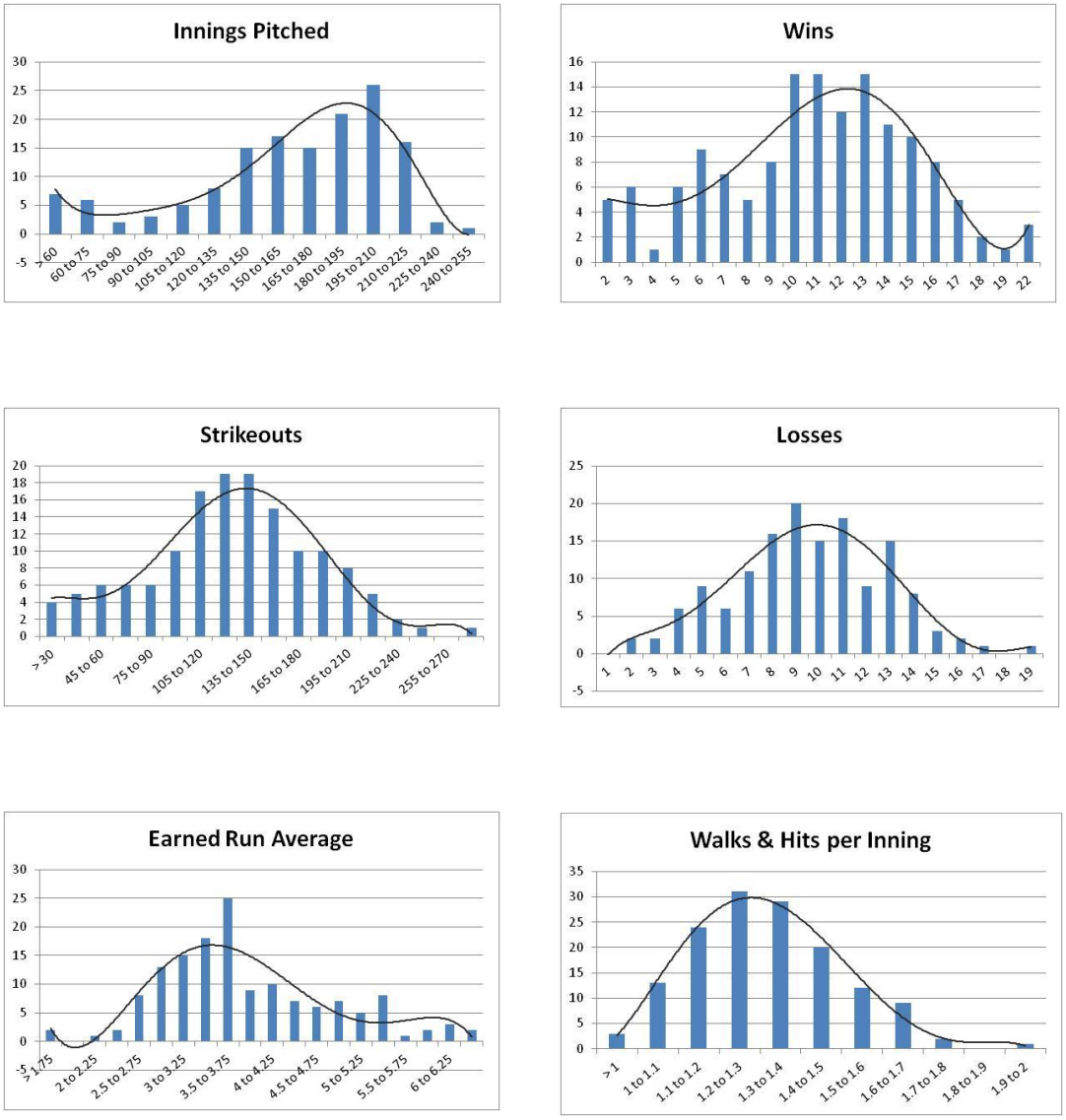

Figure 4.6. Histograms of starting pitcher statistics from the 2013 Major League Baseball season.

4.6.1 Model Accuracy

Since starting pitchers and relievers are assessed based on different statistics, we will break them out separately in this preliminary assessment beginning with the starting pitchers. The model ranks are compared with the actual rank based on the z-scores of the performance statistics, as well as the two best in class rankings of Rotoworld and Athlon sports. Table 4.6 on the following pages provides the player ranks for all starting pitchers who were ranked in the top 25 in any of the four metrics. They are sorted by actual performance so it is easier to compare across rankings quickly. The last two columns are assessments of the model rank. The column titled Relative Rank compares the model rank to the two best in class rankings. The last column titled Value compares the average rank of the three projections to the actual rank. When the second to last column indicates average, the model was in alignment with the best in class rankings. When neither of the last two columns are average, and the last two columns are in alignment, that indicates that the model did a better job on the player than the best in class rankings. When they are not in alignment the model did not do as good a job as the best in class models.

Using the last two columns as a guide, the model did a decent job projecting players compared with the best in class. Of the 43 rankings on Table 4.6 the model result agreed with the best in class rankings on only 13 and disagreed on 30 of them. When evaluating the discrepancies between the model rank and best in class, the model projected the players more accurately than the best in class models for 15 of the 30 discrepancies for a success rate of 50%. When the same analysis is performed comparing the Rotoworld ranking to the model rankings and Athlon Sports they achieve a 25% success rate. Comparing Athlon Sports rankings to the others yields a success rate of 50% accuracy. It is clear that the model rank and Athlon Sports ranks were substantively better than Rotoworld rankings, however there is not a distinguishable difference between the Athlon Sports ranks and the model ranks.

Since Athlon had a similar accuracy rate with the model it may make sense to aggregate the two results in some way. If we look at the results when Athlon and the model

were in alignment compared to the Rotoworld rankings, we find too small a sample size for analysis. Of the 43 results displayed, only seven had consistent results between only Athlon sports and the model result. Of those seven players, the model performed significantly better than Rotoworld on five. If we expand the analysis beyond Table 4.6 we find fifteen players that fit this criteria, of which the model was better than Rotoworld on nine of the players.

Table 4.6. Starting Pitcher Rankings

Player	Team	Model Rank	Actual Rank	Roto World	Athlon Sports	Relative Rank	Value
Clayton Kershaw	LAD	2	1	1	2	Average	Average
Max Scherzer	Det	50	2	25	26	Low	High
Adam Wainwright	StL	8	3	14	22	High	High
Yu Darvish	Tex	13	4	20	14	Average	High
Cliff Lee	Phi	5	5	9	10	High	High
Madison Bumgarner	SF	35	6	17	11	Low	High
Matt Harvey	NyM	34	7	66	30	Average	High
Hisashi Iwakuma	Sea	63	8	67	69	Above Ave	Very High
Jordan Zimmerman	Was	46	9	23	18	Low	High
Anibal Sanchez	Det	41	10	49	35	Average	High
Chris Sale	ChW	30	11	16	12	Low	Above Ave
Felix Hernandez	Sea	4	12	7	7	High	Below Ave
Zack Greinke	LAD	33	13	18	24	Low	High
Mike Minor	Atl	92	14	46	31	Low	High
Patrick Corbin	Ari	109	15	145	NA	High	Low
Mat Latos	Cin	15	16	28	20	High	High
Bartolo Colon	Oak	56	17	116	NA	High	High
Clay Buchholz	Bos	37	18	64	79	High	High

Francisco Liriano	Pit	62	19	69	NA	High	High
Shelby Miller	StL	59	20	118	72	High	High
Stephen Strasburg	Was	19	21	5	3	Low	Low
Justin Masterson	Cle	54	22	98	NA	High	High
Julio Teheran	Atl	130	23	322	NA	High	High
C.J. Wilson	Ana	20	24	26	37	Above Ave	Above Ave
Kris Medlen	Atl	17	25	19	16	Average	Below Ave
James Shields	KC	11	28	12	25	Above Ave	Below Ave
Justin Verlander	Det	1	32	2	1	High	Low
Matt Moore	TB	64	34	22	36	Low	Average
Gio Gonzalez	Was	14	35	15	6	Average	Low
David Price	TB	3	36	4	4	Average	Low
Cole Hamels	Phi	12	39	3	8	Low	Low
Hiroki Kuroda	NYY	25	40	30	44	High	Average
R.A. Dickey	Tor	21	46	11	19	Low	Low
Doug Fister	Det	24	49	57	40	High	Average
Jered Weaver	Ana	6	53	6	9	Average	Low
Matt Cain	SF	7	60	8	5	Average	Low
C.C. Sabathia	NYY	10	64	13	15	Low	Low
Andy Pettite	NYY	23	74	54	NA	High	Low
Johnny Cueto	Cin	26	78	21	13	Low	Low
Dan Haren	Was	22	79	34	49	High	Low
Yovani Gallardo	Mil	16	80	33	21	High	Low
Roy Halladay	Phi	9	129	10	23	Average	Very Low
Josh Johnson	Tor	18	134	44	41	High	Low

Table 4.7 depicts an analogous assessment of the top 20 relief pitchers. A slightly smaller sample of relief pitchers was used since there are fewer relevant relief pitchers for fantasy. A typical fantasy team has five to eight starting pitchers and two to four relievers. Assessing the relief pitchers similarly to what was done in the preceding paragraphs for starting pitchers, there were a total of 31 results on Table 4.7. Of those 31 results the model agreed with the best in class results on five of those and disagreed on 26 of them. Of the 26 that the model disagreed with the best in class rankings the model was better on thirteen of them, the rankings were better on eleven, and two of them were between the model and rankings and too close to determine the better result.

As in Section **??**, we will asses the average miss of the three rankings as well for both the starting pitchers and the relief pitchers. There was less differentiation between the ranks for starting pitchers, the average misses were 26.7 for our model, 28.3 for Roto World, and 25.6 for Athlon. This is consistent with the mixed results observed above, as the model finished between the two best in class rankings. The model faired slightly better when assessing the top 20 relief pitchers. The model missed by an average of 9.2, Roto World missed by 12.2, and Athlon by 16.8.

Another way to evaluate the rankings is to look at which of the three rankings is the closest. Taking that approach, our model was best with twelve, Athlon was the best in ten cases, and Rotoworld was the best in four cases. There were five instances where multiple models agreed. Similarly to the starting pitchers results discussed in the previous paragraphs, Athlon and our model performed significantly better than the Rotoworld rankings. Even in a situation where the model results here are not clearly better than all of the best in class rankings, since the best in class rankings are more similar to each other there are opportunities to maximize the efficiency of the rankings. We discuss such a strategy in Chapter 5.

4.6.2 Breakout vs Outlier

We will assess breakout pitchers in an analagous manner to how we evaluated batters in Section 3.8.2, beginning with starting pitchers and then analyzing relievers. Of the top 20

Player	Team	Model Rank	Actual Rank	Roto World	Athlon Sports	Relative Rank	Value
Craig Kimbrel	ATL	2	1	1	1	Average	Average
Greg Holland	KCA	17	2	17	15	Average	High
Joe Nathan	TEX	6	3	7	10	Above Ave	Above Ave
Aroldis Chapman	CIN	12	4	16	5	Average	High
Mariano Rivera	NYA	7	5	4	5	Average	Average
Kenley Jansen	LAN	14	6	32	38	High	Very High
Addison Reed	CHA	4	7	11	9	High	High
Jim Johnson	BAL	16	8	12	7	Low	Average
Koji Uehara	BOS	53	9	59	99	Above Ave	Very High
Sergio Romo	SFN	1	10	15	11	High	Average
Glen Perkins	MIN	25	11	8	22	Low	Average
Steve Cishek	MIA	22	12	24	27	Above Ave	High
Rafael Soriano	WAS	19	13	9	18	Average	Average
Fernando Rodney	TBA	11	14	2	6	Low	Low
Ernesto Frieri	ANA	8	15	45	99	High	High
Casey Janssen	TOR	47	16	28	21	Low	High
Grant Balfour	OAK	50	17	22	19	Low	Above Ave
Joaquin Benoit	DET	52	18	34	40	Low	High
Jim Henderson	MIL	23	19	99	99	High	High
Jason Grilli	PIT	13	20	35	25	High	Above Ave
Huston Street	SDN	21	22	19	12	Below Ave	Below Ave
Jonathan Papelbon	PHI	10	23	5	2	Low	Low
Chris Perez	CLE	24	29	21	13	Below Ave	Below Ave
Tom Wilhelmsen	SEA	5	33	14	16	Low	Low
Rafael Betancourt	COL	20	52	26	26	Above Ave	Low
Trevor Rosenthal	SLN	15	68	99	99	High	Above Ave
J.J. Putz	ARI	3	82	20	17	High	Low
John Axford	MIL	9	109	10	14	Average	Low
Drew Storen	WAS	37	112	6	4	Low	Low
Frank Francisco	NYN	18	144	28	99	High	Low
Joel Hanrahan	BOS	43	169	13	8	Low	Low

Figure 4.7. Relief Pitcher Rankings

starting pitchers in the 2013 season, nine of them were ranked by the model outside the top 40. These could easily be seen as significant model misses. However if the players were ranked consistently in our ranking compared to the best in class models, we would still potentially obtain those players. Furthermore, if our ranking was significantly higher than the best in class we would be likely to obtain them. Our model was significantly higher than the best in class for Bartolo Colon, Francisco Liriano, Patrick Corbin, and Shelby Miller while it was significantly lower than the best in class models for Max Scherzer, Jordan Zimmerman, and Mike Minor and in line with best in class for Anibal Sanchez and Hisashi Iwakuma. While these are significant misses, we still have a reasonable expectation to obtain some of these players in fantasy drafts.

Now to examine the consistency of their performances. If these players experienced a breakout performance, then we should see them maintain their improved performance to some degree in the following season (2014). Only Max Scherzer (17) and Hisashi Iwakuma (19) have remained in the top 40 rankings for starting pitchers. Jordan Zimmerman (48), Bartolo Colon (50), and Anibal Sanchez (51) remained around the top 50, in line with the model projections for 2013, while the others fell off more precipitously. Both Scherzer and Iwakuma appear to be poised to maintain their high levels of performance for years to come. While Scherzer had a number of seasons of prior data to evaluate, Iwakuma was more of an unknown having transferred to MLB from Japan the year prior and only cracked the starting rotation for the Mariners for the second half of that season. That is to say that Max Scherzer is the only breakout performer that one could have reasonable expectations that the model missed, which we could have a reasonable hope for it to identify.

While this is a worthwhile exercise the sample size is very small. To test for consistency we did the analysis in the preceding paragraphs for 2012 and 2014 as well. When expanding the view to the past three seasons there were nineteen total players that finished within the top twenty and were modeled outside the top forty. Of those nineteen we would still have at least a reasonable expectation to draft ten of them while one likely went undrafted

in most leagues (ranked outside the top 100 starting pitchers by the model and best in class rankings). That leaves only eight out of sixty that the model missed on and we would not have the opportunity to draft for a success rate of 83%. This is a good success rate, however there were seven true breakouts that the model missed, so there is room for improvement as well.

Of the top 15 relief pitchers in 2013 only Koji Uehara (53) and Glen Perkins (25) were ranked outside the top 25 by our model. Koji Uehara was actually ranked higher by our model than either of the best in class rankings, however since he was ranked so low and typically 30 to 50 relievers are drafted in leagues, he likely went undrafted in most fantasy leagues. Glen Perkins was ranked lower by our model than the best in class rankings. Both of these players remained fantasy relevant in 2014 with rankings inside the top 20 for relief pitchers (Uehara 16, Perkins 18).

Now expanding the results to include the past three seasons of results (2012-2014) we have fifteen relievers who finished in the top fifteen but ranked outside the top 25 by our rankings. More than half of these relievers went largely undrafted while twelve of the fifteen have held their value as top twenty relief pitchers. Two of the fifteen the model would recommend drafting, while the best in class models would have selected five of the fifteen, with three of them holding their value.

CHAPTER 5
FANTASY TEAM OPTIMIZATION

Projecting the performance for the upcoming season is half the battle of the fantasy baseball draft. The other half of the battle is optimally selecting a complete team based on these projections.

There are two different methods for selecting a fantasy team. The first is called a snake draft where players are selected in turn, much like the drafts done by MLB, the NBA, and the NFL. The major difference is that the order of selection is inverted in each round, so that the team (or fantasy baseball manager) who selects first in the first round will select last in the second round, first in the third round, etc. This is done in an attempt to reduce the benefit of selecting first.

The second method which has gained more and more popularity over the past few years is called the auction draft. In the auction draft, each team starts with a set amount of pretend money, with which to bid on players to create a team. Players are nominated for auction and then bid on, so that the manager who is willing to spend the most on each player will have that player on their team. Each of these methods has their benefits and drawbacks.

Of course drafting is not as simple as identifying who the best player will be for the upcoming season and then selecting that player. While the roster construction of each fantasy league varies, there are commonalities among them. We will examine the most prevalent league rules, which are by no means the only rules out there. Most commonly a team must consist of one player at each of the defensive positions on the baseball diamond: catcher, first base, second base, third base, shortstop, three outfielders, and one or two utility players who may play any position. The number of pitchers vary from league to league but in general at least two to three starting pitchers must be chosen, along with at least two relievers, and a total of six to eight pitchers (often some may be either starters or relievers). The number of teams

in each league varies as well, which adds to the complexity. Since there is so much diversity among the different leagues, we will discuss the methods by which one should evaluate the players, along with an example.

The additional challenge in developing a drafting strategy is that it is impossible to know how each of the other team managers value each player, and hence the required cost (either in terms of draft dollars or draft position) of each player. There are a number of different ways to evaluate players as compared to their peers. Ranking players as shown in Tables 3.5 and 4.6 is an industry standard, and tiering players at each position has become commonplace as well. Tiering players consists of grouping players that are identified as having comparable value to your fantasy team, and who play the same position. We will not tier players here, since we have a quantifiable measure of value. If one desires to tier their players, that would become an exercise in clustering, but provide little value in our setting since we already have a quantifiable measure of comparison.

We will attempt to optimize our fantasy team by comparing the benefit of each player over the replacement level options. Replacement can be viewed in a number of ways as well, one reasonable approach is to look at the production of a replacement player. This idea has become more common, evaluating the overall performance of a baseball player with Sabermetrics and the introduction of wins above replacement (WAR). In a twelve team league that requires one second baseman to be played, we can look at the production of the thirteenth ranked player eligible to play second base. However, in most leagues, teams not only field a starting lineup, but also have utility players and bench spots. That means that the thirteenth ranked second baseman is not likely to be available at the end of the draft, and it is difficult to ascertain exactly how many second baseman will be taken. We thus consider the replacement as an average of a group of players.

We will examine a simple example to help illustrate this concept. This example will review the drafting strategy for a five team league that drafts the following positions: catcher, first base, second base, two outfielders, one utility, and two bench spots. Keep in mind that

this is an unrealistic example in its simplicity as virtually all leagues draft at least one player at every position and almost all leagues have at least ten teams, but the concepts will be a bit easier to outline with this smaller example. We will start by ranking the top players that project to be drafted to fill out the starting lineup: five each of the catchers, first basemen, and second baseman, and ten outfielders as shown in Figure 5.1. Of the players not identified as starting in one of the aforementioned positions, we will identify the five best players and project them to be drafted as starting utility players. There will be ten bench players drafted, two for each team, and we will assign them proportionally to each position, rounding up. In this example, since there are two starting outfielders on each team, twice as many bench spots were allocated to that position in Figure 5.1. We will utilize the bench positions, along with the next best player available (the best undrafted player) as the group average for the positional baseline values. Since there are two staring outfielders, we will utilize the next two outfielders (the italicized values in Figure 5.1). The above method addresses what is known as position scarcity, or the fact that it is harder to find a quality fantasy baseball player at, say, catcher than it is to find one playing first base or in the outfield.

5.1 Literature Review

Fantasy drafts have been relatively unexplored in the research literature or other publicly available outlets. We are not aware of any previous work that utilizes similar methodologies to this dissertation to model fantasy sports drafts. We will provide a brief survey of relevant previous research.

It is likely that a contributing factor to the dearth of published work regarding fantasy baseball drafting is due to the value of the business. There are a number of patents out on draft logic, Plimi [48] and Wilcock [59] are two examples. They do not attempt to optimize draft strategy, however, so much as create an interface for drafting and utilizing rankings. Plimi [48] appears to be able to update the player rankings in his GUI. There are a few quantitative articles on player drafting. One concept that has been gaining traction in recent years is the concept of value based drafting (VBD) [23], [28], and one that we will utilize when evaluating

Name	Position	Position Rank	Index	Strategy
Prince Fielder	1B	1	5.7	1B - 1
Paul Goldschmidt	1B	2	5.6	1B - 2
Joey Votto	1B	3	5.4	1B - 3
Albert Pujols	1B	4	5.4	1B - 4
David Ortiz	1B	5	3.4	1B - 5
Eric Hosmer	1B	6	3.1	U - 3
Ike Davis	1B	7	2.8	U - 5
Adrian Gonzalez	1B	8	*2.7*	Bench
Mark Teixeira	1B	9	*2.2*	Bench
Chris Davis	1B	10	*2.2*	
Billy Butler	1B	11	1.9	
Justin Morneau	1B	12	1.3	
Brandon Belt	1B	13	1.3	
Baseline value			**2.4**	

Name	Position	Position Rank	Index	Strategy
Mike Trout	OF	1	9.5	OF - 1
Jose Bautista	OF	2	8.7	OF - 2
Andrew McCutchen	OF	3	7.3	OF - 3
Carlos Gonzalez	OF	4	7.0	OF - 4
Mike Stanton	OF	5	6.3	OF - 5
Ryan Braun	OF	6	6.1	OF - 6
Jay Bruce	OF	7	5.5	OF - 7
Yasiel Puig	OF	8	5.3	OF - 8
Adam Jones	OF	9	4.7	OF - 9
Hunter Pence	OF	10	4.7	OF - 10
Shin-Soo Choo	OF	11	4.5	U - 1
Carlos Gomez	OF	12	4.3	U - 2
Nelson Cruz	OF	13	2.9	U - 4
Jacoby Ellsbury	OF	14	*2.7*	Bench
Khris Davis	OF	15	*2.1*	Bench
Justin Upton	OF	16	*1.7*	Bench
Jason Heyward	OF	17	*1.6*	Bench
Bryce Harper	OF	18	*1.5*	
Will Myers	OF	19	*1.4*	
Baseline value			**1.8**	

Name	Position	Position Rank	Index	Strategy
Carlos Santana	C	1	3.0	C - 1
Buster Posey	C	2	2.4	C - 2
Jonathan Lucroy	C	3	0.1	C - 3
Miguel Montero	C	4	-0.8	C - 4
Russell Martin	C	5	-0.9	C - 5
Evan Gattis	C	6	*-0.9*	Bench
Joe Mauer	C	7	*-1.8*	Bench
Carlos Ruiz	C	8	*-1.8*	
Josmil Pinto	C	9	-2.1	
Baseline value			**-1.5**	

Name	Position	Position Rank	Index	Strategy
Jason Kipnis	2B	1	4.7	2B - 1
Robinson Cano	2B	2	4.6	2B - 2
Dustin Pedroia	2B	3	4.0	2B - 3
Dan Murphy	2B	4	2.7	2B - 4
Brett Lawrie	2B	5	2.3	2B - 5
Ian Kinsler	2B	6	*2.0*	Bench
Jose Altuve	2B	7	*1.8*	Bench
Dee Gordon	2B	8	*0.9*	
Howie Kendrick	2B	9	0.7	
Baseline value			**1.6**	

Figure 5.1. Example for ranking batters for a small five team draft with teams consisting of one catcher, one first baseman, one second baseman, two outfielders, and one utility player.

draft strategies. VBD, as the name suggests, values a player not just based on his performance, but also on his performance compared to the other options available at his position. While [23] and [28] discuss this concept as it pertains to a fantasy football draft, this concept is even more important when evaluating a baseball draft, as there are more positions to be considered and such a wide range of offensive performance at each of the positions.

There have been a few papers published on optimizing and assessing professional sports drafts. Carluccio [17] evaluated baseball players based on their experience and whether they were batters or pitchers. Fry [27] went into more detail evaluating the NFL draft. While optimizing production in the NFL draft is very different from a fantasy baseball draft, there are some distinct similarities. Fry outlined two basic drafting strategies in the NFL: drafting for needs and drafting the best available player. These basic strategies are true of a fantasy draft as well: in making the next pick, how much will that player pick factor in amongst the manager's previously drafted players, and for that matter what is the optimal drafting approach in that round? Fry [27] made three significant assumptions in his analysis: the other teams' valuations are known, the other teams' needs are known, and the other teams' selection strategies are known.

5.2 Player Variability

The next issue to assess is player variability. Our knowledge base is different for each player, and we are able to assess the individual likelihood of a given player to outperform or underperform his average ranking. Likewise when drafting players, there are times where it makes sense to take risks, and other times when it makes sense to play it safe. The more expensive a player is, the more important it is to be confident in his ability to produce. Another way of looking at it, is that the further a player is in production from the replacement level value mentioned in the prior section, the more difficult it will be to replace his production, and hence the more important it is to be confident in his value. Bottomline is that we would consider spending more on a player if he is capable of producing more, as compared with the replacement level, than his counterparts.

The first challenge that this idea raises is how to compare batters with pitchers. Looking at Tables 3.5 and 4.6, it is not difficult to see that pitchers are significantly more variable than batters. In the outlier analysis, 25% of the top performing batters were significant model misses as compared with 33% of starting and relief pitchers. The difference becomes more pronounced when looking at players that the model would not recommend drafting: 12% of batters, 18% of starting pitchers, and 29% of relief pitchers were significant model misses, with 18% of top performing relief pitchers going largely undrafted. When this is coupled with the fact that more batters are necessary in a fantasy lineup than pitchers, the difference becomes even more pronounced. That means that the top 20 batters are still generally in the top 20% of fantasy relevant producers while the top 20 starting pitchers make up the top 40%, and the top 15 relievers also make up the top 40% of fantasy lineups on average.

It is also important to look at the top model predictions of both batters and pitchers. The results are consistent when we look at the top 20 predictions of batters and pitchers. Let us examine these top predictions and see how many were positive outcomes. If we define a positive outcome as producing in the top 50% of all starters, the results are clear. Batters in the top 20 have a 80% success rate, while starting pitchers have only a 50% success rate, and the top ten relievers have a 60% success rate.

5.3 Drafting Algorithm

In order to optimize a draft strategy it is important to look at more than just the raw projections or rankings. Other items such as position eligibility for batters, estimate variability, competitive intelligence, and the projected performance and positions of the players already drafted need to be incorporated in the analysis.

Position Eligibility: A typical fantasy team construction consists of the following: catcher, first baseman, second baseman, shortstop, third baseman, three outfielders, two utility players, two starting pitchers, two relief pitchers, four additional pitchers, and five bench players. Using the methodology at the beginning of Chapter 5, we will need to evaluate the

average score of the 13th to 17th ranked players at each infield position, which is (-1.9, 0.3, -0.3, -0.8, -0.3) respectively. Looking at the 37th to 49th ranked outfielders (since there are three starters) yields an average score of 0. Subtracting from each of those scores the highest value (0.3) and taking the opposite yields (2.2, 0, 0.6, 1.1, 0.6, 0.3) for catcher, first baseman, second baseman, shortstop, third baseman, and outfielder respectively. Add the representative value to each player's score, first baseman remains the same, catchers are increased by 2.2, etc.

Some players also qualify for multiple positions, which is valuable especially in leagues where there is no utility position. There are three main benefits to having players that are eligible at multiple positions. First, it allows for drafting flexibility in order to achieve the optimal overall starting lineup. This is less important when there are more utility positions, but with no utility positions it is easy to see that this idea becomes crucial. If we assume the same starting lineup identified in the prior paragraph, but with no utility positions, then as soon as we draft a first baseman we have no more room for another first baseman in our starting lineup. This is true for every starting position with the exception of the outfield. This idea of narrowing the scope of draftable players for our starting lineup is important. When our selection pool is restricted, our ability to optimize overall value is also restricted, more so than when the players we are restricting ourselves from selecting are of high value. To simplify and quantify this result we will look at the next two rounds of available players. Choosing two rounds is somewhat arbitrary but the higher value players are obviously selected first in the draft and are more scarce, so the further we progress into the draft, the less of an impact the reduced player pool will have. We also want to be evaluating players that we will have the opportunity to select in the next round, so we need to select more than one round of players. Since we are looking for the highest valued players, we do not need to consider the position scarcity adjustment in this evaluation. Equation 5.1 outlines the suggested adjustment where "*players*" represents the number of players we would target for the next two rounds (twenty four selections in this example) that are eligible only at the position in question. The quantity

"*rounds*" represents the two rounds of players we are reviewing. We subtract one assuming that one player per round at the target position is expected. We divide by two because there is an opportunity cost to passing up the best player now for what might be a better opportunity in a subsequent round.

$$Player\ adjustment \quad = \quad \frac{\frac{players}{rounds} - 1}{2}. \tag{5.1}$$

The second benefit to having players eligible for multiple positions is due to teams having days off throughout the course of the season. If a player is available at multiple positions, then it allows the roster flexibility to include a bench player more often on those days off. There are 178 days during the 2015 baseball season, leaving sixteen days off for each team. If we assume independence of days off, that means that a player's replacement will have 10.88 games scheduled during those sixteen days off. If a player is able to play an additional position we will gain an additional 7.4 games played over the course of a season. The average fantasy replacement player will contribute 2.85 runs, 0.7 home runs, 2.7 runs batted in, and 0.4 steals over those seven incremental games (based on 66 runs scored, 16 home runs, 63 runs batted in, and 9 steals over the course of the season). When those statistics are quantified in our scoring rubric it generates an incremental index of 0.43.

The last benefit of having players eligible at multiple positions is to have the lineup flexibility when a player is not playing, either due to injury or a player just receiving a day of rest. This is a bit more challenging to calculate as it depends on the other players on our team that also play the position, and how much time the player will miss. For long term injuries the player can be placed on the disabled list, allowing us to pick up a replacement player without having to drop any other players from the team. Based on the analysis in the previous paragraph we can see that the incremental index value for a replacement batter over seven games is 0.43, which is an incremental value of 0.06 per game.

Estimate Variability: In the beginning, or first four to five rounds, of a draft, minimizing risk is most important. There are a number of valuable players to choose from,

and it is important to be sure that the player chosen will be an asset to the team. Later on in the draft variability can be a good thing, looking for players who could potentially have a breakout performance and play significantly better than projected. Within batters and pitchers, variability is assessed. However the largest difference in player variability is between batters and pitchers. Top ranked pitchers have a much higher likelihood to substantially under-perform their ranking than batters. In addition there are more low ranked pitchers that substantially outperform their rankings than with batters. For this reason it is better to wait until after at least the first four to five rounds of the draft before drafting a pitcher. This idea is only magnified with relief pitchers. Section 4.6.2 highlights that most seasons there were multiple relievers who were likely not drafted in most fantasy leagues and finished in the top fifteen in relief pitcher rankings. We will also see in Section 5.4 that the model ranks identify quality relievers that can be drafted in the middle to later rounds.

Competitive Intelligence: It is also important not to spend more on a player than is necessary. It is not wise to draft a player in the first round when he would likely still be available in the fifteenth round. Chris Davis is a great example of this. The highest he was ranked prior to the 2013 season was 96th among batters, and yet he finished as the third most valuable batter in fantasy. Despite the fact that he finished third among all batters, even armed with that knowledge it would be foolish to draft him in the second round of a fantasy draft. Why draft him in the second round when it is virtually guaranteed that he will still be available in the 7th round? If the model ranks the player more than ten spots higher than the best in class models, then adjust the player down in the rankings to halfway between the model rank and the highest competitive ranking. By doing this Hunter Pence would be moved down from number 13 to number 61 in the rankings, freeing us up to draft Jay Bruce (ranked 15) or Paul Goldschmidt (ranked 17) before Hunter Pence and obtaining both players on our roster.

Projected Performance: Maintaining a balanced roster by category is important as most leagues rank each category independently. In addition to the rankings it is important to track the individual statistics of each player. This is easiest to do by the z-score adjustments to

each statistic described in the batting chapter. Keep a running sum of the z-scores by category throughout the draft and it may be necessary to take a player that is ranked slightly below the best available in order to keep the categories in balance. After the first five rounds, a tweak can be made to adjust the scoring by category and helping to ensure a balanced team. In particular, we may use the following formula:

$$NewScore1 \quad = \quad \left[\frac{HighCatScore \; - \; CatScore1}{4}\right] Score1 + Score1 \qquad (5.2)$$

where HighCatScore is the highest current average z-score for any of the categories on our team as it is currently constructed, CatScore1 is the current average z-score of our team for the category we are adjusting, Score1 is the z-score for the category we are adjusting of the player we are adjusting, and NewScore1 is the new value (adjusted z-score) for the player in that category. The reason that we divide by four is to mitigate the effect of the category adjustment so we are still focusing primarily on the best available player, and only giving a slight edge to players who excel in categories where our team is currently deficient. The further along we are in the draft the more likely the magnitude of these adjustments are to increase, and the more we will focus on the areas of need rather than the best available. This is a common practice in fantasy drafts, especially for stolen bases, where a manager may find that he is deficient in a category midway through the draft and then target a couple of players late in the draft that excel in the deficient category. This is a more continuous approach where we can optimize our entire team on a balanced roster and only need to target a specific type of player late in the draft on rare occasions.

Positions of Drafted Players: Once a position has been filled by drafting a player to the fantasy team, remove the increase associated with the position outlined in position eligibility. If one player is drafted that is eligible to play second base and shortstop, do not assign him to either position until the other position has been filled. This way both positions are still identified as scarce until they are both filled, and then the increase due to the scarcity

of the position is removed for both positions simultaneously. There is little advantage to stockpiling players at one position, it is a detriment to the fantasy team until trades can be made. Once there is no longer room in the starting lineup for a player, then an assessment must be made as to how often a candidate player is likely to be in the starting lineup. First an assessment must be made as to how many players he will be able to substitute for (and if there are any other players already on the bench as options for the player). Once we determine who the player is eligible to substitute for we can assume independence between the missed games of the players (which is a reasonable assumption when the players are not on the same team; this analysis will overestimate the usefulness players on the same team in many instances) and it is fairly straight forward to evaluate how many games the player could be in the starting lineup. (E.g. player 1 plays 80% of his games and player 2 plays 70% of his teams games. The overlapping missed games is 20% times 30%, or 6% meaning that the substitute would be in the starting lineup for 54% of the season.) We may then scale that player's statistics in accordance with how many games he will be in the lineup and re-calculate his index value. The new value reflects his value to your current team as it is constructed. However this assumes that the current estimated model value is 100% accurate. There is always the chance that some of the players will not meet the estimated value either due to injury, regression, or an outlying performance. For this reason it is best to consider both the adjusted and the unadjusted values, as well as opportunities to draft highly ranked players later in the draft due to a substantial variation between our ranking and the best in class ranks.

5.4 Draft Analysis

In this section we will bring everything together and look at how these methods brought together can work in practice. Table 5.4 provides the final rankings, taking into account position scarcity and relative ranking. Integration between batters and pitchers is straightforward, since the performance was translated into a consistent, unitless measure across batters and pitchers. Using Table 5.4, we will evaluate a fantasy team draft utilizing this methodology broken down into four sections: early rounds (1-4) 5.4.1, middle rounds

(5-9) 5.4.2, late rounds (10-14) 5.4.3, and final picks (15-21) 5.4.4. For this analysis we will assume a twelve team league with each team consisting of: catcher, first base, second base, third base, shortstop, three outfielders, utility, three starting pitchers, two relief pitchers, two generic pitchers (either starters or relievers), and five bench spots.

It is common practice for draft order to be randomly generated in fantasy leagues; for this draft we were drawn fifth. Draft position highly impacts team composition in the first few rounds, but is less relevant later on. At this draft position our pick selections will be as follows:

- Early Rounds: 5, 20, 29, 44
- Middle Rounds: 53, 68, 77, 92, 101
- Late Rounds: 116, 125, 140, 149, 164
- Final Rounds: 173, 188, 197, 212, 221, 236, 245.

We will assume that we draft the player that is highest on our ranks and not ranked higher than our draft position on either of the best in class ranks, although we will call attention to other possible draft candidates.

Before we draft, it is important to look at the players that the model predicts will perform significantly better than the best in class models as potential draft targets. Doing this will assist us in the draft when we are deciding between similar players to select. For example, if we are forced with a decision between two similar players, one at shortstop and another at second base, and we see that there is a sleeper (a player who we expect to significantly outperform their draft position) at second base who we could select later in the draft, we should select the shortstop.

Table 5.1. Ranking of Results

Player	Position	Team	Model *z-score*	Model Rank	Actual Rank	RotoWorld Rank	Athlon Rank

Dustin Pedroia	2B	BOS	9.8	4	15	45	44
Jason Kipnis	2B	CLE	9.3	7	5	52	46
Adam Jones	OF	BAL	7.4	15	10	53	79
Hunter Pence	OF	SFN	7.2	16	12	110	116
Paul Goldschmidt	1B	ARI	7.1	17	4	55	81
Jay Bruce	OF	CIN	6.7	18	30	64	73
Carlos Santana	C	CLE	6.4	20	75	49	76
Brett Gardner	OF	NYA	6.2	21	80	162	223
Jimmy Rollins	SS	PHI	6.2	22	178	123	113
Brandon Phillips	2B	CIN	5.7	26	28	62	71
Jose Altuve	2B	HOU	5.3	30	27	79	105
Austin Jackson	OF	DET	5.2	31	110	78	56
Sergio Romo	RP	SFN	5.0	33	176	139	123
Ian Desmond	SS	WAS	4.8	35	32	72	78
Ben Zobrist	2B, SS, OF	TBA	4.6	38	89	114	69
Mark Trumbo	OF	ANA	4.4	40	48	194	184
Carlos Gomez	OF	MIL	4.3	41	11	232	127
Martin Prado	2B, 3B	ARI	4.3	42	57	91	111
Michael Bourn	OF	CLE	4.3	45	111	92	89
Marco Scutaro	2B, 3B	SFN	4.0	52	233	210	248
Torii Hunter	OF	DET	4.0	51	44	106	256
Eric Hosmer	1B	KCA	3.9	56	41	115	101
Jed Lowrie	2B, SS	OAK	3.4	63	49	227	160
Norichika Aoki	OF	MIL	3.3	68	107	297	220
Carlos Beltran	OF	SLN	3.3	69	64	155	175
Josh Johnson	SP	TOR	3.2	71	494	187	164

Alejandro de Aza	OF	CHA	3.1	74	55	189	358
J.J. Putz	RP	ARI	3.1	75	432	153	165
Gerardo Parra	OF	ARI	3.0	77	166	290	344
CJ Wilson	SP	ANA	3.0	78	116	118	146
Tom Wilhelmsen	RP	SEA	2.9	80	373	156	159
Coco Crisp	OF	OAK	2.9	82	38	253	300
Andy Pettitte	SP	NYA	2.7	88	304	236	368
Howie Kendrick	2B	ANA	2.6	89	85	141	149
Doug Fister	SP	DET	2.6	90	189	233	161
Ernesto Frieri	RP	ANA	2.5	96	219	425	342
Hiroki Kuroda	SP	NYA	2.5	97	163	154	181
Brandon McCarthy	SP	ARI	2.2	105	468	268	253
Tim Lincecum	SP	SFN	2.2	106	223	169	273
Jonathan Lucroy	C	MIL	2.2	108	59	242	274
A.J. Pierzynski	C	TEX	2.1	109	150	324	218
Nick Swisher	OF	CLE	2.1	110	185	250	270
Carlos Ruiz	C	PHI	2.0	116	428	470	338
Jason Vargas	SP	ANA	1.9	117	380	226	265
Ubaldo Jimenez	SP	CLE	1.9	119	137	457	349
Clay Buchholz	SP	BOS	1.5	128	106	249	304
Tim Hudson	SP	ATL	1.5	129	336	286	236
Wei-Yin Chen	SP	BAL	1.4	133	404	276	247
Wandy Rodriguez	SP	PIT	1.4	135	400	282	284
Jason Grilli	RP	PIT	1.3	136	256	231	224
Kenley Jansen	RP	LAN	1.3	139	160	221	292
Marco Estrada	SP	MIL	1.3	142	293	324	469

Daniel Murphy	2B	NYN	1.3	143	13	284	217
Nate McLouth	OF	BAL	1.3	144	83	293	357
J.P. Arencibia	C	TOR	1.3	145	374	475	244
Ryan Dempster	SP	BOS	1.2	146	395	288	289
Trevor Rosenthal	RP	SLN	1.2	148	421	499	313
Drew Stubbs	OF	CLE	1.1	150	290	468	318
Mitch Moreland	1B	TEX	1.0	152	303	345	295
Jaime Garcia	SP	SLN	0.9	163	452	297	308
Wade Davis	SP	KCA	0.8	165	474	299	263
Jim Henderson	RP	MIL	0.6	178	253	499	328
Tommy Hanson	SP	ANA	0.5	179	481	316	356
Justin Masterson	SP	CLE	0.4	183	113	344	328
Al Alburquerque	RP	DET	0.3	191	454	499	338
Edward Mujica	RP	SLN	0.3	193	261	499	348
Bronson Arroyo	SP	CIN	0.2	202	180	411	326

5.4.1 Early Rounds

In the early rounds, drafting quality players is very important. It is not nearly as important that you draft the best overall fantasy player with the first overall pick as it is that you draft a quality player that will help your team. Examining Table 5.1 for overall values and where we will need to take them is important before we even get started. As we can see on Table 5.1, position players offer a lot of value, and we can see where we can obtain six of our top 25 position players without taking any of them in the first two rounds on Table 5.2. When we include Mark Trumbo, who we have ranked 40th but could safely take in the 13th round, and Marco Scutaro, ranked 52nd and available in the 16th, we have our entire roster of position players other than catcher filled with players ranked no lower than 52nd without

drafting any of them in the first two rounds! Now as mentioned before, it is important to have a balanced team, so we do not want to neglect pitching. Since we are in a unique situation here, it makes sense to draft Clayton Kershaw with the fifth pick, given that we do not see a lot of value for pitching in Table 5.1 until the late rounds of the draft. Looking down to future rounds it is unlikely that we will be able to obtain another top 75 starting pitcher unless we draft one in these first four rounds. However it is still more important that we obtain one of the top 20 ranked batters. We will thus draft Dustin Pedroia in round 2 and Adam Jones in round 3. Then in round 4 we will be able to select Adam Wainwright providing a solid start to our pitching staff as well.

To summarize, in the first four rounds we selected: Clayton Kershaw SP, Dustin Pedroia 2B, Adam Jones OF, and Adam Wainwright SP netting three players in the top fifteen of our rankings. Of these players only Wainwright (40) finished outside the top 20 in actual fantasy rank resulting in a great start to our draft.

Table 5.2. Ranking of Results

Player	Position	Round Pick	Model *z-score*	Model Rank	Actual Rank	RotoWorld Rank	Athlon Rank
Dustin Pedroia	2B	3-29	9.8	4	15	45	44
Adam Jones	OF	4-43	7.4	15	10	53	79
Paul Goldschmidt	1B	5-53	7.1	17	4	55	81
Hunter Pence	OF	8-91	7.2	16	12	110	116
Jimmy Rollins	SS	9-101	6.2	22	178	123	113
Brett Gardner	OF	11-125	6.2	21	80	162	223

5.4.2 Middle Rounds

We have already identified Paul Goldschmidt as our choice in the fifth round, Hunter Pence in round eight, and Jimmy Rollins in round 9. We still need a catcher as well as a lot of pitchers, but it is a little bit early to totally disregard high value targets that the best in class rankings are undervaluing. Table 5.1 identifies two catchers in Jonathan Lucroy and AJ Pierzynzki ranked just outside the top 100 which will likely be available with the final picks. Since we already have our starting lineup, targeting Ben Zobrist ranked 38 on Table 5.1 provides us a player to fill in for injuries and off days since he is eligible at 2B, SS, and OF (enabling him to fill in for a total of six positions). Equations 5.3 and 5.4 outline how to evaluate the probability of overlap in missed games between two players and then two groups of players assuming independence.

$$p_{sub1} = p_1\_missed + p_2\_missed - p_1\_missed * p_2\_missed \tag{5.3}$$

$$p_{sub3} = p_{sub1} + p_{sub2} - p_{sub1} * p_{sub2} \tag{5.4}$$

Since Zobrist can sub for six players with probabilities of missing a game between 0.05 (Hunter Pence) and 0.15 (Brett Gardner), he will ultimately be in the starting lineup due to injury 48% of the time. Teams having a given day off must be considered as well, which is generally two days per week over the course of a six month season or 28% of game days being off days. These two probabilities are not entirely independent as not all teams are playing on the off days, hence the injury likelihood on an off day is slightly reduced. But since we are assuming at least one of the teams will not play on each of the off days, that positive bias will offset the negative bias of assuming independence. We thus obtain a starting rate percentage of 63%, compared with the average sub opportunity based on filling in at one position and utility of 35%. We will then select Sergio Romo RP in round 7 and Hunter Pence and Jimmy Rollins as planned in rounds 8 and 9.

All five of our selections in the middle rounds are projected to finish in the top 40, including our highest rated relief pitcher, Sergio Romo (38). We did very well in the middle

rounds as well, selecting two more top 20 overall players in Paul Goldschmidt (4) and Hunter Pence (12). Pence is a great example of our model being right on (ranked 16 and finished 12), but the best in class ranks being so far off (110 and 116) that we did not need to select him any earlier than round eight. Ben Zobrist (89) finished about in line with where we drafted him (67). Although we ranked him higher at 38, again the best in class models were lower (114, 69) so we did not have to over-reach to get him on our team. Jimmy Rollins represents the first big miss ranking-wise (22 versus an actual rank of 178), but selected 101 overall he does not hurt our team. Sergio Romo was another miss, ranked as our top relief pitcher at 33, but finished the year as the tenth ranked relief pitcher at 176.

5.4.3 Late Rounds

Now in the later rounds we are looking to find potential breakout players as well as fill out our roster to ensure we have all positions covered and that our production is balanced across the statistics. Table 5.3 shows the player indices both overall and by category before positional adjustments. We can see that our strongest category is stolen bases with an average score of 1.43 and our lowest is home runs with an average score of 0.43. This will lead us to bump up the home run score by a factor of 0.25 throughout the rest of the draft. With only three pitchers there are not enough pitchers to warrant the index comparison.

Table 5.3. Ranking of Results

Position	Player	Index	HR	Runs	RBI	Ave	SB
C	Jonathan Lucroy	-0.25	-0.21	-0.07	0.34	0.45	-0.75
1B	Paul Goldschmidt	6.23	1.38	1.56	1.9	0	1.39
2B	Dustin Pedroia	6.45	0.21	1.76	0.84	1.24	2.4
3B	Marco Scutaro	0.15	-1.12	0.55	-0.84	2.16	-0.6
SS	Jimmy Rollins	4.81	-0.31	1.06	-0.15	0.44	3.77
OF	Adam Jones	6.58	1.44	1.51	1.75	1.18	0.7
OF	Hunter Pence	6.33	1.55	1.27	2.24	0.99	0.28

OF	Brett Gardner	5.32	-0.91	1.08	-1.11	0.47	5.8
U	Mark Trumbo	3.5	1.82	0.7	1.66	-0.57	-0.11
	Average	4.35	0.43	1.05	0.74	0.71	1.43

With our first pick in the late rounds we are faced with a dilemma: select the best available player, Carlos Gomez ranked 40 overall (other than Brett Gardner whom we are planning to select in the next round), or draft starting pitcher CJ Wilson who fills a needed spot on our roster, but is ranked 78 overall according to the unadjusted values on Table 5.1. Since we have not yet drafted Brett Gardner or Mark Trumbo, but only plan to, and Josh Johnson is ranked at 71 but will likely be available a few rounds later, we will select Carlos Gomez in round 10 and then Brett Gardner as planned in round 11. Now looking at Table 5.1 for starting pitcher values we can see that Andy Pettitte, Brandon McCarthy, Ubaldo Jimenez, and Clay Buchholz will all be available in the last few rounds. We will then take JJ Putz, RP in round 12, and also identify Ernesto Frieri, ranked 96, as a late round target. We then select Mark Trumbo in round 13 as planned and if we are able to select our targeted players in the late rounds we have a bench spot to fill in round 14. The best position players to fill this spot as found on Table 5.1 are Noricha Aoki, Alejandro De Aza, Gerardo Parra, and Coco Crisp. Since each of these players offer similar benefits in average and stolen bases while being a liability in RBI and home runs we will select the top ranked player in Noricha Aoki.

In the late rounds we filled out our roster with outfielders Carlos Gomez, Brett Gardner, and Noricha Aoki. JJ Putz was our second relief pitcher and our backup first baseman is Mark Trumbo. The late round results were extraordinary as Carlos Gomez finished the season ranked 11th overall. Trumbo (48), Gardner (80), and Aoki (107) all performed well above their draft position. JJ Putz was our lone disappointment here, finishing 432 as he lost his closer job early in the season and finished with only six saves. This outcome is not surprising since relief pitchers are the most volatile position in fantasy baseball.

5.4.4 Final Picks

To begin our final picks we will select Marco Scutaro (2B,3B) and Jonathan Lucroy (C) in order to fill out our starting lineup as planned. We will then select Ernesto Frieri as our third RP, followed by Andy Pettitte, Clay Buchholz, Brandon McCarthy, and Ubaldo Jimenez as starting pitchers to round out the draft.

Marco Scutaro and Jonathan Lucroy were our last position players taken, and while Scutaro drastically underperformed his model rank (52 to 233) he was not a hinderance being taken 173rd overall. Lucroy was a great surprise, drafted 187th while being ranked at 108 and finishing 59th overall. Ernesto Frieri was taken 197th and finished 219th as a closer for the full year. Our starting pitchers had mixed results with two results in the top 150 overall (Jimenez at 137th and Buchholz at 106th). Pettitte and McCarthy finishing outside the top 300, and likely will be dropped from our team early in the season.

5.4.5 Draft Results Analysis

Our draft did not go quite in accordance with our preliminary recommendations. Our recommendation was to not draft a pitcher before the fourth round. However we wound up drafting a starting pitcher in both the first and fourth rounds due to the position player values that were projected to be available later on in the draft. The results were an overall quality team. We were able to select two top three overall starting pitchers in the first four rounds as well as obtaining two more top 35 starting pitchers in the last rounds of the draft. Our relief pitchers provided mixed results as our top rated relief pitcher wound up finishing tenth in the relief pitcher rankings and our third rated RP JJ Putz lost his closer job early in the season. We were able to draft the twelfth best RP in Ernesto Frieri late in the draft so our relievers, while not exceptional, are at least average.

We used two of our first four picks on starting pitchers, but due to the model's ability to identify sleepers late in the draft using Table 5.1, we were able to obtain a very strong offense featuring four players that finished the season in the top twelve in the rankings, and

only one starter (Marco Scutaro drafted 173rd and finished 233) that did not finish in the top 100. We were even left with two bench players in the top 110, giving us a team that would be hard to beat by any standards.

Table 5.4. Ranking of Results

Player	Position	Team	Model *z-score*	Model Rank	Actual Rank	RotoWorld Rank	Athlon Rank
Mike Trout	OF	ANA	12.9	1	2	1	6
Miguel Cabrera	3B	DET	9.9	2	1	3	2
Robinson Cano	2B	NYA	9.4	3	6	6	8
Justin Verlander	SP	DET	8.6	4	135	10	3
Clayton Kershaw	SP	LAN	7.9	5	19	5	5
Prince Fielder	1B	DET	13.0	6	50	31	18
Joey Votto	1B	CIN	9.2	7	37	12	16
Josh Hamilton	OF	ANA	9.3	8	123	75	17
Evan Longoria	3B	TBA	8.9	9	45	14	42
Andrew McCutchen	OF	PIT	8.4	10	7	41	15
David Price	SP	TBA	6.7	11	153	22	12
Jose Bautista	1B, OF	TOR	7.9	12	71	20	35
Justin Upton	OF	ATL	6.0	13	52	11	32
Adrian Gonzalez	1B	LAN	5.4	14	66	13	48
Ian Kinsler	2B	TEX	7.6	15	25	30	58
Felix Hernandez	SP	SEA	5.8	16	82	36	23
Dustin Pedroia	2B	BOS	9.8	17	15	45	44
Buster Posey	C, 1B	SFN	4.9	18	88	18	27
Jason Kipnis	2B	CLE	9.3	19	5	52	46
Cliff Lee	SP	PHI	5.7	20	53	29	30
Jered Weaver	SP	ANA	5.6	21	201	37	28
Matt Cain	SP	SFN	4.7	22	247	32	20
Carlos Gonzalez	OF	COL	3.9	23	26	7	9

Victor Martinez	C	DET	5.6	24	76	35	68
Troy Tulowitski	SS	COL	3.5	25	51	8	13
Ryan Zimmerman	3B	WAS	4.3	26	58	40	22
Adam Jones	OF	BAL	7.4	27	10	53	79
Carlos Santana	C	CLE	6.4	28	75	49	76
Craig Kimbrel	RP	ATL	4.0	29	84	39	19
Cole Hamels	SP	PHI	3.9	30	162	16	25
Paul Goldschmidt	1B	ARI	7.1	31	4	55	81
Roy Halladay	SP	PHI	4.5	32	488	34	64
Gio Gonzalez	SP	WAS	3.5	33	142	51	21
Elvis Andrus	SS	TEX	5.1	34	17	50	72
Jay Bruce	OF	CIN	6.7	35	30	64	73
Stephen Strasburg	SP	WAS	3.2	36	112	19	11
Adam Wainwright	SP	SLN	4.6	37	40	46	63
Billy Butler	1B	KCA	4.3	38	158	43	40
CC Sabathia	SP	NYA	4.3	39	269	68	41
Starlin Castro	SS	CHN	3.7	40	317	28	31
Austin Jackson	OF	DET	5.2	41	110	78	56
Brandon Phillips	2B	CIN	5.7	42	28	62	71
Yu Darvish	SP	TEX	3.7	43	42	77	39
Edwin Encarnacion	1B, 3B	TOR	4.1	44	18	48	94
Allen Craig	1B	SLN	4.2	45	78	69	52
Albert Pujols	1B	ANA	2.6	46	311	9	7
Jacoby Ellsbury	OF	BOS	2.9	47	8	23	45
Chase Headley	3B	SDN	3.3	48	291	38	49
Ben Zobrist	2B, SS, OF	TBA	4.6	49	89	114	69

Ian Desmond	SS	WAS	4.8	50	32	72	78
Adrian Beltre	3B	TEX	2.8	51	29	27	24
Jose Altuve	2B	HOU	5.3	52	27	79	105
James Shields	SP	KCA	3.9	53	125	57	67
Yoenis Cespedes	OF	OAK	2.8	54	97	26	34
Desmond Jennings	OF	TBA	4.0	55	94	71	106
Kris Medlen	SP	ATL	3.2	56	117	70	50
Giancarlo Stanton	OF	MIA	2.5	57	210	24	55
Nick Markakis	OF	BAL	3.5	58	182	61	201
Mat Latos	SP	CIN	3.4	59	103	109	60
Hunter Pence	OF	SFN	7.2	60	12	110	116
Yovani Gallardo	SP	MIL	3.3	61	333	128	61
Alex Gordon	OF	KCA	3.3	62	54	65	87
Jonathan Papelbon	RP	PHI	2.3	63	270	87	29
Martin Prado	2B, 3B	ARI	4.3	64	57	91	111
RA Dickey	SP	TOR	2.9	65	173	63	54
Michael Bourn	OF	CLE	4.3	66	111	92	89
Jimmy Rollins	SS	PHI	6.2	67	178	123	113
Joe Mauer	C	MIN	2.1	68	122	42	26
Johnny Cueto	SP	CIN	2.4	69	326	82	38
David Ortiz	1B	BOS	3.5	70	20	97	85
Chris Sale	SP	CHA	2.0	71	81	74	37
David Wright	3B	NYN	1.4	72	69	21	47
Pablo Sandoval	3B	SFN	1.7	73	228	33	120
Sergio Romo	RP	SFN	5.0	74	176	139	123
Mariano Rivera	RP	NYA	2.6	75	159	104	62

Eric Hosmer	1B	KCA	3.9	76	41	115	101
Torii Hunter	OF	DET	4.0	77	44	106	256
Shin-Soo Choo	OF	CIN	2.8	78	31	81	74
Madison Bumgarner	SP	SFN	1.7	79	63	60	33
Mike Napoli	C	BOS	2.6	80	56	76	204
Carlos Gomez	OF	MIL	4.3	81	11	232	127
Fernando Rodney	RP	TBA	2.2	82	218	100	65
Adam Dunn	OF	CHA	3.1	83	141	288	100
Matt Holliday	OF	SLN	2.0	84	23	58	88
Zack Greinke	SP	LAN	1.9	85	92	54	66
Aroldis Chapman	RP	CIN	1.4	86	136	96	51
Jose Reyes	SS	TOR	0.7	87	154	25	10
Brett Gardner	OF	NYA	6.2	88	80	162	223
Addison Reed	RP	CHA	3.0	89	167	127	108
Jason Heyward	OF	ATL	0.6	90	353	44	14
Ichiro Suzuki	OF	NYA	1.3	91	249	56	142
CJ Wilson	SP	ANA	3.0	92	116	118	146
Bryce Harper	OF	WAS	0.3	93	98	15	43
Joe Nathan	RP	TEX	2.6	94	127	116	115
John Axford	RP	MIL	2.4	95	445	108	144
Matt Wieters	C	BAL	1.5	96	121	80	99
Dan Haren	SP	WAS	2.8	97	331	124	203
Alcides Escobar	SS	KCA	2.7	98	215	179	126
Freddie Freeman	1B	ATL	1.0	99	21	59	84
Jordan Zimmerman	SP	WAS	0.8	100	73	95	53
Hanley Ramirez	SS	LAN	0.1	101	43	17	57

Jed Lowrie	2B, SS	OAK	3.4	102	49	227	160
Mark Trumbo	OF	ANA	4.4	103	48	194	184
Nelson Cruz	OF	TEX	2.5	104	128	129	193
Carlos Beltran	OF	SLN	3.3	105	64	155	175
J.J. Putz	RP	ARI	3.1	106	432	153	165
Howie Kendrick	2B	ANA	2.6	107	85	141	149
Matt Harvey	SP	NYN	1.8	108	65	241	110
Ian Kennedy	SP	SDN	2.3	109	417	174	130
Jim Johnson	RP	BAL	1.0	110	168	132	80
Josh Johnson	SP	TOR	3.2	111	494	187	164
Tom Wilhelmsen	RP	SEA	2.9	112	373	156	159
Michael Brantley	OF	CLE	2.0	113	90	125	277
Yadier Molina	C	SLN	1.1	114	61	90	90
Paul Konerko	1B	CHA	2.3	115	435	137	188
Max Scherzer	SP	DET	0.7	116	22	117	70
Kendrys Morales	1B	SEA	2.4	117	115	143	154
Erick Aybar	SS	ANA	1.3	118	169	107	138
Doug Fister	SP	DET	2.6	119	189	233	161
Hiroki Kuroda	SP	NYA	2.5	120	163	154	181
Matt Kemp	OF	LAN	-0.4	121	415	4	4
Marco Scutaro	2B, 3B	SFN	4.0	122	233	210	248
Alejandro de Aza	OF	CHA	3.1	123	55	189	358
Jake Peavy	SP	CHA	0.8	124	235	166	98
B.J. Upton	OF	ATL	0.5	125	476	89	133
Drew Storen	RP	WAS	-0.1	126	446	93	59
Rafael Soriano	RP	WAS	0.9	127	194	112	173

Anibal Sanchez	SP	DET	1.3	128	79	200	136
Salvador Perez	C	KCA	0.6	129	118	120	97
Hyun-Jin Ryu	SP	LAD	1.3	130	133	134	291
Tim Lincecum	SP	SFN	2.2	131	223	169	273
Pedro Alvarez	3B	PIT	1.7	132	74	175	153
Zack Cozart	SS	CIN	2.0	133	237	297	167
Chase Utley	2B	PHI	1.9	134	39	182	163
Jon Lester	SP	BOS	1.6	135	155	158	174
Jon Jay	OF	SLN	1.8	136	147	165	180
Norichika Aoki	OF	MIL	3.3	137	107	297	220
J.J. Hardy	SS	BAL	1.0	138	99	268	131
Kyle Seager	3B	SEA	1.5	139	91	159	176
Homer Bailey	SP	CIN	0.5	140	119	205	109
Asdrubal Cabrera	SS	CLE	0.2	141	175	94	107
Huston Street	RP	SDN	0.8	142	266	144	128
Aaron Hill	2B	ARI	-0.2	143	188	67	82
Michael Cuddyer	OF	COL	1.1	144	35	151	240
Alex Rios	OF	CHA	-0.2	145	9	86	83
Glen Perkins	RP	MIN	0.4	146	179	122	211
Greg Holland	RP	KCA	1.0	147	100	168	152
Chris Davis	1B	BAL	1.7	148	3	186	222
Matt Moore	SP	TBA	-0.2	149	140	88	140
Wilin Rosario	C	COL	0.7	150	60	146	145
Joel Hanrahan	RP	BOS	-0.2	151	489	152	95
Chris Perez	RP	CLE	0.4	152	340	173	139
Andy Pettitte	SP	NYA	2.7	153	304	236	368

Adam LaRoche	1B	WAS	-0.2	154	216	103	134
Miguel Montero	C	ARI	-0.5	155	407	136	77
A.J. Pierzynski	C	TEX	2.1	156	150	324	218
Dexter Fowler	OF	COL	-0.1	157	139	176	118
Jeremy Hellickson	SP	TBA	0.4	158	394	145	190
Coco Crisp	OF	OAK	2.9	159	38	253	300
Manny Machado	3B	BAL	1.0	160	93	234	178
Alexei Ramirez	SS	CHA	-0.3	161	62	171	102
Carl Crawford	OF	LAN	-0.6	162	262	85	132
Jason Vargas	SP	ANA	1.9	163	380	226	265
Ryan Vogelsong	SP	SFN	1.2	164	484	284	196
Will Middlebrooks	3B	BOS	-0.1	165	419	150	129
Brandon Morrow	SP	TOR	-0.4	166	487	164	104
Brian McCann	C	ATL	-0.4	167	212	195	103
Lance Lynn	SP	SLN	0.7	168	157	177	269
Jonathan Lucroy	C	MIL	2.2	169	59	242	274
Andre Ethier	OF	LAN	-0.3	170	308	133	112
Steve Cishek	RP	MIA	0.7	171	190	183	231
Brett Lawrie	2B, 3B	TOR	-0.7	172	365	99	86
Matt Garza	SP	TEX	0.1	173	277	149	155
Brandon McCarthy	SP	ARI	2.2	174	468	268	253
Jason Grilli	RP	PIT	1.3	175	256	231	224
Kenley Jansen	RP	LAN	1.3	176	160	221	292
Daniel Murphy	2B	NYN	1.3	177	13	284	217
Rafael Betancourt	RP	COL	0.9	178	410	198	226
Nick Swisher	OF	CLE	2.1	179	185	250	270

Shane Victorino	OF	BOS	0.4	180	46	240	177
Denard Span	OF	WAS	1.0	181	129	212	210
Brandon Belt	1B	SFN	0.2	182	102	239	169
Todd Frazier	3B	CIN	0.3	183	186	184	172
Tim Hudson	SP	ATL	1.5	184	336	286	236
Starling Marte	OF	PIT	0.3	185	34	172	189
Gerardo Parra	OF	ARI	3.0	186	166	290	344
Neil Walker	2B	PIT	0.1	187	146	163	166
Dan Uggla	2B	ATL	0.1	188	260	167	171
Yunel Escobar	SS	TBA	1.0	189	306	361	219
Ike Davis	1B	NYN	-2.4	190	475	66	141
Clay Buchholz	SP	BOS	1.5	191	106	249	304
Wei-Yin Chen	SP	BAL	1.4	192	404	276	247
Garrett Jones	OF	PIT	0.7	193	425	428	214
David Freese	3B	SLN	-0.8	194	366	197	114
Trevor Cahill	SP	ARI	0.6	195	420	246	209
J.P. Arencibia	C	TOR	1.3	196	374	475	244
Jonathan Broxton	RP	CIN	0.2	197	457	192	221
Frank Francisco	RP	NYN	0.9	198	464	255	235
Jayson Werth	OF	WAS	-0.6	199	24	138	268
Will Venable	OF	SDN	0.6	200	70	225	325
Carlos Marmol	RP	CHN	0.2	201	455	203	272
Anthony Rizzo	1B	CHN	-2.2	202	130	98	191
Wandy Rodriguez	SP	PIT	1.4	203	400	282	284
Justin Morneau	1B	MIN	-0.7	204	211	202	162
Tommy Milone	SP	OAK	-1.1	205	287	140	315

Trevor Plouffe	3B	MIN	0.5	206	388	469	245
Wade Davis	SP	KCA	0.8	207	474	299	263
Michael Young	3B	PHI	-0.2	208	381	206	207
Grant Balfour	RP	OAK	-0.3	209	246	188	194
Bobby Parnell	RP	NYN	-0.1	210	281	216	341
Ryan Dempster	SP	BOS	1.2	211	395	288	289
Nate McLouth	OF	BAL	1.3	212	83	293	357
Ernesto Frieri	RP	ANA	2.5	213	219	425	342
Shaun Marcum	SP	NYN	0.4	214	479	254	276
Casey Janssen	RP	TOR	-0.3	215	226	228	206
Josh Beckett	SP	LAN	-0.3	216	490	209	258
Jean Segura	SS	MIL	-1.6	217	14	355	147
Mitch Moreland	1B	TEX	1.0	218	303	345	295
Jose Veras	RP	HOU	0.1	219	343	243	285
Jarrod Parker	SP	OAK	-0.7	220	225	181	186
Stephen Drew	SS	BOS	-1.6	221	208	244	151
Carlos Ruiz	C	PHI	2.0	222	428	470	338
Sean Doolittle	RP	OAK	0.2	223	427	258	288
Jaime Garcia	SP	SLN	0.9	224	452	297	308
Trevor Rosenthal	RP	SLN	1.2	225	421	499	313
Josh Willingham	OF	MIN	-0.4	226	462	218	234
Ross Detwiler	SP	WAS	-1.5	227	478	334	168
Marco Estrada	SP	MIL	1.3	228	293	324	469
Drew Stubbs	OF	CLE	1.1	229	290	468	318
Ubaldo Jimenez	SP	CLE	1.9	230	137	457	349
Wade Miley	SP	ARI	-0.6	231	274	321	212

Andrelton Simmons	SS	ATL	-1.1	232	148	287	187
Alex Cobb	SP	TBA	-1.0	233	134	193	243
Alex Avila	C	DET	-1.2	234	416	208	185
Vinnie Pestano	RP	CLE	0.2	235	466	499	278
Mike Moustakas	3B	KCA	-2.7	236	444	191	157
Gordon Beckham	2B	CHA	-0.9	237	330	201	208
Dayan Viciedo	3B	CHA	-0.9	238	372	207	246
Josh Reddick	OF	OAK	-3.8	239	351	160	250
James Loney	1B	TBA	0.0	240	164	318	279
Francisco Liriano	SP	PIT	-0.1	241	108	278	348
Shelby Miller	SP	SLN	0.1	242	109	425	287
Tommy Hanson	SP	ANA	0.5	243	481	316	356
Hisashi Iwakuma	SP	SEA	-0.2	244	68	271	282
Jhonny Peralta	SS	DET	-2.8	245	207	292	179
Kelly Johnson	2B	TBA	-0.3	246	197	374	256
AJ Griffin	SP	OAK	-1.6	247	126	348	198
Alfonso Soriano	OF	CHN	-4.6	248	16	170	275
Jim Henderson	RP	MIL	0.6	249	253	499	328
Justin Masterson	SP	CLE	0.4	250	113	344	328
Ryan Doumit	1B	MIN	-2.4	251	386	395	202
Joaquin Benoit	RP	DET	-0.4	252	252	268	300
Adam Lind	1B	TOR	-0.6	253	120	258	271
Ervin Santana	SP	KCA	0.0	254	170	408	306
Matt Carpenter	3B	SLN	-0.3	255	33	286	369
David Murphy	OF	TEX	-1.9	256	447	215	261
Bronson Arroyo	SP	CIN	0.2	257	180	411	326

Al Alburquerque	RP	DET	0.3	258	454	499	338
Jonathon Niese	SP	NYN	-0.7	259	409	268	338
Carlos Quentin	OF	SDN	-2.8	260	397	479	215
A.J. Burnett	SP	PIT	-0.9	261	132	262	314
Edward Mujica	RP	SLN	0.3	262	261	499	348
Lorenzo Cain	OF	KCA	-2.6	263	337	247	229
Mark Reynolds	3B	CLE	-2.8	264	309	338	228
Brian Dozier	2B	MIN	-1.3	265	47	278	262
Bartolo Colon	SP	OAK	0.4	266	105	416	368
Miguel Gonzalez	SP	BAL	-0.6	267	286	499	301
Phil Hughes	SP	NYA	-1.8	268	469	499	251
Nate Jones	RP	CHA	0.2	269	438	499	358
Jarrod Saltalamacchia	C	BOS	-2.4	270	101	254	298
Colby Rasmus	OF	TOR	-1.1	271	183	357	280
Justin Ruggiano	OF	MIA	-3.1	272	292	245	328
Brandon Moss	OF	OAK	-2.2	273	77	268	309
Josh Donaldson	3B	OAK	-2.6	274	36	261	267
Edwin Jackson	SP	CHN	-0.8	275	443	499	305

CHAPTER 6
CONTRIBUTIONS AND EXTENSIONS

There has been a substantive amount of quantitative work on baseball over the past twenty years. The most exhaustive was a thesis by Brad Null [44]. This dissertation has taken much of the work done prior and expanded upon it in the following ways.

The batting model described in these pages takes the Multinomial-Dirichlet Bayesian conjugate family that has been used in the literature previously, and built in age effects and reversion in a unique manner. This methodology combines parametric and non-parametric methods to account for individual prior history, the effect of age on a player, and evaluates a player's breakout and regressive probability while also accounting for mean reversion in a new and unique manner. Prior research has not blended these three methods together. The result is the best documented forecast for individual batting performance.

There has been far less work done on modeling pitcher performance. Null [44] did utilize the Multinomial-Dirichlet Bayesian family to model pitcher performance. Integrating the new hypothesis that the team defense has a stronger impact on a ball hit into the field of play than the pitcher is unique to the analysis in this dissertation. There has also been little work done on non-parametric analysis of pitcher performance. Analyzing the model errors with decision trees to forecast breakout and regressive performance in addition to mean reversion is also unique to this work. These methods produced individual pitching performance forecasts that rival the best forecasts in the industry.

Many papers have used Markov chain routines to forecast team performance. This methodology falls short when evaluating individual performance as it is not straight forward to identify which players are probabilistically on base in the Markov chain routines. The player by player simulations performed in this algorithm are unique to this dissertation. The result allows for accurate projections of runs and RBI for batters in addition to their individual

performance. Running through the batting and the pitching models for each team iteratively allows for the batting projections to be utilized when forecasting the number of runs scored behind each pitcher's start. While the pitching model merely rivals the best in class projections, the batting model far exceeds the best in class projections, resulting in the best available projection system.

While this model provides all the necessary pieces from a fantasy perspective, there are many potential extensions of the current work. It would be straightforward to evaluate different lineup options with this methodology. A professional team could use this model to evaluate the projected outcomes as a result of utilizing different batting orders. For this analysis, it would make sense to utilize the projected performances from the models outlined here, and run the simulations via a Markov chain routine. This is a much faster method, and presumably the team would be less interested in individual player statistics here than they would be in the distribution of runs scored across many games. Another extension of this same idea is to evaluate potential Major League Baseball trades. It is straightforward to swap out individual players and re-run the simulations tracking runs scored per game, runs allowed per game, wins and losses, etc.

There is also the opportunity to integrate additional data into the methodologies. Fantasy baseball is growing, and as it grows additional metrics are being utilized in the scoring. This methodology can easily be customized to include additional statistics such as on base percentage and slugging percentage, runs created, and wins above replacement (WAR). Additional pitching metrics could also be layered in as more fantasy leagues include metrics such as holds, quality starts, FIP, and xFIP in their scoring rubric.

As of yet, we have not found minor league data to be particularly reliable in forecasting major league performance for the upcoming year, but further research could be done. There are many confounding factors in the minor league data that, when accounted for correctly, could prove to be fruitful. There are at least six different levels of minor league baseball: rookie, low A, A, high A, AA, and AAA. In addition each player progresses at a

different pace, with a different amount of prior experience and at different ages. These factors all cloud the data, but there is likely quality information that could be gleaned from it.

Additional pitching data that may prove useful is pitch fx data [1]. Pitch fx tracks the speed and trajectory of each pitch. We have not yet obtained the data for analysis, but it may prove fruitful. Tracking a significant change in a pitcher's delivery, or the outcomes of particular pitches may generate insights when it comes to evaluating the likelihood of a player to regress, break out, or even get injured.

BIBLIOGRAPHY

[1] *Baseball: Fielding the future*. https://www.sportvision.com/baseball, 2015.

[2] R. ACHARYA, A. AHMED, A. DAMOUR, H. LU, C. MORRIS, B. OGLEVEE, A. PETERSON, AND R. SWIFT, *Improving major league baseball park factor estimates*, Journal of Quantitative Analysis in Sports, 4 (2008). article 4.

[3] J. ALBERT, *A bayesian analysis of a poisson random effects model for home run hitters*, The American Statistician, 46 (1992), pp. 246–253.

[4] J. ALBERT, *A statistical analysis of hitting streaks in baseball: Comment*, Journal of the American Statistical Association, 88 (1993), pp. 1184–1188.

[5] J. ALBERT, *Hitting with runners in scoring position*, Chance, 15 (2002), pp. 8–16.

[6] J. ALBERT, *Smoothing career trajectories of baseball hitters*. Bowling Green State University, 2002.

[7] J. ALBERT, *A breakdown of a batter's plate appearance - four hitting rates*, By the Numbers, (2006). February.

[8] J. ALBERT, *Pitching statistics, talent and luck, and the best strikeout seasons of all-time*, Journal of Quantitative Analysis in Sports, 2 (2006). article 2.

[9] J. ALBERT, *Streaky hitting in baseball*, Journal of Quantitative Analysis in Sports, 4 (2008). article 3.

[10] S. ALBRIGHT, *A statistical analysis of hitting streaks in baseball*, Journal of the American Statistical Association, 88 (1993), pp. 1175–1183.

[11] S. A. BARRA, ALLEN, *The myth of clutch hitting*, Wall Street Journal - Eastern Edition, W5 (1999), pp. 234–255.

[12] B. BAUMER, *Why on-base percentage is a better indicator of future performance than batting average*, Journal of Quantitative Analysis in Sports, 4 (2008). article 3.

[13] B. BAUMER, *Using simulation to estimate the impact of baserunning ability in baseball*, Journal of Quantitative Analysis in Sports, 5 (2009). article 8.

[14] S. M. BERRY, C. S. REESE, AND P. D. LARKEY, *Bridging different eras in sports*, Journal of the American Statistical Association, 94 (1999), pp. 661–676.

[15] S. M. BERRY, C. S. REESE, AND P. D. LARKEY, *Bridging different eras in sports*, Anthology of Statistics in Sports, 16 (2005), p. 209.

[16] B. BUKIET, E. HAROLD, AND J. PALACIOS, *A markov chain apporoach to baseball*, Informs, 45 (1997), pp. 14–23.

[17] C. CARLUCCIO, *Drafting strategies in the major league baseball draft*, (2011).

[18] R. CONNOR AND J. MOSIMAN, *Concepts of independence for proportions with a generalization of the dirichlet distribution*, Journal of the American Statistical Association, 64 (1969), pp. 194–206.

[19] E. COOK, *Percentage baseball*, MIT Press, (1964).

[20] T. COVER AND C. KEILERS, *An offensive earned run average for baseball*, Operations Research, 25 (1977), pp. 729–740.

[21] R. CRAMER, *Do clutch hitters exist?*, Baseball Research Journal, 6 (1977), pp. 74–79.

[22] D. D'ESOPO AND B. LEFKOWITZ, *The distribution of runs in the game of baseball*, Optimal Strategies in Sports, (1977). New York: Elsevier-North Holland.

[23] D. DODDS AND J. BRYANT, *The principles of vbd revisited*. http://www.footballguys.com/05vbdrevisited.htm, 2005.

[24] S. DUN, G. FLEISIG, J. LOFTICE, D. KINGSLEY, AND J. ANDREWS, *The relationship between age and baseball pitching kinematics in professional baseball pitchers*, Journal of Biomechanics, 40 (2007), pp. 265–270.

[25] R. FAIR, *Estimated age effects in baseball*, Journal of Quantitative Analysis in Sports, 4 (2008). article 1.

[26] R. FREEZE, *An analysis of baseball batting order by monte carlo simulation*, Operations Research, 22 (1974), pp. 728–735.

[27] M. J. FRY, A. W. LUNDBERG, AND J. W. OHLMANN, *A player selection heuristic for a sports league draft*, Journal of Quantitative Analysis in Sports, 3 (2007).

[28] C. HARRIS, *How to make vbd work for you*. http://sports.espn.go.com/fantasy/football/ffl/story?page=nfldk2k12_ vbdwork, 2012.

[29] N. HIROTSU AND M. WRIGHT, *A markov chain approach to optimal pinch hitting strategies in a designated hitter rule baseball game*, Journal of Operations Research Society of Japan, 46 (2003), pp. 353–371.

[30] N. HIROTSU AND M. WRIGHT, *Modeling a baseball game to optimise pitcher substitution strategies incorporating handedness of players*, Journal of Management Mathematics, 16 (2005), pp. 179–194.

[31] N. HIROTSU AND M. WRIGHT, *Modeling a baseball game to optimize pitcher*,

Economics, Management and Optimization in Sports, (2013), p. 131.

[32] R. HOWARD, *Dynamic programming and markov processes*, MIT Press, (1960).

[33] B. JAMES, *The Bill James Baseball Abstract, 1984*, Ballantine Books New York, 1984.

[34] S. JENSEN, B. MCSHANE, AND A. WYNER, *Hierarchical bayesian modeling of hitting performance in baseball*, Bayesian Analysis, 4 (2009), pp. 631–652.

[35] S. T. JENSEN, B. B. MCSHANE, A. J. WYNER, ET AL., *Hierarchical bayesian modeling of hitting performance in baseball*, Bayesian Analysis, 4 (2009), pp. 631–652.

[36] S. T. JENSEN, K. E. SHIRLEY, AND A. J. WYNER, *Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball*, The Annals of Applied Statistics, (2009), pp. 491–520.

[37] M. LEWIS, *Moneyball*, W.W. Norton and Co., New York, 2004.

[38] V. MCCRACKEN, *Pitching and defense: How much control do hurlers have?* http://baseballprospectus.com/article.php?articleid=878, 2001.

[39] E. MILLS AND H. MILLS, *Player win averages*, AS Barnes & Co., Cranbury, NJ, (1970).

[40] MISC, *Play by play statistics*. www.retrosheet.org. accessed September, 2013.

[41] J. MOSIMANN, *On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions*, Biometrika, 49.

[42] H. NOBUYOSHI, *Reconsideration of the best batting order in baseball*, Journal of Quantitative Analysis in Sports, 7 (2011). article 13.

[43] B. NULL, *Modeling baseball player ability within a nested dirichlet distribution*, Journal of Quantitative Analysis in Sports, 5 (2009). Article 5.

[44] B. NULL, *Stochastic Modeling and Optimization in Baseball*, PhD thesis, Department of Management Science and Engineering, Stanford University, 2009.

[45] J. PAISLEY AND L. CARIN, *Dirichlet process mixture models with multiple modalities*, in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, IEEE, 2009, pp. 1613–1616.

[46] M. PANKIN, *Finding better batting orders*. http://www.pankin.com/markov/btn1191.htm, 1991.

[47] M. B. B. A. PIETTE, JAMES AND S. JENSEN, *A point-mass mixture random effects model for pitching metrics*, Journal of Quantitative Analysis in Sports, 6 (2010).

[48] J. PLIMI, *Fantasy sports league pre-draft logic method*, Feb. 23 2006. US Patent App. 11/208,112.

[49] F. QUINTANA, P. MÜLER, G. ROSNER, AND M. MUNSELL, *Semi-parametric bayesian inference for multi-season baseball data*, Bayesian analysis (Online), 3.

[50] R DEVELOPMENT CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[51] E. SCHALL AND G. SMITH, *Do baseball players regress toward the mean?*, The American Statistician, 54 (2000), pp. 231–235.

[52] R. SCHULZ, D. MUSA, J. STASZEWSKI, AND R. SIEGLER, *The relationship between age and major league baseball performance: Implications for development*, Psychology and Aging, 9 (1994), pp. 274–286.

[53] J. SKELLAM, *A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials*, Journal of the Royal Statistical Society. Series B (Methodological), 10.

[54] J. SOKOL, *A robust heuristic for batting order optimization under uncertainty*, Journal of Heuristics, 9 (2003), pp. 353–370.

[55] H. STERN, *A statistician reads the sports page: Baseball by the numbers*, Chance, 10 (1997), pp. 38–41.

[56] G. STEVENS, *Bayesian Statistics and Baseball*, PhD thesis, Pomona College, 2013.

[57] T. TAKEI, S. SEKO, AND K. ANO, *Improved optimal batting order with several effects for baseball*, Kyoto University Bulletin, 1194 (2001), pp. 87–96.

[58] J. THORN AND P. PALMER, *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and its Statistics*, Doubleday, Garden City, New York, 1984.

[59] J. WILCOCK, *System and method for conducting a fantasy sports draft*, 2005. US Patent App. 10/765,684.

[60] K. WOOLNER, *Baseball Between the Numbers*, Basic Books, New York, New York, 2006.

APPENDIX A
SIMULATION ROUTINE AND COMPUTATIONAL EFFICIENCY

SIMULATION ROUTINE AND COMPUTATIONAL EFFICIENCY

This chapter will discuss in more detail the algorithms discussed in the prior chapters as well as look at potential methods for improving the computational efficiency. Currently each seasonal run of each team is run sequentially, which leads to almost 50,000 games being simulated in order to achieve the final results. The simulations were run on an Intel quad Core i3 processor at 2.40 GHz with 6.0 GB of RAM. It takes just over half a second to simulate one game, so the entire simulation takes just over one day to complete.

As outlined in the algorithm, each team's batting and pitching results are simulated concurrently. Ten seasons are simulated and the results of each season is tracked independently in order to obtain a measure of the player's stability as well as their expected outcome. These results are then utilized in order to rank the players projected productions for use in fantasy baseball leagues.

A.1 A BEGINNERS GUIDE TO PARALLEL COMPUTING

As this paper and my expertise are focused more on the statistical elements than computational, this section will only skim the surface of what parallel computing has to offer. It is clear that this modeling structure would benefit greatly from the ability to simulate games in parallel rather than in series. If computational power were not lost in running in parallel one could theoretically simulate all 50,000 games in about half a second, rather than in just over one day. In order for this to be achieved, however, one would have to have 50,000 cores or CPUs available for the process.

In general parallel computing takes processes that are being run in series and assigns tasks to different CPUs or cores on a multi-core machine in order to increase efficiency. A manual approach to parallel computing in this example would be to open up four instances of

R on a four core machine. One could then run the simulations for seven teams in two of the R instances and eight teams on the other two instances. R is built to utilize different cores with each instance of R, when possible, so the result is that the computational time will be divided by four.

R [50] has packages that are designed to automatically assign these tasks to different cores, thus significantly increasing the scalability of the aforementioned process as well as streamlining the output. A cluster is defined as a group of independently defined machines referred to as nodes, working together. R can leverage a computer cluster in order to cut the amount of processing time required down to minutes rather than over a day, if the cluster is big enough.

Of course with increased speed also comes increased complexity. In order for the machines to work together communication protocols must be created, and the communications also take time and computing power. To maximize the computational efficiency one should minimize the ratio of data transmitted between the machines, to data processed by the individual machines. For example in this context it would be most efficient, if utilizing 30 cores, to assign one baseball team's simulations to each core, rather than passing back and forth the data for multiple baseball teams to each core.

The CPUs/GPUs can work together in a number of different methodologies. The methodologies below, as well as the R package framework, will utilize the master and slave architecture. One master process controls the other slave nodes and communication is done through message passing methods. These methods for communication supported by the R package in the following chapter are:

- Sockets
- Message Passing Interface
- NetWork Spaces
- Parallel Virtual Machine

There are advantages and disadvantages of each of these methods, diving into these methodologies is beyond the scope of this paper.

A.2 Parallel Computing in R

While R was built with sequential computing in mind, there have been a few packages developed in order to incorporate parallel computing and the efficiencies that it creates. One of the simpler packages to utilize is the Snowfall package. The Snowfall package is an enhancement to the Snow package that preceded it in order to make the functionality more accessible and user friendly.

The Snowfall package is set up in such a way that the same code can be run either on a multicore machine or a cluster. A simple command creates a cluster using the communication methodology that is specified from the list in the previous section, socket mode is the default. Using this methodology one can take advantage of a cluster in order to run the simulations in a fraction of the time.

San Diego State's Computational Science Research Center provides a number of clusters which are available for use. One of these clusters has 96 CPU cores. Using this methodology and the clusters available at SDSU the process that takes over a day running in series on a single machine can be completed in under 30 minutes on the cluster. This provides the advantage of testing algorithm changes rapidly as well as enabling more data runs, thus providing more precise results.