A MIXED EFFECTS MULTINOMIAL LOGISTIC-NORMAL MODEL FOR FORECASTING BASEBALL PERFORMANCE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Eric A. E. Gerber

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Bruce Craig, Chair

Department of Statistics, Purdue University

Dr. Hyonho Chun

Department of Mathematics and Statistics, Boston University

Dr. George McCabe

Department of Statistics, Purdue University

Dr. Vinayak Rao

Department of Statistics, Purdue University

Approved by:

Dr. Hao Zhang

Head of the Department of Statistics, Purdue University

ACKNOWLEDGMENTS

There are so many people without whom it would not have been possible to complete this work. My advisor: Dr. Bruce Craig, whose patient guidance and subtle humor I have benefited from over the last four years. Through our many meetings, he has had a profound impact on me as both a researcher and educator. My committee: Dr. Vinayak Rao, Dr. George McCabe, and Dr. Hyonho Chun, who each provided advice and encouragement over the years. My Japanese baseball data provider Data Stadium Inc. for patiently awaiting my irregular reports on my progress, and Dr. Jim Albert for connecting me with them. Former department head, now Dean Rebecca Doerge, who would sit me down and tell me what I needed, instead of what I wanted, to hear. Laura and Ce-Ce, who gave me freedom to develop my teaching skills. Doug, without whom my inefficient code would still be running to this day.

I would also like to thank my peers. My roommates: John, Tanner, Qi, Deborah, Whitney, and Ryan, who each in their unique way made home worth coming home to, but especially Tim, my friend and roommate of six years, who is the reason I applied to Purdue in the first place. My office mates, especially Min for providing just the right balance of distraction and motivation. My academic brothers, Zach and Will. Nathan and Mohit, who got me through the first few years, and a host of others I have counted among my best friends; Lawlor, Jeremy, T-Mike, Evidence, Bertjan, Jiasen, Li Cheng, Sophie, Ayu, Emery, Jiapeng, with special places in my heart for Yumin and AMT. My friends from Ag Econ; Sofi, Jeff, Brian and Christi, David, Stacy and Ray, who helped me forget about statistics every once in a while. My girlfriend, Jillian, who has helped make my last year of graduate school the most romantic time of my life.

Finally, my family. My sister Nicole, who always tells me I got this. My brother Chad, who is my best friend and inspires me more than anyone. My mother, who is so full of love and support, and my father, who taught me to love baseball and, more importantly, to pursue passion and excellence in whatever I do.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Gerber, Eric A. E. Ph.D., Purdue University, August 2019. A Mixed Effects Multinomial Logistic-Normal Model for Forecasting Baseball Performance. Major Professor: Bruce A. Craig.

Prediction of player performance is a key component in the construction of baseball team rosters. Traditionally, the problem of predicting seasonal plate appearance outcomes has been approached univariately. That is, focusing on each outcome separately rather than jointly modeling the collection of outcomes. More recently, there has been a greater emphasis on joint modeling, thereby accounting for the correlations between outcomes. However, most of these state of the art prediction models are the proprietary property of teams or industrial sports entities and so little is available in open publications.

This dissertation introduces a joint modeling approach to predict seasonal plate appearance outcome vectors using a mixed-effects multinomial logistic-normal model. This model accounts for positive and negative correlations between outcomes both across and within player seasons. It is also applied to the important, yet unaddressed, problem of predicting performance for players moving between the Japanese and American major leagues.

This work begins by motivating the methodological choices through a comparison of state of the art procedures followed by a detailed description of the modeling and estimation approach that includes model fit assessments. We then apply the method to longitudinal multinomial count data of baseball player-seasons for players moving between the Japanese and American major leagues and discuss the results. Extensions of this modeling framework to other similar data structures are also discussed.

1. MEASURING BASEBALL PLAYER PERFORMANCE

1.1 Introduction

Baseball is one of the most attended spectator sports in the world. Both Major League Baseball (MLB) of the United States and Nippon Professional Baseball (NPB) of Japan are the most attended sporting leagues in their respective countries. This industry works with billions of US dollars and the smallest decisions, whether on the field or in management, can have significant impacts on team and player performances and profit. Since the reorganization of the NPB into its current form in 1950, nearly every year at least one baseball player has transitioned between the two major baseball leagues.

There is inherent risk involved in transitioning between the two leagues for both player and team. MLB players transitioning from the US to Japan tend to cost NPB teams more in salary than Japanese players. Players moving from NPB to MLB are more often among the strongest in Japanese baseball, yet are coming from a league most consider weaker than MLB. They also can cost MLB teams a non-trivial amount simply for the rights to negotiate with them. In both directions of transition, it is important for the team management to determine the worth and level of the added investment.

Additionally, there are distinct cultural differences in how the two countries approach the game. This "culture shock" presents another risk to players making the transition. Japanese teams are known for adhering to a much stricter practice routine than their American counterparts. This differing level of expectations, as well as (in both cases) not necessarily speaking the primary language of the country, can lead

to discomfort on the player's part that could affect performance. More practically, the language barrier becomes a larger issue for pitchers and catchers, who must communicate throughout the game. Both MLB and NPB teams provide translators for foreign players on their rosters, to help mitigate this issue. However, many players do not choose to stay in the foreign league long simply due to not having the comforts of home readily available.

Because of these risks, it becomes invaluable for teams from both leagues to have reliable tools for predicting a player's performance when moving between the United States and Japan. Certainly the baseball operations departments of each team have their own methods for predicting performance and access to proprietary data sets that track not only the traditional baseball statistics but also advanced measures, such as pitch spin and angle, batting launch angle, or spatial fielding metrics. To this point, however, the question of predicting baseball player's performance when moving between MLB and NPB has not been addressed in an academic setting with data that are publicly available.

The goal of this dissertation is to develop novel statistical methodology that will serve, among other applications, to predict performance of baseball players moving between the MLB and NPB leagues. In the remaining sections of this chapter, the current state of baseball prediction literature is discussed and some traditional approaches to modeling the data are presented. In the second chapter, our mixed-effects multinomial logistic-normal model is introduced, along with a Gibbs sampling algorithm for estimation, tools for assessing model fit, prediction and assessing uncertainty about those predictions. The third chapter describes our data, assesses our model-fitting algorithm via a simulation study, and then applies the methodology to the real data set that contains all players who have played in both Japan's NPB and the United States' MLB leagues. The final chapter summarizes the statistical contributions of this work and presents extensions and avenues for future research.

1.2 Review of Baseball and Sports Performance Prediction Literature

Statistics have played a major role in the growth and popularity of baseball. They have also evolved over time. What were once just simple summaries (e.g. runs, hits, batting average) available in the newspaper box scores the following day are now highly intricate and almost instantaneously available. For example, in a current televised broadcast of a baseball game, before each pitch, statistics describing a player's (pitcher and/or batter) performance in that situation are reported to the viewer. The intention is to give the consumer a sense, using data, of the talent level of the players they are watching and what might be expected from the players on the next pitch. Statistics, like these, but often far more intricate, are used by the head coaches (called managers) to make in-game decisions, such as player substitutions, pitch type and location, or defensive shifts. Statistics are also used by team ownership to make economic decisions; how large a contract to offer players, or to make their case in salary arbitration.

Beginning with Frederick Mosteller's (1952) work on predicting the winner of Major League Baseball's World Series, the sport of baseball has been fertile ground in academia for the development of statistical forecasting systems. Most work in sports prediction focuses on univariate and binary on-field performance metrics. One of the earliest examples of player prediction comes from Lindsey (1959), who utilizes proportion tests to predict a baseball player's batting average based on the handedness of the batter-pitcher match-up. Subsequently, in a seminal paper, Efron and Morris (1975) use Stein's estimator via empirical Bayes to predict batting averages for baseball players for the remainder of the 1970 season.

The focus on single continuous, or binary, outcomes continued for many years and is still a motivating problem in sports, especially baseball. Bennet and Flueck (1983) analyze run production metrics in terms of prediction of baseball team scoring uti-

lizing linear model-building. Both Albert (1994) and Berry et al. (1999) use various forms of logistic regression to predict the success of individual at-bats in baseball. Albright (1993) compares first-order homogeneous Markov chain models and logistic regression models to predict baseball at-bats as binary outcomes. Rosner, Mosteller, and Youtz (1996) use truncated binomial regression to model baseball pitcher performance in the form of the number of runs allowed. These examples serve to illustrate the state of baseball prediction in academia over the 50 years since Mosteller's work.

More recently, there has been a broader and more rich tapestry of statistical models applied to the problem of prediction in baseball and other sports, including a greater emphasis on hierarchical and non-parametric models. Berry et al. (1999) use a hierarchical model to compare home run and batting average abilities across eras. In a non-parametric setting, Chandler and Stevens (2012) use random forests to project future success of minor-league baseball players, where their metric for success was simply a binary outcome of playing in greater than 320 Major League games. Models with similar goals of predicting performance in sports other than baseball have been developed for predicting squash victories via a first-order Markov chain model of match-play (McGarry and Franks, 1994), modeling NFL quarterback performance with negative binomial regression (Wolfson et al., 2011), and predicting average runs for batsmen in cricket via a hierarchical linear mixed model (Wickramasinghe, 2014).

Within this realm of published academic research, however, there has been little focus on developing models that seek to predict the entire vector of plate appearance outcomes. In other words, not just predicting whether the player gets a hit, but whether the hit is a single, double, triple, or home run, whether the out is a strike out, out in play and whether the player got a sacrifice or was walked, which are all possible in baseball. Null (2009) defines 14 distinct outcomes for a baseball players plate appearance and assumes each batter has some "true" probability vector for experiencing each outcome. Null's work also serves as an introduction to the class of

nested Dirichlet models, used as an alternative to the Dirichlet or Generalized Dirichlet for modeling multinomial data. Finally, Albert develops a Bayesian random effects model for estimating component rates of batting outcomes in order to better predict batting average (Albert, 2016). Albert's work does treat the outcomes as multinomial in nature, but factorizes the multinomial likelihood into a product of conditional binomial likelihoods. However, of these two, only Null attempts to include covariates, using a fixed effects regression model to attempt capturing an age effect.

Outside the realm of academic journals, there is an ample and active sector of the internet where baseball projection systems are developed and implemented. The rise of fantasy baseball as a betting and leisure activity has given cause for not only large corporations, but private individuals, to develop models for forecasting player performance. These models seek to predict a player's entire seasonal performance. Most of the results of the projections from these models are published for free on the internet, although betting fantasy websites require payment to access their predictions, and are often concerned more with daily performance than seasonal performance.

Unfortunately, only one of the most common projection systems is completely open source, with all others maintaining some proprietary ownership of the methodology used. Not only that, but there is a relative dearth of uncertainty quantification associated with the resulting projections, which would allow for comparing the "floor" and "ceiling" of different players. In a fantasy baseball setting, users often must consider the trade-off between rewards for high- or low-risk moves, but this risk is entirely based on the user's perception of players and not on any quantity provided with projections being used. In a team management setting, uncertainty facilitates direct comparison of players when considering signing or trade moves.

Any discussion of forecasting systems for seasonal baseball performance would do well in starting with the self-described most basic forecasting system. Claiming to

use as little intelligence as possible, The Marcel the Monkey Forecasting System (the Marcels) was developed by Tom Tango and first published in 2001. The process for predicting a player's performance based on the Marcels is completely available online, and consists mainly of a weighted average that utilizes the last three seasons of a player's career, including an age factor, and regressing to the mean for each component by creating the same weighted average of the player's league average and then adding 1200 plate appearances at that weighted league average rate to the weighted plate appearances from the player's last three seasons (Tango, 2004).

Despite Tango declaring that he does not stand by the system, the Marcels tend to perform competitively with many of the proprietary systems. Additionally, the Marcels is one of the only systems with some level of uncertainty assessment, via a singular value which describes the "reliability" of a prediction; simply the proportion of the prediction which comes from player data as opposed to the regression to the mean (e.g. a player who had a weighted total of 7000 plate appearances over three years would have a reliability score of $\frac{7000}{7000+1200} = 0.854$). As Tango describes on his website, players with no data, including minor league players and players coming from Japan, will be predicted at the mean and have a reliability score of 0 (Tango, 2012).

More complicated projection systems also annually publish predictions for the upcoming baseball season. While some of the methodology of these systems, or at least choices which guide the methodology, are occasionally discussed on the internet, none of the systems have been published and thus are maintained as proprietary. Among the most popular of the projection systems is Steamer, originally developed in 2008 and currently maintained by Jared Cross, Dash Davidson, and Peter Rosenbloom. PECOTA, or the Player Empirical Comparison and Optimization Test Algorithm, is the proprietary system of Baseball Prospectus originally developed by Nate Silver in 2002-2003. ZiPS, or the sZymborski Projection System, was developed by Dan Szymborski while writing at the website Baseball Think Factory. Recently, in 2010

Brian Cartwright developed Oliver, a relatively basic forecasting system more in the vein of the Marcels, except that Cartwright developed his own Major League Equivalencies for minor league players. KATOH, developed by Chris Mitchell in 2014, is another proprietary system focused on forecasting major league hitting with minor league statistics.

These represent some of the more popular and important projection systems that have been developed. All MLB teams, and likely most NPB league teams have also developed their own in-house projection systems. There are many more systems which currently exist, and many more will certainly be developed. The main advantages to proposing the projection system described in this research are the open source nature of the methodology, in which it is free for anyone to critique, tinker with, and improve upon, and the ability to assess risk, the uncertainty quantification about the predictions being made.

There has been minimal work in the literature of sports prediction in providing prediction intervals, or accuracy scores for the outcomes being predicted. Mentioned previously, the Marcels is one of the only projection systems to include a reliability metric for the predictions, which he specifically notes gives a score of 0 for minor league players and players coming from Japan (Tango, 2012) and thus the Japanese player is predicted at the league average for their age. Will Larson, as described in Section 2.5.3, has done some of the most work in assessing the predictive accuracy of the results from online projection systems. In 2011, he published an article on Fangraphs where he calculates the variance in forecasts of six projection system results published online (Larson, 2011). He compares that variance's ability to predict the absolute error for a particular players consensus forecast (defined as the average among the six different forecasts) with the Marcels' reliability measure.

For these key reasons (lack of open source methods and lack of uncertainty quantification) we feel adding our method to the lexicon of sports prediction is a useful contribution. The following section begins laying the foundation for the methodological choices that drive the development of our proposed model. We describe the multinomial random variable distribution and how the distribution itself is related to baseball data, but fails to accurately describe such data by itself. We then briefly describe three alternative hierarchical models that may be considered for predicting baseball data.

1.3 Modeling Baseball Player Performance Data

The most readily available baseball performance data takes the form of seasonal performance vectors of counts. Each time a batter or pitcher has a plate appearance (for pitchers, instead of plate appearance the count is called "batter faced"), there are a finite number of distinct outcomes that can occur. The outcome of a plate appearance has an immediate impact on the state of the game. While baseball statisticians have only recently, in the relative terms of the sport, begun to focus on outcomes on the singular pitch level, historically analysis has been performed on a plate appearance level. In assessing the performance of baseball players, it is natural to consider the outcomes from the player's seasons worth of plate appearances. It is usually on this level that metrics for player valuation occur.

In predicting how a player will perform in the ensuing season, there are several strategies, each aligning with a different goal. Singular metrics of player value are mostly meant for ranking and comparing players in an inferential sense, not necessarily a predictive context. Among these metrics, Wins Above Replacement (WAR) is the most popular, though there are many different versions of WAR (Baumer et al., 2015). These metrics convert the player's performance in terms of outcomes first into how many runs the individual player contributed to his team, and then into wins based on

the number of wins a run is worth, and finally scaled so that a value of 0 represents a "replacement" player. A more simple, but perhaps more popular metric, is On-Base plus Slugging Percentage (OPS), where the calculation is simply the sum of two ratios.

In projection systems these singular metrics, while capable of being calculated, are not the main goal. Instead the goal is the vector of outcome counts for the ensuing season; the number of home runs, strikeouts, walks, or other potential outcomes of each player. While the granularity of the number of outcomes can take many forms, arguably the most relevant list of outcomes is made up of ten outcomes for a single plate appearance:

- Base hits: (Single, Double, Triple, Home Run).
- Free passes: (Base on Balls, Hit-by-Pitch).
- Sacrifices: (Sacrifice Bunt, Sacrifice Fly).
- Outs: (Strikeout, Out-in-Play).

There are several projection models, that are published each year in the vein of predicting a vector's worth of plate appearances. As summarized in Section 1.2, two major drawbacks of the existing projection models are that many are proprietary, and rarely do they provide a form of uncertainty quantification.

Baseball is as much a business as it is a game. Developing and providing some level of uncertainty about predictions of player performance can greatly assist management in evaluating the risk associated with paying players. Especially in a comparative setting, if management is deciding between two players to pursue, a point estimate of each player's performance (their value) may suggest one player is better. However, an estimate of uncertainty (their risk) may show that the other player has a higher potential floor.

Our method departs from many other methods by modeling additional uncertainty about a player's outcome probability vector. Specifically, given some covariates, we assume there is still a distribution about each player's probability vector. This additional variability factors in components such as injury and what ball parks he primarily plays in that could affect seasonal performance. This additional uncertainty also allows us more flexibility in modeling the relationships among outcomes.

Based on estimates of this model for a predicted season, intervals describing the uncertainty may be formed. Some major goals of this research are to contribute a model in which prediction accuracy is competitive with existing models, is completely open source, and provides credible intervals for the seasonal counts being predicted which are comparable between players. To do so, we begin by defining a distributional assumption about the seasonal player performance count vector. We then proceed to developing a consistent narrative for modeling those data.

1.3.1 Multinomial Distributed Data

Probability theory provides a natural distribution, the multinomial distribution, for modeling a vector of outcome counts, conditional on a probability vector associated with those counts. The multinomial distribution in its most basic form relies on two parameters, the number of independent trials N and the probability vector $\Pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$, where K represents the number of unique outcomes. We denote a random vector W that follows the multinomial distribution as:

$$W \sim Multi_K\left(N, \Pi\right)$$

Definition. A vector of counts, W following the multinomial distribution implies the following joint probability mass function over the non-negative count vector, W:

$$f\left(w_1, ..., w_K | N, \Pi\right) = \frac{N!}{w_1! \cdots w_K!} \prod_{k=1}^{K} \pi_k^{w_k} \tag{1.1}$$

In this function, K is the number of categories, N is the number of trials, and $\Pi = (\pi_1, ..., \pi_K)$ is the vector of probabilities associated with these unique categories such that $\sum_{k=1}^{K} \pi_k = 1$. The multinomial distribution can also be described in terms of each independent trial; $W_t \sim Multi_K\,(1, \Pi)$, where $W = \sum_{t=1}^{N} W_t$.

Several properties of the multinomial are straightforward extensions of binomial distribution properties. For example, the first and second moments of each count w_k produce familiar terms for the expected value and variance for the counts of each outcome:

$$E\,[w_k] = N\pi_k \tag{1.2}$$

$$Var\,(w_k) = N\pi_k\,(1 - \pi_k) \tag{1.3}$$

The remainder of the variance-covariance matrix ($k \neq k'$) is made up of the off-diagonal covariance terms,

$$Cov\,(w_k, w_{k'}) = -N\pi_k\pi_{k'} \tag{1.4}$$

Resulting in correlations:

$$Cor\,(w_k, w_{k'}) = -\sqrt{\frac{\pi_k\pi_{k'}}{(1 - \pi_k)\,(1 - \pi_{k'})}} \tag{1.5}$$

When the vector of probabilities depend on some covariates, modeling of multinomial data via multinomial logisitic regression, or the multinomial logit model, is most common. Given K states, the probabilities of those states are described using $K - 1$ logit equations, where the inverse additive logistic transformation of the probabilities produces log-odds, relative to a chosen baseline outcome (we will assume the final category, K) which are each linked to a covariate vector X:

$$log\left(\frac{\pi_k}{\pi_K}\right) = X\beta_k \text{ for } k = 1, ..., K - 1$$

For the multinomial logit, this means that each of the probabilities is function of the covariates as follows:

$$\pi_k = \frac{e^{X\beta_k}}{1+\sum_{k=1}^{K-1} e^{X\beta_k}}$$

1.3.2 Baseball Data as Multinomial Data

Some arguments could be made that a baseball player's plate appearances over the course of a season are not independent. In each plate appearance, there are many different factors which impact a batter or pitcher's success; the specific opposition being the most obvious, but other factors such as player health or weather. However, there have been several studies examining streaky hitting in baseball (Albright, 1993; Albert, 2008; Albert, 2013) with the general consensus that there is not a great deal of difference between a player's behaviour and a random model assuming constant probability of success. This leads us to assume that it is not unreasonable to treat a player's seasonal plate appearances as a set of independent trials.

In general baseball terms, it is possible to define any level of granularity to the outcomes, the dimensionality of the multinomial count vector, K. Traditionally in predicting player performance, the problem is reduced to a binomial setting; predicting a single outcome. For predicting overall player performance, predicting each outcome in this setting does not make practical sense, as it ignores correlations between outcomes which may contribute significant information that may improve predictive accuracy. We also wish to account for potential variability in the probability vector over seasons.

Note that within a player-season, assuming a given probability vector and multinomial distribution, there is necessarily a negative correlation between each pair of outcomes (Equation 1.5). As will be discussed in detail later in this chapter, there are several outcomes in baseball that are known to be positively correlated within a player-season. Even more basically, while we use covariates to allow the mean probability to

differ across player-seasons, we expect there to be some variability about that mean, or overdispersion. Addressing these issue drives, in large part, the methodological choices in developing our model for baseball performance data. To account for varying probabilities, we describe a model which augments the base multinomial model via a hierarchical structure where Π_{ij} follows a distribution.

1.3.3 Hierarchical Structure of Baseball Data

The desire to account for variability in Π_{ij}, including potentially positive associations among outcomes in the count vector, leads us to the following hierarchical structure of the data model for player i season j:

$$W_{ij}|\Pi_{ij} \sim Multi_K\left(N, \Pi_{ij}\right)$$

$$\Pi_{ij} \sim f\left(\Pi_{ij}\right)$$

In order to determine a reasonable distribution $f\left(\Pi_{ij}\right)$, we turn to compositional data analysis. Compositional data consists of non-negative vectors whose elements sum to one; in essence, probability vectors. The modern discussion of modeling compositional data begins with John Aitchison, who in the 1980s developed many of the tools used today. In his original works Aitchison (1982, 1986) discusses two main classes of models for modeling compositional data, the Dirichlet and logistic-normal classes. A key difference arises in the modeling space. Dirichlet models for compositional data conduct analysis directly on the simplex space $\mathbb{S}^K$ where, due to the sum to one constraint, a compositional vector exists. Logistic-normal models propose transforming the compositional vector via a log-transformation, converting from the simplex to $\mathbb{R}^{K-1}$ Euclidean space. Aitchison leans towards the logistic-normal approach (Aitchison, 1982), due to the strong internal independence structure of the Dirichlet. We discuss this structural limitation in the next section.

Keeping in mind the hierarchical structure of our data model, a strictly multinomial model for the counts does not allow for positive correlations, nor does the multinomial-Dirichlet distribution for counts that arises from modeling $f(\Pi_{ij}) \sim Dirichlet$. The distribution of counts produced from modeling a multinomial probability vector with the logistic-normal distribution, $f(\Pi_{ij}) \sim LogitNormal$, does allow for positive correlations.

As early as 1966 (Good, 1966) the Dirichlet was being considered as a distribution for compositional data. Connor and Mosimann (1969) developed a generalization of the Dirichlet distribution in biological applications of modeling bone composition in rats and scute growth in turtles. Later, Campbell and Mosimann (1987) more carefully examined the process of modeling of compositional data with the Dirichlet via the inclusion of covariates. Multinomial-Dirichlet regression has also been used in the context of microbiome data analysis (Chen and Li, 2013). In a Bayesian framework, the Dirichlet distribution has been popular as a conjugate prior to the multinomial distribution (Good, 1966; Wong, 1998; Ocampo et al., 2019). In terms of multinomial data, the multinomial-Dirichlet distribution is a marginal compound distribution for count data, which results from modeling the probability vector of the multinomial via a Dirichlet distribution. A related distribution is the nested Dirichlet, developed by Null (2009), which is a generalized version of the Generalized Dirichlet distribution developed by Connor and Mosimann. The nested Dirichlet was developed specifically to account for some limitations of the regular Dirichlet. Both the Dirichlet and the nested Dirichlet are discussed in more detail in Section 1.4.

Our work focuses on the logistic-normal distribution for modeling the compositional vector that is the underlying probability of a player's seasonal outcome vector. Use of the logistic-normal to describe compositional data originates with Aitchison (1982) and subsequent hierarchical modeling approaches have followed. Albert and Chib (1993) discuss applying a Gibbs sampling approach to the multinomial probit model

introduced by Aitchison and Bennet (1970). Logistic-normal regression has been utlized in such fields as household purchasing behavior (Allenby and Lenk, 1994), species composition (Billheimer et al., 2001), chemical markers in plants (Brewer et al., 2005), and is very popular in modeling in the context of microbiome data analysis (Xia et al., 2013; Grantham et al., 2017; Ren et al., 2017; Li et al., 2018). Similar to the multinomial-Dirichlet, the marginal distribution for multinomial count vectors where the probabilities are modeled via the logistic-normal is the multinomial logistic-normal distribution. Our main result concerns the multinomial logistic-normal being the most appropriate model for baseball data.

The major deficiency in existing logistic-normal models for compositional data involves hardly any emphasis on developing models focused on prediction, especially in the context of longitudinal data. The flexibility in the logistic-normal to capture potentially positive correlations between components of a compositional vector has been discussed since Aitchison's initial work, and we examine that in more detail, specifically in the context of baseball outcomes, in Section 1.5. When data take on a longitudinal structure, however, it is also important to account for correlation between components over time. In a species composition framework, it could be possible that the proportion of one species at time t might be correlated with the proportion of another species at time $t+1$. In baseball terms, home runs and strikeouts may not only be correlated within a player-season, but also across a player's career.

In baseball specifically, several outcomes are significantly positively correlated, leading to the logistic-normal as a more natural choice for modeling the underlying composition than the Dirichlet. The next section describes the Dirichlet, nested Dirichlet, and logistic-normal distributions as possible distributions for Π_{ij}, including discussion on the implied correlation among counts corresponding to each distribution. The succeeding section uses a real data example to compare the three distributions in their

ability to capture positive correlation between components of the composition (the multinomial probability vector).

1.4 Hierarchical Models for Baseball Count Data

In this section we discuss three candidate hierarchical models for describing baseball outcome data. Over the course a season, an individual player accrues plate appearances the outcomes of which can be treated as draws from a multinomial distribution given a probability vector representing that player's ability. We express this as:

$$w_{ij1}, ..., w_{ijK} \sim Multi_K\left(PA_{ij}, \Pi_{ij}\right)$$

where PA_{ij} is the number of plate appearances for player i's season j, and K represents the total number of outcomes. Because we believe Π_{ij} to not be fixed, we introduce the hierarchical structure with Π_{ij} described by some distribution.

The main focus in this section deals with how different distributions for the underlying probability vector impact the distribution of the counts. For simplicity, we consider the player-seasons to be independent and we only briefly discuss the framework for including covariates.

1.4.1 Multinomial-Dirichlet Distribution

Modeling the underlying probability vector Π according to the Dirichlet allows for additional variability in W over that of the strictly multinomial distribution.

Definition. A probability vector following the Dirichlet distribution with parameter vector α, $\Pi \sim Dir(\alpha_1, ..., \alpha_K)$, implies the following joint probability distribution function over the probability vector:

$$f\left(\pi_1, ..., \pi_K | \alpha_1, ..., \alpha_K\right) = \prod_{k=1}^{K} \frac{\pi_k^{\alpha_k - 1}}{B(\alpha)} \tag{1.6}$$

where K is the number of categories, $\alpha = (\alpha_1, ..., \alpha_K)$, $B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$, and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

It is sometimes useful (e.g., in simulation) to consider the Dirichlet in terms of K independently distributed Gamma distributions. Specifically, if

$$A_k \sim Gamma\left(\alpha_k, \theta\right)$$

and

$$A_0 = \sum_{k=1}^{K} A_k$$

then

$$\Pi = \left(\frac{A_1}{A_0}, ..., \frac{A_K}{A_0}\right) \sim Dirichlet\left(\alpha_1, ..., \alpha_K\right)$$

The Dirichlet has the following properties in terms of expected value, variance and covariance (when $k \neq k'$) of the probability vector terms:

$$E\left[\pi_k\right] = \frac{\alpha_k}{\alpha_0} \tag{1.7}$$

$$Var\left(\pi_k\right) = \frac{\alpha_k\left(\alpha_0 - \alpha_k\right)}{\alpha_0^2\left(\alpha_0 + 1\right)} \tag{1.8}$$

$$Cov\left(\pi_k, \pi_{k'}\right) = \frac{-\alpha_k \alpha_{k'}}{\alpha_0^2\left(\alpha_0 + 1\right)} \tag{1.9}$$

When considering $W \sim Multi(N, \Pi)$ and Π follows the Dirichlet distribution, the resulting distribution of the counts is the multinomial-Dirichlet (Mosimann, 1962).

Definition. For a vector of counts, W, distributed according to the multinomial distribution, with probability vector Π distributed according to the Dirichlet distribution, implies the following probability distribution function:

$$f\left(w_1, ..., w_K | \alpha_1, ..., \alpha_K\right) = \frac{\left(N!\right)\Gamma\left(\alpha_0\right)}{\Gamma\left(N+\alpha_0\right)} \prod_{k=1}^{K} \frac{\Gamma\left(w_k+\alpha_k\right)}{\left(w_k!\right)\Gamma\left(\alpha_k\right)} \tag{1.10}$$

where K, α, and α_0 are defined as in the Dirichlet, and the following expected values, variances and covariances:

$$E\left[w_k\right] = N\left(\frac{\alpha_k}{\alpha_0}\right) \tag{1.11}$$

$$Var\left(w_k\right) = N\left(\frac{\alpha_k}{\alpha_0}\right)\left(1-\frac{\alpha_k}{\alpha_0}\right)\left(\frac{N+\alpha_0}{1+\alpha_0}\right) \tag{1.12}$$

$$Cov\left(w_k, w_{k'}\right) = -N\left(\frac{\alpha_k \alpha_{k'}}{\alpha_0^2}\right)\left(\frac{N+\alpha_0}{1+\alpha_0}\right) \tag{1.13}$$

The most important of these results for our purposes is that the correlation between any two categories under the multinomial-Dirichlet distribution is necessarily negative.

$$Cor\left(w_k, w_{k'}\right) = -\sqrt{\frac{\alpha_k \alpha_{k'}}{\left(\alpha_0-\alpha_k\right)\left(\alpha_0-\alpha_{k'}\right)}} \tag{1.14}$$

In fact, we can see by comparing the multinomial-Dirichlet correlation (Equation 1.14) to the multinomial correlation (Equation 1.5), that the correlation is the same. Modeling the probabilities according to a Dirichlet provides for increased variation but does not alter the correlation structure.

Essentially, this negative correlation is entirely due to the sum to one constraint of the probability vector. A probability vector of length K which follows a Dirichlet distribution is equivalent to a composition formed from K independent gamma distributed components. As Aitchison (1982) describes, this is the strong independence assumption the Dirichlet has for all outcomes. The sum to one constraint combined

with the independence structure of the gamma components produces negative correlations among all components. Regardless of the true relationship between any two categories in the data, the multinomial-Dirichlet model cannot account for positive correlations between outcomes.

Inclusion of covariates for modeling purposes can be approached by directly modeling the Dirichlet parameters with fixed covariates via a log-link, which is equivalent to modeling the log-odds in the multinomial logit model described in Section 1.3.1.

$$log\,(\alpha_k) = X\beta_k \text{ for } k = 1, ..., K$$

The coefficients under the multinomial-Dirichlet covariate model have a slightly more natural interpretation than under the nested Dirichlet or logistic-normal models. However, the resulting maximisation of the likelihood defined by the inclusion of covariates does not have a closed form (Campbell and Mosimann, 1987).

Campbell and Mosimann (1987) suggest that covariate models of the Dirichlet may provide more flexibility in describing covariances between outcomes. As an alternative within the Dirichlet family, apart from the nested Dirichlet described in the next section, Campbell and Mosimann also suggested finite mixtures of Dirichlet distributions can account for any pairwise covariance, but noted that finite mixtures of Dirichlet distributions are not identifiable.

The Dirichlet, although a conjugate prior for the multinomial distribution, is a very restrictive distribution. In a Bayesian context, Albert (2016) showed that placing independent beta priors on each component probability worked better in predicting batting average than using a multinomial-Dirichlet model, noting the same lack of flexibility of a Dirichlet prior that motivated Null (2009) to use the nested Dirichlet.

1.4.2 Multinomial Nested Dirichlet Distribution

Null (2009) proposes the nested Dirichlet as an alternative model for the probabilities in multinomial data. While the nested Dirichlet retains the desirable properties of the regular Dirichlet, it also improves the flexibility of modeling relationships between the outcomes by introducing a nesting structure which adds an additional n variables under which original outcomes are nested. It thus adds an additional n parameters to the K parameters already present in the simple Dirichlet.

Null suggests the simplest way to visualize the difference between the nested and regular Dirichlet was via a nesting tree. In Figure 1.1, the right tree represents the nesting for a regular Dirichlet distribution with $K = 3$, while the left adds an additional $n = 1$ nesting variable under which two of the original outcomes are nested. This structure implies that the pair (π_2, π_3) have a Dirichlet distribution conditional on π_4, and the pair (π_1, π_4) have an unconditional Dirichlet distribution.

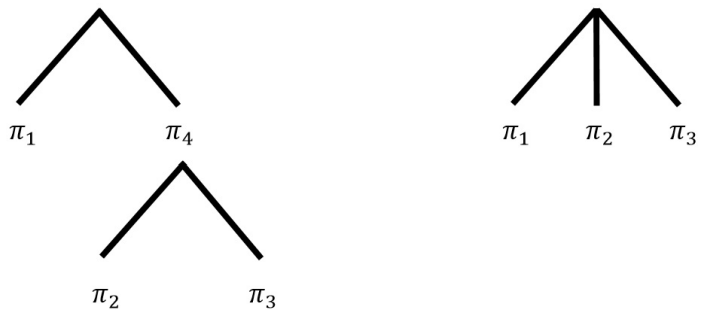

Figure 1.1. Comparing nested and regular Dirichlet distributions with categories $K = 3$

Based on the ordering of the original outcomes in the first tree, there are a total of 3 possible nesting structures for the three category case, not including the regular Dirichlet which could be considered a subset of the nested Dirichlet. The left tree in Figure 1.1 is also a representation of the Generalized Dirichlet. The Generalized

Dirichlet is a subset of the nested Dirichlet where there are two variables in each nesting and at least one is an original outcome (Connor and Mosimann, 1969). While in the three category case, the Generalized and nested Dirichlet are the same, as more categories are included more nesting structures become possible in the nested Dirichlet. This provides the lower bound for the number of possible nestings, $\frac{K!}{2}$ for $K \geq 3$.

Definition. As defined by Null, a nested Dirichlet distribution implies the following joint probability distribution function over the probability vector:

$$f\left(\pi_1, ..., \pi_K \middle| \, \alpha_1, ..., \alpha_K, \alpha_{K+1}, ..., \alpha_{K+n}\right) = \frac{\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \prod_{k'=1}^{n} \pi_{K+k'}^{\alpha_{K+k'} - \alpha_{0k'}}}{\prod_{k'=0}^{n} B\left(\alpha_{k'}^*\right)} \tag{1.15}$$

where $\alpha_{k'}^*$ is the vector of parameters for the k'th nesting, with α_0^* the vector of parameters for the unnested outcomes, and $\alpha_{0k'}$ is the sum of all parameters in vector $\alpha_{k'}^*$.

For the specific nesting structure of the nested Dirichlet defined by the Generalized Dirichlet, the covariance is straightforward to compute. Since we will be focusing on the three category case for our example, and the Generalized and nested Dirichlet share a common structure in that case, it suffices to describe general moment functions (Wong, 1998) for three categories in the structure of the left nesting in Figure 1.1.

$$E\left[\pi_1^{r_1} \pi_2^{r_2}\right] = \prod_{k=1}^{2} \frac{\Gamma(\alpha_{0k}) \Gamma\left(\alpha_k + r_k\right) \Gamma\left(\gamma_k + \delta_k\right)}{\Gamma\left(\alpha_k\right) \Gamma\left(\gamma_k\right) \Gamma\left(\alpha_{0k} + r_k + \delta_k\right)} \tag{1.16}$$

$$E\left[\pi_1^{r_1} \pi_3^{r_3}\right] = \prod_{k \in \{1,3\}} \frac{\Gamma\left(\alpha_{0k}\right) \Gamma\left(\alpha_k + r_k\right) \Gamma\left(\gamma_k + \delta_k\right)}{\Gamma\left(\alpha_k\right) \Gamma\left(\gamma_k\right) \Gamma\left(\alpha_{0k} + r_k + \delta_k\right)} \tag{1.17}$$

where $\gamma_1 = \alpha_4$, $\gamma_2 = \alpha_3$, $\gamma_3 = \alpha_2$, and $\delta_k = \sum_k r_k$. And, for the joint expectation of the two nested outcomes:

$$E\left[\pi_1^{r_1} \pi_2^{r_2} \pi_3^{r_3}\right] = \prod_{k=1}^{3} \frac{\Gamma\left(\alpha_{0k}\right) \Gamma\left(\alpha_k + r_k\right) \Gamma\left(\gamma_k + \delta_k\right)}{\Gamma\left(\alpha_k\right) \Gamma\left(\gamma_k\right) \Gamma\left(\alpha_{0k} + r_k + \delta_k\right)} \tag{1.18}$$

where $r_1 = 0$, $r_2 = r_3 = 1$, $\delta_1 = 2$, $\delta_2 = \delta_3 = 0$. Extending to more categories, and nestings that are not equivalent to the nesting structure of the Generalized Dirichlet, complicates moment derivation. The general moment functions defined in Equations 1.16, 1.17, and 1.18 are enough to calculate expected values, variances, covariances and correlations among three outcomes in a three category nested Dirichlet framework (Connor and Mosimann, 1969; Wong, 1998) based on Figure 1.1.

For $k = 1, 2, 3$, $\gamma_1 = \alpha_4$, $\gamma_2 = \alpha_3$, and $\gamma_3 = \alpha_2$:

$$E[\pi_k] = \frac{\alpha_k}{\alpha_k + \gamma_k} \prod_{l=1}^{k-1} \frac{\gamma_l}{\alpha_l + \gamma_l}$$

$$Var(\pi_k) = E[\pi_k] \left(\frac{\alpha_k + 1}{\alpha_k + \gamma_k + 1} \prod_{l=1}^{k-1} \frac{\gamma_l + 1}{\alpha_l + \gamma_l + 1} - E[\pi_k] \right)$$

$$Cov(\pi_k, \pi_{k'}) = E[\pi_k] \left(\frac{\alpha_k}{\alpha_k + \gamma_k + 1} \prod_{l=1}^{k-1} \frac{\gamma_l + 1}{\alpha_l + \gamma_l} - E[\pi_k] \right)$$

The resulting marginal multinomial nested Dirichlet distribution of the counts will, depending on the nesting structure chosen, allow for positive correlations between outcomes. One of the key features that Null (2009) espouses as a benefit of the nested Dirichlet distribution is its conjugacy to the multinomial distribution. This is a necessarily Bayesian result, implying that the resulting posterior distribution of counts given the nested Dirichlet parameters is also of nested Dirichlet form. Null (2009) does not provide formulation of this multinomial nested Dirichlet distribution. Since the Generalized Dirichlet is a subset of the nested Dirichlet, and that distribution is well established (Connor and Mosimann, 1969; Wong, 1998; Bouguila, 2008), we can define the multinomial Generalized Dirichlet distribution to give an idea of what the multinomial nested Dirichlet looks like.

Definition. For a vector of counts, W, distributed according to the multinomial distribution, with probability vector Π distributed according to the Generalized Dirichlet distribution, implies the following probability distribution function:

$$f\left(w_1,...,w_K|\alpha_1,...,\alpha_{K-1},\gamma_1,...,\gamma_{K-1}\right) = \frac{N!}{w_1!\cdots w_k!}\times \prod_{k=1}^{K-1}\frac{\Gamma\left(\alpha_k+w_k\right)}{\Gamma\left(\alpha_k\right)}\frac{\Gamma\left(\gamma_k+z_{k+1}\right)}{\Gamma\left(\gamma_k\right)}\frac{\Gamma\left(\alpha_k+\gamma_k\right)}{\Gamma\left(\alpha_k+\gamma_k+z_k\right)}$$

where in the context of the nested Dirichlet defined before, the values of γ are values of α dependent upon the nesting structure, $N=\sum_{k=1}^{K} w_k$, and $z_k=\sum_{l=k}^{K} w_l$.

Covariate inclusion is less straightforward in the nested Dirichlet than in the Dirichlet. Null (2009) addressed modeling of the probabilities with a covariate by using a least squares approach, directly relating the covariate to the $K+n$ vector of probabilities defined by the nesting structure, $\tilde{\pi}$:

$$\tilde{\pi}_k = X\beta_k + \varepsilon \text{ for } k = 1,...,K+n$$

One of the complications Null identified was that the model allowed for probabilities outside the viable range of $[0,1]$, indicating that the conjugacy of the nested Dirichlet is lost via the inclusion of covariates. Null also noted that, under the proposed covariate model, outcomes nested together will have opposite effects; in other words a one unit increase in the covariate will impact outcomes nested together in an exactly opposite way (Null, 2009).

1.4.3 Multinomial Logistic-Normal Distribution

The final data model we consider is the multinomial logistic-normal distribution. The logistic-normal family of distributions provides a considerably more flexible covariance structure due to the increase in number of unknown parameters, inherited from

the multivariate logistic-normal distribution. This, like certain nestings of the nested Dirichlet, is capable of preserving positive correlations between outcomes. Applying the inverse additive logistic transformation to the probabilities, we move to the $\mathbb{R}^{K-1}$ space instead of the simplex $\mathbb{S}^K$. This mapping from the simplex to real space is a one-to-one mapping, thus there is not too much interpretability lost. Most importantly there is no longer the strong independence between all outcomes, and the unconstrained covariance structure of the multivariate logistic-normal allows for modeling of positive pairwise correlations between outcomes. In the context of the logistic-normal, this inverse additive logistic transformation of the underlying multinomial probabilities produces log-odds which follow a multivariate normal distribution.

Definition. A probability vector following the logistic-normal distribution, $\Pi \sim LogitNormal_{K-1}(\mu, \Sigma)$, may be transformed into log-odds which follow the multivariate normal distribution:

$$log\left(\frac{\pi_1}{\pi_K}\right), ..., log\left(\frac{\pi_{K-1}}{\pi_K}\right) \sim N_{K-1}(\mu, \Sigma)$$

where if we call Y the vector of log-odds which we expand to K dimensions by defining the K^{th} coordinate to be 1, then we redefine the probability vector

$$\Pi = \left(\frac{exp(y_1)}{1 + \sum_{k=1}^{K-1} exp(y_k)}, ..., \frac{exp(y_{K-1})}{1 + \sum_{k=1}^{K-1} exp(y_k)}, \frac{1}{1 + \sum_{k=1}^{K-1} exp(y_k)}\right) \tag{1.19}$$

It suffices to describe the joint probability distribution function of the probabilities in terms of the multivariate normally distributed log-odds, $Y \sim N_{K-1}(\mu, \Sigma)$:

$$f(y_1, ..., y_{K-1}|\mu, \Sigma) = \frac{exp\left(-\frac{1}{2}(Y-\mu)^T \Sigma^{-1}(Y-\mu)\right)}{\sqrt{(2\pi)^{K-1}|\Sigma|}} \tag{1.20}$$

The unconstrained covariance matrix allows for a much more flexible pairwise correlation structure between outcomes, including allowing for positive correlation. The

logistic-normal distribution's ability to model positive correlation makes it an attractive alternative to the Dirichlet for certain applications. However, there is no analytic solution for the mean and covariance matrix of the logistic normal distribution; the integral expressions for moments of all positive orders are not reducible to any simple form. Moments may instead be defined in terms of the logarithms of the ratio of the logistic-normal variates (Aitchison and Shen, 1980), for categories $k = 1, ..., K$:

$$E\left[log\left(\frac{\pi_k}{\pi_{k'}}\right)\right] = \mu_k - \mu_{k'}$$

$$cov\left[log\left(\frac{\pi_k}{\pi_{k''}}, \frac{\pi_{k'}}{\pi_{k''}}\right)\right] = \sigma_{kk'} + \sigma_{k'k''} - \sigma_{kk''} - \sigma_{k''k'}$$

where σ_{jk} denotes the $(j,k)^{th}$ element of Σ and with $\mu_K = 0$ and $\sigma_{jK} = 0$. Though this formulation does not naturally describe the moments in terms of the probabilities. For this, we may use simulation to show that for a fixed value of the mean, there is a direct mapping of the logistic-normal correlation to the compositional correlations.

Consider the $K = 3$ case, where the resulting log-odds vector, y, is bivariate normal:

$$Y = \left(log\left(\frac{\pi_1}{\pi_3}\right), log\left(\frac{\pi_2}{\pi_3}\right)\right) \sim N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) \tag{1.21}$$

For fixed $\mu = (\mu_1, \mu_2)$ and σ_1, σ_2, we can vary ρ from -1 to 1, simulate a large number of bivariate normal random variables at each value of ρ, transform them into compositions via the additive logistic transformation, and assess their correlation. While the values of the mean vector μ and the standard deviations σ_1 and σ_2 will affect the resulting correlations, there will always be a direct mapping from the logistic-normal correlation to the compositional correlations.

More formally, our simulation proceeds as follows:

1. For some choice of μ, σ_1 and σ_2, and for each of $\rho = \{\rho_1, \rho_2, ..., \rho_{999}, \rho_{1000}\} = \{-1, -.998, ..., .998, 1\}$, simulate

$$Y_{ij} \sim N_2(\mu, \Sigma_i)$$

where $\Sigma_i = \begin{pmatrix} \sigma_1^2 & \rho_i\sigma_1\sigma_2 \\ \rho_i\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ and $j = 1, ..., 1000$. This generates 1000 bivariate normal samples of each of 1000 different correlations along the sequence from -1 to 1.

2. For each correlation i and sample j, transform according to equation (1.19):

$$\pi_{ij} = \left(\frac{e^{y_{ij1}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{e^{y_{ij2}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{1}{1 + e^{y_{ij1}} + e^{y_{ij2}}} \right)$$

3. Plot ρ versus each pairwise correlation of π and fit a smooth curve

For the two plots (Figures 1.2 and 1.3) presented we set the standard deviations equal to 1. The different figures correspond to different choices of the μ vector. While the mean and variance will have an impact on the shape, the figures show that the logistic-normal is capable of modeling positive correlation between compositional outcomes.

Definition. For a vector of counts, W, conditionally multinomially distributed with probability vector Π distributed according to the logistic-normal distribution, implies the following probability distribution function:

$$f(w_1, ..., w_K | y_1, ..., y_{K-1}, \mu, \Sigma) \propto \prod_{k=1}^{K} \left[\frac{exp(y_k)}{1 + \sum_{l=1}^{K} exp(y_l)} \right]^{w_k} \times exp\left[-\frac{1}{2} (Y - \mu)^T \Sigma (Y - \mu) \right]$$

Since the likelihood above has no closed form, there is once again no analytic solution to the moments of the multinomial logistic-normal distribution. A similar process to

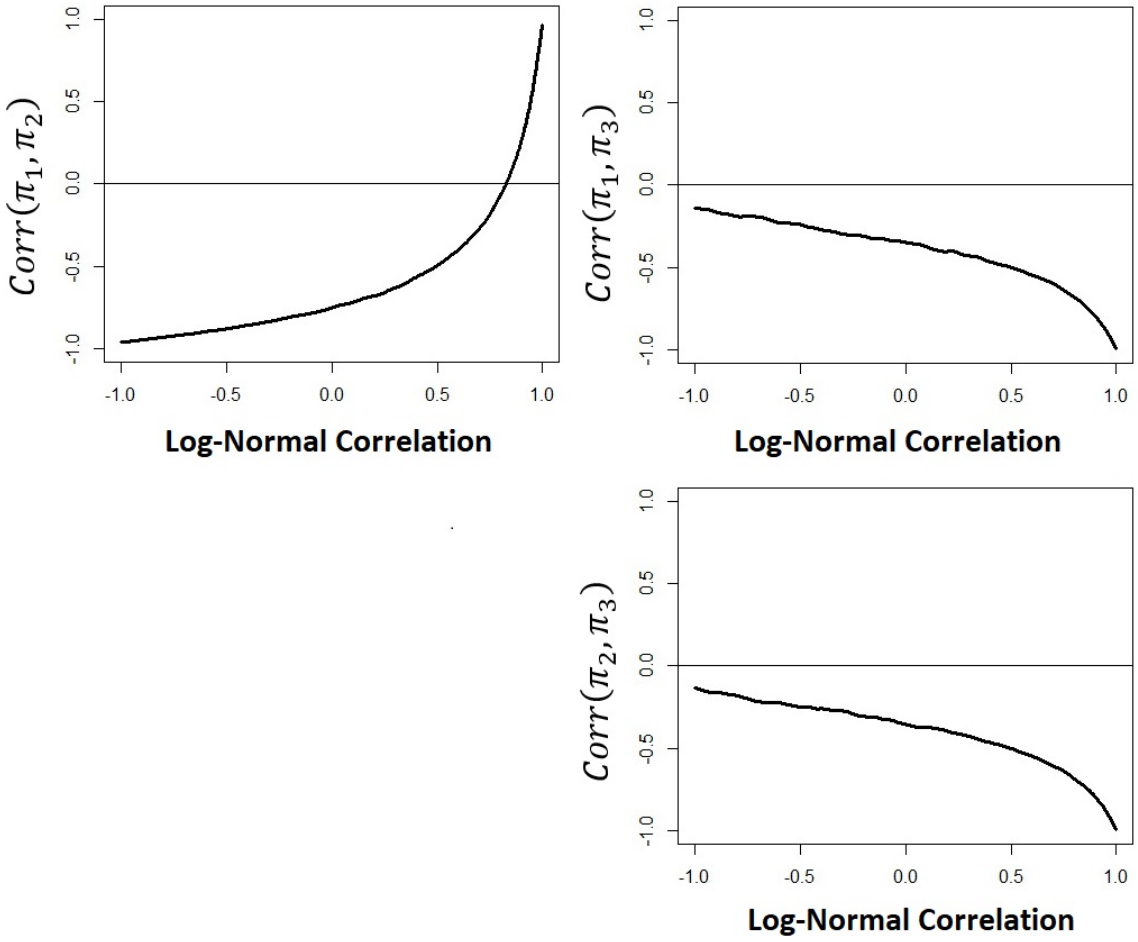

Figure 1.2. Resulting compositional correlations from simulating log-odds from bivariate normal with mean vector $(0, 0)$ and standard deviations $\sigma_1 = \sigma_2 = 1$

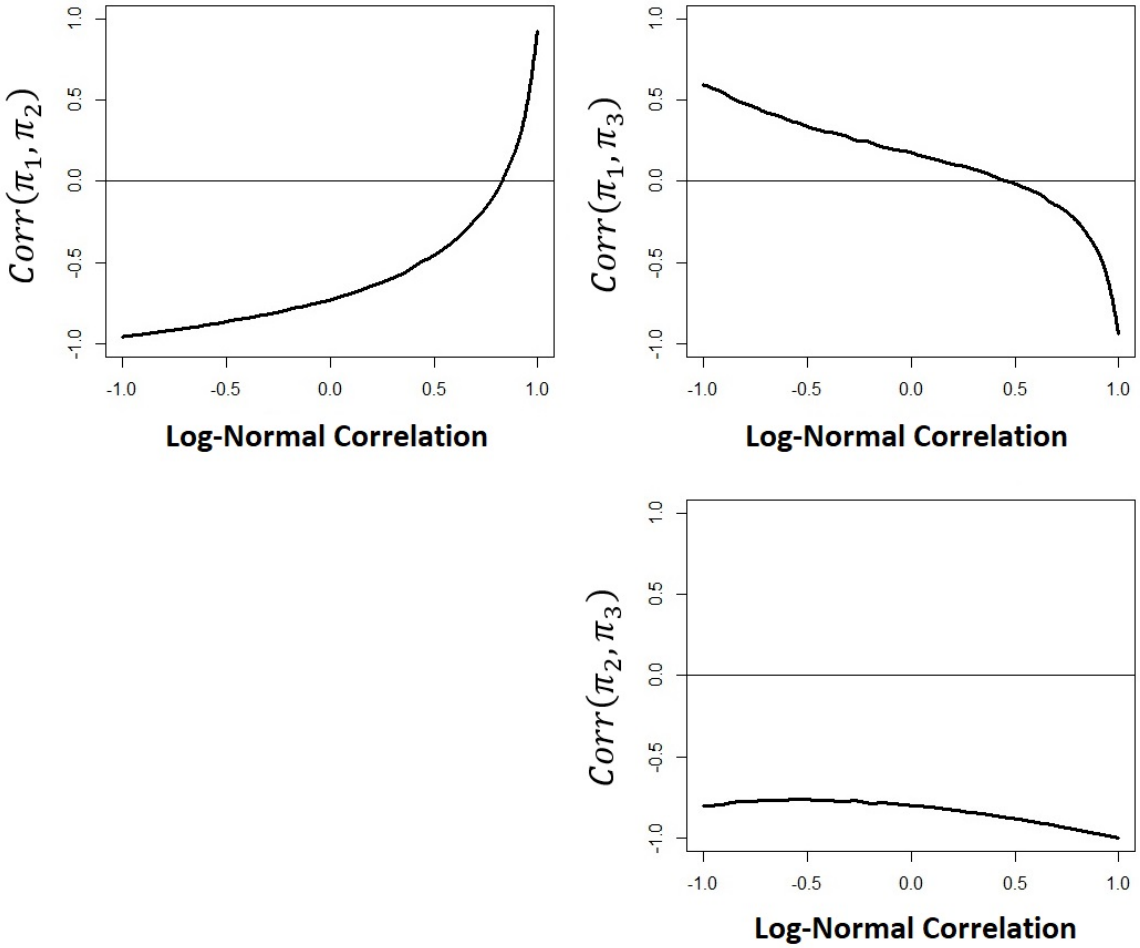

Figure 1.3. Resulting compositional correlations from simulating log-odds from bivariate normal with mean vector $(-1, 1)$ and standard deviations $\sigma_1 = \sigma_2 = 1$

the one used for the logistic-normal (Figures 1.2 and 1.3) may be used to simulate an implied correlation structure for the distribution of counts which follow a multinomial logistic-normal distribution. The multinomial aspect of the model will slightly diminish correlation from the logistic-normal distribution for the probabilities, but given a large number of plate appearances, this difference is minimal.

Finally, inclusion of covariates follows the traditional multivariate regression setting, where the covariates are regressed on the multivariate normal log-odds.

$$y_k = X\beta_k + \varepsilon \text{ for } k = 1, ..., K-1$$

or, formatted another way,

$$Y \sim N_{K-1}(X\beta, \Sigma)$$

where X is a $n \times p$ matrix of covariates, β is an $p \times (K-1)$ matrix of coefficients, and Σ is the $(K-1) \times (K-1)$ covariance matrix. This is the equivalent covariate model to the multinomial logit covariate model described in Section 1.3.1 and the multinomial-Dirichlet covariate model described in Section 1.4.1.

Just as in the Dirichlet covariate model, since there is no closed form of the resulting likelihood, we must use some approximation method to optimize the likelihood. The unconstrained covariance matrix allows for better modeling of positive correlations between outcomes than the multinomial-Dirichlet model. In comparison to the multinomial nested Dirichlet, there is no need to impose a possibly suboptimal nesting structure which may struggle to describe relationships between outcomes, and there are subsequently no restrictions on the covariate effects, where in the nested Dirichlet there were opposite effects within pairs of nested outcomes.

1.5 Model Comparison Example in Baseball Player Performance

In this section we present a simplified example to illustrate the differences in ability of the three models described in Section 1.4 to recover correlations. We simplify the problem by ignoring the multinomial structure. In other words, we focus on modeling the proportion (or rates) of each category. We also do not consider covariates. Our argument here is that if we can show that the logistic-normal model for Π is preferred over the Dirichlet and nested Dirichlet in modeling correlations among the probabilities in this setting, we expect the multinomial logistic-normal model to be preferred when considering the counts.

Using Major League Baseball data consisting of all qualified batters from 1961-2013 (Friendly, 2015), we focus on the categories home runs (HR), strike outs (SO), and other. Those categories are chosen due to the common belief that home runs and strike outs are positively correlated. We restrict this analysis to qualified batters, (plate appearances or $N > 501$), to reduce the impact of differing plate appearances on the rates. Figure 1.4 and Table 1.1 show that home run and strikeout rates are positively correlated across player-seasons.

Of all 6437 qualified player-seasons in this data set, 95 resulted in no home runs. Since both Dirichlet and logistic-normal analysis take place with the log-link, and $log(0) = -\infty$, we can remove these player-seasons from the data set before analysis. The removal did not change any of the correlation coefficients of our three outcomes by more than a percent in either direction. There are alternatives to removal however, namely adding some negligible number ϵ to all zero counts, or "compressing" the data symmetrically via transformation (Smithson and Verkuilen, 2006). In data sets with a larger occurrence of zero counts, these can be employed to avoid too much loss of information. A summary of the difference in choices for handling zero counts is provided in Section 2.5.

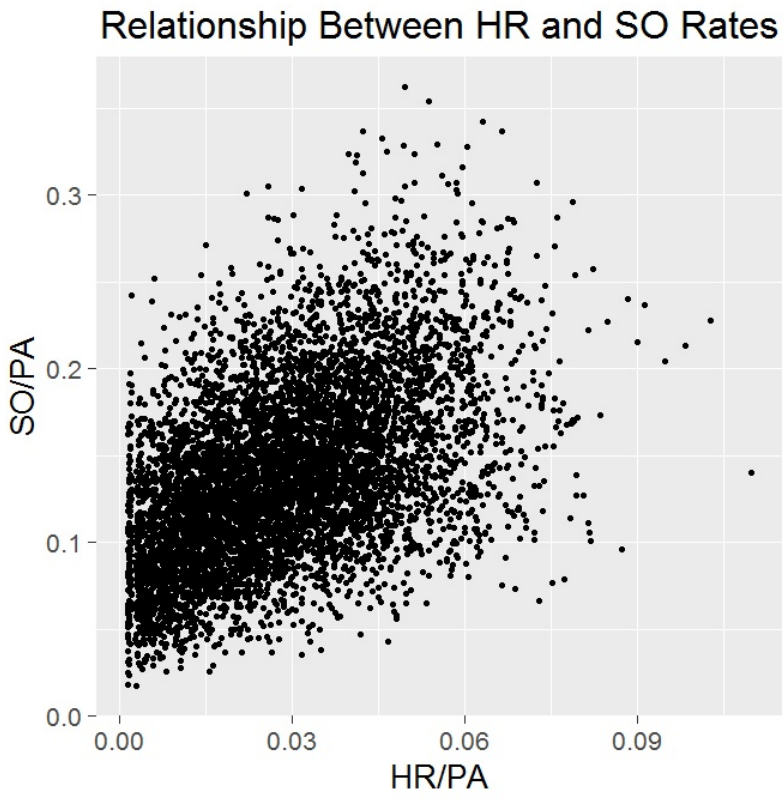

Figure 1.4. Positive Correlation between HR and SO rates

Table 1.1.
Observed correlation matrix for rates of three batting outcomes

outcome	**HR**	**SO**	**OTHER**
HR	1	0.49	−0.69
SO	0.49	1	−0.97
OTHER	−0.69	−0.97	1

1.5.1 Implied Correlation Coefficients and Model Fits

In R, there are readily available packages to calculate maximum likelihood estimates of the parameters for both the Dirichlet and multivariate normal distributions. For the nested Dirichlet, since each nesting is conditionally independent (Null, 2009), we can fit a Dirichlet distribution for each nesting to estimate the respective parameters. In R the relevant functions are dirichlet.mle (Robitzsch, 2016) and mlest (Gross and Bates, 2012). It is also straightforward under each distribution to calculate the log-likelihood under the maximum likelihood estimates of the model parameters, and subsequently the Akaike information criterion (AIC) and Bayesian information criterion (BIC) in order to compare the model fits.

Dirichlet

Fitting the Major League Baseball data of the three outcome rates with the Dirichlet distribution results in three α parameters, one for each of the categories. Table 1.2 gives the maximum likelihood estimates for the Dirichlet distribution fit using the dirichlet.mle function. It is straightforward to calculate the implied correlations (Table 1.3) between the three outcome probabilities using Equation 1.14.

Table 1.2.
Maximum likelihood estimates for Dirichlet distribution fit to Major League Baseball data of three categories

estimate	**value**
α_{HR}	1.6595
α_{SO}	7.4384
α_{OTHER}	44.5838

Table 1.3.
Implied correlation matrix for rates of three batting outcomes under Dirichlet model

outcome	**HR**	**SO**	**OTHER**
HR	1	−0.07	−0.40
SO	−0.07	1	−0.89
OTHER	−0.40	−0.89	1

When we compare the implied Dirichlet correlation coefficients to the observed coefficients from the data (Table 1.1), we see the Dirichlet is woefully inadequate at capturing the relationship between home run and strikeout rates, our two positively correlated variables.

Nested Dirichlet

For the nested Dirichlet, since we are dealing with only three categories, there are only $\frac{3!}{2} = 3$ possible nesting structures. The three possible nestings include nesting π_{HR} and π_{SO} together, π_{HR} and π_{OTHER}, or π_{SO} and π_{OTHER} together. For sake of brevity we refer to each nesting structure by its nested pair: $\langle \pi_{HR}, \pi_{SO} \rangle$, $\langle \pi_{HR}, \pi_{OTHER} \rangle$, $\langle \pi_{SO}, \pi_{OTHER} \rangle$ respectively. All nestings are also representations of the Generalized Dirichlet and, as Connor and Mosimann (1969) note in their original paper, the unnested probability will always be negatively correlated with any other nested probability. This property makes the choice of nesting in this simple case obvious, though we will present results for all three for comparisons sake. Since we know home runs and strikeouts are positively correlated, the most appropriate nesting pairs those and leaves the other category unnested; $\langle \pi_{HR}, \pi_{SO} \rangle$.

It is important to note that the choice becomes immediately less obvious as soon as a single new outcome is added, and becomes progressively less feasible to intuit as the dimension increases. This is why Null (2009) imposes constraints on nesting selection, and even then admits derivation of a new algorithm to search through nestings would be advisable.

For each of our three nestings, maximum likelihood estimation proceeds as follows:

1. Fit a Dirichlet distribution on the unnested outcome probability, π_1, and the sum of the nested probabilities, $\pi_4 = \pi_2 + \pi_3$, this estimates nested Dirichlet parameters α_1 and α_4

2. Fit a Dirichlet distribution on the nested outcome probabilities, π_2 and π_3, this estimates nested Dirichlet parameters α_2 and α_3

Note that fitting a Dirichlet distribution to the two category case is equivalent to fitting a Beta distribution. The Generalized Dirichlet is in essence a collection of nestings conditionally distributed as Beta distributions. For more categories, some nestings in the nested Dirichlet may have more than two outcomes, thus we say we fit a Dirichlet, even in the three category case where we are really fitting two conditionally independent Beta distributions.

The resulting maximum likelihood estimates for the three trees are presented in Table 1.4, while the implied correlation matrices, calculated by using Equations 1.16 to 1.18 with each set of estimates from the three nesting structures, $\langle \pi_{HR}, \pi_{SO} \rangle$, $\langle \pi_{HR}, \pi_{OTHER} \rangle$, and $\langle \pi_{SO}, \pi_{OTHER} \rangle$ are presented in Table 1.5, Table 1.6, and Table 1.7 respectively.

Table 1.4.
Maximum likelihood estimates for nested Dirichlet distribution fit to Major League Baseball data of three categories, three different nesting structures

estimate	$\langle \pi_{HR}, \pi_{SO} \rangle$	$\langle \pi_{HR}, \pi_{OTHER} \rangle$	$\langle \pi_{SO}, \pi_{OTHER} \rangle$
α_{HR}	3.0942	1.9911	2.1479
α_{SO}	15.6546	6.2536	5.9466
α_{OTHER}	29.7058	56.6326	35.2388
α_4	6.0136	38.4926	72.9318

Table 1.5.
Implied correlation matrix for rates of three batting outcomes under nested Dirichlet model with $\langle \pi_{HR}, \pi_{SO} \rangle$ nesting structure

outcome	**HR**	**SO**	**OTHER**
HR	1	0.54	−0.56
SO	0.54	1	−0.96
OTHER	−0.56	−0.96	1

Table 1.6.
Implied correlation matrix for rates of three batting outcomes under nested Dirichlet model with $\langle \pi_{HR}, \pi_{OTHER} \rangle$ nesting structure

outcome	**HR**	**SO**	**OTHER**
HR	1	−0.09	0.08
SO	−0.09	1	−0.93
OTHER	0.08	−0.93	1

Table 1.7.
Implied correlation matrix for rates of three batting outcomes under nested Dirichlet model with $\langle \pi_{SO}, \pi_{OTHER} \rangle$ nesting structure

outcome	**HR**	**SO**	**OTHER**
HR	1	−0.05	−0.30
SO	−0.05	1	0.02
OTHER	−0.30	0.02	1

While we see here the nested Dirichlet's ability to account for positive correlations between outcomes, we also begin to see how the choice of nesting can significantly impact analysis. The only implied correlation matrix from our three nested Dirichlet fits which resembles the observed matrix is from our intuited most appropriate nesting structure, $\langle \pi_{HR}, \pi_{SO} \rangle$. This structure does a good job of capturing the relationships between the outcomes. However, the other two nesting structures are almost useless at describing the relationships between the outcomes, and subsequently produce worse model fits. As we expand the number of categories the choice of nesting structure quickly becomes a more pressing, and more difficult to handle, concern.

Logistic-Normal

Fitting the logistic-normal to our data set of three outcomes will result in $\frac{1}{2}(3-1)(3+2) = 5$ total parameters, a mean vector, μ of length 2 and a 2×2 covariance matrix, Σ. μ corresponds to the mean of the log-odds of the first two outcomes in relation to the last, in this case $OTHER$, the most prevalent outcome. To calculate maximum likelihood estimates for the logistic normal, we first transform the three outcomes into the two log-odds; $log\left(\frac{\pi_{HR}}{\pi_{OTHER}}\right)$ and $log\left(\frac{\pi_{SO}}{\pi_{OTHER}}\right)$ which for sake of conciseness we denote henceforth as $l_{H/O}$ and $l_{S/O}$. We then calculate bivariate normal maximum likelihood estimates for μ and Σ using the mlest function (Table 1.8).
Assessing the implied correlation between the three outcomes requires some simulation as there is no explicit formula for the covariance of the outcome probabilities based on the multivariate normal covariance matrix. Given maximum likelihood estimates of μ and Σ we proceed as follows to assess the implied correlation (Table 1.9), similar to the process outlined in Section 1.4.3:

1. Simulate 10000 bivariate normal random variables from $N_2\left(\hat{\mu}, \hat{\Sigma}\right)$ where $\hat{\mu}$ and $\hat{\Sigma}$ are the maximum likelihood estimates.

2. Convert the resulting log-odds to probabilities via the additive logistic transformation and find the correlation matrix.

We see here that the logistic-normal model works about as well as the optimal nesting structure in the nested Dirichlet model, as far as recovering the correlation coefficients between the three outcomes.

Via simulation, we have some evidence that it is also significantly more useful in describing these relationships than the Dirichlet model or the suboptimal nestings of the nested Dirichlet. When comparing AIC and BIC across the Dirichlet, best nested Dirichlet, and logistic-normal (Table 1.10), we can see that in this simplified setting that the logistic-normal is the best candidate model.

Table 1.8.
Maximum likelihood estimates for logistic-normal distribution fit to Major League Baseball data of three categories, on log-odds scale

estimate	**value**
$\mu_{l_{H/O}}$	-3.6111
$\mu_{l_{S/O}}$	-1.8516
$\sigma_{l_{H/O}}$	0.8557
$\sigma_{l_{S/O}}$	0.4622
ρ	0.5778

Table 1.9.
Implied approximate correlation matrix for rates of three batting outcomes under logistic-normal model

outcome	**HR**	**SO**	**OTHER**
HR	1	0.39	-0.70
SO	0.39	1	-0.93
OTHER	-0.70	-0.93	1

Table 1.10.
AIC and BIC for model fit of rates of three batting outcomes under Dirichlet, best nested Dirichlet, and logistic-normal models

measure	**DIR**	**NDIR**	**LN**
AIC	-53570.15	-55564.76	-63083.56
BIC	-53549.88	-55537.74	-63049.78

1.6 Discussion

This chapter began with a review of the relevant literature surrounding prediction in baseball, from which there are several takeaways. In academic research, there has long been a focus on predicting singular outcomes as opposed to the entire vector of seasonal outcomes. Most projection systems which provide results for the new season are developed with proprietary methodology. Very few of these projections systems also offer some form of uncertainty assessment. The chapter also discusses the value of modeling seasonal outcome counts jointly, as well as the value of providing uncertainty about the prediction.

A description of the multinomial distribution, and how it helps describe baseball data, follows. This is succeeded by a brief introduction of compositional data analysis and the history of approaches there-in. We introduce three data models based on a hierarchical structure for multinomial data with a varying probability vector, including the multinomial-Dirichlet, multinomial nested Dirichlet and multinomial logistic-normal distributions. The chapter discusses the need for allowing for positive correlation between outcomes in baseball data, and compares three different models based on the Dirichlet, nested Dirichlet, and logistic-normal distributions for the probability vector Π. Based on this preliminary work, we posit that a multinomial logistic-normal model is the most appropriate for joint modeling of baseball outcomes. The multinomial-Dirichlet model is not capable of describing positive correlations among outcomes and the multinomial nested Dirichlet model relies on a predetermined nesting structure the choice of which becomes non-trivial with more outcomes.

The main part of the discussion in this chapter also simplified the problem considerably, ignoring the longitudinal nature of baseball data and largely ignoring handling of covariate effects. In the following chapter, we introduce a Bayesian hierarchical multinomial-logistic normal model which includes player-specific covariates, and ran-

dom effects which serve to account for the longitudinal aspect of baseball player data and model correlation over time. This is followed by a description of a corresponding Gibbs sampling algorithm for fitting the model. It concludes by developing methods for prediction, uncertainty assessment about predictions, and assessment of model fit.

2. A MULTINOMIAL LOGISTIC-NORMAL MODEL FOR FORECASTING BASEBALL PERFORMANCE

2.1 Introduction

In Chapter 1, we described several hierarchical models that could be used to fit baseball player season outcome vectors and argued that the multinomial logisitic-normal model is the most appropriate for our data and prediction goals. It allows for the straightforward inclusion of covariates, and we illustrated its ability to account for positive and negative correlations between outcome categories.

Our data (e.g., multiple seasons per player), however, are also longitudinal, suggesting the possibility of positive correlations over time. To account for this, we move to a mixed effects framework. In this chapter, we describe this mixed-effects multinomial logistic-normal hierarchical model, our Bayesian framework for estimation and prediction, as well as some novel diagnostic tools for assessing model fit.

The inclusion of random effects not only accounts for the longitudinal nature of the data, but also makes sense practically. When considering a set of outcomes, or just a single outcome, we would expect a distribution of player abilities in the sense that certain players would be consistently above or below the average over the course of their careers. With the fixed effects model, there is player-season variability, but it assumes the same distribution for all players with the same set of covariates. In other words, all players with the same covariates have the same ability. Thus, each season a player is equally likely to have an above- or below-average season. With random effects, players remain above/below average in ability across seasons beyond what is captured in the covariates. This approach is similar in concept to the random effects

model built for home runs by Berry et al. (1999) which seeks to characterize the talent of each player with respect to the average home run rate.

The fixed-effects version of the multinomial logistic-normal model has been previously used to model counts of microbial taxa. Xia et al. (2013) focus on the problem of variable selection, proposing to use the multinomial logistic-normal model to relate covariates to bacterial composition. Because their main interest is in selecting the covariates (nutrients) that are associated with the composition, they propose a MCEM group ℓ_1 penalized likelihood estimation method. They report simulation study results indicating their variable selection method and model outperforms multinomial logistic and multinomial-Dirichlet models. Since the focus of the paper is on developing a variable selection technique for gut bacterial composition, there is no emphasis placed on prediction.

Mixed-effect versions have been used to model arthropod assemblages as well as microbiome data. Grantham et al. (2017) propose a Bayesian mixed-effects model called MIMIX (MIcrobiome MIXed model) for use in designed experiments where the goal is variable selection of covariates and inference of treatment effects on microbial taxa. A spike-and-slab prior is placed on the fixed effects, leading to the ability to perform a posterior test of the effect of treatment. Due to the high dimensionality of microbiome compositions, Bayesian factor analysis (Rowe, 2002) is used to lower the dimensional representation of the fixed and random effects. The application in the paper identified $K = 2662$ operational taxonomic units (OTUs, or the number of categories). The Bayesian factor analysis step reduces the dimension K, something that is not necessary in our situation.

The focus of MIMIX is on inference of treatment effects and not prediction. It was also developed with experimental data from a Randomized Complete Block design (Grantham et al., 2017). Random effects were introduced to model the dependence

structure of the taxa within blocks. The paper states MIMIX is not suited for handling data from longitudinal studies. However, one could consider the samples within blocks as repeated measures, and the corresponding covariance structure as building equal correlation over time. Still, MIMIX relies on balanced experimental design data and is focused on tests for treatment effects. Grantham et al. (2017) choose the multinomial logistic-normal, as Xia et al. (2013) did, to allow for a more flexible dependence structure among microbial taxa. It is not immediately clear whether MIMIX could be used in our application; whether MIMIX would allow for an unbalanced "design" where samples (seasons) of blocks (players) range from 2 to 27, and whether prediction under the MIMIX framework is straightforward.

The MIMIX mixed-effects framework is based on an earlier paper by Billheimer et al. (2001) where random effects are introduced to the multinomial logistic-normal model for analyzing experimental data concerning arthropod assemblages. In that paper the goal is to also estimate treatment effects in a designed experiment setting. The inclusion of random effects accommodates overdispersion. As opposed to MIMIX, Billheimer et al. (2001) specify the random effects covariance structure instead of estimating it in the model framework, a simplification which speeds up computation time. The overall covariance matrix describes spatial dependencies, and the fixed random covariance structure allows for overdispersion, but does not model dependencies. Billheimer et al. (2001) follows work from an earlier technical report (Billheimer and Guttorp, 1995) in which a multinomial logistic-normal state-space model is described for modeling the ecological condition of the Delaware Bay estuary. In the 1995 report, a conditional autoregressive Markov random field prior distribution is specified for the compositions, and the inclusion of covariate effects, within the context of their model, is demonstrated. This model more directly accounts for spacial dependencies in a multinomial logistic-normal setting.

In addition to a different focus (estimating treatment effects), there is also a great deal of difference in the nature of the data across these studies and our baseball data. Microbiome data typically involve a very large K, upwards of 2,700 taxa (Grantham et al., 2017), with many counts being null and the total count in the thousands. The arthropod data described in Billheimer et al. (2001) concerned itself with only three categories of arthropod, with observed counts ranging from a minimum of 7 to a maximum of 34. Baseball data are somewhere in between, with a relatively reasonable limit on the dimension of the outcome vector (usually $K \leq 14$) and count totals that are often zero and almost never over 500.

Also, in microbiome analysis, variable selection of covariates and dimension reduction of K are important facets of the methodology due to the immense number of covariates, which often exceed the sample size of the data. In our application, there is a reasonable number of covariates of interest. The arthropod study (Billheimer et al., 2001) discusses interpretation of fixed effects, but does not discuss estimation nor include them in the final model, though Billheimer's technical report on invertebrate compositions in Delaware Bay does (Billheimer and Guttorp, 1995).

None of these methodological papers concerning this hierarchical model have been concerned with the prediction of future counts, in any context. There has also been little done in developing measures for diagnostic tools for overall model fit. This chapter proceeds with a description of our mixed-effects multinomial logistic-normal hierarchical model, which accounts for the longitudinal nature of baseball performance data and the inclusion of covariates. The estimation via a Bayesian framework and subsequent development of predictions and model fit diagnostics follow.

2.2 A Mixed Effects Multinomial Logistic-Normal Model

In Chapter 1, we used a simplified baseball outcomes example to argue that the logistic-normal distribution is most appropriate for modeling the underlying probabilities due to its ability to allow for positive and negative correlations among outcome probabilities and to easily include covariates. The only other distribution that allowed positive correlations was the nested Dirichlet but the nesting structure is difficult to determine and the inclusion of covariates is not straightforward. We now propose the inclusion of random effects to correlate a player's outcome vectors over time. The resulting mixed-effects multinomial logistic-normal distribution of counts for player-seasons constitutes our model for analysis.

We begin our description of this model with some notational additions to the one introduced in Section 1.4.3. We still assume each observed K-dimensional count vector, for player $i \in \{1, ..., m\}$ and season $j \in \{1, ..., n_i\}$, follows a multinomial distribution, with parameters PA_{ij}, the known number of plate appearances, and Π_{ij}, the unknown probability vector.

What differentiates this model from the one described in Section 1.4 is how we model the probability vector, Π_{ij}. In the fixed model, this probability vector follows the logistic-normal distribution, meaning the log-odds, Y_{ij}, calculated by applying the additive log-ratio transformation function, $alr\left(\Pi_{ij}\right)$, follow a multivariate normal distribution:

$$Y_{ij} = \left(log\left(\frac{\pi_{ij1}}{\pi_{ijK}}\right), ..., log\left(\frac{\pi_{ij(K-1)}}{\pi_{ijK}}\right)\right) \sim N_{K-1}\left(X_{ij}\beta, \Sigma\right)$$

In the random effects version, we include a random effect vector ψ_i in the determination of Y_{ij}, specifically:

$$Y_{ij} \sim N_{K-1}\left(X_{ij}\beta + \psi_i, \Sigma\right) \tag{2.1}$$

Thus our mixed-effects multinomial logistic-normal model can be expressed

$$W_{ij} \sim Multi_K\left(PA_{ij}, \Pi_{ij}\right)$$

where

$$\Pi_{ij} = alr^{-1}(Y_{ij})$$

and

$$Y_{ij} = X_{ij}\beta + \psi_i + \varepsilon_{ij}$$

with

$$\psi_i \sim N_{K-1}\left(0, \Phi\right)$$

$$\varepsilon_{ij} \sim N_{K-1}\left(0, \Sigma\right)$$

The inclusion of the random effect allows us to account for a player's above- or below-average talent level over time, which has been done in a univariate setting for baseball (Berry et al., 1999).

Considering all sets of latent log-odds vectors from player i,

$$Y_i = [Y_{i1}, Y_{i2}, \cdots Y_{in_i}]^T$$

our model says they are multivariate normal:

$$Y_i|X_{ij}, \beta, \Sigma, \Phi \sim N_{(K-1)n_i}\left(\begin{bmatrix} \vec{X_{i1}\beta} \\ \vec{X_{i2}\beta} \\ \vdots \\ \vec{X_{in_i}\beta} \end{bmatrix}, \begin{bmatrix} \Phi+\Sigma & \Phi & \dots & \Phi \\ \Phi & \Phi+\Sigma & & \Phi \\ & \vdots & \ddots & \vdots \\ \Phi & \Phi & \dots & \Phi+\Sigma \end{bmatrix}\right) \tag{2.2}$$

This description better describes the assumed covariance structure among a player's seasonal outcome probabilities.

By explicitly defining a random effects design matrix, Z, multiple random effects could be included. Call q the number of random effects of interest and, for notational simplicity, move to denoting $K-1$ as d, the dimensionality of the latent log-odds, we

can then define a random effects vector of length dq for each player. The change in dimensionality, and some of the notation, follows:

$$Y_{ij} = X_{ij}\beta + Z_{ij}\psi_i + \varepsilon_{ij}$$

$$vec(\psi_i) \sim N_{dq}\left(0, \Phi\right)$$

Thus, given q player-specific random covariates, the log-odds latent variables have the multivariate normal distribution:

$$Y_{ij}|X_{ij}, \beta, \psi_i, \Sigma \sim N_d\left(X_{ij}\beta + Z_{ij}\psi_i, \Sigma\right)$$

And, unconditionally:

$$vec(Y_i)|X_{ij}, \beta, \Sigma \sim N_{dn_i}\left(vec(X_i\beta), Q_i^{-1}\right)$$

Where $Q_i^{-1} = \left(Z_i \otimes I_d\right)\Phi\left(Z_i \otimes I_d\right)^T + \left(I_{n_i} \otimes \Sigma\right)$.

2.2.1 Implied Correlation Structure in the Counts

One of the drawbacks of the multinomial logistic-normal model is the lack of direct translation from the covariance matrix on the log-odds scale to the count scale. It is possible, however, to estimate the implied correlation matrix on the simplex space, which represents the counts for a baseline set of covariates (defined by the fixed intercept), from the covariance matrix on the real space where the log-odds reside.

We can use simulation to estimate the correlation structure of the outcomes based on the true covariance matrices; $Corr(W^\Lambda)$, for $\Lambda \in \{\Sigma, \Phi\}$:

1. Simulate a large number of multivariate normal random variables:

$$Y^\Lambda \sim N_d(\beta_0, \Lambda)$$

2. Transform via inverse-additive-logistic-ratio transformation to probabilities:

$$\Pi^{\Lambda} = \left(\frac{e^{y_1^{\Lambda}}}{1 + \sum_{k=1}^{d} e^{y_k^{\Lambda}}}, \frac{e^{y_2^{\Lambda}}}{1 + \sum_{k=1}^{d} e^{y_k^{\Lambda}}}, ..., \frac{1}{1 + \sum_{k=1}^{d} e^{y_k^{\Lambda}}} \right)$$

3. Simulate counts from a multinomial distribution for exposure PA and each of the simulated probabilities:

$$W^{\Lambda} \sim Multi_{d+1}\left(PA, \Pi^{\Lambda}\right)$$

4. Compute the sample correlation matrix using the resulting counts.

While each of Σ and Φ are informative in their own right, of greater interest is likely the unconditional correlation structure implied for the $2(d+1)$ vector of counts within and between two seasons for player i: $(W_{ij}, W_{ij'})$, which can be found using the same process as above, with $\Lambda = ((I_d \otimes \Sigma) + (\mathbf{1}_d \otimes \Phi))$ (see Equation 2.2).

2.2.2 Bayesian Mixed-Effects Multinomial Logistic-Normal Model

We take a Bayesian inferential approach to the estimation of this model. Bayesian inference is a very popular and natural estimation approach for multilevel (hierarchical) models (Lindley and Smith, 1972; Smith, 1973; Kass and Steffey, 1989; Gelman, 2006). In the Bayesian setting, we can also simplify calculations by treating the latent logit values as additional unknowns. Finally, a Bayesian framework allows for us to easily incorporate prediction uncertainty using the model and joint posterior distribution of the parameters.

To establish our model in a Bayesian context, we choose priors for the parameters of the model. For the fixed effects (covariate) matrix, we consider an improper uniform prior:

$$p(\beta) \propto 1$$

For the two covariance matrices, we place noninformative inverse-Wishart priors, with fixed hyperparameters:

$$\Sigma \sim Wishart^{-1}\left(\nu_1 = d, \Lambda_1 = I_d + \mathbf{1}_d\mathbf{1}_d^T\right)$$

$$\Phi \sim Wishart^{-1}\left(\nu_2 = d, \Lambda_2 = I_d + \mathbf{1}_d\mathbf{1}_d^T\right)$$

2.3 Estimation via Metropolis within Gibbs Algorithm

Due to the intractability of the multinomial logistic-normal conditional posterior likelihood for the latent variables (see Section 1.4.3), a Metropolis or Metropolis-Hastings algorithm in the Gibbs step is used for updating the log-odds. Multiple proposal schemes for this algorithm are possible, and three options with their advantages and drawbacks are discussed. We begin this section by describing the full posterior likelihood of the multinomial logistic-normal model and proceed by outlining the Gibbs algorithm used for estimation.

2.3.1 Full Posterior Log-Likelihood

Due to the absence of a closed form, we incorporate a Gibbs algorithm to sample from the conditional posterior log-likelihood. However, given parameter values, it is at least relatively straightforward to calculate the full posterior likelihood of the multinomial logistic-normal model, at least proportional to the product of the data likelihood and the priors on the parameters.

$$\begin{aligned} f\left(Y, \Psi, \beta, \Phi, \Sigma | W\right) &\propto p\left(W | Y, \Psi, \beta, \Phi, \Sigma\right) \times f\left(Y, \Psi, \beta, \Phi, \Sigma\right) \\ &\quad Multi_{d+1}\left(W|Y\right) N_d(Y|\beta, \Phi, \Sigma)\, Wishart_d^{-1}\left(\Phi, \Sigma\right) \end{aligned}$$

By taking the log, this can be broken into a sum of distinct parts. At each iteration, the parameter values may be used to evaluate the full posterior log-likelihood. The

first part of the full posterior log-likelihood, the data likelihood, can be written in terms of the data given the latent variables:

$$l\left(W|Y,\Psi,\beta,\Phi,\Sigma\right) = log\left(p\left(W|Y\right)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} log\left(C_{ij}\prod_{k=1}^{d+1}\Pi\left(y_{ijk}\right)\right)$$

where $C_{ij} = \frac{PA_{ij}!}{w_{ij1}!...w_{ij(d+1)}!}$ is the multinomial coefficient and

$$\Pi(Y_{ij}) = \left(\frac{e^{y_{ij1}}}{1+\sum_{k=1}^{d}e^{y_{ijk}}}, \frac{e^{y_{ij2}}}{1+\sum_{k=1}^{d}e^{y_{ijk}}}, ..., \frac{1}{1+\sum_{k=1}^{d}e^{y_{ijk}}}\right)$$

The subsequent log-likelihoods for the different parameters can be broken up further:

$$l\left(Y,\Psi,\beta,\Phi,\Sigma\right) = l(Y|X,\Psi,\beta,\Sigma) + l(\Psi|\Phi) + l(\beta) + l(\Phi) + l(\Sigma)$$

In the multinomial logistic-normal setting, we define the distribution on the log-odds to be a multivariate normal distribution. The form of the other parameter log-likelihoods will depend on the defined prior distributions of those parameters; in our context, the normal prior on the random effects, the improper uniform prior on the fixed effects, and the inverse-Wishart priors on the covariance matrices.

2.3.2 Outline of Gibbs Sampler

We first describe the framework for the Gibbs sampler, allowing for a single random intercept, $q = 1$. When multiple random effects are included in the model ($q > 1$), some of the expressions are slightly complicated. These expressions are presented after the simpler single random intercept framework.

(1) Initialize starting values, $t-1=0$, Y_{ij}^0, β^0, ψ_i^0, Σ^0 and Φ^0.

(2) Update of Y_{ij} (Metropolis step):

$$Y_{ij}^t|X_{ij},\beta^{t-1},\psi_i^{t-1},\Sigma^{t-1},W_{ij} \propto p(W_{ij}|Y_{ij}^t) \times f(Y_{ij}^t|X_{ij},\theta_i^{t-1}=(\beta^{t-1},\psi_i^{t-1},\Sigma^{t-1}))$$

$$p(W_{ij}|Y_{ij}^t) \times f(Y_{ij}^t|X_{ij},\theta_i^{t-1}) \sim Multi_{d+1}(PA_{ij},\Pi(Y_{ij}^t)) \times N_d(X_{ij}\beta^{t-1}+\psi_i^{t-1},\Sigma^{t-1})$$

where Y_{ij}^t is proposed from a normal approximation to the beta distribution (see Section 2.3.3).

(3) Joint Update of β, ψ_i (Gibbs step):

$$\beta^t,\psi_i^t|X_i,\Sigma^{t-1},\Phi^{t-1},Y_i^t \propto f(\beta^t|X_i,\Sigma^{t-1},\Phi^{t-1},Y_i^t) \times f(\psi_i^t|\beta^t,X_i,\Sigma^{t-1},\Phi^{t-1},Y_i^t)$$

(a) Posterior of β without conditioning on ψ_i:

$vec(\beta^t)|X_i,\Sigma^{t-1},\Phi^{t-1},Y_i^t \sim N_{d(p+1)}(\mu_\beta = vec(A^{-1}b),\Sigma_\beta = A^{-1})$

where:

$$A = \sum_{i=1}^{m}\left((X_i^TX_i)\otimes(n_i\Phi^{t-1}+\Sigma^{t-1})^{-1}\right) \tag{2.3}$$

$$b = A\left(\sum_{i=1}^{m}X_i^TX_i\right)^{-1}\left(\sum_{i=1}^{m}X_i^TY_i^t\right) \tag{2.4}$$

(b) Conditional posterior of ψ_i:

$vec(\psi_i^t)|\beta^t, X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t \sim N_d(\tilde{\psi}_i, U_i)$

where:

$$\tilde{\psi}_i = U_i \left(\Sigma^{(t-1)^{-1}} \otimes \mathbf{1}_{n_i}\right) vec\left(Y_i^t - X_i\beta^t\right) \tag{2.5}$$

$$U_i = \left(\Phi^{(t-1)^{-1}} + \left(\Sigma^{(t-1)^{-1}} \otimes \mathbf{1}_{n_i}\mathbf{1}_{n_i}^T\right)\right)^{-1} \tag{2.6}$$

(4) Update of Φ (Gibbs step):

$\Phi^t|\Psi^t \sim Wishart^{-1}\left(\nu_2' = \nu_2 + m, \Lambda_2' = \Lambda_2^{-1} + \Psi^t\Psi^{t^T}\right)$

where:

$$\Psi^t = (\psi_1^t, ..., \psi_m^t)^T \tag{2.7}$$

(5) Update of Σ (Gibbs step):

$\Sigma^t|\Psi^t, Y^t \sim Wishart^{-1}\left(\nu_1' = \nu_1 + N, \Lambda_1' = \Lambda_1^{-1} + \hat{\varepsilon}\hat{\varepsilon}^T\right)$

where:

$$Y^t = (Y_1^t, ..., Y_m^t)^T \tag{2.8}$$

$$\hat{\varepsilon}_i = Y_i^t - X_i\hat{\beta}^t - \psi_i^t \tag{2.9}$$

$$\hat{\beta}^t = (Y^t - Z\Psi^t)X^T(XX^T)^{-1} \tag{2.10}$$

(6) Set $t = t + 1$, repeat steps (2) - (5) until convergence criteria met.

When multiple random effects are of interest, it is straightforward to define a random covariate design matrix, Z, made up of vectors Z_i for players $i = 1, ..., m$. Most

of the above algorithm can be represented the same way with this simple inclusion. However, one of the steps is complicated. The expression in step (3)(a), the posterior of β without conditioning on ψ_i, is a little more complex. Equation (2.3) must be rewritten as an expression where $\mathbf{1}_{n_i}$ is replaced by Z_i:

$$A = \sum_{i=1}^{m} \left(X_i \otimes I_d\right)^T Q_i \left(X_i \otimes I_d\right)) \tag{2.11}$$

$$Q_i^{-1} = \left(Z_i \otimes I_d\right) \Phi^{(t-1)} \left(Z_i \otimes I_d\right)^T + \left(I_{n_i} \otimes \Sigma^{(t-1)}\right) \tag{2.12}$$

Convergence and Computation Time

While the individual parameters themselves may not converge at the same rate as the full posterior log-likelihood, monitoring the posterior log-likelihood can provide a useful tool for assessing convergence. When small changes in the parameters no longer have a large impact on the posterior likelihood, it is reasonable to say the algorithm has converged. Many papers, which implement MCMC sampling algorithms, use visual inspection of the time series of realized values to determine if convergence has been reached. Gelman et al. (2004) provide some stricter strategies for assessing convergence, including computing potential scale reduction factors and effective sample sizes.

Currently, we graphically assess the convergence of the posterior log-likelihood (see Section 2.3.1), afterward setting a predetermined fixed number of iterations for the algorithm. This is a rather ad hoc approach, and not ideal. Still, generally we see convergence for relatively large data sets (at least 4000 player-seasons/observations) stabilize in convergence after a few thousand iterations, regardless of dimensionality of the count vector. After discarding burn-in and thinning we generally achieve good mixing. As an example, in one simulation study, we ran our model algorithm for 2500 iterations, discard the first 20% and thin every 2^{nd}, leaving us with 1000 MCMC real-

izations of each parameter. The effective sample sizes of the parameter chains ranged between approximately 173 and 885. Gelman et al. (2004) give a default rule of having at least 10 independent draws per sequence.

Computing resources will impact computation time, but Gibbs sampling is a generally inefficient Monte Carlo sampling method, as is the Metropolis algorithm. We briefly discuss a potential alternative to using Metropolis, Hamiltonian Monte Carlo, as a future improvement in Chapter 4. A run of the code (which will be published in an R package subsequent to publication of this thesis) of 10000 iterations in R version 3.1.2, on an Intel Core i7-3770 3.40GHz processor takes approximately 15 hours for the $m = 500$, $N = 4991$, $p = 13$, $q = 1$, and $d = 9$. The same data set, with $d = 2$, takes approximately 10 hours. Since our main purpose is to predict performance for new players, it would be somewhat prohibitive to refit the algorithm each time we wish to make a prediction about a new player. We will discuss how we circumvent this in Section 2.4.

2.3.3 Metropolis Normal Approximation to Beta Proposal

Due to the properties of the multinomial logistic-normal distribution, the conditional posterior likelihood of the latent variables, $f\left(Y_{ij}^t | X_{ij}, Z_{ij}, \beta^{t-1}, \psi_i^{t-1}, \Sigma^{t-1}, W_{ij}\right)$, has no closed-form expression. This motivates the use of a Metropolis or Metropolis-Hastings algorithm in the Gibbs step. In such algorithms, values of the target distribution are proposed from a different distribution, and a Metropolis acceptance ratio is calculated based on the ratio of likelihoods (or difference in log-likelihoods). There are numerous potential proposal distributions which could be used in step (2) of the Gibbs Sampler described in the previous subsection. Here we describe our Metropolis proposal distribution of choice, while alternatives are described in more detail in Appendix B.

Described in Appendix B.1, our initial choice for proposal distribution was a normal random walk scheme where new values of the log-odds Y_ij were proposed from a multivariate Gaussian centered at the previous value, with a proposal covariance equal to a scaled version of Σ. This proposal scheme tended to produce poor mixing of Y_ij. Instead we consider proposing based on probabilities the latent variables imply. For the purposes of proposing, we begin by considering the counts W_{ij} as being generated through a series of conditionally binomial draws:

$$w_{ijk}|\left\{w_{ijl}\right\}_{l\notin\{k,d+1\}} \sim binomial\left(PA_{ij} - \sum_{l\notin\{k,d+1\}} w_{ijl}, \pi^*_{ijk} = \frac{\pi_{ijk}}{\pi_{ijk}+\pi_{ij(d+1)}}\right)$$

In Appendix B.2, we discuss a Metropolis-Hastings scheme for proposing new probabilities π^*_{ijk} from a beta proposal distribution, based on the current values of the parameters, and then converting to log-odds. In essence, we assume *a priori* that π^*_{ijk} follows a beta distribution based on the current values of the parameters:

$$\pi^*_{ijk} = \frac{e^{y_{ijk}}}{1+e^{y_{ijk}}} \sim beta\left(\alpha^*_{ijk}, \beta^*_{ijk}\right) \tag{2.13}$$

with $\alpha^{*^{t-1}}_{ijk} = \frac{1+e^{\mu^{t-1}_{ijk}}}{\sigma^{2^{t-1}}_k} + \frac{e^{\mu^{t-1}_{ijk}}}{1+e^{\mu^{t-1}_{ijk}}}$, $\beta^{*^{t-1}}_{ijk} = \alpha^{*^{t-1}}_{ijk} e^{-\mu^{t-1}_{ijk}}$, $\mu^{t-1}_{ijk} = X_{ij}\beta^{t-1}_k + \psi^{t-1}_{ik}$, and $\sigma^{2^{t-1}}_k$ is the $(k,k)^{th}$ element of Σ^{t-1}.

And thus given the data and current values of the parameters:

$$\pi^{*^t}_{ijk}|W_{ij}, \theta^{t-1}_{ik} \sim beta\left(w_{ijk} + \alpha^{*^{t-1}}_{ijk}, w_{ij(d+1)} + \beta^{*^{t-1}}_{ijk}\right)$$

This Metropolis-Hastings beta proposal mixed much better than the normal random walk proposal, but took more computation time. The proposal we settled on moves back to a symmetric, Metropolis, proposal setting by proposing the logit of the beta distributed probabilities (Equation 2.13) with a normal distribution.

In essence, since we know $y^t_{ijk} = logit(\pi^{*^t}_{ijk})$, and, since $\pi^{*^t}_{ijk}$ is beta distributed, then:

$$E\left[logit(\pi_{ijk}^{*^t})\right] = \psi_0\left(w_{ijk} + \alpha_{ijk}^{*^{t-1}}\right) - \psi_0\left(w_{ij(d+1)} + \beta_{ijk}^{*^{t-1}}\right) \tag{2.14}$$

$$var\left[logit(\pi_{ijk}^{*^t})\right] = \psi_1\left(w_{ijk} + \alpha_{ijk}^{*^{t-1}}\right) + \psi_1\left(w_{ij(d+1)} + \beta_{ijk}^{*^{t-1}}\right) \tag{2.15}$$

where $\psi_0(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function.

We can then construct a proposal distribution such that:

(1) Propose $Y_{ij}^t|\mu_{\pi^*}^{t-1}, \Sigma_{\pi^*}^{t-1} \sim N_d\left(\mu_{\pi^*}^{t-1}, c\Sigma_{\pi^*}^{t-1}\right)$

where $\mu_{\pi^*}^{t-1} = \left(E\left[logit(\pi_{ij1}^{*^t})\right], ..., E\left[logit(\pi_{ijd}^{*^t})\right]\right)$, $\sigma_{\pi^*k}^{2^{t-1}} = var\left[logit(\pi_{ijk}^{*^t})\right]$ is the $(k,k)^{th}$ element of the diagonal matrix $\Sigma_{\pi^*}^{t-1}$, c is a scalar used to control the "step-size" of the random walk, and $\alpha_{ijk}^{*^{t-1}}$ and $\beta_{ijk}^{*^{t-1}}$ are defined as previously.

(2) Accept with probability $\alpha = min\left\{1, \frac{f\left(Y_{ij}^t|X_{ij},\theta_i^{t-1},W_{ij}\right)}{f\left(Y_{ij}^{t-1}|X_{ij},\theta_i^{t-1},W_{ij}\right)} \times \frac{q\left(\pi_{ij}^{t-1}|X_{ij},\theta_i^{t-1},W_{ij}\right)}{q\left(\pi_{ij}^{t}|X_{ij},\theta_i^{t-1},W_{ij}\right)}\right\}$

Where $q\left(Y_{ij}^t|X_{ij},\theta_i^{t-1},W_{ij}\right) \sim N_d\left(\mu_{\pi^*}^{t-1}, c\Sigma_{\pi^*}^{t-1}\right)$.

2.4 Prediction of a New Player

Suppose an NPB team is in need of a catcher and is considering Kevin Plawecki, former Purdue catcher and current catcher with the Cleveland Indians. Because he has only played in the major leagues, he is not in our data set meaning we do not have information regarding his ability (ψ_i) and therefore cannot make an accurate prediction of his performance if he were to transition to NPB. This section discusses how we approach this given that we have already generated samples from the joint posterior distribution of β, Σ, and Φ. In this situation, we have a new individual, i' who is not present in the original data, and thus whose past latent variables and subsequent random effect are unknown; $Y_{i'1}, ..., Y_{i'j'}, \psi_{i'}$. We could add his output vectors to our data set are rerun our Metropolis-within-Gibbs sampler, but that is going to take time.

A faster alternative is to rerun the Metropolis-within-Gibbs sampler holding the β, Σ, and Φ quantities fixed at their posterior means. In other words, using our algorithm to only fit the latent variables and random effect for player i'. By using our algorithm, we obtain a set of posterior values $\psi_{i'}^t$ for this new player and can use the posterior mean for our prediction. This is a useful approach when we want to quickly generate predictions for new players. A comparison of this approach with the posterior mean of $\Psi_{i'}$ based on refitting the model with this player's historical data revealed little difference (for an illustration, see Figure 3.7 in the next Chapter). Thus, one might consider updating the model annually and using this approach to make predictions for players of interest during the year.

2.5 Methods for Assessing Model Fit and Prediction Uncertainty

Model diagnostics for Bayesian hierarchical models is an open research area (Crespi and Boscardin, 2009; Yuan and Johnson, 2012). The most common form of model checking is the use of posterior predictive distributions (Gelman et al., 2004; Gelman and Hill, 2006). Graphical model fit assessments are also practical and useful.

In this section, we discuss a method for assessing overall model fit using squared Mahalanobis distances of the log-transformed seasonal proportion vectors. We also discuss model selection via the use of the deviance information criterion (DIC) (Gelman et al., 2004). Finally, we discuss methods for assessing predictive accuracy and outline processes of quantifying uncertainty, in the form of prediction or credible intervals, for our multinomial logistic-normal model.

2.5.1 Overall Model Fit

A convenient way to approach model fit diagnostics is to define a form of residual and assess if it behaves as it should under the model being fit. For response data, which exist on the simplex, $\mathbb{S}^{d+1}$, many traditional distance measures can be misleading

(Aitchison, 1986). Aitchison (1986) pioneered many of the techniques for measuring distances on the simplex, including Aitchison's distance (see Appendix C). These measures generally involve a transformation of the composition or count vector via some log-ratio, but are expressly designed to compare two points.

In our Bayesian framework, we are interested the joint behavior of the data relative to the multinomial logistic-normal model. Because each count vector has a different posterior predictive distribution, we are interested in whether the data jointly seem to fit these distributions. In the univariate normal error case, this is accomplished by computing the residuals, or deviation from the estimated mean, and comparing their distribution to the normal distribution. For our case, we propose computing the squared Mahalanobis distances (Mahalanobis, 1936) of the transformed counts into log-odds and comparing their distribution to what is expected under our model.

We consider the Mahalanobis distance because it takes into account the covariance structure among the counts, or transformed log-odds. The general form for the squared Mahalanobis distance of a vector Y is:

$$MD^2 = ((Y - M)^T C(Y - M)) \tag{2.16}$$

where M is the true or estimated location (mean) vector and C is the true or estimated covariance matrix.

If our model implied the transformed count data were multivariate normal, we would compare these squared distances to a chi-square distribution (Filzmoser et al., 2005). For large N, this is approximately true under our model but it would be risky to make that assumption here. Instead, we propose using simulation to form the empirical distribution of squared Mahalanobis distances for each data vector and use those distributions to assess if the data behave as expected.

In the context of our framework, suppose we have fit the model and have parameter estimates. For player i season j, we then generate L new count vectors $W_{ij}^1, ..., W_{ij}^L$, under the fitted model and transform them to the log-odds scale:

$$\tilde{Y}_{ij}^l = \left(log\left(\frac{w_{ij1}^l}{w_{ij(d+1)}^l}\right), ..., log\left(\frac{w_{ijd}^l}{w_{ij(d+1)}^l}\right)\right), l = 1, 2, \ldots, L \tag{2.17}$$

Methods described in the following subsection can be used if there is the possibility of zero counts.

Our goal is to determine how well the data are fitting the model so we calculate the squared Mahalanobis distances of all model-generated log-odds, $\left\{\tilde{Y}_{ij}\right\}$, and our observed log-odds, Y_{ij}^{obs}, using:

$$MD_{ij}^2 = \left(\left(Y_{ij} - \bar{\tilde{Y}}_{ij}\right)^T \hat{\Sigma}_{\tilde{Y}_{ij}}^{-1} \left((Y_{ij} - \bar{\tilde{Y}}_{ij}\right)\right) \tag{2.18}$$

where Y_{ij} is either Y_{ij}^{obs} or $\tilde{Y}_{ij}$, $\bar{\tilde{Y}}_{ij}$ is the sample mean of model-generated log-odds and $\hat{\Sigma}_{\tilde{Y}_{ij}}$ is their sample covariance.

Using the model-generated MD^2 values, we generate the empirical cumulative distribution function, $\mathcal{F}$, and compute the percentile of the observed squared Mahalanobis distance, after correcting for the discreteness of the multinomial counts and the fact that our empirical distribution is based on L values. If the data fit the model, we would then expect these percentiles to be distributed uniformly.

Because the response vector are counts, the distribution of squared Mahalanobis distances is discrete in nature. To consider the percentiles as uniform, some form of continuity adjustment is needed. One option, explained in Dunn and Smyth (1996), is to use randomization to smooth away the discreteness; we can randomly generate a uniform random variable for the percentiles of the observed MD_{ij}^2 where the minimum and maximum depend on the observed value's ordered location among the model-based $\widetilde{MD}_{ij}^2$:

(1) If the observed $MD_{ij}^2 \leq min\left(\widetilde{MD}_{ij}^2\right)$

$$u_{ij} \sim Unif\left(0, \mathcal{F}\left(min\left(\widetilde{MD}_{ij}^2\right)\right)\right)$$

(2) If the observed $MD_{ij}^2 > max\left(\widetilde{MD}_{ij}^2\right)$

$$u_{ij} \sim Unif\left(\mathcal{F}\left(max\left(\widetilde{MD}_{ij}^2\right)\right), 1\right)$$

(3) Otherwise,

$$u_{ij} \sim Unif\left(\mathcal{F}\left(MD_{ij}^{2*}\right), \mathcal{F}\left(\text{observed } MD_{ij}^2\right)\right)$$

$$\text{where } MD_{ij}^{2*} = max\left(\widetilde{MD}_{ij}^2 < \text{observed } MD_{ij}^2\right)$$

Dunn and Smyth (1996) also propose back-transforming these percentiles to standard normal values to serve as our residuals, r_{ij}. Normal quantile plots and residual plots can the be used to assess fit.

This process can be seen as generating a form of Dunn-Smyth residual (Dunn and Smyth, 1996). A R package, DHARMa (Hartig, 2019), has been published which extends this idea to creating residuals for results from fitted generalized linear mixed models in a univariate context. Our multivariate extension, using the squared Mahalanobis distance, is straightforward to implement and provides a way to assess model fit under our model; thus creating "conditional" residuals. It is also possible to generate the L new count vectors from the marginal distribution of $\tilde{Y}_{ij}$, though these would not take into account the random effect's impact on the mean. This Dunn-Smyth type residual could also be used to assess prediction of a marginal count, if only one outcome is of interest.

When calculating the residuals based off of new players i' and new seasons j', held out of the initial run, we can view the squared Mahalanobis distances as a form of residual where the independence assumption is more fully realized than when looking at the residuals of the player-seasons used to fit the model.

One drawback with this method is obvious; the data are necessarily finite, and there will certainly be the occasional zero count. In the presence of a single zero count within a vector of posterior predicted counts, there is no way to compute the associated log-odds without somehow adjusting said count vector. Given enough player-seasons with large exposure, we expect the impact these zero counts have to be somewhat diminished.

Handling Zero Counts

In some cases, posterior predictive draws will return zero counts for some outcomes or the observed player-season outcome vector will have zero counts. There are several corrections which are well established in the literature for handling zero counts logistic regression settings. One simple correction is to add a small error term, ε, to each zero count, which does not dramatically effect the observed probability, and in re-scaling when calculating the log-odds, allows for non-infinite values. For a count vector W, simply replace all zero counts with ε:

$$W^{\varepsilon} = W\left[w_k = 0 \rightarrow w_k = \varepsilon\right] \tag{2.19}$$

The above approach however introduces the same bias for vectors with zero counts regardless of the overall exposure, and some tuning to determine the optimal ε is almost certainly needed. The correction we use instead is based on one proposed by Smithson and Verkuilen (2006). It is possible to compress the data symmetrically around $\frac{1}{K}$, penalizing vectors with less information. In other words, for a count vector W of length K, with exposure N and in the presence of zero counts:

$$W^{comp} = \frac{W(N-1) + \frac{1}{K}}{N} \tag{2.20}$$

This is the same correction we can perform for the player-seasons before applying the multinomial logistic-normal methodology in estimating the initial values of the log-odds from the data. The benefits of this approach over the first approach are

twofold. Firstly it does not require any tuning. Secondly, by penalizing vectors with lower exposure, it limits the impact those vectors may have on the estimation of parameters, allowing those vectors with more information to be the driving force in model estimation. Obviously, with an abundance of zeros, a different strategy for handling excessive sparsity would be necessary.

2.5.2 Comparing Models via DIC

Information criterion have played a large role historically in the comparison of model fit. The most commonly used model selection criteria are the Akaike information criterion (AIC) and the Bayesian (also called Schwarz) information criterion (BIC). BIC, as described in Gelman et al. (2004) has a slightly different goal than AIC. While AIC is an estimation of predictive fit, BIC is meant to approximate the marginal probability density of the data. Since we are interested in predictive fit, an AIC like criterion would be more appropriate.

Gelman et al. (2004) describes alternatives to AIC that can be calculated based on posterior samples of the parameters. Deviance information criterion (DIC) is a somewhat Bayesian version of AIC, while the Watanabe-Akaike information criterion ($WAIC$) is more fully Bayesian. Gelman et al. prefer $WAIC$ to DIC based on it not conditioning on a point estimate. However, $WAIC$ requires partitioning the data, which becomes extremely non-trivial in longitudinal data settings such as ours. For this reason, we focus on using the DIC as our measure of comparative model fit.

DIC, like AIC, is calculated from estimating the log posterior density as well as the effective number of parameters, which serves as a penalty for over-fitting the data. Essentially, any models which have produced posterior chains of the parameters for the same data set may be compared using DIC, with the smaller model DIC indicating a more appropriate model fit.

The calculation of DIC begins with estimating the number of effective parameters, p_{DIC}, based on posterior samples $t = 1, ..., l$:

$$p_{DIC} = 2\left(l\left(W|\hat{\theta}\right) - \frac{1}{l}\sum_{t=1}^{l} l\left(W|\theta^t\right)\right) \tag{2.21}$$

where W is the data, θ is the vector of parameters for which we have posterior samples, and $\hat{\theta}$ is the vector of posterior means of said parameters, $\hat{\theta} = \frac{1}{l}\sum_{t=1}^{l}\theta^t$.

The DIC is then defined in terms of the deviance, similarly to AIC:

$$DIC = -2l\left(W|\hat{\theta}\right) + 2p_{DIC} \tag{2.22}$$

Direct comparison of model fit to data sets follows immediately. We will use DIC in Section 3.3 to compare data simulated under the multinomial logit and multinomial logistic-normal settings. This will establish how DIC can help show the importance of allowing for the additional variability the multinomial logistic-normal has over the multinomial logit.

2.5.3 Predictive Accuracy and Uncertainty

In the previous two sections, we describe a process for assessing overall model fit via Dunn-Smyth type residuals based on squared Mahalanobis distances, and the Deviance information criterion for discriminating between models. As mentioned in the previous section DIC, an AIC like criterion, is an estimation of overall predictive fit, though is only useful in a comparative setting. The Dunn-Smyth residual plots may assess if predictions are reasonable under the model fit. Neither are used for assessing the accuracy of predictions. We discuss two measures traditionally used in assessing predictive accuracy of compositional data in Appendix C, but neither of these measures provide measures of uncertainty about a single prediction.

In baseball projection systems specifically, there has already been some work done online in terms of comparing existing projection systems. Starting in 2015, Will Larson collected numerous projection results from different systems across the web at The Baseball Projection Project (Larson, 2015b) and reviewed their performance on the website Fangraphs (Larson, 2015a). In these reviews, Larson tested important outcomes from each projection system received for error and predictive power via root mean square error ($RMSE$) and R^2. This is a fairly naïve approach that would nonetheless be straightforward to use in comparing this model to existing projection systems. Additionally, many comparisons of projections focus on comparing summary level statistics, especially on-base plus slugging percentage, or OPS (Smith, 2006; Tango et al., 2007) via metrics such as $RMSE$. Since it is straightforward to calculate OPS from the posterior predictive means of our model fit, it would be possible to compare this model to existing projection systems in this way.

However, none of the existing projection systems provide a useful level of uncertainty quantification about individual projections of player-seasons. While direct comparison of our posterior predictive means to other projection results is possible, it is not possible to adequately factor in the variability about the projection for competing systems. We believe it to be extremely valuable to provide prediction or credible intervals about the entire vector of predicted counts for individual player-seasons, especially in terms of assessing the relative risks involved in a team's management choosing between players, depending on the team's needs.

The following subsection discusses building of prediction and credible intervals based on samples from the posterior predictive distribution generated from fitting the model. Since the intervals are based on samples from a distribution, they can be built not only for individual outcomes, but for summary level statistics as well, which are often the baseball fan's preferred method of immediately assessing a player's worth.

Uncertainty Quantification

There has been little work done in quantifying the uncertainty of predictions in baseball. What work has been done has been of the singular metric variety, providing a single measure for the certainty of a projection. Part of this dearth of uncertainty quantification comes from the simple truth that many intervals in a prediction setting are not quite useful in baseball terms. Simultaneous intervals for multinomial count vectors are often too wide as to be useful. However, when comparison between the uncertainty of players is of interest, even intervals that would be useless on their own provide some value.

Given the choice between similar players, i.e. players of similar age and position and with similar point predictions, it is reasonable to assume a team would prefer the player accompanied by less uncertainty about that point prediction. For our purposes, it is straightforward to provide that quantification based on the posterior predictive distributions of the player-seasons being predicted.

Point estimates are estimated in a straightforward manner as the mean of the MCMC samples, $t = 1, ..., l$, of counts from the posterior predictive distribution, $W^1_{i'j'}, ..., W^l_{i'j'}$. There can similarly be a point estimate for the underlying probability vector for the posterior predictive distribution by taking the mean of the MCMC samples of the probabilities defined by the log-odds, $Y^1_{i'j'}, ..., Y^l_{i'j'}$. Based on this mean posterior predictive probability, a 95% prediction interval for the new player-season can be constructed of the convex hull of the 95% most likely posterior predictive samples under a multinomial distribution with probability vector equal to the mean posterior predictive probability. While the resulting intervals will sometimes be too wide to be very informative when direct comparison of two players is desirable, they are adequate for assessing predictive accuracy on a single player-season basis. The issue of confidence intervals in multinomial settings being overly conservative has been discussed

before (Sison and Glaz, 1995), and the hierarchical structure of our model only adds additional variability.

An alternative to the prediction interval created from the posterior predictive distribution of counts would be a credible interval for the mean prediction, formed on the based on samples from the posterior distribution of the random effects, $\psi_{i'}^1, ..., \psi_{i'}^l$. This, in essence, ignores seasonal variability and asks what the interval about a player's expected performance would be. A 95% credible interval for a new player-season is constructed via selecting the 95% most likely ($t^* = 1, ..., l^*$) random effects for a new player under the posterior distribution of random effects $\psi_{i'}^{t^*}$, calculating the expected log-odds for the new player-season $X_{i'j'}\beta^{t^*} + \psi_{i'}^{t^*}$, and finally transforming those into probabilities. Multiplying by the exposure represents an interval for the vector of expected counts of a new player-season. These credible intervals are narrower, and more useful for assessing risk between players. However, by ignoring the seasonal variability that is present for each vector of counts, we cannot expect the intervals to cover the observed counts 95% of the time.

The same process may also be used to create 95% credible intervals for singular metrics, which might be of interest as a summary of player ability. One may take the same set of 95% most likely posterior samples of random effects which produced the credible intervals of all the categories and calculate the singular metric, be it batting average (AVG), on-base plus slugging percentage (OPS), or other metrics which may be calculated from the original vector of counts. This results in what can be considered samples from a posterior predictive distribution of the metric, from which a credible interval can be constructed.

2.6 Discussion

In this chapter we introduced a mixed-effects Bayesian hierarchical multinomial logistic-normal model for baseball performance data. We discussed methodological choices in what features to include in the model, and pointed to adjustments which could be made. Subsequently we described a Metropolis-within-Gibbs algorithm for sampling from the conditional posterior distributions of our latent variables and parameters which tends to converge within a few thousand iterations. We discussed the process of forming predictions and assessing uncertainty about those predictions via prediction or credible intervals constructed from posterior samples. We also spent some time developing methods for assessing overall model fit and model diagnostic procedures, including introducing a Dunn-Smyth type residual which conditions on the model fit using squared Mahalanobis distances for transformed multivariate count response.

The ensuing chapter begins by providing context for the importance of our application, followed by discussion of the data we have to work with. It continues with the application of the model described in this chapter first in a simulation study, and then to a real data set of all baseball players to have played in both Japan's NPB and MLB leagues up to and including the year 2014. In the simulation study, we show that in the presence of sufficient additional uncertainty about probabilities, the mixed effects multinomial logistic-normal model is preferred to the mixed effects multinomial logit, and illustrate how predictions and uncertainty assessment proceeds. We follow with the real data analysis and discuss the model fit of the real data and its predictive performance.

3. PREDICTING PERFORMANCE OF PLAYERS MOVING BETWEEN NPB AND MLB

3.1 Introduction

Masanori Murakami was the first Japanese player to play in Major League Baseball. His NPB team, the Nankai Hawks, sent him to the San Francisco Giants as a minor league exchange player in 1964. Murakami was promoted to the Major Leagues that year, and pitched for two seasons in MLB before returning to his NPB club.

Murakami, however, was not the first player to have played in both leagues. Over a decade earlier, in 1953, MLB pitchers Leo Kiely and Phil Paine played in games for the Mainichi Orions and Nishitetsu Lions, respectively, while completing their military service in Japan. Both pitchers would return to the United States in 1954 after their military service ended. The year 1953 also saw the first position player move from the United States to Japan. Larry Raines left the Negro Leagues to play two seasons for the Hankyu Braves. In 1954, he became the first American player to win a batting title in Japan. When Raines eventually played in the Major Leagues, he made his debut with the Cleveland Indians in 1957, seven years before Murakami joined the Giants.

The purpose of relating this history is to establish that players have been moving between NPB and MLB for over sixty years. While most of the transitions across the Pacific Ocean are MLB players moving to NPB, some of the most dynamic Japanese players have also moved, including all-stars and at least one player, in Ichiro Suzuki, who is all but certain to be inducted into MLB's Hall of Fame.

In both directions, money is a huge factor. NPB teams operate on modest budgets compared to their MLB counterparts, and are only allowed an allotment of four foreign players on their active roster. In 2014, Andruw Jones $3.8 million contract made up 14% of the Rakuten Golden Eagles' entire payroll. For Japanese teams, recruiting foreign players who fit their roster, are interested in playing in Japan, and can provide appropriate value for their contract, is vital for success.

Meanwhile, Japanese players have three main avenues for coming to play in the United States. The first is possible when they are young. Before entering the NPB draft, they may sign as an amateur free agent with MLB teams. In doing so, the player is forced to sit out two to three years before being able to return to play for an NPB club. One notable player, Junichi Tazawa, took this route and remains in MLB today. The second avenue is by accruing enough service time in NPB, nine years, to be considered a free agent. While several successful players, such as Hideki Matsui, have come to MLB in this way, they are often older and in less demand. The final avenue is via a posting system, negotiated between NPB and MLB, where players who have yet to accrue the nine years of service time are posted by their NPB teams. MLB teams pay a release fee to the NPB team following negotiations with the player. While the exact nature of the fee has changed over the years, this release fee has exceeded $50 million, as was the case in 2006 when the Boston Red Sox paid to negotiate with Daisuke Matsuzaka.

In Chapter 1, the current state of publicly available baseball projections was discussed. Many of the available projection systems provide estimates of seasonal counts for players coming to MLB from NPB, but we are not aware of any that do the same going from MLB to NPB. Furthermore, of all the available projections for MLB players, only a few provide an estimate of uncertainty. Most are proprietary and the process by which the uncertainty is quantified is unknonw. For the Marcels' reliability score, which is open source, Japanese players always have a score of 0 and are predicted at the league average for their age.

In this chapter, we use simulation studies to demonstrate our model's ability to predict player counts and to construct informative credible intervals about the player's expected performance for years when the player switched leagues. We then apply our mixed-effects Bayesian hierarchical multinomial logistic-normal model to predict performance of players moving between the NPB and MLB leagues. We begin with a description of the data used, then follow with results, including assessing model fit, model comparison, and predictive performance with uncertainty quantification. We end the chapter with a discussion of the model's performance and use in this context.

3.2 Data Description

The data used in this project come from two main sources. Nippon Professional Baseball player-season data were provided by Data Stadium, Inc., S-GATE Akasaka, 6-2-4, Akasaka, Minato-ku, Tokyo 107-0052. Major League Baseball player-season data for relevant players, as well as their corresponding covariates, were taken from the Lahman database available in R (Friendly, 2015). Some data augmentation was necessary to collate the two data sources; from small matters such as converting weight in pounds to weight in kilograms, to slightly more involved tasks such as identifying discrepancies in recorded birth dates. Some artifacts of the different data sources, such as slight differences in height and weight for the same players, still remain. We also consider players switching between leagues in the middle of a season as two seasons of data.

The data consist of all players who have played in both NPB and MLB before 2015. For the batting data, this includes 605 players with 4991 player-seasons. While it is possible to fit a pitching outcome model using our hierarchical approach (i.e., replace at bats with batters faced), we focus only on modeling hitter outcomes here (so this includes pitchers as batters).

There is substantial variability in the length of careers for the 605 players, as well as the amount of time spent in each league, and the number of plate appearances in each season. All told, there were an average of 8.25 seasons per player, with an average of 5.62 MLB seasons per player and 2.63 NPB seasons per player, with 225.3 plate appearances per season. Players moving from MLB to NPB averaged 5.82 years before switching and 2.42 years after switching, while players moving from NPB to MLB averaged 7.36 years before the switch and 3.40 years after the switch.

Tables 3.1 and 3.2 give nine rows of the collated batting data set to give a more concrete taste of what the data look like. While Phil Paine was a pitcher and teams would not be particularly interested in his bat, Paine and Larry Raines serve as an illustration of the problem of interest. Paine moved from the National League to Pacific League in 1953, Raines from the Pacific League to American League in 1957. At the time, neither the Nishitetsu Lions nor the Cleveland Indians could know exactly what performance to expect from the players, respectively. After sixty years, and an average of 10 players switching leagues per year, there should be enough information to better answer the question of player performance in another league.

3.2.1 Practical Considerations

Real data analysis is not conducted without making some assumptions or dealing with some limitations. Among the covariates available in the data set, those chosen were based on what might most help explain variability in performance. Height, weight, age, and batting handedness all could reasonably impact the performance of certain types of players; heavier players, for example, will be less likely to be slap hitters looking to take advantage of their speed. A quadratic age term was included; there have been studies that suggest the career trajectories of players' performance are quadratic in nature (Albert, 2002; Fair, 2008), though occasionally a more complex aging effect is modeled, such as Berry et al. (1999) use of random spline curves.

Table 3.1.
Covariate values over the careers of two of the first players to switch leagues between NPB and MLB, representative of the real data set

name	**height (cm)**	**weight (kg)**	**bats**	**year**	**age**	**league**	**position**
Phil Paine	188	81	R	1951	21	MLB-NL	P
Phil Paine	185	81	R	1953	23	NPB-PL	P
Phil Paine	188	81	R	1955	25	MLB-NL	P
Phil Paine	188	81	R	1958	28	MLB-NL	P
Larry Raines	179	75	R	1953	23	NPB-PL	IF
Larry Raines	179	75	R	1954	24	NPB-PL	IF
Larry Raines	178	74	R	1957	27	MLB-AL	3B
Larry Raines	178	74	R	1958	28	MLB-AL	2B
Larry Raines	179	75	R	1962	32	NPB-PL	IF

Table 3.2.
Outcome counts over the careers of two of the first players to switch leagues between NPB and MLB, representative of the real data set

name	1B	2B	3B	HR	SH	SF	BB	HBP	SO	OIP
Phil Paine	0	0	0	0	0	0	0	0	1	3
Phil Paine	5	1	0	1	0	0	2	0	9	8
Phil Paine	1	0	0	0	0	0	0	0	0	2
Phil Paine	2	0	0	0	1	0	0	0	1	4
Larry Raines	99	21	16	8	3	0	41	2	53	306
Larry Raines	120	38	8	18	7	4	25	6	60	302
Larry Raines	48	14	0	2	2	0	19	1	40	140
Larry Raines	0	0	0	0	0	0	0	0	5	4
Larry Raines	43	6	1	5	6	2	18	0	46	117

Including league as a fixed effect makes sense in that these are the only leagues we are interested in players moving between (no leagues unaccounted for) and we expect league to not be independent of age; players moving to NPB from MLB tend to be older than the reverse, as indicated by the average age of 30.97 years for NPB player-seasons and 27.89 years for MLB player-seasons in the data set. Finally, although we did not include it in the simulation studies, we would expect that position could explain some variability, especially given a typical batting line for a pitcher.

Choosing the granularity of the count vectors depends on the nature of the data collected and the goals of the research. We think that our ten main outcomes are those which on the whole give enough information to describe a player's performance, while being sufficient for calculating summary statistics, such as batting average or OPS. They are, however, different than what others have used. Null (2009), for instance, considers hit location (Ground Ball, Fly Ball) and the rare catcher's interference event in his outcome list, but not the more common sacrifice.

3.3 Simulation Results

Before applying the methodology to the real data problem, it is important to establish the performance of the method via simulation (i.e., proof of concept). In this section, we summarize the results of our model fit to data simulated data under the multinomial logistic-normal (MLN) model with known parameter settings, as well as under the multinomial logit (M-Logit) model. In doing so, we describe how well our approach recovers the various parameters, performs in terms of prediction, and indicates good or poor model fit to the data. We conducted several simulation studies involving different parameter settings and different values of K, but only report results from one involving the most basic three-category vector.

3.3.1 Three-Category Simulation Model

As we did in our earlier example to justify the use of the logistic-normal distribution (see Section 1.5), we consider a simulation study with $K = 3$ outcomes where two of the three outcomes having a positive correlation. Our choice here is $(HR, BB, OTHER)$, which is slightly different from the one used earlier: $(HR, SO, OTHER)$.

To ensure the simulated data are as similar as possible in form to the real data, we use the player covariates and season at bats from the real data set and base the parameter settings on a trial run of our algorithm fit to the real data. Covariates included are: $height, weight, age, age^2, handedness$, and $league$. For model fit assessment, we also simulate data using the multinomial logit model (see Appendix A). By fitting both of these models to data sets generated under each model, we can compare model fits and assess the robustness of each model to varying conditions.

The process of simulation for the MLN and multinomial logit models unfold like so:

- Set global parameters, Φ, Σ, β. Note that Σ is only used for the MLN model.
- For each player i, simulate a random effect vector; $\psi_i \sim N_2(0, \Phi)$.
- For each season, $j = 1, ..., n_i$, use the player-season covariates to compute the log-odds vector $Y_{ij} = X_{ij}\beta + \psi_i$. If generating data for the MLN model, simulate an error vector; $\varepsilon_{ij} \sim N_2(0, \Sigma)$, and add this to the log-odds vector.
- Convert the resulting log-odds into probabilities Π_{ij}
- Simulate data W_{ij} from a multinomial distribution given plate appearances PA_{ij} and probability vector Π_{ij}.

The two data archetypes are easily delineated by how the Π_{ij} are modeled. The multinomial logistic-normal model allows for variability in Π_{ij} based on the inclusion

of error terms, while the multinomial logit model assumes these logits are constant. The corresponding models in terms of the probability vector for player i, season j are:

- Mixed effects multinomial logistic-normal:

$$\Pi_{ij} = \left(\frac{e^{y_{ij1}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{e^{y_{ij2}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{1}{1 + e^{y_{ij1}} + e^{y_{ij2}}} \right)$$

$$Y_{ij} = X_{ij}\beta + \psi_i + \varepsilon_{ij}$$

- Mixed effects multinomial logit:

$$\Pi_{ij} = \left(\frac{e^{y_{ij1}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{e^{y_{ij2}}}{1 + e^{y_{ij1}} + e^{y_{ij2}}}, \frac{1}{1 + e^{y_{ij1}} + e^{y_{ij2}}} \right)$$

$$Y_{ij} = X_{ij}\beta + \psi_i$$

Under the same parameter settings of β and Φ, the data archetypes have the same mean probability vector for each player-season. The only difference is whether the Π_{ij} are random or not.

Under each data archetype, we use the real data design matrix to simulate 605 players and 4991 player-seasons. Due to the relative abundance of data, we fit our mixed effects multinomial logistic-normal model to 500 of the players, leaving out 105 players and corresponding careers for a validation approach to assessing model fit and prediction. While we have simulated data for every year of their career, for the 105 out of sample players, only their career up until switching leagues from NPB to MLB or vice versa is used to generate posterior chains of random effects, since this will mimic the data we would like to predict in practice. The goal is then to generate posterior predictive distributions of the 105 player-seasons representing the player's switch in leagues. These seasons will be predicted using the multinomial logistic-normal model

for both of the data archetypes, while the in-sample 500 player careers will be used for fitting the same MLN model. For the multinomial logistic-normal data, we will examine the MCMC posterior chains for the parameters to assess the performance in terms of parameter recovery, as well as construct prediction and credible intervals for a selected pair of players. Results for a single data set of each type fit under each model are presented, and were indicative of results across several simulation studies.

Parameter Settings

Parameter settings were determined by a short trial run of the real data with the Metropolis-within-Gibbs algorithm used to fit the mixed effects multinomial logistic-normal model. Values of the parameters are below, but initially it is helpful to use an example to gain intuition on how to interpret the fixed parameters.

Counts are simulated based on six covariates: $height, weight, age, age^2, handedness$, $league$, and the single player random effect. The fixed intercept, β_0, is chosen so a baseline player represents a player with average height, weight, age, is right handed and in the American League of MLB. In essence, the continuous predictors are scaled, while a baseline category is chosen for the categorical predictors. A change of 1 unit of the continuous covariates thus represents the change in log-odds for a player a standard deviation from the mean of those predictors. The indicator variables representing the categorical predictors can be seen as the shift in log-odds for a baseline player when moving between handedness or league.

The player random effect can be interpreted as the individual player's ability above or below the average player with their covariates. While it is not unilaterally true, a good rule of thumb for interpreting the random effects is if the player has a positive random effect, they are above average in the category corresponding to the numerator of the log-odds.

For example, consider the fixed intercept, $\beta_0 = (-3.6269, -2.1937)$. Since this is on the log-odds scale, it is often convenient to know what this baseline player's outcome composition looks like on the probability scale;

$$\beta_0 = (-3.6269, -2.1937) \implies (HR_0, BB_0, Other_0) = (0.023, 0.098, 0.879)$$

Now imagine a baseline player moving from the MLB American League to the NPB Central League. Say; $\beta_{CL} = (.1813, -.1664)$. The corresponding resulting log-odds for that player would be $\beta_0 + \beta_{CL} = (-3.4456, -2.3601)$. The Central League effect thus shifts the outcome probability vector to be on average $(0.028, 0.084, 0.888)$. In a 500 plate appearance season, this would translate to approximately 2 more home runs and 7 fewer walks.

Tables 3.3 through 3.6 describe the full framework of parameter settings, including the fixed effects, covariance matrices, and first four random effects (corresponding to the first four players in the data set).

Table 3.3.
Within player career covariance matrix, Φ, setting for three-category simulation study, based on MLN model fit to real data

log-odds	$log\left(\frac{HR}{OTHER}\right)$	$log\left(\frac{BB}{OTHER}\right)$
$log\left(\frac{HR}{OTHER}\right)$	0.2032	0.0658
$log\left(\frac{BB}{OTHER}\right)$		0.0945

Table 3.4.
Across player-season covariance matrix, Σ, setting for three-category simulation study, based on MLN model fit to real data

log-odds	$log\left(\frac{HR}{OTHER}\right)$	$log\left(\frac{BB}{OTHER}\right)$
$log\left(\frac{HR}{OTHER}\right)$	0.1006	0.0389
$log\left(\frac{BB}{OTHER}\right)$		0.1681

Table 3.5.
Fixed effects, β, settings for three-category simulation study, based on MLN model fit to real data

parameter	**vector value**
β_0	$(-3.6269, -2.1937)$
β_{height}	$(0.0992, 0.0006)$
β_{weight}	$(0.1537, 0.0366)$
β_{age}	$(0.2041, 0.1955)$
β_{age^2}	$(-0.1201, -0.1070)$
$\beta_{hand=left}$	$(0.0519, -0.1167)$
$\beta_{hand=switch}$	$(0.0715, -0.2479)$
$\beta_{league=CL}$	$(0.1813, -0.1664)$
$\beta_{league=NL}$	$(-0.2030, -0.0377)$
$\beta_{league=PL}$	$(0.2915, -0.0248)$

Table 3.6.
First four random effects, ψ_i's, settings for three-category simulation study, simulated via $\psi_i \sim N_2(0, \Phi)$

player random effect	**vector value**
ψ_1	$(0.2125, 0.3929)$
ψ_2	$(-0.0993, 0.0761)$
ψ_3	$(0.3194, -0.0810)$
ψ_4	$(0.9468, 0.1454)$

Results

For the 500 players simulated in the data set, consisting of 4159 player-seasons, both algorithms showed convergence of the log-likelihood within 10000 iterations and showed adequate mixing after discarding burn-in and thinning the posterior chains.

We begin by assessing parameter estimation under the multinomial logistic-normal (MLN) model fits for the MLN data and multinomial logit (M-Logit) data. It is important to note that minor shifts in log-odds have even less an impact on the resulting probability vector. This is apparent in the recovery of the fixed intercept term. The graphs of the joint posterior distribution show that the MLN algorithm recovers the intercept term well based on MLN data, but does not do so as well with M-Logit data. A slight bias in the recovery of the parameters however, does not lead to a great deal of difference in recovery when transformed to the simplex (Figures 3.1 and 3.2).

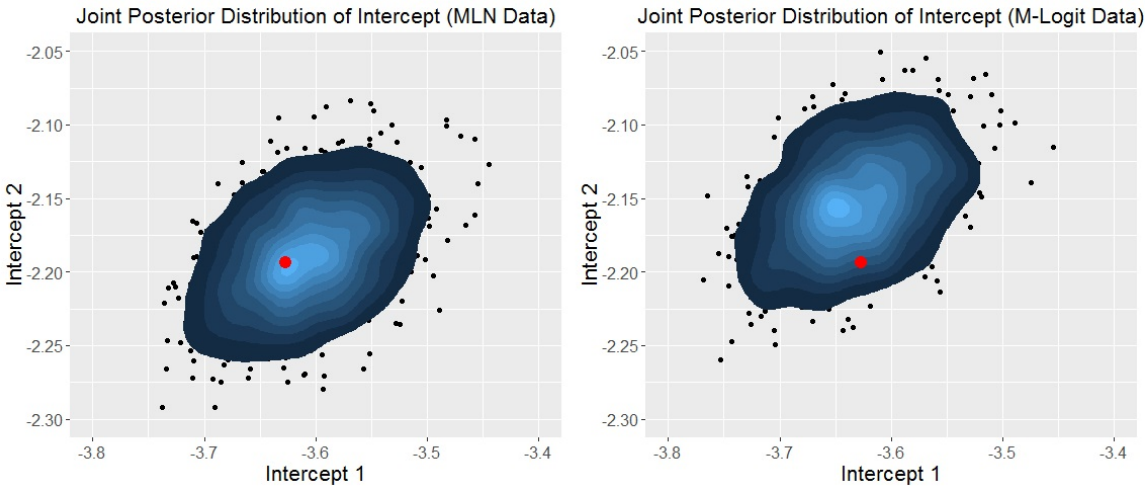

Figure 3.1. Recovery of fixed intercept for MLN and M-Logit data under MLN model fit, showing contours based on MCMC samples of the joint posterior distribution. The red point represents the true value of β_0 as defined in the parameter settings section

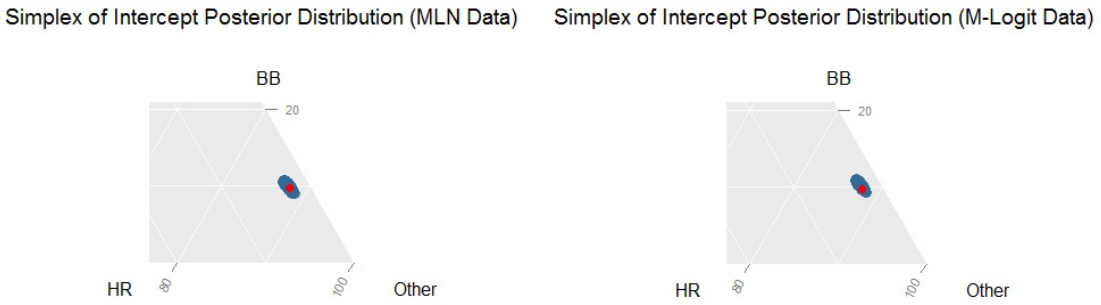

Figure 3.2. Recovery of fixed intercept, transforming posterior samples from log-odds to probability scale, for MLN and M-Logit Data under MLN model fit. The red dot represents the true values of the baseline probability vector, as defined in the parameter settings section.

Table 3.7 reports the recovery in terms of comparing true values to the posterior means of the MCMC samples (after discarding 35% burn-in) for the fixed effects, β and covariance matrix terms, $(\sigma_1, \sigma_{12}, \sigma_2) \in \Sigma$ and $(\phi_1, \phi_{12}, \phi_2) \in \Phi$, for both the MLN and M-Logit data fit with the MLN model. The only apparent difference is in the recovery of the age effects in terms of the fixed effects. The Σ covariance matrix terms are also much smaller, as to be expected, for the M-Logit fit. Across ten different runs with different data sets (five each of MLN and M-Logit simulated data), there was a wide variability in the posterior mean of age effects recovered on the log-odds scale.

However, on the probability scale, comparing the home run rate age trend for a baseline player under the true parameter values versus the recovered MLN posterior means shows there is not a great deal of difference, except perhaps at the tails (Figure 3.3). Over five hundred plate appearances, the difference in rate for a 20 year old baseline player is approximately 1 home run, indicating that even with the relatively poor recovery, the recovered values do not shift the resulting probability from the truth a great deal. This may explain why the effect struggled to be recovered across different data sets.

To consider recovery of the covariance matrices on the probability scale, we can compare the implied correlation structure of the outcome counts, $W = (HR, BB, Other)$, based on Σ vs. $\hat{\Sigma}$, and Φ vs. $\hat{\Phi}$, as described in Section 2.2.2. First for the implied correlations in the count based on the true values:

$$Corr(W^{\Sigma}) = \begin{pmatrix} 1 & 0.07 & -0.29 \\ 0.07 & 1 & -0.97 \\ -0.29 & -0.97 & 1 \end{pmatrix} \quad Corr(W^{\Phi}) = \begin{pmatrix} 1 & 0.28 & -0.60 \\ 0.28 & 1 & -0.94 \\ -0.60 & -0.94 & 1 \end{pmatrix}$$

Table 3.7.
Comparing posterior means of MLN fit with true values for MLN and M-Logit simulated data

parameter	**true value**	**MLN post mean**	**M-Logit post mean**
β_0	(-3.63, -2.19)	(-3.61, -2.19)	(-3.63, -2.15)
β_{height}	(0.10, 0.00)	(0.05, -0.01)	(0.07, 0.01)
β_{weight}	(0.15, 0.04)	(0.15, 0.03)	(0.14, 0.01)
β_{age}	(0.20, 0.20)	(-0.05, -0.22)	(0.28, -0.08)
β_{age^2}	(-0.12, -0.11)	(0.13, 0.29)	(-0.20, 0.16)
$\beta_{hand=left}$	(0.05, -0.12)	(-0.09, -0.17)	(-0.05, -0.19)
$\beta_{hand=switch}$	(0.07, -0.25)	(-0.13, -0.35)	(-0.11, -0.33)
$\beta_{league=CL}$	(0.18, -0.17)	(0.16, -0.13)	(0.13, -0.15)
$\beta_{league=NL}$	(-0.20, -0.04)	(-0.13, -0.04)	(-0.10, -0.09)
$\beta_{league=PL}$	(0.29, -0.02)	(0.35, 0.00)	(0.36, -0.02)
$(\sigma_1^2, \sigma_{12}, \sigma_2^2)$	(0.11, 0.04, 0.17) MLN only	(0.12, .05, 0.17)	(0.06, 0.02, 0.03)
$(\phi_1^2, \phi_{12}, \phi_2^2)$	(0.20, 0.07, 0.09)	(0.19, 0.06, 0.08)	(0.19, 0.06, 0.09)

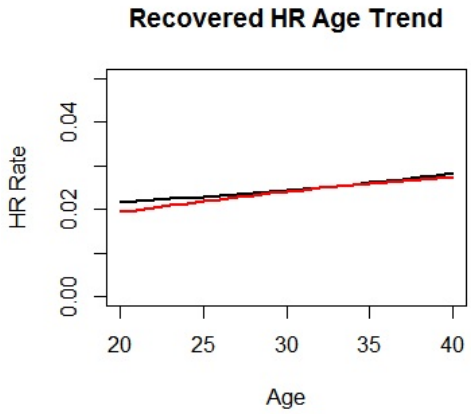

Figure 3.3. Baseline player home run rate age trend. Black line represents age trend based on posterior mean of β from MLN fit to MLN data, while the red line represents the age trend based on the true value of β

Then for the MLN data fit with MLN:

$$Corr(W^{\hat{\Sigma}}) = \begin{pmatrix} 1 & 0.11 & -0.34 \\ 0.11 & 1 & -0.97 \\ -0.34 & -0.97 & 1 \end{pmatrix} \quad Corr(W^{\hat{\Phi}}) = \begin{pmatrix} 1 & 0.29 & -0.59 \\ 0.29 & 1 & -0.95 \\ -0.59 & -0.95 & 1 \end{pmatrix}$$

And finally for the M-Logit data fit with MLN, where although Σ does not truly exist, the MLN model assumes it does:

$$Corr(W^{\hat{\Sigma}}) = \begin{pmatrix} 1 & 0.11 & -0.50 \\ 0.11 & 1 & -0.92 \\ -0.50 & -0.92 & 1 \end{pmatrix} \quad Corr(W^{\hat{\Phi}}) = \begin{pmatrix} 1 & 0.28 & -0.59 \\ 0.28 & 1 & -0.94 \\ -0.59 & -0.94 & 1 \end{pmatrix}$$

However, in order to properly interpret Φ, which builds correlation across years of a player's career, it is more relevant to look at the recovery of the block matrix with $\Sigma + \Phi$ on the diagonal and Φ on the off-diagonals; i.e. the covariance matrix of the unconditional log-odds (Equation 2.2) which we will denote Λ; and call the implied correlation structure of the corresponding counts, $Corr(W^{\Lambda}_{jj'})$. This larger matrix for the MLN fit is supplied in Appendix D.

Generally, the MLN model does a good job of recovering the correlation structure in the outcomes, both across player-seasons and within a player's career. While the individual covariance matrix terms may not be recovered perfectly (Table 3.7), our goal of accounting for a possibly positive correlation structure is achieved. For the M-Logit data, even with the MLN model estimating parameters, Σ, that were not used to generate the data, the relationship implied by Φ is recovered fairly well.

Comparing Model Fits

In this simulation study, we have separated the data into a training and test set of 500 and 105 players respectively. We can examine model fit with either of the data sets; the 4159 player-seasons of the training data set used to fit either the MLN or M-Logit models will have posterior predictive distributions we can compare observed values to. We can also use the test set of 105 player-seasons (focusing on those seasons where the players moved between leagues for the first time).

In this subsection, we conduct model diagnostics for the three-category simulation study on both the MLN and M-Logit data sets, both training and test sets. We begin by comparing the overall model fit of all four (MLN training, M-Logit training, MLN test, M-Logit test) data sets with both methods via the residual plots of the Dunn-Smyth type residual diagnostic detailed in Section 2.5.1.

The first set of residual plots (Figures 3.4 and 3.5) compare the MLN and M-Logit model fits to the 4159 player-seasons of the training data for both the MLN and M-Logit data archetypes. Under a good model fit, we expect these residuals to be standard normal. M-Logit data when fit with the incorrect MLN model shows some small lack of fit. The difference here is not terribly large, but the M-Logit data residuals tend to have a lighter right tail than the MLN data residuals. This is an indication that the observed quantiles for the M-Logit data are not large enough for those under MLN. The real difference comes in examining model fit of the data under the M-Logit fit, where the MLN data is clearly not a good fit.

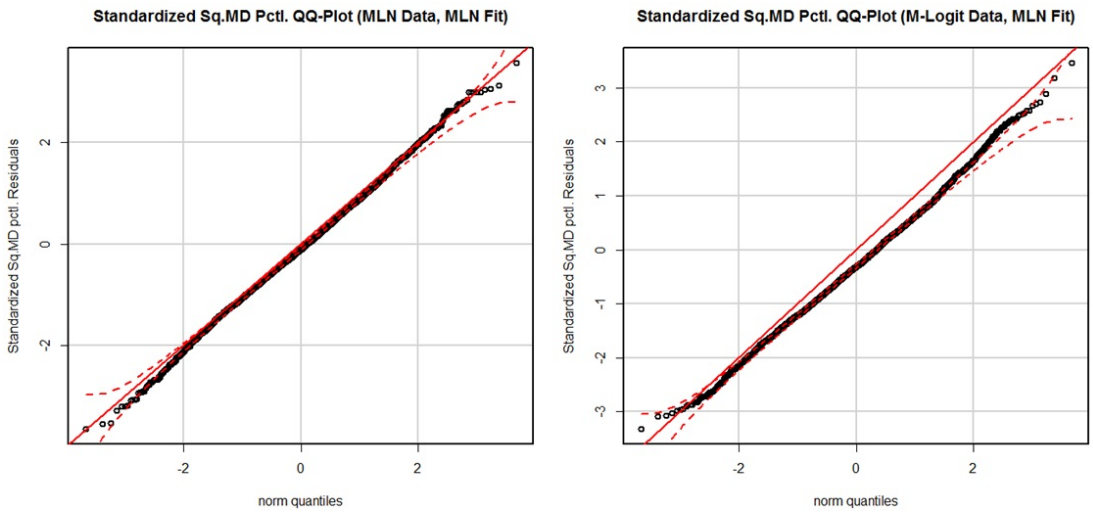

Figure 3.4. Standard normal QQ-plot of standardized MD^2 percentile residuals for MLN and M-Logit training data sets fit with MLN

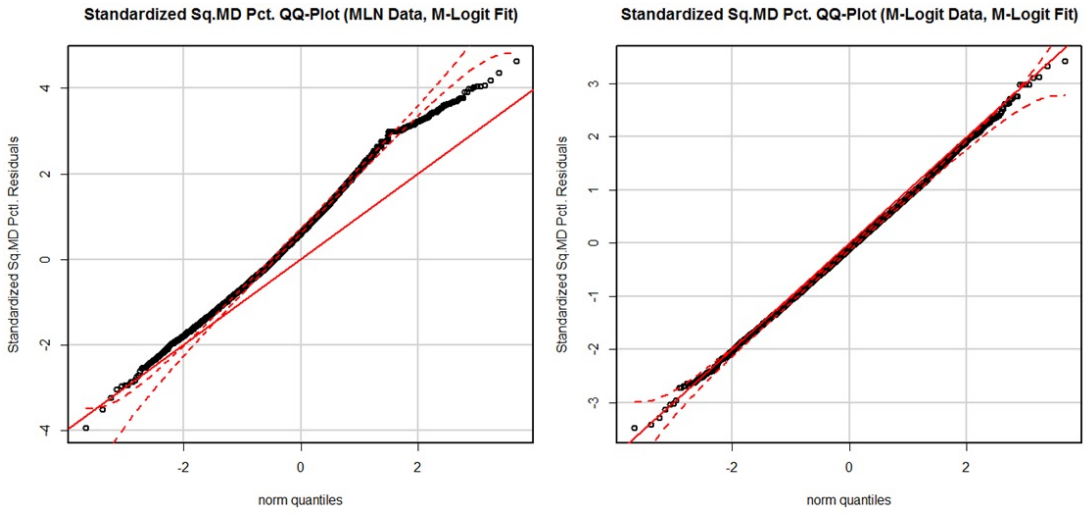

Figure 3.5. Standard normal QQ-plot of standardized MD^2 percentile residuals for MLN and M-Logit training data sets fit with M-Logit

While similar plots may be constructed for the 105 player-seasons to be predicted, due to the shrinkage in estimating $\psi_{i'}$ for players who did not have much career data, poorly predicted seasons will have an undue impact on the fit, whereas the QQ-plots for the overall model fit illustrate our method works well on aggregate.

The deviance information criterion, or DIC (Section 2.5.2) is a method for comparing model fit. Taking advantage of posterior chains of parameters, the DIC may be calculated for any model fit, and includes the log-likelihood based on the posterior mean of the parameters and an estimate of a model's effective number of parameters as part of its calculation. For both MLN and M-Logit training data sets, we calculate the posterior log-likelihood estimate, $l\left(W|\hat{\theta}\right)$, effective number of parameters, p_{DIC}, and DIC for each model fit, and present p_{DIC} and DIC (Tables 3.8 and 3.9).

Table 3.8.
Effective number of parameters, p_{DIC}, for training data simulated under MLN and M-Logit models and fit with both

p_{DIC}	**MLN Data**	**M-Logit Data**
MLN Fit	3764.99	2313.38
M-Logit Fit	736.20	711.97

Table 3.9.
Deviance information criterion, DIC, for training data simulated under MLN and M-Logit models and fit with both

DIC	**MLN Data**	**M-Logit Data**
MLN Fit	41661.35	36306.78
M-Logit Fit	48344.03	35297.63

A smaller DIC indicates a more appropriate model, and the results (Table 3.9) suggest that the most appropriate model for each data set is the true model under the data generating process. The additional parameters introduced by the presence of Σ (Table 3.8) in the MLN model are unnecessary for M-Logit data. The M-Logit model does not fit the MLN data particularly well, since it is unable to account for the additional variability that is present in the MLN.

We also calculated two more measures of predictive accuracy based on the model fit of the training data sets, Aitchison's R^2, or aR^2, and the Sum of Compositional Errors SCE. These results are presented in Appendix C for both the training and test sets. The main drawback behind these measures is that neither take into account the covariance structure; the key difference between the MLN and M-Logit models.

Model diagnostics for multinomial Bayesian hierarchical models will continue to be an open problem, but the Dunn-Smyth residual type diagnostic presented here does show promise in being able to discriminate between models and assess model fit in a multivariate context. The residual plots capture the difference in data having the additional uncertainty that the MLN model allows over that of the M-Logit.

Predictions and Uncertainty

Our main goal is to form predictions and uncertainty about those predictions for new player-seasons. First we examine recovery of the random effects for the 105 new players, the posterior samples of which were generated based on their careers up until the season they move and the posterior samples of the parameters from the initial run of the algorithm. We can see (Figure 3.6) that based on the posterior means of the random effects for the two log-odds, the recovery is not unreasonable, though clearly shrunk towards zero.

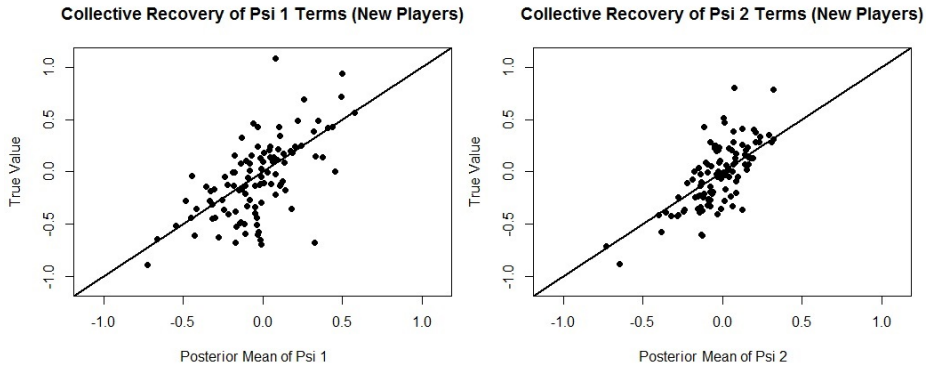

Figure 3.6. Recovery of $\psi_{i'1}$ and $\psi_{i'2}$ for new players $i' = 1, ..., 105$ in MLN test data, based on player's career to the point of switching leagues and posterior chains of other parameters under MLN model fit

As mentioned in Chapter 2, estimating the random effects of the held out players in the prediction set by using the posterior means of the fitted parameters of the initial run, instead of refitting them with the training data, saves time and does not greatly impact recovery of the random effects, based on Bland Altman plots of the player's random effects (Figure 3.7).

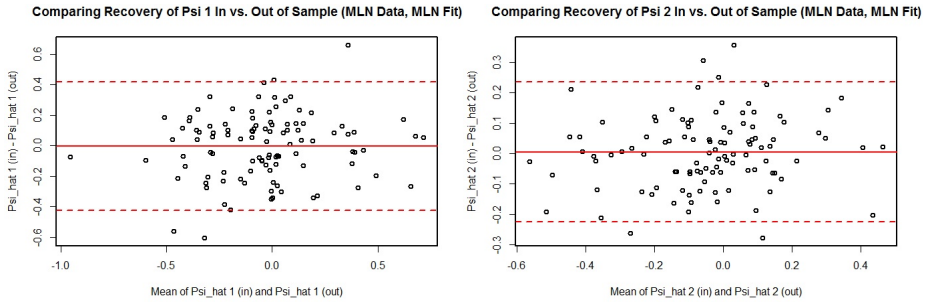

Figure 3.7. Bland Altman plots for posterior means of $\psi_{i'1}$ and $\psi_{i'2}$ for new players $i' = 1, ..., 105$ in MLN test data under (in): MLN fit to combined data set of training set and new players' careers before switching, and (out): Estimating new players' random effects based on posterior means of parameters from initial training set MLN fit

The posterior chain of the random effects is then used in tandem with the posterior chains of the parameters β, Σ, Φ to generate posterior predictive distributions. Sampling from the posterior predictive distributions for individual player-seasons is described in Section 2.4. Since there are 105 different players, we will focus first on two of them, with relatively similar covariate values yet different lengths of career. For players $i' = 1, 2$ we will be predicting the count vector for seasons $j' = 10$ and $j' = 4$, respectively (Tables 3.10 and 3.11).

Table 3.10.
Player 1's career up to and including moving from MLB to NPB

j'	**height (cm)**	**weight (kg)**	**age**	**hand**	**league**	**HR**	**BB**	**Other**
1	188	88	22	switch	AL	0	0	2
2	188	88	23	switch	AL	2	0	41
3	188	88	24	switch	AL	8	18	444
4	188	88	25	switch	AL	5	8	491
5	188	88	26	switch	AL	3	12	411
6	188	88	27	switch	AL	1	12	399
7	188	88	28	switch	AL	0	0	18
8	188	88	28	switch	NL	3	11	161
9	188	88	29	switch	AL	0	2	64
10	190	95	30	switch	CL	2	5	243

The two players are not a perfect match on covariates, but do share similarities. They are close in height and weight, the same age when switching leagues, both move from the American League to the Central League, and have similar numbers of

plate appearances in their first year in NPB. Where they drastically differ is in their histories, used to estimate their random effect; player 1 has 9 years of data, averaging 235 PA/year, while player 2 has only 3 years, averaging 18 PA/year.

Table 3.11.
Player 2's career up to and including moving from MLB to NPB

j'	**height (cm)**	**weight (kg)**	**age**	**hand**	**league**	**HR**	**BB**	**Other**
1	185	83	23	right	AL	0	3	9
2	185	83	24	right	AL	0	1	30
3	185	83	28	right	AL	0	2	9
4	189	92	30	right	CL	7	12	260

What we see in terms of the posterior predictive distributions is that the simplex of player 1's predicted counts shows a smaller spread than the simplex of player 2 (Figures 3.8 and 3.9), while the posterior predictive mean is also closer to the true observed count for player 1 (Tables 3.12 and 3.13).

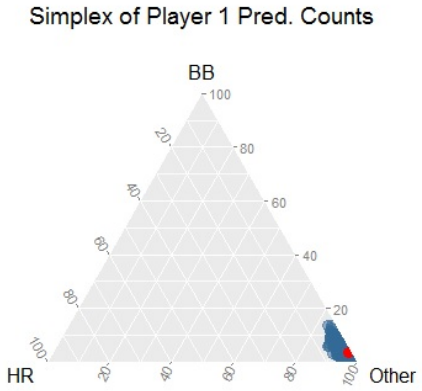

Figure 3.8. Posterior predictive samples of W, on the simplex, for player 1's 10th season, the red dot represents the true value

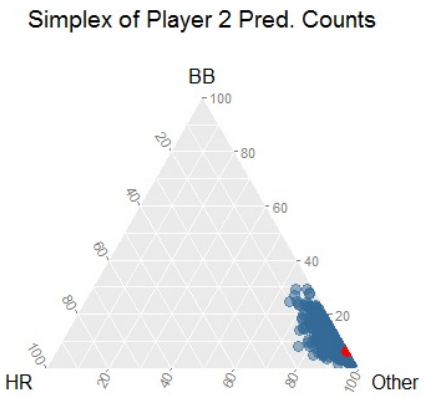

Figure 3.9. Posterior predictive samples of W, on the simplex, for player 2's 4th season, the red dot represents the true value

Table 3.12.
Player 1's 10th season posterior predicted mean versus observed counts

	HR	**BB**	**Other**
post pred mean	4.42	9.82	235.76
true value	3	9	238

Table 3.13.
Player 2's 4th season posterior predicted mean versus observed counts

	HR	**BB**	**Other**
post pred mean	9.37	26.97	242.65
true value	2	18	259

It is generally the case, though not a rule, that more data used to estimate the random effect results in less variability in the posterior predictive distribution. The random effect is only one source of variability. In this case, the players were similar enough in other aspects that the largest source of the disparity is the amount of information.

While the posterior predicted mean serves as the point prediction for the new player-season, we are also interested in the uncertainty about that prediction. From the plots it is easy to see that in comparing these two players, player 1 has more certainty about their prediction than player 2. However, it is useful to provide some quantitative measure of that uncertainty. For this, we construct two types of intervals about the posterior predicted mean, as described in Section 2.5.3. While interpreting the intervals, it is important to realize that the lower bound and upper bound will not necessarily sum to the exposure, as they are simply the marginal limits for the 95% most likely outcomes based on a joint predictive posterior distribution. The convex hull from which intervals will be formed contains combinations that sum to the number of plate appearances in the season being predicted.

The first type of interval is a prediction interval about the posterior predictive distribution. In some cases, especially when moving to additional categories, these intervals may be too wide to be useful for direct comparison between players, but do serve as a measure of predictive accuracy on a single player basis. Several papers have discussed that simultaneous confidence intervals about multinomial counts tend to be overly conservative (Quesenberry and Hurst, 1964; Sison and Glaz, 1995), and our hierarchical model adds additional variability. Construction of the interval proceeds as described in Section 2.5.3; 95% of the most likely posterior predictive count samples are chosen to form the convex hull that defines the interval. We do not present prediction intervals here, though they are provided in Appendix E. In terms of direct comparison of player's ability, it may be more useful to concern ourselves with the expected performance of the players, rather than the predicted count itself.

We can construct credible intervals from samples of the posterior predictive distribution of the log-odds for the new player-season. The credible intervals on the log-odds differ from the prediction interval on the counts in that are describing uncertainty about a player's expected performance. Narrower intervals can be more useful in direct comparison of players, as opposed to assessing predictive power. We present the 95% credible intervals for player 1 and 2 (Tables 3.14 and 3.15) which we can compare to their observed and posterior mean counts (Tables 3.12 and 3.13).

Table 3.14.
95% credible interval for the expected performance of player 1's 10th season

95% **cred int**	**HR**	**BB**	**Other**
lower bound	2	6	230
upper bound	8	14	241

Table 3.15.
95% credible interval for the expected performance of player 2's 4th season

95% **cred int**	**HR**	**BB**	**Other**
lower bound	3	14	224
upper bound	20	42	260

While the additional variability in player 2 is still obvious, it is now much simpler to draw direct comparisons between the two players based on the bounds of the credible intervals as compared to the prediction intervals. For a team looking for a player who can hit home runs or walk, player 1 does not represent a strong candidate. The upper bounds for those two categories on the credible interval represent rather low

numbers, especially for the walk rate, which is exactly the same as the lower bound of player 2. Player 1 had 9 years to establish their performance level, and in terms of home runs and walks it has been uninspiring. With the lower bounds for home run rate being nearly identical, the intervals suggest that player 2 may be worth the risk. In terms of actual performance, the home runs slightly favor player 1, while player 2 under-performed there but did walk at a much higher rate.

3.3.2 Ten-Category Simulation Study

We also completed a simulation run in a higher dimensional setting; data sets simulated with the dimensionality of the real data, $K = 10$. While the $K = 3$ setting was useful for examining the viability of the model in the most basic framework often, as in our case, more outcomes are of interest. The baseball outcomes for the ten-category setting are as described in Section 1.3, with abbreviations $(1B, 2B, 3B, HR, SH, SF, BB, HBP, SO, OIP)$. From these ten outcomes, most singular value metrics used in baseball may be calculated.

The process of simulation remains the same. Parameter settings are based on a trial run of the algorithm. Data are generated from both the MLN and M-Logit models. The only difference is in the dimensionality; for example, when simulating random effects for both data sets and error terms for the MLN data, they are now $\psi_i \sim N_9(0, \Phi)$ and $\varepsilon_{ij} \sim N_9(0, \Sigma)$, respectively. We use the same covariates as in the three-category case, with the same sample of 500 players representing the training data sets.

Results were much in line with those from the three-category case. The algorithms converged within 10000 iterations, with decent mixing after discarding burn-in and thinning of the chains. Parameters tended to be reasonably recovered, and the DIC measure indicated that the data set generated under each of the models was more appropriately fit by the corresponding model. Because of the quite large increase in

parameters, from 26 to 180 (not including random effects and the log-odds), we do not present all the results here. Computation time increased a small amount over that of the three-category simulation study.

3.4 Real Data Results

As with the simulation studies in the previous section, we fit both a lower dimensional, three-category model to the real data as well as a higher dimensional ten-category model. Both the multinomial logistic-normal and multinomial logit mixed-effects models are considered. In addition to the covariates used in the simulation studies, we also include position as a covariate. Since some of the player-seasons do not distinguish player's primary position outside of infield, outfield, designated hitter or pitcher, those are the four levels of the covariate we utilize.

For the three-category model, we again consider $(HR, BB, Other)$, since we expect home runs and walks to be positively correlated. When discussing results for the ten-category model, we occasionally narrow in on the relationship between HR and BB for comparison purposes. In both cases, we will discuss interpretation of the intercept and covariate effects.

This section will proceed by describing and discussing results from the three-category analysis first, followed by the ten-category. In each subsection, we assess the posterior log-likelihood convergence for the data under both the multinomial logistic-normal (MLN) and multinomial logit (M-Logit) models and for both levels of dimensionality, as well as discussing Markov chain mixing and parameter estimation of the same. We then compare model fit of the MLN and M-Logit fits for both levels of dimensionality, and finally assess predictive performance and discuss utility of the credible intervals about the predictions.

3.4.1 Three Categories

The three-category real data analysis produced results not too dissimilar from the simulation study. Under the three-category setting, the multinomial logistic-normal model fit the data well, and was a better fit than the multinomial logit model. As in the simulation study, we separate the data into 500 players and 4159 player-seasons which will be used as a training set to fit the model, and 105 out of sample players, whose 105 player-seasons corresponding to their first year playing in a different league will be used as a test set.

Convergence and Model Fit

Both the MLN and M-Logit model fits to the three-category training data set appeared to converge well before 10000 iterations (Figure 3.10). After discarding 30% burn-in and thinning the chains, both MCMC chains show decent mixing, though the MLN algorithm mixes slightly better than the M-Logit.

To assess and compare model fits, we use the same metrics as in the simulation study. We provide standardized squared Mahalanobis distance percentile residual plots of the MLN fit, and calculate the deviance information criterion (DIC) under both MLN and M-Logit fits.

The residual QQ-plot for the training data (Figure 3.11), shows a relatively good fit to the MLN model, although there is a very slight tendency for smaller than expected percentiles/residuals. On the other hand the fit to the M-Logit model (Figure 3.12) shows larger than expected residuals and more deviation from the expected trend.

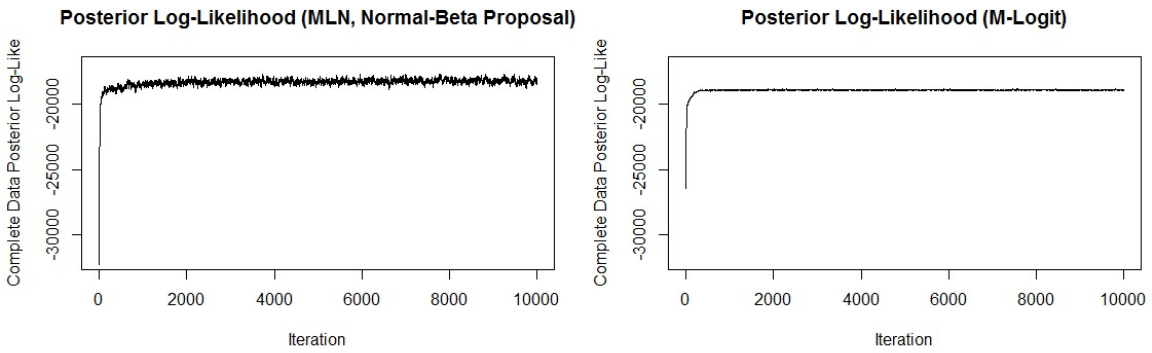

Figure 3.10. Convergence of complete data posterior log-likelihood of three-category real data fit with both MLN and M-Logit models

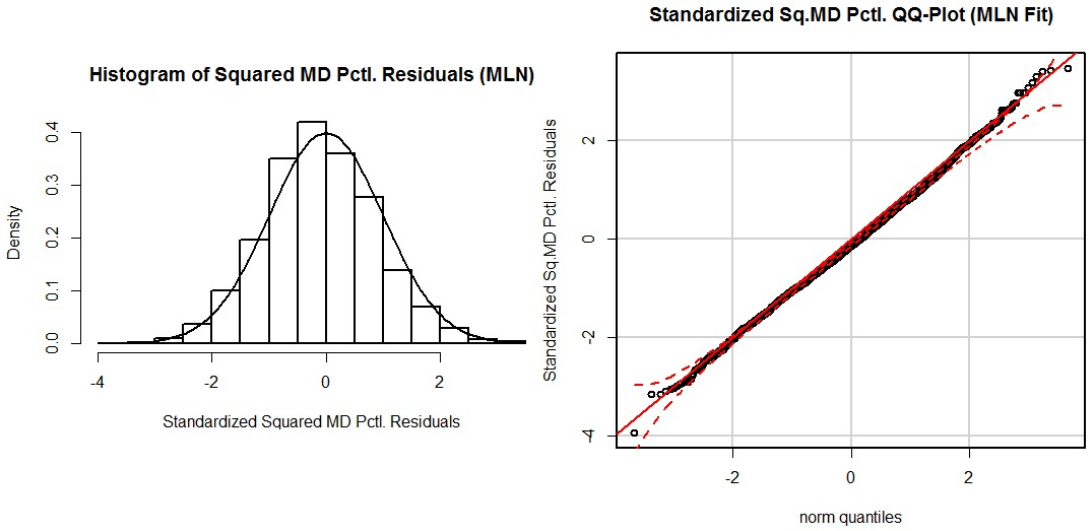

Figure 3.11. Histogram and standard normal QQ-plot of standardized MD^2 percentile residuals for MLN fit of three-category real data set

Table 3.16.
Effective number of parameters and deviance information criterion for three-category real data training set fit under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
p_{DIC}	3014.15	682.76
DIC	37229.52	37808.39

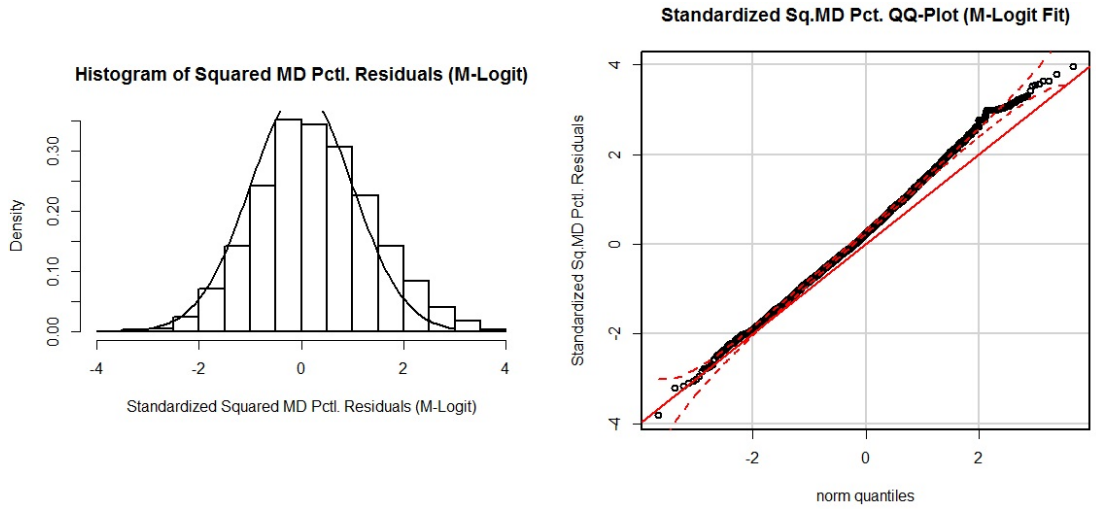

Figure 3.12. Histogram and standard normal QQ-plot of standardized MD^2 percentile residuals for M-Logit fit of three-category real data set

In addition, a comparison on models using DIC shows the MLN is a better fit. While the MLN model has a significantly larger effective number of parameters, the improved fit means DIC favors MLN over M-Logit (Table 3.16). Because the MLN model is favored under both assessments, we only summarize results for that fit in the next subsection.

Estimation and Prediction

In terms of parameter estimates, some of the results are quite a bit different from what we saw in the simulation study (Table 3.17). The largest covariate effect in magnitude is the indicator variable for the pitcher position, which makes some sense; the coefficient is negative and pitchers are generally much worse hitters than position players. Age and switching to the NPB leagues are the next largest effects based on posterior means. However, examining the 95% credible ellipsoids around each of the effects, four are not significantly away from zero: the Age effects, NL effect, and DH effect (Figure 3.13). The rest of the effects all had ellipsoids which did not include zero, with the pitcher effect, CL and PL effects being the farthest from $(0, 0)$, respectively.

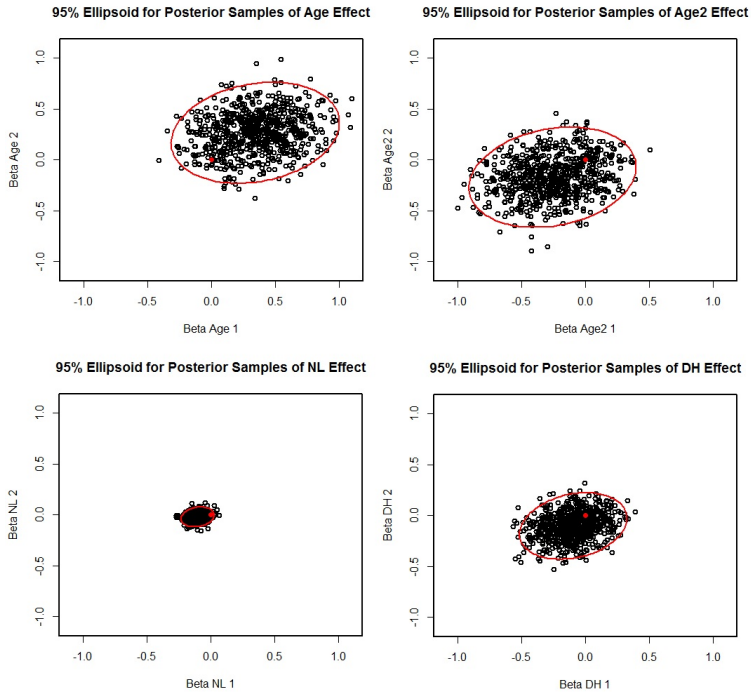

Figure 3.13. Fixed effects with 95% credible ellipsoids including zero for MLN fit of MLN simulated training data

Table 3.17.
Posterior means of parameters under MLN model fit, discarding first 30% of samples, thinning every 10^{th}

parameter	**MLN post mean**
β_0	$(-3.60, -2.41)$
β_{height}	$(0.14, 0.02)$
β_{weight}	$(0.16, 0.05)$
β_{age}	$(0.34, 0.27)$
β_{age^2}	$(-0.26, -0.17)$
$\beta_{hand=left}$	$(-0.06, 0.14)$
$\beta_{hand=switch}$	$(-0.15, 0.22)$
$\beta_{league=CL}$	$(0.32, -0.18)$
$\beta_{league=NL}$	$(-0.10, -0.02)$
$\beta_{league=PL}$	$(0.32, -0.03)$
$\beta_{pos=of}$	$(0.11, -0.06)$
$\beta_{pos=dh}$	$(-0.10, -0.10)$
$\beta_{pos=p}$	$(-1.93, -0.95)$
$(\sigma_1^2, \sigma_{12}, \sigma_2^2)$	$(0.115, 0.037, 0.060)$
$(\phi_1^2, \phi_{12}, \phi_2^2)$	$(0.127, 0.039, 0.087)$

Interpreting the covariate effects is as simple as transforming from the log-odds scale to the probability scale. We first look at the intercept in order to understand the baseline player's expected outcome composition. A baseline player is one of average height, weight and age, who bats right handed in the AL and is an infielder:

$$\beta_0 = (-3.60, -2.41) \implies$$
$$(HR_0, BB_0, Other_0) = (0.024, 0.081, 0.895)$$

Over 500 plate appearances, this means that a baseline player is expected to hit 12 HR and receive 40 walks.

Table 3.18.
Estimated covariate effects on the baseline player in terms of expected perturbation from baseline over 500 plate appearances for a one unit increase in each covariate. For height and weight, one unit is one standard deviation above the mean height or weight

parameter effect	**HR**	**BB**	**Other**
baseline	12	40	448
height	14	41	445
weight	14	42	444
left	11	46	443
switch	10	50	440
CL	17	34	449
NL	11	40	449
PL	17	39	444
OF	14	38	448
DH	11	37	452
P	2	17	481

Table 3.18 summarizes the covariate effects (except for age), in terms of deviation from a baseline player (based on a one unit increase in the covariate). We see the covariate that is most directly related to our problem of predicting player performance moving between NPB and MLB, League, implies that the AL and NL are more difficult leagues to succeed in (in terms of HR and BB) as a hitter. We can also see the dramatic reduction in HRs and BBs for a pitcher. Age effects for the baseline player is presented as a plot of expected HRs over the range of ages (Figure 3.14).

Figure 3.14. Posterior distribution (grey) and mean (black) of baseline player home run rate age trend per 500 plate appearances for MLN fit of three-category real training data set

We next examine the implied correlation structure of the outcomes based on the two covariance matrices. Since the covariates (mean) also factor into these covariances (see Section 1.4), we consider the implied correlation matrices for a baseline player (using the intercept as the mean). Using the posterior mean of each element, the correlation matrices are:

$$Corr(W^{\hat{\Sigma}}) = \begin{pmatrix} 1 & 0.21 & -0.58 \\ 0.21 & 1 & -0.92 \\ -0.58 & -0.92 & 1 \end{pmatrix} \quad Corr(W^{\hat{\Phi}}) = \begin{pmatrix} 1 & 0.16 & -0.52 \\ 0.16 & 1 & -0.93 \\ -0.52 & -0.93 & 1 \end{pmatrix}$$

To better understand the correlation structure both within and across outcomes, we combine the Σ and Φ matrices for a pair of outcomes. This results in a block matrix with $\Sigma + \Phi$ on the diagonal and Φ on the off-diagonals (Equation 2.2). We denote this matrix Λ; and call the implied correlation structure of the corresponding counts, $Corr(W^{\Lambda}_{jj'})$. The implied temporal correlation structure recovers positive correlations

between home runs and walks both within ($\rho = 0.22$) and between ($\rho = 0.11$) seasons for a baseline player.

$$Corr(W_{jj'}^{\hat{\Lambda}}) = \begin{pmatrix} 1 & 0.22 & -0.43 & 0.38 & 0.11 & -0.23 \\ 0.22 & 1 & -0.70 & 0.11 & 0.48 & -0.41 \\ -0.43 & -0.70 & 1 & -0.23 & -0.41 & -0.04 \\ 0.38 & 0.11 & -0.23 & 1 & 0.22 & -0.43 \\ 0.11 & 0.48 & -0.41 & 0.22 & 1 & -0.70 \\ -0.23 & -0.41 & -0.04 & -0.43 & -0.70 & 1 \end{pmatrix}$$

It is important to keep in mind the purpose of this methodology is prediction. For assessing prediction, we have approached this real data analysis from a cross validation type framework. While the model estimation is done with 500 randomly selected players, the main intention is to predict the counts for players moving to a new league in the midst of their careers. The 105 players left out of the model estimation step will thus have their careers split, with their careers leading up to their changing leagues used to estimate their random effects and their first year in a new league used to assess predictive accuracy. All seasons in a player's career after their initial season in a new league (either MLB or NPB) are discarded.

Practically, this approach mirrors the process of predicting for a group of players who might be changing leagues in the ensuing season. Given a model fit, random effects for any number of players potentially moving may be estimated and their performance in the ensuing season predicted. For NPB and MLB teams deciding which players in the opposing league to pursue, this process makes the most sense.

We apply the predictive assessment techniques described in Chapter 2 to the real data in both three and ten-category frameworks, to illustrate our methods use in pre-

dicting baseball player performance for players moving between the leagues. We have already seen that the Dunn-Smyth type, standardized squared Mahalanobis distance percentile, residuals for the out of sample players appear to approximately follow the standard normal distribution, lending some validity to the predictions.

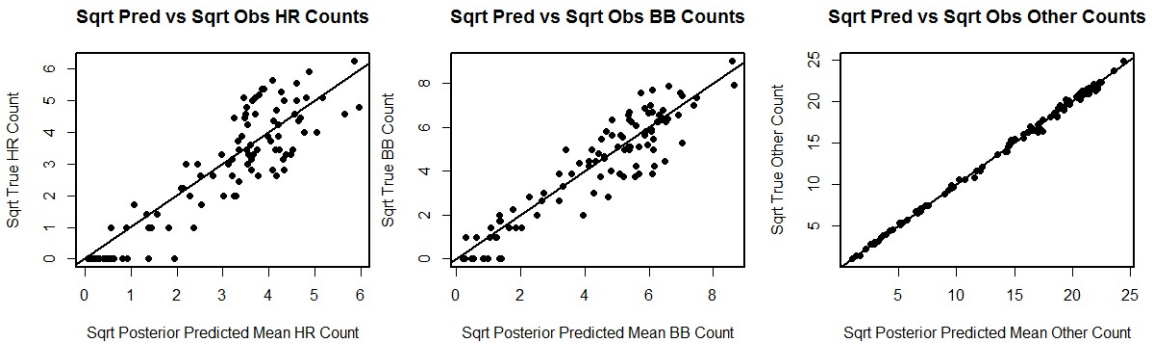

Figure 3.15. Predicted versus observed counts of each category for out of sample players from MLN fit

In terms of predictive accuracy, while not a perfect metric by any means, we may plot the square root of the posterior predictive mean of the new player-seasons counts versus the square root of their true value (Figure 3.15). This has been done to assess predictions in similar contexts (Xia et al., 2013). Taking the square root helps visualize the recovery of small counts which might otherwise look well recovered due simply to low exposure. We see that the baseline outcome, $OTHER$, is seemingly predicted well across the board, while home runs and walks are not unreasonable, though struggle with predicting very small counts. While this is encouraging, we are more interested in where the observed points fall in the respective posterior predictive distributions.

It is also useful to look at individual players. The two players whose covariates were used as examples from the test set in the simulation study were Mike Young (player 1) and Rod Allen (player 2), chosen because of the similarity in covariates but dissim-

ilarity in career length and information prior to switching leagues. Young played nine years as an outfielder in MLB, mostly for the Baltimore Orioles in the AL, averaging 235 PA/year, before moving to the Hiroshima Carp in the CL in 1990 at age 30. Rod Allen played three years in the American League mostly as an outfielder, averaging 18 PA/year, before moving to the CL to play for the Carp in 1989, at the age of 30. In 1989, Allen hit 11 home runs and 25 walks in 279 plate appearances, while in 1990 Young hit 11 home runs and walked 26 times in 250 plate appearances. We sampled from both these player's posterior predictive distributions for their first year in the Central League, and constructed 95% credible intervals about their expected performance.

Looking at the posterior predictive distribution of counts, it is not readily apparent from the plots whether Rod Allen's 1989 season or Mike Young's 1990 season has a larger spread (Figures 3.16 and 3.17). However, their 95% credible intervals are more informative (Tables 3.19 and 3.20), and we see that the Young's credible intervals is much tighter than Allen's, reflecting how much more information went into predicting his performance compared to Allen. Both intervals covered the true values.

Table 3.19.
95% Credible interval, posterior mean, and observed value for Mike Young's 1990 season, based on MLN fit of three-category real test data

	HR	**BB**	**Other**
CI lower bound	8	18	203
post pred mean	12.64	25.17	212.19
true value	11	26	213
CI upper bound	18	32	222

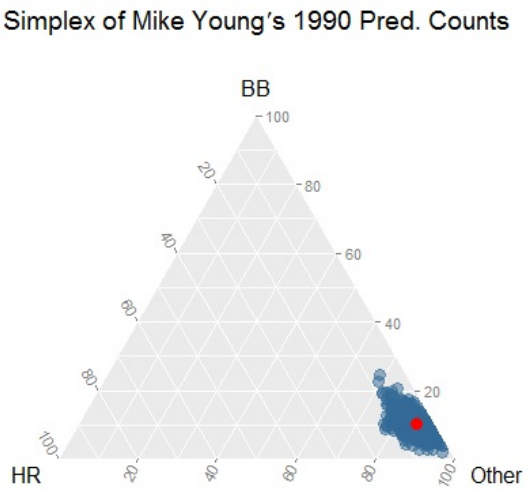

Figure 3.16. Posterior predictive samples of W, on the simplex, for Mike Young's 1990 season, based on MLN fit of three-category real test data, the red dot representing the true value

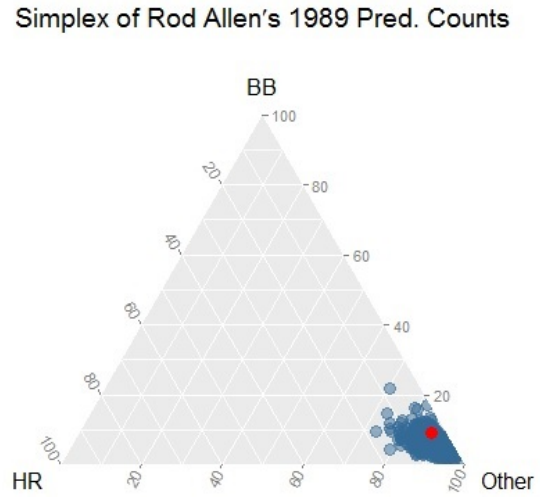

Figure 3.17. Posterior predictive samples of W, on the simplex, for Rod Allen's 1989 season, based on MLN fit of three-category real test data, the red dot representing the true value

Table 3.20.
95% Credible interval, posterior mean, and observed value for Rod Allen's 1989 season, based on MLN fit of three-category real test data

	HR	**BB**	**Other**
CI lower bound	6	10	234
post pred mean	13.38	17.78	247.84
true value	11	25	243
CI upper bound	22	30	261

The main goal of these credible intervals is for player comparison. What we see here is that, based on the bounds of the intervals, a team might be more likely to choose Young over Allen; in fewer plate appearances, Young gives higher floors in both HR and BB, and a higher ceiling in BB with a similar ceiling in HR on a rate basis, when compared to Allen. In this case, focusing on three categories, the choice of Young over Allen would be justified; though their numbers look similar, Young hit the same number of home runs and walked once more than Allen, in 29 fewer plate appearances. While this is only one example, it illustrates the power of having this form of uncertainty assessment about a prediction.

3.4.2 Ten Categories

While the three-category study shows promise, we are interested in predicting not only a player's counts for a season, but also statistics, such as batting average or OPS, from those counts. Moving to the ten categories identified in Section 1.3, allows us to do this. We produce much of the same output as in the three-category study, except due to the increase in dimensionality, some of the plotting becomes more difficult. For example, it is not possible to plot all ten outcomes on a simplex, though groups of three may be plotted at a time.

Convergence and Model Fit

Both complete data log-likelihoods converge quickly. One thing of note is how quickly the MLN fit seemed to converge compared to the M-Logit, though to a lower value; the MLN fit reached its maximum posterior log-likelihood at iteration 304 and levels off. One explanation is the MLN model being sensitive to tuning parameters, such as step-size for the Metropolis proposal step. While the average acceptance ratio over the iterations is not unreasonable (Figure 3.19), a higher rate, and thus more exploration of the distribution space, might allow the log-likelihood under MLN to converge to something higher.

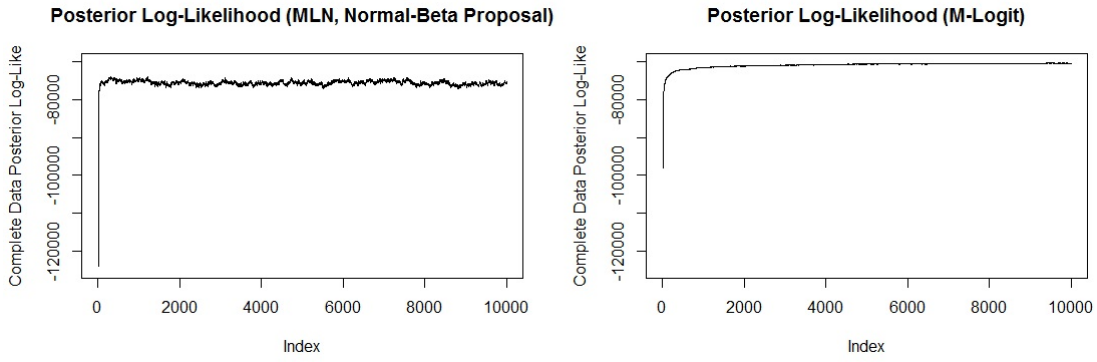

Figure 3.18. Convergence of complete data posterior log-likelihood of ten-category real data fit with mixed effects MLN Metropolis-within-Gibbs algorithm

The ten-category results are less conclusive, but generally favor the M-Logit fit for the data. First we examine the residual plots under the MLN fit based on our Dunn-Smyth type residuals (Figure 3.20). In this case we see a much lighter right tail, more in line with the M-Logit simulated data results, where now the pointwise confidence envelope falls entirely below the reference line. However, the M-Logit fit (Figure 3.21) does not look particularly good either, showing the same trend as MLN data fit with M-Logit as in the simulation study. These results would to indicate that the data is neither quite MLN or M-Logit in nature.

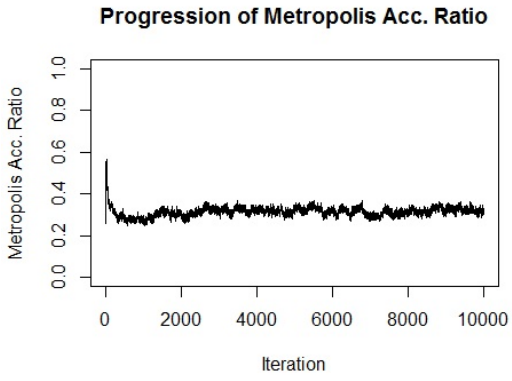

Figure 3.19. Average Metropolis acceptance ratio for all player-seasons log-odds (latent variables, Y_{ij}'s) over the MLN algorithm, ten-category real data training set

The DIC supports the M-Logit as the better model (Table 3.21). The MLN model in this case involves many more covariance parameters in Σ. With ten categories, this produces over four times the number of effective parameters. If the additional variability is relatively small in some outcome categories, or there are several weak relationships among the categories, the added model complexity is not worth it.

Table 3.21.
Effective number of parameters and deviance information criterion for ten-category real data training set fit under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
p_{DIC}	10439.57	2225.30
DIC	145797.5	141057.0

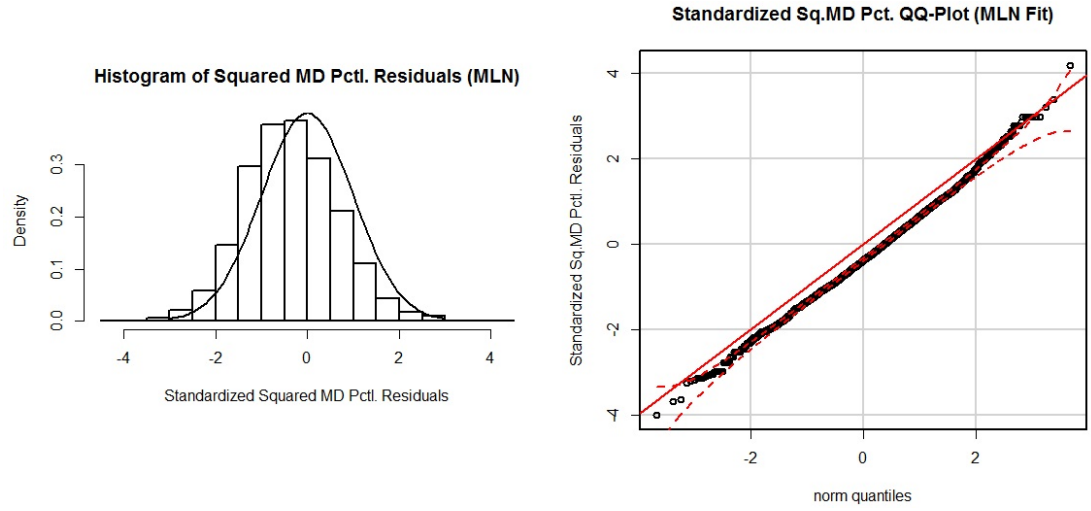

Figure 3.20. Histogram and standard normal QQ-lot of standardized MD^2 percentile residuals for MLN fit of ten-category real data set

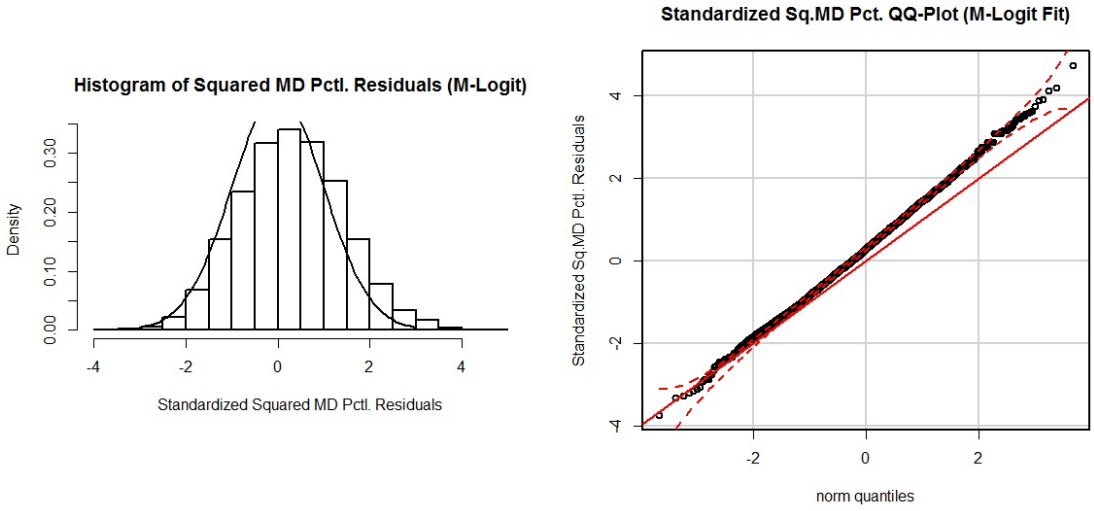

Figure 3.21. Histogram and standard normal QQ-lot of standardized MD^2 percentile residuals for M-Logit fit of ten-category real data set

However, based on the estimation of Σ, discussed in the next subsection, most of the covariance terms recovered have posterior distributions that do not include zero. Since we adopt the MLN model to help describe the correlation among outcomes, it will also be useful to investigate what the covariance matrices imply about those correlations. In the next section, we see there appears to be positive associations to be recovered between outcomes.

Since DIC suggests the M-Logit fits the training data better, but the residual plots are less conclusive, we will continue comparison of the models through the estimation and prediction results.

Estimation and Prediction

With $p = 13$ (including the intercept) and $d = 9$, there are 117 covariate effects for each of the model fits. Since it is unreasonable to delve into too much detail, we will simply assess the recovery of the fixed effects in terms of the intercept and two covariates, League and Age. Both 13×9 β fixed effect matrices for the MLN and M-Logit fits to ten-category data are available as tables in the Appendix D.

Table 3.22 shows the intercept vector and Central League effect vector for both model fits. The intercepts imply similar probability vectors for a baseline player, and not too dissimilar implied values for HR and BB when compared with the three-category results. Over 500 plate appearances, a baseline player in the MLN model will have on average the following counts: $(1B, 2B, 3B, HR, SH, SF, BB, HBP, SO, OIP) = (76, 21, 2, 12, 2, 4, 41, 3, 86, 253)$, and under the M-Logit model, the baseline player's expected performance is almost the same: $(1B, 2B, 3B, HR, SH, SF, BB, HBP, SO, OIP) = (78, 21, 2, 12, 2, 4, 39, 3, 89, 250)$.

Table 3.22.
Posterior mean of intercept and CL fixed effects over both the MLN and M-Logit algorithms, ten-category real data training set

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
MLN Intercept	-1.21	-2.49	-4.99	-3.01	-5.10	-4.24	-1.82	-4.43	-1.08
MLN CL	0.03	-0.12	-0.49	0.27	-0.35	-0.29	-0.19	0.14	0.04
M-Logit Intercept	-1.17	-2.48	-4.79	-3.04	-4.73	-4.23	-1.85	-4.45	-1.03
M-Logit CL	0.06	0.07	-0.39	0.60	-0.42	-0.08	0.04	0.43	0.10

The Central League effect is similar to what it was in the three-category results; moving from the AL to CL provides an expected increase in home runs. However, the overall performance increase is not as obvious with ten categories, since in both model fits, there is a corresponding increase in expected strikeouts, and decrease in triples. The corresponding expected outcomes over 500 plate appearances for baseline players moving from AL to CL are, for MLN: $(78, 19, 1, 16, 1, 3, 34, 3, 90, 254)$, and for M-Logit: $(78, 21, 1, 21, 1, 3, 39, 4, 93, 238)$. One way to summarize the results is with OPS. A baseline AL player under MLN would be expected to have an OPS of 0.720 which would increase only a little to 0.723 in the CL. For the M-Logit, the baseline OPS is 0.721 which increases more drastically to 0.812 in the CL.

Since age also plays an important role in our model, we will look at what the quadratic age term implies about the trajectory of home runs and strikeouts over a baseline player's career in both models. The age effect on HR was almost exactly the same as in the three-category models, and both models imply a convex shape for age's impact on strikeouts, but MLN estimates a much steeper overall decline in strikeouts from age 20 to age 40 (Figures 3.22 and 3.23).

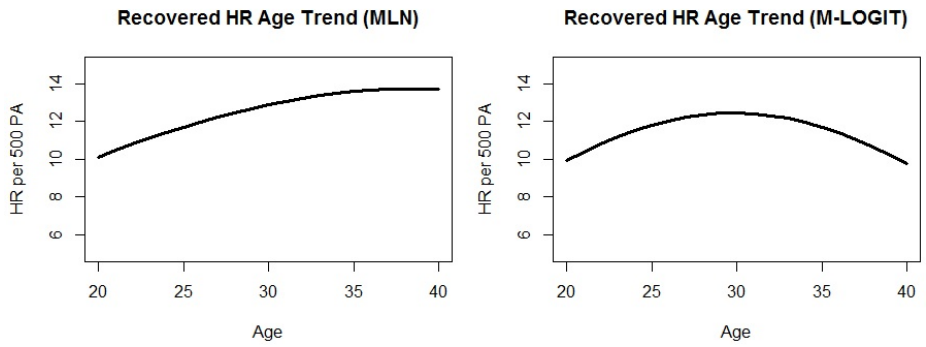

Figure 3.22. Comparison of baseline player home run rate age trend per 500 plate appearances for MLN and M-Logit fits of ten-category real training data set

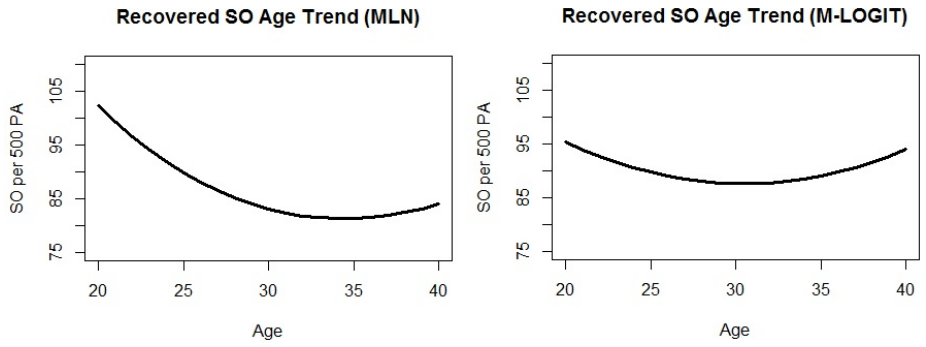

Figure 3.23. Comparison of baseline player strikeout rate age trend per 500 plate appearances for MLN and M-Logit fits of ten-category real training data set

Instead of reporting all the covariance matrices, we will focus first on the recovered posterior mean of the Σ matrix from the MLN fit, and then look at some implied correlations among counts for both fits. By focusing on Σ, we can hone in on the only difference between MLN and M-Logit simulated data, and perhaps gain some understanding of which model better suits this data. DIC prefers the M-Logit fit, possibly due to some of the small in magnitude covariance terms we see in the MLN

fit for Σ, and the presence of random effects allows some variability that Σ might have described to be explained by Φ. However, 26 of the 36 Σ covariance terms posterior distributions did not include zero. This gives evidence in favor of Σ perhaps providing some benefit to helping describe the covariance structure of the data.

$$\hat{\Sigma} = \begin{pmatrix} 0.02 & 0.01 & 0.02 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.00 \\ & 0.05 & 0.04 & 0.04 & -0.01 & 0.03 & 0.02 & 0.04 & 0.01 \\ & & 0.27 & 0.04 & 0.09 & 0.06 & 0.02 & 0.04 & 0.01 \\ & & & 0.15 & -0.06 & 0.07 & 0.05 & 0.09 & 0.03 \\ & & & & 0.57 & -0.01 & -0.02 & 0.01 & 0.02 \\ & & & & & 0.14 & 0.03 & 0.06 & 0.01 \\ & & & & & & 0.08 & 0.04 & 0.03 \\ & & & & & & & 0.29 & 0.03 \\ & & & & & & & & 0.06 \end{pmatrix}$$

In terms of the implied correlation structure in the counts, the above matrix results in several implied positive correlations among counts. We can investigate this for both fits. For sake of brevity, we will focus in on a select few relationships that are implied by the recovered covariance matrices, again only for a baseline player. First, to compare against the three-category results, we will look at the recovered implied relationship between HR and BB for each fit, with MLN fit correlations on the left, M-Logit on the right:

$$Corr(W_{HR}^{\hat{\Sigma}}, W_{BB}^{\hat{\Sigma}}) = 0.19$$

$$Corr(W_{HR}^{\hat{\Phi}_{MLOGIT}}, W_{BB}^{\hat{\Phi}_{MLOGIT}}) = 0.21$$

$$Corr(W_{HR}^{\hat{\Phi}_{MLN}}, W_{BB}^{\hat{\Phi}_{MLN}}) = 0.11$$

$$Corr(W_{jHR}^{\hat{\Lambda}}, W_{jBB}^{\hat{\Lambda}}) = 0.19$$

The implied correlation is diminished somewhat, but not dissimilar to the three-category results. In the absence of Σ, the M-Logit algorithm relies on the random effects covariance matrix to account for that additional variability, and it does tend to be larger under the M-Logit fit than the MLN fit. We can look at other pairwise correlations of interest, and determine if they make sense in the context of our data.

Another major pair of outcomes which are commonly thought to be relatively strongly positively correlated in baseball, in addition to home runs and walks, are home runs and strikeouts (see Section 1.5). Meanwhile, triples and home runs are thought to be strongly negatively correlated, as are singles and strikeouts. To further flesh out our understanding of the estimation of the two models, we will investigate the recovered association of these pairs.

$$Corr(W_{HR}^{\hat{\Phi}_{MLOGIT}}, W_{SO}^{\hat{\Phi}_{MLOGIT}}) = 0.17 \qquad Corr(W_{3B}^{\hat{\Phi}_{MLOGIT}}, W_{HR}^{\hat{\Phi}_{MLOGIT}}) = -0.15$$

$$Corr(W_{jHR}^{\hat{\Lambda}}, W_{jSO}^{\hat{\Lambda}}) = 0.12 \qquad Corr(W_{j3B}^{\hat{\Lambda}}, W_{jHR}^{\hat{\Lambda}}) = -0.04$$

$$Corr(W_{1B}^{\hat{\Phi}_{MLOGIT}}, W_{SO}^{\hat{\Phi}_{MLOGIT}}) = -0.59$$

$$Corr(W_{j1B}^{\hat{\Lambda}}, W_{jSO}^{\hat{\Lambda}}) = -0.52$$

In large part we see the relationships we expect, though perhaps not as strong a positive correlation between home runs and strikeouts. In the unconditional implied correlation structure based on Λ (and thus only in the MLN fit), the strongest within season correlation is the -0.69 between (SO, OIP), followed by the -0.52 between (1B, SO). The strongest positive correlations are 0.19 between (HR, SF) and (1B, OIP).

These make quite a bit of sense; striking out and putting the ball in play are exactly opposite outcomes, and players who hit a lot of singles are likely putting the ball in play often as well, instead of striking out. Meanwhile, players often attempt to hit

home runs when runners are on third base, assuming that even if they don't get all of the ball, a sacrifice fly will net the team a run. It seems, with the random effects covariance matrix, that the M-Logit model is capable of describing the positive relationships much better for the ten-category setting than it was for the three-category. This may be a major reason the M-Logit is the preferred fit for the training data; Σ may not be as necessary to describe the positive associations which are present in the data.

A collection of plots similar to Figure 3.15, where the square root of predicted counts is compared to the square root of true counts, can be found in Appendix E; it shows the MLN fit struggles somewhat with rarer events, such as triples and sacrifice hits, but appears to do well with singles, doubles and strikeouts. On an individual player basis, we will once again focus on Mike Young and Rod Allen's first seasons in the Central League in Japan. In this case, credible intervals are still useful for direct player comparison, but we may also gain some utility by comparing summary statistics, such as OPS.

Table 3.23.
95% Credible interval, posterior mean, and observed value for Mike Young's 1990 season, based on MLN fit of ten-category real test data

	1B	**2B**	**3B**	**HR**	**SH**	**SF**	**BB**	**HBP**	**SO**	**OIP**
CI LB	28	6	0	6	0	1	17	1	47	94
post pred mean	33.89	8.96	0.57	10.59	0.50	1.43	24.21	2.67	59.84	107.34
true value	30	11	0	11	0	0	26	2	65	105
CI UB	42	12	1	17	1	2	35	5	75	121

Table 3.24.
95% Credible interval, posterior mean, and observed value for Rod Allen's 1989 season, based on MLN fit of ten-category real test data

	1B	**2B**	**3B**	**HR**	**SH**	**SF**	**BB**	**HBP**	**SO**	**OIP**
CI LB	28	7	0	5	0	1	9	1	30	113
post pred mean	42.02	10.86	0.74	11.90	0.75	1.55	18.41	3.01	56.21	133.54
true value	50	12	1	11	0	3	25	4	47	126
CI UB	58	17	1	21	3	3	31	11	86	161

OPS is calculated by adding on-base percentage, $OBP = \frac{H+BB}{AB}$, and slugging percentage, $SLG = \frac{TB}{AB}$, and is popular as a summary of batting ability which incorporates both plate discipline and ability to hit for extra base hits. With ten categories, it is straightforward to calculate OPS, since $H = 1B + 2B + 3B + HR$, $TB = 1B + 2(2B) + 3(3B) + 4(HR)$, and $AB = H + SO + OIP$. We present credible intervals and posterior predictive means for both Mike Young and Rod Allen's first CL seasons.

Table 3.25.
Example of two predicted player-season OPS and their 95% credible intervals, under MLN model fit

player-season	**OPS LB**	**OPS PPM**	**OPS True**	**OPS UB**
Mike Young '90	0.642	0.789	0.784	0.948
Rod Allen '89	0.572	0.775	0.891	0.939

The posterior means for HR and BB for both of these players are closer to the truth now than they were in just the three-category results, though Allen's BB rate is still predicted at much lower than the truth. In this case, this has an impact on the OPS estimate. While Young's OPS is predicted very well, Allen's undershot. Solely based

on the predicted OPS and corresponding interval, Young seems to be the preferred player, since he has a higher floor, mean, ceiling, and a narrower interval. While Allen ended up with a better OPS than Young, his observed OPS was not outside the range of the credible interval.

While we have discussed utilizing the credible intervals as a form of risk assessment when pursuing two players, all we have done so far is to compare the marginal bounds of the outcomes. A singular metric, like OPS, is closer to what might be useful in terms of immediately comparing players. A more formal exploration of risk would involve assessing the range in monetary value a player might be worth.

3.5 Discussion

The results in this chapter at once show that our methodology is promising in its ability to provide accurate predictions for baseball player seasons after switching leagues, while also indicating that there is still work to be done to refine and improve the model. When comparing models in terms of DIC, we found that the dimensionality of the response had an impact on whether our model was preferred to the simpler multinomial logit model. Even though DIC found the additional covariance matrix terms to be unnecessary in the ten category study, the recovery of more than half of the covariance posterior distributions were significantly different from zero in the multinomial logistic-normal fit. The Dunn-Smyth residual QQ-plots agreed with DIC in the three category model, but there was a less clear delineation of model fit when moving to ten categories, showing that neither of our candidate models seemed to fit the data perfectly. For these reasons, and the added flexibility gained from the inclusion of the Σ covariance matrix, we still believe our model to be useful in this context. Additionally, the framework we have developed for providing uncertainty quantification for predictions of baseball seasonal count outcomes is a novel contribution to the study of baseball performance.

The application in this chapter is itself untouched, and we established the benefit of focusing on the movement of players between NPB and MLB. The estimation results tend to coincide with traditional thinking about the difference in leagues; that it is easier to succeed as a power hitter in the Japanese Leagues. In part because of this, there could be a bias introduced in terms of a player's estimated random effect when only estimated based on their performance in another league. Including this, there are certainly other practical considerations and improvements that could be discussed and we will attempt to flesh out in the final chapter.

4. SUMMARY AND FUTURE DIRECTIONS

4.1 Summary

Baseball is a sport that can be broken down into a series of events (at bats) that have a finite set of unique outcomes (e.g., strike out, single, or home run). Predicting a season of these at bats (or the underlying probability vector) is of great interest to those who construct baseball team rosters. In academia, however, predicting seasonal performance has been primarily done on a univariate level. Furthermore, proprietary baseball projection systems that work on the multivariate level do not usually supply uncertainty quantification. This dissertation fills those voids. We also focus on the unexplored problem of predicting players moving from Japan's NPB to Major League Baseball, as well as in reverse.

In the modeling of multivariate count data, the multinomial logistic-normal distribution has been shown to be a very flexible model because of its ability to describe both positive and negative correlations between the counts. The multinomial logistic-normal distribution has yet to be utilized in a prediction context with longitudinal data, and there has been little discussed on ways of assessing model fit. We use a Bayesian version of this model to predict a season of counts in a longitudinal setting and provides uncertainty quantification in the form of credible intervals.

Chapter 1 focuses on summarizing the state of the art in sports prediction, especially baseball, and discusses why there is a place for this type of projection system. Specifically, one that is open source and provides a richer framework of uncertainty assessment. Of the published projection results, only one has a methodology that is open source or any form of uncertainty quantification. In the context of our specific

problem, that system, the Marcels, only predicts Japanese players coming to MLB at the league average with a reliability score of 0. To this point, there is no published system we are aware of which provides projections for players moving from MLB to Japan's NPB.

Also in Chapter 1, we discuss the need for hierarchical models to describe baseball outcome data. In Section 1.4 we describe three candidate models, the multinomial-Dirichlet, multinomial-nested Dirichlet, and multinomial logistic-normal, and discuss their relative correlation structures for the counts. We compare the fit of these models to some baseball data in Section 1.5 and find that the multinomial logistic-normal distribution fits the data better than either the multinomial-Dirichlet or multinomial-nested Dirichlet. The multinomial-Dirichlet is incapable of modeling the positive correlations among outcomes. The multinomial-nested Dirichlet is more flexible but it relies on predetermining the nesting structure, which is non-trivial, especially for a large number of outcomes. For these reasons, we focus on the multinomial logistic-normal as the hierarchical model of choice for our application.

Chapter 2 introduces our mixed effects multinomial logistic-normal model and subsequent Bayesian framework for estimation. The inclusion of a random effect vector to model autocorrelation is explained and our Metropolis or Metropolis-Hastings within Gibbs algorithm for estimation is outlined. Chapter 2 continues by outlining the various strategies for making predictions under the model framework, assessing model fit, predictive accuracy, and uncertainty about predictions.

The third chapter begins with description and presentation of results from a simulation study where data are simulated under both the mixed effects multinomial logistic-normal and mixed effects multinomial logit. After fitting both data sets with both models, we show that the deviance information criterion is capable of discerning the appropriate model fit, and that in the presence of additional variability in

the underlying probabilities, the multinomial logistic-normal model is a better fit for the data than the multinomial logit. We also give examples from the model fits of predictions and credible intervals for the expected predicted count vector.

Chapter 3 continues with our real data application of the model, using collated data from NPB and MLB of all players to have played in both leagues from before 2015. We discuss the importance and novelty of addressing the problem of predicting performance for those players switching between the leagues, followed by results from fitting both our multinomial logistic-normal model and the multinomial logit model to the data given two levels of dimensionality in the response vector of counts. We found that in the lower dimension setting, the multinomial logistic-normal model provides a better fit than the multinomial logit when two of the three outcomes of interest are positively correlated. In the higher, ten dimensional setting, the difference is less stark, though signs in evaluating the predictive performance point to the multinomial logistic-normal giving a better estimate of uncertainty about expected predicted performance than the multinomial logit.

We conclude in this chapter by discussing some of the work yet to be done in fully realising the potential of this approach, including improvements to the methodology as described in the earlier chapters, and extensions both within baseball and other sports, and also other disciplines as well. We believe the model has merits in its current form that further the state of prediction in baseball and contributes to the development of hierarchical models for count vectors by describing the multinomial logistic-normal distribution's use in prediction of longitudinal data.

4.2 Future Work/Directions

In this section we describe some possible improvements to our methodology as it currently stands, followed by application extensions which may benefit from use of the method.

4.2.1 Improvements to Methodology

Individual Aging Effects

In terms of our application to baseball prediction, there could be reason to believe that different players age in different ways over the course of their careers, leading to the natural question of including individual age effects instead of an overall effect as we have done in this work.

In some baseball prediction research (Berry et al., 1999; Fair, 2008) it has been suggested that players do not all age similarly in ability. In our application we restricted aging to have a fixed quadratic relationship with the outcomes of interest. Our random effect term allows players to be constantly above or below this average age trend, but does not allow the shape of the trend to differ from player to player. Berry et al. (1999) assumed a different aging function for each player i, where the function of a player's age is modeled with a flexible random curve. In their paper, they denote $g(a)$ the mean aging curve for the singular metric, for instance home run rate, and define a peak age $\breve{a}$ where they assume $g(\breve{a}) = 0$, and their aging function follows:

$$f_i(a) = \begin{cases} g(a)\psi_{1i} & \text{if } a < a_M \\ g(a) & \text{if } a_M \leq a \leq a_D \\ g(a)\psi_{2i} & \text{if } a > a_D \end{cases} \tag{4.1}$$

where $\psi_i = (\psi_{1i}, \psi_{2i})$ represents player i's variation from the mean aging curve, and the values a_M and a_D correspond to cutoff ages for a "maturing" and "declining" phase in the player's career. In this way, players are assumed to mature at different rates until they reach the peak age, and decline at different rates until the end of their career. This does, however, imply at least an additional 2 parameters per player per log-odds in the context of our multinomial logistic-normal model, and a corresponding non-trivial addition to the computation time required to fit the model. However, if that shortcoming could be overcome, the potential benefits of more flexibly modeling the aging trajectory of player's performance could prove to be quite valuable in improving prediction.

Adjusting Random Effects

In our specific problem's context, there are limitations in using a single random effect vector to describe a player's performance relative to the mean performance. It is commonly accepted that the baseball played in NPB is on a slightly lower level than that played in MLB, and our results in terms of the recovered league effects bear that out to some extent. It is perhaps not reasonable to assume that a player whose ability is above average in one league should similarly be equally above average in ability in the other. While we expect there to be a fixed effect of league on player performance, however, there may be some interaction between player and league.

This is especially a potential issue in the prediction of new players when switching leagues. A single random effect vector per player describes that player's performance relative to the mean over their career, whether in one league or another. When predicting new player's however, they have not yet moved into a new league, so an estimated random effect will only represent their performance relative to their initial league. Our model thus assumes identical players moving to the same league from different leagues will maintain their relative performance, and might tend to project

the player from the weaker league to perform better than his identical counterpart. We still believe this to be a more informative approach than not using a random effect to predict new league performance. Generally, the league covariate effects are larger in magnitude than the player random effects, so the changing in league drives most of the change in performance while the random effect still does the job of building in correlation among outcomes over time.

It could still be useful to investigate alternative approaches, one option could be to initially estimate separate random effects for the players used to fit the model, whose careers in both leagues are used. After the initial model is fit, and each player has the number of random effects corresponding to each league, it could be possible to identify which of those players are most similar to the out-of-sample player whose season in a new league is being predicted. Once this "distribution" of similar players is identified, one can define a subsequent distribution of random effects for the new league based on these similar players from which the new player's random effect in the new league may be estimated. There are several options for identifying these similar players, whether via some non-parameteric clustering algorithm, or something as simple as a similarity score, such as one developed by Bill James (James, 1994), which is currently used in baseball circles already to rank similar MLB players (Baseball-Reference, 2019).

Another potential option would be to include a fixed inflation factor attached to league that is modeled as part of a player's random effect after they switch leagues. In the simplest case, this would mean including two extra parameters, one for players moving from MLB to NPB, and one for NPB to MLB. Instead of a constant ψ_i, the random effect in player-seasons after switching leagues would become $\gamma_{MLB}\psi_i$ for NPB players moving to MLB and $\gamma_{NPB}\psi_i$ for the converse.

Thus γ_{League} would be an inflation or deflation factor that is modeled on their performance after switching leagues. This way, after a player's random effect is estimated based on their performance in their original league, the fixed inflation factor that was estimated in the model could be brought in to scale the random effect corresponding to the switch being made. The simplest case of MLB-NPB, NPB-MLB would bring in only two new parameters, while accounting for moving between (or within) both groups of sub-leagues (AL, NL) and (CL, PL) would mean eight (or 12).

Additional Covariates/Interactions

A simpler option to account for the possible effect discussed in the previous subsection is to include a "home" covariate in the model, which identifies where the player comes from. The idea being that where a player was trained, either in the NPB or MLB, could have some impact on their performance throughout their career. We tried this for the three category real data and the posterior distribution for the fixed effect was significantly different from zero, while it had an impact on the significance of the age effect terms. This is worth investigating further. This is just one of the additional covariates that might be useful to include in a more fully realised addressing of the problem.

Another interesting covariate could be an indicator representing whether a player has previous international experience. This could help address the uneasily tested, but certainly present culture shock issue. If a player has at some point in their life spent time in another country, it could impact their ability to more seamlessly adapt to a new cultural environment and aid in their success on the field. There is obviously some limit to the data collection in terms of measuring this variable.

One option could be to simply look at whether a player has spent time in another foreign league. Occasionally, both NPB and MLB teams send young players to play

baseball over the winter either in the Dominican Republic, Mexico or Puerto Rico. There are also professional and semi-professional leagues in many countries across the globe in which players may have played before. Identifying players who have prior international experience, and may be more inclined to be comfortable in another country, could be beneficial to identifying future success.

Finally, the model fit in Chapter 3 is a strictly additive model, with the only non-linear term being the quadratic age effect. It could be useful to investigate possible interactions among the covariates, especially concerning the three covariates age, league and position. It makes some sense that players age may impact performance differently in different leagues or at different positions, and that players of certain positions are more likely to succeed in a different league. The inclusion of additional covariates and interactions necessarily increases computation time, however, which is not a theoretical concern, but for the time being is a practical one.

Improving Computational Efficiency

Improving computational efficiency of Bayesian multinomial logistic-normal models themselves is an active area of research. To fit the model, we currently utilize a Metropolis-within-Gibbs algorithm. On their own, both the Metropolis (or Metropolis-Hastings) and Gibbs sampling algorithms, while guaranteed to converge, can be quite inefficient, depending on a properly tuned step-size parameter to adequately explore the entire distribution space. Hamiltonian Monte Carlo (HMC) was devised (Duane et al., 1987) to reduce correlation between samples in MCMC and improve mixing and efficiency. The reduction in correlation comes from the time evolution of the Markov chain as defined by a set a set of derivative equations, which form the Hamiltonian. We have not yet explored the feasibility of calculating the Hamiltonian in our mixed-effects multinomial logistic-normal setting, though HMC has been used in similar models (Äijö et al., 2017; Silvernamn et al., 2018), and there

remains the potential that moving to a Hamiltonian MCMC approach instead of the current Metropolis approach will speed up convergence of the overall chain. However, there may be an even more efficient alternative to HMC.

A current working paper (Silverman et al., 2019) describes a framework for writing MLN models as marginally latent matrix-t process (LTP) models. In their real data analysis, they show their LTP approach with a Laplace approximation provided similar estimates to a HMC approach, and did so more than 1000 times faster. This was with a subset from data set studying the microbiome in Crohn's disease patients, and included 83 samples, 49 taxa (categories) and 4 covariates. In terms of our application, baseball data involves data sets with greater number of observations and more covariates (though fewer categories). Silverman et al. (2019) showed via simulation study that the gains in efficiency of their Laplace approximation approach were consistent as the number of samples, covariates and dimension of categories increased. This LTP model framework thus promises to vastly improve computational efficiency so that, in our methodology, prediction can occur at a faster rate.

4.2.2 Extensions in Sports Research

Minor and Other Foreign Baseball Leagues

Our main goal in application is to predict performance of baseball players moving between NPB and MLB leagues while including quantification of uncertainty about that prediction. A natural extension is to broaden the scope to include players from the Minor Leagues in America, as there are many players who move to NPB, or other foreign leagues, without ever playing in MLB. Japanese teams also often target the best players in the minor leagues to recruit. Minor league players serve to make more money in Japan than in the minor leagues, while playing at a higher level and in front of more fans. NPB teams would certainly find a model which includes minor league players to be beneficial. Given data access, this model would be easily extended to

include the minor leagues, or even other foreign leagues such as the Korean Baseball Organization (KBO) or Cuban Nacional Series (CNS). The top major leagues in those two countries, both of which have seen star players move to MLB and succeed (examples include Hyun-Jin Ryu from KBO and Jose Abreu from CNS), and within which, especially the KBO, former MLB players have found great success (including Eric Thames, who has since moved back to MLB from KBO).

Cricket

The man credited with inventing baseball's box score, Henry Chadwick, adapted it from a cricket scorecard. This underlies the connection between the two most popular bat-and-ball games in the world; the only mainstream ball games where the defense begins a possession with the ball. As in baseball, for each ball bowled (equivalent to a pitch), the batsmen or striker attempts to strike the ball out of the reach of fielders so his team may score runs. In cricket, there are fewer possible outcomes per ball faced (the equivalent to a plate appearance) than in baseball, though the nature of the game means that any number of runs may be scored per strike of the ball, as opposed to baseball where the limit is four.

This is the main difference that may make this methodology's application to cricket less useful. Since any number of runs may be scored per ball faced, the main statistics kept on batsmen deal with the number of runs per ball, which do not face the same constraint of adding up to the number of balls faced (consider RBI in baseball). It would still be possible to define outcomes with this constraint for batsmen, but they may not be as useful as they are in baseball. For bowlers (or the equivalent of the pitcher) however, there is a limit of ten ways in which a batter may be dismissed (be put out, in baseball terms), and these directly impact the style and performance of bowlers.

There are, as in baseball, numerous leagues for Cricket players around the world, with varying levels of skill in each. Predicting performance for players moving between them during league off-seasons benefits the teams in those leagues.

Timed Sports

With sports that necessarily rely on a clock to determine length of play, and which are less obviously discrete in nature than baseball or cricket, there is still a framework within which outcomes could be defined. Consider in basketball, on a singular possession for a player in a game, there are a discrete set of outcomes which could occur: score 2 points, score 3 points, missed shot, assist a score, other pass, turnover, commit offensive foul, draw defensive foul. Data of this granularity may be difficult to come by, but even on a play-by-play basis, instead of possession, a discrete set of outcomes may be defined. However, the clock is a major factor in the sport, and a discrete set of mutually exclusive outcomes fails to incorporate time.

Likewise American football, the most popular sport in the United States, may be delineated on a play-by-play basis by certain outcomes, but also has a time component that would not be capable of being modeled in our context. In addition, each play in football produces a spatial result, in the form of yardage. In single yard increments, it would be possible to include these in a discretization of football outcomes, but it would greatly expand the number of possible outcomes to account for the 100 yard length of the football field.

4.2.3 Extensions to Other Disciplines

Clinical: Microbiome Composition

The field in which the multinomial logistic-normal distribution receives the most attention, without a doubt, is the study of microbial taxa and their composition

in the gut. Numerous papers in this research area have already been mentioned, including some very recent and state of the art studies (Xia et al., 2013; Grantham et al., 2017; Ren et al., 2017; Li et al., 2018; Silverman et al., 2019). The multinomial logistic-normal model's place in this field is firmly in place, but most of the research in microbial composition analysis is focused on inference, even when data is collected in a time series context (Äijö et al., 2017). Prediction is a secondary concern when the main interest is examining treatment or covariate effects on compositions which have a direct impact on health, but there could be a place for prediction of the microbiome based on application of new treatments. Especially if hoping to predict the effect a treatment may have on a patient's microbial gut composition when moving from an original treatment, our prediction scheme could serve as a stepping point for improving that area of research.

Ecological: Species' Habitat Composition

Much of the research done in modeling habitat composition has been focused on cross-sectional experimental data (Billheimer and Guttorp, 1995; Billheimer et al., 2001; Brewer et al., 2005). We provide a framework for making predictions in a longitudinal observational data setting, which could easily be applied to the problem of modeling species' habitat composition. There are many factors which may have an impact on the composition of species in an area of land or sea, which could change over time; weather and proximity of human habitation being most obvious. The multinomial logistic-normal model has already been used in this context due to its ability to handle overdispersion of counts which is often found in ecological systems (Billheimer et al., 2001). There are ways it may be useful to make predictions about such data, including determining the ideal timing for species reintroduction to a habitat. Our prediction and uncertainty quantification framework could be used to compare the range of predicted species' counts in two different ecological systems, which may help make decisions related to human use of those systems.

REFERENCES

Äijö, T., Müller, C. L., and Bonneau, R. (2017). Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. *Bioinformatics*, 34:372–380.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, 29 West 35th Street, New York, NY 10001.

Aitchison, J. and Bennett, J. A. (1970). Polychotomous Quantal Response by Maximum Indicant. *Biometrika*, 57:253–262.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, pages 267–281.

Albert, J. (1994). Exploring Baseball Hitting Data: What about those Breakdown Statistics? *Journal of the American Statistical Association*, 89(427):1066–1074.

Albert, J. (2002). Smoothing career trajectories of baseball hitters. page August 22.

Albert, J. (2008). Streaky Hitting in Baseball. *Journal of Quantitative Analysis in Sports*, 4(1):Article 3.

Albert, J. (2013). Looking at Spacings to Assess Streakiness. *Journal of Quantitative Analysis in Sports*, 9(2):151–163.

Albert, J. (2016). Improved component predictions of batting and pitching measures. *Journal of Quantitative Analysis in Sports*, 12(2):73–85.

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):699–679.

Albright, S. C. (1993). A Statistical Analysis of Hitting Streaks in Baseball. *Journal of the American Statistical Association*, 88(424):1175–1183.

Allenby, G. M. and Lenk, P. J. (1994). Modeling Household Purchase Behavior with Logistic Normal Regression. *Journal of the American Statistical Association*, 89(428):1218–1231.

Baseball-Reference (2019). Similarity Scores. https://www.baseball-reference.com/about/similarity.shtml. Accessed 2019-06-29.

Baumer, B. S., Jensen, S. T., and Matthews, G. J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84.

Bennet, J. M. and Flueck, J. A. (1983). An Evaluation of Major League Baseball Offensive Performance Models. *The American Statistician*, 37(1):76–82.

Berry, S. M., Reese, C. S., and Larkey, P. D. (1999). Bridging Different Eras in Sports. *Journal of the American Statistical Association*, 194(447):661–676.

Billheimer, D. and Guttorp, P. (1995). Spatial Models for Discrete Compositional Data. Technical Report, University of Washington, Seattle, Department of Statistics.

Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association*, 96(456):1205–1213.

Bouguila, N. (2008). Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brewer, M. J., Filipe, J. A. N., Elston, D. A., Dawson, L. A., Mayes, R. W., Soulsby, C., and Dunn, S. M. (2005). A Hierarchical Model for Compositional Data Analysis. *Journal of Agriculture, Biological, and Environmental Statistics*, 10(1):19–34.

Campbell, G. and Mosimann, J. E. (1987). Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*, pages 10–17.

Chandler, G. and Stevens, G. (2012). An Exploratory Study of Minor League Baseball Statistics. *Journal of Quantitative Analysis in Sports*, 8(4):1–26.

Chen, J. and Li, H. (2013). Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *The Annals of Applied Statistics*, 7(1):418–442.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64:194–206.

Crespi, C. M. and Boscardin, W. J. (2009). Bayesian Model Checking for Multivariate Outcome Data. *Computational Statistics and Data Analysis*, 53(11):3765–3772.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B.*, 195(2):216–222.

Dunn, K. P. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:1–10.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, 70:311–319.

Fagerland, M. W. and Hosmer, D. W. (2012). A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models. *The Stata Journal*, 12(3):447–453.

Fair, R. C. (2008). Estimating Age Effects in Baseball. *Journal of Quantitative Analysis in Sports*, 4(1):Article 1.

Filzmoser, P., Garrett, R. G., and Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, 31:579–587.

Friendly, M. (2015). *Lahman: Sean Lahman's Baseball Database*. R package version 4.0-1.

Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, 48(3):432–435.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, The Edinburgh Building, Cambridge CB2 8RU, UK.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487.

Good, I. J. (1966). How to Estimate Probabilities. *J. Inst. Maths Applics*, 2:364–383.

Grantham, N. S., Reich, B. J., Borer, E. T., and Gross, K. (2017). MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments. *ArXiv e-prints*.

Gross, K. and Bates, D. (2012). *mvnmle: ML estimation for multivariate normal data with missing values*. R package version 0.1-11.

Hartig, F. (2019). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.2.4.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, 1:81–102.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22:1433–1446.

Hijazi, R. H. and Jernigan, R. W. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics*, 4:77–91.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

James, B. (1994). *The Politics of Glory: How Baseball's Hall of Fame Really Works*. Macmillan.

Kass, R. and Steffey, D. (1989). Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *Journal of the American Statistical Association*, 84(407):717–726.

Kopka, H. and Daly, P. W. (1995). *A Guide to LaTeX : Document Preparation for Beginners and dvanced Users*. Addison-Wesley, Reading Massachusetts, second edition.

Larson, W. (2011). Projecting Uncertainty. https://community.fangraphs.com/projecting-uncertainty/. Accessed 2019-06-15.

Larson, W. (2014). Evaluating 2013 Projections. https://community.fangraphs.com/evaluating-2013-projections/. Accessed 2019-06-05.

Larson, W. (2015a). 2014 Projection Review (Updated). https://community.fangraphs.com/2014-projection-review-updated/. Accessed 2019-06-05.

Larson, W. (2015b). The Baseball Projection Project. http://www.bbprojectionproject.com/. Accessed 2019-06-05.

Li, Z., Lee, K., Karagas, M. R., Madan, J. C., Hoen, A. G., O'Malley, A. J., and Li, H. (2018). Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Statistics in Biosciences*, 10(3):587–608.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society*, 34(1):1–41.

Lindsey, G. R. (1959). Statistical Data Useful for the Operation of a Baseball Team. *Operations Research*, 7(2):197–207.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

McGarry, T. and Franks, I. M. (1994). A stochastic approach to predicting competition squash match-play. *Journal of Sports Sciences*, 12(6):576–584.

Minka, T. (2010). Bayesian linear regression. Technical Report, MIT Media Lab.

Mosimann, J. E. (1962). On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions. *Biometrika*, 49(1/2):65–82.

Mosteller, F. (1952). The World Series Competition. *Journal of the American Statistical Association*, 47(259):355–380.

Neal, D., Tan, J., Feng, H., and Wu, S. (2010). Simply Better: Using Regression Models to Estimate Major League Batting Averages. *Journal of Quantitative Analysis in Sports*, 6(3).

Null, B. (2009). Modeling Baseball Player Ability with a Nested Dirichlet Distribution. *Journal of Quantitative Analysis in Sports*, 5(2).

Ocampo, S. R., Garcia, H., and Uy, M. (2019). Dirichlet-Multinomial Estimation of Small Area Proportions of Socio-Economic Classes. In Kor, L.-K., Ahmad, A.-R., Idrus, Z., and Mansor, K. A., editors, *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, pages 317–323, Singapore. Springer Singapore.

Pettitt, A. N., Tran, T. T., Haynes, M. A., and Hay, J. (2006). A Bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants. *Journal of the Royal Statistical Society: Series A*, 169(1):97–114.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Polya-gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349.

Quesenberry, C. P. and Hurst, D. C. (1964). Large Sample Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 6(2):191–195.

Ren, B., Bacallado, S., Favaro, S., Vatanen, T., Huttenhower, C., and Trippa, L. (2017). Bayesian nonparametric mixed effects models in microbiome data analysis. *arXiv preprint arXiv:1711.01241*.

Robitzsch, A. (2016). *sirt: Supplementary Item Response Theory Models*. R package version 1.10-0.

Rosner, B., Mosteller, F., and Youtz, C. (1996). Modeling Pitcher Performance and the Distribution of Runs per Inning in Major League Baseball. *The American Statistician*, 50(4):352–360.

Rowe, D. B. (2002). *Multivariate Bayesian statistics: models for source separation and signal unmixing*. CRC Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shen, S. M. (1982). A method for discriminating between models describing compositional data. *Biometrika*, 69(3):587–596.

Silverman, J. D., Roche, K., Holmes, Z. C., David, L. A., and Mukherjee, S. (2019). Bayesian multinomial logistic normal models through marginally latent matrix-t processes.

Silvernamn, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S., and David, L. A. (2018). Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(202).

Sison, C. P. and Glaz, J. (1995). Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. *Journal of the American Statistical Association*, 90(429):366–369.

Skrondal, A. and Rabe-Hesketh, S. (2003). Multilevel Logistic Regression for Polytomous Data and Rankings. *Psychometrika*, 68(2):267–287.

Smith, A. F. M. (1973). A General Bayesian Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(1):67–75.

Smith, C. (2006). Projection system championships. http://lanaheimangelfan.blogspot.com/2006/12/projection-system-championships.html. Accessed 2019-06-05.

Smithson, M. and Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables. *Psychological Methods*, 11(1):54–71.

Tango, T. (2004). Tango on Baseball. http://www.tangotiger.net/archives/stud0346.shtml. Accessed 2019-04-01.

Tango, T. (2012). Marcel 2012. http://www.tangotiger.net/marcel/. Accessed 2019-04-01.

Tango, T., Silver, N., and Others (2007). Forecast evaluations. http://www.insidethebook.com/ee/index.php/site/comments/forecast_evaluations/. Accessed 2019-06-05.

Wickramasinghe, I. (2014). Predicting the performance of batsmen in test cricket. *Journal of Human Sport and Exercise*, 9(4):744–751.

Wolfson, J., Addona, V., and Schmicker, R. H. (2011). The Quarterback Prediction Problem: Forecasting the performance of college quarterbacks selected in the NFL Draft. *Journal of Quantitative Analysis in Sports*, 7(3):1–19.

Wong, G. and Mason, W. (1985). The Hierarchical Logistic Regression Model for Multilevel Analysis. *Journal of the American Statistical Association*, 80(391):513–524.

Wong, T. T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97:165–181.

Wright, D. and London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, 62:439–456.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics*, 69:121–139.

Yuan, Y. and Johnson, V. E. (2012). Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models. *Biometrics*, 68(1):156–164.

A. MIXED-EFFECTS MULTINOMIAL LOGIT MODEL

A.1 Bayesian Mixed-Effects Multinomial Logit Model

In order to properly benchmark the Bayesian mixed-effects multinomial logistic-normal model framework described in this chapter, we establish an equivalent model for data with a single random effect which does not include the additional variability described by Σ. Data are generated under a mixed-effects multinomial logit model following the same set-up as in the multinomial logistic-normal, except that the model for the log-odds differs:

$$Y_{ij} = X_{ij}\beta + \psi_i \tag{A.1}$$

This may seem a trivial change from the single random effect multinomial logistic-normal model, but the removal of the ε error term in some ways makes computation a less efficient task than in the multinomial logistic-normal. Mixed-effects multinomial logistic regression has been addressed previously in both frequentist (Hartzel et al., 2001; Hedeker, 2003) and Bayesian (Pettitt et al., 2006) settings. In order to properly compare our multinomial logistic-normal model, we present a corresponding Bayesian hierarchical model for the mixed-effects multinomial logit, with a single random effect.

The removal of Σ presents some issues in the sense that both fixed and random effects may no longer be updated jointly, as they are the only two components that make up the underlying log-odds. Due to this, it is convenient to treat the intercept as a separate entity from the other fixed effects, and set it as the prior mean of the random effects. We retain the improper uniform prior on the β parameters, and the disperse inverse-Wishart prior on the random effects covariance matrix Φ, and set the prior

mean of the random effects to be the overall intercept, which we will denote μ. So, the log-odds are modeled with β not including an intercept:

$$\psi_i \sim N_d\left(\mu, \Phi\right)$$

$$p(\mu, \beta) \propto 1$$

$$\Phi \sim Wishart^{-1}\left(\nu = d, \Lambda = I_d + \mathbf{1}_d\mathbf{1}_d^T\right)$$

Similar to the mixed-effects multinomial logistic-normal, either numerical integration or MCMC methods are required to sample from the conditional posteriors of the random and fixed effects.

A.2 Outline of Gibbs Sampler for Mixed-Effects Multinomial Logit Model

We describe a Metropolis-within-Gibbs sampler for fitting the mixed-effects multinomial logit model with a single random effect. This algorithm is used in comparing model fit between data simulated under the two data archetypes, multinomial logistic-normal versus multinomial logit, in the simulation studies in Chapter 3.

(1) Initialize starting values, $t-1=0$, μ^0, β^0, ψ_i^0, and Φ^0.

(2) Update of ψ_i:

$$\psi_i^t|X_i, \mu^{t-1}, \beta^{t-1}, \Phi^{t-1}, W_i \propto p(W_i|X_i, \mu^{t-1}, \beta^{t-1}, \psi_i^t) \times f(\psi_i^t|X_{ij}, \theta_i^{t-1} = (\mu^{t-1}, \beta^{t-1}, \Phi^{t-1}))$$

via Metropolis algorithm.

$$p(W_i|X_i, \mu^{t-1}, \beta^{t-1}, \psi_i^t) \times f(\psi_i^t|X_{ij}, \theta_i^{t-1}) \sim \prod_{j=1}^{n_i} Multi_{d+1}(PA_{ij}, \Pi(Y_{ij}^t)) \times N_d(\mu^{t-1}, \Phi^{t-1})$$

Where $Y_{ij}^t = \mu^{t-1} + X_{ij}\beta^{t-1} + \psi_i^t$, and

$$\Pi(Y_{ij}^t) = \left(\frac{e^{y_{ij1}^t}}{1 + \sum_{k=1}^{d} e^{y_{ijk}^t}}, ..., \frac{1}{1 + \sum_{k=1}^{d} e^{y_{ijk}^t}} \right) \tag{A.2}$$

With multivariate normal proposal distribution:

$$\psi_i^t \sim N_d \left(\psi_i^{t-1}, c_\psi \Phi^{t-1} \right)$$

Where c_ψ is a scalar used to control the "step-size" of the proposal distribution.

(3) Update of Φ:

$\Phi^t | \Psi^t \sim Wishart^{-1} \left(\nu_2' = \nu_2 + m, \Lambda_2' = \Lambda_2^{-1} + \Psi^t \Psi^{t^T} \right)$

Where:

$$\Psi^t = (\psi_1^t, ..., \psi_m^t)^T \tag{A.3}$$

(4) Update of μ:

$\mu^t | \Psi^t, \Phi^t \sim N_d \left(\bar{\Psi}^t, \frac{\Phi^t}{m} \right)$

Where:

$$\bar{\Psi}^t = \frac{1}{m} \sum_{i=1}^{m} \psi_i^t \tag{A.4}$$

(5) Update of β:

For each vector β_k in the matrix β update given β_{-k}, the matrix with column k removed, corresponding to the conditional likelihood (Holmes and Held, 2006):

$$l \left(\beta_k^t | \beta_{-k}^{t-1}, W_{ijk}, \psi_i^t \right) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left(\frac{e^{\eta_{ijk}^{t-1}}}{1 + e^{\eta_{ijk}^{t-1}}} \right)^{W_{ijk}} \left(\frac{1}{1 + e^{\eta_{ijk}^{t-1}}} \right)^{PA_{ij} - W_{ijk}} \tag{A.5}$$

Where:

$$\eta_{ijk}^{t-1} = X_{ij}\beta_k^{t-1} + \psi_i^t - C_{ijk}^{t-1} \tag{A.6}$$

$$C_{ijk}^{t-1} = log \left(1 + \sum_{l \neq k} e^{X_{ij}\beta_l^{t-1} + \psi_i^t} \right) \tag{A.7}$$

via Metropolis algorithm with multivariate normal proposal distribution of dimension p, the number of covariates (not including the intercept):

$$\beta_k^t \sim N_p\left(\beta_k^{t-1}, c_\beta \Sigma_\beta^{t-1}\right)$$

Where c_β is a scalar used to control the "step-size" of the proposal distribution Σ_β^{t-1} can be an proposal covariance matrix. One option is to use Pólya-Gamma auxiliary variables (Polson et al., 2013) to estimate a reasonable shape of the covariance matrix:

$$\Sigma_\beta; t-1 = X^T \Omega_k^{t-1} X \tag{A.8}$$

Where $\Omega_k^{t-1} = diag\left(\{\omega_{ijk}^{t-1}\}\right)$, and ω_{ijk}^{t-1} are Pólya-Gamma auxiliary variables given β_k^{t-1}:

$$\omega_{ijk}^{t-1} \sim PG\left(PA_{ij}, \eta_{ijk}^{t-1}\right)$$

(6) Set $t = t + 1$, repeat steps (2) - (5) until convergence criteria met.

In practice, this Gibbs algorithm is somewhat slow mixing in the parameters, but the complete data posterior log-likelihood tends to converge quickly.

B. ALTERNATIVE PROPOSAL SCHEMES

B.1 Random Walk Normal Proposal

The first proposal scheme falls under a Metropolis framework, in which due to symmetry of the proposal distribution, detailed balance is natively satisfied. In essence, each player-season's latent variables are following a multivariate Gaussian random walk, where the next value of Y_{ij} is determined only by the previous value. The proposal scheme proceeds:

(1) Propose $Y_{ij}^{t}|Y_{ij}^{t-1}, \Sigma^{t-1} \sim N_d\left(Y_{ij}^{t-1}, c\Sigma^{t-1}\right)$

Where c is a scalar used to control the "step-size" of the random walk.

(2) Accept with probability $\alpha = min\left\{1, \frac{f\left(Y_{ij}^{t}|X_{ij},Z_{ij},\theta_i^{t-1},W_{ij}\right)}{f\left(Y_{ij}^{t-1}|X_{ij},Z_{ij},\theta_i^{t-1},W_{ij}\right)}\right\}$.

The advantages of this proposal scheme involve the conceptual simplicity and a slight computation time advantage over the schemes that follow. However, the gains in computation time do not tend to make up for losses in correlation among the samples from the target distribution. Auto-correlation functions can show that under the same data set and same settings, this Normal Random Walk proposal scheme tends to mix poorly.

B.2 Static Beta Proposal

Ignoring the correlation between outcomes when proposing, which the normal random walk proposal does not do, or instead proposing based on probabilities the latent variables imply rather than the latent variables themselves, can be shown to improve mixing at a small cost of computation time. Thus the final two proposal schemes deal with proposing log-odds or their corresponding probabilities based on independent beta distributions for each probability.

In the first case, we recognize that we can, for the purposes of proposing, redefine each outcome count as having come from a conditional binomial distribution, and the resulting probability from a beta distribution.

$$w_{ijk}|\left\{w_{ijl}\right\}_{l\notin\{k,d+1\}} \sim binomial\left(PA_{ij} - \sum_{l\notin\{k,d+1\}} w_{ijl}, \pi^*_{ijk} = \frac{\pi_{ijk}}{\pi_{ijk}+\pi_{ij(d+1)}}\right)$$

$$\pi^*_{ijk} = \frac{e^{y_{ijk}}}{1+e^{y_{ijk}}} \sim beta\left(\alpha^*_{ijk}, \beta^*_{ijk}\right) \tag{B.1}$$

This allows us to define our second proposal scheme, which will operate under a Metropolis-Hastings framework to account for the non-symmetry of the beta proposal distribution:

(1) Propose $\pi^{*^t}_{ijk}|X_{ij}, Z_{ij}, \theta^{t-1}_{ik}, W_{ij} \sim beta\left(w_{ijk}+\alpha^{*^{t-1}}_{ijk}, w_{ij(d+1)}+\beta^{*^{t-1}}_{ijk}\right)$
Where $\alpha^{*^{t-1}}_{ijk} = \frac{1+e^{\mu^{t-1}_{ijk}}}{c\sigma^{2^{t-1}}_k} + \frac{e^{\mu^{t-1}_{ijk}}}{1+e^{\mu^{t-1}_{ijk}}}$, $\beta^{*^{t-1}}_{ijk} = \alpha^{*^{t-1}}_{ijk} e^{-\mu^{t-1}_{ijk}}$, $\mu^{t-1}_{ijk} = X_{ij}\beta^{t-1}_k + Z_{ij}\psi^{t-1}_{ik}$, $\sigma^{2^{t-1}}_k$ is the $(k,k)^{th}$ element of Σ^{t-1}, and c is a scalar used to control the "step-size" of the random walk.

(2) Convert $y^t_{ijk} = logit(\pi^{*^t}_{ijk})$.

(3) Accept with probability $\alpha = min\left\{1, \frac{f\left(Y^t_{ij}|X_{ij},Z_{ij},\theta^{t-1}_i,W_{ij}\right)}{f\left(Y^{t-1}_{ij}|X_{ij},Z_{ij},\theta^{t-1}_i,W_{ij}\right)} \times \frac{q\left(\pi^{t-1}_{ij}|X_{ij},Z_{ij},\theta^{t-1}_i,W_{ij}\right)}{q\left(\pi^t_{ij}|X_{ij},Z_{ij},\theta^{t-1}_i,W_{ij}\right)}\right\}$
Where $q\left(\pi^t_{ij}|X_{ij},\theta^{t-1}_i,W_{ij}\right) \sim \prod_{k=1}^d beta\left(w_{ijk}+\alpha^{*^{t-1}}_{ijk}, w_{ij(d+1)}+\beta^{*^{t-1}}_{ijk}\right)$.

C. AITCHISON'S R^2 AND SUM OF COMPOSITIONAL ERRORS

C.1 Formulation

It is possible to define distance metrics on the simplex, which can be used for assessing predictive accuracy of competing models. Aitchison (1986) suggests two measures for assessing the amount of explained variability in a model for compositional data. The first is an R^2 type measure which attempts to estimate the percentage of variation explained by a proposed model. The second is based on Aitchison's distance, a distance metric defined on the simplex, called the Sum of Compositional Errors (SCE). In both cases, a log-transform is used on odds of observed data, necessitating an adjustment for zero counts.

Aitchison's R^2, which we will denote aR^2, is based on a measure of total variability dependent on the variation matrix of the transformed log-odds:

$$T(W) = [\tau_{kk'}] = \left[var\left(log\left(\frac{w_k}{w_{k'}} \right) \right) \right] \tag{C.1}$$

where W is the data. This matrix is symmetric with zero diagonal elements. The total variability measure Aitchison describes is defined:

$$t(W) = \frac{1}{d} \sum_{k<k'} \tau_{kk'} = \frac{1}{2d} \sum T(W) \tag{C.2}$$

where d is the dimensionality of the compositional vector. Finally, Aitchison's R^2 measure is simply the ratio of total variability of the observed data and fitted data:

$$aR^2 = \frac{t(\hat{W})}{t(W)} \tag{C.3}$$

where $\hat{W}$ is the fitted data. This measure is attractive in that the interpretation is equivalent to that of R^2 in traditional ordinary least squares.

The Sum of Compositional Errors, in contrast, is the analogue to the Sum of Squared Errors (SSE) in traditional linear modeling. It is based on Aitchison's distance, a measure of distance between two compositional vectors on the simplex, which could be considered as an alternative residual to the squared Mahalanobis distance, between observed data W and fitted data $\hat{W}$:

$$\Delta_s\left(W, \hat{W}\right) = \left[\sum_{k=1}^{d}\left(log\left(\frac{\hat{w}_k}{g(\hat{W})}\right) - log\left(\frac{w_k}{g\left(W\right)}\right)\right)^2\right]^{\frac{1}{2}} \tag{C.4}$$

where $g\left(W\right) = \left(w_1 w_2 \cdots w_d\right)^{1/d}$ is the geometric mean of the vector W. Thus, if this measure can be considered the residual for player i in season j, the sum for all player-seasons would be the analogue to the SSE, called SCE, which serves as a measure of unexplained variability in the model:

$$SCE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Delta_s\left(W_{ij}, \hat{W}_{ij}\right) \tag{C.5}$$

In each of the above, the result will be the same whether working with the count vector W or a probability vector Π. Based on posterior predictive mean count vectors for each player-season being predicted, the above measures may be treated as measures of predictive accuracy. Unlike the squared Mahalanobis distances described in Chapter 2, these measures do not take into account the correlation which exists between the outcomes. This has the potential to under-represent models which seek to better account for that correlation, such as the multinomial logistic-normal.

C.2 Simulation Study

Tables C.1 through C.4 report Aitchison's measure of total variability for the observed and fitted data, $t(W)$ and $t(\hat{W})$, Aitchison's R^2, aR^2 and the Sum of Compositional Errors, SCE respectively for the two training data sets and the two model fits.

Table C.1.
Estimated total variability in observed counts, $t(W)$ for training data simulated under MLN and M-Logit models

$t(W)$	**MLN Data**	**M-Logit Data**
True Variability	0.7124	0.6206

Table C.2.
Total variability in posterior predicted mean counts, $t(\hat{W})$ for training data simulated under MLN and M-Logit models and fit with both

$t(\hat{W})$	**MLN Data**	**M-Logit Data**
MLN Fit	0.1129	0.1196
M-Logit Fit	0.1553	0.1342

Table C.3.
Aitchison's R^2, aR^2, for training data simulated under MLN and M-Logit models and fit with both.

aR^2	**MLN Data**	**M-Logit Data**
MLN Fit	0.1585	0.1927
M-Logit Fit	0.2180	0.2163

Table C.4.
Sum of Compositional Errors, SCE, for training data simulated under MLN and M-Logit models and fit with both.

SCE	**MLN Data**	**M-Logit Data**
MLN Fit	2907.82	2457.86
M-Logit Fit	2894.52	2452.80

These two measures of predictive accuracy do not find a great deal of difference between the predictive accuracy of the posterior predicted means for the MLN or M-Logit training data sets, though in both cases the measures tend to favor the M-Logit fit of the data sets. However, for all the above, we are examining model fit and predictive accuracy of data that was used to fit the models in question. A much more interesting problem is how the above metrics perform with out-of-sample data, namely the test sets constructed of the 105 player's held out of the initial model fitting.

Finally, we can also investigate whether the Aitchison's R^2 and the Sum of Compositional Errors metrics tell us something about the predicted player-seasons.

Table C.5.
Aitchison's R^2, aR^2, for the posterior predictive mean of 105 predicted seasons of test data simulated under MLN and M-Logit models

aR^2	**MLN Data**	**M-Logit Data**
MLN Fit	0.1140	0.1712
M-Logit Fit	0.2077	0.2356

Table C.6.
Sum of Compositional Errors, SCE, for the posterior predictive mean of 105 predicted seasons of test data simulated under MLN and M-Logit models

SCE	**MLN Data**	**M-Logit Data**
MLN Fit	66.68	54.38
M-Logit Fit	68.44	55.94

There is not a great deal of difference between the results of the two metrics for the predicted player-seasons when compared to the results on the training data, except perhaps that Aitchison's R^2 is somewhat larger for the M-Logit data being predicted in both the MLN and M-Logit fits, and that the SCE for the both sets of data being predicted is now smaller under the MLN fit than under the M-Logit fit. Recall that these last two measures, aR^2 and SCE, do not take into account the difference in variability in the models the way DIC and the QQ-plot methods do.

C.3 Real Data Analysis

For the training data sets under both fits:

Table C.7.
Aitchison's R^2 and Sum of Compositional Errors for three category real data training set fit under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
aR^2	0.434	0.475
SCE	3232.73	3216.08

Using the posterior predicted means of the out-of-sample player-seasons, we may also calculate aR^2 and SCE for the two model fits.

Table C.8.
Aitchison's R^2 and Sum of Compositional Errors for three category real data test set under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
aR^2	0.652	0.815
SCE	70.97	74.96

For the ten category training data:

Table C.9.
Aitchison's R^2 and Sum of Compositional Errors for ten category real data training set fit under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
aR^2	0.324	0.441
SCE	11222.54	11040.75

Ten category test data:

Table C.10.
Aitchison's R^2 and Sum of Compositional Errors for ten category real data test set under MLN and M-Logit models

	MLN Model Fit	**M-Logit Model Fit**
aR^2	0.345	0.388
SCE	273.48	282.35

D. ADDITIONAL MODEL ESTIMATION RESULTS

D.1 Simulation Study

Implied Correlation Matrices

This represents, unconditional on the random effects, the covariance matrix of the vector $(W_{ijHR}, W_{ijBB}, W_{ijOther}, W_{ij'HR}, W_{ij'BB}, W_{ij'Other})$, for player i and seasons j and j'.

$$Corr(W^{\Lambda}_{jj'}) = \begin{pmatrix} 1 & 0.21 & -0.46 & 0.64 & 0.16 & -0.32 \\ 0.21 & 1 & -0.96 & 0.16 & 0.34 & -0.35 \\ -0.46 & -0.96 & 1 & -0.32 & -0.35 & 0.40 \\ 0.64 & 0.16 & -0.32 & 1 & 0.21 & -0.46 \\ 0.16 & 0.34 & -0.35 & 0.21 & 1 & -0.96 \\ -0.32 & -0.35 & 0.40 & -0.46 & -0.96 & 1 \end{pmatrix}$$

$$Corr(W^{\hat{\Lambda}}_{jj'}) = \begin{pmatrix} 1 & 0.20 & -0.46 & 0.57 & 0.15 & -0.29 \\ 0.20 & 1 & -0.96 & 0.15 & 0.34 & -0.35 \\ -0.46 & -0.96 & 1 & -0.29 & -0.35 & 0.40 \\ 0.57 & 0.15 & -0.29 & 1 & 0.20 & -0.46 \\ 0.15 & 0.34 & -0.35 & 0.20 & 1 & -0.96 \\ -0.29 & -0.35 & 0.40 & -0.46 & -0.96 & 1 \end{pmatrix}$$

D.2 Real Data Analysis

Fixed Effects

Below we present first the posterior means for the fixed effects under the MLN model fit, followed by the posterior means of the fixed effects under the M-Logit model fit, for the ten category real data training set.

Table D.1.
Posterior mean of fixed effects under MLN fit of ten-category real data training set

	$\beta_{.1}$	$\beta_{.2}$	$\beta_{.3}$	$\beta_{.4}$	$\beta_{.5}$	$\beta_{.6}$	$\beta_{.7}$	$\beta_{.8}$	$\beta_{.9}$
$\beta_{0.}$	-1.21	-2.49	-4.99	-3.01	-5.10	-4.24	-1.82	-4.43	-1.08
$\beta_{height.}$	-0.01	0.02	-0.02	0.16	-0.19	0.08	0.05	0.00	0.10
$\beta_{weight.}$	0.01	0.09	-0.15	0.17	-0.32	0.06	0.09	0.17	0.19
$\beta_{age.}$	-0.22	0.02	-0.68	0.19	-0.22	0.01	0.12	0.95	-0.41
$\beta_{age^2.}$	0.23	0.00	0.51	-0.12	0.05	0.04	-0.03	-0.92	0.37
$\beta_{left.}$	-0.01	-0.09	0.10	-0.09	-0.15	0.00	0.12	-0.34	-0.03
$\beta_{switch.}$	0.02	0.02	0.28	-0.16	0.16	0.11	0.26	-0.11	0.10
$\beta_{cl.}$	0.03	-0.12	-0.49	0.27	-0.35	-0.29	-0.19	0.14	0.04
$\beta_{nl.}$	0.02	0.02	0.06	-0.11	0.03	-0.04	-0.02	0.01	-0.03
$\beta_{pl.}$	0.03	-0.01	-0.49	0.36	-0.67	-0.17	-0.05	0.27	-0.05
$\beta_{of.}$	0.03	0.04	0.22	0.11	-0.05	-0.02	-0.06	0.12	0.02
$\beta_{dh.}$	0.01	0.11	-0.18	-0.07	-2.08	0.07	-0.04	0.18	0.36
$\beta_{p.}$	-0.15	-0.40	0.47	-1.12	3.87	-0.05	-0.57	-0.29	0.93

Table D.2.
Posterior mean of fixed effects under M-Logit fit of ten-category real data training set

	$\beta_{.1}$	$\beta_{.2}$	$\beta_{.3}$	$\beta_{.4}$	$\beta_{.5}$	$\beta_{.6}$	$\beta_{.7}$	$\beta_{.8}$	$\beta_{.9}$
$\beta_{0.}$	-1.17	-2.48	-4.79	-3.04	-4.73	-4.23	-1.85	-4.45	-1.03
$\beta_{height.}$	-0.02	0.03	-0.02	0.18	-0.19	0.08	0.05	0.09	0.16
$\beta_{weight.}$	0.03	0.06	-0.11	0.05	-0.22	-0.03	0.06	0.10	0.09
$\beta_{age.}$	0.03	0.13	-0.15	0.58	-1.18	0.18	0.29	0.20	-0.20
$\beta_{age^2.}$	-0.04	-0.16	-0.09	-0.59	1.03	-0.14	-0.27	-0.21	0.19
$\beta_{left.}$	0.00	-0.10	0.09	-0.09	-0.22	0.03	0.07	-0.16	-0.07
$\beta_{switch.}$	-0.03	-0.06	0.06	-0.19	0.01	-0.03	0.10	-0.10	-0.13
$\beta_{cl.}$	0.06	0.07	-0.39	0.60	-0.42	-0.08	0.04	0.43	0.10
$\beta_{nl.}$	-0.01	0.04	0.02	0.00	-0.14	0.02	0.07	0.09	0.03
$\beta_{pl.}$	0.01	0.09	-0.43	0.58	-0.53	-0.08	0.14	0.47	0.02
$\beta_{of.}$	0.00	-0.01	0.11	0.03	-0.08	-0.05	-0.06	0.04	0.01
$\beta_{dh.}$	0.00	0.02	-0.06	-0.11	-0.18	0.10	0.08	0.07	0.14
$\beta_{p.}$	-0.24	-0.62	-1.16	-1.58	3.62	-1.19	-0.71	-1.03	0.82

E. ADDITIONAL PREDICTION RESULTS

E.1 Simulation Study

Prediction intervals for player 1 and 2 are presented in table and radar chart form.

Table E.1.
95% prediction interval for the posterior predicted mean of new player $i' = 1$ season $j' = 10$

95% **pred int**	**HR**	**BB**	**Other**
lower bound	0	1	222
upper bound	13	22	248

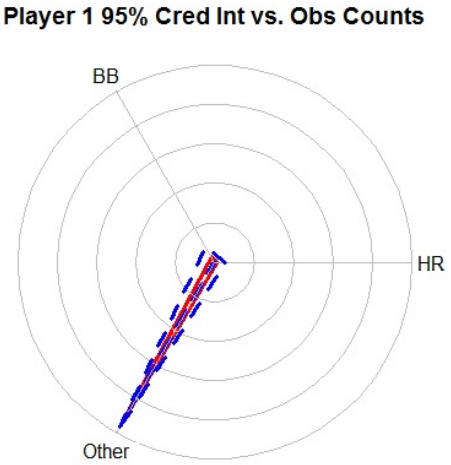

Figure E.1. Radar plot showing 95% prediction interval (blue dashed lines) for the posterior predicted mean versus the observed count (solid red line) of player $i' = 1$

Table E.2.
95% prediction interval for the posterior predicted mean of new player $i' = 2$ season $j' = 4$

95% **pred int**	**HR**	**BB**	**Other**
lower bound	0	5	209
upper bound	29	57	270

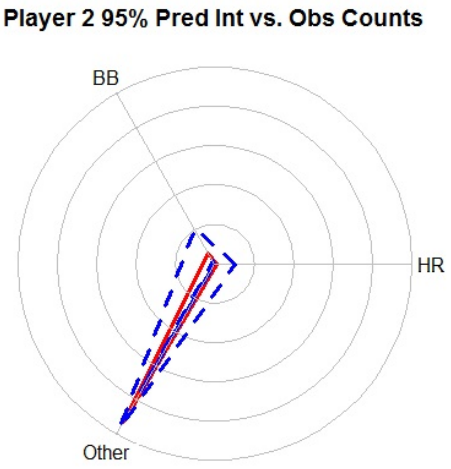

Figure E.2. Radar plot showing 95% prediction interval (blue dashed lines) for the posterior predicted mean versus the observed count (solid red line) of player $i' = 2$

For both players, the prediction interval covers the true value, and it is plain to see the difference in uncertainty. However, the width of the interval is not terribly useful; telling a manager that a player is predicted to hit between 0 and 29 home runs over 279 plate appearances this season is somewhat reductive. There is also quite a bit of overlap between the two player's prediction intervals, making comparing the relative uncertainty difficult. This motivates our predominant use of the credible intervals about the player's expected performance.

E.2 Real Data Analysis

Predicted versus Observed plots for the ten category study:

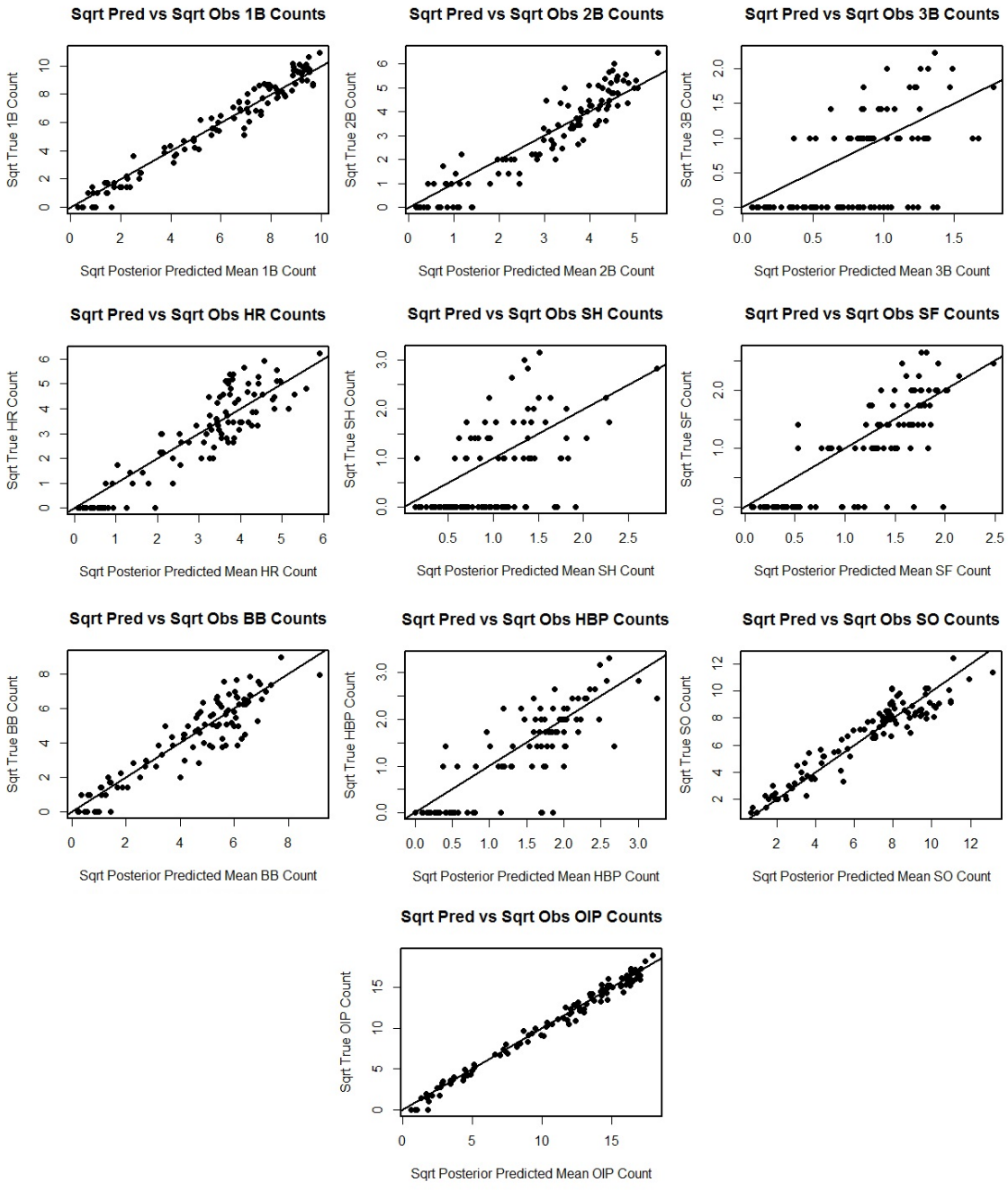

Figure E.3. Predicted versus observed counts of each category for out of sample players from MLN fit of ten categories

VITA

Eric Gerber was born in Chapel Hill, North Carolina. He received a B.A. in Mathematics and International Studies from University of North Carolina at Asheville in December 2012. In August 2013, he entered the Statistics graduate program at Purdue University. He graduated with his M.S. in Statistics in May 2016, and then continued in the PhD program. After receiving his PhD in August 2019, he will transition to an Assistant Professor position at California State University, Bakersfield.