

# Modeling Baseball Player Ability with a Nested Dirichlet Distribution

Brad Null \*  
Stanford University

October, 2008

## Abstract

In this paper we introduce the nested Dirichlet probability distribution and propose a method of using it to model Major League Baseball (MLB) player abilities. To do so, we define fourteen distinct outcome types for any typical plate appearance (excluding intentional walks and bunt attempts), and we assume that every player has an underlying fourteen dimensional ability vector,  $x$ , where each element represents the probability that the player will experience the corresponding outcome type in any typical plate appearance. We then use the method of maximum likelihood to fit a nested Dirichlet prior joint distribution on  $x$  for all MLB batters (excluding pitchers) over the period from 2003-2006.

As the nested Dirichlet (like the Dirichlet distribution) is conjugate prior to multinomial data, this model yields a nested Dirichlet posterior distribution for all players as well. We also present extensions to incorporate age effects and year-to-year variance in player underlying abilities to improve the model's predictive power while maintaining a nested Dirichlet posterior, leading to surprising new evidence that the underlying abilities of players (not just their statistical performances) are mean-reverting in some sense. We also evaluate the posteriors generated by this extended model as a forecasting tool versus 2007 results, showing that the model's accuracy is competitive with popular projection systems, and that the model demonstrates a reasonable estimate of posterior uncertainty. Finally, we discuss some further ideas for extending the model as well as some key applications.

---

\*Department of Management Science and Engineering, Stanford University. Email: null@stanford.edu .

# 1 Introduction

The game of baseball (like many sports) is rife with areas for stochastic modeling and optimization. On the modeling front a few values of particular interest include the likelihood of different play outcomes, the likelihood of different game outcomes, and the distribution of season results for players and teams. With respect to decision making, there is potential value in evaluating play related decisions (e.g. bunts, steals, etc.), deciding when and who to substitute, and evaluating trades and free agent signings. To address any of these problems with respect to a specific team or situation, we must first address a more fundamental question: How do we model and estimate the ability of a specific player? It is this fundamental question we concern ourselves with in this paper.

There are many readily available systems of player forecasts (see [13] for a comparison of the performance of a representative sample of them). Unfortunately though, none of these systems provides all of the elements we would like to have in order to address each of the modeling and optimization problems above. In addition to the two key objectives of any model: to provide accurate predictions and avoid bias; in order to address the problems stated in the previous paragraph, we would like the model to have the following:

1. Granularity - We need a likelihood for specific event types (not just aggregate statistics like batting average). To model plate appearances, we need to know about all possible outcomes.
2. Prior and posterior distribution for all players (not just a point estimate) - This is especially important when modeling the distribution of outcomes over entire seasons.
3. A model of the relationship between event types - If two abilities (e.g. strikeout rate and walk rate) are correlated, we should account for that.
4. Transparency - We want to know what the model is doing. This way we can continue to discuss it, diagnose it, and improve it.
5. Updatability - The model should be easily updated. The state of information during a baseball season changes every day, so we would like to be able to update the model every day as well.

Given a model with each of these features, we can not only address each of the modeling and optimization problems above, but we would also be equipped with a model of value in its own right to teams, fantasy players, consumers and gamblers. The nested Dirichlet based model we present here is such a model.

## 1.1 Our approach

We begin by identifying several possible event types for any plate appearance. This is by no means a definitive list, and it is open for debate whether or not this is the “right” way to segment event outcomes. The event types we work with here are:

1. Int - the play resulted in catcher’s interference.
2. HBP - the play resulted in the batter being hit by the pitch.
3. BB - the play resulted in a base on balls.
4. K - the play resulted in a strikeout.

5. HR,FB - the play resulted in a fly ball home run.
6. 3B,FB - the play resulted in a fly ball triple.
7. 2B,FB - the play resulted in a fly ball double.
8. 1B,FB - the play resulted in a fly ball single.
9. out,FB - the play resulted in a fly ball out.
10. HR,GB - the play resulted in a ground ball home run.
11. 3B,GB - the play resulted in a ground ball triple.
12. 2B,GB - the play resulted in a ground ball double.
13. 1B,GB - the play resulted in a ground ball single.
14. out,GB - the play resulted in a ground ball out.

Given such a list of types, we assume that each batter has some “true” probability of experiencing each event type. We represent these probabilities via an ability vector of variables for each player, and we assume there is some joint prior distribution over these ability variables.

For our purposes, we will assume the prior distribution of these abilities follows a nested Dirichlet joint distribution. This distribution is introduced in [11]. It is an extension of the Dirichlet and Generalized Dirichlet distributions, and like those distributions is conjugate prior to multinomial evidence. We also develop two extensions to the nested Dirichlet distribution to model how a player’s underlying abilities change over time according to an aging process and a random noise process (with mean-reversion) respectively.

Using historical data, we parameterize our underlying distribution as well as the stochastic process extension using the method of maximum likelihood. We parameterize the aging model using minimum squared error estimation. We then use the model and data to represent our understanding of a player’s ability via a nested Dirichlet posterior distribution.

## 1.2 Our contribution

We present a family of nested Dirichlet distributions which are more flexible and natural than both the Dirichlet and Generalized Dirichlet, while remaining conjugate prior to multinomial data, as well as a method for partially optimizing over alternative nestings by maximizing likelihood. We also show how to extend a model using this prior distribution to incorporate deterministic and random effects in the variables over time while still remaining conjugate to the data.

In applying the model to baseball, our method of estimating prior and posterior distributions of player abilities as well as age and noise effects on a granular level leads to an improved understanding of ability distribution and age and noise effects in baseball both in general and on this granular level. Most notably, we present surprising new evidence that the underlying abilities of players (not just their statistical performances) are mean-reverting in some sense. Finally, evaluating the model versus test data indicates that its posterior distributions are unbiased in several important senses, and forecasts derived from the model are competitive with the best available proprietary systems.

### 1.3 Related work and organization

Only recently have Bayesian approaches been undertaken to attempt to separate the inherent randomness of statistical performance from true player ability in baseball. In particular, two approaches most directly parallel the approach of this paper. Albert [2, 3] models player abilities using nested Beta distributions (which are a de facto form of nested Dirichlet modeling). Unlike our approach, Albert uses a full Bayesian hierarchical model as opposed to the method of maximum likelihood (although given the quantity of data used here, this difference should have minimal impact). Also, Albert uses a static nesting (as opposed to evaluating over a class of potential nestings) and a much smaller set of event types. Tango et al. [17] model the prior distributions of player ability using normal distributions. In their Marcel model, there is no evidence of nesting (everything appears to be modeled marginally) and a heuristic parametrization scheme is used to fit the distribution.

Numerous other approaches have been taken to predict future performance of players or teams. One interesting direction among these involves examining the correlation coefficients of player performance from year-to-year with respect to different metrics, see for example [4]. These coefficients can be seen as aggregating the effects of the distribution of player ability, the variance in a player's ability over time, and variance with respect to binomial observations of actual plate appearances. Our analysis by comparison attempts to decompose these effects and account for each of these influences separately.

One of the extensions parameterized in section 5 involves incorporating age effects. Several researchers have observed how ability and performance change somewhat predictably with age. The methods proposed to quantify these aging effects include the Delta method [14], fixed-effects asymmetric models [7], and Bayesian quadratic models [1]. We use the model of [7], with the difference that rather than estimating these effects on the level of aggregate statistics, we estimate conditional effects for each event type separately.

The rest of this paper is organized as follows. Section 2 introduces the nested Dirichlet distribution. Section 3 presents one application of this distribution to baseball and section 4 evaluates forecasts derived from this model. Based upon these results, section 5 introduces a couple of extensions to the model and reviews forecasts derived from these versions. Finally, section 6 extends the model to predict all player statistics and compares these versus popular forecasting systems, and section 7 presents a few concluding remarks.

## 2 The nested Dirichlet distribution

Using prior and posterior distributions to model player ability enables us to not only derive point estimates of a player's ability at any point in time, but some understanding of the uncertainty associated with that estimate as well, even if we have no prior evidence for that player. Additionally, the prior helps explain the regression to the mean phenomenon frequently observed in baseball and elsewhere (see [12] for a discussion of regression to the mean in baseball). The main issue then is to figure out the best model for this prior. There are limitless alternatives, but one attractive property is a a prior distribution that is conjugate to the data.

## 2.1 Conjugate prior and the Dirichlet distribution

For a conjugate prior distribution, the posterior distribution (after incorporating the evidence) is of the same family as the prior distribution. For example, the evidence data in this application is multinomial (e.g. 10 strikeouts, 5 walks, 5 home runs). The most common conjugate prior to multinomial data is the Dirichlet distribution. The joint probability distribution function of the Dirichlet distribution is:

$$D(x_1, \dots, x_n; \alpha_1, \dots, \alpha_n) = \frac{1}{B(\alpha)} \prod_{i=1}^n x_i^{\alpha_i - 1}$$

where the  $x$  vector represents the underlying probability of each event type for any plate appearance, the  $\alpha$  vector represents the parameters, and  $B(\alpha)$  is the multinomial beta function:  $B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(A)}$  with  $A = \sum_{i=1}^n \alpha_i$ .

Given this distribution, and an  $n$  dimensional evidence vector  $e$  for a given player, the posterior distribution for that player is the Dirichlet with parameter vector  $(\alpha + e)$ . This posterior is still Dirichlet (i.e. conjugate) and is also very easy to update.

### 2.1.1 Limitations of the Dirichlet

The Dirichlet does have a few limitations however. Namely, it severely restricts the covariance matrix of the variables in the prior distribution. For example, any two variables with the same mean must have the same variance, and in general, there is a direct relationship between the mean of a variable,  $\pi_i$ , and its variance,  $\sigma_i^2$ , represented by the equation:

$$\sigma_i^2 = \frac{\pi_i(1 - \pi_i)}{A + 1}$$

Thus, the Dirichlet does not always fit reality well. For example, using Major League data from 2003-2006 an analysis of the marginal distribution of strikeout (K) rates of batters indicates a mean of 18.41% and a standard deviation of 5.74%. Likewise, marginal analysis indicates that the underlying distribution of fly ball singles (1B,FB) has mean approximately 8.51% and standard deviation approximately 1.27%.<sup>1</sup> If we were to model these as part of a Dirichlet distribution and wanted to fit these mean and variances, we must have  $\sigma_4^2 = \frac{\pi_4(1-\pi_4)}{A+1}$  and  $\sigma_8^2 = \frac{\pi_8(1-\pi_8)}{A+1}$ , or in other words  $A = 44.6 = 480.7$ , which we obviously cannot do. Figures 1 and 2 indicate the actual models of these distributions via Dirichlet versus marginal parametrization. The  $A$  value corresponding to this Dirichlet distribution is 166.7. Observe how the Dirichlet “splits the difference” in fitting the variance for these two abilities.

Another issue with the Dirichlet is that it will model negative covariance between all pairs of distinct variables (i.e.  $i \neq j$ ), according to the formula:

$$\text{Cov}(x_i, x_j) = \frac{-\pi_i \pi_j}{A + 1}$$

Unfortunately, this is not always the case. For example, figures 3 and 4 compare Dirichlet model implied correlations for batters between some of the variables versus an estimate of correlation calculated by looking at the correlation coefficient of the actual rates of each type for players with

---

<sup>1</sup>The marginal analysis was done by fitting a Beta distribution to this event type alone. We have fit both of these distributions using the method of maximum likelihood discussed in more detail in [11].

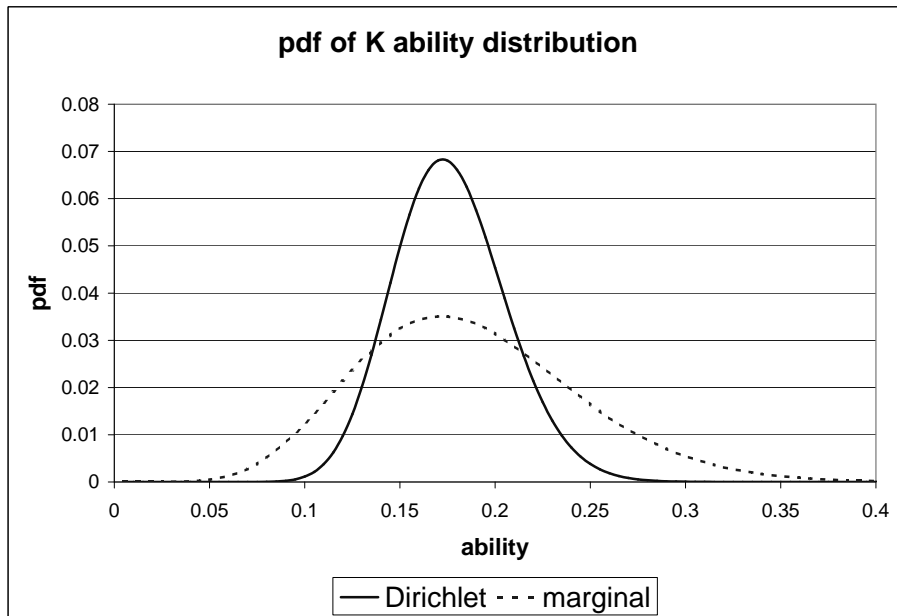


Figure 1: Comparison of marginal Dirichlet and marginal Beta fitted distributions for strikeouts

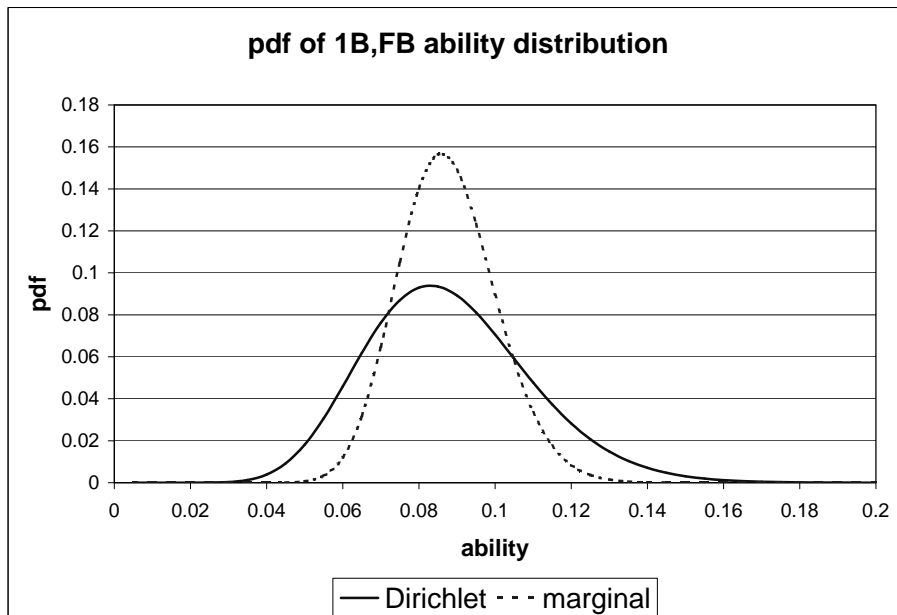


Figure 2: Comparison of marginal Dirichlet and marginal Beta fitted distributions for fly ball singles.

a sufficient number of plate appearances. The data indicate that strikeouts, walks, and fly ball home runs are all positively correlated. However, the Dirichlet is unable to model this.

type	BB	K	HR,FB
BB	1	-0.13	-0.05
K	-0.13	1	-0.07
HR,FB	-0.05	-0.07	1

Figure 3: Dirichlet model implied correlation coefficient for batters

type	BB	K	HR,FB
BB	1	0.16	0.33
K	0.16	1	0.33
HR,FB	0.33	0.33	1

Figure 4: Observed correlation for batters

## 2.2 The nested Dirichlet distribution

Fortunately, the Dirichlet is not the only conjugate prior for multinomial data. A more flexible prior distribution can be created using nested Dirichlet distributions. A nested Dirichlet distribution modifies the Dirichlet distribution by adding a series of variables,  $x_{n+1}, \dots, x_{n+k}$ , to the distribution. We denote these as *nesting variables*. For every nesting variable, some subset of variables is nested underneath that variable in a conditional Dirichlet distribution.

When nesting variable  $x_{n+a}$  is added, we represent by  $X_a$  the set of all variables nested under  $x_{n+a}$ . The new variable,  $x_{n+a}$ , then replaces  $X_a$  in the original Dirichlet distribution. The new variable  $x_{n+a} = \sum_{i \in X_a} x_i$ . Additionally, for all  $i \in X_a$  we define  $\tilde{x}_i = \frac{x_i}{x_{n+a}}$ . We also define  $\tilde{X}_a = \{\tilde{x}_i : i \in X_a\}$ .  $\tilde{X}_a$  is then distributed according to a Dirichlet distribution independent of all other conditional Dirichlets.

It is easiest to visualize the nesting structure of a nested Dirichlet distribution using a *nesting tree*. The tree on the right of figure 5 represents the nesting tree for a simple nested Dirichlet distribution. Here  $x_4$  is a nesting variable which aggregates the original variables  $x_2$  and  $x_3$ . The distribution of the three original variables is now represented by two nested Dirichlet (or in this case Beta) distributions, one on each level of the tree, as opposed to a single Dirichlet according to the tree on the left.

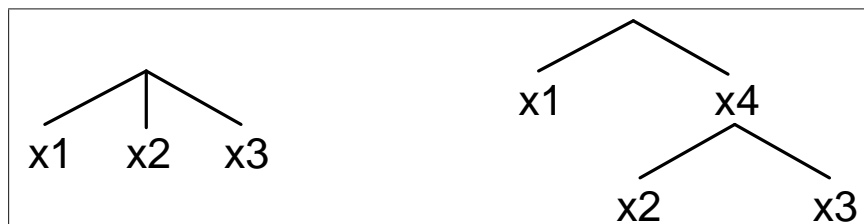


Figure 5: Comparison of the Dirichlet and nested Dirichlet distributions

### 2.2.1 Characterizing the nested Dirichlet distribution

We use the following additional notation to characterize the distribution:

- $n$  original variables (referred to as variables  $1, \dots, n$ )
- $k$  nesting variables, (referred to as variables  $n + 1, \dots, n + k$ )
- $\Theta$  -  $k + 1$  sets indexed  $\theta_0, \dots, \theta_k$ .
  - $\theta_0$  indicates the indices of all unnested variables.
  - $\theta_i$  for  $i > 0$  indicates variables nested under the  $i$ th nesting variable.
- $A$  -  $k + 1$  vectors indexed  $A_0, \dots, A_k$ , where  $A_i$  corresponds to the vector of parameters for the  $i$ th nesting (in the same order as in  $\theta_i$ ).

We will also indicate by  $\alpha_i$  the parameter associated with the  $i$ th variable and  $I_i$  the nesting variable that variable  $i$  is nested under. If  $I_i = 0$ , variable  $i$  is unnested. Note that  $I_i = m$  iff  $i \in \theta_{m-n}$ . The following summarizes some key notation and properties of the distribution:

- $\tilde{x}$  - an  $n + k$  dimensional conditional probability vector such that  $\tilde{x}_i = x_i/x_{I_i}$  if  $I_i > 0$  and  $\tilde{x}_i = x_i$  otherwise.
- $X_j = \{x_i : I_i = j\}$
- $\tilde{X}_j = \{\tilde{x}_i : I_i = j\}$
- $f(X_0) = f(\tilde{X}_0) = D(X_0; A_0)$
- For  $i \in \{1, \dots, k\}$ ,  $f(\tilde{X}_i) = D(\tilde{X}_i; A_i)$
- If  $I_i = j > 0$ ,  $x_i = \tilde{x}_i * x_j$
- If  $I_i \neq I_j$  then  $\tilde{x}_i$  and  $\tilde{x}_j$  are independent.

The joint pdf of the nested Dirichlet distribution can be represented as:

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

where  $\bar{A}_k = \sum_{j \in \theta_k} \alpha_j$ . Recall that  $x_n, x_{n+1}, \dots, x_{n+k}$  are degenerate. Further, if  $I_i = 0$ , the mean and variance of  $x_i$  follow straight from the general Dirichlet distribution as  $E[x_i] = \frac{\alpha_i}{A_0}$  and

$$\sigma_i^2 = \frac{\alpha_i(\bar{A}_0 - \alpha_i)}{A_0^2(\bar{A}_0 + 1)}.$$

If  $I_i = k > 0$ :

$$E[x_i] = E[\tilde{x}_i]E[x_{n+k}] \tag{1}$$

$$\sigma_i^2 = \tilde{\sigma}_i^2 \sigma_{n+k}^2 + \tilde{\sigma}_i^2 E[x_{n+k}]^2 + E[\tilde{x}_i]^2 \sigma_{n+k}^2 \tag{2}$$

Derivations of these values as well as the covariance and other values related to the variables in the nested Dirichlet distribution are presented in [11].

### 2.2.2 Nested Dirichlet versus Dirichlet and Generalized Dirichlet

The nested Dirichlet distribution is closely related to another conjugate prior extension of the Dirichlet known as the Generalized Dirichlet. The Generalized Dirichlet was derived to allow for a more general covariance matrix (including positive covariance) and counter some of the deficiencies we saw with the Dirichlet in section 2.1.1. (See [6] and [18] for more on the Generalized Dirichlet.)

The Generalized Dirichlet is in fact a subset of the nested Dirichlet. Where the nested Dirichlet allows for any sort of nesting tree, the Generalized Dirichlet allows for only nesting trees where there are two variables in each nesting and at least one is an original variable. Figure 6 depicts two typical nesting trees. While both are valid nested Dirichlet trees, only the tree on the left represents a Generalized Dirichlet distribution. Thus, the nested Dirichlet allows for a more natural and flexible modeling of the relationship of the variables. In addition, the nested Dirichlet allows for an even more general covariance matrix. Results in [11] have shown that estimation using a nested Dirichlet distribution can improve results versus the Dirichlet twice as much as the Generalized Dirichlet.<sup>2</sup>

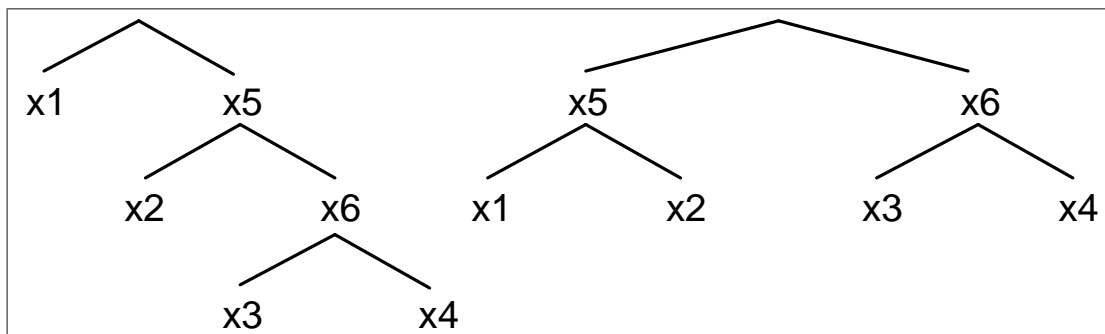


Figure 6: Generalized Dirichlet and non-Generalized Dirichlet trees

One interesting result from [6] is the observation that any Dirichlet distribution can be represented by a Generalized Dirichlet distribution. In [11] this result is extended to show that any nested Dirichlet distribution can be equivalently realized with a nested Dirichlet with  $n - 2$  nesting variables and all nestings conditionally distributed as Beta distributions. This is important because in our application, as is frequently the case, the nesting tree is not known a priori. This generalization greatly simplifies the number of possible nestings to consider. Thus, we will view all distributions as such from here on out.

### 2.2.3 How to fit the distribution

We fit the distribution to data using the method of maximum likelihood. See [11] for details. Basically, for each nesting, we estimate the best parameters and find the conditional likelihood of that nesting. As each nesting is conditionally independent, the likelihood of the entire distribution is the product of the likelihoods of each nesting.

The traditional approach to creating the sorts of hierarchies in the Generalized and nested Dirichlet distributions is to choose a reasonable structure a priori. However, one of our goals is to attempt to mine the data in an effort to find new ways of looking at the relationships between

<sup>2</sup>When the underlying data is generated from a nested Dirichlet distribution.

the various event types in baseball. Thus, as mentioned above, we have intentionally adopted an agnostic view with respect to nesting structure here.

To select among possible nestings we will extend the principle of maximum likelihood by comparing the likelihood of all nestings and choosing the one with maximum likelihood. Experimental results have shown this to be a promising approach [11]. However even with the simplification above, this still leaves more than  $14!$  nestings to compare.

### 3 Application of the nested Dirichlet to baseball

#### 3.1 Our simplification

For our purposes, we narrow the field of possible nestings we will consider by constraining our nestings to include nesting variables representing “fly ball” and “ground ball”. This is to say that among our nesting variables two of them,  $a$  and  $b$ , will be classified as “fly balls” and “ground balls” such that  $x_a = \sum_{j=5}^9 x_j$  and  $x_b = \sum_{j=10}^{14} x_j$ .

Further, we break the 14 original variables plus these two nesting variables into three sets. The first set consists of the first four original variables plus the two added variables. The second set consists of all fly ball component variables (types 5-9), and the final set is all ground ball component variables (types 10-14). Within each of these groupings we assume a Generalized Dirichlet distribution.

To calculate the maximum likelihood estimators for the entire distribution, we first calculate the maximum likelihood estimators and attendant likelihood for each Generalized Dirichlet distribution corresponding to every possible ordering of the conditional variables within a group. For example, given an evidence set consisting of the five columns of “fly ball” component results, we calculate the maximum likelihood estimator of the eight Generalized Dirichlet parameters for each of the  $5!/2$  orderings of these variables (note that we divide by 2 because it does not matter how we order the final two variables). Comparing the likelihood of the evidence data with respect to each ordering and its corresponding set of estimators for both batters and pitchers, we select the ordering with maximum combined likelihood (i.e. the likelihood for batters is multiplied by the likelihood for pitchers) and its corresponding estimators as the parametrization of this subtree of our overall nested Dirichlet distribution.<sup>3</sup>

For simplicity’s sake, we parameterize this as three separate nested Dirichlet distributions,  $O$ ,  $F$ , and  $G$  with the following ordered event types. For  $O$ :

1. Int - catcher’s interference
2. HBP - hit by pitch
3. BB - base on balls
4. K - strikeout
5. FB - fly ball
6. GB - ground ball

---

<sup>3</sup>The question of whether or not to use the same nesting for batters and pitchers is an interesting theoretical issue in its own right, which we do not examine in depth here. For our purposes, we have added this constraint so as to allow for a broader range of models of batter/pitcher interaction, which we analyze in the sequel to this paper.

For  $F$ :

1. HR,FB - fly ball home run
2. 3B,FB - fly ball triple
3. 2B,FB - fly ball double
4. 1B,FB - fly ball single
5. out,FB - fly ball out

And for  $G$ :

1. HR,GB - ground ball home run
2. 3B,GB - ground ball triple
3. 2B,GB - ground ball double
4. 1B,GB - ground ball single
5. out,GB - ground ball out

We represent the distribution this way for convenience only. Our underlying  $x$  variables will still represent fully unconditional event type abilities. Our evidence set consists of all plate appearances for the 2003 through 2006 Major League Baseball season in which the batter was a position player (i.e. not a pitcher) and the outcome of the play did not come about as a result of a bunt attempt or intentional walk.<sup>4</sup> For each player we build the evidence set by counting the number of times he experiences each event type.<sup>5</sup>

### 3.2 Optimized nesting and parameters

Using this method, our maximum likelihood parametrization for this restricted nested Dirichlet for pitcher and batters respectively (we use the superscript  $P$  to indicate parameters for pitchers and  $B$  to indicate those for batters) is as follows.

$$\theta_O = \begin{bmatrix} 1 & 7 \\ 2 & 8 \\ 5 & 9 \\ 6 & 10 \\ 3 & 4 \end{bmatrix}, \quad A_O^P = \begin{bmatrix} 0.9 & 9800.7 \\ 7.3 & 691.0 \\ 36.7 & 50.2 \\ 18.1 & 13.9 \\ 12.3 & 22.5 \end{bmatrix}, \quad A_O^B = \begin{bmatrix} 0.1 & 1045.5 \\ 2.3 & 227.9 \\ 42.4 & 62.5 \\ 14.1 & 10.9 \\ 8.1 & 18.6 \end{bmatrix}.$$

$$\theta_F = \begin{bmatrix} 5 & 6 \\ 3 & 7 \\ 1 & 8 \\ 4 & 2 \end{bmatrix}, \quad A_F^P = \begin{bmatrix} 132.0 & 83.1 \\ 304.8 & 858.5 \\ 38.7 & 118.1 \\ 41.0 & 738.6 \end{bmatrix}, \quad A_F^B = \begin{bmatrix} 104.7 & 66.5 \\ 77.5 & 218.6 \\ 3.3 & 11.9 \\ 2.5 & 44.1 \end{bmatrix}.$$

---

<sup>4</sup>The play-by-play data used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at “[www.retrosheet.org](http://www.retrosheet.org)”.

<sup>5</sup>In this application, we define a switch hitter as two different players, one for each side of the plate. It remains to be seen whether or not this is the best way to approach this issue.

$$\theta_G = \begin{bmatrix} 5 & 6 \\ 4 & 7 \\ 3 & 8 \\ 2 & 1 \end{bmatrix}, \quad A_G^P = \begin{bmatrix} 626.6 & 200.2 \\ 281.2 & 26.3 \\ 219.1 & 10.9 \\ 293.2 & 2.5 \end{bmatrix}, \quad A_G^B = \begin{bmatrix} 388.6 & 121.1 \\ 152.1 & 15.1 \\ 10.0 & 0.5 \\ 985.7 & 8.4 \end{bmatrix}.$$

Figure 7 gives the tree representation of this nesting. We emphasize at this point that by no means do we intend to claim that this is *the* distribution of player abilities, and in fact the discussion of correlation coefficients below shows that we still have work to do, but our results do lead to some interesting insights. For example, whereas traditional models of event type relationships in baseball have always branched off BB and K at the top of the tree and then look at batted balls conditionally, this approach indicates that it is perhaps more effective to do the analysis the other way around.

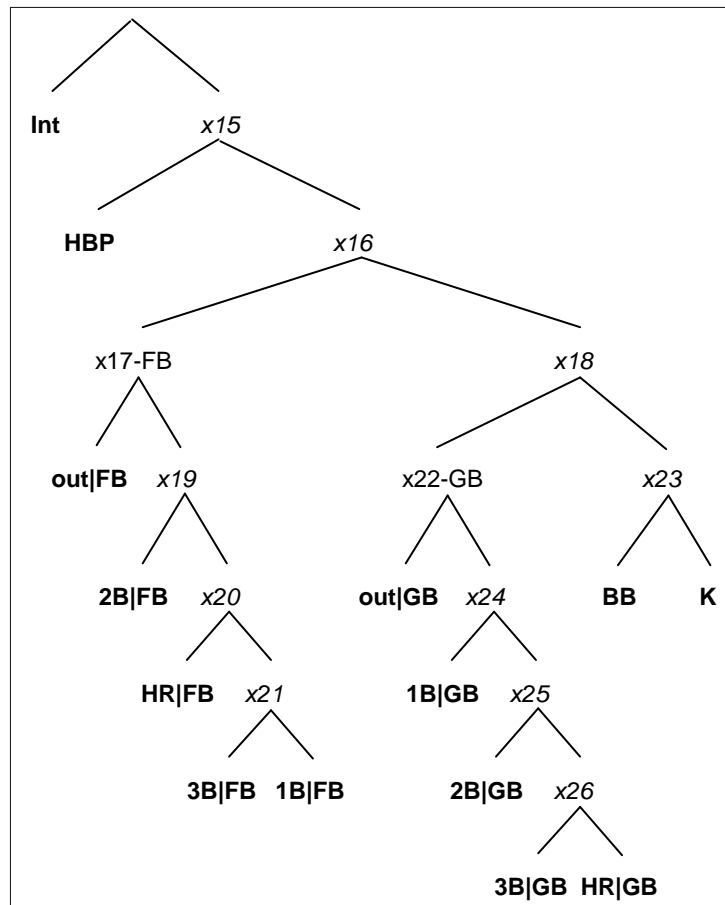


Figure 7: Tree representation of MLE nesting

We summarize the unconditional prior means and standard deviations of the original variables in the nested Dirichlet distribution in figure 8.<sup>6</sup> We also present the standard error of these estimates in

<sup>6</sup>There has been much discussion recently concerning which results batters and pitchers have control over (see [3, 2, 10]). The distributions derived here and the summarizing statistics in figure 8 should have relevance to that discussion, though we do not delve into the debate in detail here.

figure 8 using Cramer-Rao bounds (the methodology for calculating these is also discussed in [11]). Notice that the Cramer-Rao bounds are very tight.

type	model implied		CR implied SE	
	mean	SD	mean	SD
Int	0.01%	0.02%	0.00%	0.01%
HBP	0.99%	0.65%	0.03%	0.03%
BB	7.81%	2.97%	0.11%	0.07%
K	17.95%	4.85%	0.18%	0.10%
HR,FB	2.51%	1.25%	0.05%	0.04%
3B,FB	0.48%	0.31%	0.01%	0.01%
2B,FB	4.07%	0.74%	0.04%	0.02%
1B,FB	8.50%	1.77%	0.07%	0.03%
out,FB	24.48%	3.26%	0.13%	0.08%
HR,GB	0.00%	0.00%	0.00%	0.00%
3B,GB	0.03%	0.05%	0.00%	0.00%
2B,GB	0.68%	0.23%	0.01%	0.01%
1B,GB	7.17%	1.51%	0.06%	0.03%
out,GB	25.31%	4.88%	0.17%	0.10%

Figure 8: Summary statistics for batters

**Comparison of correlation coefficients** Figure 9 displays a subset of the model implied correlation coefficients for our variables. Notice that unlike the Dirichlet, the nested Dirichlet implies a positive correlation between walks and strikeouts. Unfortunately though, our model was unable to match a positive correlation between these two variables and the fly ball home run variable. This indicates that perhaps the subset of nestings we have selected is too restrictive. We present some ideas for how to improve the model’s performance with respect to this issue in section 7.

type	BB	K	HR,FB
BB	<b>1</b>	<b>0.18</b>	<b>-0.05</b>
K	<b>0.18</b>	<b>1</b>	<b>-0.07</b>
HR,FB	<b>-0.05</b>	<b>-0.07</b>	<b>1</b>

Figure 9: Nested Dirichlet model implied correlation coefficients for batters

## 4 Forecasting

We now evaluate the predictive power of the nested Dirichlet model applied to baseball by using data on all results from the 2007 season as a test set. Given our evidence set and distribution parameters, we have a posterior joint probability distribution of the ability vector for any player,  $i$ , parameterized by the  $n+k$  dimensional vector  $\alpha^i$  (again see [11] for computational details). We will summarize the test data in a matrix  $f$  where the  $i, j$ th element  $f_{ij}$  represents the number of times player  $i$  experiences event type  $j$ . We define these values for nesting variables as well. We will let

$\pi_j^i = \frac{\alpha_j^i}{\sum_{m \in X_{I_j}} \alpha_m^i}$ ,  $F(i) = \sum_{j=1}^n f_{ij}$ , and  $F_j(i) = \begin{cases} F(i) & \text{if } I_j = 0 \\ f_{i,I_j} & \text{else} \end{cases}$  (i.e.  $F(i)$  is the total number of plate appearances for player  $i$  in the test set and  $F_j(i)$  is the total number of plate appearances for player  $i$  with a result in the sub-nesting tree rooted at  $I_j$ ).

To make a forecast using the model we begin by taking  $F(i)$  as given for all  $i$  (i.e. we know the number of plate appearances for all players). We then estimate the expectation and variance of the quantity of each event type  $j$  for each player  $i$ ,  $E[f_{ij}|F(i)]$  and  $\sigma^2[f_{ij}|F(i)]$  respectively. The formulae for these are:

$$E[f_{ij}|F(i)] = \pi_j^i E[F_j(i)] \quad (3)$$

$$\sigma^2[f_{ij}|F(i)] = \pi_j^i (1 - \pi_j^i) \frac{E[F_j(i)^2] + E[F_j(i)](\bar{A}_j)}{\bar{A}_j + 1} + (\pi_j^i)^2 \sigma^2[F_j(i)] \quad (4)$$

where  $\bar{A}_j = \sum_{m \in X_{I_j}} \alpha_m^i$  and  $[F_j(i)]$  implies  $[F_j(i)|F(i)]$ . These values can be calculated iteratively down the nesting tree. Observe that the first part of  $\sigma^2[f_{ij}|F(i)]$  is the mean of the conditional variance, while the second part is the variance of the conditional mean.

Barry Bonds (i=225); F(225)=434										
type	j1	j2	I <sub>j</sub>	$\alpha_{j1}^i$	$\alpha_{j2}^i$	$\pi_{j1}^i$	$E[f_{ij1} F(i)]$	$\sigma[f_{ij1} F(i)]$	$E[f_{ij2} F(i)]$	$\sigma[f_{ij2} F(i)]$
Int	1	15	0	0.1	2969.3	0.00	0.0	0.2	434.0	0.0
HBP	2	16	15	31.3	1688.0	0.02	7.9	3.1	426.1	3.1
FB	17	18	16	727.4	838.4	0.46	197.9	11.7	228.1	11.7
GB	22	23	18	352.1	448.9	0.44	100.3	9.9	127.9	10.7

Figure 10: Forecasted results for Barry Bonds ( $j = 225$ )

Figure 10 summarizes this information for a selection of variables pertaining to Barry Bonds. The  $j1$  and  $j2$  columns indicate both the index of the noted variable as well as the index of the variable nested with it in figure 7.  $I_j$  indicates the index of the variable they are nested underneath. Note that all of these values are the same for all players, while the remaining information is player specific. Each of the variables listed here are nested directly underneath one of the variables on the row above. Thus, the mean and standard deviations calculated on the righthand side of each row are functions of the  $\alpha$  and  $\pi$  terms on that row as well as the rightmost mean and standard deviation terms on the row above.

#### 4.1 Forecasts of event type counts by player

In Figure 11 we analyze the weighted root mean squared error (RMSE) of the nested Dirichlet model for each event type where the error for any player  $i$  and event type  $j$  is determined as  $\frac{f_{ij} - E[f_{ij}]}{F(i)}$  and these are weighted by  $F(i)$ . So for example, since Barry Bonds had 3 HBP's in 2007, his HBP error is calculated as  $\frac{3-7.9}{434} = -1.1\%$ .

In addition to our nested Dirichlet, we calculate these values for a naive model with all players projected to perform with the population historical average (Hist) and a model with all players projected to their personal historical average (or population average if no experience) (Player). The nested Dirichlet model significantly outperforms the naive approaches.

type	actual%	min E RMSE	ND E RMSE	RMSE		
				ND	Hist	Player
Int	0.01%	0.06%	0.07%	0.07%	0.07%	0.06%
HBP	0.97%	0.63%	0.72%	0.75%	0.90%	1.17%
BB	8.13%	1.71%	2.10%	2.31%	3.24%	2.96%
K	16.89%	2.41%	3.05%	3.83%	5.99%	4.43%
HR,FB	2.75%	1.03%	1.19%	1.31%	1.71%	1.38%
3B,FB	0.48%	0.45%	0.49%	0.53%	0.57%	0.62%
2B,FB	4.41%	1.27%	1.36%	1.38%	1.43%	1.80%
1B,FB	8.73%	1.79%	2.00%	1.93%	2.11%	2.33%
out,FB	25.29%	2.75%	3.16%	3.65%	4.91%	4.22%
HR,GB	0.00%	0.01%	0.01%	0.00%	0.00%	0.00%
3B,GB	0.04%	0.12%	0.12%	0.15%	0.15%	0.20%
2B,GB	0.65%	0.53%	0.55%	0.52%	0.52%	0.65%
1B,GB	7.32%	1.65%	1.80%	1.99%	2.47%	2.41%
out,GB	24.32%	2.77%	3.34%	3.88%	5.36%	4.82%
FB	41.67%	3.13%	3.76%	4.08%	5.54%	4.72%
GB	32.33%	2.99%	3.82%	4.55%	6.82%	5.46%
average	10.87%	1.456%	1.721%	1.933%	2.612%	2.328%

Figure 11: RMSE of player event type predictions

In order to further benchmark our results, we have also estimated the minimum expected RMSE (min E RMSE) for each event type and the nested Dirichlet model implied expected RMSE for each type (ND E RMSE). To calculate these expected errors, we observe first that the expected squared error for each player is simply the variance of his event frequency given his total number of plate appearances or  $\frac{\sigma^2[f_{ij}|F(i)]}{F(i)}$ . Calculating this for each player, we then add them together, weighting by  $F(i)$  and divide this sum by the total number of plate appearances,  $\sum_i F(i)$ , to get the expected mean squared error. Taking the square root gives us the median (not the mean) expected root mean squared error.

The approximate minimum expected error provides a rough lower bound on the performance with respect to this metric (in this circumstance, an error close to zero is obviously unreasonable), though the bound is not necessarily tight. Perhaps more important is the model implied expected error which provides an indication of how reasonable the model is. The observed error is systematically higher than the model implied error indicating that the model systematically underestimates the posterior variance of player abilities.

## 4.2 Segmentation

To check for further bias in the model, we segment the population by age and then by ability and aggregate the forecasts for players in each segment. Our results indicate forecasts from this model exhibit bias with respect to each of these segmentations.

### 4.2.1 Age bias

In figure 12 we have separated players into five groups by age. In order to get a clearer picture of age effects, we have compared model implied performance according to the aggregate metric OPS versus actual OPS.<sup>7</sup> The results in figure 12 clearly indicate that the model under-predicts performance for young players and over-predicts performance for older ones. Further, using the model implied variance of the expected OPS for each group, we calculate that actual performance for each group is at least two standard deviations away from the predicted performance.

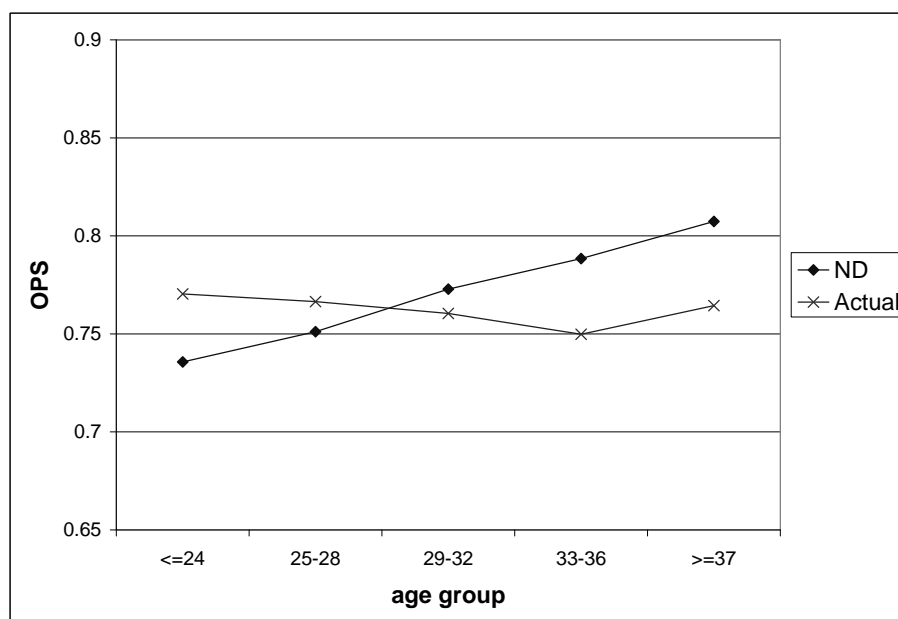


Figure 12: Actual versus predicted OPS by age

### 4.2.2 Ability bias

The second potential area of bias we examine is whether or not the nested Dirichlet model over- or under-estimates the performance of good or bad players. The general formulation of the model should take care of the regression to the mean phenomenon in actual statistics, but there may still be some sort of mean-reversion in underlying player ability. We examine this effect by sorting all players into five groups according to their predicted OPS. We then look at the difference between the predicted and actual OPS for players in each group. Figure 13 shows the actual performance of each group alongside the predicted performance of each group using the nested Dirichlet model as well as using the two historically based prediction methods.<sup>8</sup> Observe that the nested Dirichlet model seems to capture some but not all of the regression to the mean from historical player stats.

<sup>7</sup>OPS represents on-base percentage plus slugging percentage or  $OBP\% + SLG\%$ , where  $OBP\% = \frac{HBP\%+BB\%+HR\%+3B\%+2B\%+1B\%}{1-Int\%}$  and  $SLG\% = \frac{4*HR\%+3*3B\%+2*2B\%+1B\%}{1-(Int\%+K\%+FB\%+GB\%)}$ .

<sup>8</sup>In order to avoid a selection bias caused by the fact that players that perform well are likely to both be better than we expected and to play more than average, we actually update parameters for these data sets daily for the purpose of these forecasts.

The model implied expected difference between the performance of group 1 and group 5 is .1923. The actual difference is .1718, 1.8 standard deviations away. While this is not generally considered statistically significant in and of itself, we have observed similar effects over several other data sets.<sup>9</sup>

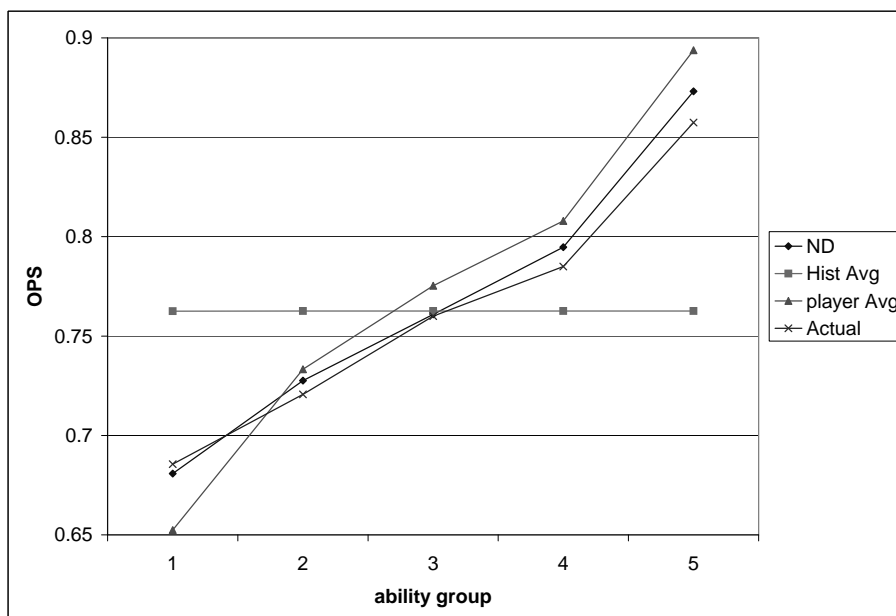


Figure 13: Predicted OPS by estimated ability level

## 5 Model extensions

The results of the previous section indicate three shortcomings of the general nested Dirichlet model:

1. Variance underestimation - the model underestimates uncertainty in player predictions.
2. Age bias - the model over-predicts performance of older players and under-predicts performance of younger players.
3. Ability bias - the model underestimates regression to the mean.

In this section, we propose two model extensions to address these shortcomings and improve the model’s forecasting accuracy.

### 5.1 Age effects

We will first address aging effects. The fact that aging effects appear in the data is not a revelation. In fact, they have been discussed for some time (for instance see [7, 14, 1]). In this section, we will use a fixed effects asymmetric quadratic additive model popularized by Fair [7]. “Fixed” in this context means that every player has the same effect, while “asymmetric” means that the rates at

<sup>9</sup>This is based upon similar analysis of 2005 and 2006 results. For details, please contact the author.

which abilities increase and decrease with age need not be the same. In this model, for a given underlying variable, the ability for player  $i$  in year  $t$  is:

$$\tilde{x}_{ia} = \begin{cases} \tilde{x}'_i + \beta_1 a + \gamma_1 a^2 + \epsilon_{ia} & a \leq \delta \\ \tilde{x}'_i + C + \beta_2 a + \gamma_2 a^2 + \epsilon_{ia} & a \geq \delta \end{cases}$$

where  $\tilde{x}'_i$  is a measure of player  $i$ 's underlying ability,  $C$  is a constant,  $a$  is player  $i$ 's age at the time in question, the  $\beta$  and  $\gamma$  terms are the quadratic coefficients,  $\delta$  is the peak-performance age (could be a maximum or a minimum depending upon the event type), and  $\epsilon_{ia}$  is an error term or random noise. Observe that the underlying abilities we model are the conditional ability variables which were originally of the form  $\tilde{x}_{ij}$  to represent conditional likelihood of event type  $j$  for player  $i$  and now are represented  $\tilde{x}_{ija}$  to represent this ability when  $i$  is age  $a$  with  $\tilde{x}'_{ij}$  representing some baseline ability. In this section we assume  $j$  is understood and drop the notation for the event type unless it is necessary for clarification. We also require that the curve is continuous with zero derivative at  $\delta$ . This further implies that:

$$\begin{aligned} \beta_1 &= -2\gamma_1\delta \\ \beta_2 &= -2\gamma_2\delta \\ C &= (\gamma_2 - \gamma_1)\delta^2 \end{aligned}$$

Although we could use the method of maximum likelihood here as we did with the original model, it is greatly advantageous for us to be able to incorporate significantly more data over a much longer time horizon. Thus, we estimate the parameters using a least squares approach assuming fixed  $\tilde{x}'_i$  values.<sup>10</sup> Thus, if there are  $n$  players in the group, it suffices to estimate the  $n + 3$  parameters:  $\gamma_1, \gamma_2, \delta, \tilde{x}'_i$ . For our purposes, we will be less concerned with these parameters and more concerned with representing player  $i$ 's ability at age  $a$ ,  $\tilde{x}_{ia}$  as a function of his ability at age 28 (the age we have chosen as the normalizing age). We will represent this with an additive factor  $R_a$  such that  $\tilde{x}_{ia} = \tilde{x}_{i,28} + R_a$ .

We fit this model by taking all players from 1956-2005 that have at least 8 seasons of 300+ plate appearances. Then we use those seasons to generate observed  $\hat{x}_{ia}$  terms where for event type  $j$ ,  $\hat{x}_{ia} = \frac{e_{ija}}{e_{iI_j a}}$  where  $e_{ija}$  represents the observed number of times player  $i$  experience event type  $j$  at age  $a$ . We then pick the  $\delta$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$  terms that minimize the squared error, subject to the constraint that the curve is continuous at  $\delta$ .<sup>11</sup> Also, to prevent the model from choosing a relatively extreme  $\delta$  value and subsequently overfitting the few points on the outside of this value, we have modified the algorithm slightly to not permit a  $\delta$  that cuts off less than 40 observations. We do however, allow the model to select a symmetric curve with implied  $\delta$  outside this range. The resulting calculated  $R$  factors for the nested Dirichlet model for several significant event types are in figure 14.<sup>12</sup> Note that, as with strikeouts and walks, variables nested together will have opposite age effects.

<sup>10</sup>One other complication we have glossed over is that as represented, this model allows for probabilities outside the viable range of  $[0, 1]$ . It would be preferable to include some boundary in the model, such as disallowing any  $\tilde{x}_{ia}$  value that moves too close to this boundary. Doing so however would greatly increase the difficulty in estimating the model parameters. Thus, as in this application this issue does not significantly alter any of our results, we will ignore it here.

<sup>11</sup>For more specifics on the algorithm, see [7].

<sup>12</sup>We should also note that for some of the event types, some of the old data is missing some elements, so we have had to approximate the counts. Specifically until 1987 and from 1999 to 2002 we do not have reliable indication of

age	HBP	BB	K	HR,FB	3B,FB	2B,FB	1B,FB	out,FB	1B,GB	out,GB	FB	GB
20	-0.01%	-10.88%	10.88%	-5.89%	2.72%	-1.22%	-2.72%	1.60%	-0.16%	-0.09%	-3.97%	1.50%
21	-0.01%	-8.77%	8.77%	-4.78%	2.36%	-1.03%	-2.36%	0.83%	-0.12%	-0.06%	-3.01%	0.31%
22	-0.01%	-6.87%	6.87%	-3.78%	2.01%	-0.85%	-2.01%	0.28%	-0.08%	-0.05%	-2.17%	0.29%
23	-0.01%	-5.18%	5.18%	-2.88%	1.66%	-0.68%	-1.66%	-0.05%	-0.06%	-0.03%	-1.48%	0.28%
24	-0.01%	-3.71%	3.71%	-2.09%	1.32%	-0.52%	-1.32%	-0.16%	-0.03%	-0.02%	-0.92%	0.25%
25	0.00%	-2.46%	2.46%	-1.41%	0.98%	-0.37%	-0.98%	-0.15%	-0.02%	-0.01%	-0.49%	0.20%
26	0.00%	-1.43%	1.43%	-0.83%	0.65%	-0.24%	-0.65%	-0.12%	-0.01%	0.00%	-0.19%	0.15%
27	0.00%	-0.60%	0.60%	-0.36%	0.32%	-0.11%	-0.32%	-0.07%	-0.01%	0.00%	-0.03%	0.08%
28	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
29	0.00%	0.39%	-0.39%	0.26%	-0.32%	0.10%	0.32%	0.09%	0.02%	0.01%	-0.02%	-0.09%
30	0.00%	0.56%	-0.56%	0.41%	-0.63%	0.19%	0.63%	0.20%	0.06%	0.03%	-0.06%	-0.20%
31	0.00%	0.56%	-0.56%	0.45%	-0.94%	0.27%	0.94%	0.34%	0.10%	0.07%	-0.12%	-0.32%
32	0.00%	0.51%	-0.51%	0.43%	-1.24%	0.34%	1.24%	0.49%	0.16%	0.12%	-0.20%	-0.45%
33	0.01%	0.41%	-0.41%	0.37%	-1.54%	0.40%	1.54%	0.66%	0.23%	0.18%	-0.30%	-0.59%
34	0.01%	0.26%	-0.26%	0.28%	-1.83%	0.44%	1.83%	0.85%	0.32%	0.25%	-0.43%	-0.74%
35	0.01%	0.07%	-0.07%	0.15%	-2.12%	0.48%	2.12%	1.07%	0.41%	0.34%	-0.57%	-0.91%
36	0.01%	-0.16%	0.16%	-0.01%	-2.40%	0.50%	2.40%	1.30%	0.52%	0.44%	-0.74%	-1.09%
37	0.01%	-0.44%	0.44%	-0.20%	-2.68%	0.51%	2.68%	1.55%	0.64%	0.55%	-0.93%	-1.29%
38	0.01%	-0.77%	0.77%	-0.44%	-2.95%	0.51%	2.95%	1.83%	0.78%	0.68%	-1.14%	-1.49%
39	0.00%	-1.14%	1.14%	-0.70%	-3.22%	0.44%	3.22%	2.12%	0.92%	0.82%	-1.38%	-1.71%
40	-0.01%	-1.56%	1.56%	-1.01%	-3.48%	0.30%	3.48%	2.44%	1.08%	0.97%	-1.63%	-1.94%
41	-0.02%	-2.02%	2.02%	-1.34%	-3.74%	0.09%	3.74%	2.77%	1.25%	1.13%	-1.91%	-2.18%
42	-0.05%	-2.53%	2.53%	-1.71%	-3.99%	-0.20%	3.99%	3.13%	1.44%	1.31%	-2.20%	-2.44%

Figure 14: Age factors for batters

These factors are with respect to the conditional variables  $\tilde{X}$ , thus they are unlikely to compare directly with estimates of unconditional age effects. To provide such a comparison, figure 15 compares the expected OPS curve for a median player using this model alongside Fair’s results calculating these sorts of effects using OPS directly [7].

It has been noted (see [15] for a representative discussion of the issue) that isolating only players with long careers may lead to selection bias, as these players may tend to have different age effects than the general population or may tend to perform disproportionately well at young and/or old ages. These effects would likely lead the model to under-predict age effects. Our results here do not confirm such a result (perhaps due to the granular nature of our analysis). Nonetheless, this concern does seem warranted, and in section 7 we indicate a slight change in the method to account for any such effect.<sup>13</sup>

**Incorporating age adjustments into a nested Dirichlet model** Given these  $R$  values, we can incorporate age effects into the nested Dirichlet model by adjusting all evidence to a normalized age (in this case we have chosen 28). More formally, if  $e_{ija}$  represents the number of times player  $i$  experiences event type  $j$  at age  $a$  in the evidence set, we let  $e'_{ija} = e_{ija} - R_a e_{iIja}$ . We then use the method of maximum likelihood to model the distribution of player ability at age 28 given the data in  $e'$ . To estimate a distribution of player  $i$ ’s conditional ability at age  $a_f$ , with respect to event type  $j$ ,  $\tilde{x}_{ija_f}$ , we start with player  $i$ ’s posterior distribution over  $\tilde{x}_{ij(28)}$  given  $e'$ , denoting

---

whether hits were on ground balls or fly balls. Thus we have approximated these numbers by finding the historical conditional probability that each type of hit was on a fly ball or ground ball, and splitting up the hits in the data accordingly. This method has likely muted the factors for these event types, although we have not performed a thorough analysis of this impact.

<sup>13</sup>It is also left to future analysis to evaluate whether such an additive model of age effects is more reasonable than other alternatives.

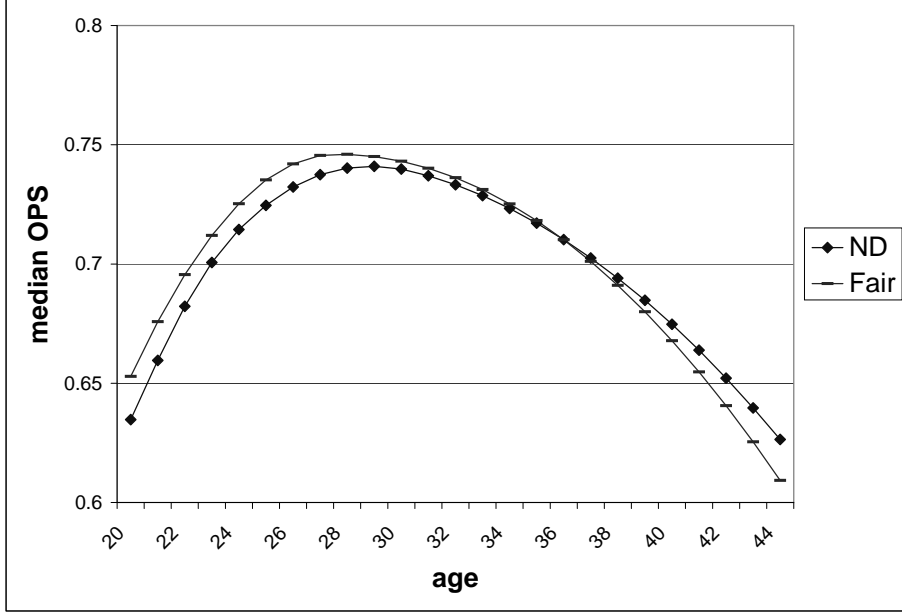


Figure 15: Comparison of expected age effects for a median player

the beta parameters for the distribution by  $\alpha_j^i, \beta_j^i$ . (We will take  $i$  and  $j$  as given in the rest of this discussion and let  $\alpha'$  and  $\beta'$  represent  $\alpha_j^i$  and  $\beta_j^i$  respectively. The mean and variance of this distribution are  $\pi' = \frac{\alpha'}{N'}$  and  $\sigma^{2'} = \frac{\pi'(1-\pi')}{N'+1}$  where  $N' = \alpha' + \beta'$ . To estimate the posterior distribution for the player at age  $a_f$ , we note that  $\pi_f = \pi' + R_f$  and assume  $\sigma_f^2 = \sigma^{2'}$ . We then approximate the posterior at age  $a_f$  by a Beta distribution with mean  $\pi'$  and variance  $\sigma^{2'}$ . To do this we use the relations  $N_f = \frac{\pi_f(1-\pi_f)}{\sigma_f^2} - 1$ ,  $\alpha_f = \pi_f N_f$ , and  $\beta_f = (1 - \pi_f)N_f$ .

## 5.2 A model of player noise and mean-reversion

In modeling random disturbances to player ability, we would like a better estimate of player variance as well as regression to the mean, while still maintaining a nested Dirichlet structure.

In order to do this we consider the following class of models. First, with respect to the conditional ability variables of the form  $\tilde{x}$ , we will represent player  $i$ 's conditional ability at time  $t$  as  $\tilde{x}_{it}$  (leaving the event type to be understood). For any player  $i$ , his conditional ability with respect to a specific event type in year  $t + 1$  ( $\tilde{x}_{i,t+1}$ ) is related to his ability in year  $t$  ( $\tilde{x}_{i,t}$ ) via a function of the following form:

$$\tilde{x}_{i,t+1} = a_1 \tilde{x}_{i,t} + (1 - a_1)\mu + a_2(\epsilon_{i,t+1} - \mu) \quad (5)$$

where  $a_1$  and  $a_2$  are constants,  $\mu$  is the mean of the population prior distribution ( $\frac{\alpha}{N}$ ), and  $\epsilon_{i,t+1}$  is a random variable drawn from a distribution with the same mean and variance as the population prior distribution. We assume  $x_{i,1}$  is drawn from the original population prior.

For simplicity, we would like the model to have the property that the prior for  $x_{i,t}$  is identical

for all  $i$  and  $t$ .<sup>14</sup> The first two terms on the righthand side of equation 5 ensure that  $E[x_{i,t+1}] = E[x_{i,t}]$ . Since  $\sigma_{i,t+1}^2 = a_1^2\sigma_{i,t}^2 + a_2^2\sigma^2$  where  $\sigma^2$  is the initial prior distribution variance, setting  $a_2 = \sqrt{1 - a_1^2}$  ensures that  $\sigma_{i,t+1}^2 = \sigma_{i,t}^2 = \sigma^2$ . Thus, the mean and variance match. To ensure that the distributions match, we ascribe a non-parametric implicit distribution to  $\epsilon_{i,t}$  such that the convolution in 5 converts to a Beta distribution.

We fit this model using the method of maximum likelihood on  $\alpha$ ,  $\beta$ , and  $a_1$  simultaneously, where  $\alpha$  and  $\beta$  are the parameters of the Beta prior distribution. Given prior parameters for a given player,  $i$ , in year  $t$  of  $\alpha_{i,t}^{init}$  and  $\beta_{i,t}^{init}$ , and relevant evidence of  $h_{it} = e_{ijt}$  and  $m_{it} = e_{iIjt} - e_{ijt}$  we get posterior parameters for player  $i$  in year  $t$  of  $\alpha_{i,t}^{fin} = \alpha_{i,t}^{init} + h_{i,t}$  and  $\beta_{i,t}^{fin} = \beta_{i,t}^{init} + m_{i,t}$ .

The posterior mean and variance are then  $\mu_{i,t}^{fin} = \frac{\alpha_{i,t}^{fin}}{\alpha_{i,t}^{fin} + \beta_{i,t}^{fin}}$  and  $\sigma_{i,t}^{2(fin)} = \frac{\mu_{i,t}^{fin}(1 - \mu_{i,t}^{fin})}{\alpha_{i,t}^{fin} + \beta_{i,t}^{fin} + 1}$ . For year  $t + 1$  we then have:

$$\begin{aligned}\mu_{i,t+1}^{init} &= a_1\mu_{i,t}^{fin} + (1 - a_1)\mu \\ \sigma_{i,t+1}^{2(init)} &= a_1^2\sigma_{i,t}^{2(fin)} + a_2^2\sigma^2 \\ N_{i,t+1}^{init} &= \frac{\mu_{i,t+1}^{init}(1 - \mu_{i,t+1}^{init})}{\sigma_{i,t+1}^{2(init)}} - 1 \\ \alpha_{i,t+1}^{init} &= \mu_{i,t+1}^{init}N_{i,t+1}^{init} \\ \beta_{i,t+1}^{init} &= (1 - \mu_{i,t+1}^{init})N_{i,t+1}^{init}\end{aligned}$$

Using these formulae and observing that  $\alpha_{i,1}^{init} = \alpha$  and  $\beta_{i,1}^{init} = \beta$ , given a multi-year set of evidence for a player, we can calculate the likelihood of the observed evidence for that player given any initial  $\alpha, \beta, a_1$  triple (see [11] for details on the formula for likelihood of an evidence set given an  $\alpha, \beta$  prior). As all players are independent conditioned on these parameters, the likelihoods for a set of players is simply the product of their individual likelihoods. We then solve for the triple that maximizes the likelihood of the entire evidence set. We can compare over nestings as well, but for our purposes, we will take the nesting in figure 7 as given.

**Using this model with age-adjusted data** Finally, to use age-adjusted data with this model we can follow the same methodology as above with the modified model:

$$\tilde{x}_{i,t+1} - R_{g_{t+1}} = a_1(\tilde{x}_{i,t} - R_{g_t}) + (1 - a_1)\mu + a_2(\epsilon_{i,t+1} - \mu)$$

where  $g_t$  and  $g_{t+1}$  are the ages in the two years and  $R$  is the vector of age effects to convert performance at each age to the benchmark age (28). Using data for batters from 1988-1998, we get the values in the second column of figure 16 as our  $a_1$  estimates.<sup>15</sup> Though a thorough estimate of the precision of these figures is justified, preliminary analysis, by way of estimating these values using different subsets of this data gives us the interval of confidence represented in the third and fourth columns of figure 16. For most of the more frequent event types these values vary by less than 3% and on all event types they vary by less than 10%.

<sup>14</sup>While it is quite likely that this prior actually changes over time, we err in favor of simplicity for the present analysis.

<sup>15</sup>As mentioned earlier, parts of the data do not fully specify hits were on ground balls or fly balls. Thus we use this period as the consecutive set of years that does.

type	a_1 - decay rate	a_1 - range		remaining differential			
				5yr residual	10yr residual	50yr residual	100yr residual
Int	99.7%	80.1%	99.7%	98.3%	96.6%	83.9%	70.4%
HBP	96.2%	94.3%	97.6%	82.3%	67.7%	14.2%	2.0%
BB	95.2%	94.5%	95.3%	78.1%	61.0%	8.4%	0.7%
K	95.2%	94.5%	95.4%	78.1%	61.0%	8.5%	0.7%
HR,FB	95.2%	94.7%	95.2%	78.0%	60.9%	8.4%	0.7%
3B,FB	99.5%	99.3%	100.0%	97.4%	94.9%	77.1%	59.4%
2B,FB	85.0%	83.2%	86.7%	44.4%	19.7%	0.0%	0.0%
1B,FB	99.5%	99.3%	100.0%	97.6%	95.3%	78.6%	61.8%
out,FB	90.9%	88.6%	90.9%	62.0%	38.5%	0.8%	0.0%
HR,GB	98.6%	97.8%	100.0%	93.4%	87.2%	50.4%	25.4%
3B,GB	100.0%	98.6%	100.0%	100.0%	100.0%	100.0%	100.0%
2B,GB	98.5%	88.7%	98.5%	92.5%	85.6%	46.0%	21.2%
1B,GB	93.2%	84.7%	93.2%	70.2%	49.2%	2.9%	0.1%
out,GB	92.7%	83.5%	92.7%	68.5%	46.9%	2.3%	0.1%
FB	92.4%	91.4%	92.4%	67.3%	45.3%	1.9%	0.0%
GB	93.6%	93.6%	93.9%	71.7%	51.4%	3.6%	0.1%

Figure 16: Comparison of noise parameters

**Do abilities really mean revert?** Although most of the mean-reversion rates in figure 16 are small, they are still significant, which leads us to question: Do player abilities really mean-revert? It is hard to imagine that some day Barry Bonds will be no more likely to be the home run champion than Juan Pierre, even though, as this model indicates (by looking at the residuals in the right side of figure 16) that day may be fifty years in the future. That said, how do we test whether this is likely the case?

First, let's consider some alternative hypotheses for why the nested Dirichlet model underpredicts regression to the mean. Most simply, perhaps our estimation procedure under-estimates all of the nested Dirichlet parameters. Increasing these parameters would increase the expected regression to the mean. However, this would also increase the model implied confidence in estimates of player ability, which we already know (see section 4) the nested Dirichlet model over-estimates. So this must not be the problem.

It certainly seems there is some year-to-year variance in player ability, so what if we assume it is not mean-reverting. Doing this would cause  $\sigma_{it}^2$  to increase with  $t$ , and the year to year prior to become more diffuse.<sup>16</sup> It would also not account for the observed mean-reversion in player ability. So it seems player ability must mean-revert to something. If not the population mean, then what?

Lets consider a model where each player has some long-run ability  $\tilde{x}_i^0$ , and the model for his ability in any year is

$$\tilde{x}_{i,t+1} = a_1 \tilde{x}_{i,t} + (1 - a_1) \tilde{x}_i^0 + a_2 (\epsilon_{i,t+1} - \mu) \quad (6)$$

This model would capture some additional mean-reversion without forcing every player to converge to the population mean. The major downside of this model is that it is much more expensive to fit and update than the model in 5. There is also no simple way to estimate every player with a

<sup>16</sup>As a caveat, we should note that it is quite possible that player abilities do become more diffuse over time, but without a foundation of evidence for such a phenomenon, we stick with the simpler assumption here that they do not.

nested Dirichlet (or conditional Beta) posterior distribution. Another potential issue is that when we consider this distribution, it seems reasonable to assume that these long-run abilities change over time as well. If we decide to factor this into the model we are back to the point where either everyone mean-reverts to the population mean or a player’s distribution becomes more diffuse over time.

### 5.3 Forecasting with the extended model

In figure 17, we evaluate player error forecasts using data from 2003-06 and the extended nested Dirichlet model incorporating both age and ability effects, replicating the weighted player RMSE calculations from figure 11. Observe that the average accuracy has improved slightly (by .02%). More interesting though is the fact that the observed error is now in line with the model expected error.

		min E	Ext ND E	RMSE
type	actual%	RMSE	RMSE	Ext ND
Int	0.01%	0.06%	0.07%	0.07%
HBP	0.97%	0.64%	0.75%	0.75%
BB	8.13%	1.71%	2.36%	2.27%
K	16.89%	2.42%	3.53%	3.78%
HR,FB	2.75%	1.02%	1.28%	1.28%
3B,FB	0.48%	0.46%	0.50%	0.52%
2B,FB	4.41%	1.27%	1.40%	1.37%
1B,FB	8.73%	1.79%	2.13%	1.94%
out,FB	25.29%	2.76%	3.47%	3.63%
HR,GB	0.00%	0.01%	0.01%	0.00%
3B,GB	0.04%	0.12%	0.12%	0.15%
2B,GB	0.65%	0.52%	0.55%	0.52%
1B,GB	7.32%	1.65%	1.88%	1.99%
out,GB	24.32%	2.77%	3.85%	3.83%
FB	41.67%	3.13%	4.27%	4.03%
GB	32.33%	2.99%	4.56%	4.46%
average	10.87%	1.457%	1.922%	1.912%

Figure 17: RMSE of player predictions in the extended model

In figure 18, we examine predicted versus actual OPS by age group of both the original nested Dirichlet model and the extended model incorporating both the age effect and player noise models of this section. As opposed to the original model, which is off by at least two standard deviations on all groups, the extended model has significant error for the 24 and under group only. One theory behind this is that players with significant experience at an early age are conditionally more likely to be better players.<sup>17</sup>

In figure 19 we examine predicted minus actual OPS by expected ability group. The extended model has no error greater than .01 points or approximately one standard deviation. Examining the predicted OPS differential between the high and low ability groups directly, whereas the observed difference was 1.8 standard deviations away from that predicted by the original model, the extended

<sup>17</sup>This leads to another interesting idea to extend the model, conditionally updating a player’s prior distribution given the year of his debut.

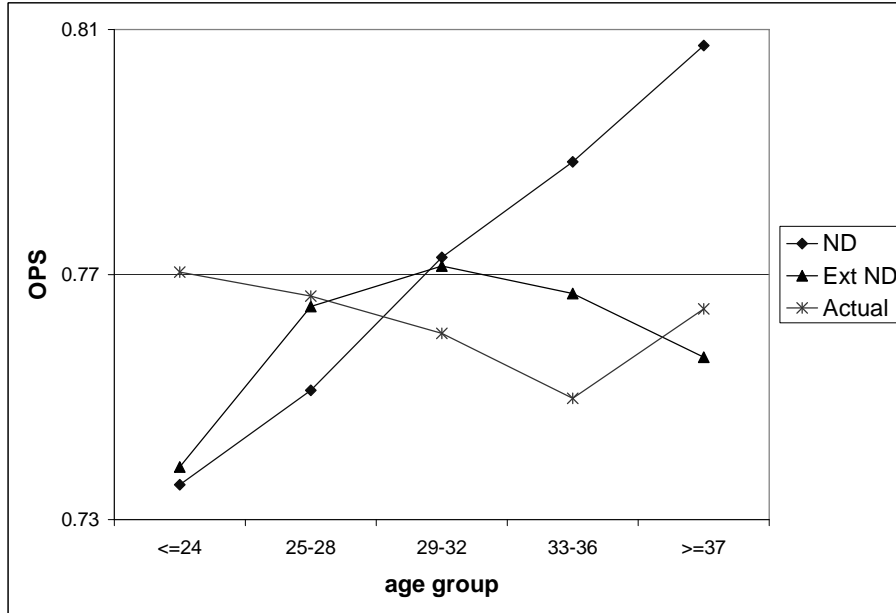


Figure 18: Predicted OPS by age

model implies an expected difference of .1807. The observed difference for this model was .1889, .8 standard deviations away.<sup>18</sup> Thus, the extended model is much more consistent with the results.

## 6 Comparison with proprietary forecasting systems

To properly benchmark the predictive power of these models, we would like to compare their predictions against those of other forecasting systems. To do so, we will need to perform additional modeling. Recall that in building these nested Dirichlet models, we separated batters by handedness and removed all bunt attempts and intentional walks from the data. In order to compare with other forecasting systems, we must aggregate our predictions for switch hitters and incorporate these other possible outcomes. We will factor these in by looking at historical data and fitting a series of Beta prior distributions as appropriate. These distributions will represent a higher level nested Dirichlet model, but for this model we are given the nesting tree in figure 20. With respect to this tree, we first predict the frequency of left and right handed appearances for all switch hitters (we assume only one-side for non-switch hitters), then we predict the frequency of intentional walks for all batters, then the frequency of bunts, followed by the frequency of sacrifice bunts and non-sacrifice bunts and the related outcome.

Given this aggregate model, we can now forecast actual player statistics and compare these with other forecasting systems. We will benchmark versus the popular Marcel models (see [16]), one such representative system for which predictions are freely available. Marcel is an admittedly simple system, but previous analysis (see [13]) has shown that its accuracy is comparable with that of other touted prediction systems. Because Marcel does not include every player that actually played in 2007, we will limit the comparison to those that Marcel does include.

<sup>18</sup>The values differ because the models place different players in the five groups.

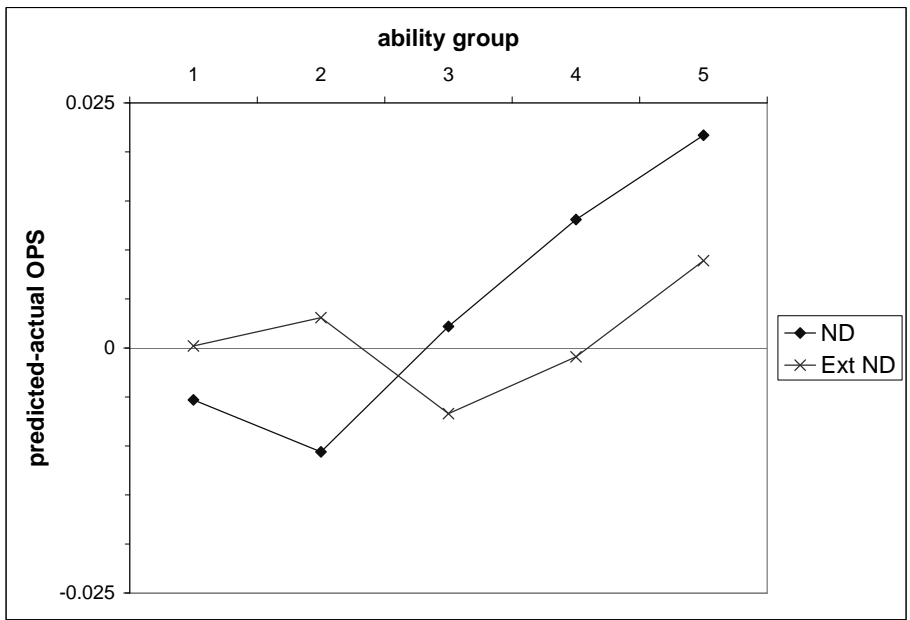


Figure 19: Predicted-Actual OPS by estimated ability level

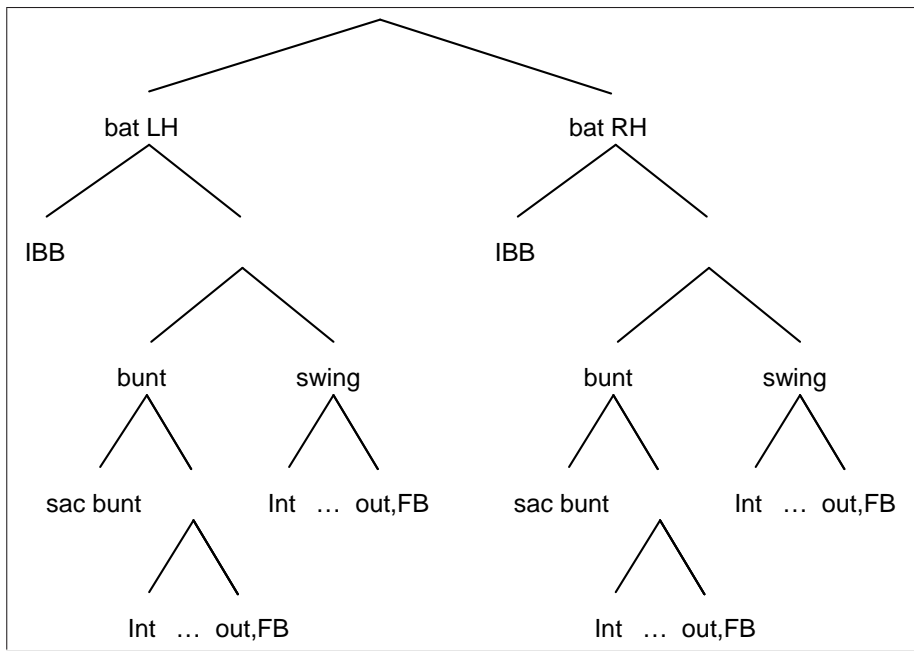


Figure 20: Aggregate probability tree for a batter

Further, since Marcel, like most publicly available forecasting systems, and unlike the modeling we have performed predicts plate appearances and aggregate totals of events, we will limit ourselves to rate metrics, explicitly OPS.<sup>19</sup> This will further help us level the playing field, since we do not have predictions from Marcel for each of the 14 event types that we have used. In figure 21 we evaluate the weighted root mean squared error (w RMSE OPS) and the mean absolute error (w AE OPS) of the predicted player OPS for the basic nested Dirichlet model and the three extensions as well as for Marcel.

	ND	ND w age	ND w ab	Ext ND	Marcel
AE OPS	0.0722	0.0688	0.0700	0.0675	0.0713
RMSE OPS	0.0946	0.0912	0.0930	0.0906	0.0950

Figure 21: Performance of our models and Marcel

Silver [13] compared the performance of Marcel and several other systems in predicting 2007 batter OPS. In his analysis, he evaluated the unweighted root mean squared error and absolute error, limiting the comparison to players with more than 250 plate appearances. Figure 22 shows the *estimated* relative performance of the extended nested Dirichlet model (Ext ND) versus the systems observed by Silver.<sup>20</sup> The nested Dirichlet model performed comparably to the best alternative systems with respect to these metrics.

	AE	RMSE
Ext ND	0.060	0.0785
CHONE	0.061	0.077
ZIPS	0.061	0.077
PECOTA	0.061	0.078
Marcel	0.063	0.081
THT	0.063	0.081
RotoTimes	0.067	0.086
RotoWire	0.067	0.087
ESPN	0.068	0.087

Figure 22: Relative performance versus several popular forecasting systems

## 7 Conclusions

The analysis here is intended to serve as more of an introduction and motivation for a new model of baseball player ability as opposed to a rigorous testing of it. Nonetheless, our results indicate that the nested Dirichlet model with extensions is a promising and flexible model for baseball player ability. Not only is it competitive with the most accurate publicly available models, it also has a

<sup>19</sup>The formula for OPS is presented in section 4.

<sup>20</sup>We do not know the exact set of players Silver used. Thus, our values are estimated by comparing the performance of the extended nested Dirichlet model versus Marcel on an approximately equivalent data set.

reasonable estimate of posterior variance by player and is flexible enough to allow for extensions to incorporate additional useful information (as we have done here with age effects and year-to-year variance). Additionally, the model allows for granular predictions and is easily updatable as it is still conjugate prior to multinomial data. As a result, the model is eminently valuable as a player forecasting tool or as an input for predicting particular plate appearances (via the well-known log5 method<sup>21</sup> or another method) as well as game and season results.

Though our final analyses have focused on player error and OPS, specifically because this allowed us to benchmark against other systems, we should not overlook the potential value in the granularity of the model. For example, the model has been designed to accurately estimate a likelihood for any player for any event type. Improved accuracy in these estimates could have significant impacts in modeling the efficacy of different players in different game situations. For example, a home run hitter and a contact hitter with the same OPS would have very different potentials to change the game when batting late in a tie game with nobody on base. Further, an accurate understanding of the uncertainty related to our understanding of player abilities can have a significant impact when modeling games and seasons.<sup>22</sup> This ability to model individual event types well is perhaps the most significant characteristic of this model, and a key advantage of the model over a simpler distribution such as the Dirichlet, particularly when the goal is to apply such a model towards evaluating play and game results.

## 7.1 Possible improvements

Nonetheless, there are still numerous potential extensions that may improve the model. These include:

- **Test for bias in maximum likelihood estimators** The maximum likelihood estimators we have used are sensitive to selection bias in the sense that if better players get more plate appearances, parameter estimates will be skewed by those players. Thus, we should examine carefully the potential magnitude of this bias present in the current estimates, and how that affects potential applications of these results.
- **Find a better nesting** We have seen that the nesting in figure 7 still leaves something to be desired in terms of how it models the covariance of the variables. Thus, we would like to try more nestings to see if we can model the distribution better. Given the large number of possible nestings, this means that we may need to alter our definition of event types or derive a new algorithm to search through nestings. This could include defining partially observed event types.
- **Mixture models** It is quite likely that different pitchers and hitters are drawn from different distributions. We have already anticipated this in the analysis above by examining only the performance of non-pitchers batting, but more (and more sophisticated) segmentation would probably be useful. For instance, in section 5 we observed that players who debut early may be drawn from a different distribution. There may also be segmentation that is not so easy to see. To incorporate this we might build a mixture model similar the method of [5] to

---

<sup>21</sup>This method, credited to Dallas Adams, was presented by Bill James in the out-of-print 1983 Baseball Abstract. See [9] for a representative discussion of the method.

<sup>22</sup>More on this in the sequel to this research.

determine not only how likely it is that each player is drawn from a certain distribution, but what the parameters of that distribution are.

- **Recalculate age effects using representative data** As we mentioned in section 5, we should examine our model of age effects more closely to see if there is selection bias. If so, we can counter this by randomly selecting players (in proportion to their experience) and estimating age effects over this random sample using a maximum likelihood based technique.
- **Get more data (e.g. Minor Leagues)** Our priors could certainly be improved, especially for inexperienced players, if we factored in more information about the new players in the league, particularly minor league data, although we would need to translate this data to account for the differences in the leagues (as we did to account for age effects in section 5).
- **Additional alteration of inputs** Excepting for age effects, we implicitly assumed that every plate appearance was under “normal” conditions, or at least that the conditions averaged out in some way. However, given that different players may play a disproportionate number of their games in certain locations or against righthanded pitchers for example, it may be quite beneficial to “normalize” these inputs, much as we have done in age-adjusting data.
- **From fixed effects to a Bayesian model** Players probably don’t all have the same year-to-year variance in ability or the same age effects. Perhaps a Bayesian analysis to estimate each player’s individual degree effects along the lines of [1] would improve the model.
- **Year-to-year population-wide effects** Finally, in section 5 we observed that there seem to be year-to-year changes in the overall likelihood of different event types. Thus, we might consider incorporating a jump parameter into the model to account for this.

## Acknowledgements

I would like to thank Sam Chiu, Tom Cover, and Hervé Kieffel for helpful discussions concerning this research.

## References

- [1] Albert, J. 2002. Smoothing Career Trajectories of Baseball Hitters, August 22.
- [2] Albert, J. 2006. Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time. *Journal of Quantitative Analysis in Sports*, 2, 1, 2.
- [3] Albert, J. 2006. A Breakdown of a Batter's Plate Appearance - Four Hitting Rates. *By the Numbers*, February, 23-30.
- [4] Baumer, B. S. 2008. Why On-Base Percentage is a Better Indicator of Future Performance than Batting Average: An Algebraic Proof, *Journal of Quantitative Analysis in Sports*, 4, 2, 3.
- [5] Bouguila, N. 2008. Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4), 462-474.
- [6] Connor, R.J. and Mosiman, J.E. 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194-206.
- [7] Fair, R.C. 2008. Estimated Age Effects in Baseball. *Journal of Quantitative Analysis in Sports*, 4, 1, 1.
- [8] James, B. 1982, *The Bill James Baseball Abstract 1982*. Ballantine.
- [9] Levitt, D. The Batter/Pitcher Matchup. [http://www.baseballthinkfactory.org/btf/scholars/levitt/articles/batter\\_pitcher\\_matchup.htm](http://www.baseballthinkfactory.org/btf/scholars/levitt/articles/batter_pitcher_matchup.htm)
- [10] McCracken, V. 2001. Pitching and Defense: How Much Control Do Hurlers Have? *Baseball Prospectus*, <http://baseballprospectus.com/article.php?articleid=878>.
- [11] Null, B. 2008, *The Nested Dirichlet Distribution: Properties and Applications*. Working Paper.
- [12] Schall, E. and Smith, G. Do Baseball Players Regress Toward the Mean? <http://www.economics.pomona.edu/GarySmith/BBregress/baseball.html>.
- [13] Silver, N. 2007 Hitter Projection Roundup. <http://www.baseballprospectus.com/unfiltered/?p=564>.
- [14] Tango, T. 2002, Tango on Baseball Archives - Aging Patterns. <http://www.tangotiger.net/archives/artAging.shtml#1013>
- [15] Tango, T. 2007, Peak Offensive Age. [http://www.insidethebook.com/ee/index.php/site/comments/peak\\_offensive\\_age/](http://www.insidethebook.com/ee/index.php/site/comments/peak_offensive_age/)
- [16] Tango, T. Marcel 2008. <http://www.tangotiger.net/marcel/>.
- [17] Tango, T., Lichtman, M., and Dolphin, A. 2006. *The Book - Playing the Percentages in Baseball*, TMA Press.
- [18] Wong, T.T. 1998. Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97, 165-181.