Bayesian Analysis (2009) **4**, Number 4, pp. 631–652

Hierarchical Bayesian Modeling of Hitting Performance in Baseball

Shane T. Jensen*, Blakeley B. McShane† and Abraham J. Wyner‡

Abstract. We have developed a sophisticated statistical model for predicting the hitting performance of Major League baseball players. The Bayesian paradigm provides a principled method for balancing past performance with crucial covariates, such as player age and position. We share information across time and across players by using mixture distributions to control shrinkage for improved accuracy. We compare the performance of our model to current sabermetric methods on a held-out season (2006), and discuss both successes and limitations.

Keywords: baseball, hidden Markov model, hierarchical Bayes

1 Introduction and Motivation

There is substantial public and private interest in the projection of future hitting performance in baseball. Major league baseball teams award large monetary contracts to top free agent hitters under the assumption that they can reasonably expect that past success will continue into the future. Of course, there is an expectation that future performance will vary, but for the most part it appears that teams are often quite foolishly seduced by a fine performance over a single season. There are many questions: How should past consistency be balanced with advancing age when projecting future hitting performance? In young players, how many seasons of above-average performance need to be observed before we consider a player to be a truly exceptional hitter? What is the effect of a single sub-par year in an otherwise consistent career? We will attempt to answer these questions through the use of fully parametric statistical models for hitting performance.

Modeling and prediction of hitting performance is an area of very active research within the quantitatively-oriented baseball community. Popular current methods include PECOTA (Silver 2003) and MARCEL (Tango 2004). PECOTA is considered a "gold-standard" tool in the sabermetrics community and its predictions are billed by *Baseball Prospectus* as being "deadly accurate". It is a sophisticated commercial product managed by a team of statisticians which incorporates proprietary data, minor league histories, and detailed injury reports. Since PECOTA is proprietary, we cannot

*Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:stjensen@wharton.upenn.edu

†Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:mcshaneb@wharton.upenn.edu

‡Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:ajw@wharton.upenn.edu

 DOI:10.1214/09-BA424

say exactly what methods they use though we know the general method is based on matching a player's past career performance to the careers of a set of comparable major league ballplayers. For each player, their set of comparable players is found by a nearest neighbor analysis of past players (both minor and major league) with similar performance at the same age. Once a comparison set is found, the future performance prediction for the player is based on the historical performance of those past comparable players. Factors such as park effects, league effects and physical attributes of the player are also taken into account. PECOTA also makes use of substantial manual curation both to the matching process and to introduce new information as it becomes available. We have observed that the pre-season PECOTA predictions are adjusted on a daily basis as news (*e.g.*, injury information, pre-season performance, etc.) is released.

In contrast, our focus is on a model-based approach to prediction of hitting performance which is fully-automated and based on publicly available data. Thus, a more appropriate benchmark for our analysis is MARCEL, a publicly available prediction engine based on the same freely available dataset (Lahman 2006) as our model. MARCEL is a simple two-stage system for prediction. First, MARCEL takes a weighted average of the performance of the player over the previous three years, giving more weight to the most recent seasons. Then, it shrinks this weighted average to the overall league mean based on the number of plate appearances. Thus, the more data for a given player, the less shrinkage. Over several seasons, MARCEL has performed well against more elaborate competitors (Tango 2004), but should be outperformed by our principled approach. Although it is less of a fair benchmark, we will also compare with PECOTA in order to assess how well our model does against the best available proprietary commercial product.

In Section 2, we present a Bayesian hierarchical model for the evolution of hitting performance throughout the careers of individual players. Bayesian or Empirical Bayes approaches have recently been used to model individual hitting events based on various within-game covariates (Quintana et al. 2008) and for prediction of within-season performance (Brown 2008). We are addressing a different question: how can we predict the course of a particular hitter's career based on the seasons of information we have observed thus far? Our model includes several covariates that are crucial for the accurate prediction of hitting for a particular player in a given year. A player's age and home ballpark certainly has an influence on their hitting; we will include this information among the covariates in our model. We will also include player position in our model, since we believe that position is an important proxy for hitting performance (*e.g.*, second basemen have a generally lower propensity for home runs than first basemen). Finally, our model will factor past performance of each player into future predictions. In Section 3, we test our predictions against a hold out data set, and compare our performance with several competing methods. A major advantage of our model-based approach is the ability to move beyond the point predictions offered by other engines to the incorporation of variability via calibrated predictive intervals. We examine our results not only in terms of accuracy of our point predictions, but also the quality the prediction intervals produced by our model. We also investigate several other interesting aspects of our model in Section 3 and then conclude with a brief discussion in Section 4.

2 Model and Implementation

Our data comes from the publicly-available Lahman Baseball Database (Lahman 2006), which contains hitting totals for each major league baseball player from 1871 to the present day, though we will fit our model using only seasons from 1990 to 2005. In total, we have 10280 player-years of of information from major league baseball between 1990 and 2005 that will be used for model estimation. Within each season j, we will use the following data for each player i:

1. Home Run Total : Y_{ij}
2. At Bat Total : M_{ij}
3. Age : A_{ij}
4. Home Ballpark : B_{ij}
5. Position : R_{ij}

As an example, Barry Bonds in 2001 had $Y_{ij} = 73$ home runs out of $M_{ij} = 476$ at bats. We excluded pitchers from our model, leaving us with nine positions: first basemen (1B), second basemen (2B), third basemen (3B), shortstop (SS), left fielder (LF), center fielder (CF), right fielder (RF), catcher (C), and the designated hitter (DH). There were 46 different home ballparks used in major league baseball between 1990 and 2005. Player ages ranged between 20 and 49, though the vast majority of player ages were between 23 and 44.

2.1 Hierarchical Model for Hitting

Our outcome of interest for a given player i in a given year (season) j is their home run total Y_{ij}, which we model as a Binomial variable:

$$Y_{ij} \sim \text{Binomial}(M_{ij}, \theta_{ij}) \tag{1}$$

where θ_{ij} is a player- and year-specific home run rate, and M_{ij} are the number of opportunities (at bats) for player i in year j. Note that by using at-bats as our number of opportunities, we are excluding outcomes such as walks, sacrifice flies and hit-by-pitches. We will assume that the number of opportunities M_{ij} are fixed and known so we focus our efforts on modeling each home run rate θ_{ij}. The i.i.d. assumption underlying the binomial model has already been justified for hitting totals within a single season (Brown 2008), and so seems reasonable for hitting totals across an entire season.

We next model each unobserved player-year rate θ_{ij} as a function of home ballpark $b = B_{ij}$, position $k = R_{ij}$ and age A_{ij} of player i in year j.

$$\log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \alpha_k + \beta_b + f_k(A_{ij}) \tag{2}$$

The parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_9)$ are the position-specific intercepts for each of the nine player positions. The function $f_k(A_{ij})$ is a smooth trajectory of A_{ij}, that is different for each position k. We allow a flexible model for $f_k(\cdot)$ by using a cubic B-spline (de Boor 1978) with different spline coefficients $\boldsymbol{\gamma}$ estimated for each position. The age trajectory component of this model involves the estimation of 36 parameters: four B-spline coefficients per position $\times$ nine different positions.

We call the parameter vector $\boldsymbol{\beta}$ the "team effects" since these parameters are shared by all players with the same team and home ballpark. However, these coefficients $\boldsymbol{\beta}$ can not be interpreted as a true "ballpark effect" since they are confounded with the effect of the team playing in that ballpark. If a particular team contains many home run hitters, then that can influence the effect of their home ballpark. Separating the effect of team versus the effect of ballpark would require examining hitting data at the game level instead of the seasonal level we are using for our current model.

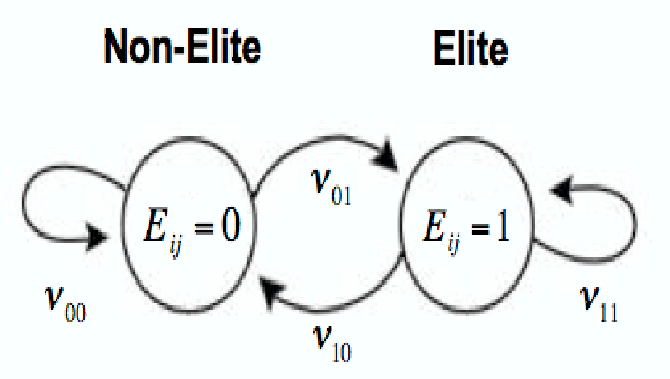
There are two additional aspects of hitting performance that are not captured by the model outlined in (1)-(2). Firstly, conditional on the covariates age, position, and ballpark, our model treats the home run rate θ_{ij} as independent and identically-distributed across players i and years j. However, we suspect that not all hitters are created equal: we posit that there exists a sub-group of elite home run hitters within each position that share a higher mean home run rate. We can represent this belief by placing a mixture model on the intercept term α_k dictated by a latent variable E_{ij} in each player-year. In other words,

$$\alpha_k = \begin{cases} \alpha_{k0} & \text{if} \quad \mathrm{E}_{ij} = 0 \\ \alpha_{k1} & \text{if} \quad \mathrm{E}_{ij} = 1 \end{cases}$$

where we force $\alpha_{k0} < \alpha_{k1}$ for each position k. We call the latent variable E_{ij} the elite status for player i in year j. Players with elite status are modeled as having the same shape to their age trajectory, but with an extra additive term (on the log-odds scale) that increases their home run rate. However, we have a different elite indicator E_{ij} for each player-year, which means that a particular player i can move in and out of elite status during the course of his career. Thus, the elite sub-group is maintained in the player population throughout time even though this sub-group will not contain the exact same players from year to year.

The second aspect of hitting performance that needs to be addressed is that the past performance of a particular player should contain information about his future performance. One option would be to use player-specific intercepts in the model to allow each player to have a different trajectory. However, this model choice would involve a large number of parameters, even if these player-specific intercepts were assumed to share a common prior distribution. In addition, many of these intercepts would be subject to over-fitting due to small number of observed years of data for many players. We instead favor an approach that involves fewer parameters (to prevent over-fitting) while still allowing different histories for individual players. We accomplish this goal by building the past performance of each player into our model through a hidden Markov model on the elite status indicators E_{ij} for each player i. Specifically, our probability model of the elite status indicator for player i in year $j+1$ is allowed to depend on the

Figure 1: Hidden Markov Model for Elite Status



elite status indicator for player i in year j:

$$p(\mathrm{E}_{i,j+1} = b | \mathrm{E}_{ij} = a, R_{ij} = k) = \nu_{abk} \qquad a, b \in \{0, 1\} \tag{3}$$

where E_{ij} is the elite status indicator and R_{ij} is the position of player i in year j. This relationship is also graphically represented in Figure 1. The Markovian assumption induces a dependence structure on the home run rates $\theta_{i,j}$ over time for each player i. Players that show elite performance up until year j are more likely to be predicted as elite at year $j + 1$. The transition parameters $\boldsymbol{\nu}_k = (\nu_{00k}, \nu_{01k}, \nu_{10k}, \nu_{11k})$ for each position $k = 1, \ldots, 9$ are shared across players at their position, but can differ between positions, which allows for a different proportion of elite players in each position. We initialize each player's Markov chain by setting $\mathrm{E}_{i0} = 0$ for all i, meaning that each player starts their career in non-elite status. This initialization has the desired consequence that young players must show consistently elite performance in multiple years in order to have a high probability of moving to the elite group.

In order to take a fully Bayesian approach to this problem, we must specify prior distributions for all of our unknown parameters. The forty-eight different ballpark coefficients $\boldsymbol{\beta}$ in our model all share a common Normal distribution,

$$\beta_l \sim \mathrm{Normal}(0, \tau^2) \qquad \forall \quad l = 1, \ldots, 48 \tag{4}$$

The spline coefficients $\boldsymbol{\gamma}$ needed for the modeling of our age trajectories also share a common Normal distribution,

$$\gamma_{kl} \sim \mathrm{Normal}(0, \tau^2) \qquad \forall \quad k = 1, \ldots, 9, \, l = 1, \ldots, L \tag{5}$$

where L is the number of spline coefficients needed in the modeling of age trajectories for $f(A_{ij}, R_{ij})$ for each position. In our latent mixture model, we also have two intercept coefficients for each position, $\boldsymbol{\alpha}_k = (\alpha_{k0}, \alpha_{k1})$, which share a truncated Normal distribution,

$$\boldsymbol{\alpha}_k \sim \mathrm{MVNormal}(\mathbf{0}, \tau^2 \boldsymbol{I}_2) \cdot \mathrm{Ind}(\alpha_{k0} < \alpha_{k1}) \qquad \forall \quad k = 1, \ldots, 9 \tag{6}$$

where $\mathbf{0}$ is the 2×1 vector of zeros and $\boldsymbol{I}_2$ is the 2×2 identity matrix. This bivariate distribution is truncated by the indicator function $\text{Ind}(\cdot)$ to ensure that $\alpha_{k0} < \alpha_{k1}$ for each position k. We make each of the prior distributions (4)-(6) non-informative by setting the variance hyperparameter τ^2 to a very large value (10000 in this study). Finally, for the position-specific transition parameters of our elite status $\boldsymbol{\nu}$, we use flat Dirichlet prior distributions,

$$\begin{aligned} (\nu_{00k}, \nu_{01k}) &\sim \text{Dirichlet}(\omega, \omega) \quad \forall \quad k = 1, \ldots, 9, \\ (\nu_{10k}, \nu_{11k}) &\sim \text{Dirichlet}(\omega, \omega) \quad \forall \quad k = 1, \ldots, 9. \end{aligned} \tag{7}$$

These prior distributions are made non-informative by setting ω to a small value ($\omega = 1$ in this study). We also examined other values for ω and found that using different values had no influence on our posterior inference, which is to be expected considering the dominance of the data in equation (9). Combining these prior distributions together with equations (1)-(3) give us the full posterior distribution of our unknown parameters,

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E} | \boldsymbol{X}) \quad \propto \quad & \prod_{i,j} p(Y_{ij} | M_{ij}, \theta_{ij}) \cdot p(\theta_{ij} | R_{ij}, A_{ij}, B_{ij}, E_{ij}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ & \cdot p(E_{ij} | E_{i,j-1}, \boldsymbol{\nu}) \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}). \end{aligned} \tag{8}$$

where we use $\boldsymbol{X}$ to denote our entire set of observed data $\boldsymbol{Y}$ and covariates $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{R})$.

2.2 MCMC Implementation

We estimate our posterior distribution (8) by a Gibbs sampling strategy (Geman and Geman 1984). We iteratively sample from the following conditional distributions of each set of parameters given the current values of the other parameters:

1. $p(\boldsymbol{\alpha} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E}, \boldsymbol{X}) = p(\boldsymbol{\alpha} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{E}, \boldsymbol{X})$
2. $p(\boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E}, \boldsymbol{X}) = p(\boldsymbol{\beta} | \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{E}, \boldsymbol{X})$
3. $p(\boldsymbol{\gamma} | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\nu}, \boldsymbol{E}, \boldsymbol{X}) = p(\boldsymbol{\gamma} | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{E}, \boldsymbol{X})$
4. $p(\boldsymbol{\nu} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{E}, \boldsymbol{X}) = p(\boldsymbol{\nu} | \boldsymbol{E})$
5. $p(\boldsymbol{E} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E}, \boldsymbol{X})$

where again $\boldsymbol{X}$ denotes our entire set of observed data $\boldsymbol{Y}$ and covariates $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{M}, \boldsymbol{R})$. Combined together, steps 1-3 of the Gibbs sampler represent the usual estimation of regression coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in a Bayesian logistic regression model. The conditional posterior distributions for these coefficients are complicated and we employ the common strategy of using the Metropolis-Hastings algorithm to sample each coefficient (see, e.g. Gelman et al. (2003)). The proposal distribution for a particular coefficient is a Normal distribution centered at the maximum likelihood estimate of that coefficient. The

variance of this Normal proposal distribution is a tuning parameter that was adaptively adjusted to provide a reasonable rejection/acceptance ratio (Gelman et al. 1996). Step 4 of the Gibbs sampler involves standard distributions for our transition parameters $\boldsymbol{\nu}_k = (\nu_{00k}, \nu_{01k}, \nu_{10k}, \nu_{11k})$ for each position $k = 1, \ldots, 9$. The conditional posterior distributions for our transition parameters implied by (8) are

$$\begin{aligned} (\nu_{00k}, \nu_{01k})|\boldsymbol{E} &\sim \text{Dirichlet}\left(\mathrm{N}_{00k} + \omega, \mathrm{N}_{01k} + \omega\right) \\ (\nu_{11k}, \nu_{10k})|\boldsymbol{E} &\sim \text{Dirichlet}\left(\mathrm{N}_{11k} + \omega, \mathrm{N}_{10k} + \omega\right) \end{aligned} \tag{9}$$

where $\mathrm{N}_{abk} = \sum_i \sum_{t=1}^{n_i} \mathrm{I}(\mathrm{E}_{i,t} = a, \mathrm{E}_{i,t+1} = b)$ over all players i in position k and where n_i represents the number of years of observed data for player i's career. Finally, step 5 of our Gibbs sampler involves sampling the elite status E_{ij} for each year j of player i, which can be done using the "Forward-summing Backward-sampling" algorithm for hidden Markov models (Chib 1996). For a particular player i, this algorithm "forward-sums" by recursively calculating

$$\begin{aligned} p(\mathrm{E}_{it}|\boldsymbol{X}_{i,t}, \boldsymbol{\Theta}) &\propto p(X_{i,t}|\mathrm{E}_{it}, \boldsymbol{\Theta}) \cdot p(\mathrm{E}_{it}|\boldsymbol{X}_{i,t-1}, \boldsymbol{\Theta}) \\ &\propto p(X_{i,t}|\mathrm{E}_{it}, \boldsymbol{\Theta}) \sum_{e=0}^{1} p(\mathrm{E}_{it}|\mathrm{E}_{i,t-1} = e, \boldsymbol{\Theta})\, p(\mathrm{E}_{i,t-1} = e|\boldsymbol{X}_{i,t-1}, \boldsymbol{\Theta}) \end{aligned} \tag{10}$$

for all $t = 1, \ldots, n_i$ where $\boldsymbol{X}_{i,k}$ represents the observed data for player i up until year k, $X_{i,k}$ represents only the observed data for player i in year k, and $\boldsymbol{\Theta}$ represents all other parameters. The algorithm then "backward-samples" by sampling the terminal elite state E_{i,n_i} from the distribution $p(\mathrm{E}_{i,n_i}|\boldsymbol{X}_{i,n_i}, \Theta)$ and then sampling $\mathrm{E}_{i,t-1}|\mathrm{E}_{i,t}$ for $t = n_i$ back to $t = 1$. Repeating this algorithm for each player i gives us a complete sample of our elite statuses $\boldsymbol{E}$. We ran multiple chains from different starting values to evaluate convergence of our Gibbs sampler. Our results are based on several chains where the first 1000 iterations were discarded as burn-in. Our chains were also thinned, taking only every eighth iteration, in order to eliminate autocorrelation.

2.3 Model Extension: Player-Specific Transition Parameters

In Section 2.1, we introduced a hidden Markov model that allows the past performance of each player to influence predictions for future performance. If we infer player i to have been elite in year t ($\mathrm{E}_{i,t} = 1$), then this inference influences the elite status of that player in his next year, $\mathrm{E}_{i,t+1}$ through the transition parameters $\boldsymbol{\nu}_k$. However, one potential limitation of these transition parameters $\boldsymbol{\nu}_k$ is that they are shared globally across all players at that position: each player at position k has the same probability of transitioning from elite to non-elite and vice versa. This model assumption allows us to pool information across players for the estimation of our transition parameters in (9), but may lead to loss of information if players are truly heterogeneous with respect to the probability of transitioning between elite and non-elite states. In order to address this possibility, we consider extending our model to allow player-specific transition parameters in our hidden Markov model.

Our proposed extension, which we call the PSHMM, has player-specific transition parameters $\boldsymbol{\nu}^i = (\nu^i_{00}, \nu^i_{01}, \nu^i_{10}, \nu^i_{11})$ for each player i, that share a common prior distribution,

$$\begin{aligned}(\nu^i_{00}, \nu^i_{01}) &\sim \text{Dirichlet}\,(\omega_{00k}\,,\,\omega_{01k})\\ (\nu^i_{11}, \nu^i_{10}) &\sim \text{Dirichlet}\,(\omega_{11k}\,,\,\omega_{10k})\end{aligned} \tag{11}$$

where k is the position of player i. Global parameters $\boldsymbol{\omega}_k = (\omega_{00k}, \omega_{01k}, \omega_{11k}, \omega_{10k})$ are now allowed to vary with flat prior distributions. This new hierarchical structure allows for transition probabilities $\boldsymbol{\nu}^i$ to vary between players, but still imposes some shrinkage towards a common distribution controlled by global parameters $\boldsymbol{\omega}_k$ that are shared across players with position k. Under this model extension, the new conditional posterior distribution for each $\boldsymbol{\nu}^i$ is

$$\begin{aligned}(\nu^i_{00}, \nu^i_{01})|\boldsymbol{E} &\sim \text{Dirichlet}\left(\mathrm{N}^i_{00} + \omega_{00k}\,,\,\mathrm{N}^i_{01} + \omega_{01k}\right)\\ (\nu^i_{11}, \nu^i_{10})|\boldsymbol{E} &\sim \text{Dirichlet}\left(\mathrm{N}^i_{11} + \omega_{11k}\,,\,\mathrm{N}^i_{10} + \omega_{10k}\right)\end{aligned} \tag{12}$$

where $\mathrm{N}^i_{ab} = \sum_{t=1}^{n_i - 1} \mathrm{I}(E_{i,t} = a, E_{i,t+1} = b)$.

To implement this extended model, we must replace step 4 in our Gibbs sampler with a step where we draw $\boldsymbol{\nu}^i$ from (12) for each player i. We must also insert a new step in our Gibbs sampler where we sample the global parameters $\boldsymbol{\omega}_k$ given our sampled values of all the $\boldsymbol{\nu}^i$ values for players at position k. This added step requires sampling $(\omega_{00k}, \omega_{01k})$ from the following conditional distribution:

$$p(\omega_{00k}, \omega_{01k}|\boldsymbol{\nu}) \propto \left[\frac{\Gamma(\omega_{00k} + \omega_{01k})}{\Gamma(\omega_{00k})\Gamma(\omega_{01k})}\right]^{n_k} \times \left[\prod_{i=1}^{n_k} \nu^i_{00}\right]^{\omega_{00k}-1} \times \left[\prod_{i=1}^{n_k} \nu^i_{01}\right]^{\omega_{01k}-1} \tag{13}$$

where each product is only over players i at position k and n_k is the number of players at position k. We accomplish this sampling by using a Metropolis-Hastings step with true distribution (13) and Normal proposal distributions: $\omega^{prop}_{00k} \sim \mathrm{N}(\hat{\omega}_{00k}, \sigma^2)$ and $\omega^{prop}_{01k} \sim \mathrm{N}(\hat{\omega}_{01k}, \sigma^2)$. The means of these proposal distributions are:

$$\hat{\omega}_{00k} = \overline{\nu}_{00k}\left(\frac{\overline{\nu}_{00k}(1-\overline{\nu}_{00k})}{s^2_{0k}} - 1\right) \quad \text{and} \quad \hat{\omega}_{01k} = (1-\overline{\nu}_{00k})\left(\frac{\overline{\nu}_{00k}(1-\overline{\nu}_{00k})}{s^2_{0k}} - 1\right) \tag{14}$$

with

$$\overline{\nu}_{00k} = \sum_{i=1}^{n_k} \nu^i_{00}\,/\,n_k \qquad \text{and} \qquad s^2_{0k} = \sum_{i=1}^{n_k} (\nu^i_{00} - \overline{\nu}_{00k})^2\,/\,n_k$$

where each sum is over all players i at position k and n_k is the number of players at position k. These estimates $\hat{\omega}_{00k}$ and $\hat{\omega}_{01k}$ were calculated by equating the sample mean $\overline{\nu}_{00k}$ and sample variance s^2_{0k} with the mean and variance of the Dirichlet distribution (13). Similarly, we sample $(\omega_{11k}, \omega_{10k})$ with the same procedure but with obvious substitutions.

3 Results and Model Comparison

Our primary interest is the prediction of future hitting events, $Y^{\star}_{t+j}$ for years $j = 1, 2, \ldots$ based on our model and observed data up to year t. We estimate the full posterior distribution (8) and then use this posterior distribution to predict home run totals $Y^{\star}_{i,2006}$ for each player i in the 2006 season. The 2006 season serves as an external validation of our method, since this season is not included in our model fit. We use our predicted home run totals $\boldsymbol{Y}^{\star}_{2006}$ for the 2006 season to compare our performance to several previous methods (Section 3.2) as well as evaluate several internal model choices (Section 3.1). In Section 3.3, we present inference for other parameters of interest from our model, such as the position-specific age curves.

3.1 Prediction of 2006 Home Run Totals: Internal Comparisons

We can use our posterior distribution (8) based on data from MLB seasons up to 2005 to calculate the predictive distribution of the 2006 hitting rate $\theta_{i,2006}$ for each player i.

$$\begin{aligned} p(\theta_{i,2006}|\boldsymbol{X}) \quad = \quad & \int p(\theta_{i,2006}|R_{i,2006}, A_{i,2006}, B_{i,2006}, \mathrm{E}_{i,2006}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ & \cdot p(\mathrm{E}_{i,2006}|\boldsymbol{E}_i, \boldsymbol{\nu}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E}_i|\boldsymbol{X})\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma}\, d\boldsymbol{\nu}\, d\boldsymbol{E} \end{aligned} \tag{15}$$

where $\boldsymbol{X}$ represents all observed data up to 2005. This integral is estimated using the sampled values from our posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{E}_i|\boldsymbol{X})$ that were generated via our Gibbs sampling strategy.

We can use the posterior predictive distribution (15) of each 2006 home run rate $\theta_{i,2006}$ to calculate the distribution of the home run total $Y^{\star}_{i,2006}$ for each player in the 2006 season.

$$p(Y^{\star}_{i,2006}|\boldsymbol{X}) \quad = \quad \int p(Y^{\star}_{i,2006}|M_{i,2006}, \theta_{i,2006}) \cdot p(\theta_{i,2006}|\boldsymbol{X})\, d\theta_{i,2006} \tag{16}$$

However, the issue with prediction of home run totals is that we must also consider the number of opportunities $M_{i,2006}$. Since our overall focus has been on modeling home run rates $\theta_{i,2006}$, we will use the true value of $M_{i,2006}$ for the 2006 season in equation (16). Using the true value of each $M_{i,2006}$ gives a fair comparison of the rate predictions $\theta_{i,2006}$ for each model choice, since it is a constant scaling factor. This is not a particularly realistic scenario in a prediction setting since the actual number of opportunities will not be known ahead of time.

Based on the predictive distribution $p(Y^{\star}_{i,2006}|\boldsymbol{X})$, we can report either a predictive mean $\mathrm{E}(Y^{\star}_{i,2006}|\boldsymbol{X})$ or a predictive interval $C^{\star}_i$ such that $p(Y^{\star}_{i,2006} \in C^{\star}_i|\boldsymbol{X}) \geq 0.80$. We can examine the accuracy of our model predictions by comparing to the observed home run totals $Y_{i,2006}$ for the 559 players in the 2006 season, which we did not include in our model fit. We use the following three comparison metrics:

1. **RMSE**: root mean square error of predictive means,

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum_i (\mathrm{E}(Y^\star_{i,2006}|\boldsymbol{X}) - Y_{i,2006})^2}$$

2. **Interval Coverage**: fraction of 80% predictive intervals $C^\star_i$ covering observed $Y_{i,2006}$

3. **Interval Width**: average width of 80% predictive intervals $C^\star_i$

In Table 1, we evaluate our full model outlined in Section 2.1 relative to several simpler modeling choices. Specifically, we examine a simpler version of our model without positional information or the mixture model on the α coefficients. We see from Table 1 that our full model gives proper coverage and a substantially lower RMSE than the version of our model without positional information or the elite/non-elite mixture model. We also examine a truly simplistic strawman, which is to take last years home run totals as the prediction for this years home run totals (ie. $Y^\star_{i,2006} = Y_{i,2005}$). Since this strawman is only a point estimate, that comparison is made based solely on the RMSE. As expected, the relative performance of this strawman model is terrible, with a substantially higher RMSE compared to our full model. Of course, this simple strawman alternative is rather naive and in Section 3.2, we compare our performance to more sophisticated external prediction approaches.

Table 1: Internal Comparison of Different Model Choices. Measures are calculated over 559 Players from 2006 season.

Model	RMSE	Coverage of 80% Intervals	Average Interval Width
Full Model	5.30	0.855	9.81
No Position or Elite Indicators	6.87	0.644	6.56
Strawman: $Y^\star_{i,2006} = Y_{i,2005}$	8.24	NA	NA
Player-Specific Transitions	5.45	0.871	10.36

We also considered an extended model in Section 2.3 with player-specific transition parameters for the hidden Markov model on elite status, and the validation results from this model are also given in Table 1. Our motivation for this extension was that allowing player-specific transition parameters might reduce the interval width for players that have displayed consistent past performance. However, we see that the overall prediction accuracy was not improved with this model extension, suggesting that there is not enough additional information in the personal history of most players to noticeably improve the model predictions. Somewhat surprisingly, we also see that the width of our 80% predictive intervals are not actually reduced in this extended model. The reason is that, even for players with long careers of data, the player-specific transition

parameters $\boldsymbol{\nu}^i$ fit by this extended model are not extreme enough to force all sampled elite indicators $E_{i,2006}$ to be either 0 or 1, and so the predictive interval is still wide enough to include both possibilities.

3.2 Prediction of 2006 Home Run Totals: External Comparisons

Similarly to Section 3.1, we use hold-out home run data for the 2006 season to evaluate our model predictions compared to the predictions from two external methods, PECOTA (Silver 2003) and MARCEL (Tango 2004), both described in Section 1. We view MARCEL as the primary competitor of our approach, as it also is a fully-automated method based on publicly available data. However, out of general interest we also compare our prediction accuracy to the proprietary and manually-curated PECOTA system. For a reasonable comparison set, we focus our external validation on hitters with an empirical home run rate of least 1 home run every 40 at-bats in at least one season up to 2005 (minimum of 300 at-bats in that season). This restriction reduces our dataset for model fitting down to 118 top home run hitters who all have predictions from the competing methods PECOTA and MARCEL. As noted above, our predicted home run totals for 2006 are based on the true number of at bats for 2006. In order to have a fair comparison to external methods such as PECOTA or MARCEL, we also scale the predictions from these methods by the true number of at bats in 2006.

Our approach has the advantage of producing the full predictive distribution of future observations (summarized by our predictive intervals). However, the external methods do not produce comparable intervals, so we only compare to other approaches in terms of prediction accuracy. We expand our set of accuracy measures to include not only the root mean square error (RMSE), but also the median absolute error (MAE). In addition to comparing the predictions from each method using overall error rates, we also calculated "% BEST" which is, for each method, the percentage of players for which the predicted home run total $Y^{\star}_{i,2006}$ is the closest to the true home run total among all methods. Each of these comparison statistics are given in Table 2. In addition to giving these validation measures for all 118 players, we also separate our comparison for young players (age $\leq$ 26 years in 2006) versus older players (age $>$ 26 years in 2006). The age cut-off of 26 years was used in order to isolate the small subset of players that were just beginning their careers and for which each player had little personal history of performance. It is worth noting that only 8 out of the 118 players (around 7%) in our 2006 test dataset were classified as young by this criterion, so the vast majority (110 out of 118) of players are in the "older" category.

We see from Table 2, that our model is extremely competitive with the external methods PECOTA and MARCEL. When examining all 118 players, our model has the smallest median absolute error and the highest "% Best" measure, suggesting that our predictions are superior on these absolute scales. Our performance is more striking when we examine only the small subset of young players in our dataset. We have the best prediction on 62% of all young players, and for these young players, both the RMSE and MAE from our method is substantially lower than either PECOTA or MARCEL. We credit this superior performance to our sophisticated hierarchical approach that builds in

Table 2: Comparison of our model to two external methods on the 2006 predictions of 118 top home run hitters. We also provide this comparison for only young players (age $\leq$ 26 years) versus only older players (age $>$ 26 years).

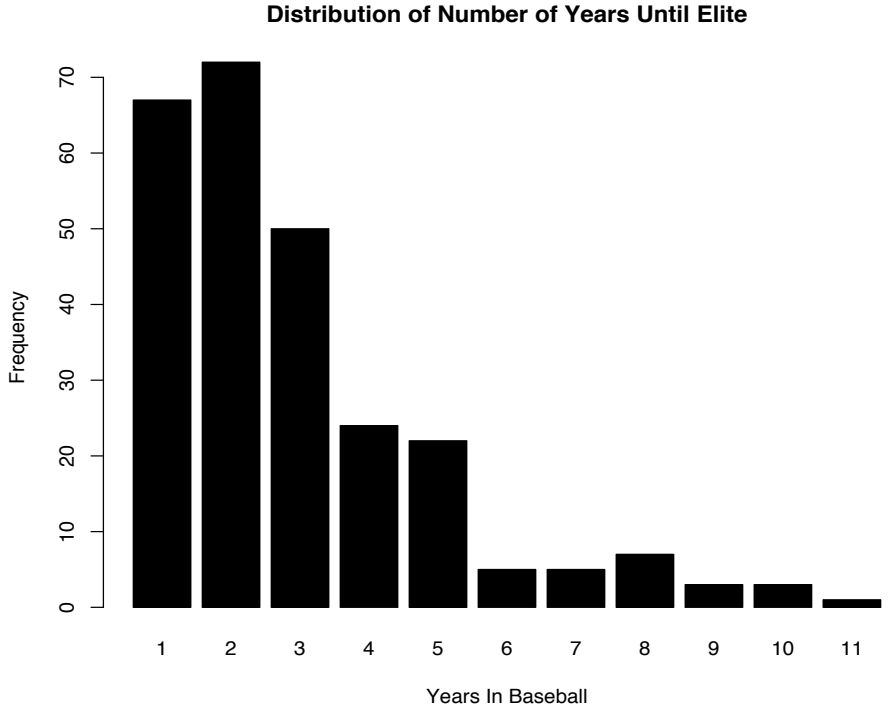
Method	**All Players**			**Young Players**			**Older Players**		
	RMSE	MAE	% BEST	RMSE	MAE	% BEST	RMSE	MAE	% BEST
Our Model	7.33	4.40	41 %	2.62	1.93	62%	7.56	4.48	39%
PECOTA	7.11	4.68	28 %	4.62	3.44	0%	7.26	4.79	30%
MARCEL	7.82	4.41	31 %	4.15	2.17	38%	8.02	4.57	31%

information via position instead of relying solely on limited past personal performance. All eight young players had played three seasons or less before 2006, and six of the eight players had two seasons or less before 2006. For these players, very little past information is available about their performance and so the model must rely heavily on position, where information is shared between players.

However, our method is not completely dominant: we have a larger root mean square error than PECOTA for older players (and overall), which suggests that our model might be making large errors on a small number of players. Further investigation shows that our model commits its largest errors for players in the designated hitter (DH) position. This is somewhat expected, since our model seems to perform best for young players and DH is a position almost always occupied by an older player. Beyond this, the model appears to be over-shrinking predictions for players in the DH role, perhaps because this player position is rather unique and does not fit our model assumptions as well as the other positions. Also, PECOTA is a manually-curated system that can account for the latest information in terms of injuries and playing time adjustments, which can greatly benefit their predictions. Overall, the validation results are generally very encouraging for our approach compared to our nearest competitor, MARCEL, as well as the proprietary system PECOTA. Our performance is especially good among younger players where a principled balance of positional information with past performance is most advantageous.

We further investigate our model dynamics among young players by examining how many years of observed performance are needed to decide that a player is an elite home run hitter. This question was posited in Section 1 and we now address the question using our elite status indicators E_{ij}. Taking all 559 available players examined in Section 3.1, we focus our attention on the subset of players that were determined by our model to be in the elite group ($P(E_{ij} = 1) \geq 0.5$) for at least two years in their career. For each elite home run hitter, we tabulate the number of years of observed data that were needed before they were declared elite. The distribution of the number of years needed is given in Figure 2. We see that although some players are determined to be elite based on just one year of observed data, most players (74%) need more than one year of observed performance to determine that they are elite home run hitters. In fact, almost half of players (46%) need more than two years of observed performance to determine that they are elite home run hitters.

Figure 2: Distribution of number of seasons of observed data needed to infer elite status ($\mathrm{P}(E_{ij} = 1) \geq 0.5$) among all players determined by our model to be elite during their career. Note that increasing the cut-off for elite states (e.g. $\mathrm{P}(E_{ij} = 1) \geq 0.75$) shifts the distribution towards a higher number of seasons needed, whereas decreasing the cut-off for elite states (e.g. $\mathrm{P}(E_{ij} = 1) \geq 0.25$) shifts the distribution towards a lower number of seasons needed.



We also investigated our model dynamics among older players by examining the balancing of past consistency with advancing age, which was also posited as a question in Section 1. Specifically, for the older players (age ≥ 35) in our dataset, we examined the differences between the 2006 home run rate predictions $\hat{\theta}_{i,2006} = \mathrm{E}(\theta_{i,2006}|\boldsymbol{X})$ from our model versus the naive prediction based entirely on the previous year $\tilde{\theta}_{i,2006} = Y_{i,2005}/M_{i,2005}$. Is our model contribution for a player (which we define as the difference between our model prediction $\hat{\theta}_{i,2006}$ and the naive prediction $\tilde{\theta}_{i,2006}$) more a function of advancing age or past consistency of that player? Both age and past consistency (measured as the standard deviation of their past home run rates) were found to be equally good predictors of our model contribution, which suggests that both sources of information are being evenly balanced in the predictions produced by our model.

3.3 Age Trajectory Curves

In addition to validating our model in terms of prediction accuracy, we can also examine the age trajectory curves that are implied by our estimated posterior distribution (8). We will examine these curves on the scale of the home run rate θ_{ij} which is a function of age A_{ij}, ball-park b, and elite status E_{ij} for player i in year j (with position k):

$$\theta_{ij} = \frac{\exp\left[(1-\mathrm{E}_{ij})\cdot\alpha_{k0} + \mathrm{E}_{ij}\cdot\alpha_{k1} + \beta_b + f_k(A_{ij})\right]}{1+\exp\left[(1-\mathrm{E}_{ij})\cdot\alpha_{k0} + \mathrm{E}_{ij}\cdot\alpha_{k1} + \beta_b + f_k(A_{ij})\right]}. \tag{17}$$

The shape of these curves can differ by position k, ballpark b and also can differ between elite and non-elite status as a consequence of having a different additive effect α_{k0} vs. α_{k1}. In Figure 3, we compare the age trajectories for two positions, DH and SS, for both elite player-years ($E_{ij}=1$) vs. non-elite player-years ($E_{ij}=0$) for an arbitrary ballpark. Each graph contains multiple curves (100 in each graph), each of which is the curve implied by the sampled values $(\boldsymbol{\alpha},\boldsymbol{\gamma})$ from a single iteration of our converged and thinned Gibbs sampling output. Examining the curves from multiple samples gives us an indication of the variability in each curve.

We see a tremendous difference between the two positions DH and SS in terms of the magnitude and shape of their age trajectory curves. This is not surprising, since home run hitting ability is known to be quite different between designated hitters and shortstops. In fact, DH and SS were chosen specifically to illustrate the variability between position with regards to home run hitting. For the DH position, we also see that elite vs. non-elite status show a substantial difference in the magnitude of the home run rate, though the overall shape across age is restricted to be the same by the fact that players of both statuses share the same $f_k(A_{ij})$ in equation (17). There is less difference between elite and non-elite status for shortstops, in part due to the lower range of values for shortstops overall. Not surprisingly, the variability in the curves grows with the magnitude of the home run rate.

We also perform a comparison across all positions by examining the elite vs. non-elite intercepts $(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)$ that were allowed to vary by position. We present the posterior distribution of each elite and non-elite intercept in Figure 4. For easier interpretation, the values of each α_{k0} and α_{k1} have been transformed into the implied home run rate θ_{ij} for very young (age = 23) players in our dataset. We see in Figure 4 that the variability is higher for the elite intercept in each position, and there is even more variability between positions. The ordering of the positions is not surprising: the corner outfielders and infielders have much higher home run rates than the middle infielder and centerfielder positions.

For a player at a specific position, such as DH, our predictions of his home run rate for a future season is a weighted mixture of elite and non-elite DH curves given in Figure 3. The amount of weight given to elite vs. non-elite for a given player will be determined by the full posterior distribution (8) as a function of that player's past performance. We illustrate this characteristic of our model in more detail in Figure 5 by examining six different hypothetical scenarios for players at the 2B position. Each plot in Figure 5 gives several seasons of past performance for a single player, as well

Figure 3: Age Trajectories $f_k(\cdot)$ for two positions and elite vs. non-elite status. X-axis is age and Y-axis is Rate $= \theta_{ij}$

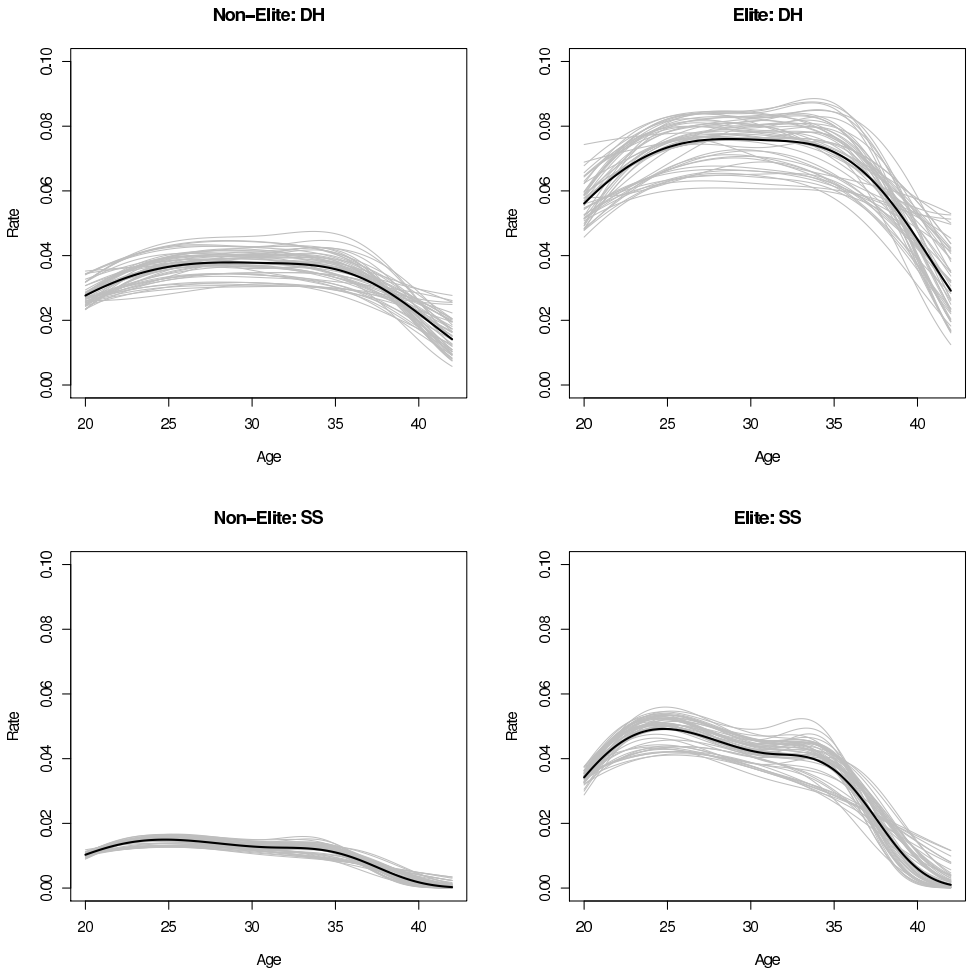
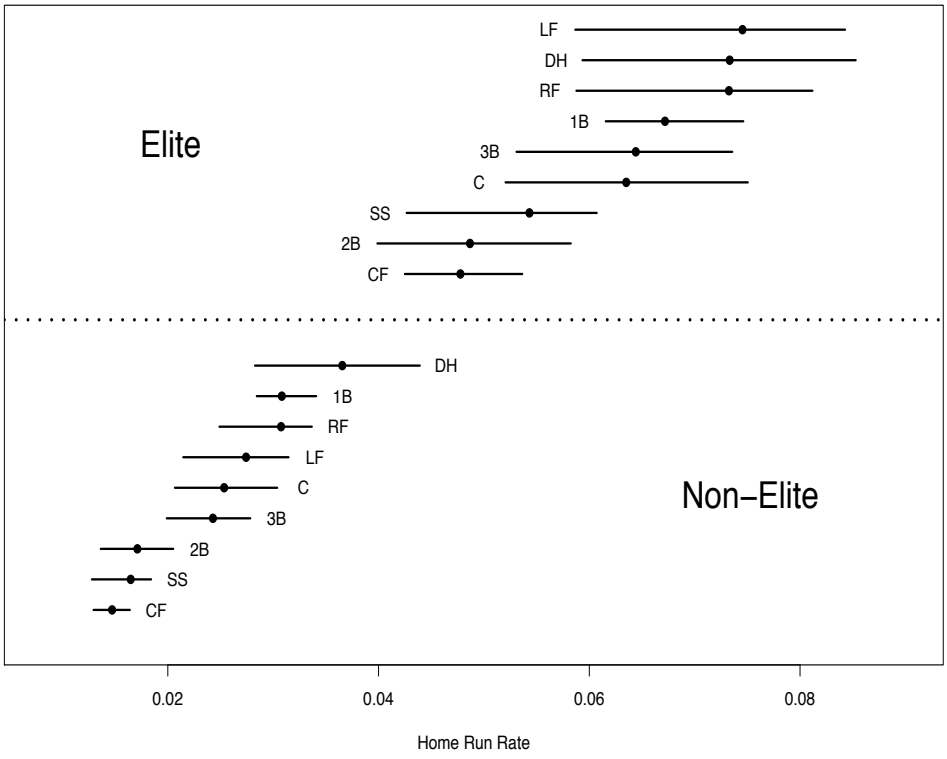


Figure 4: Distribution of the elite vs. non-elite intercepts $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ for each position. The distributions of each $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ are presented in terms of the home run rate θ_{ij} for very young (age = 23) players. The posterior mean is given as a black dot, and the 95% posterior interval as a black line.



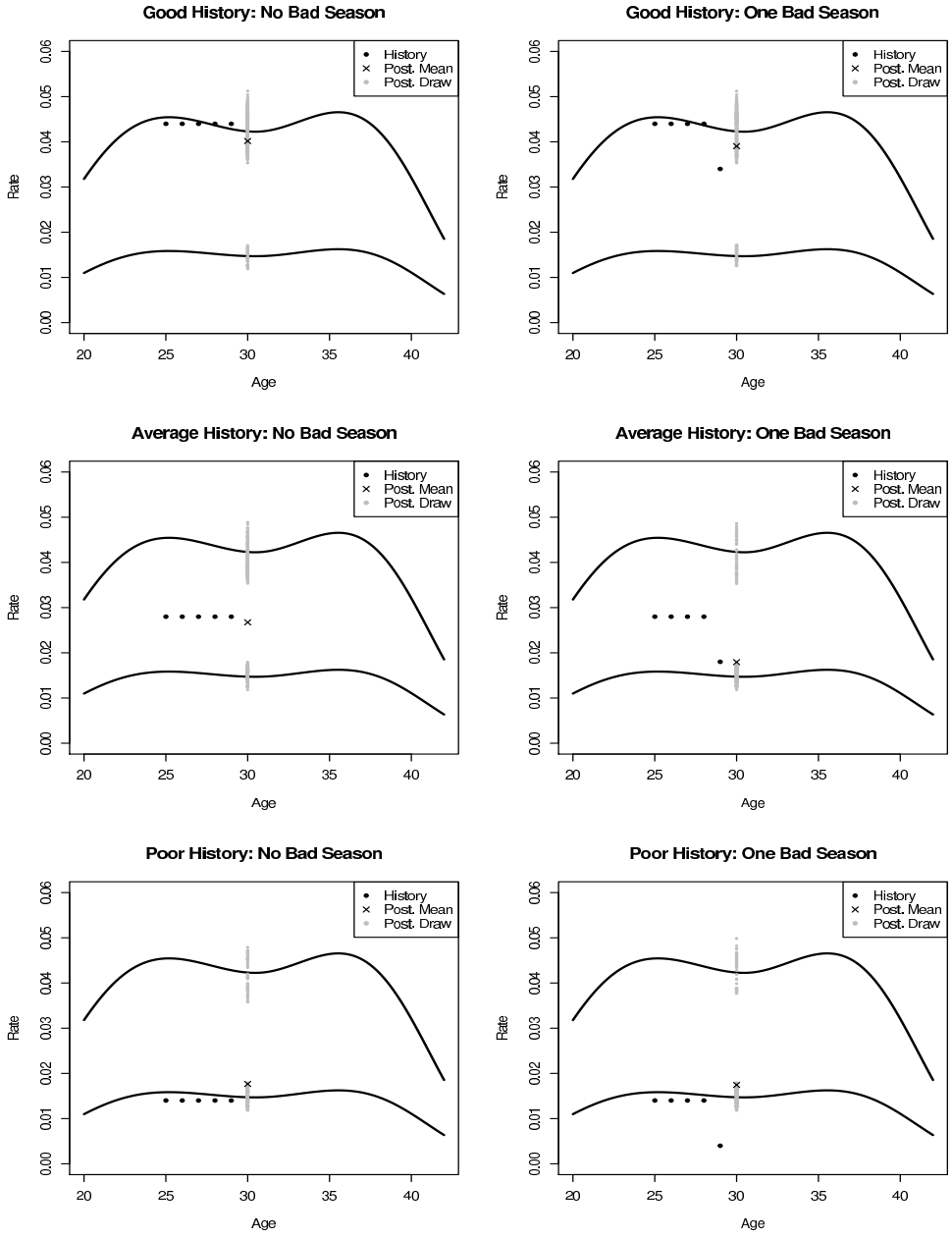
as predictions for an additional season (age 30). Predictions are given both in terms of posterior draws of the home run rate as well as the posterior mean of the home run rate. The elite and non-elite age trajectories for the 2B position are also given in each plot. We focus first on the left column of plots, which shows hypothetical players with consistently high (top row), average (middle row), and poor (bottom row) past home run rates. We see in each of these left-hand plots that our posterior draws (gray dots) for the next season are a mixture of posterior samples from the elite and non-elite curves, though each case has a different proportion of elite vs. non-elite, as indicated by the posterior mean of those draws (black $\times$).

Now, what would happen if each of these players was not so consistent? In Section 1, we asked about the effect of a single sub-par year on our model predictions. The plots in the right column show the same three hypothetical players, but with their most recent past season replaced by a season with distinctly different (and relatively poor) home run hitting performance. We see from the resulting posterior means in each case that only the average player (middle row) has his predictions substantially affected by the one season of relatively poor performance. Despite the one year of poor performance, the player in the top row of Figure 5 is still considered to be elite in the vast majority of posterior draws. Similarly, the player in the bottom row of Figure 5 is going to be considered non-elite regardless of that one year of extra poor performance. The one season of poor performance has the most influence on the player in the middle row, since the model has the most uncertainty with regards to the elite vs. non-elite status of this average player.

4 Discussion

We have presented a sophisticated Bayesian hierarchical model for home run hitting among major league baseball players. Our principled approach builds upon information about past performance, age, position, and home ballpark to estimate the underlying home run hitting ability of individual players, while sharing information across players. Our primary outcome of interest is the prediction of future home run hitting, which we evaluated on a held out season of data (2006). When compared to the previous methods, PECOTA (Silver 2003) and MARCEL (Tango 2004), we perform well in terms of prediction accuracy, especially our "% BEST" measure which tabulates the percentage of players for which our predictions are the closest to the truth. Our prediction accuracy completely dominates the MARCEL procedure which represents our closest natural competitor, since it is also a fully-automated and based on publicly-available data. Our prediction accuracy is also competitive with the proprietary PECOTA system which is especially impressive given that PECOTA is manually curated based on the latest information about injuries and playing time. Our approach does especially well among young players, where a principled balance of positional information with past performance seems most helpful. In addition, our method has the advantage of estimating the full posterior predictive distribution of each player, which provides additional information in the form of posterior intervals. Beyond our primary goal of prediction, our model-based approach also allows us to answer interesting supplemental questions

Figure 5: Six different hypothetical scenarios for a player at the 2B position. Black curves indicate the elite and non-elite age trajectories for the 2B position. Black points represent several seasons of past performance for a single player. Predictions for an additional season are given as posterior draws (gray points) of the home run rate and the posterior mean of the home run rate (black ×). Left column of plots gives hypothetical players with consistently high (top row), average (middle row), and poor (bottom row) past home run rates. Right column of plots show the same hypothetical players, but with their most recent past season replaced by a relatively poor home run hitting performance.



such as the ones posed in Section 1.

We have illustrated our methodology using home runs as the hitting event since they are a familiar outcome that most readers can calibrate with their own anecdotal experience. However, our approach could easily be adapted to other hitting outcomes of interest, such as on-base percentage (rate of hits or walks) which has become a popular tool for evaluating overall hitting quality. Also, although our procedure is presented in the context of predicting a single hitting event, we can also extend our methodology in order to model multiple hitting outcomes simultaneously. In this more general case, there are several possible outcomes of an at-bat (out, single, double, etc.). Our units of observation for a given player i in a given year j is now a vector of outcome totals $\boldsymbol{Y}_{ij}$, which can be modeled as a multinomial outcome: $\boldsymbol{Y}_{ij} \sim \text{Multinomial}(M_{ij}, \boldsymbol{\theta}_{ij})$ where M_{ij} are the number of opportunities (at bats) for player i in year j and $\boldsymbol{\theta}_{ij}$ is the vector of player- and year-specific rates for each outcome. Our underlying model for the rates θ_{ij} as a function of position, ball-park and past performance could be extended to a vector of rates $\boldsymbol{\theta}_{ij}$. Our preliminary experience with this type of multinomial model indicates that single-event predictions (such as home runs) are not improved by considering multiple outcomes simultaneously, though one could argue that a more honest assessment of the variance in each event would result from acknowledging the possibility of multiple events from each at-bat.

An important element of our approach was the use of mixture modeling of the player population to further refine our estimated home run rates. Sophisticated statistical models have been used previously to model the careers of baseball hitters (Berry et al. 1999), but these approaches have not employed mixtures for the modeling of the player population. Our internal model comparisons suggest that this mixture model component is crucial for the accuracy of our model, dominating even information about player position. Using a mixture of elite and non-elite players limits the shrinkage towards the population mean of consistently elite home run hitters, leading to more accurate predictions. Our fully Bayesian approach also allows us to investigate the dynamics of our elite status indicators directly, as we do in Section 3.2.

In addition to our primary goal of home run prediction, our model also estimates several secondary parameters of interest. We estimate career trajectories for both elite and non-elite players within each position. In addition to evaluating the dramatic differences between positions in terms of home run trajectories, our fully Bayesian model also has the advantage of estimating the variability in these trajectories, as can be seen in Figure 3. It is worth noting that our age trajectories do not really represent the typical major league baseball career, especially at the higher values of age. More accurately, our trajectories represent the typical career conditional on the player staying in baseball, which is one reason why we do not see dramatic dropoff in Figure 3. Since our primary goal is prediction, the fact that our trajectories are conditional is acceptable, since one would presumably only be interested in prediction for baseball players that are still in the major leagues. However, if one were more interested in estimating unconditional trajectories, then a more sophisticated modeling of the drop-out/censoring process would be needed.

Our focus in this paper has been the modeling of home run rates θ_{ij} and so we have made an assumption throughout our analysis that the number of plate appearances, or opportunities, for each player is a known quantity. This is a reasonable assumption when retrospectively estimating past performance, but when predicting future hitting performance the number of future opportunities is not known. In order to maintain a fair comparison between our method and previous approaches for prediction of future totals, we have used the future number of opportunities, which is not a reasonable strategy for real prediction. A focus of future research is to adapt our sophisticated hierarchical approach to the modeling and prediction of plate appearances M_{ij} in addition to our current modeling of hitting rates θ_{ij}.

References

Berry, S. M., Reese, S., and Larkey, P. D. (1999). "Bridging Different Eras in Sports." *Journal of the American Statistical Association*, 94: 661–686. 649

Brown, L. D. (2008). "In-Season Prediction of Batting Averages: A Field-test of Simple Empirical Bayes and Bayes Methodologies." *Annals of Applied Statistics*, 2: 113–152. 632, 633

Chib, S. (1996). "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics*, 75: 79–97. 637

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag. 634

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC, 2nd edition edition. 636

Gelman, A., Roberts, G., and Gilks, W. (1996). "Efficient Metropolis jumping rules." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 5*, 599–608. Oxford University Press. 637

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6: 721–741. 636

Lahman, S. (2006). "Baseball Archive." *Lahman's Baseball Database*, Version 5.5. URL http://www.baseball1.com/ 632, 633

Quintana, F. A., Mueller, P., Rosner, G. L., and Munsell, M. (2008). "Semi-parametric Bayesian Inference for Multi-Season Baseball Data." *Bayesian Analysis*, 3: 317–338. 632

Silver, N. (2003). "Introducing PECOTA." *Baseball Prospectus*, 2003: 507–514. 631, 641, 647

Tango, T. (2004). "Marcel The Monkey Forecasting System." *Tangotiger.net*, March 10, 2004.
URL http://www.tangotiger.net/archives/stud0346.shtml 631, 632, 641, 647

Acknowledgments

We would like to thank Dylan Small and Larry Brown for helpful discussions.

Bayesian Analysis (2009) 4, Number 4, pp. 653–660

Comment on Article by Jensen et al.

Jim Albert* and Phil Birnbaum†

1 Introduction

Prediction of future batting performance is an important problem in baseball. Due to trades and the free agent system, there is a good movement of players between teams in the "hot-stove league" (the baseball off-season) and teams will acquire new players with the hope that they will achieve particular performances in the following season. The authors propose a Bayesian hierarchical modeling framework for estimating home run hitting probabilities and making predictions of future home run hitting performance. Generally, this is an attractive methodology, especially when one is collecting data from many players who have similar home run hitting abilities. By use of hierarchical modeling, the estimates of the home run probabilities shrink or adjust the observed rates towards a combined regression estimate. One attractive feature of the Bayesian approach is that it is straightforward to obtain predictions from the posterior predictive distribution and the authors test the value of their method by comparing it with two alternative prediction systems MARCEL and PECOTA. It is straightforward to fit these hierarchical models by MCMC algorithms and the authors provide the details of this fitting algorithm.

Although we admire the authors' paper from a Bayesian modeling/computation perspective, it seems deficient from the application (baseball perspective). There is a substantial research on home run hitting and in the modeling of career trajectories of ballplayers and we believe this research should be helpful in defining relevant covariates and proposing realistic models for trajectories. In the following comments, we discuss several concerns with the basic modeling framework, focus on the choice of suitable adjustments and suggest a more flexible framework for modeling career trajectories.

2 Data

The authors use data from the Lahman database where the counts of home runs and at-bats are collected for each player for each season in the period 1990 and 2005. Although this is a rich dataset, we are puzzled that the authors did not use the more detailed play-by-play data available from the Retrosheet organization (`www.retrosheet.org`). This dataset is easy to access and manipulate. As will be seen shortly, this richer dataset would allow for the inclusion of suitable covariates in the adjustment of the home run rates.

*Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, http://www-math.bgsu.edu/~albert/

†Society of American Baseball Research, http://philbirnbaum.com/

 DOI:10.1214/09-BA424A

3 Adjustments for Home Run Rates

In comparing baseball hitters across eras, Schell (2005) explains the importance of adjusting home run rates for the era of play, the distribution of league-wide talent, the ballpark effect, and a player's late-career decline. Adjustments for league-wide talent and the ballpark are also crucial in the modeling of a player's hitting trajectory and the prediction of future performance. There have been dramatic changes in home run hitting from 1990 to 2005. The overall major league home run rate increased by 26% between 1992 and 1993 and the rate has shown a 50% increase in this 15 year period. Schell documents the significant impact of ballparks on the pattern of home run hitting. In the current baseball season, it appears to be much easier to hit home runs in the new Yankee stadium in New York. The park factor for the new Yankee Stadium is currently 1.295, which means that the rate of home run hitting in the Yankee home games is about 30% higher than the rate of home run hitting in the Yankees away games.

One can understand changes in league-wide hitting talent by the fitting of a random effects model. For a given season, we observe the number of home runs and at-bats (y_i, n_i) for all batters. We assume that y_i is binomial(n_i, p_i) and then we assume the home run probabilities $\{p_i\}$ follow a beta distribution with shape parameters a and b. The fitted values $\hat{a}$ and $\hat{b}$ are informative about the location and shape of the home run abilities of the batters. This random effects model is fit separately for each season, obtaining estimates $\hat{a}_j$ and $\hat{b}_j$ for season j. The top graph in Figure 1 displays the median home run ability of the players for the seasons 1990 to 2005, and the bottom graph plots the interquartile spread of the home run ability distribution against season. This figure shows dramatic changes in the location and spread of talent of hitting home runs in this 15 year period. One way of adjusting a player's season home run rate compares his rate relative to the distribution of home run rates hit for that particular season. Specifically, one can compute a predictive standardized score as in Albert (2009) using the average and standard deviation of the predictive distribution.

This paper does include some adjustments in their regression model (2), specifically, covariates for the home ballpark and fielding position. As the authors explain, the data does not break down a player's home run data by home and away games and so the "home ballpark" covariate actually confounds two variables, the ballpark effect and the team hitting ability. One could define a true ballpark effect by using the Retrosheet data. We are puzzled by the inclusion of the fielding position covariate. Although there are some tendencies, for example, first-basemen tend to hit more home runs than second-basemen, modern hitters of all non-pitching positions are proficient in hitting home runs. Why do the authors believe that fielding position is an important covariate? More importantly, why do the authors believe that players of different positions have different home run trajectories?

Another possible regression adjustment is the number of opportunities AB. There is a general positive correlation between AB and home run rate – players with more at-bats tend to hit a higher rate of home runs. Also, if a young player has a limited number of AB one season, it is more likely that he will have a small number of home runs and be sent back to the minors the following season. Also the number of AB and

the player's career trajectory provides a good prediction of the player's AB in a future season. (The authors assume that the player's 2006 AB is the same as the AB in the previous season.)

4 Elite/Non-Elite Players

The authors introduce a latent elite variable in their model with the justification that "that there exists a sub-group of elite home run hitters within each position that share a higher mean home run rate". The authors do not present any evidence in the paper that home run rates cluster in two groups of non-elite players and elite players. In our exploration of these data, there appears to be a continuum of home run ability that is right skewed with a few possible large outliers. It seems that the latent elite variable is introduced not because the data suggests the two clusters, but rather to induce some dependence in the home run rates for the same player. There is a more straightforward way to model this dependence, specifically to assume that each player has a unique trajectory, where the individual player regression coefficient vectors are assumed to follow a common distribution. This comment relates to the authors' approach for modeling trajectories which will be described next.

5 Modeling Career Trajectories

In the motivation for the career trajectories, the authors say that they "favor an approach that involves fewer parameters (to prevent over-fitting)". But they make the very restrictive assumption that players of a particular fielding position share the same career trajectory. This assumption does not reflect the variable trajectory patterns of home run hitting. To illustrate the variability in trajectories, consider the home run hitting patterns of the Hall of Fame players Mickey Mantle and Hank Aaron (both who played the same outfield position) who played in the same era. Figure 2 plots standardized home run rates for both players as a function of age, where the rates have been standardized using the predictive distribution as described above. Note that Mantle peaked in his late 20's and declined quickly until retirement. In contrast, Aaron peaked in home run hitting ability much later in his career and showed a more gradual decline towards the end of his career.

It can be difficult to estimate the player trajectories individually using regression models due to the high variability of the observed rates as shown in Figure 1. But one can obtain good smoothed estimates of the individual trajectories by use of a multilevel model. If the vector of regression coefficients for the ith player is represented by β_i, then one can assume that the $\{\beta_j\}$ are a random sample from a common normal distribution with mean vector β and variance-covariance matrix Σ, and the hyperparameters β, Σ are assigned a vague prior at the second state. The posterior estimates smooth the individual trajectory estimates towards a common trajectory. This multilevel model is shown to be successful in smoothing trajectories of WHIP (walk and hit) rates for pitchers in Albert (2009). We have also used it for estimating trajectories of batter on-

base percentages, and we would expect similar good results for estimating trajectories of home run rates. This analysis would lead to more realistic estimates of career trajectories and likely better predictions of future home run hitting. Certainly, one should make different predictions for the home run hitting for a 35-year old Mickey Mantle and a 35-year old Hank Aaron since their patterns of decline were very different.

6 A Sabermetrics Perspective

Sabermetrics is the scientific search for objective knowledge about baseball, and the search for better predictions of future performance is certainly something that sabermetricians – especially those who may be employed by major league clubs - are interested in. But they are concerned with more than just accurate predictions; they are concerned with what it is the projection reveals about players and changes in their performance.

Bill James, in a discussion about the existence of clutch hitting in James (1984), says "How is it that a player who possesses the reflexes and the batting stroke and the knowledge and the experience to be a .260 hitter in other circumstances magically becomes a .300 hitter when the game is on the line? How does that happen? What is the process? What are the effects? Until we can answer those questions, I see little point in talking about clutch ability." Likewise, sabermetricians are interested in the process that leads to a prediction of home run hitting.

Sabermetricians are unsatisfied with mere predictions, no matter how accurate. Given an accurate prediction of future performance, they ask, "what is it about that prediction that makes it accurate? What does it tell us about the relationship of past performance to future performance?"

One attractive feature of MARCEL is that it gives us clues to what might be going on. Tango (2004) gives the full MARCEL algorithm, in which we can see the assumptions that went into the formula. We see how it weights recent performance relative to more distant performance, how much one should regress to the mean, and how one adjusts the predictions to adjust to changes to league norms. These individual assumptions can be adjusted in order to minimize prediction error, and, in so doing, we would come closer to learning objective information about player hitting.

The Bayesian modeling approach presented in this paper, however, is more complex and opaque. It performs only marginally better than MARCEL, while using more information such as home team scoring and player position. It is uncertain what an experienced sabermetrician would learn from the Bayesian process, and it is uncertain whether the (marginally) improved predictions are the result of a better model, or simply the result of the additional information being used.

Further, while the Bayesian model has shown itself to be successful in predicting, certain of its assumptions are almost certain to be false. As has been noted, the classification of hitters into only two categories – elite and non-elite – is certainly false, as home-run-hitting ability appears to be a continuum; there is no evidence that the distribution of home run rates, even by position, is bimodal.

The fact that the Bayesian model gives reasonable estimates cannot be taken as evidence that the assumptions are correct. For instance, a black-box model that predicts swine-flu infection rates is valuable, but, if the assumptions that went into the model are correct, this model is useful in predicting future outbreaks. If the assumptions are incorrect, the predictions based on the model may be inaccurate.

Sabermetricians would be very interested in the success of the Bayesian model in predicting home run rates for younger hitters; as Table 2 of the paper shows, the Bayesian algorithm beats MARCEL 62% of the time, and beats PECOTA 100% of the time. We note, however, that this is based only on a sample of eight players. Still, one could discover possible attributes of the prediction methodology by a case-by-case exploration. It would useful to see the full list of players and their estimates, along with a discussion of what kinds of players, such as power hitters or high-average players, are better estimated than others types of players. This would provide a useful comparison of the methods, and provide a direction for future research to improve the knowledge that the field of sabermetrics has compiled about the aging process.

As it stands now, the Bayesian method has made sabermetrics aware that slight improvements over MARCEL are possible, but, without further exploration, we are left with little understanding of where the improvements came from, where MARCEL is weak, what assumptions need to be refined, or, indeed, how the aging process in baseball can better be explained.

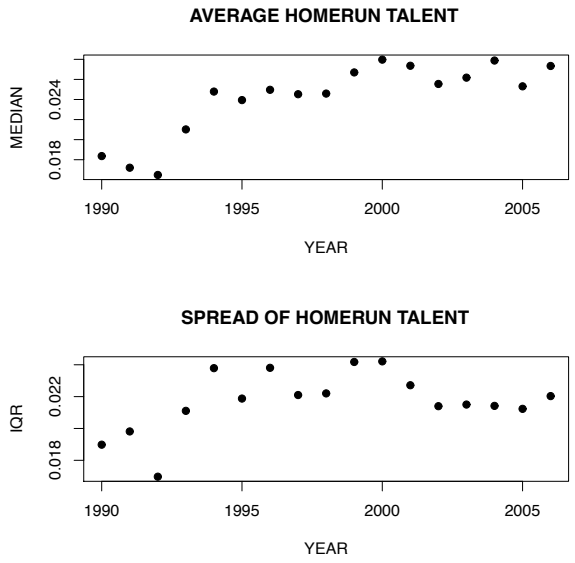


Figure 1: Fitted home run talent distributions for the seasons 1990 to 2005. The top graph displays the median home run ability and the bottom graph displays the interquartile range of the talent distribution.

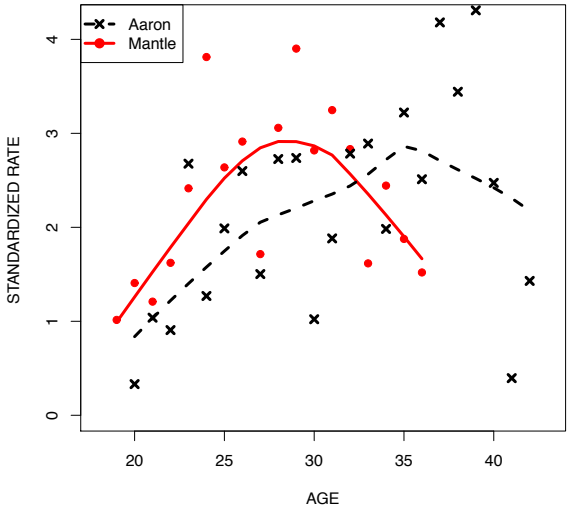


Figure 2: Standardized home run rates for Mickey Mantle and Hank Aaron ploted as a function of age. The lowess smooths show that the home run trajectories of the two players were significantly different.

7 Summing Up

The authors have proposed a useful hierarchical modeling framework and illustrated the potential benefits of Bayesian modeling in predicting future home run counts. But we believe the methods could be substantially improved by the proper adjustment of the home run rates, the inclusion of useful covariates, and more realistic modeling of the career trajectories. From the viewpoint of a baseball general manager, the prediction of a particular player's future performance is very important and it seems that this prediction has to allow for the player's unique career trajectory pattern. For the problem of individual predictions, we don't believe this methodology will be very helpful, since all players of a particular fielding position are assumed to have the same trajectory and lumped into the broad elite/non-elite classes. But we do believe that this general approach, with the changes described above, can be used to make helpful predictions of offensive performance.

References

Albert, J. (2009). "Is Roger Clemens' WHIP Trajectory Unusual." *Chance*, 22: 8–22.

James, B. (1984). *The 1984 Baseball Abstract*. Ballentine Books.

Schell, M. (2005). *Baseball's All-Time Best Sluggers: Adjusted Batting Performance*

from Strikeouts to Home Runs. Princeton University Press.

Tango, T. (2004). "Marcel the Monkey Forecasting System." Http://www.tangotiger.net/archives/stud0346.shtml.

Bayesian Analysis (2009) **4**, Number 4, pp. 661–664

Comment on Article by Jensen et al.

Mark E. Glickman*

I offer my congratulations to Jensen, McShane and Wyner (hereafter JMW) on their paper modeling home run frequencies of Major League Baseball (MLB) players. It is always refreshing to read such a clearly written, well-organized paper on a topic of interest to a broad audience and one that illustrates cutting edge modeling and computational tools in Bayesian Statistics. It is also worth noting that the first author is becoming an accomplished researcher in quantitative aspects of baseball, most recently having developed complex statistical models for evaluating fielding (Jensen et al. 2009). The current paper adds to his accruing and impressive list of work on Statistics in sports.

In the current paper, the authors develop and investigate a model for home run frequencies for MLB seasons from 1990 through 2005 based on publicly available data. The data contains player performance information aggregated by season, so examining within-season variation is not possible. Home run frequencies for a player within a season are modeled as binomial counts (out of the total number of at-bats, appropriately defined), and the probability of a home run during a season is a function of the player's position, team, and age. The authors make some interesting specific assumptions that result in a unique model. First, they posit that the effect of age on the log-odds of the probability of a home run follows a cubic B-spline relationship for a given field position. Second, they assume a latent categorization of each player in a given season as elite versus non-elite, essentially treating a player's home run frequency as a mixture of two binomial components with different probabilities. Third, the latent elite status for each player is assumed to follow a Markov process with transition probabilities that are common for all players at the given field position. The authors also investigate a generalization of their basic model in which the transition probabilities can vary by player through model components specific to players at that position. The entire model is fit via MCMC simulation from the posterior distribution, and performance of their approach is evaluated through measures that compare model predictions in 2006 to observed home run frequencies. They conclude that their basic model fares well against existing competitor approaches that are not nearly as sophisticated. The authors deserve credit for constructing a model that is competitive with one that makes use of data obtained on a daily basis. It is also particularly impressive that their model predicts well given the paucity of covariate information.

One can raise minor quibbles with the authors' approach, but many of the concerns are an artifact of the constraints on the data available to them. For example, the ability to account for within-season variation strikes me as a clear deficiency in modeling home run probabilities. Given that players are generally improving from year to year in their twenties, it is not unreasonable to speculate that some of this improvement is occurring within a season rather than between seasons. Because the data JMW use is aggregated

*Boston University School of Public Health, Boston, MA, mailto:mg@bu.edu

 DOI:10.1214/09-BA424B

by season, it is impossible to infer such changes. The authors also incorporate a team indicator in their model, which ostensibly is a proxy for playing half of the time in their own ballpark, though this does not account for minor artifacts such as within-season player trades. As JMW note, this team parameter may be difficult to interpret when it applies to a whole season of games. If individual game-specific data were available, then the impact of the actual ballpark could be incorporated into the model which may have a profound effect on inferences. My own bias is to wonder whether modeling and predicting home run frequencies is a question that baseball front office staff or other professionals really want answered. While forecasting home run probabilities seems like an interesting theoretical question, various metrics to measure hitting rates might be of greater practical utility. The authors do mention at the conclusion of the paper their interest in pursuing such activities. I also found curious that the expanded model involving Markov transition probabilities that varied by player produced worse predictions than the simpler model in which the transition probabilities were constrained to vary only by player-position. This may suggest some combination of a model not sufficiently capturing important features of the data, or an expanded model that is too highly parameterized.

To me, the most interesting aspect of the paper is the decision to incorporate a latent indicator of elite status into the model, and the accompanying stochastic process. On the one hand, JMW are able to account for variation in home run rates and improve predictions by introducing a 2-state hidden Markov model (HMM). One clear benefit of incorporating this model component is that it allows answering questions about when certain players can be considered elite versus non-elite. On the other hand, I wonder whether a 2-state Markov model is the most appropriate and most flexible for predicting home run frequencies. The authors consider a HMM in which players at the same position share the same transition probabilities, and another in which the transition probabilities vary by player but are centered at position-specific distributions. In both cases, the size of the effect of being elite for all players at the specified position is the same. I realize that JMW are focused on keeping the model as simply parameterized as possible, but the question arises whether accuracy (especially predictive accuracy, one of the main implied goals of the paper) is being sacrificed. Given that all the parameters of the HMM are integrated out of the posterior distribution in making predictions, it is the *structure* of the HMM that is most crucial, and not inferences about any of the HMM parameters.

The authors' HMM assumes that players at any given time are in one of two states, once accounting for age, position and team. However, it strikes me that player effects (beyond the effect of age, position and team) more justifiably fall on a continuum. A natural way to modify JMW's model is to assume

$$\text{logit}\ \theta_{ijkb} = \alpha_k + \beta_b + f_k(A_{ij}) + \delta_{ijk} \tag{1}$$

where θ_{ijkb} is the home run probability for player i with home ballpark b in season j at position k; α_k, β_b and $f_k(A_{ij})$ are as defined in JMW; and δ_{ijk} is a player-specific effect following a stochastic process with a continuous state-space, such as

$$\delta_{ijk} \sim \text{N}(\delta_{i,j-1,k},\ \psi^2), \tag{2}$$

where initial player effects may be assumed drawn from a common distribution centered at a position-specific model component,

$$\delta_{i1k} \sim \mathrm{N}(\eta_k,\ \phi^2) \tag{3}$$

with position-specific effects η_k. This model assumes that, beyond the effects of ballpark, position and age, an individual player effect in a given season is drawn from a distribution centered at last season's mean, thus inducing a time-correlation particular to that player. Such an approach can represent trajectories of not only elite players, but also better-than-average players as well as worse-than-average players. Similar models for binomial data in a game/sports context have been examined by Fahrmeir and Tutz (1994) and Glickman (1999), among others, though these approaches do not include an additive spline component for age. Various changes to the assumptions in (2) and (3) could be considered, such as having the innovation variance, ψ^2, depend on player position (that is, ψ_k^2), the transition model could be heavy-tailed, such as a t-distribution instead of normal (which would account for occasional bursts of improvement in home run probability), or having the prior variance, ϕ^2, depend on the player position (that is, ϕ_k^2).

An advantage to a continuous state-space compared to a 2-state system is that it recognizes varying degrees of improvement and worsening over time beyond what is captured by age-specific effects. Substituting the HMM in the authors' framework with that in (2) should involve straightforward modifications to the MCMC algorithm, so the computational details ought to involve tractable calculations. Again, because the parameters of a continuous state-space model are integrated out of the posterior distribution to obtain predictive inferences, or even age-curve estimates, the richer structure compared to the 2-state HMM may result in more reliable inferences. The richer structure may also more appropriately calibrate the levels of uncertainty in predictions which appear overly conservative as evidenced in Table 1 of their paper. Of course, one needs to fit such a model to the data to be convinced of such speculation.

Notwithstanding some of my suggestions for alternative directions the authors could take in further refining their model, I think that their approach makes an important contribution to a growing literature on sophisticated methods in analyzing sports data. Modeling the effect of age through a cubic B-spline is a nice feature of their approach, and accounting for time dependence in home run rates through a hidden Markov model is a novel addition, even though my feeling is that a continuous state-space Markov model may be more promising. I look forward to the continued success and insightful work from this productive group of researchers.

References

Fahrmeir, L. and Tutz, G. (1994). "Dynamic stochastic models for time-dependent ordered paired comparison systems." *Journal of the American Statistical Association*, 89: 1438–1449. 663

Glickman, M. E. (1999). "Parameter estimation in large dynamic paired comparison

experiments." *Applied Statistics*, 48: 377–394. 663

Jensen, S. T., Shirley, K., and Wyner, A. J. (2009). "Bayesball: a Bayesian hierarchical model for evaluating fielding in major league baseball." *Annals of Applied Statistics*, 3: 491–520. 661

Bayesian Analysis (2009) **4**, Number 4, pp. 665–668

Comment on Article by Jensen et al.

Fernando A. Quintana* and Peter Müller †

1 Introduction

We congratulate Shane T. Jensen, Blake McShane and Abraham J. Wyner (henceforth JMW) for a very well written and interesting modeling and analysis of hitting performance for Major League Baseball players. JMW proposed a hierarchical model for data extracted from the Lahman Baseball Database. They model the player/year-specific home run rate using covariate information such as the player's age, home ballpark, and position. The proposed approach successfully strikes a balance of parsimonious assumptions where detail does not matter versus structure where it is important for the underlying decision problem. An interesting feature of the model is the time-dependence that is induced by assuming the existence of a hidden Markov chain that drives the transition of players between "elite" and "non-elite" conditions. In the former case, JMW postulate that the home run rate is increased by a certain position-dependent quantity. The model is used to predict home run totals for the 2006 season, and the results compared to some external methods (MARCEL and PECOTA). The comparison gives some mixed results, with the proposed method rating generally well, compared to their competitors.

2 Some general comments

Inference for the Lahmann baseball data raises a number of practical challenges. The data include records on over 2,000 players, but for many of them there is information for only a couple of years. In many cases there are several years with missing information. As usual in sports data, there is tremendous heterogeneity and unbalance among the experimental units (players). We suspect this is partly the reason why the focus is on predictions for a subset of players. However, this opens the question of whether the model actually provides a good fit for *all* the players. We believe an interesting challenge is to extend the modeling approach to larger subsets, and maybe all players. For such extended inference the model needs to be extended to properly reflect the increased heterogeneity across all players. We propose a possible approach below. Also, the inference focus would shift from prediction to more emphasis on an explanatory model.

Model (2) and the proposed variations, have the interesting feature of incorporating in the home run rates θ_{ij} an explicit dependence on player position k, home ballpark

*Departamento de Estadística, Pontificia Universidad Católica de Chile, Chile, mailto:quintana@mat.puc.cl

†Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX, mailto:pmueller@mdanderson.org

 DOI:10.1214/09-BA424C

b and a smooth position-specific age trajectory, expressed as an hypothesized linear combination in the logit scale. The smooth function of age seems to capture interesting nonlinear features of the home run rates evolution on time, as seen in Figures 3 and 5. One may even venture the existence of an "optimal" age for hitting, and a natural decay in abilities with progressing age. In fact, such conclusions have been reached elsewhere, and even if not the target of this work, it is a nice feature of the analysis that the same kind of findings are uncovered.

The hidden Markov model for "elite" status is the model component that is responsible for introducing dependence across seasons for a given player. The extended model allows for player-specific transition parameters, i.e., individual trajectories for the binary elite indicator variables. Concretely, JMW assume the parameters (ν^i_{00}, ν^i_{11}) controlling these transitions to be a priori independent and Beta-distributed, with conditional independence across players sharing a same position k. These assumptions imply flexibility in the evolution of the $\{E_{ij}\}$ elite indicators, which are well defined regardless of missing data patterns along the sequences of home runs. Looking at the results of the analysis, it is quite remarkable that a large number of players achieve elite status after only one or two major league seasons, as seen in Figure 2. Intuitively one would have expected a peak more likely around 3-5 years. JMW seem to be equally surprised at such findings, when they comment that the sum over years 2 through 11 still represents 75% of the cases considered.

Another consequence of the elite/non-elite model is that the effect on home run rates θ_{ij} is only through a position-specific added term $\alpha_k = \alpha_{k0}(1 - E_{ij}) + \alpha_{k1}E_{ij}$ on the logit scale. While this has the advantage of borrowing strength across players with the same position, it may be not flexible enough to capture highly heterogeneous home run profiles.

3 Extending the proposed approach

The latent elite indicator E_{ij} defines a mixture model for the observed home run totals. The use of E_{ij} is an elegant way to formalize inference about top players. The model balances parsimony with sufficient structure to achieve the desired inference. The authors correctly point out some of the remaining limitations. Perhaps the most important limitation is that the model reduces the heterogeneity of the population of all players to a mixture of only two homogeneous subpopulations. This is particularly of concern in the light of the underlying decision problem. The resulting inference only informs us about the probability of a player being in the elite group. Some evidence for more heterogeneity beyond the mixture of only two subpopulations is seen in Figure 4. The wide separation of the credible intervals suggests scope for intermediate performance groups in the model. The population of players is highly heterogeneous, but not in such a sharply bimodal fashion. It is also interesting to note in the same figure the almost preserved ordering across positions between elite and non-elite groups.

A minor extension of the model could generalize the mixture to a random partition into H subpopulations, which could help closing the gap just pointed out. Each cluster

could have a cluster-specific set of intercepts α_{kh}, $h = 0, \ldots, H-1$ for the logistic regression prior (2) of player-season home run rates θ_{ij}. Like in JMW's model, the intercepts remain ordered $\alpha_{kh} \leq \alpha_{k,h+1}$, $k = 1, \ldots, 9$. This allows us to interpret the clusters labels $h = 0, \ldots, H-1$ as latent player performance.

Formally the model extension would replace (2) by

$$\text{logit}(\theta_{ij}) = \alpha_{ih} + \beta_b + f_k(A_{ij}), \tag{1}$$

where β_b and $f_k(A_{ij})$ are as earlier, and $h = E_{ij}$ is the imputed cluster membership for player i in season j. The prior for $\boldsymbol{\alpha_k} = (\alpha_{kh},\ h = 0, \ldots, H-1)$ is similar to (9), now for the $H-$dimensional vector $\boldsymbol{\alpha_k}$. The prior for the latent cluster membership E_{ij} remains as in (3), extended to transitions between H states. The number of transition parameters ν_{rs} remains unchanged with prior probability ν_{01} for upgrades in elite level, prior probability ν_{10} for downgrades and ν_{00} for the probability of remaining in state $E_{ij} = 0$ and ν_{11} for the probability of remaining in a performance state $E > 0$. Like in (7) the transition probabilities are position-specific.

The number of states H would itself be treated as unknown, with a geometric prior $p(H) = (1-p)^{H-1}p$ and a hyperparameter p. The only additional step in the MCMC implementation is a transition probability to change H. We consider two transitions, "birth" of an additional performance level by splitting an existing level h into two new levels and the reverse "death" move. This could be implemented as a reversible jump move.

The generalized model defines a random partition of the player-years (ij) into performance clusters $h = 0, \ldots, H-1$. The unique features of this random partition model would be the ordering of the clusters and the dependence across j. Both features are naturally accommodated by the outlined model-based clustering. We see it as an interesting and challenging application of model-based clustering. In contrast to much of the of clustering models in the recent Bayesian literature, the use of standard clustering models such as the ubiquitous Polya urn would be inappropriate. The Polya urn model does not naturally allow the desired ordering of cluster-specific parameters and time-dependence of cluster membership indicators.

4 Final words

We realize the above proposal can be extended/modified in many different ways, the main point being the possibility of improving on the analysis and model proposed by JMW. Our aim here was not to criticize the model but to help improve it. We indeed think the hidden Markov component is a very nice feature, which combined with a flexible extension, could motivate further analysis of the data under a more general framework.

Acknowledgments

Fernando Quintana was partially funded by Fondecyt grant 1060729.

Bayesian Analysis (2009) **4**, Number 4, pp. 669–674

Rejoinder

Shane T. Jensen*, Blakeley B. McShane† and Abraham J. Wyner‡

We thank each discussant for his insightful comments and suggestions for improvement. We are pleased by the positive reception of our current endeavor towards model-based prediction of hitting performance. It is our belief that academic statisticians can serve a leadership role in the transition of quantitative analysis of baseball from simple tabulations to sophisticated model-based approaches.

1 Alternative Models for Latent Variables

A clear theme of this discussion is the flexibility of the Bayesian hierarchical framework as a principled means for prediction in this application. Of course, the other side of that coin is that our model can always be improved by more sophisticated extensions. The discussants offer several great suggestions for improvements to our methodology. A first step in this effort is suggested by multiple discussants: extensions of the latent "elite" mixture model. These proposals are great directions for future research, and we briefly discuss the prospects of each below.

Albert & Birnbaum question our employment of a latent mixture model, citing the fact that these mixture components are not self-evident from the raw home-run rate distributions. However, they also note the presence of skewness and outliers. We argue that latent mixture models are a common strategy for addressing skewness and outliers. In fact, our original motivation for a latent mixture model was the observation that hitters with consistently high home run rates were over-shrunk in a model that did not allow for subpopulations of extreme home run performance.

Both Quintana & Müller and Glickman discuss the limitation of our mixture model to two latent states. In our original analysis, we experimented with the addition of a third latent state which was intended to capture players that showed inferior performance relative to their position. However, the estimated models that included this third state did not show any greater predictive power than the two-state model.

Quintana & Müller suggest a more comprehensive amelioration of our mixture model: allowing the number of latent states to be unknown and estimated. Certainly, this proposal is the most natural extension of the current approach and would help address the concerns raised by the discussants about the imposed "elite" vs. "non-elite" framework. The hurdle would be implementation of this more complicated model, as the reversible-jump approach proposed by Quintana & Muller could be complicated in practice.

*Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:stjensen@wharton.upenn.edu

†Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:mcshaneb@wharton.upenn.edu

‡Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, mailto:ajw@wharton.upenn.edu

 DOI:10.1214/09-BA424REJ

Glickman proposes a model extension that is further afield. Instead of a discrete state space model, he proposes a latent state that evolves continuously in an autoregressive fashion. In our opinion, this continuous state space model would perform well for players with a long and consistent history of performance. However, we are skeptical there would be enough autocorrelated signal for younger players with very little personal history. For these cases with sparser information, we believe our simpler model is better able to pool information between players.

We have a similar concern about Albert & Birnbaum's proposal to fit random effects for each player. We concede that players (at the same position) can have very different trajectories, as illustrated by their comparison of Mickey Mantle and Hank Aaron. However, although there is enough information to model players with long careers in this way, we suspect that these random effects would be too variable for players who have only played a few seasons. For such players, the enforced shrinkage of our model is beneficial.

Furthermore, while the selection of Mantle and Aaron nicely illustrate the benefits of modeling trajectories individually, it also illustrates some of the pitfalls. Though Mantle and Aaron were both towering sluggers of their era, we contend that both players are unusually deviant from what is generally observed and their careers represent extreme points in the space of individual trajectories. Mantle suffered a precipitous decline due to debilitating injury while Aaron had an almost miraculously steady and lengthy career.

Thus, we are not sure it is a criticism to point out that we would have failed to predict Aaron's unusual performance into his forties or Mantle's steep early decline, unaided by health information. For the purposes of prediction, discounting unusual individual career trajectories and being guided mainly by position is a sound strategy, and we remind the reader that center fielders like Mantle are more likely to experience sharp declines in production than corner outfielders like Aaron. That said, the random effects framework is a great idea, and we are currently investigating extending our model to allow more flexible trajectories within each position.

There are of course many other generalizations and improvements not raised by discussants which we will consider in future work. Most promising is the extension of the usual first order Markov model to higher order or even variable order. This direction has the potential to more accurately model an individual player's trajectory.

2 Position and Other Potential Covariates

Beyond the latent mixture model, the discussants provide several suggestions for additional data and/or covariates that could further improve our predictions. Specifically, Albert & Birnbaum suggest the retrosheet database which provides more detailed within-season information for each player. We agree that the additional detail within the retrosheet database could improve our modeling efforts. One immediate advance, as proposed by Albert & Birnbaum, would be to divide each hitter's season into home

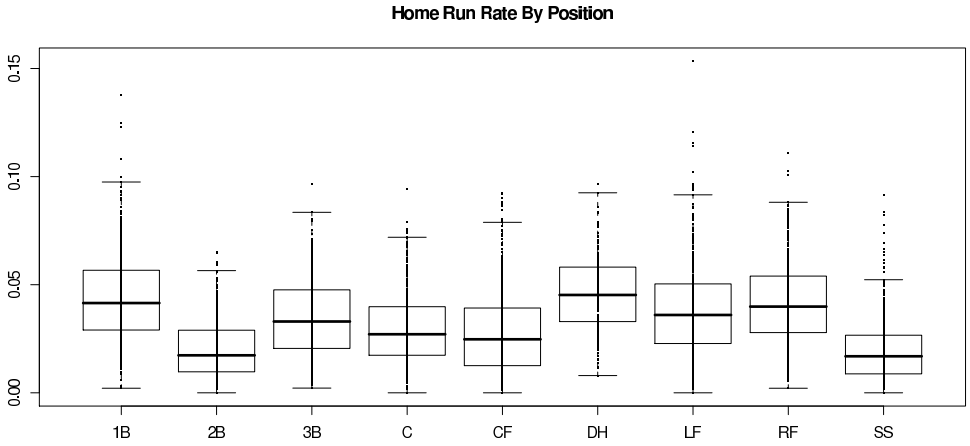


Figure 1: Boxplots of empirical home run rates by position. Each point gives HR/AB for a given player-season for all player-seasons with 300 or more at-bats from 1990-2006.

Table 1: Analysis of Variance Table

Source	DF	Sum Sq.	Mean Sq.	F Ratio	Prob > F
Position	8	0.31486	0.03936	120.6345	<2e-16
Year	16	0.05922	0.00370	11.3446	<2e-16
Age	25	0.00670	0.00027	0.8214	0.7178
Residuals	3801	1.24009	0.00033		

versus away games, thus enabling the estimation of true ballpark effects. We would favor estimation of ballpark effects in this way rather than the use of external park factors, which is also proposed by Albert & Birnbaum. In our experience, external park factors are highly inconsistent from year to year and do not seem to contain much signal except in some extreme cases (e.g., Coors Field or Citizens Bank Park).

Albert & Birnbaum question the use of position as a covariate in our model, claiming that it is not immediately evident what information is being added by position. They are correct to assert that there is heterogeneity of home run talent within each position, but there is large variation in home run rates across position as can be seen in Figure 1. In fact, we perform an analysis of the variance of home run rates by the nine positions, seventeen years, and twenty-six ages in our dataset in Table 1. Position accounts for 20% of the total variation in home run rates, far more than any other factor.

These results suggest that position is a very informative covariate for home run ability. In our view, position serves as a proxy variable for several player characteristics, such as body type and speed, that cannot be directly observed from the data. Scouts and

managers incorporate many of these unobserved variables into their personnel decisions in terms of where to place players. By assigning a particular player to traditional power positions such as first base, managers are adding information about that player's propensity to hit home runs. We think this information is especially important for younger players who have less performance history upon which to base predictions.

Albert & Birnbaum also point out that our model does not address major shifts in hitting performance between different eras in baseball. We do not argue the point, as it was not the goal of our paper (though we note that Table 1 shows that the year factor accounts for a modest 3.6% of the variance in home run rates). Our priority is the prediction of future hitting performance, which motivated our focus on the current era. The comparison of hitting performance in different eras is also an interesting question, and has been addressed in the past with sophisticated Bayesian approaches (Berry et al. 1999).

We did investigate, somewhat indirectly, the possible effects of different eras on our predictions. We fit our full model on a larger dataset consisting of all seasons from 1970 to 2005, in addition to our presented analysis based on seasons from 1990 to 2005. We saw very little difference in the predictions between these two analyses, suggesting that any large-scale changes in hitting dynamics over the past forty years do not have a major impact on future hitting predictions.

Albert & Birnbaum also suggest using at-bats as a covariate for the modeling of home run rates. This is a good suggestion and we have investigated the modeling of at-bats as a means for improving the prediction of hitting totals. However, we need to correct one statement made by Albert & Birnbaum: we do not assume that each player's 2006 at-bats are the same as the at-bats in the previous season. Rather, we scale the predictions of hitting rates from our model (and the two external methods) by the actual 2006 at-bat totals in our comparisons.

3 Focus on Prediction

Glickman suggests that home run totals may not be the most interesting outcome to people in baseball. We certainly agree that home runs are not the best measure of overall hitting performance, and we emphasize that our methodology can be adapted to any other hitting event. Home runs were chosen for illustration since we believe that most readers have a good intuition about the scale and variation of home run totals. We also have experimented with a multinomial extension of our procedure that would model each hitting outcome (i.e., singles, doubles, etc.) simultaneously, and this remains an area of future research.

More generally, Albert & Birnbaum call for greater focus on model interpretation. Despite our emphasis on prediction, there are elements of our model that are interesting in their own right. The position-specific aging curves provide an interesting contrast in the aging process between players at these different positions. Our "elite" versus "non-elite" indicators also provide a means for separating out consistently over-performing

players relative to their position.

Quintana & Müller also inquire about the predictive power of our model for *all* players, not just the subset of players examined in our analysis. Our primary motivation was to have a set of common players for comparison with the external methods. However, we concede that the players excluded from our analysis probably represent an even tougher challenge for prediction. Albert & Birnbaum also suggest that extra insight would be gained from a case-by-case exploration and comparison of our predictions. To this end, we have made available the entire set of our predictions for the 2006 season at the following website: http://stat.wharton.upenn.edu/~stjensen/research/predictions.2006.xlsx

References

Berry, S. M., Reese, S., and Larkey, P. D. (1999). "Bridging Different Eras in Sports." *Journal of the American Statistical Association*, 94: 661–686. 672