The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Predicting Home Run Production in Major League Baseball Using a Bayesian Semiparametric Model

Gilbert W. Fellingham & Jared D. Fisher

To cite this article: Gilbert W. Fellingham & Jared D. Fisher (2017): Predicting Home Run Production in Major League Baseball Using a Bayesian Semiparametric Model, The American Statistician, DOI: 10.1080/00031305.2017.1401959

To link to this article: https://doi.org/10.1080/00031305.2017.1401959

Predicting Home Run Production in Major League Baseball Using a Bayesian Semiparametric Model

Gilbert W. Fellingham*
Department of Statistics, Brigham Young University
and
Jared D. Fisher
Department of Information, Risk, and Operations Management, University of Texas

October 21, 2017

Abstract

This paper attempts to predict home run hitting performance of Major League Baseball players using a Bayesian semiparametric model. Following Berry, Reese and Larkey (1999) we include in the model effects for era of birth, season of play, and home ball park. We estimate performance curves for each player using orthonormal quartic polynomials. We use a Dirichlet process prior on the unknown distribution for the coefficients of the polynomials, and parametric priors for the other effects. Dirichlet process priors are useful in prediction for two reasons: (1) an increased probability of obtaining more precise prediction comes with the increased flexibility of the prior specification (Fellingham, Kottas and Hartman, 2015), and (2) the clustering inherent in the Dirichlet process provides the means to share information across players (Dahl and Newton, 2007). Data from 1871 to 2008 were used to fit the model. Data from 2009 to 2016 were used to test the predictive ability of the model. A parametric model was also fit to compare the predictive performance of the models. We used what we called 'pure performance' curves to predict future performance for 22 players. The nonparametric method provided superior predictive performance.

Keywords: Performance curves, Dirichlet process prior, orthonormal polynomials, prediction

*The authors gratefully acknowledge Kris Young for help with the data files.

1 Introduction

In this paper we evaluate Major League Baseball players' career home run hitting performance using a Bayesian semiparametric model. Performance curves based on player age were fit to players' home run hitting data, and then, for a subset of the data, extended to examine future performance. To account for effects of playing in different circumstances, we added effects for era of birth, season of play and home ballpark to the model following the paper by Berry et al. (1999). Berry et al. (1999) used their model to facilitate across era player comparisons. While our model has the same capability, we focus on borrowing strength across players to predict future performance in a subset of contemporary players. We use a binomial likelihood and model the log odds of hitting a home run, using performance curves based on orthonormal quartic polynomials. By using a Dirichlet process prior for the unknown distributions of the coefficients of these curves, we introduce the nonparametric Bayesian framework. Parametric normal distributions were used for the priors of the season, era, and ballpark effects. We also fit a fully parametric Bayesian model to compare predictive performance.

The Dirichlet process (Ferguson, 1973) is a nonparametric Bayesian method that allows prior distributions to be more flexible than traditional parametric Bayesian methods (see also Sethuraman (1994), Müller and Rosner (1998), Carlin and Louis (2009), and Frigyik, Kapila and Gupta (2010)). Nonparametric Bayesian methods also induce a clustering on posterior distributions with nonzero probability, thus adding insight to posterior inference (Ghahramani, 2013).

We hypothesized that the different clusters induced using the nonparametric prior would contain players who developed similarly throughout their careers, despite differences in era of birth, season of play, or home ballpark. This methodology then would give us the ability to identify what we call 'pure performance' by examining performance curves after ballpark, season, and era effects are removed. By assuming that players who have yet to retire would follow the pattern of other players in their performance cluster, we could then make predictions about their future home run hitting performance.

2 Previous Work

Albright (1993) looked at hitting streaks in baseball using a handful of techniques to determine

that, overall, the occurrence of hitting streaks appears to be random. One model used is a logistic regression with covariates for the right/left handedness of the pitcher, the pitcher's ERA, the difference of the number of runs scored in the current game, inning of the game, and other situational variables, as well as a custom created measure of the player's recent success.

Schell (1999) compared players' batting averages with the assumption that batting averages change across eras, and thus are only directly comparable to other contemporaries. Using a set of adjustments to batting average, Tony Gwynn was identified as the best hitter of all time.

Quintana, Müller, Rosner and Munsell (2008) used Bayesian random effects to model an at bat for a given player in a given season as a success (hit, walk, sacrifice) or a failure. All of the at bats for a given player in a given season were modeled with a Markov model, the transition probabilities of which are found with autologistic regression using lagged responses and 13 covariates, including pitcher's ERA and indicator variables for turf versus grass, presence of runners on base, if in the 7th inning or later, etc. The coefficients of the lagged responses were drawn from a Dirichlet process mixture of normal models. They concluded that even with a sophisticated model there is substantial variability within each player and for players across seasons.

Jensen, McShane and Wyner (2009) used a Bayesian logistic regression framework to model Major League Baseball players' home run rate as a function of home ballpark, position and age. "Position" constituted position-specific intercepts, and "age" was modeled using B-splines, with distinct coefficients for each position. They also estimated a latent variable for a player being considered an "elite" player in a given season, and the transitions to and from elite status were determined using a hidden Markov model. If elite, there was an effect added to the position-specific intercepts. This means nine performance curves were estimated, each with two possible intercepts, one representing typical performance and the other representing elite performance. They concluded that their approach does particularly well predicting the performance of young players.

Berry et al. (1999) used hierarchical random curves in a Bayesian setting to model baseball, ice hockey and golf athletes and compared them to other players of their sports from different eras. This was done by including parameters for the eras in the model, where eras were defined as the different decades in which players were born. To model Major League Baseball play-

ers hitting ability, both for home runs and for average, they used parametric effects for decade of birth, season of play and home ballpark. They concluded that the best home run hitter per at bat is Mark McGwire. The work we present mirrors the Berry et al. (1999) paper using a semiparametric Bayesian model to estimate the performance curves.

Fellingham et al. (2015) demonstrate the effectiveness of Bayesian nonparametrics in prediction. They compare the parametric model to the nonparametric model and show that the added flexibility of the DP prior increases predictive performance using a number of metrics.

3 Methodology

3.1 The Dataset

The data used to build the model consist of all Major League Baseball hitting data from 1871 to 2008, from the Lahman Baseball Database (Lahman, 2016). We used the data from 1871 to 2008 to build the model, and data from 2009 to 2016 to test the predictive ability of the model. For our purposes, the data are limited to players who played a minimum of six seasons, with at least 50 at bats per season. All player data is cutoff after 45 years of age (to eliminate outlier data points from special "comeback" appearances). The final dataset included the names, birth year, and number of seasons played for 3735 players. Further, for each season and each player, we also have:

- total home runs,
- total at bats,
- age,
- team played for, and
- season of play.

Note that the player's age for a season is calculated as the difference between year of play and year of birth, hence the player's age at the end of the year played. Based on these data, two new variables were created. The ballpark variable is an integer from 1 to 222 which indexes the home

ballpark of the player that season. In the event that a player played in multiple home ballparks in a single season (because the player changed teams or his team changed home ballparks) the ballpark in which the player played the most home games was chosen. Since few players were born in the first three decades, and since the game did not change as much in its formative years, we placed all players born in these decades in one group. Thus, the decade variable is an integer from 1 to 14 that indexes the decade of birth, where 1 represents the 1830-1850's and 14 represents the 1980's.

3.2 Nonparametric Model description

We assumed the number of home runs in a season is distributed as a binomial random variable. That is,

$$h_{ij}|\pi_{ij} \sim \text{Binomial}(ab_{ij}, \pi_{ij}),$$

where, for player i in his j^{th}th season, h_{ij} is the number of home runs hit, ab_{ij} is the number of "at bats", and π_{ij} is the probability of hitting a home run in a given at bat. We used a logistic regression model as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_i + \theta_i + \delta_j + \xi_p,$$

where $\boldsymbol{\beta}_i$ is a vector of coefficients for an orthonormal quartic polynomial vector $\boldsymbol{x}_{ij}$ based on the player's age in his j^{th} season. The $\boldsymbol{\beta}_i$ vector contains coefficients for linear through quartic terms. The intercept term is subsumed in the other parameters. Fourth-order (quartic) polynomials provide a flexible fit for the curves. We settled on quartic orthonormal polynomials after examining other possible forms of the performance curves. In preliminary tests on small subsets of the data, quartics provided better performance than fits using piecewise linear, quadratic and cubic polynomials. Other possibilities for the performance curves, such as Gaussian processes, were not examined.

We assumed a Dirichlet process (DP) prior on the unknown distribution for the coefficients of the polynomial. The Dirichlet process (Ferguson, 1973) is a stochastic process with sample paths that can be interpreted as distributions G on a sample space Ω. Thus, with a Dirichlet process prior, the posterior can flexibly model data without parametric constraints. A DP is defined in terms of two parameters: a parametric 'base' or 'centering' distribution G_0 on Ω, and a positive

scalar 'concentration' parameter α. The larger the value of α, the closer a realization of the process is to G_0. The priors for the coefficients were as follows:

$$\begin{aligned} \boldsymbol{\beta}_{1-4,i} \mid G_{1-4} &\overset{\text{iid}}{\sim} G_{1-4},\ i = 1, ..., n_{\text{players}} \\ G_{1-4} &\overset{\text{ind.}}{\sim} \text{DP}(\alpha, G_{1-4,0}) \\ \alpha &\sim \text{Gamma}(2, .2) \\ G_{1-4,0} &\sim \text{MVN}(\boldsymbol{\mu}_{\beta_{1-4}}, \Sigma) \end{aligned}$$

where

$$\boldsymbol{\mu}_{\beta_{1-4}} = (0.25, -3.0, 0.0, 0.0)'$$

and

$$\Sigma = \begin{bmatrix} 30 & -24 & 20 & -20 \\ -24 & 30 & -20 & 20 \\ 20 & -20 & 20 & -16 \\ -20 & 20 & -16 & 20 \end{bmatrix}.$$

The parameterization we are using for the Gamma yields an expected value of 10 for the prior for α. An α of 10 yields approximately 60 clusters for 3735 players (Escobar and West, 1995), which seemed to us to be a reasonable number. We chose μ_{β_1} to be positive in order to model growth at the beginning of a career, and μ_{β_2} to be negative to model the expected peak and decline later in the career. μ_{β_3} and μ_{β_4} were set to zero because we felt cubic and quartic effects would generally be small. The values in Σ are relatively large and reflect our uncertainty about the coefficients. However, we did expect the linear and quadratic coefficients would have a fairly large negative correlation, as would the linear and quartic coefficients, while the linear and cubic coefficients would be positively correlated. Also, we expected the quadratic coefficients would be negatively correlated with the cubic coefficients, and the cubic coefficients would be negatively correlated with the quartic coefficients. Those expectations are reflected in the patterns of positive and negative signs seen in Σ.

Following Berry et al. (1999), we defined parametric effects for the season of play, era of play, and home ballparks.

- θ_i - Effect of the decade of birth for a player, i $\in \{1, \ldots, 3735\}$

- δ_{j} - Effect of the season of play, $\mathrm{j} \in \{1,\ldots,138\}$
- ξ_{p} - Effect of the home ballpark, $\mathrm{p} \in (1,\ldots,222)$.

We assumed a normal prior for the season and ballpark effects.

$$\begin{aligned}\delta_{\mathrm{j}} &\overset{\text{iid}}{\sim} N(-2.5, \text{var} = 4), \mathrm{j} \in \{1,...,138\} \\ \xi_{\mathrm{p}} &\overset{\text{iid}}{\sim} N(-2.5, \text{var} = 4), \mathrm{p} \in \{1,...,222\}.\end{aligned}$$

The prior specification differs for the era effects in that we specify a hierarchical prior, where d is an index that references a decade.

$$\begin{aligned}\theta_i &\overset{\text{iid}}{\sim} N\left(\mu_{d_i}, \sigma^2_{d_i}\right),\ d_i \in \{1,\ldots,14\}, i \in \{1,\ldots,3735\} \\ \mu_{d_i} &\overset{\text{iid}}{\sim} N(-2.5, \text{var} = 1),\ d_i \in \{1,\ldots,14\} \\ \sigma^2_{d_i} &\overset{\text{iid}}{\sim} IG(3,3),\ d_i \in \{1,\ldots,14\}\end{aligned}$$

We centered decade, season and home ballpark effects at -2.5 since the probability of hitting a home run is generally small. In the data we modeled (MLB data up through 2008) 2.03% of at bats resulted in home runs. Before 1920, only 0.58% of at bats resulted in home runs. A coefficient of -2.5 for all three effects yields a probability of roughly 0.0006. Thus, home run hitting ability in the early years would be due to individual performance only. The variances were chosen to reflect our belief that we did not expect the coefficients to vary by more than ± 6. It is our experience that noninformative priors on variances can be problematic. Since we didn't expect θ_i to vary by more than ± 6, and IG(3,3) (E[IG(3,3)] = 3/(3-1) = 1.5) gave fairly good coverage of a variance up to 4, we felt comfortable with this prior.

3.3 Parametric Model description

The parametric model mirrored the nonparametric model except for the distributions of the $\beta_{1-4,i}$. A hierarchical model was now assumed where:

$$
\begin{aligned}
\boldsymbol{\beta}_{1-4,i} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), i = 1, ..., n_{\text{players}} \\
\mu_1 &\sim N(0.5, 16) \\
\mu_2 &\sim N(-0.5, 16) \\
\mu_3 &\sim N(0.8, 9) \\
\mu_3 &\sim N(0.1, 9) \\
\sigma_1^2 &\sim \text{Gamma}(15, .5) \\
\sigma_2^2 &\sim \text{Gamma}(15, .5) \\
\sigma_3^2 &\sim \text{Gamma}(10, .5) \\
\sigma_4^2 &\sim \text{Gamma}(10, .5)
\end{aligned}
$$

3.4 Computation

3.4.1 Nonparametric Model

We adopt a Bayesian approach, and as such, make inference from the posterior distributions of the parameters in question. In the next sections, we discuss details of generating the posterior distributions. Code for the sampling process was written in FORTRAN.

3.4.2 Drawing Samples of the Parametric Effects

When sampling from the distributions of the nonhierarchical parametric effects, the first parameter of each of these sequences was set equal to the mean of the prior distribution as a baseline, that is $\delta_1 = \xi_1 = -2.5$. Samples of all the other parametric effects as well as the hyper parameters of the era effects were drawn using the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953), Hastings (1970)).

Since we anticipated the era effects and the season effects would be correlated, their candidates were drawn from a series of multivariate normal distributions, by dividing the seasons into 10 year blocks, though the last (14th) block only has 7 years instead of 10. Thus,

$$
\begin{aligned}
\mu_{d_1}, \delta_{2-11} &\sim MVN(\boldsymbol{\mu}, \Sigma_1) \\
\mu_d, \delta_{[10(k-3)+2]-[10(k-2)+1]} &\sim MVN(\boldsymbol{\mu}, \Sigma_d), d \in \{2, ..., 13\} \\
\mu_{d_{14}}, \delta_{132-138} &\sim MVN(\boldsymbol{\mu}, \Sigma_{14}),
\end{aligned}
$$

where $\boldsymbol{\mu}$ is a vector of the previous iteration's values of μ_d and δ. Once the candidates were drawn, the candidate draws were then accepted or rejected individually.

The candidate values for the ξ_p, θ_i, and σ_d^2 were drawn independently from normal distributions centered at the previous iteration value, with standard deviation σ_c. Here, each of the σ_c were initialized at 1. Then after every 1000 MCMC iterations, each σ_c was automatically adjusted. If the acceptance rate for the parameters for the previous 1000 iterations was less than 20%, then σ_c was set to $.5\sigma_c$. If the acceptance rate was greater than 35%, then σ_c was set to $1.5\sigma_c$. The adjustments to the σ_c were only made during the burn in, and once the burn was completed, the σ_c values did not change.

3.4.3 Drawing Samples for the Polynomial Coefficients

Algorithm 6 from Neal (2000) was used to draw MCMC samples from the Dirichlet process. All $\boldsymbol{\beta}_i$, $i = 1, ..., n_{\text{players}}$, were initialized at the mean of the centering distribution. Then for all iterations and for all players, a candidate value, $\boldsymbol{\beta}_i^*$ was drawn from the distribution

$$\frac{1}{\alpha + n_{\text{players}} - 1} \sum_{j \neq i} \delta(\boldsymbol{\beta}_j) + \frac{\alpha}{\alpha + n_{\text{players}} - 1} G_0$$

where $\delta(\boldsymbol{\beta}_j)$ is a point mass at the single vector $\boldsymbol{\beta}_j$. This candidate was then accepted with probability

$$\min \left[1, \frac{F(y_i, \boldsymbol{\beta}_i^*)}{F(y_i, \boldsymbol{\beta}_i^t)} \right],$$

where $\boldsymbol{\beta}_i^t$ is the current value of $\boldsymbol{\beta}_i$, and F is the binomial likelihood as previously specified. Note that this sampler was run 50 times for each player in each iteration to aid convergence. For more on computation of the Dirichlet process, see Blackwell and MacQueen (1973), MacEachern (1998), MacEachern and Müller (1998).

3.4.4 MCMC Draws

To obtain draws from the posterior distribution, 100,000 iterations were first used as a burn in. Twenty random starting points were then selected from the posterior distributions of these draws and used as starting points for twenty parallel MCMC samplers (Gelman and Rubin, 1992). Each sampler was further burned in 100,000 iterations and followed by another 500,000 iterations with every 500^{th} draw saved as a sample from the posterior. This resulted in 20,000 draws from the

posterior distribution, with the 1000 draws from each sampler appended together. Figure 1 shows trace plots for a few parameters. Clearly, mixing was slow. Since multiple chains were used, we could compute the $\hat{R}$ statistic (Gelman and Rubin (1992)) for the parameters. Since we will be predicting for 22 players, we computed the $\hat{R}$ statistic for the 88 β's for these players. All the $\hat{R}$ statistics for the β's were less than 1.02. All the $\hat{R}$ statistics for the ξ's were less than 1.06. All the $\hat{R}$ statistics for the δ's were less than 1.36. All the $\hat{R}$ statistics for the θ's were less than 1.25, except 3 less than 2.11, and 1 at 3.58. All the $\hat{R}$ statistics for the μ_{d_i}'s and $\sigma^2_{d_i}$ were less than 1.23.

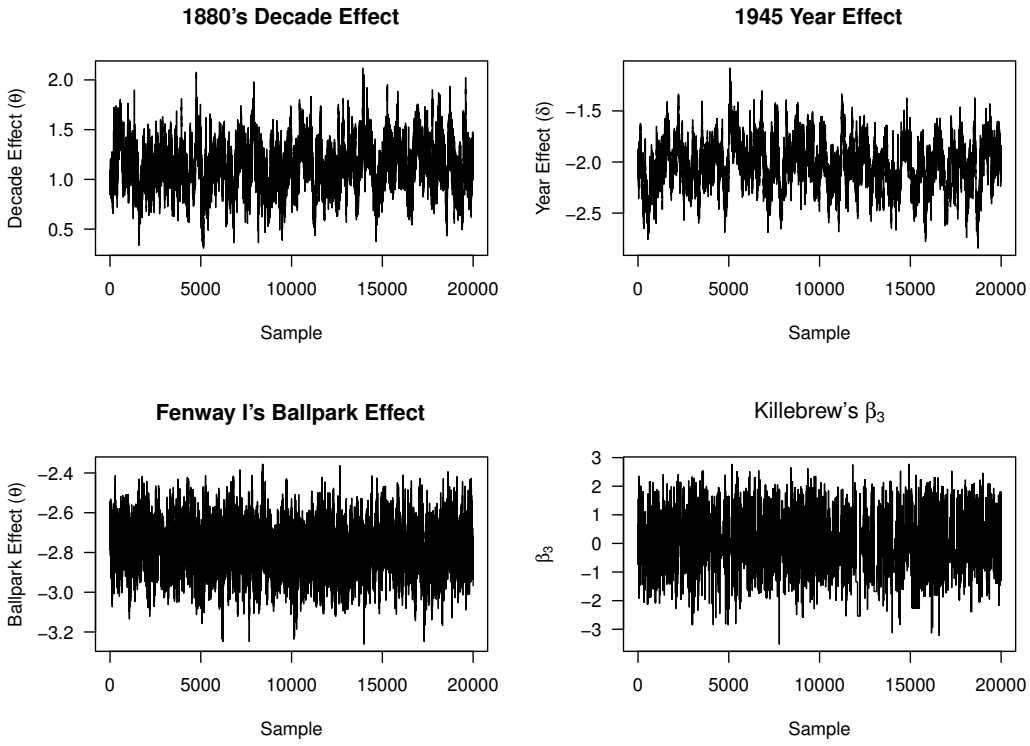

Figure 1: Trace plots of the 20,000 draws from the posterior distributions of the four types of parameters of interest. Note that each of the 1000 draws from the twenty parallel runs were appended together, such that the first 1000 draws are from the first sampler, and the second 1000 from the next, and so on. As the polynomial coefficients have a nonparametric prior, we expect to see some 'sticky' parts in the trace plot of Killebrew's β_3.

3.4.5 Hierarchical Model

The code for this model was written in and compiled using Nimble (NIMBLE Development Team (2017)). We used 200,000 iterations keeping every 20^{th}. The first 2000 iterations (after thinning) were discarded. Since we only ran one chain, we report the Raftery-Lewis diagnostic (Raftery and Lewis (1992) for the 88 β's used in prediction. All the Raftery-Lewis diagnostics were under 3.0, and 77 of the 88 (87.5%) were under 2.0.

3.5 Inference

The performance curves were used to identify players that developed similarly in their careers, despite differences in era and team, using both the nonparametric and the hierarchical models. The nonparametric model was used to identify year-to-year, era-to-era, and ballpark-to-ballpark differences in home run hitting. Using mean performance curves, we examined the predictive ability of both models on eight seasons (2009-2016) played since the data used to build the model were produced.

4 Results

4.1 Performance Curves

Figure 2 shows the posterior performance curves generated using the nonparametric model, 95% credible intervals on those curves, and the original data points, for four players who had prolific careers hitting home runs. The outlier point from Barry Bonds career is the 2001 season, where he set the single season home run record with 73 home runs. As the underlying quartic polynomial is by nature a smooth function, the jagged nature of the curves shown is due to the season and ballpark effects, as they can change year to year. Note the general shape of Hank Aaron's performance over his career. It was examination of his performance that led us to consider orthonormal quartic polynomials as potential performance curves.

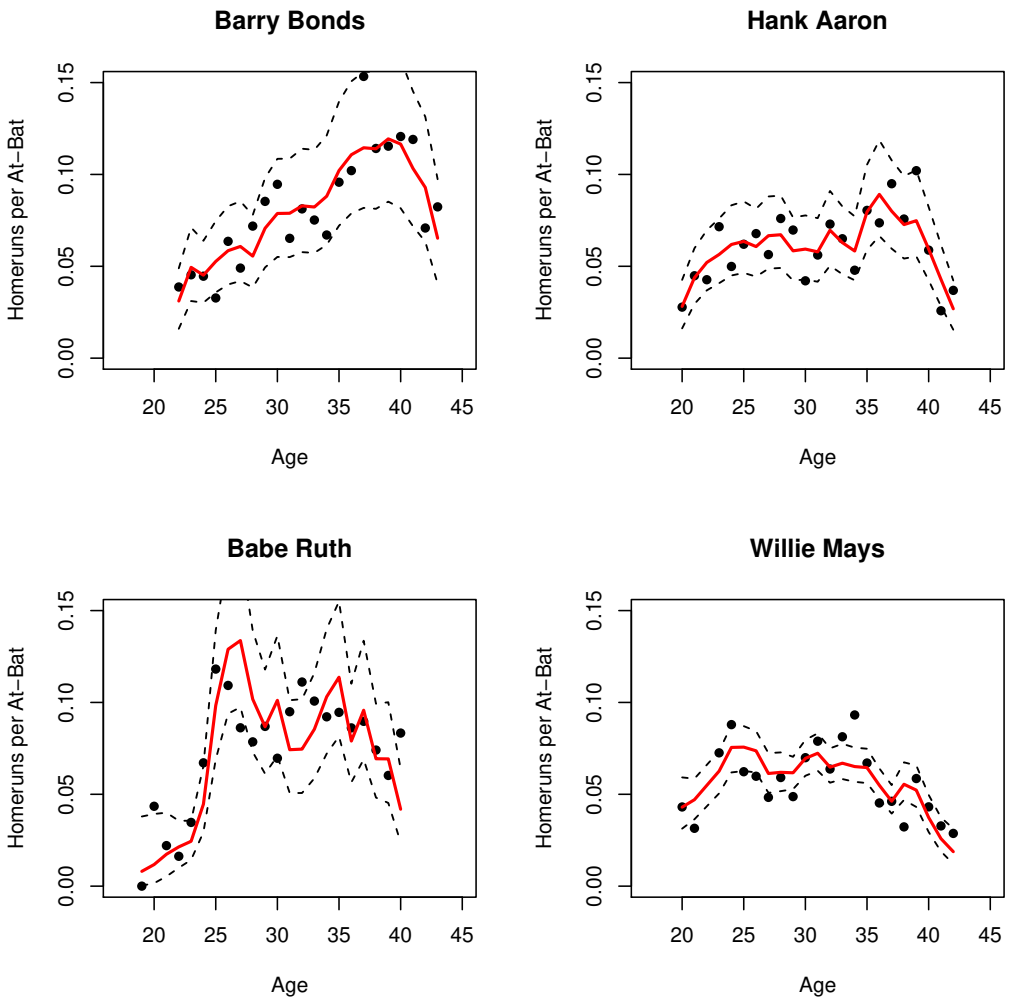

Figure 2: Mean performance curves, with effects, of the top 4 all-time home run hitters. The solid line marks the posterior mean of the curve, while the dashed lines show the 95% credible interval about the mean curve.

4.2 Era, Season, and Ballpark Effects

As noted previously, all estimates for these parameters come from the nonparametric model. The posterior distributions of the era effects, defined by the decade of birth, are summarized in Figure 3. Again, note that as a baseline, players born in the 1830's and 1840's were grouped in with those born in the 1850's. Babe Ruth, born in 1895, seemed to be the end of one era and the beginning of the next, based on the large increase in home run hitting (increase in μ_d) between Ruth's era, the 1890's, and the following era. The "Golden Age" for home run hitting seems to

be for players born in the 1920's. However, one should keep in mind the interaction between era and season effects, and note that the season effects for later eras are clearly increasing (see Figure 4). Players born in this decade generally had their most productive years following World War II. The first African American player in the Major Leagues, Jackie Robinson, was born in 1919. We found that African American players began to dominate home run hitting with the birth decade of the 1930's.

Figure 4 shows the size of the season effects, with significant rule changes and events marked to show their impact. Again, note that, as a baseline, the 1871 season's effect was set equal to -2.5, that is $\delta_1 = -2.5$. Larger effect sizes for a given season mean that home runs were hit more frequently that season. These season effects show that making the baseball's center cork, outlawing the spitball, lowering the mound, shrinking the strike zone and league expansion all increased the frequency of home runs. However, restricting the mound height in 1904 was not followed by an increase in home run hitting, but a decrease in the effect size. World War II also brought a drop in home run hitting.

Figure 5 shows distributions of ballpark effects for 30 modern ballparks with 10 of the ballparks noted by name, either because they were harder of easier to hit home runs in, or because they are famous for some other reason. Note the variances increase to the right of the plot, as these ballparks are newer and have less data to estimate the effects.

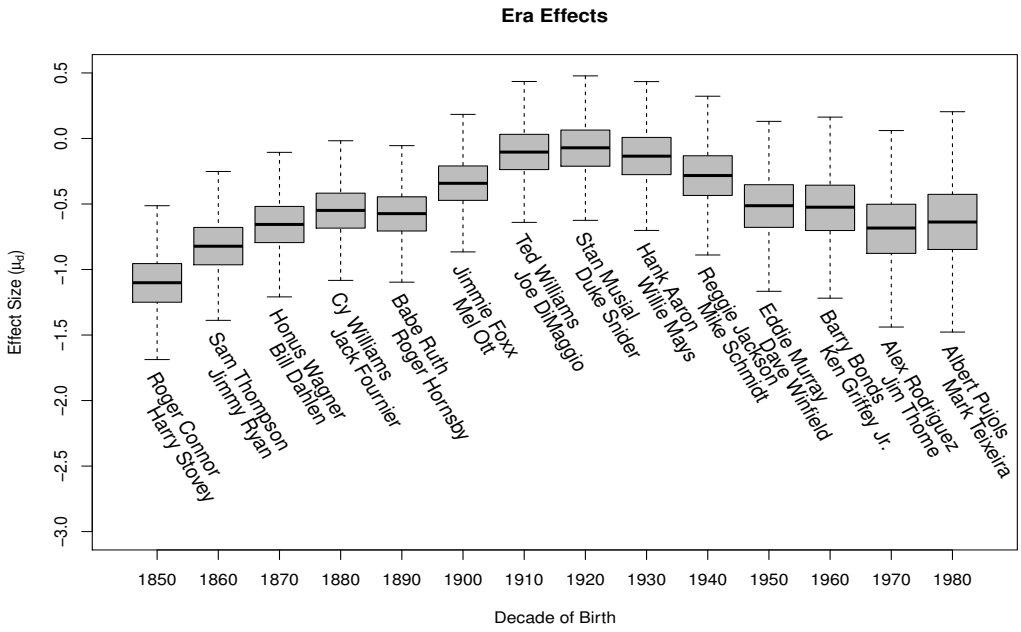

Figure 3: Boxplots of posterior distributions of decade or era effects, with two notable players from that birth decade.

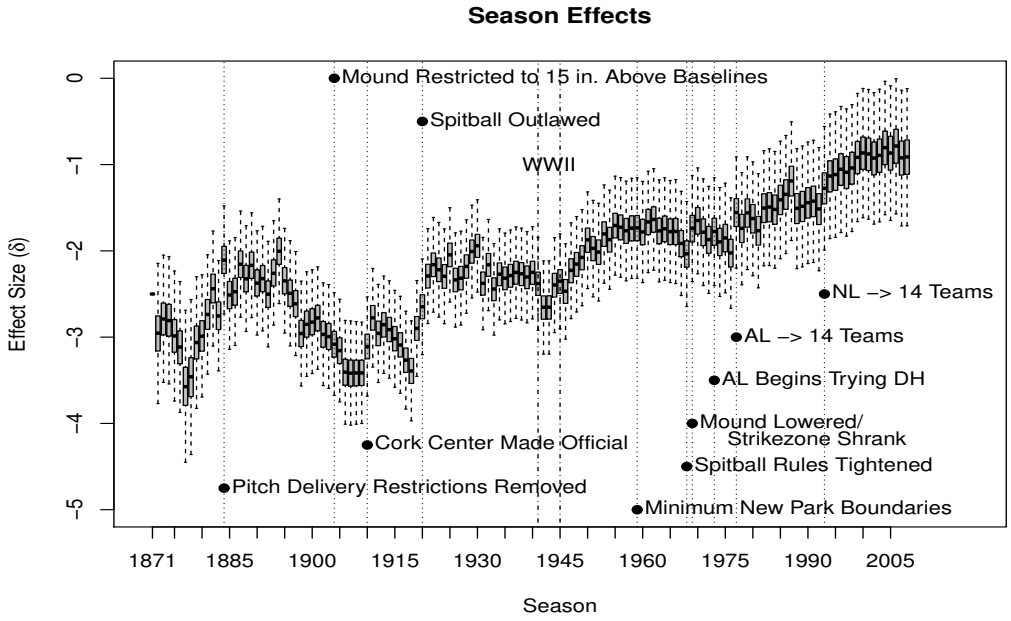

Figure 4: Boxplots of posterior distributions of season effects with some seasons highlighted due to rule changes that occurred prior to those seasons.

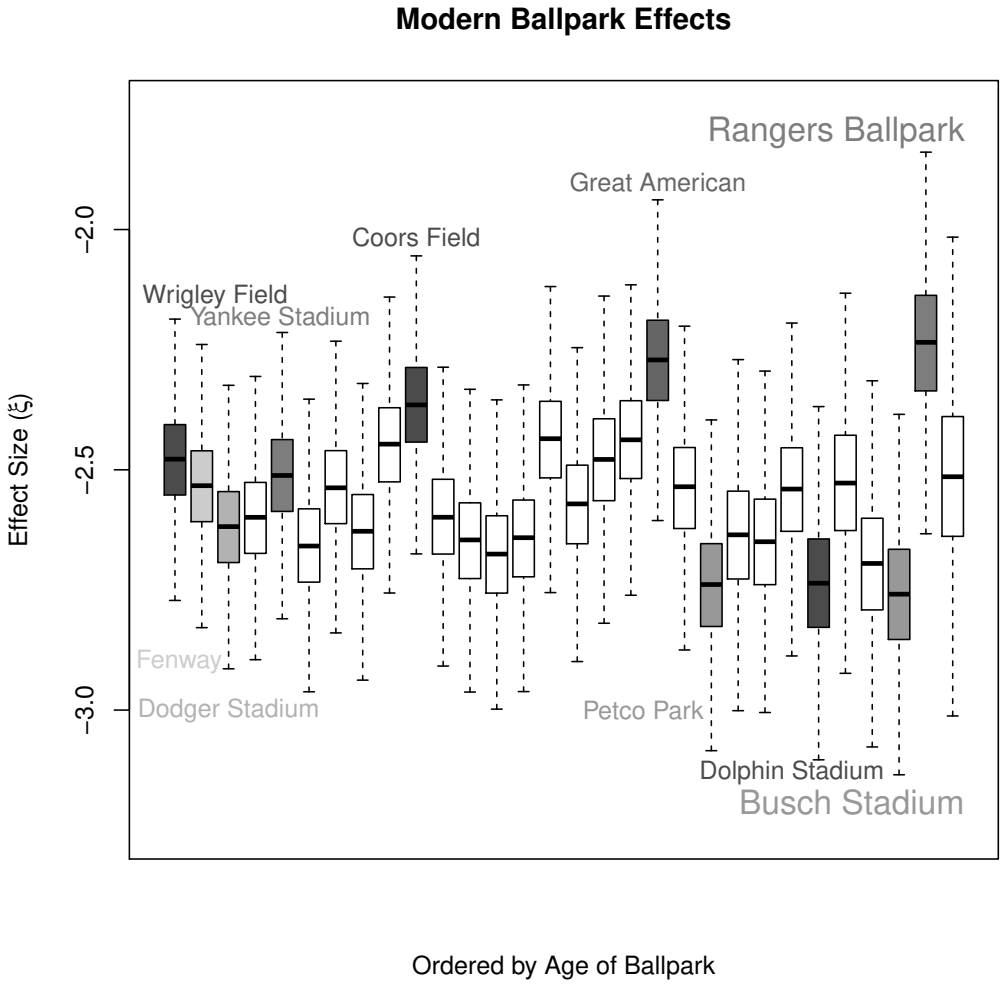

Figure 5: Boxplots of posterior distributions of modern home ballpark effects. Parks are highlighted either because they were the three hardest to hit home runs in, the three easiest to hit home runs in, or they were one of four parks notable for other reasons.

4.3 'Pure Performance'

By removing ballpark, season and era effects from the mean curve on home runs per at-bat, we are left with a new metric that we call the 'pure performance' curve. We use these pure performance curves to predict future batting performance. The clustering induced by the Dirichlet process priors and the hierarchical nature of the fully parametric model allows for a borrowing of strength among similar players to obtain performance curves that may more accurately represent a player

of a certain type.

The means of the posterior distributions of the polynomial coefficients, $\hat{\beta}_i$, are used as the coefficients to produce a mean performance curve using both the nonparametric and hierarchical models. As players' home ballparks of the future are not predictable, and because season effects of future seasons are unknown, we estimate all of the effects by estimating an overall effect, ϕ_i where

$$\underset{\phi_i}{\operatorname{argmin}} \sum_{j=1}^{n_i} \left(\frac{h_{ij}}{ab_{ij}} - \hat{\pi}_{ij} \right)^2$$

and

$$log\left(\frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}} \right) = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}_i + \phi_i.$$

So

$$\hat{\pi}_{ij} = exp(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}_i + \phi_i)/(1 + exp(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}_i + \phi_i)).$$

Performance plots and predictions are presented in terms of the $\hat{\pi}_{ij}$. Although across era comparisons were not the purpose of the analysis, we note that Babe Ruth is the best all time home run hitter using pure performance curves generated from the nonparametric model, while Mark McGwire is second.

4.4 Prediction of Future Performance

The efficacy of these methods was tested using data from eight seasons (Baseball-Reference.com, 2017) that have been played since the data used to build the model were produced. As the archetypal desire for prediction is the future performance of young players, we looked at the three youngest groups of players in the dataset. Because players with fewer than 6 seasons experience through 2008 were removed, we look at players with 6 - 8 seasons experience through the 2008 season, and thus 11-16 seasons of experience at the end of the 2016 season. Specifically, we selected players from this subset who were also in the current top 100 active players in total home runs at the conclusion of the 2008 season. This left 22 players. Of these, 2 players played only 4 of the 8 seasons, 6 played 5 seasons, 4 played 6 seasons, 2 played 7 seasons, and 8 played all 8 seasons. (The 22 players we considered are: Albert Pujols, Mark Teixeira, Carlos Pena, Travis Hafner, Chase Utley, Brandon Inge, Victor Martinez, Lyle Overbay, Mike Cuddyer, Adam

Dunn, Miguel Cabrera, Jason Bay, Justin Morneau, Juan Uribe, Ty Wigginton, Eric Hinske, Austin Kearns, Bill Hall, Jhonny Peralta, Miguel Olivo, Juan Rivera, and Carl Crawford.)

As noted previously, we believe the clustering process engendered by the nonparametric portion of the model is useful in prediction. The number of clusters will vary from iteration to iteration, but in our model the 3735 players tended to be grouped in roughly 55 clusters. In figure 6 we show a histogram of the posterior number of clusters at each iteration.

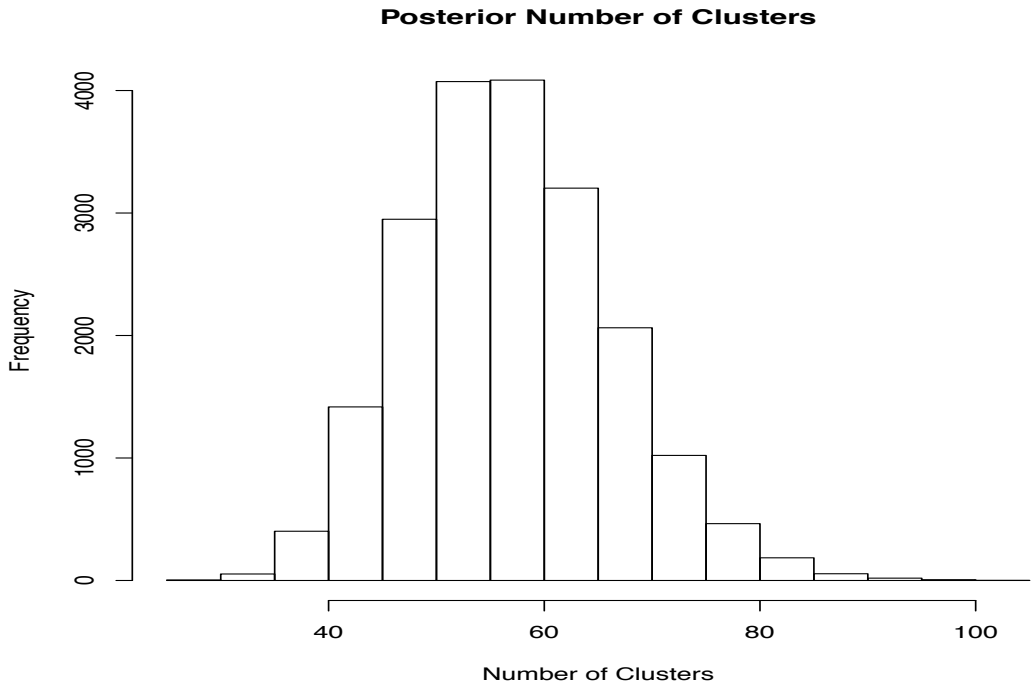

Figure 6: Histogram of the number of clusters at each iteration of the posterior sampling.

Using the method of Dahl and Newton (2007) we found an 'optimal' cluster for the purpose of examining the players who tended to be grouped together. In figure 7 we show the average performance curves for players in four of these clusters in gray or in color, along with the performance curve for that particular cluster in black. It is important to note that the curves we show are average curves for the players. The single black curve is the curve produced for this single iteration of this cluster which meets the criterion of 'optimal' by the definition of Dahl and Newton (2007). There are in each cluster some curves that are quite different from one another. However, there is clearly a grouping of players that perform similarly over their careers. In this figure, each of the clusters has one to three of the players whose future performance we predict printed in color. It can be seen how the future performance of these players could reasonably be

thought to follow the performance of the majority of the other players in the cluster.

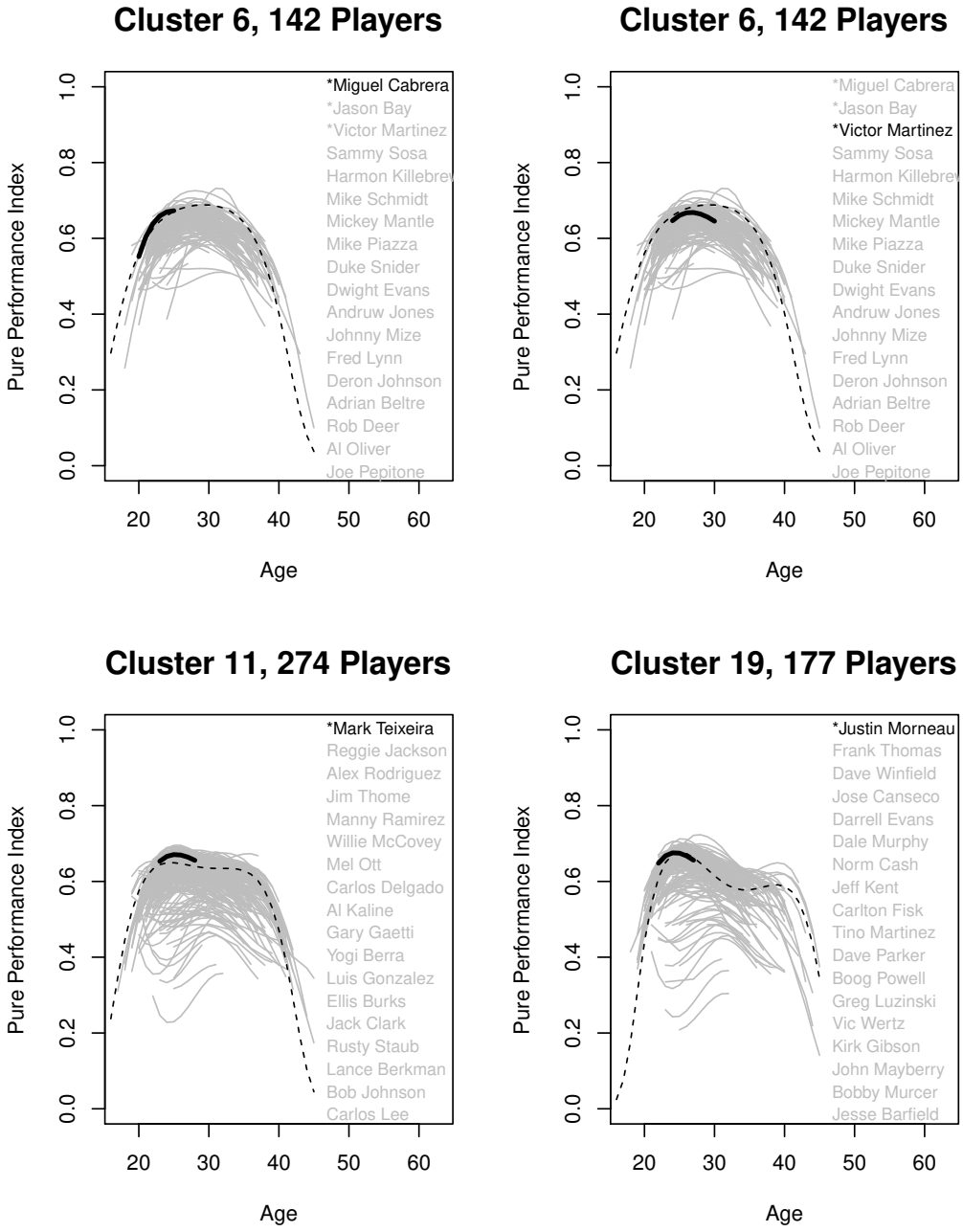

Figure 7: Four clusters from the 'optimal' grouping using the method of Dahl and Newton (2007). Gray curves are average curves for the players found in that cluster. Solid curves are average curves of players for whom we predict future performance. The dashed curve is the performance curve from this single iteration.

The 95% posterior predictive intervals were calculated for all 22 players using both the non-parametric and hierarchical models. We rate performance of the methods by the number of seasons the 95% posterior prediction intervals contain the actual season output of home runs. Since we would not know in advance how many at bats to expect for each player, we used 500 at bats for all players. This number is in general somewhat high, as most players do not achieve 500 at bats in a given season. The choice of 500 at bats makes the 95% prediction intervals conservative. Good predictive capability in this conservative situation would obviously be just as good with a lower figure for number of at bats.

First, we note that in general, the predictive performance of the models is quite good. In figure 8 we show the three players for whom the 95% posterior predictive intervals included all future performance.

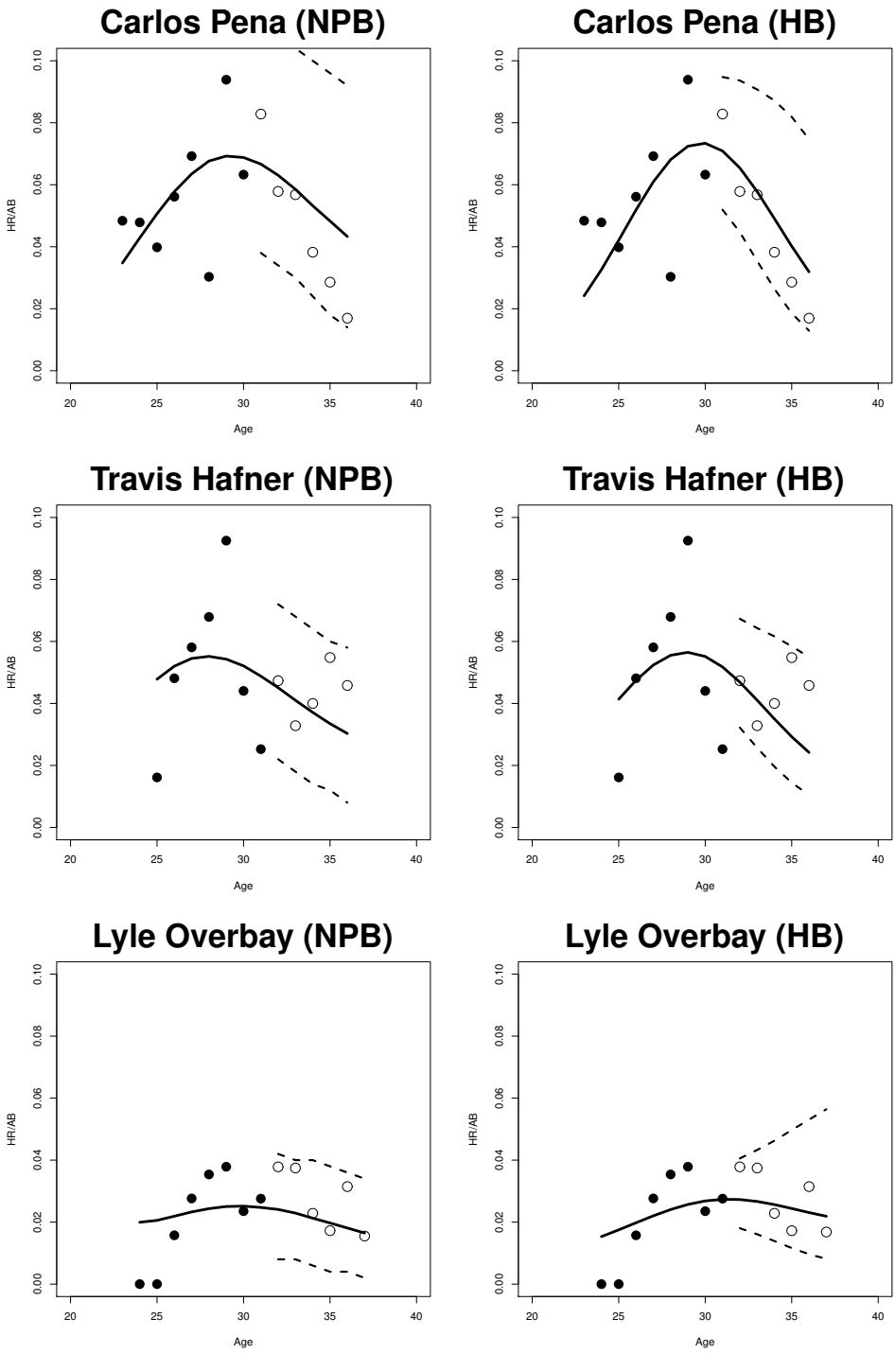

Figure 8: The three players whose performance was always included in the 95% posterior predictive intervals for both the nonparametric Bayes (NPB) and the hierarchical Bayes (HB) models. The open circles represent the seasons which were not included in the model. The solid line is the 'pure performance' curve. The dashed lines are 95% posterior prediction intervals.

There were, however, some individuals whose performance was not predicted as well. For example, Mike Cuddyer in Figure 9 was significantly under predicted using both models. There were two data points that seemed to influence the model downward. In his first year in the majors, 2001, he hit 0 home runs in 20 at bats (Baseball-Reference.com (2017)). In his eighth year, 2008, he hit only 3 home runs in 249 at bats. He injured his hand early in that season, and never really recovered. It appears that this poor year may have influenced our pure performance curve unduly, in that the curve itself seems to be a reasonable fit for the first 8 seasons, but under predicts the next 7 using both methods.

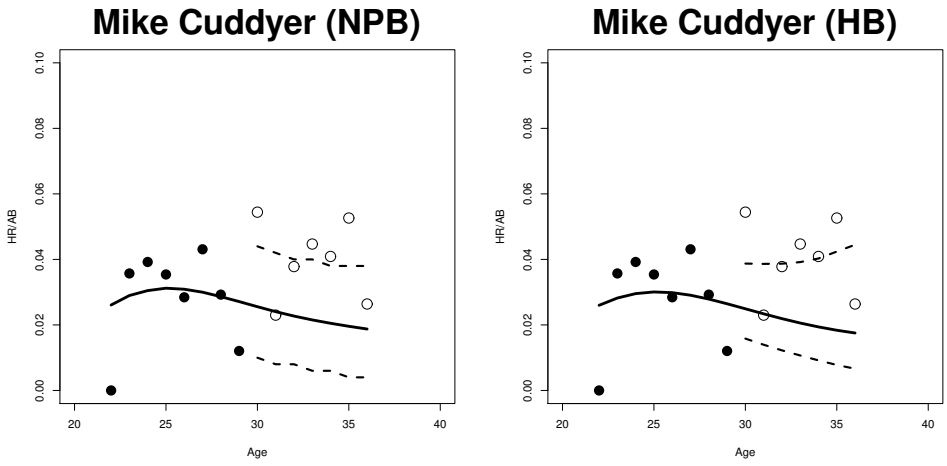

Figure 9: Posterior predictive curves for Mike Cuddyer. Both models underpredicted his performance. The open circles represent the seasons which were not included in the model. The solid line is the 'pure performance' curve. The dashed lines are 95% posterior prediction intervals.

In general, the nonparametric method out performed the hierarchical model. Of the 22 players for whom we predicted future performance, 11 players were predicted better using the nonparametric model, while 2 players were predicted better using the hierarchical model. For 9 players, the prediction performance was equivalent. In figure 10 we show three of the players whose predictions were better using the nonparametric model.

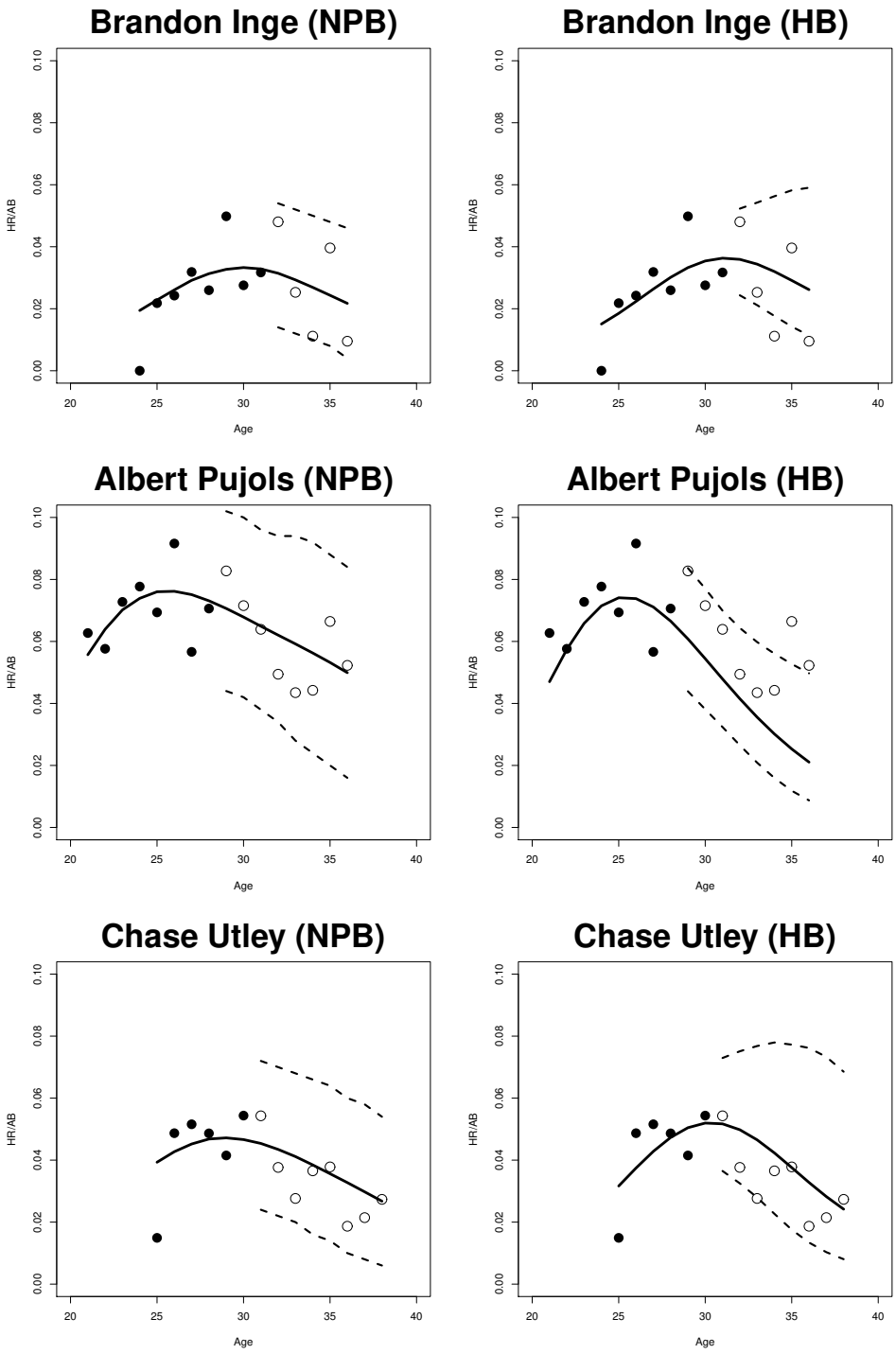

Figure 10: Three of the 11 players whose future performance was predicted better by the non-parametric model. The open circles represent the seasons which were not included in the model. The solid line is the 'pure performance' curve. The dashed lines are 95% posterior prediction intervals.

In figure 11 we show the 2 players whose predictions were better using the hierarchical model. It should be noted that the prediction intervals are fairly wide for both methods. For 500 at bats, a typical prediction interval would be from 15 to 35 home runs. This is quite a wide spread. But we also see that the data have quite a bit of year-to-year variation.

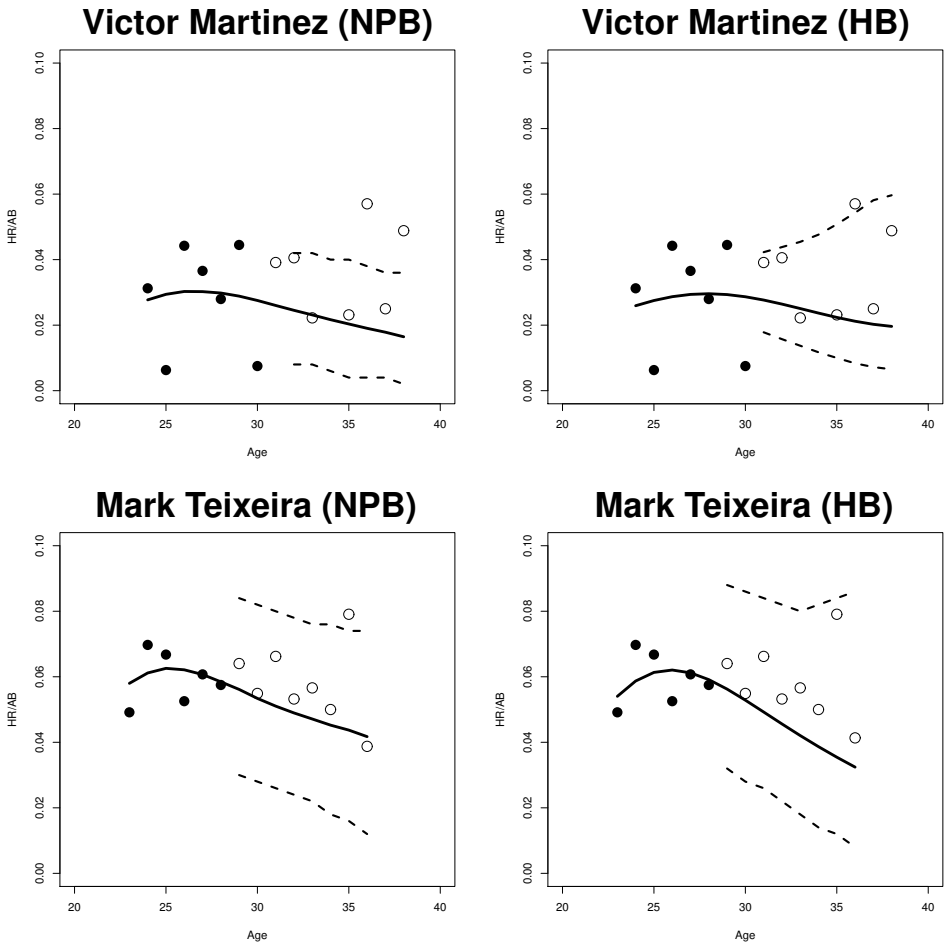

Figure 11: The two players whose future performance was predicted better by the hierarchical model. The open circles represent the seasons which were not included in the model. The solid line is the 'pure performance' curve. The dashed lines are 95% posterior prediction intervals.

For these 22 players, home run output was within the prediction interval 118 times of 131 seasons (90.1%) using the nonparametric model. For the 22 players, 105 of 131 seasons (80.2%) had actual output within the 95% posterior predictive inntervals of the hierarchical model. If all seasons where at bats were less than 100 and Mike Cuddyer's performance were removed from consideration, those numbers become 116 of 124 (93.5%) for the nonparametric model, and 102

of 124 (82.3%) for the hierarchical model.

We also computed the root mean square error (RMSE) and the median absolute deviation (MAD) for each model. We used the 124 predicted points that did not include Mike Cuddyer or any season with fewer than 100 at bats. RMSE was 0.01485 for the HB model and 0.01482 for the NPB model. MAD was 0.00989 for the HB model and 0.00829 for the NPB model.

We are not aware of any other work that has attempted to predict performance over such a significant time frame. Being able to use past information to predict performance for eight years in the future would seem to be a useful tool.

5 Discussion

We have demonstrated a methodology that allows for flexible modeling of functional data while accounting for latent effects. Fellingham et al. (2015) showed the advantages of using nonparametric Bayesian methods to predict future performance. These advantages have been verified by this paper. The natural clustering induced by the method means that predicted performance is, in a sense, still accomplished within the range of the data. That is, even though we have no data for the players for whom we predict, we do have data for players whose lifetime performance is similar.

While we believe this method is very useful for predicting, there are some areas where the methodology could be improved. Large data sets particularly lend themselves to this methodology since the potential for borrowing strength is increased. However, the computational speed of the proposed method needs to be increased for the methodology to be seen as a reasonable alternative if time is an issue. We believe algorithms that speed conversion of posterior distributions would be extremely useful in this context. Implementing the work of Polson, Scott and Windle (2013) in this setting would seem to be a viable option. Also, in other work we have found algorithm 8 of Neal (2000) to be superior to algorithm 6 used in this example when evaluating sampling variability.

It is not clear what effect including other covariates in the model might have in the predictive power of the method. If covariates are available, it may be that the efficacy of the Bayesian semiparametric method is decreased. However, we are not aware of anyone who has examined this question.

We found our predictions for the 22 young players chosen to be reasonably accurate. Thus, we believe that this methodology could be used to guide decision makers as they seek to determine which players are most likely to be productive as their careers progress.

References

Albright, S. (1993), 'A statistical analysis of hitting streaks in baseball', *Journal of the American Statistical Association* **88**, 1175–1183.

Baseball-Reference.com (2017), 'Baseball encyclopedia of players', `http://www.baseball-reference.com/players/`.

Berry, S. M., Reese, C. S. and Larkey, P. D. (1999), 'Bridging different eras in sports', *Journal of American Statistical Association* **94**(447), 661–684.

Blackwell, D. and MacQueen, J. B. (1973), 'Ferguson distributions via Polya urn schemes', *The Annals of Statistics* **1**(2), 353–355.

Carlin, B. P. and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, 3rd edn, Chapman & Hall/CRC.

Dahl, D. B. and Newton, M. A. (2007), 'Multiple hypothesis testing by clustering treatment effects', *Journal of the American Statistical Association* **102**, 517–526.

Escobar, M. D. and West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**, 577–588.

Fellingham, G. W., Kottas, A. and Hartman, B. M. (2015), 'Bayesian nonparametric predictive modeling of group health claims', *Insurance: Mathematics and Economics* **60**, 1–10.

Ferguson, T. (1973), 'Bayesian analysis of some nonparametric problems', *Annals of Statistics* **1 (2)**, 209–230.

Frigyik, B. A., Kapila, A. and Gupta, M. R. (2010), Introduction to the Dirichlet distribution and related processes, Technical Report UWEETR-2010-0006, Department of Electrical Engineering, University of Washington.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

Ghahramani, Z. (2013), 'Bayesian non-parametrics and the probabilistic approach to modelling', *Philisophical Transactions of the Royal Society* **371**(UNSP 20110553).

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57 (1)**, 97–109.

Jensen, S. T., McShane, B. B. and Wyner, A. J. (2009), 'Hierarchical Bayesian modeling of hitting performance in baseball', *Bayesian Analysis* **4**(4), 631–652.

Lahman, S. (2016), 'Lahman baseball database', `http://www.seanlahman.com/baseball-archive/statistics/`.

MacEachern, S. N. (1998), Computational methods for mixture of Dirichlet process models, *in* D. Dey, P. Müller and D. Sinha, eds, 'Practical Nonparametric and Semiparametric Bayesian Statistics', Vol. 133 of *Lecture Notes in Statistics*, Springer New York, pp. 23–43.

MacEachern, S. N. and Müller, P. (1998), 'Estimating mixture of Dirichlet process models', *Journal of Computational and Graphical Statistics* **7**(2), 223–238.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21 (6)**, 1087–1092.

Müller, P. and Rosner, G. (1998), Semiparametric pk/pd models, *in* D. Dey, P. Müller and D. Sinha, eds, 'Practical Nonparametric and Semiparametric Bayesian Statistics', Vol. 133 of *Lecture Notes in Statistics*, Springer New York, pp. 323–337.

Neal, R. M. (2000), 'Markov chain sampling methods for Dirichlet process mixture models', *Journal of Computational and Graphical Statistics* **9**, 249–265.

NIMBLE Development Team (2017), 'Nimble: An r package for programming with bugs models, version 0.6-5'.
URL: *http://r-nimble.org*

Polson, N. G., Scott, J. G. and Windle, J. (2013), 'Bayesian inference for logistic models using Polya-gamma latent variables', *Journal of the American Statistical ASsociation* **108**(504), 1339–1349.

Quintana, F. A., Müller, P., Rosner, G. L. and Munsell, M. (2008), 'Semi-parametric Bayesian inference for multi-season baseball data', *Bayesian Analysis* **3**(2), 317–338.

Raftery, A. E. and Lewis, S. M. (1992), 'One long run with diagnostics: Implementation strategies for markov chain monte carlo', *Statistical Science* **7**, 493–497.

Schell, M. (1999), *Baseball's All-Time Best Hitters*, Princeton, NJ: Princeton University Press.

Sethuraman, J. (1994), 'A constructive definition of of Dirichlet priors', *Statistica Sinica* **4**, 639–650.