Summary of “Predicting Home Run Production in Major League Baseball Using a Bayesian Semiparametric Model”, by Gilbert W. Fellingham and Jared D. Fisher

Since the conception of the game of baseball, numbers have defined the game. As a result, the ability to predict these numbers has become arguably a more important skill than actually producing those numbers. In this article, Fellingham and Fisher attempt to use Bayesian analysis to predict the production of Major League Baseball players, specifically in the case of home runs, while taking into effect multiple other factors, such as era/time frame and ballpark factors. However, in baseball, there is no clear, definitive prior probability distribution for anything, which is where Bayesian analysis runs into roadblocks. Fellingham and Fisher attack this issue using a Dirichlet process to build a non-parametric prior to perform more accurate analysis on predicting future production.

Before going into the mathematics of the paper, there are a few terms that may need clarification. The paper outlines the use of a Dirichlet process as a way of building a prior. In these non-parametric problems, we need a way of creating a prior distribution. With the Dirichlet process, the prior is taken to be a set of probability distributions on a given sample space[1]. Essentially, to put it incredibly simply, the Dirichlet process is a Markov Chain Monte Carlo-esque stochastic process that, instead of sampling from a set of known prior parameters to build the unknown joint distribution with posterior hyperparameters, builds the prior probability distribution[2][3]. While that is not exactly what is happening mathematically in this situation, the logic of the stochastic process follows similarly enough to where this can be an effective analogy to further the understanding of the Dirichlet process.

Using this Dirichlet process (further abbreviated DP) to build a prior probability distribution has many benefits. By introducing a non-parametric prior, the distribution, since it is no longer constrained by specific values, becomes increasingly more flexible, while still providing sufficient information to maintain a meaningful, informative prior. Furthermore, the DP

[1] Ferguson, T. (1973), “Bayesian Analysis of Some Nonparametric Problems,” Annals of Statistics, 1, 209–230.
[2] Neal, R. M. (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” Journal of Computational and Graphical Statistics, 9, 249–265.
[3] Ferguson, T. (1973), “Bayesian Analysis of Some Nonparametric Problems,” Annals of Statistics, 1, 209–230.

introduces clustering in the posterior distributions. In this particular study, the clusters serve as a way of easily comparing players with similar career development and trajectory, which allows for more feasible and accurate predictions of future performance. Also, by using the non-parameterized prior, the analysis is able to isolate potential influential factors outside of player performance, such as era, age, or ballpark. By accounting for these factors, it is possible to isolate what can be interpreted as the pure production of the player, allowing for comparison across all time frames and home ballpark advantages.

The data for this project derives from the Lahman Baseball Database, which contains all Major League Baseball hitting data from when the numbers were first tracked in 1871 to the present day. Because the goal of this paper is predictive analysis, the dataset was then limited to players with substantial data available. Thus, the dataset was condensed to only include players with at least six seasons of service time, with at least 50 at bats in each of those six or more seasons. Finally, to account for extreme outliers, the data was limited to players younger than 45. All in all, the data includes 3735 players from 1871-2016. The seasons 1871-2008 are used to "train" the model, while 2009-2016 are used to "test" the predictive ability of the model. Variables include: total number of home runs, total number of at bats, team, age, and year. A factor variable is also introduced to account for the home ballpark of each player, since some fields are easier to hit home runs at than others, as well as a factor for whichever decade the player was born in, since era could be heavily correlated to the amount of home runs being hit.

The likelihood function for predicting home runs is given as a binomial random variable, expressed as: $h_{ij}|\pi_{ij} \sim \text{Binomial}(ab_{ij}, \pi_{ij})$ where h_{ij} is the number of home runs hit by player *i* in 1, 2, . . ., 3735 in their *j*th season, ab_{ij} is the number of at bats for player *i* in season *j*, and π_{ij} is the probability of hitting a home run. This is a fairly intuitive formula, with *h* representing the number of successes in *ab* Binomial trials with probability of success π. Then, the predictive model uses logistic regression to model the log odds of hitting a home run as: $\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}_i + \theta_i + \delta_j + \xi_p$

with $\boldsymbol{\beta}_i$ representing a vector of coefficients for the 4th degree polynomial $\mathbf{x}_{ij}$ based on the age of player *i* in season *j*. θ_i is the variable that takes into account the effect of the decade of birth for a particular player, δ_j is the effect of the particular season, or year, and ξ_p is the ballpark factor.

The distribution of $\boldsymbol{\beta}_i$ is where this idea of the DP comes into play, assuming a DP prior on the unknown coefficients of the 4th degree polynomial. While we use the DP prior as a way of building a non-parameter prior, the DP still depends on two parameters: a "centering" distribution G_0 and a scaling parameter $\boldsymbol{\alpha}$, where larger values of $\boldsymbol{\alpha}$ correspond to the process being closer to G_0. Knowing this, the prior for the coefficients in $\boldsymbol{\beta}_i$ are as follows:

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

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 subscript 1-4 on $\boldsymbol{\beta}$ and *G* correspond to the different degree terms of the 4th degree polynomial. The parameters 2 and 0.2 of the Gamma distribution yield an expected value of 10, which in turn corresponds to approximately 60 clusters in our posterior distribution derived from the DP prior. The positive value of $\boldsymbol{\beta}_1$ reflects the predicted increase in development and production for players early in their careers, while the value of $\boldsymbol{\beta}_2$ reflects players inevitably regressing later in their careers. The relatively large values contained within Σ reflect the lack in certainty in our coefficients of the vector $\boldsymbol{\mu}$. There is a definitive pattern in the positive/negative values within Σ: values in columns/rows with the same parity of polynomial order (a number being odd or even) are positively correlated whereas opposite parities have negative correlations, which is not explained further but should feel somewhat intuitive.

The distributions of our remaining parameters are as follows:

$$\delta_j \overset{\text{iid}}{\sim} N(-2.5, \text{var} = 4), \quad j \in \{1, \ldots, 138\}$$

$$\xi_p \overset{\text{iid}}{\sim} N(-2.5, \text{var} = 4), \quad p \in \{1, \ldots, 222\}$$

$$\theta_i \overset{\text{iid}}{\sim} N\left(\mu_{d_i}, \sigma^2_{d_i}\right), \quad d_i \in \{1, \ldots, 14\}, i \in \{1, \ldots, 3735\}$$

$$\mu_{d_i} \overset{\text{iid}}{\sim} N(-2.5, \text{var} = 1), \; d_i \in \{1, \ldots, 14\}$$

$$\sigma^2_{d_i} \overset{\text{iid}}{\sim} IG(3, 3), \quad d_i \in \{1, \ldots, 14\}$$

The distributions of δ_j and ξ_p are simple Normal distributions. Their negative means reflect the idea that the probability of hitting home runs is generally low. The distribution of θ_i is more complex. θ_i follows what is called a hierarchical prior distribution, where the parameters of θ_i are themselves random variables. The IG distribution of $\boldsymbol{\sigma}_d^2$ is the inverse gamma distribution.

For the sake of comparison, the paper also fits a parametric model. The distributions of θ_i, δ_j, and ξ_p are identical in the parametric model as they are in the non-parametric, however since $\boldsymbol{\beta}$ followed the DP prior, this distribution changes drastically:

$$\begin{aligned} \boldsymbol{\beta}_{1-4,i} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), i = 1, \ldots, n_{\text{players}} & \sigma_1^2 &\sim \text{Gamma}(15, .5) \\ \mu_1 &\sim N(0.5, 16) & \sigma_2^2 &\sim \text{Gamma}(15, .5) \\ \mu_2 &\sim N(-0.5, 16) & \sigma_3^2 &\sim \text{Gamma}(10, .5) \\ \mu_3 &\sim N(0.8, 9) & \sigma_4^2 &\sim \text{Gamma}(10, .5) \\ \mu_4 &\sim N(0.1, 9) \end{aligned}$$

This new distribution is similar to the hierarchical prior seen in θ_i. Also, as before, the subscripts of 1-4 represent the different degrees of polynomial in the 4th degree polynomial.

The computation of the posterior distribution is completed using the Metropolis-Hastings algorithm for Markov Chain Monte Carlo. This particular algorithm has been known to take numerous iterations,[4] which is very much representative of this paper. Something to note is that the samples taken for the MCMC for the era effect and the season effect are taken from a joint multivariate normal distribution to account for the correlation between those two factors. Another note is that the two non hierarchical parametric variables, δ_j and ξ_p have their initial values set to their means of -2.5, but all other parameters and hyper parameters are calculated using the Metropolis-Hastings algorithm. Once the MCMC process started, the values of ξ_p, θ_i, and $\boldsymbol{\sigma}^2$ are independently sampled from a normal distribution with mean equal to the previous iterated value and standard deviation $\boldsymbol{\sigma}_c$, which becomes adjusted every 1000 iterations based on the percentage of acceptances/rejections present in the previous 1000 iterations, until the burn-in period is finished and $\boldsymbol{\sigma}_c$ becomes fixed.

[4] Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," Journal of Computational and Graphical Statistics, 9, 249–265.

To produce the posterior distribution, the MCMC process first began with a 100,000 iteration burn-in period. Then, for each of the parameters, 20 values were randomly selected to create 20 parallel MCMC sampling chains. Next, these 20 chains for each parameter were burned for another 100,000 iterations before running 500,000 iterations, only extracting every 500^{th} value. Then, the 20 parallel sample chains for each parameter get appended together to make one continuous chain for each parameter. So, all in all, each parameter was run for 12,100,000 iterations, with only 20,000 being extracted to become the posterior distribution. The process for the parameterized model was very similar, except only one chain was run for 200,000 iterations, with every 20^{th} value drawn to become the posterior. With only one chain being run, the diagnostics were reported using the Rafferty-Lewis criterion.

While these simulations provide posterior distributions for the predictors of era and ballpark, one of the main goals of this analysis is to remove those factors to isolate pure performance of individual players. These performance curves, modeled by the 4^{th} degree polynomials examined throughout the analysis, are clustered as a result of the DP prior, causing players with similar career development and trajectory to be grouped together, leading to easier, more reliable and accurate prediction models. The scope was narrowed to active players with at least 6 years of experience as of 2008 who were also in the active top 100 in career home runs as of 2008, leaving a subset of 22 players. After calculating a 95% prediction interval for each player using both the hierarchical and non-parametric models, it was clear to see that the non-parametric model performed better. 20 of the 22 players were predicted as accurately or more accurately under the non-parametric model, with the future performance of 3 players being 100% within the prediction intervals. All in all, when using the non-parametric model, predicted home run output was within the interval 90.1% of the time (118 of 131 total seasons), which increased to 93.5% (116 of 124 seasons) when removing outliers, which was significantly better than the hierarchical model, which had 80.2% (105/131) and 82.3%(102/124) accuracy, respectively.