Improving Batter-Pitcher At-Bat Modeling Using K-Means Clustering and Mixture Models

Abstract

In baseball, a common refrain is that a batter "hits this pitcher well" due to their success against the pitcher. Given the typically small sample sizes of batter-pitcher matchup history, this sentiment rarely holds predictive value. The study aims to increase sample sizes by grouping similar pitchers together so that the question is whether a batter "hits this *type of* pitcher well." This study uses k-means clustering and mixture models to group pitchers to answer this question, while also comparing the two different grouping methods. The findings of the study argue that looking at a batter's history against similar pitchers is far more predictive than looking at a batter's history against a single pitcher. Between the two grouping methods, mixture models produced more predictive groupings and provide a framework for future analytic research.

Keywords: Baseball, Sabermetrics, Mixture Models, K-Means, Beta-Binomial, Logistic Regression, ANOVA

1 Introduction

1.1 Overview

In baseball, a pitcher throws a variety of pitch types towards a batter in an attempt to fool the batter into missing the pitch or inducing weak contact. The batter is trying to make solid contact with the ball to generate a hit. When teams can bunch their hits together, this will oftentimes lead to runs, which is how the game is scored. The mean batting average ($\frac{hits}{AB}$) in Major League Baseball is typically around .250 with a standard deviation of .027. For each pitch thrown, hundreds of data points are generated by high-speed cameras that gauge everything from the spin axis of the ball to the exact millimeter the ball is released from. The following statistics are used in the pitcher grouping:

Pitch Rate: The percentage of all pitches a pitcher throws of a given type.

Velocity: The miles per hour of the ball at its release speed.

Spin Rate: Spin rate is a measure of how often a ball spins during the time (in rotations per minute) it is thrown to the plate.

Extension: Extension is the distance from the center of the pitching rubber to where the pitcher releases the ball. For this data, I used the Pythagorean x, z distance ($\sqrt{x^2 + z^2}$) although if the x-coordinate is negative, I kept that to represent the difference between righties and lefties. A negative extension represents a right-handed pitcher.

1.2 Model Overview

There are many ways to group multivariate data. In unsupervised learning, there are hundreds of different strategies to say that one observation in a dataset is similar to another.[1] This study will compare two different methods: k-means clustering and finite mixture models.[2] Using a baseball dataset, this study aims to group observations using the two methods and then compare the benefits that each model offers. Each algorithm will also be compared to a control group in which each observation is treated as its own group.

[1]The term "observation" will refer to a pitcher, rather than an at-bat.

[2]When the word "cluster" is used, it will be used to refer to the k-means algorithm, although mixture models can also be used for clustering. "Grouping" will be used to refer to both methods.

1.2.1 Past Model Work

K-means K-means clustering has a long history as the preeminent unsupervised learning technique to group together data. K-means takes an input "k" number of clusters and then assigns a random observation to each cluster. K-means then assigns each future observation to the cluster of which the centroid of the cluster is most similar to the observation. The centroid is the theoretical mean observation of a given cluster but does not necessarily exist within the dataset, although after initializing the clusters with k-observations, the centroid would be the initial point of each cluster. K-means measures the similarity between two points by using the sum of the squared Euclidian distances for each observation and each centroid and assigning the observation to the cluster that minimizes that distance. After each observation has been assigned to its initial cluster, the observations are tested again to see if an updated centroid is a better fit. This process repeats until no observation switches clusters. K-means provides a quick and dirty method that does a fairly accurate job of assigning clusters although it is not entirely optimal (Hartigan & Wong 1979).

Mixture Models Mixture models were initially used to identify separate groups of univariate data. A mixture model assumes that there is an underlying set of distributions that our data belongs to, and that each data point belongs to one of those distributions. A normal mixture model is the most common mixture model and assumes that each data point belongs to some unknown normal distribution. The mixture model works to find the best potential normal distributions and assigns that point to a given mixture. Unlike k-means, mixture models assign points to mixtures probabilistically, recognizing that each point may have a reasonable chance of belonging to multiple mixtures. An example of the utility of mixture models can be seen below. Mixture A is a set of 10,000 random observations generated from a normal distribution with a mean of 2 and a standard deviation of 1. Mixture B follows a normal distribution with a mean of 5 and a standard deviation of 2. It can be difficult to ascertain that there are two distinct distributions from the leftmost plot, but with overlapping density plots, it can be more clear. The overlap between the density plots can also help to show the need for probabilistic groupings. While many values can be safely assumed to be in one mixture or the other, other points could reasonably belong

to either cluster as the values are inside the density plots of both mixtures (McLachlan & Peel 2000).

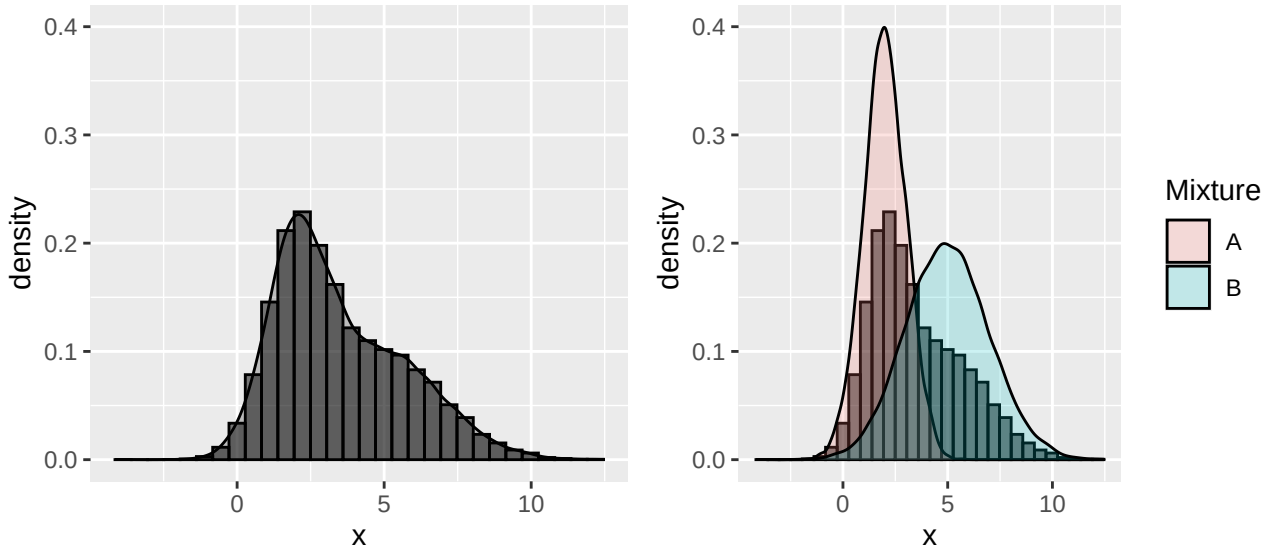

Figure 1: Visualizing Mixture Models

With a known number of components, mixture models use an expectation-maximization algorithm to help identify potential parameters for the different distributions. The expectation-maximization works similarly to k-means by initially assigning samples of the data to "k" (a user input value) to k mixtures and then assuming that each remaining observation has a uniform likelihood of belonging to each mixture. Next, the probability of each remaining observation belonging to each mixture is found, using probability theory. Finally, each mixture's distribution is updated by maximizing the expected conditional log-likelihood with respect to the assumed distribution (Melnykov & Maitra 2010). For multivariate mixture modeling, the same principles apply as univariate mixture modeling, but use a covariance matrix for each observation for assumed distributions to define the relationship between variables (Nagode & Fajdiga 2011). The shapes of the models must also be defined in terms of the orientation of the eigenvectors of the mixtures. To simplify the problem, we can add constraints to the groupings that we look for in the covariance matrix, allowing us to look at fewer parameters to estimate and provide easier interpretations of the mixture outputs. The different parameters are the volume, shapes, and orientation of the mixtures. (Hennig

2011).

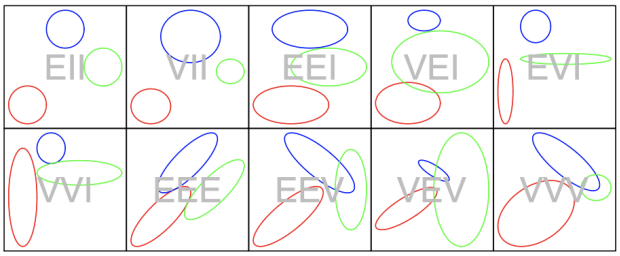

Figure 2: Mixture Model Shapes

1.3 Baseball Overview

An oft-heard refrain while watching a baseball game is that a batter hits a given pitcher well. Most of the time, this is a comment on a batter's history of success against that pitcher. With few exceptions, each batter only has small sample sizes against each pitcher they face, as each batter likely only faces a specific pitcher a few times each season if at all. Occasionally, over many years, a batter may face a pitcher many times. Over 2015-2019, the most common batter-pitcher at-bat was Lorenzo Cain versus Jose Quintana, as they had 71 plate appearances against each other. Lorenzo Cain had 2943 plate appearances over that span and Jose Quintana faced 3971 batters. Only 2.4% of Cain's plate appearances were against Quintana and 1.8% of Quintana's at-bats were against Cain. Even in the most common matchup in all of baseball, the sample size is quite small. Frequent batter-pitcher matchups are even rarer in the current 30-team era with more pitchers per game and more evenly distributed scheduling (Lindbergh 2021). On average, for any given at-bat, the batter faced the pitcher just 2.25 times between 2015-2019, excluding that at-bat itself. The distribution of the at-bats against a pitcher can be seen below.

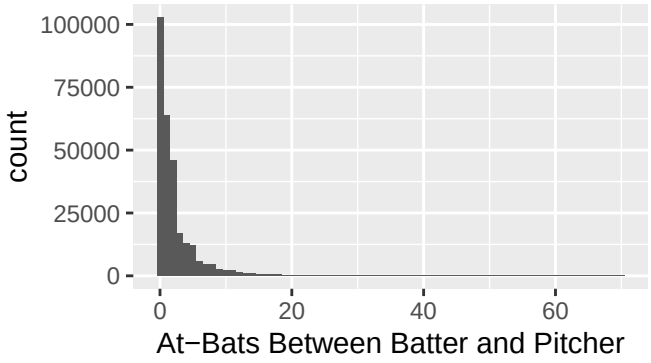

Figure 3: The Infrequency of Repeated Batter-Pitcher Matchups

In short, due to the small sample sizes, it can be difficult to make meaningful claims from just a batter's history against a given pitcher, such as that a batter hits a specific pitcher well. If we assume that the results of a series of at-bats follow a binomial distribution with true probability p, even if a batter is 4 for 9 against a pitcher, good for a .444 batting line, this does not indicate a strong skill against that pitcher. If the true expected batting average against that pitcher for this batter is .250, that batter will get 4 or more hits almost 17% of the time per 9 at-bats. Likewise, the batter will get no hits or just one hit over 30% of the time. The sample sizes are simply too small to make any claim of value for most situations. Therefore, this study aims to narrow the binomial test confidence interval by expanding our sample size, while also looking to balance the tradeoff between increasing sample sizes and decreasing the value of the information, as the greater our grouping sizes are (and thus the greater our sample sizes would be), the less similar each pitcher will be.

This study seeks to remedy this problem by increasing the sample sizes. Instead of looking at just a batter's history versus a given pitcher, this study looks at a batter's history against a given *type of* pitcher. While all pitchers are unique, some pitchers are more unique than others. From the way they pitch, to the pitches they throw, to how they utilize their various pitches, each pitcher's pitching style tells a story awash in data. This data can be utilized to group pitchers together.

1.3.1 Past Baseball Work

Numerous past studies have looked to measure the value of batter-pitcher matchup history. Renowned baseball statistician Tom Tango first looked at this in a 2006 study, finding that he could not conclude that batter-pitcher matchup history held any meaning in predicting future outcomes (Tango et al. 2014). Former FanGraphs writer and San Diego Padres analyst Dave Cameron took another approach to assess batter-pitcher history. He uses the example of the Mariners starting 41-year-old Raúl Ibañez in 2013 against Yankees ace C.C. Sabathia. Ibañez had very good career numbers against Sabathia in 52 plate appearances, hence why the Mariners chose to play Ibañez that day, despite his normally not starting. As Cameron further breaks down Ibañez's history against Sabathia, he notes that much of Ibañez's success came in 2002 and 2003, and Ibañez was fairly mediocre against Sabathia in the years afterward. In 2002, Ibañez was in his prime as an above-average outfielder, whereas Sabathia was 21 years old, in his second year in the majors, and still learning how to pitch at a Major League level. While Ibañez undoubtedly had plenty of experience against Sabathia, the notion that Ibañez's history over ten years ago would be relevant against Sabathia in 2013, despite Ibañez's lack of recent success, demonstrates that even extensive batter-pitcher matchup histories oftentimes do not hold predictive merit (Cameron 2013).

2 Methods

2.1 Data Collection

The entirety of this data was scraped from Baseball Savant using the `baseballR` package. The original dataset consisted of every pitch in the Major League Baseball regular season from 2015 to 2019. These years were chosen, as at the time of data collection these were all past full seasons with Statcast data available. Each pitcher who threw at least 100 pitches of each of their pitch types over the 5 seasons was included in the dataset. Due to small sample sizes, eephuses, knuckleballs, forkballs, and screwballs were not included, as no more than 4 total pitchers in the entire dataset threw any of those pitches. In the end, there were 1204 pitchers in the dataset. For modeling pitch outcomes, each at-bat from 2015-2019 was included. For modeling purposes, only at-bats in which both the hitter had

more than 30 plate appearances and the pitcher had more than 30 batters faced, were kept. Using this methodology, 925,098 at-bats were kept.

2.2 Methods Overview

To best measure the validity of different grouping approaches, there are a series of steps to take. Firstly, the pitchers need to be separated into groups and the number of groups for each method must be determined. Ideally, the group count of both methods is relatively similar, so that comparisons can be more easily made and have similar values. Once the groupings are complete, comparable models must be created using the various groupings (control, k-means, mixture). Each model should account for the differences in sample sizes between players. A player who is 30/100 against a cluster or mixture versus 3/10 should have different weights, as the 3/10 player should have more shrinkage and regression to the mean, whereas the player who has hit 30/100 should more heavily weight that data. Finally, the models should be compared to test the strengths and weaknesses of each grouping algorithm.

2.3 Preliminary Analysis of Dataset

To best understand the outputs, we should have some understanding of our data. The first thing to look at is the overall pitch rates in the dataset. The first column shows the rate of pitchers who throw each pitch. The second column shows the overall frequency of each pitch among all pitches in the dataset. The third column shows the mean frequency of each pitch, among all pitchers who throw that pitch. For example, 23.5% of pitchers throw a sinker, but only 10.7% of all pitches are sinkers. However, among the 23.5% of pitchers who throw a sinker, those pitchers throw 45.6% sinkers on average.

Table 1: Pitch Rates

Pitch Type	% of Pitchers	% of All Pitches	% Pitches Among Pitchers
2-Seam FB	0.355	0.081	0.227
4-Seam FB	0.881	0.431	0.489
Changeup	0.486	0.077	0.158
Curve	0.398	0.065	0.165
Cutter	0.213	0.045	0.211
Knuckle-Curve	0.068	0.013	0.197
Sinker	0.235	0.107	0.456
Slider	0.638	0.169	0.266
Splitter	0.061	0.011	0.186

Digging further into the numbers, velocity is a key component of each pitch. Each pitch type is, on average, thrown at a different speed. When looking at the barplot, the means are calculated by removing any pitcher who does not throw that specific pitch from the mean calculation; whereas in the violin plot, a pitcher who does not throw the pitch is calculated as a 0.

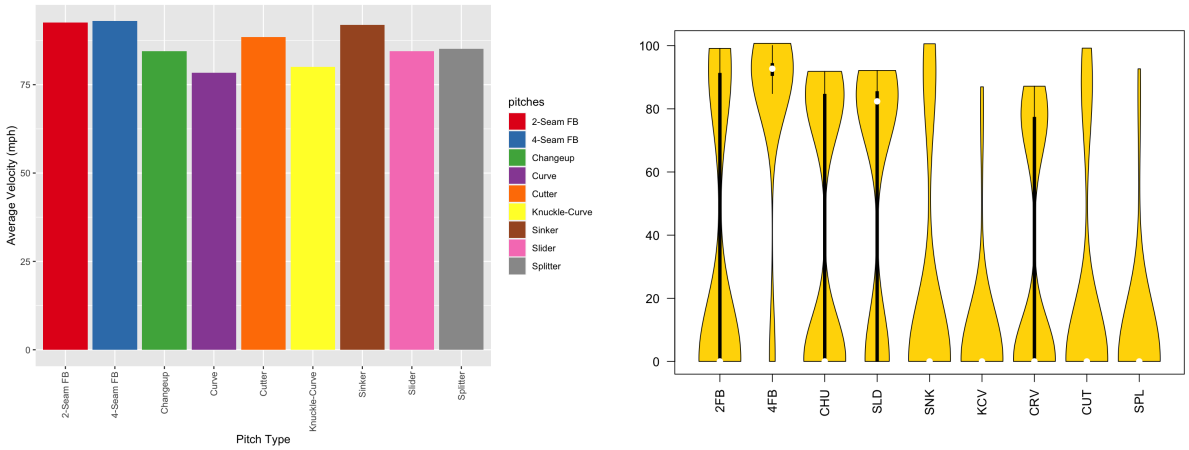

Figure 4: A Look At Velocities

Next, we will analyze spin rates, which track how the ball moves through the air. Like the velocity plots, when looking at the barplot, the means are calculated by removing any pitcher who does not throw the pitch from the mean calculation; whereas in the violin plot, a pitcher who does not throw the pitch is calculated as a 0.

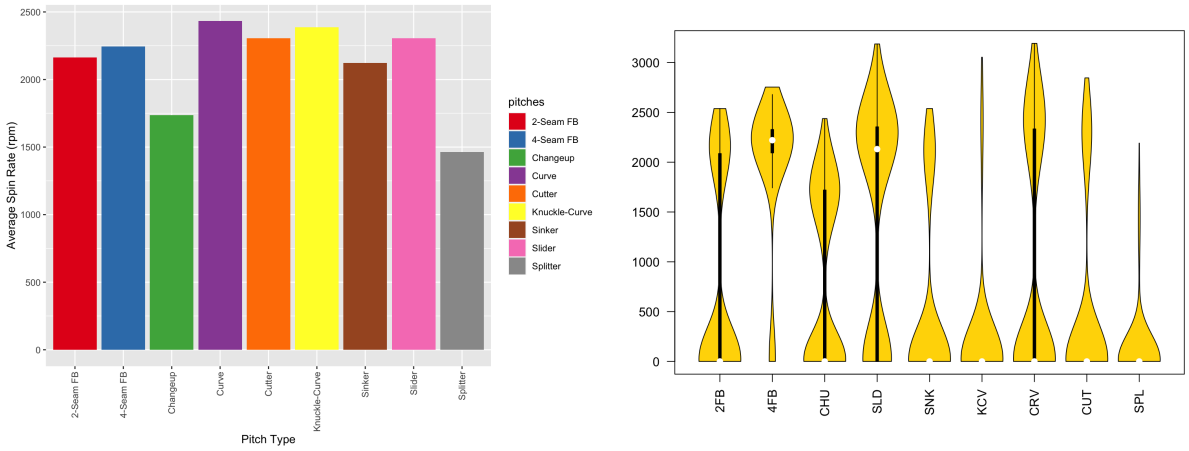

Figure 5: A Look At Spin Rates

The last, and potentially most tricky part of the dataset is the release extensions. In this dataset, extension is defined as the total Euclidian distance from the center of the mound in terms of vertical and horizontal height. Righthanders were programmed as negative extension. For pitchers who do not throw a specific pitch, their release point for that pitch is 0.

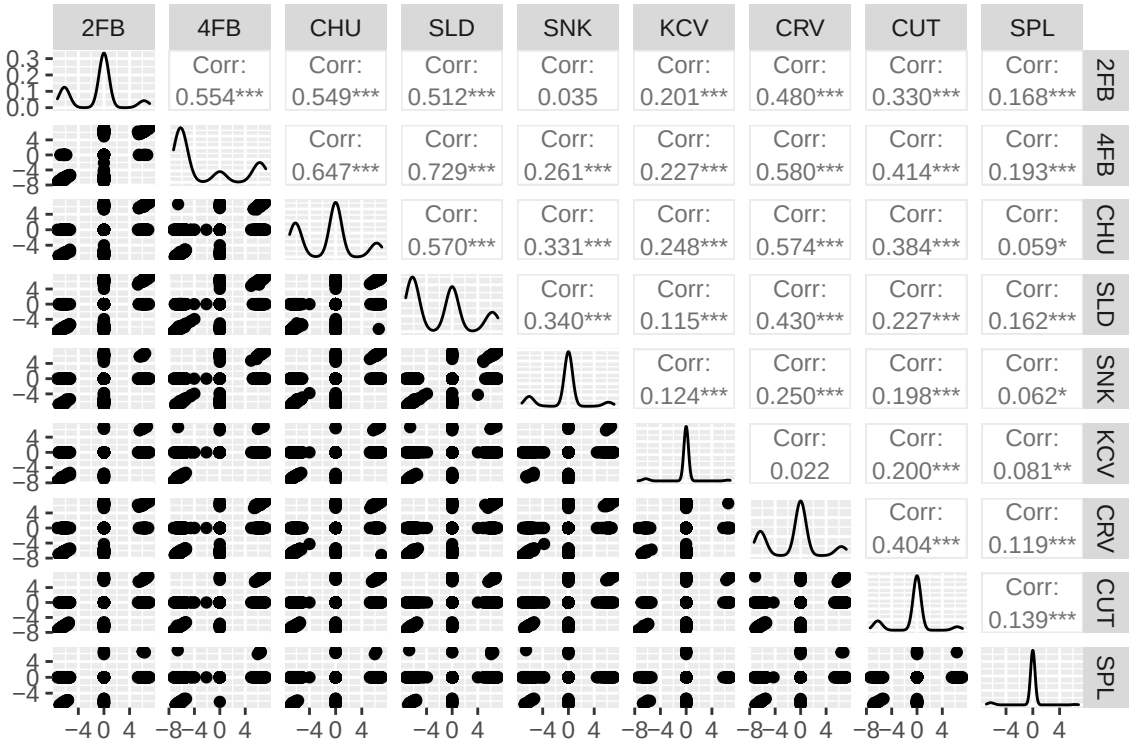

Figure 6: A Look At Release Extensions

The distributions are relatively normal, with two caveats: firstly, much of the density is for players who do not throw the pitch (more density towards 0 indicates that the pitch is less common), and secondly, the distributions are bimodal due to the righties and lefties. If the extension data did not include 0, and did not adjust the data by handedness, then it would be fairly normal. This can be seen in this histogram; the low correlations are partially a product of pitchers who do not throw a pitch having an extension of 0, and the difference between righties and lefties.

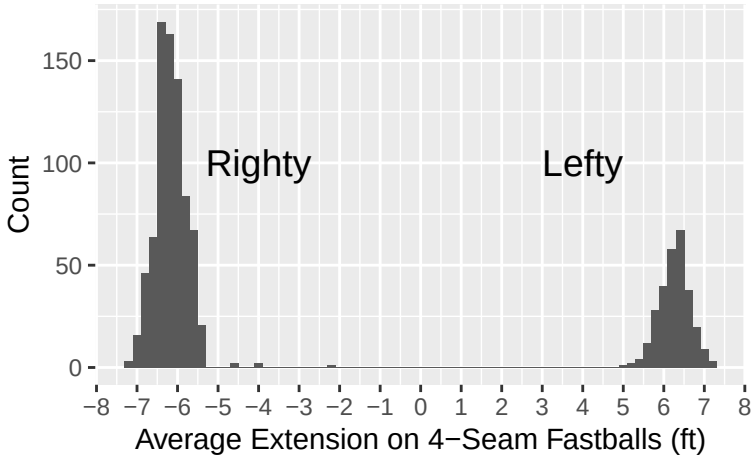

Figure 7: Comparing Righties to Lefties

2.4 Cluster Models

Now that we have assessed the data, the k-means clusters must be created. To determine the number of clusters to use, we will look at a scree plot and silhouette plot, and choose the cluster count that best approximates our data. Testing for every potential size of clusters from 1 to 100, 48 was chosen as the best cluster size. In the silhouette plot, 48 was the cluster count that produced the best average silhouette width. Looking at the scree plot, 48 seems like a reasonable choice that can be said to be close to the "elbow" of the plot.

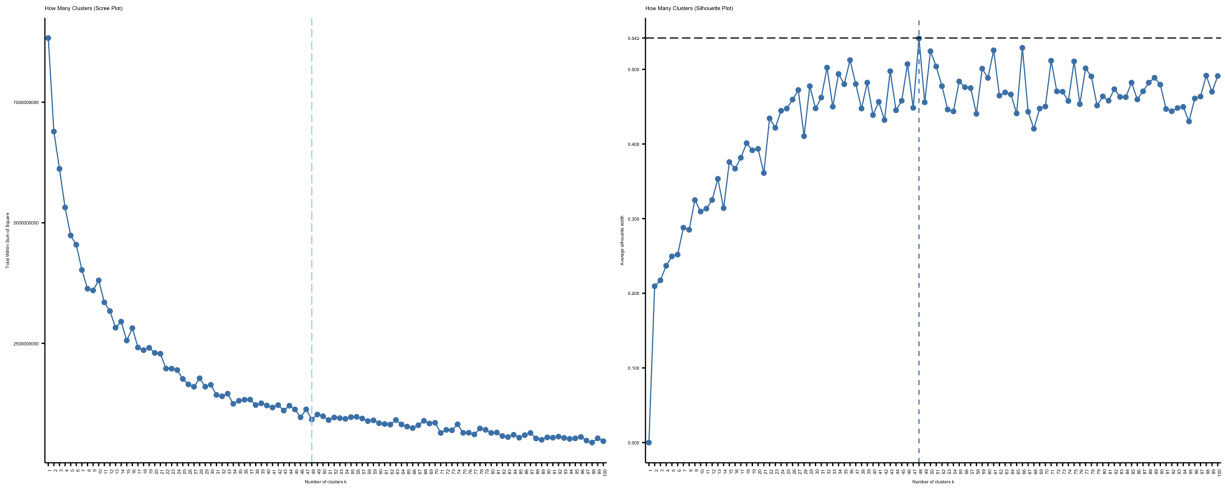

Figure 8: How Many Clusters?

Having now created the clusters, we employed a decision tree to help explain what is happening under the hood of the cluster generation, although the decision tree is only an approximate look at the clustering and does not fully account for every cluster nor perfectly classify every pitcher.

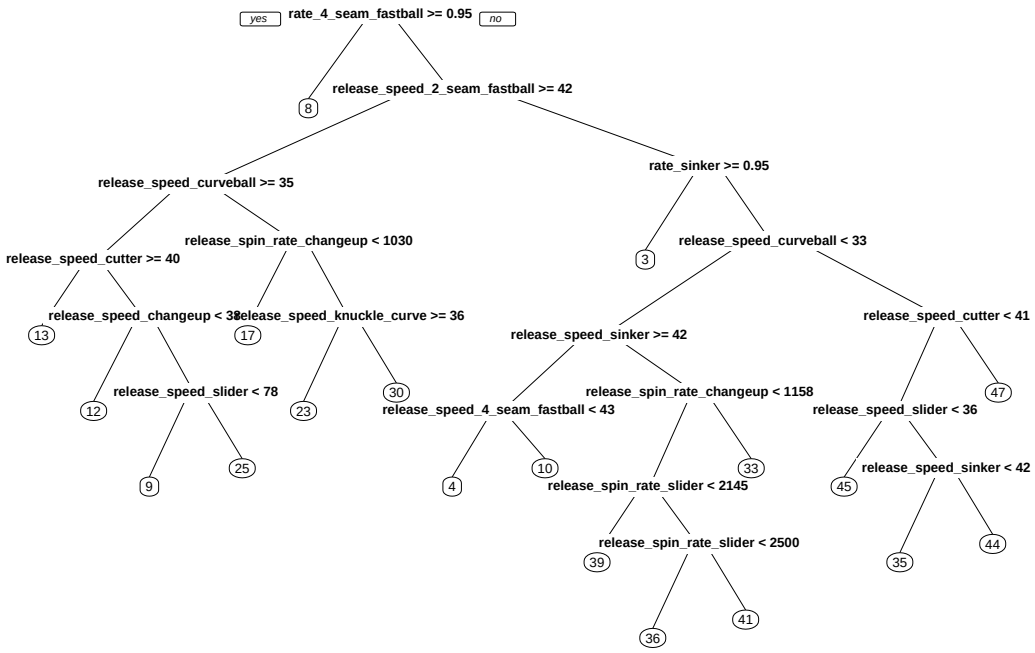

Figure 9: K-Means Decision Tree

2.5 Mixture Models

The same basic principles that are used in clustering apply to mixture model group creation. Like clustering, we want to test for some large number of potential mixture numbers and then select the mixture count that creates the strongest mixtures. To do this, we will use Bayes Information Criterion, looking for the lowest score (McLachlan & Peel 2000, 209). Below is a plot showing the BIC scores for different potential mixture algorithms. As mentioned above, each potential option provides different constraints to the potential mixture model distributions. We will use *EEI* (equal volume, round shape, spherical covariance) as our mixture algorithm with 38 mixtures.

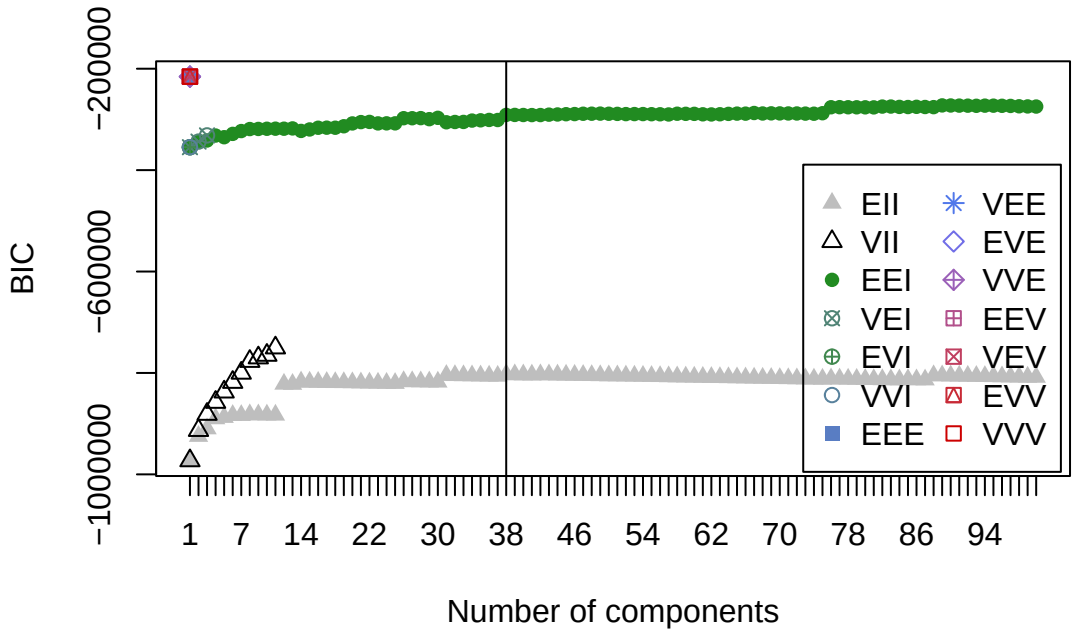

Figure 10: How Many Mixtures?

Looking at the now-created mixtures, the mixtures seem mostly focused on handedness, before any other attribute. Using a decision tree once again, we can see that the mixture model quickly separates the righties who throw those pitches into their own groups. It seems like the biggest splits are whether the pitcher is a righty who throws changeups or a lefty who throws sliders. Almost all the splits look to use righty and lefty extensions, grouping the pitchers by handedness and pitch types. The groups average about 32 pitchers per mixture, with a maximum of 90 and a minimum of 4.

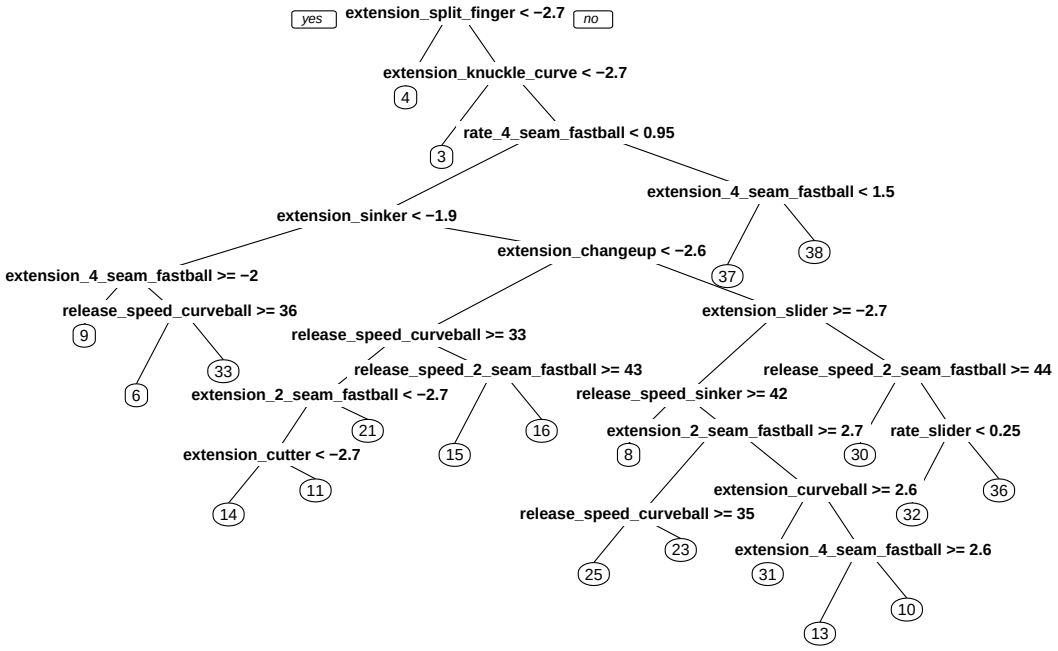

Figure 11: Mixture Model Decision Tree

2.6 Model Comparison

When comparing the two grouping algorithms, a few things stand out: firstly, the group sizes are far less neatly distributed for the k-means clustering algorithm than the mixture algorithm. Secondly, the k-means clustering algorithm relies far more on pitch types, rather than pitch handedness, focusing far more on the rates of individual pitches. From looking through each cluster and mixture, the vast majority of mixtures are entirely, or very close to entirely, right-handed or left-handed, whereas most clusters are a mix of both, although some clusters are more exclusively one or the other.

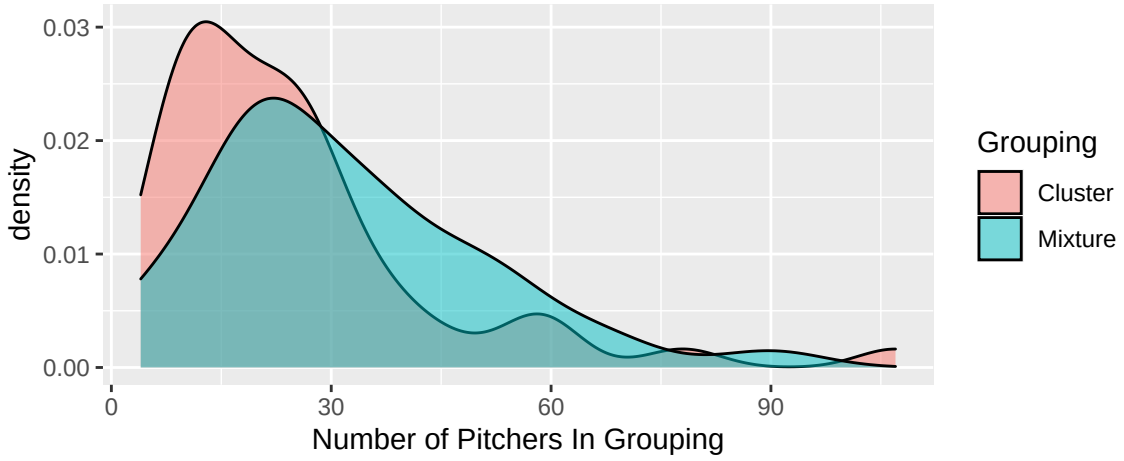

Figure 12: A Look At Group Sizes

2.7 Results Preview

To test the strength of the grouping algorithms, a variety of methods will be used to compare three different inputs: batter versus pitcher history, batter versus cluster history, and batter versus mixture history. Mathematically, batter versus pitcher history could be considered the same as treating each observation as their own mixture or cluster and helps to serve as a baseline model against which to compare the other models. To help account for the differences in sample sizes, a beta-binomial distribution, using Bayesian inference, will be used to help produce a future expected batting average. A beta-binomial distribution uses a prior that a given value follows a beta distribution and then takes the outcome of a binomial situation to create a posterior distribution. More formally, the beta-binomial distribution shows that if $X \sim \text{Beta}(\alpha, \beta)$, and we are given $Y \sim \text{Bin}(n, p)$, then $X|Y \sim \text{Beta}(\alpha + np, \beta + n(1-p))$. For predicting future batting averages, we will use a prior that $BA \sim \text{Beta}(14, 34)$, which may seem to be an arbitrary choice (and it is), but the $\text{Beta}(14, 34)$ distribution closely follows the batting average distribution of our dataset, as seen below for each batter in the dataset. It may seem like a $\text{Beta}(14, 34)$ distribution produces too high of a median batting average, but the batting average of this dataset is much higher than the overall MLB, as only players who meet a minimum at-bat threshold are kept in the dataset, which dramatically increases the standard of batters. To give an example of how the distribution would play out in this scenario, the binomial

information given would be a player's history against a pitcher, cluster, or mixture. Say a batter was 35/100 against a given pitcher's cluster. The posterior would now follow a $\text{Beta}(14+35, 34+100-35) = \text{Beta}(49, 99)$ distribution. This example can be seen below on the right, with the black line showing the new information of a .350 batting average in 100 at-bats.

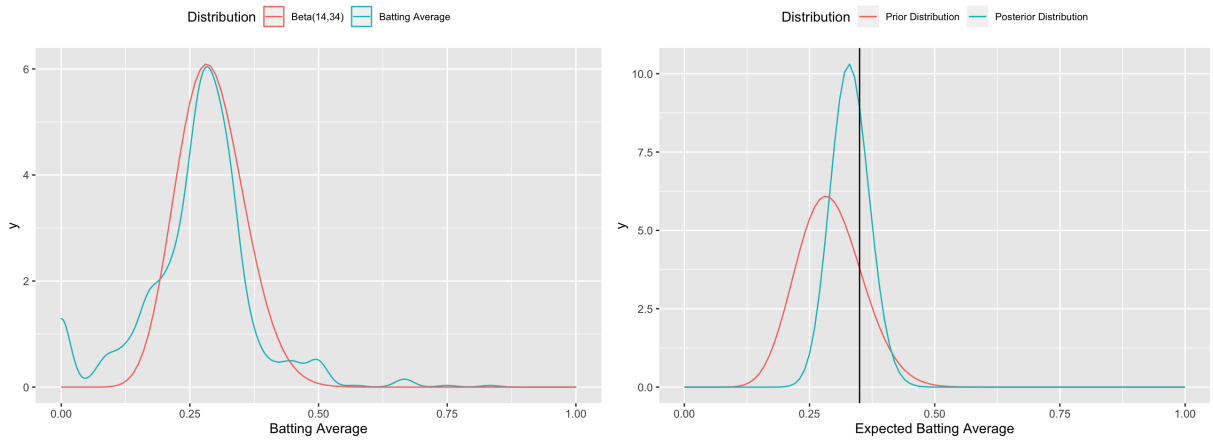

Figure 13: Bayesian Inference Using the Beta-Binomial Conjugate

To best understand the results, we will build three separate logistic regression models (using an 80-20 training-testing split), each meant to predict the outcome of an individual at-bat between a specific batter and pitcher. Each model will take in the pitcher's season batting average against (the batting average the pitcher allows to opposing hitters), the batter's season batting average, and the batter's history against the grouping method that the model is testing. For each of the different inputs, including the batting averages, the result of the plate appearance being modeled has been removed. From there, a variety of tests will be performed to compare the models.

3 Results

While there is no perfect way to compare logistic regression models, there are a variety of ways to use the models to compare the value of batter versus pitcher data, batter versus cluster data, and batter versus mixture data. The first method compares the three logistic regression models to a baseline model containing just knowledge about the batter's batting

average and pitcher's batting average against. Using an ANOVA likelihood ratio test, we examine the value of each additional predictor to the baseline model. In this case, the additional predictor would be the mean of the posterior beta distribution from the pitcher, cluster, and mixture. All three additional inputs are statistically significant (P<.05) compared to the baseline model, although the deviances are different. The additional batter versus pitcher input has a deviance of 4.339 and a P-value of .0372. The additional batter versus cluster input has a deviance of 11.93 and a P-value of 0.000552. The additional batter versus mixture input has a deviance of 30.01 and a P-value of 4.29e-08. This indicates that we can be confident that each of the three inputs provides statistically significant value to the model, and the differences in deviance and p-values suggest that the mixture model provides the most additional value, followed by the cluster, and then pitcher. A pseudo-R^2 value for the three models shows the mixture model as the most predictive. Finally, a Hosmer and Lemeshow goodness of fit test for each of the three methods finds that the mixture model produces the best fit, followed by the cluster model, and then the pitcher-only model.

This conclusion is confirmed by analyzing a model that has all three inputs, as well as the batter's batting average and pitcher's batting average against, and then looking at a variable importance metric of the various inputs into the model. The variable importance metrics find that the value of the batter-pitcher and batter-cluster variable is relatively negligible, whereas the value of the batter versus mixture history is quite high. This conclusion is supported by the z-scores in the model summary. The z-score of batter-pitcher is -0.05 (P-value: .96), batter-cluster is 0.74 (P-value: 0.46), and batter-mixture is 4.26 (P-value: 0.000021). In other words, when given access to information from all three models, a logistic regression model found that only the mixture model output was statistically significant.

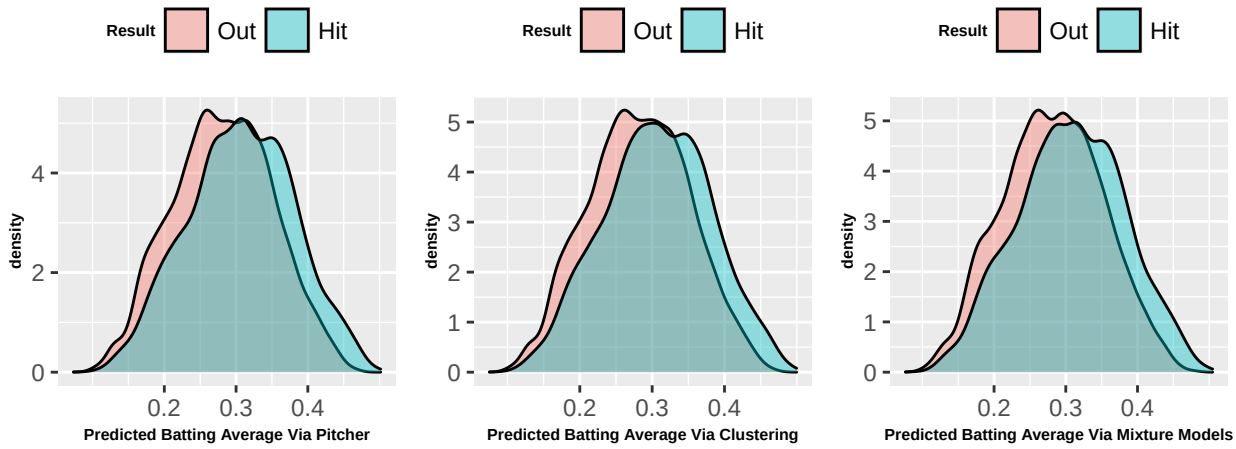

Figure 14: Predictions By Grouping Method

Finally, looking at the testing data can also provide a helpful method for comparing the various inputs. In the testing dataset, there were 7306 hits. Using just the posterior distribution from the batter-pitcher matchups, we would expect 7206 hits, from the batter-cluster matchups, we would expect 7267 hits, and from the batter-mixture matchups, we would expect 7261 hits.

4 Discussion

The results of the study all point to one conclusion: that grouping pitchers together provides predictive value in projecting the outcomes of at-bats. Furthermore, this method proves that there is more value than using just batter-pitcher history and that, between methods, using mixture modeling is a more effective grouping method than k-means clustering, at least on this dataset.

4.1 Model Implications

The results of this study illuminate the differences in the strengths and weaknesses of mixture models versus clusters. Mixture models clearly outperformed cluster models and more closely matched the baseball intuition that handedness should be an important factor in grouping pitchers. Both models outperformed the baseline of treating every observation

as their own individual group, which supports the notion that grouping via clustering and mixture models can improve sample size quality. As mentioned in the introduction, the mean at-bats between batter and pitcher in this dataset is 2.25, however, the mean at-bats between pitcher and cluster is 17.08 and the mean at-bats between pitcher and mixture is 20.96. The distributions can be seen below (note the different y-axis scales).

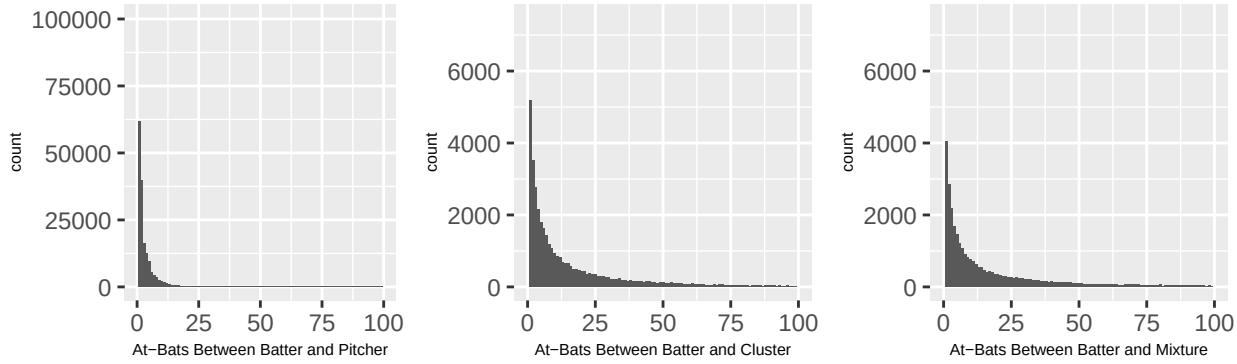

Figure 15: Distribution of At-Bats between Batter and Group

4.1.1 Future Work

While this data set produces mixtures in which almost every observation fits neatly into a specific group, one of the key strengths of the mixture models is the probabilistic nature of the groupings. While formulating the question for this study, one of the expected challenges was that past histories against a given mixture were going to need to be weighted by the likelihood of an observation belonging to each mixture. A future study with a dataset that produces less clear mixtures would provide another opportunity to test the strengths of mixture modeling versus k-means clustering, as k-means does not compute probabilistic groupings, and instead assigns each observation to the cluster of the closest centroid.

4.1.2 Limitations & Assumptions

There are a variety of limitations that should be mentioned for how this study can be extrapolated. First and foremost, while this study may have many theoretical implications, the study's findings can only fully be applied to this baseball dataset towards this baseball question. We cannot say that mixture models should always be preferred to k-means

clustering when grouping observations. This study also relies upon certain assumptions made by the researchers, such as setting a minimum number of plate appearances in a season to be able to stay in the dataset, which could affect the outcomes of the grouping methods.

The other major limitation is the lack of a clear test to compare non-nested logistic regression models. While there are many ways to test the fit of a logistic regression model or compare nested models, there is not a clear statistical method that definitively proves that one model is superior to another. Accordingly, this study relies on a variety of tests, each with its imperfections, to draw its conclusion.

4.2 Baseball Implications

The baseball implications of this study are extensive. Knowing a batter's history against a group of similar pitchers can help to make decisions from lineup construction to betting. If a manager is debating which batter to start when the first batter has good numbers against the pitcher and the second batter has good numbers against the entire pitcher's "type," this study would advocate for starting the second batter. Furthermore, the inverse of this example can also be true; if a manager is debating which reliever to call in, they could consider using the pitcher who belongs to the group that the batter has historically performed worst against. Similarly, if a gambler is looking for a team or player in daily fantasy, that historically has had good results against a pitcher's mixture, the findings of this study could provide an edge for the gambler.

4.2.1 Future Work

There are a variety of ways this study could be expanded. Firstly, having access to minor league data would dramatically improve the quality of the dataset. While there are 30 MLB teams, there were 206 minor league teams and 14 leagues below the MLB from 2015-2019. Having access to this dataset would exponentially increase the sample sizes in this dataset, add predictive value, and provide a large enough sample size to use an only forward-thinking dataset.

Another extension is to test the same premise beyond just simple batting averages.

While batting average is important, looking at more advanced statistics such as wOBA, wRC+, exit velocity, and more would either enhance or show the limitations of the study. Almost any statistic that is the result of a batter-pitcher matchup could be measured by the grouping method and could test the same premise. A simple extension could be to project on-base percentage, which, at a minimum, would allow us to use all plate appearances, not just at-bats.

4.2.2 Limitations & Assumptions

While this study has many benefits, there are also some clear limitations from a baseball perspective. Like the Raúl Ibañez example, pitchers are not always consistent over long periods. The data used was from over five years of data, and not all pitchers will remain the same, even over just a few years. Some pitchers will get injured and lose velocity, while other pitchers will learn a new pitch or change their repertoire. This study looks at the aggregate pitching distribution for a given pitcher, but there will be many situations where a pitcher looks very different in 2019 compared to 2015. As we continue to get more Statcast data, the best method would be to continue to expand our dataset, while more heavily weighting more recent information. Another potential limitation of this study is that the logistic regression models use season-long batting averages with a minimum plate appearance threshold; but at the low end of the plate appearance threshold, a player still might have a decent variation from their true talent level, meaning that some of the projections on the extremes of the dataset may produce particularly high or low projections compared to the true talent level of the player, although this would be consistent for all the model types and therefore not impact the comparison of the models. In other words, the projections from the model are probably not overly predictive on their own, but still provide a useful way of measuring the values of the grouping algorithms.

This study also relies on a potentially problematic statistical quandary: that the models are trained using some datapoints in which the data uses information from the future to predict the past. While the data has been adjusted to not include the tested at-bat in history, a data point in the testing set may rely upon a batter being 4 for 9 against a pitcher or group, in which all of those at-bats take place after the data point in question.

The data could be adjusted to be only forward-thinking so that each modeled at-bat only knows the history against the pitcher, but it would dramatically reduce the sample size. For two reasons, we decided to prioritize sample size: firstly, if we prove that the theory works, then it should still apply to future examples, even if there are not as many players with detailed histories against pitchers or groupings. Secondly, in an ideal world, this study would have more data, including minor league data, and then be entirely forward-thinking, but this study already uses the maximum public Statcast data fully available.

4.3 Conclusion

The findings of this study provide a compelling argument for the value of mixture models as a tool for clustering, particularly compared to a more traditional k-means model. From a baseball perspective, this study helps to show that saying a batter hits a *type of* pitcher well is far more predictive than arguing that a batter hits a specific pitcher well. This study also shows how an approach using a beta-binomial distribution and Bayesian inference can help to weight different sample sizes of batter outcomes. Overall, we hope that this study will provide an instructive and novel look at batter-pitcher matchups and provide more insight on potential grouping methods.

References

Cameron, D. (2013), 'The absurdities of batter/pitcher match-up numbers'.
URL: *https://blogs.fangraphs.com/the-absurdities-of-batterpitcher-match-up-numbers/*

Hartigan, J. A. & Wong, M. A. (1979), 'Algorithm applied statistics 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

Hennig, C. (2011), 'Clustering with the gaussian mixture model'.
URL: *http://www.homepages.ucl.ac.uk/ ucakche/presentations/ercimtutorial.pdf*

Lindbergh, B. (2021), 'Effectively wild: Take the money and pun'.

McLachlan, G. J. & Peel, D. (2000), *Finite mixture models*, Wiley.

Melnykov, V. & Maitra, R. (2010), 'Finite mixture models and model-based clustering', *Statistics Surveys* **4**(none).

Nagode, M. & Fajdiga, M. (2011), 'The rebmix algorithm for the multivariate finite mixture estimation', *Communications in Statistics - Theory and Methods* **40**(11), 2022–2034.

Tango, T. M., Lichtman, M. G. & Dolphin, A. E. (2014), *The book: Playing the percentages in baseball*, TMA Press.