Trouble with the Curve: Identifying Clusters of MLB Pitchers using Improved Pitch Classification Techniques

Michael A. Pane

May 1, 2013

Abstract

The PITCHf/x database, which records the location, velocity, and trajectory of every pitch thrown in Major League Baseball (MLB), has allowed the statistical analysis of MLB to flourish since its introduction in late 2006. Using PITCHf/x, pitches have been classified by hand, requiring considerable effort, or using neural network clustering and classification, which is often difficult to interpret. We use model-based clustering with a multivariate Gaussian mixture model and an adjusted Bayesian Information Criterion to determine the number of different clusters. We verify these results via cross validation, validation by prediction strength, and through visual inspection. Furthermore, we use our method to cluster pitchers into groups with similar characteristics via k-means clustering and the Fisher-wise criterion. Our method builds a strong foundation towards addressing many open MLB research questions, including preventing pitcher injury.

I would like to dedicate my Senior honors thesis to David and Allison Pane, Joseph Pane, and Kelly Ross. Without their constant support and encouragement, I would not have made it to where I am today.

Acknowledgements

I would like to thank Samuel L. Ventura, Rebecca C. Steorts, and Andrew C. Thomas for their constant daily support and mentoring throughout the completion of my thesis. I am incredibly grateful for the opportunity to learn under them and see my knowledge and work grow over the past year.

I also would like to thank Dan Rozenson and Harry Pavlidis for pointing me to "Baseball on a Stick" and the updated PITCHf/x database code.

Finally, I would like to thank my parents because I would not be where I am today without their constant love and support. I am forever grateful for them giving me the opportunity to attend Carnegie Mellon, and supporting my goals and dreams.

Contents

1 Introduction **10**

2 PITCHf/x and Pitch Types **13**
2.1 The PITCHf/x Database 14
2.2 Transformation of Data 14
2.3 Pitch Type Definitions 15
2.4 Classification Of Pitch Clusters 17

3 Cluster Analysis of Pitches **21**
3.1 Previous Analysis 21
3.2 K-means Clustering 23
3.3 Hierarchical Clustering: Average Linkage 25
3.4 Model-Based Clustering 26
3.5 Calculating the Number of Clusters 29
3.5.1 Adjusted Bayesian Information Criterion 30
3.5.2 Validation by Prediction Strength 31
3.5.3 Compare and Contrast Cluster Number Selectors 33
3.6 Stability of Cluster Membership 37

4 Assigning Pitchers to Subgroups Using Clustering Methods **39**
4.1 Motivating Applications of Pitcher Clustering 39
4.2 Data Manipulation 41
4.3 Pitcher Clustering: K-means 42

5 Conclusions and Future Work **47**
5.1 Conclusions 47
5.2 Future Work 48

List of Figures

2.1 Cliff Lee (2010 and 2011 seasons): Figure 2.1a (left) displays MLB's classification system developed via a neural network classification system. As before, there are obvious misclassifications, but our method's clusters are clearly defined and classified with no other obviously misclassified pitches. Figure 2.1b (right) displays the updated clustering method we introduce in Chapter 4. The classification algorithm separates and labels the two-seam (grey) and four-seam (red) fastballs by assigning the four-seam fastball to the fastest pitch with the least amount of back and top spin relative to Lee's other offspeed pitches. It is important to note that the clustering method we introduce in 3.4 breaks Cliff Lee's two-seam fastball into two separate pitches, but our classification system labels the clusters as the same pitches. This method appears to agree with the manually corrected data from Brooks Baseball. . . 20

3.1 The pitches thrown by Barry Zito. Figure 3.1 displays the MLB Advanced Media classification developed via a neural network system. There are obvious misclassifications: a small number of four-seam fastballs (which should be red) are classified as sliders (brown), as are some curveballs (black) and changeups (green). 23

3.2 Barry Zito (2010 and 2011 seasons): K-means with $k = 5$ has obvious misclassifications, and does not perform better than current methods. 25

3.3 Barry Zito (200 pitches): Average Linkage has obvious misclassifications because many pitches are clearly not near their assigned cluster, and does not perform better than current methods. 26

3.4 Displays Barry Zito's 2010 and 2011 seasons. Figure 3.4a displays the MBC model (using BIC_{adj}) and tightly clusters the pitches, as well as splits up the four-seam and sinker (light blue) clusters in an empirically sensible way. Figure 3.4b displays the MBC model using BIC as the cluster selector. The clustering is incorrect and thus the colors do not correspond to the labels in Table 3.3. 29

3.5 Mitchell Boggs (RP) 2010: 3.5a (left) displays MBC using BIC_{adj}, and Figure 3.5b (right) displays MBC using validation by prediction strength. Empirically, we observe that Boggs throws two pitches and BIC_{adj} correctly identifies the number of clusters. Note: In Figure 3.5a black corresponds to a four-seam fastball and red a slider. Figure 3.5b does not have any pitch labels. 34
3.6 C.J. Wilson (SP) 2010: Figure 3.6a (left) displays MBC using BIC_{adj}, and Figure 3.6b (right) displays MBC using validation by prediction strength. Empirically, we observe that Wilson throws 6 pitches as displayed in Figure 3.6a, and Figure 3.6b is incorrect. This example displays BIC_{adj} choosing a greater cluster number than validation by prediction strength, but still resulting in the correct number of clusters. 35
3.7 The pitches thrown by Jon Lester classified using model-based clusters, which shows two distinct clusters for curveballs (purple and black) corresponding to a change from the 2010 to 2011 seasons. The difference between the two clusters is the speed and horizontal break of the pitch. This is one example of how the MBC model can detect subtle but important pitch evolution differences. 36
3.8 The pitches thrown by Tim Wakefield classified using MBC. MBC can also detect and classify pitches with relatively higher variances, like the pseudo-spins for Tim Wakefield's knuckleball (orange), compared to the fastball (red) and curveball (black). 37
3.9 80% and 20% Stability: Percent of pitches in the same cluster in subset and full data set, across 20 replications of the procedure. For both the 80% and 20% subsets, pitchers have 80% or more of their pitches clustered in the same cluster on average. We find that this stability holds on sample sizes as low as 100 pitches. 38

4.1 Pitcher Clustering for Right handed pitcher: CH-Index Cluster Number Selector 42
4.2 Pitcher Cluster: Each data point is an individual pitcher, in an individual year between 2010-2012. The two clusters are referenced as red and black due to their lack of classification label. Section 4.3 analyzes and describes the relationships and results. A subset of the clustering variables are visualized in Figure 4.2, and are defined in Section 4.2. 46

List of Tables

2.1 Pitch type summary: RR corresponds to a right-handed pitcher with horizontal movement from left to right, LL corresponds to a left-handed pitcher with horizontal movement from right to left, RL corresponds to a right-handed pitcher with horizontal movement from right to left, and LR corresponds to a right-handed pitcher with horizontal movement from left to right. 18
2.2 Pitch type names with corresponding colors for Cliff Lee. 20

3.1 Pitch type names with corresponding colors for Barry Zito. 22
3.2 Pitch type names with corresponding colors for Barry Zito. 25
3.3 Pitch type names with corresponding colors for Figure 3.4a (Barry Zito Adjusted BIC). 29
3.4 Comparing the difference in number of clusters (pitch types) chosen by BIC_{adj} and validation by prediction strength. A positive number represents BIC_{adj} choosing a higher number of clusters (k) for a pitcher than validation by prediction strength. 33
3.5 Pitch type names with corresponding colors for Figure 3.6a. 35
3.6 Pitch types with corresponding colors for Jon Lester and Tim Wakefield. . . 36

Chapter 1

Introduction

Within the past 10 years, nearly every aspect of a Major League Baseball (MLB) game has become available as machine-readable data. Not only can we know the speed and movement of every pitch thrown during every game recorded, but we can calculate the spin and trajectory as the baseball leaves a pitchers hand. These variables are collectively called "PITCHf/x," developed by the company Sportvision. Perhaps more than any other technological contribution in baseball, the deployment of the PITCHf/x system has proven to be an invaluable resource to both teams and fans in their statistical analyses of baseball, both in the data's original form and as augmented by estimates of pitch spin parameters (known as "pseudo-spin").

End users of the system, particularly Brooks Baseball and MLB Advanced Media (MLBAM), have classified the pitches recorded into common pitch types, both by hand and using neural network classification methods. We hypothesize that the current method for classifying pitches in the database can be refined due to evidence from graphical exploration and previous research (Foster 2012). Improving pitch clustering and classification is a popular area of research in sports statistics for investigating pitcher clustering, predicting pitcher

injury, among other baseball studies.

Every pitcher has his own unique combination of pitch types, which have a set of relevant physical characteristics. Correctly identifying the type of every pitch is important in the analysis of any pitcher since a pitcher's pitch-type characteristics are what uniquely quantifies his performance on the field. The current database attempts to classify pitches with a separate neural network classification system for each pitcher developed by MLB-AM. However, specifics pertaining to the method, model, and training data used are not publicly available (Foster 2012). Most previous methods have either not significantly improved the current neural network system, or have not been publicly released. Brooks Baseball PITCHf/x classifications are considered the most accurate and publicly available, largely because their pitch classifications are assigned by variety of methods. Moreover, these classifications are thoroughly and manually cross-checked. Since we do not have pitch by pitch classification data and only a summary of each pitch's characteristics, Brooks Baseball serves as the best source when gauging accuracy at present.

We propose an alternative method for pitch clustering, using model-based clustering with a multivariate Gaussian mixture model (MBC). This method has several advantages: examining all pitch types, including identifying pitches with differing variances, following the evolution of a pitch across time, and illustrating pitches with similar characteristics. Even though ground truth is unknown, complicating evaluation of clustering methods, we show through empirical investigation, cross validation, and comparison to Brooks Baseball cluster summaries that our method is a significant improvement over the current database's neural network classification system.

In Chapter 2, we discuss the PITCHf/x database, data transformations, current clustering method being used, and our re-classification algorithm. In Chapter 3, we improve the current pitch classification system by comparing the current neural network classification results to those from various statistical clustering methods, such as k-means, hierarchical clustering, and MBC. We also address implementation of the MBC method as well as the selection of the clusters and the stability of this clustering method, and refining pitch clustering and classification. In Chapter 4, we attempt to cluster pitchers based on the refined pitch characteristics. Finally, we summarize our work in Chapter 5 and propose extensions to our proposed methods.

Chapter 2

PITCHf/x and Pitch Types

The PITCHf/x system was introduced in 2006 by Sportvision and has tracked pitches from all MLB ballparks. The raw PITCHf/x database contains information on each pitch thrown by every pitcher, including acceleration, velocity, and pitch position. We introduce variable transformations to calculate the spin of each pitch, and from previous research (Nathan 2007) we determine which of the the newly calculated spin variables are important in clustering pitches. Next, we describe the varying pitch types a pitcher may throw. Each pitch has different characteristics that are controlled by the way the pitcher grips and releases the baseball, and it is important to define all possible pitches. We conclude Chapter 2 by introducing a pitch classification algorithm. Since we do not have a ground truth data set, we introduce our own classification algorithm that label our pitch clusters as this is vital in evaluating the quality of our clustering method.

2.1 The PITCHf/x Database

PITCHf/x data is available from several sources, and code exists to automatically download, build, and maintain the database[1]. Our data subset consists of pitches thrown by roughly 900 pitchers in the 2010, 2011, and 2012 seasons; we exclude data from before the 2010 season due to reported inconsistencies within the PITCHf/x database. In order to efficiently and effectively handle our data we create multiple `R` scripts that take the raw PITCHf/x data and create a separate `R` data file for each pitcher and each year they pitched between 2010 and 2012 (e.g. `BarryZito.2010.RDtata`, `BarryZito.2011.RData`, and `BarryZito.2012.Rdata`).

2.2 Transformation of Data

The raw database contains trajectory information on each pitch. This includes acceleration and velocity, though not the spin of each pitch which is useful in pitch classification because it can help distinguish between different pitch types. Some spin variables can be estimated using physics and a number of simplifying assumptions (Nathan 2007); the name "pseudo-spin" is given to these quantities due to this. We use the approaches from Nathan (2007) to calculate pseudo-spin in this work.

There are several important variable definitions that we use throughout our paper and in our figures. The pitch types mentioned below are defined in Section 2.3:

- The pitch's **start speed** is measured in miles per hour (mph) at the release point. This is measured using radar guns and is commonly acknowledged as the speed of the

[1]The data can be downloaded from the following websites: `http://www.wantlinux.net/2009/10/pitch-fx-data-with-pitch-type/` and `http:/www.MLB.com.` Furthermore, multiple Python scripts from `http://sourceforge.net/projects/baseballonastic/` were used to download the data into an SQL database.

pitch.

- The **back spin** of the baseball is measured in radians per second (rps). A positive number represents back spin. Most fastballs have back spin, while off-speed pitches have a tendency to have more "top-spin", or negative back spin.

- The **side spin** of the baseball is measured in radians per second (rps). A positive number represents left-to-right spin, or a left-handed pitcher's curveball or slider. A negative number represents right-to-left spin, or the direction of a right-handed pitcher's curveball or slider.

2.3 Pitch Type Definitions

Before clustering pitches, it is important to define the various pitch types a pitcher can possibly throw, as well as their characteristics. Each pitch has different characteristics such as speed and movement that are controlled by the way the pitcher grips and releases the baseball.

The most common type of pitch thrown is a fastball, but fastballs are typically broken up into four different types of subcategories: Four-seam fastball, two-seam fastball, sinker, and cut fastball. In terms of the way the pitch behaves, the four-seam fastball is more or less a "standard" fastball. It is generally the fastest pitch thrown, and it usually has very minimal vertical or horizontal movement. Unlike other pitches, almost every pitcher throws a four-seam fastball. In contrast, a two-seam fastball is slightly slower and tends to break slightly downward and toward the handedness of the pitcher (e.g., to the right for a right-handed pitcher). It can be seen as a tradeoff between velocity and movement. As most pitches, the type of movement on a two-seam fastball does vary from pitcher to pitcher. A sinker is

very similar to the two-seam fastball although a sinker tends to have slightly more downward movement than the two-seam fastball. Between the four and two-seam fastballs/sinker, these are the most common types of fastballs thrown.

The fourth type is a cut-fastball or "cutter". A cutter breaks in the opposition direction of a pitcher's handedness (e.g., to the left for a right-handed pitcher). This tends to be slightly slower than a two-seam fastball but also has slightly more movement.

Any pitch that is not a fastball is generally called an "off-speed pitch". Most pitchers throw various combinations of off-speed pitches, and very few throw all of the different varieties. One type of off-speed pitch is a "changeup". A changeup looks similar to a four-seam fastball from the batter's perspective as it has a similar trajectory, but the ball is traveling substantially slower causing the batter to swing too soon. Most changeups move relatively straight with very little movement, but one variety of the changeup is a "circle changeup". The circle changeup is named for its unique grip and is slow like a standard changeup but moves horizontally in the same direction as the pitcher's handedness.

Despite being named a fastball, the split-finger fastball or "splitter" is not classified as a fastball but rather an off-speed pitch. The splitter has its own distinctive grip. It travels at a speed much slower than a standard fastball, but it breaks sharply down, causing the batter to swing over top of the ball. This pitch is rarely thrown and also is thought to be harmful to a pitcher's arm.

One of the most common off-speed pitches, the "curveball" is slow (like a changeup) but breaks sharply, usually downward but it can also break slightly in the opposite direction of the pitcher's handiness. The classic curveball is called a "12 to 6 curveball" (referring to

the face of a clock), although many right handed pitchers throw a curveball that is more accurately described as "1-to-7" or "2-to-8" instead (same for left-handed pitchers, just the opposite side of the clock).

Another common pitch, the "slider", is usually defined as halfway between a fastball and a "1-7" or "2-to-8" curveball. The speed of a slider is slightly faster than a curveball and consequently breaks less than a curveball. At times it can also have greater horizontal movement than a curveball.

In fact, there is a continuum of pitches going from four-seam fastball to cut fastball to slider to curveball. Some pitchers throw pitches that would be best described as in between two pitches defined above. For example, sometimes a pitcher can throw a "slurve" (between a curveball and slider), or while others can throw fastballs with more movement. These pitch names are merely discrete categorizations. At times, pitchers believe they are throwing one pitch, but the speed and movement say they are throwing a different pitch. Table 2.3 summarizes the characteristics of each pitch. This table shows that there is a tradeoff between speed and movement (vertical and horizontal break) across the continuum of pitches.

2.4 Classification Of Pitch Clusters

In order to have a fully automated pitch clustering and classification system, we propose a simple classification algorithm, which improves pitch clustering by assigning sensible pitch labels to the respective clusters from our heuristic. Although ground truth is unknown, we make comparisons with the current classification labels in the PITCHf/x database, Brooks Baseball classifications, and graphical visualization to determine how well our method performs. We find that our classification algorithm has many strengths, including consistent

Table 2.1: Pitch type summary: RR corresponds to a right-handed pitcher with horizontal movement from left to right, LL corresponds to a left-handed pitcher with horizontal movement from right to left, RL corresponds to a right-handed pitcher with horizontal movement from right to left, and LR corresponds to a right-handed pitcher with horizontal movement from left to right.

Pitch Name	**Speed**	**Vertical Break**	**Horizontal Break**
Four-seam fastball	Fastest	None	None
Two-seam fastball	Fast	Down (slight)	RR/LL (slight)
Cut fastball	Medium-Fast	Down (slight)	RL/LR (moderate)
Split-finger fastball	Medium	Down (a lot)	None
Changeup	Slow	Down (slight)	None or RR/LL (slight)
Curveball	Slow	Down (a lot)	None or RL/LR (moderate)
Slider	Medium/Slow	Down (moderate)	RL/LR

performance, labeling multiple clusters as the same pitch type when appropriate (such as the Jon Lester curveball evolution from 2010 to 2011, referenced in Section 3.6 and visualized in Figure 3.7), and correcting mistakes from our clustering algorithm (such as Cliff Lee's pitches, visualized in Figure 2.1). Our pitch classifications appear to have results similar to those from Brooks Baseball (which is the closest to the "truth" against which we can compare).

Our algorithm classifies and names a key subset of pitches: four-seam fastballs, two-seam fastballs, sinkers, change-ups, cut-fastballs, sliders, curveballs, and knuckleballs. It splits up the three dimensional field of three important variables we introduced in Section 2.2 and use in our cluster analysis in Chapter 3: start speed, top spin, and side spin. The algorithm assigns pitch labels based on where each cluster mean falls. For each classification, the algorithm begins by assigning the cluster mean with the highest starting velocity as a four-seam fastball. For each additional cluster, the algorithm goes through a series of constraints that are derived from each cluster's mean and variance to determine the label for each cluster. For example, if a cluster mean has the same side spin direction as the four-seam fastball

then it checks if the speed differs from the four-seam fastball by more than 6 MPH and if the side spin varies less than 60 rotations per second from the four-seam fastballs. If it does, it assigns the cluster as a change-up. If not, it determines if the side or back spin difference from the four-seam fastball is greater. If the side spin difference is the greater of the two, it assigns the cluster as a two-seam fastball; if the back spin difference is larger, it assigns the cluster as a sinker. The rest of the classification algorithm follows similar decision processes.

We evaluate how well our classification algorithm performs by taking a sample of various starting and relief pitchers with differing pitch repertoires. We find after analyzing and comparing the results to Brooks Baseball and the neural network classification system by comparing the pitch classifications for each pitcher, in 46 of the 50 scenarios, our classification algorithm empirically appears correct and has similar classification results as Brooks Baseball. Examples of two pitchers that the algorithm does not classify correctly are Barry Zito and Derek Lowe. It interchanges their sinker and four-seam fastball clusters because in the data sample, the measured speed of the sinkers are greater than those for the four-seam fastball, a characteristic that is not commonly true for other pitchers.

In Figure 2.1a, we visualize Cliff Lee's MLB-AM neural network classification compared to our method. Of particular note, our method separates and labels the two-seam and four-seam fastball by assigning the four-seam fastball to the fastest pitch with the least amount of back and top spin relative to Lee's other off-speed pitches. Our method splits the two fastballs similar to Brooks Baseball, but Brooks Baseball instead labels the two-seam fastball as a sinker. Both pitches are very similar in their characteristics. It is important to note that our clustering method breaks Cliff Lee's two-seam fastball into two separate pitches, but our classification system labels the clusters with the same pitch type. Thus, our classification algorithm "fixes" a mistake in our clustering algorithm. The neural network classification

has obvious misclassification in the slider cluster with pitches labeled as changups, curveballs, and cut-fastballs. Our method clearly improves both the clustering and classification of Cliff Lee compared to the neural network classification, and also can resolve mistakes in our clustering algorithm by merging two clusters into the same level.

Table 2.2: Pitch type names with corresponding colors for Cliff Lee.

Pitch Name	Four-Seam	Two Seam	Cut Fastball	Changeup	Curveball	Slider
Color	Red	Grey	Blue	Green	Black	Brown

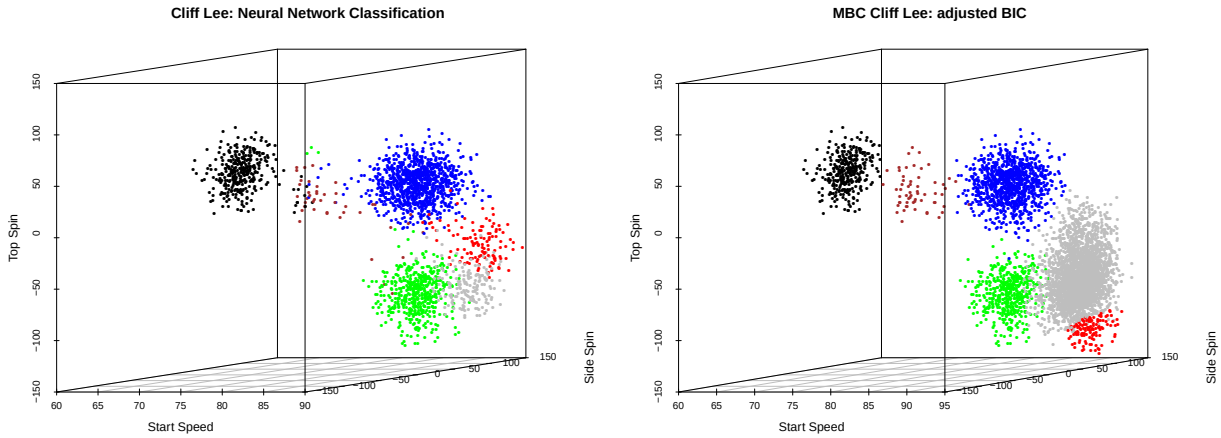

Figure 2.1: Cliff Lee (2010 and 2011 seasons): Figure 2.1a (left) displays MLB's classification system developed via a neural network classification system. As before, there are obvious misclassifications, but our method's clusters are clearly defined and classified with no other obviously misclassified pitches. Figure 2.1b (right) displays the updated clustering method we introduce in Chapter 4. The classification algorithm separates and labels the two-seam (grey) and four-seam (red) fastballs by assigning the four-seam fastball to the fastest pitch with the least amount of back and top spin relative to Lee's other off-speed pitches. It is important to note that the clustering method we introduce in 3.4 breaks Cliff Lee's two-seam fastball into two separate pitches, but our classification system labels the clusters as the same pitches. This method appears to agree with the manually corrected data from Brooks Baseball.

Chapter 3

Cluster Analysis of Pitches

Improving pitch clustering and classification enhances our analysis when we approach pitcher clustering and ultimately predicting and preventing pitcher injury. Since ground truth is not available and pitchers throw different subsets of pitches with varying characteristics, we develop a clustering method, and then run the classification algorithm developed in Section 2.4 to label our clusters. We propose improving the current pitch clustering and classification system, and we compare the current PITCHf/x and neural network classification methods to (i) k-means, (ii) hierarchical clustering, (iii) and model-based clustering with a multivariate Gaussian mixture model (MBC). We conclude that MBC combined with our pitch classification algorithm yield an improvement over the current neural networks classification method.

3.1 Previous Analysis

We illustrate issues in the current neural network method used in the PITCHf/x database via a motivating example. Figure 3.1 plots top spin, side spin, and start speed for every pitch thrown by Barry Zito in 2010-2011. These pitches have been classified with the system

developed via a neural network classification system for each pitcher developed by MLB-AM. Specifics pertaining to the method, model, and training data used are not publicly available (Foster 2012). Brooks Baseball only publishes their classified pitches and does not make their data or neural network method publicly available. Hence, checking the validity of our method versus theirs in a statistically sound manner is challenging.

Based on Figure 3.1 and according to Brooks Baseball, we hypothesize that Barry Zito threw 5 types of pitches in 2010–2011. Based on our graphical observations and Brooks Baseball, the five pitches are a four-seam fastball, sinker, slider, curveball, and changeup. Moreover, we speculate that the pitches in the two-seam fastball cluster should instead be classified as sinkers since Brooks Baseball classifies these pitches as such. The neural network classifies Brooks Baseball's sinker cluster as two-seam fastballs. Using the method from Section 2.4, we choose to classify this as a sinker as well, in agreement with Brooks Baseball and against MLB-AM, suggesting that these problems may be easily fixable.

In addition, the neural networks classification appears to have obvious misclassifications. Four-seam fastballs are sometimes classified as curveballs, while curveballs and changeups are often classified as sliders, which are clearly labeled in the incorrect clusters in Figure 3.1 because many pitches are clearly not near their assigned cluster. Through further empirical investigation these obvious clustering appear in the majority of pitch classifications in the current PITCHf/x database.

Table 3.1: Pitch type names with corresponding colors for Barry Zito.

Pitch Name	Four Seam	Two Seam	Sinker	Changeup	Curveball	Slider	Intentional
Color	Red	Grey	Light Blue	Green	Black	Brown	Yellow

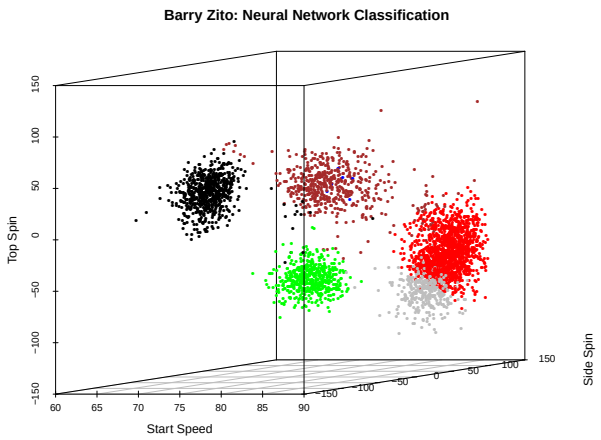

Figure 3.1: The pitches thrown by Barry Zito. Figure 3.1 displays the MLB Advanced Media classification developed via a neural network system. There are obvious misclassifications: a small number of four-seam fastballs (which should be red) are classified as sliders (brown), as are some curveballs (black) and changeups (green).

3.2 K-means Clustering

We explore a variety of methods for clustering pitches. We begin with k-means, a cluster analysis method that partitions n observations into k clusters, where each observation belongs to the cluster with the nearest mean.

We calculate the Fisher-wise criterion (CH-index) (Calinski and Harabasz 1974) to determine the number of clusters to use along with k-means. The CH-Index is defined as:

$$\frac{\mathrm{B}(k)/(k-1)}{\mathrm{W}(k)/(n-k)},$$

where $\mathrm{B}(k)$ is the between-cluster variation and $\mathrm{W}(k)$ is the within-cluster variation. The

between-cluster variation and within-cluster variation are defined respectively to be

$$\mathrm{B}(k) = \sum_{k=1} n_k ||\bar{x}_k - \bar{x}||^2 \quad \text{and} \quad \mathrm{W}(k) = \sum_{k=1} \sum_{C(i)=k} ||x_i - \bar{x}_k||^2,$$

where n_k is the number of pitches for the kth pitcher, x_k is the kth pitch, $\bar{x}$ is the mean of each cluster, and k is the number of clusters.

Ideally, we want a larger between-cluster variation and a smaller within-cluster variation. We want to find the value of k that maximizes $\mathrm{B}(k)$ and minimizes $\mathrm{W}(k)$ simultaneously.

We use the optimal value of k from the CH-index to run the k-means algorithm, which iterates between two steps: determining the closest center to each point, and calculating the new means after the points are assigned to their respective clusters. The process repeats until nothing changes in the latest iteration (Hastie et al. 2009). We show a motivating example in Figure 3.2. We run k-means with 5 clusters (determined by CH-index) and do not see an improvement from the initial PITCHf/x classification system. There are many fastballs/sinkers (black and light blue) that are classified as changeups (red), sliders (green) classified as curveballs (blue), and sliders classified as fastballs. Through our initial investigation we determine k-means does not perform well on many other pitches because many pitches are clearly not near their assigned cluster, and thus conclude that k-means does not improve the PITCHf/x neural networks classification. We investigate CH-index as a selector for the number of clusters in Section 3.5 and determine it is not optimal in comparison to other methods.

Table 3.2: Pitch type names with corresponding colors for Barry Zito.

Pitch Name	Four Seam	Two Seam	Sinker	Changeup	Curveball	Slider	Intentional
Color	Red	Grey	Light Blue	Green	Black	Brown	Yellow

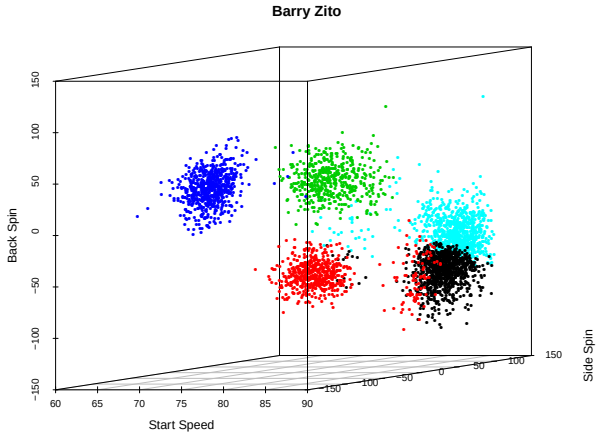

Figure 3.2: Barry Zito (2010 and 2011 seasons): K-means with $k = 5$ has obvious misclassifications, and does not perform better than current methods.

3.3 Hierarchical Clustering: Average Linkage

We next investigate another cluster analysis method, hierarchical clustering, which builds a hierarchy of clusters either from a bottom-up or top-down approach. Average linkage is defined with the following equation:

$$\frac{1}{N_A N_B} \sum_{i \varepsilon A} \sum_{i' \varepsilon B} d_{ii'},$$

where N represents the number of observations in clusters A and B, and d represents the individual pairwise dissimilarities using Euclidean distance in the space of variables considered between two points/pitches (Hastie et al. 2009).

Figure 3.3 illustrates average linkage hierarchical clustering on Barry Zito's pitches, and we do not see an improvement from the initial PITCHf/x classification system. A collection of changeups are classified in the same cluster as four-seam fastballs, and sinkers are classified as sliders. After checking this on a number of other pitchers, we determine that hierarchical clustering is not an improvement over previous methods because many pitches are clearly not near their assigned cluster.

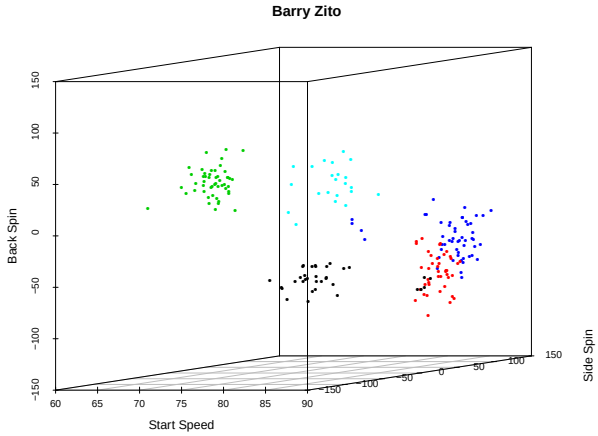

Figure 3.3: Barry Zito (200 pitches): Average Linkage has obvious misclassifications because many pitches are clearly not near their assigned cluster, and does not perform better than current methods.

3.4 Model-Based Clustering

The last clustering method we explore is model-based clustering with a Gaussian mixture model (MBC). We illustrate multiple motivating examples comparing MBC to the neural networks method, assess an optimal cluster number selector, and test the stability of the method. Using these criterion, we conclude that MBC, paired with our pitch classification

algorithm, is an improvement over the current neural networks classifications system.

Mixture models for clustering rely on a straightforward generative premise: there is a series of simple probabilistic models for how an event can be generated and a weight on which model is used to generate the observation. This description lines up directly with pitcher intent: while we as observers may not know what type of pitch is intended, the pitcher himself makes a choice of a specific pitch type (fastball, slider, curveball, etc) with a basic profile: a grip and arm motion that gives the ball a desired speed, spin and trajectory.

Implementing a multivariate Gaussian model for any particular pitcher profile makes intuitive sense. Each pitcher's coordinate has a mean value; for example, a typical four-seam fastball could have an initial velocity of 95 mph, a back spin of 100 rps, and side spin of 10 rps. The resulting pitch is noisy, both in the pitcher's delivery and due to other external factors, such as wind. Such noise can affect multiple pitch characteristics simultaneously. The resulting noisy pattern in multiple dimensions is a hyper-ellipsoid. In Figures 2.1b, 3.4a, 3.5a, 3.6a, 3.7, and 3.8, we visualize the MBC clustering results in three dimensions for a variety of pitchers and observe that the clusters are ellipsoids and the MBC clustering performs optimally. One particular advantage to this approach is that we can detect when two clusters overlap, since geometry, and not just proximity, is an important factor in MBC.

Given a set of Gaussian clusters and weights, MBC determines the probability that each observed pitch belongs to a given cluster by calculating the relative probability density for the pitch if it were a member of each cluster, factoring in each cluster's relative weight. Most pitchers, for example, throw more fastballs than other off-speed pitches, and this is taken into account directly via the weights. In our pitch classifications, we declare a pitch to belong to the class with the highest probability under the model.

We use the `mclust` library in the `R` programming language to perform our clustering operations, with some modifications that we describe to account for additional information. Given a pre-selected number of clusters, we use the Expectation-Maximization (EM) algorithm to calculate the maximum likelihood estimates (MLEs) for each cluster location, shape, and weight. Once we run this across a range of cluster counts, we use the Bayesian Information Criterion (BIC) to determine the optimal number of clusters in the model, which attempts to maximize the likelihood of the data while penalizing excessive numbers of parameters.

Empirically, MBC creates clusters of pitches that are more tightly confined than current pitch classifications, suggesting there are very few curveballs, changeups, or sliders misclassified. A major concern is the difference between the four-seam fastball and the sinker/two-seam fastball clusters. We expect two-seam fastballs/sinkers to have a slightly slower start speed than four-seam fastballs because a four-seam fastball is known to be a pitcher's fastest pitch, but this is not the case with the PITCHf/x neural networks classification method. However, MBC finds the velocity to be an important factor between the two clusters and splits them accordingly.

Figure 3.4a shows the MBC results for the pitches of Barry Zito in 2010 and 2011. We assess the clustering results and observe the red cluster is the faster pitch and light blue cluster is the slightly slower pitch with vertical spin closer to Zito's other off-speed pitches. MBC's potential four-seam fastball cluster (red) has the smallest number of pitches, but other sources suggest that Zito threw his four-seam fastball most often, leading us to suspect that while the labels (from Section 2.4) may need adjusting, the proper clusters are still being detected. This also may indicate there is not much difference between Zito's four-seam and two-seam/sinker pitches from the batter's perspective. To further support subtle

advantages of MBC over the neural networks classification, we evaluate the overall stability of our clustering method in Section 3.6.

Table 3.3: Pitch type names with corresponding colors for Figure 3.4a (Barry Zito Adjusted BIC).

Pitch Name	Four Seam	Two Seam	Sinker	Changeup	Curveball	Slider	Intentional
Color	Red	Grey	Light Blue	Green	Black	Brown	Yellow

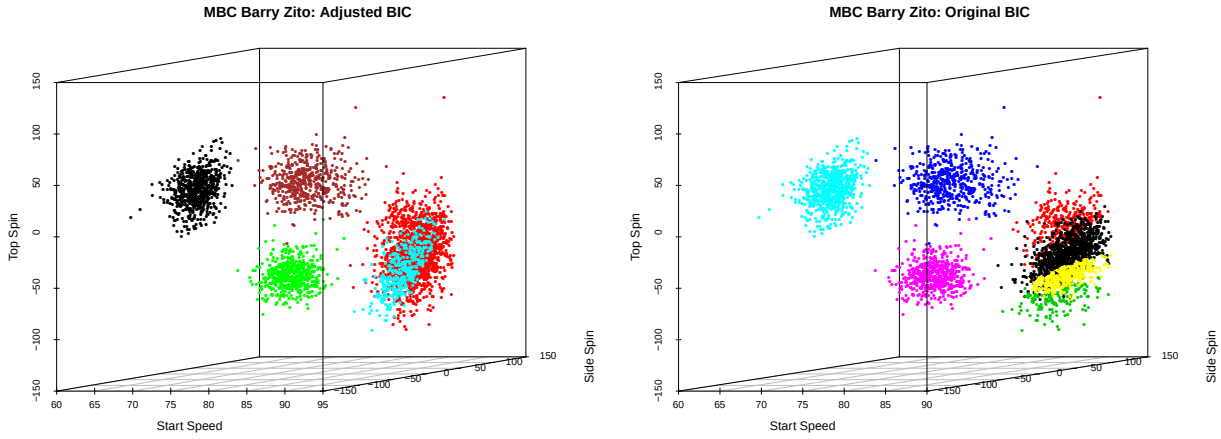

Figure 3.4: Displays Barry Zito's 2010 and 2011 seasons. Figure 3.4a displays the MBC model (using BIC_{adj}) and tightly clusters the pitches, as well as splits up the four-seam and sinker (light blue) clusters in an empirically sensible way. Figure 3.4b displays the MBC model using BIC as the cluster selector. The clustering is incorrect and thus the colors do not correspond to the labels in Table 3.3.

3.5 Calculating the Number of Clusters

In its original form, MBC tends to overestimate the number of clusters for the pitchers we have tested. On inspecting the clusters produced, it is clear that the method is favoring relatively "thin" clusters or many small clusters (as visualized in Figure 3.4b), with high

internal correlations between variables, which is highly unrealistic for the physical examples we consider. Limiting the model to a smaller number of pitches is not generally feasible since the number of correct pitches varies from pitcher to pitcher. While manual inspection is possible for individual pitchers, it would be far more preferable to automate the method to remove this issue for all pitchers. We explore three possible solutions: CH-index (introduced in Section 3.2), an adjusted Bayesian Information Criterion (BIC_{adj}) term (3.5.1) and cluster validation by prediction strength (3.5.2). Through initial investigation we rule out CH-index due to its inconsistencies and obvious underperformance relative to the other two proposed methods. We ultimately determine BIC_{adj} performs optimally amongst the methods we test.

3.5.1 Adjusted Bayesian Information Criterion

We develop our own criterion for choosing the number of clusters, the adjusted Bayesian Information Criterion (BIC_{adj}). Since we observe that most pitch clusters are close to spherical and our prior knowledge suggests that flat ellipsoidal clusters are unlikely, we are motivated to constrain the creation of clusters to have low intra-cluster correlations between the three variables of interest. Currently, each cluster k has three parameter sets: μ_k, the cluster mean; σ_k, the standard deviation of each dimension; and Σ_k, the intra-cluster correlation matrix. Ideally, the off-diagonal terms of Σ_k should be kept small in absolute value, indicating little correlation exits.

To account for this, we develop an additional penalty term for the current BIC formula that adds a value proportional to each intra-cluster correlation term, penalizing for high correlations. Using BIC_{adj}, if the model finds $k = 6$ has high intra-cluster correlations, compared to $k = 5$ which has low intra-cluster correlations, then the correlation penalty term will be large, and BIC_{adj} will be smaller and choose $k = 5$. Since BIC does not factor in a penalty

term for high intra-cluster correlations, it chooses $k = 6$. We choose our k based off of the minimum BIC or BIC_{adj}.

$$\begin{aligned}\text{BIC}_{\text{adj}} = &-2\log(f(Y|\hat{p})) - 2\lambda \sum_i log(f(c_i)) \\ &+ [k \cdot (j + j(j-1)/2 + (k-1)] \cdot \log(n), \end{aligned} \tag{3.1}$$

where $f(Y|\hat{p})$ is the probability of the parameters given the data, $f(c_i)$ is the sum of the upper off diagonal elements of Σ_k, and j is the number of clustering variables ($j = 3$ in our work).

The additional penalty term has a tuning parameter λ. To determine the best value of λ we use cross validation, using all of the 2010 data as training data and all of 2011 as test data. We test values of 0.25, 0.5, 1, 1.5, and 2. For each of these values, we run MBC on the 2010 data set and determine 0.5 is the optimal lambda value by empirically observing a subset of clustering results and finding $\lambda = 0.5$ results in the lowest number of pitch misclassifications. We then validate the clustering results using 2011 as validation data and determine that $\lambda = 0.5$ is the optimal weight for the additional correlation penalty term in BIC_{adj} using the same method.

3.5.2 Validation by Prediction Strength

In addition to the adjusted BIC method, we investigate choosing the number of clusters via cluster validation by prediction strength (Tibshirani and Walther 2009). We first split the data into two separate subsets: training and validation data. After experimentation with

various subset sizes we split the data equally among the two groups due to no obvious advantage in performance, and because a 50% split performs optimally for pitchers with lower sample sizes.

We cluster the test and training data into k clusters via MBC. Next, we calculate how well the training set clustering means predict co-membership in the test set. Co-membership is defined as the total number of points x_i in the test data set that fall into the same cluster as when they are clustered via MBC. If it is not in the same cluster, it receives a 0. Cluster membership for the test data set is determined by finding the training data cluster mean with the minimum euclidean distance for each point in the test data. We then calculate prediction strength for each possible cluster number. We define prediction strength by:

$$\frac{1}{N_{test} \cdot (N_{test} - 1)}(M \cdot k),$$

where N_{test} is the number of pitches in the test data, M is the number of successful co-memberships ($M = 1$), and k is the number of clusters.

We calculate the prediction strength statistic defined above for each of the possible total cluster numbers (k). The minimum and maximum number of possible pitches we assume an individual pitcher can throw is 3 to 7 and is therefore is our range of k. This is held consistent throughout our analyses. We then determine which value of k maximizes prediction strength, and choose the selected k as our clustering number for each individual pitcher.

3.5.3 Compare and Contrast Cluster Number Selectors

Next, we evaluate if the BIC_{adj} or the cluster validation by prediction strength is the better cluster selector and determine if either method outperforms BIC. We run each method for each pitcher that threw over 200 pitches for each year of the 2010-2012 MLB seasons. Preliminary results are shown in Table 3.4. $K_{\text{BIC}_{\text{adj}}}$ is the number of clusters chosen for a pitcher using the BIC_{adj} method and $K_{\text{pred.strength}}$ is the number of clusters chosen for a pitcher using validation by prediction strength.

Table 3.4: Comparing the difference in number of clusters (pitch types) chosen by BIC_{adj} and validation by prediction strength. A positive number represents BIC_{adj} choosing a higher number of clusters (k) for a pitcher than validation by prediction strength.

$K_{\text{BIC}_{\text{adj}}} - K_{\text{pred.strength}}$	-3	-2	-1	0	1	2	3
Count	36	261	352	263	0	110	29

Since ground truth is unknown, we continue using empirical observation and Brooks Baseball to determine the accuracy of our methods. We initially notice in table 3.4 that the BIC model tends to choose more clusters than our prior information suggests, and after further investigation we determine that both of the proposed methods, on average choose lower k than does original BIC. Through empirical analysis and comparisons to Brooks Baseball, we determine that both (BIC_{adj}) and validation by prediction strength outperform BIC. We visualize the inaccuracy of BIC in Figure 3.4b as BIC chooses 7 clusters, while BIC_{adj} (and validation by prediction strength) correctly choose 5 clusters. There are many other examples of BIC underperforming which we omit for simplicity.

We continue our analysis by determining the optimal cluster number selector of our two proposed methods. An observation made from Table 3.4 is BIC_{adj} tends to choose a lower k than does validation by prediction strength. We also find that our two methods only choose the same number of clusters 263/1054 or $\approx 25\%$ of the time. In order to determine which

method is correct (if any), we randomly select 50 pitchers to compare the clustering results to empirical investigation and Brooks Baseball's number of pitch types. Empirically, we find that in 45 of the 50 cases, BIC_{adj} outperforms validation by prediction strength based on the comparison to BIC. BIC_{adj} is also a substantial improvement over BIC in this analysis, so we use it going forward to choose the number of pitch clusters in MBC. Figures 3.5 and 3.6 illustrate BIC_{adj} outperforming validation by prediction strength both when BIC_{adj} chooses the higher and lower number of clusters and for both relief (RP) and starting (SP) pitchers.

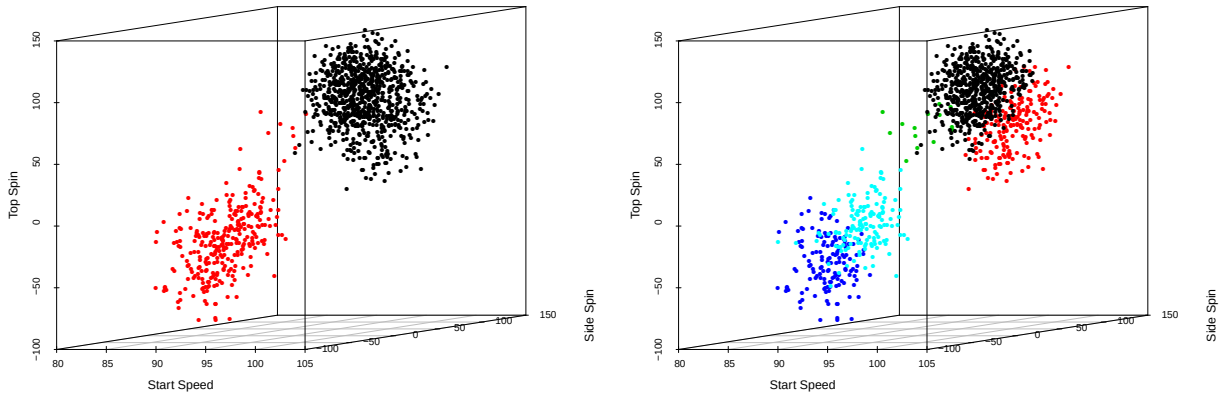

Figure 3.5: Mitchell Boggs (RP) 2010: 3.5a (left) displays MBC using BIC_{adj}, and Figure 3.5b (right) displays MBC using validation by prediction strength. Empirically, we observe that Boggs throws two pitches and BIC_{adj} correctly identifies the number of clusters. Note: In Figure 3.5a black corresponds to a four-seam fastball and red a slider. Figure 3.5b does not have any pitch labels.

Table 3.5: Pitch type names with corresponding colors for Figure 3.6a.

Pitch Name	Four Seam	Two Seam	Cut Fastball	Changeup	Curveball	Slider
Color	Light Blue	Pink	Blue	Green	Black	Red

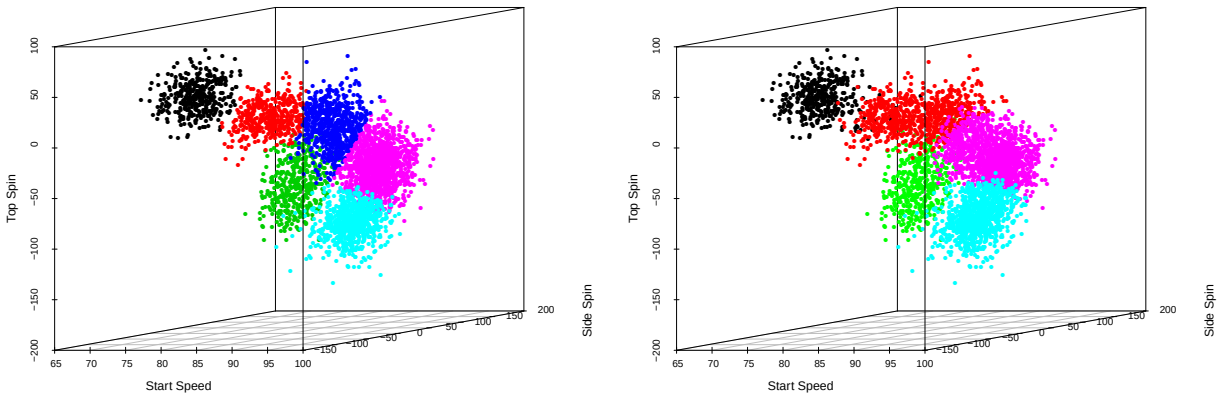

Figure 3.6: C.J. Wilson (SP) 2010: Figure 3.6a (left) displays MBC using BIC_{adj}, and Figure 3.6b (right) displays MBC using validation by prediction strength. Empirically, we observe that Wilson throws 6 pitches as displayed in Figure 3.6a, and Figure 3.6b is incorrect. This example displays BIC_{adj} choosing a greater cluster number than validation by prediction strength, but still resulting in the correct number of clusters.

We use MBC for all pitchers during the 2010 and 2011 seasons, and show a few illustrations from this large analysis below. Figure 3.7 shows how our method is ideal for detecting how a pitch can evolve over time. The purple and black clusters represent Jon Lester's curveballs. The difference between the two clusters is the speed and horizontal break of the pitch. In 2010, Jon Lester's curveball averaged 78 MPH (back cluster), while in 2011 it averaged 76 MPH with slightly more horizontal break (red cluster). This subtle yet important difference reflects a change in Lester's curveball from 2010 and 2011, which our MBC model detects. Figure 3.8 visualizes how our method can cluster pitches with high variance, such as Tim Wakefield's knuckleball (orange). The pseudo-spins, corresponding to additional break, are considerably more variable than other pitches due to the unpredictable nature of stitch

position in a knuckleball's trajectory.

Table 3.6: Pitch types with corresponding colors for Jon Lester and Tim Wakefield.

Pitch Name	Four Seam	Sinker	Cut Fastball	Changeup	Curveball	Knuckleball
Color	Red	Light Blue	Blue	Green	Black and Purple	Orange

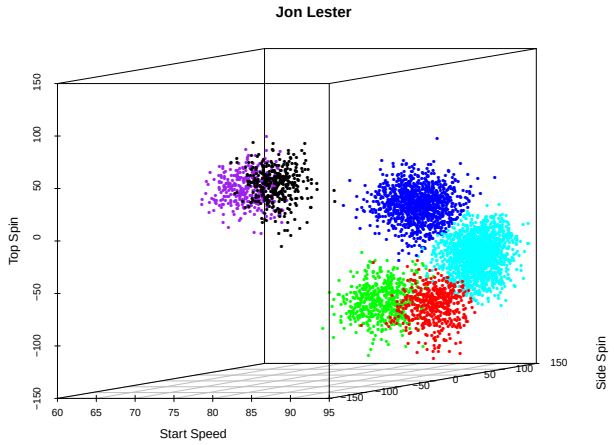

Figure 3.7: The pitches thrown by Jon Lester classified using model-based clusters, which shows two distinct clusters for curveballs (purple and black) corresponding to a change from the 2010 to 2011 seasons. The difference between the two clusters is the speed and horizontal break of the pitch. This is one example of how the MBC model can detect subtle but important pitch evolution differences.

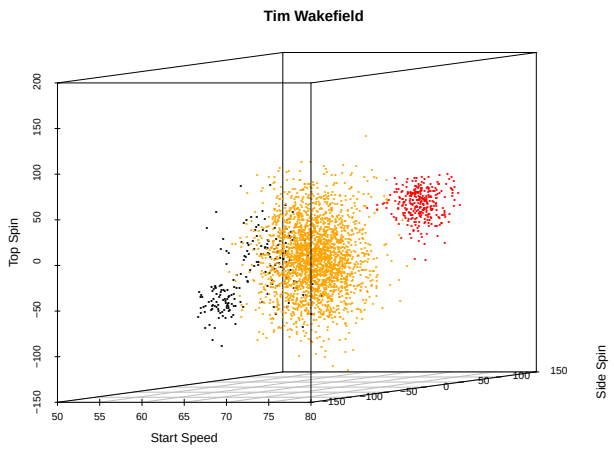

Figure 3.8: The pitches thrown by Tim Wakefield classified using MBC. MBC can also detect and classify pitches with relatively higher variances, like the pseudo-spins for Tim Wakefield's knuckleball (orange), compared to the fastball (red) and curveball (black).

3.6 Stability of Cluster Membership

We evaluate quality of the MBC method by assessing the *stability* of the method in terms of how sensitive our model is to smaller sample sizes. To do this, we take the data for each pitcher and run the clustering on an 80% subset of the data, the remaining 20% subset, and then on the full data. Next, we calculate the number of pitches in both the 80% and 20% subsets that do not change clusters compared to the full data set. In order to be confident in our results, we repeat this subsetting process 20 times for each pitcher and find the mean and standard error. We find that after 20 samples our standard error is sufficiently small and we are confident with our stability sample mean estimates.

Figure 3.9 shows the two distributions of all pitchers for the 80% and 20% subsets, and the proportion of pitches that are in the same cluster as the full data for all pitchers. Overall, for both the 80% and 20% subsets, pitchers have 80% or more of their pitches clustered in the same cluster on average which is substantially higher than that of k-means and hierarchical

clustering. We find that this stability holds on sample sizes as low as 100 pitches. It is important to note that we kept the number of clusters chosen (k) the same on both subsets as when we ran it on the full data.

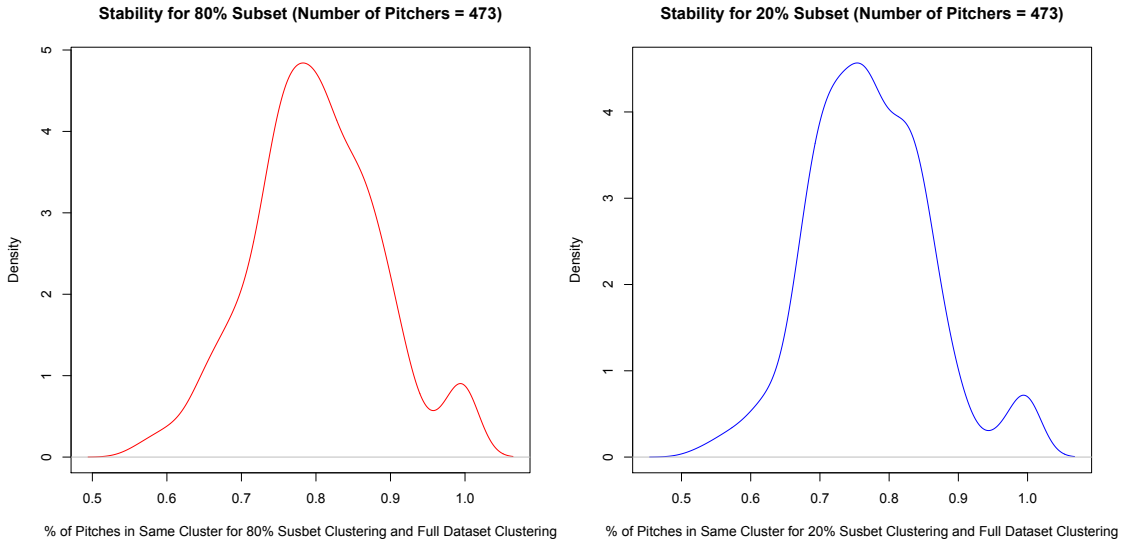

Figure 3.9: 80% and 20% Stability: Percent of pitches in the same cluster in subset and full data set, across 20 replications of the procedure. For both the 80% and 20% subsets, pitchers have 80% or more of their pitches clustered in the same cluster on average. We find that this stability holds on sample sizes as low as 100 pitches.

Chapter 4

Assigning Pitchers to Subgroups Using Clustering Methods

Using our new methods, we now have an improved classification method for all pitchers. Accurate pitch types equips us with the necessary tools to tackle countless other baseball research questions, including pitcher clustering. We begin by introducing two motivating applications of pitcher clustering: increasing power when evaluating pitcher match-ups and predicting pitcher injury. Next, we cluster pitchers into similar groups using a variety of data manipulation and clustering techniques. In addition, we build a framework to cluster pitchers in greater detail. We illustrate this via two motivating examples.

4.1 Motivating Applications of Pitcher Clustering

Clustering pitchers can help us answer several open baseball research questions. One is how we can assign pitchers themselves to clusters, a problem highly motivated by the need to assess a pitcher's likely performance against unfamiliar players or teams. In these situations, most hitters likely have never faced the current pitcher, and thus, there is no or very little

data allowing inference to be made about the hitter's history against the current pitcher. For example, the team facing the pitcher may have never played against him before, but they have faced a different pitcher that is also, say, a hard throwing, side-armed lefty with a good slider. The team can then use the information from the similar pitcher in their analysis, to determine how well their players will perform against the new pitcher. Placing pitchers into similar groups and looking for relationships with groups of batters would be invaluable to MLB. This information can lead to advanced and more detailed scouting reports of what type of pitches are a hitter's strength or weakness.

Another motivation is predicting pitcher injury. Professional baseball teams invest millions of dollars in players; how can baseball teams best protect their expensive assets? Most major preventative decisions are not made with any known successful analyses. Decisions on pitcher workload are made with reasoning that has "worked" in the past, and what is arbitrarily seen around the sport as the appropriate way to handle pitchers. For example, the Washington Nationals were heavily scrutinized for shutting down their top starting pitcher, Stephen Strasburg, during their playoff run in the 2012 season. Strasburg had elbow surgery the season before, and was told he would have a limit of 160 innings because the previous year they put the same restriction on another pitcher recovering from the same injury, which did not recur. Approaching the issue of workload and ultimately injury prevention more systematically and quantitatively can save MLB teams millions of dollars. One way to attempt to quantify and ultimately predict pitcher injury determining if there are certain characteristics among clusters of pitchers that indicate a pitcher is going to be injured in the future.

Although we are not able to directly address the two proposed frameworks above, we explore a useful method that we will continue in future work.

4.2 Data Manipulation

To simplify our work and avoid lurking variables and effects, we limit our analysis to only right handed pitchers from the 2010-2012 seasons. We consider the following variables for pitcher clustering (The names in parenthesis are how the variables are represented in Figure 4.2, and only a subset of the clustering variables are visualized in Figure 4.2):

- Overall mean and variance of release point in the z direction, or how high off the ground the ball is when released from the pitcher's hand, as calculated by the raw PITCHf/x database. (release.z.mean, release.z.var)

- Overall mean and variance of release point in the x direction, or how far left or right on the pitchers mound the ball is when released from the pitcher's hand, as calculated by the raw PITCHf/x database. (release.x.mean, release.x.var)

- Maximum range of velocity between two pitch types, as determined by our PITCHf/x classifications. (velocity.range)

- Maximum start speed pitch cluster or the fastest pitch cluster, as determined by our PITCHf/x classifications. (max.velocity)

- Mean of the clusters with the maximum and minimum side-spin, as determined by our PITCHf/x classifications. (max.x and min.x)

- Mean of the clusters with the maximum and minimum top-spin, as determined by our PITCHf/x classifications. (max.z and min.z)

4.3 Pitcher Clustering: K-means

We begin clustering pitchers by exploring k-means and CH-index. We use the variables in Section 4.2 in our clustering as they uniquely distinguish characteristics of pitchers.

Figure 4.1 displays the results of the CH-Index cluster number selector for k-means clustering, and determines two clusters is the optimal choice of k. Our k-means clustering problem has many variables and thus making the high dimensional variable relationships and clustering hard to visualize. We display the results in Figure reffig:ynplot-pitcher, and visualize each possible two dimensional relationship, where the colors red and black represent the two clustering groups. There are obvious variables carrying the cluster separation as well as interesting relationships among the possible bivariate combinations. We are able to visualize slight separation between the two clusters in the velocity range vs minimum top spin (min.z) plot, velocity range vs minimum side spin (min.x) plot, and a variety of the other bivariate relationships in Figure 4.2.

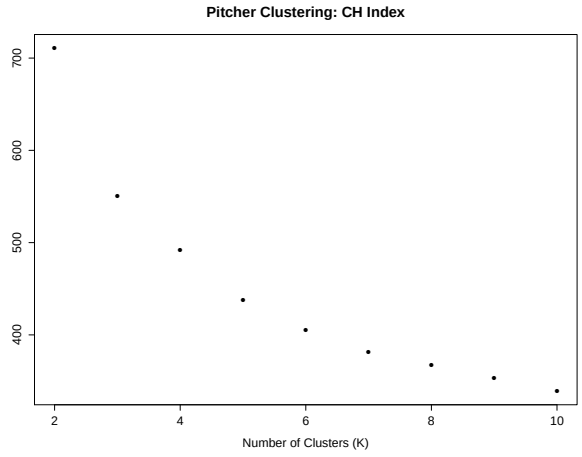

Figure 4.1: Pitcher Clustering for Right handed pitcher: CH-Index Cluster Number Selector

Our clustering system displays two separate clusters, and because these clusters do not have classification names, we reference them as the red and black clusters. Velocity range, minimum top spin, and minimum side spin are the variables most influencing the cluster memberships. Based on the results, and what we know about the characteristics of right-handed pitchers, we describe an average pitcher in each of the two clusters and then choose two motivating examples to support our description of each cluster membership.

An average pitcher in the red cluster has a negative minimum side spin, negative top spin, and a velocity range of 15 mph or higher. We can describe an average right-handed pitcher in the red cluster from our knowledge of pitch characteristics described in Section 2.3. An average pitcher in the red cluster throws a curveball/slider with a lot of top and side spin (and a lot of movement in the horizontal and vertical directions). An average pitcher in the black cluster has a positive minimum side sin, positive top spin, and velocity range of lower than 15 mph. We can describe a black cluster's average pitcher as a player that either throws no curveball/slider/sinker or a curveball/slider/sinker with minimal spin in either direction. Through our preliminary cluster analysis, there are no other obvious characteristics that are separating the two clusters we found.

Displaying the same results in examples, a randomly chosen pitcher from the red cluster is Jarrod Parker who follows the description we described for the red cluster as he throws a sinker and curveball that have a large amount of back and top spin (and movement in both the vertical and horizontal directions). A randomly chosen pitcher from the black cluster is Jordan Zimmerman. Zimmerman also follows the description we described as he throws a slider and an occasional curveball but both pitches have minimal spin and movement relative to pitchers in the red cluster.

Beyond cluster relationships, we also observe overall trends through bivariate relationships in Figure 4.2. We highlight two relationships, the first is consistent with our prior beliefs, while the second relationship is not what we hypothesized before our analysis.

As the difference between a pitcher's fastest and slowest pitch decreases, the minimum top spin increases. This is consistent with our prior beliefs because if a right-handed pitcher's minimum top spin is positive, then he most likely does not throw a curveball/slider (slower pitches) with a lot of spin or movement. As a result, it has a higher velocity, and ultimately leads velocity range to be small. The pitcher may also not throw a curveball or slider and this would lead to the same relationship.

Another relationship we observe through the bivariate clustering relationships in Figure 4.2 is release point variation in the x direction increases, maximum side spin increases. In other words, the larger the variance in how far left or right the right-handed pitchers release point is, the larger the maximum side spin spin (and horizontal movement). The maximum side-spin of a right-handed pitcher corresponds to how much side-spin their fastball clusters has. This implies the larger the variation in horizontal release point, the more right to left side-spin their fastballs have. Although release point variation ultimately is not a strong factor in separating our two pitcher clusters, it does display an interesting relationship with the side-spin of pitches that we will investigate in future work.

Overall our initial cluster analysis separates two types of pitchers: pitchers with curveballs, sliders, and sinkers with high spin and movement, and pitchers that throw less extreme, faster off-speed pitches or do not throws curveballs, sliders, or sinkers all together. Although an interesting result, in order to answer our motivating questions of increasing sample size for hitter vs pitcher match-ups and predicting pitcher injury, we would like to see more sep-

arate clusters. In future work we can continue exploring other possible variables that help distinguish pitchers such as the pitch type sequence (the order a pitcher throws each pitch type during a game), if a pitcher has been injured before, percentage of pitches thrown for strikes, percent of each pitch type a pitcher throws, etc. We can use a variety of additional statistical methods including principle component analysis to decrease variable dimension and variable selection techniques. In addition to clustering, we can calculate pitcher similarity scores to determine what pitchers are most similar and different than others.

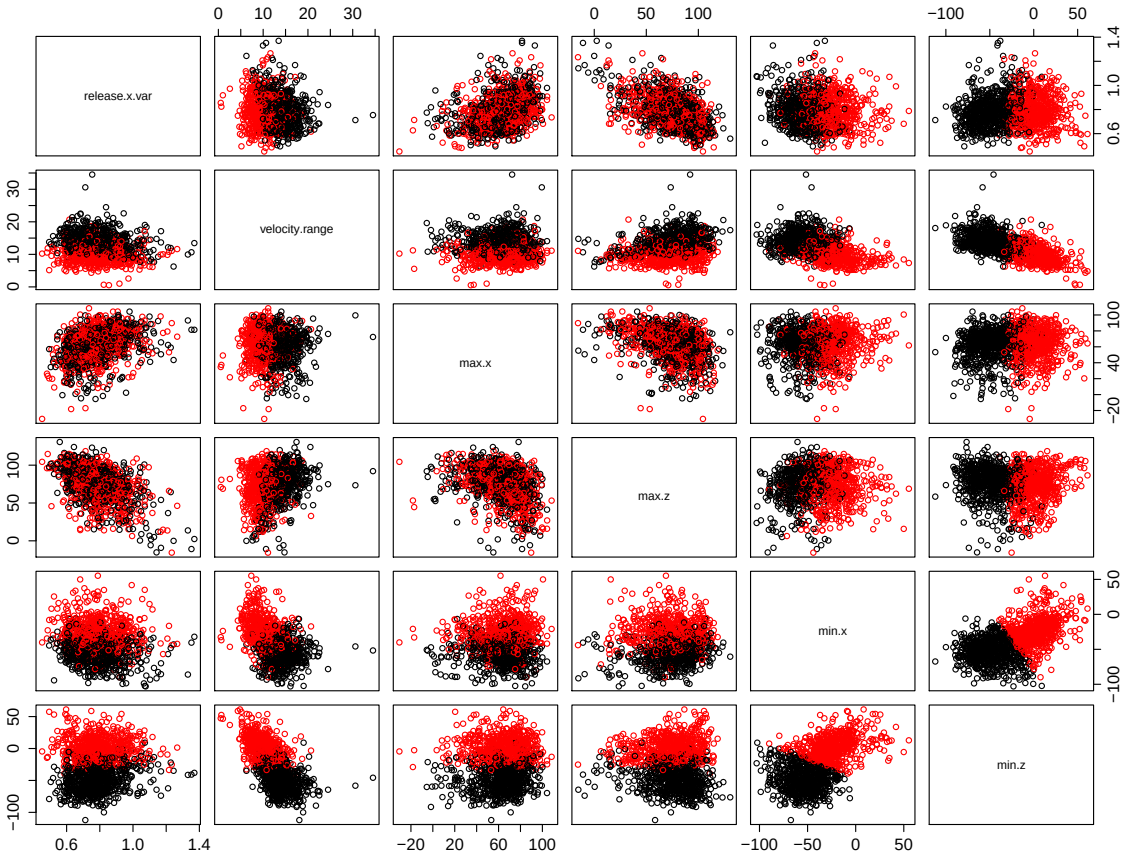

Figure 4.2: Pitcher Cluster: Each data point is an individual pitcher, in an individual year between 2010-2012. The two clusters are referenced as red and black due to their lack of classification label. Section 4.3 analyzes and describes the relationships and results. A subset of the clustering variables are visualized in Figure 4.2, and are defined in Section 4.2.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

We propose a new clustering method and pitch classification algorithm and test this approach using the PITCHf/x database for the 2010-2012 MLB seasons. Our analysis illustrates better pitch clustering and classifications than the current neural network method based on empirical evidence in our plots and the stability our method shows when tested on multiple subsets. Furthermore, using our model-based clustering approach's results, we design a simple algorithm to classify each pitch based on the individual pitcher's clustering results, which performs well for most pitchers. Its strongest features are correcting any obvious misclassification in the neural network model, accounting for pitch evolution over time, and correctly identifying pitches with high internal variances. Our method performs well in a highly debated topic in MLB analysis, namely distinguishing two-seam and four-seam fastballs.

We then use our improved clustering and classification methods to improve pitcher clustering. We do not use variables that are independent of team decisions (i.e. pitcher selection order),

but instead incorporate variables that physically define a pitcher, such as the characteristics of their pitch clusters which we found while improving pitch clustering with MBC. We found two clusters that separate pitchers based on the type and characteristics of the off-speed pitches they throw, and we have ultimately built a strong foundation to continue further pitcher clustering analysis.

5.2 Future Work

Our new clustering and classification method performs sufficiently well, but there are alternative approaches we can use to find clusters of pitches. For example, we could incorporate prior information about each pitcher, such as whether or not they are a relief pitcher, and use a fully Bayesian method for detecting clustering of pitches. Furthermore, research needs to be extended to cluster pitchers in a successful and useful manner. Specifically, looking at more variables that fully encompass a pitcher's arsenal of pitches, pitch sequence, percentage of pitches thrown for strikes, and others. Additionally, we propose using data on pitcher injuries to predict potential future injuries. We propose utilizing additional statistical methods including principle component analysis and advanced variable selection techniques to handle possible high dimensional clustering issues.

Bibliography

Calinski, R. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3, Pages 1-27.*

Foster, A. (2012). Scouting with PITCHf/x.
URL http://www.baseballprospectus.com/article.php?articleid=17327.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* 2nd ed. Springer-Verlag, New York.

Nathan, A. M. (2007). Effect of the magnus force in the PITCHf/x tracking system.
URL http://baseball.physics.illinois.edu/Magnus.pdf.

Tibshirani, R. and Walther, G. (2009). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 511-528.*