Fixing the Curve: Improving Major League Baseball Pitch Classification with Model-Based Clustering

December 17th, 2018

Abstract: Clustering is a popular unsupervised learning method used for classifying data points into specific groups. It can be utilized in a variety of applications, from locating continental faults in earthquake studies to identifying particular customer bases for targeted marketing strategies. In theory, data points belonging to the same group should have similar characteristics, while data points belonging to different groups should have very different characteristics. K-means and hierarchical clustering are two of the most popular methods for implementing this machine learning technique, while methods such as model-based clustering are not utilized as frequently. In this paper, we engage an in-depth study of clustering methods for pitch classification from a baseball dataset. While there is more work to be completed in this field, our results indicate that model-based clustering is an attractive clustering method due to its ability to automate the selection of the number of clusters.

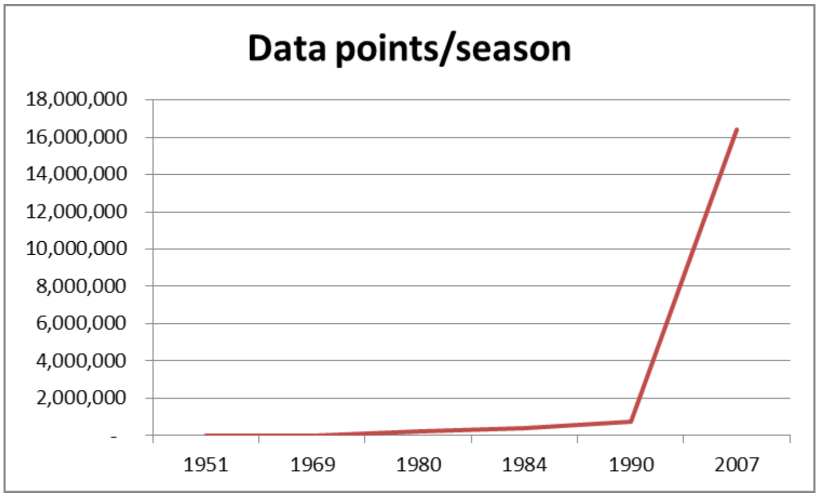


Figure 1: Number of data points collected per baseball season from 1951 to 2007, the year PITCHf/x was introduced.

Introduction

When you think about big data, the sport of baseball usually isn't the first thing to come to mind. Baseball has been around for nearly 200 years, but only recently has the sport entered the realm of data analysis. It all began in 1951, when Hy Turkin and S.C. Thomson published *The Official Encyclopedia of Baseball*, the first of its kind.[1] With basic statistics such as batting average outlined in the text, roughly 1,800 data points were collected each season. Then an improved version of Turkin and Thomson's encyclopedia, *The Baseball Encyclopedia: The Complete and Official Record of Major League Baseball*, was published by The Macmillan Company in 1969.[2] With this came a collection of about 12,000 data points per season. Finally in 1980, a new name entered the conversation, effectively paving the way for baseball analysts for years to come: Bill James. His early work, *The Bill James Historical Baseball Abstract*, had just begun to gain traction, introducing various new statistics to the baseball vocabulary.[3] This resulted in the collection of over 200,000 data points per season, a significant leap compared to the previous decade. Among the statistical innovations attributable to James are runs created and win shares. Soon after his first publication, James also proposed the creation of Project Scoresheet, a network of fans that worked together to collect and distribute play-by-play data of baseball games. Shortly after, pitch-by-pitch data became available and by 1990 nearly 800,000 data points were collected per season. Since then new technologies have been developed and introduced, increasing data collection to around 16,000,000 data points per season. This progression of increasing data collection is represented in Figure 1.[4]

The first big technological advancement that baseball experienced was PITCHf/x, a system that tracks the speed and trajectory of every pitched baseball during a game.[5] It was first implemented in 2007 and was installed in every Major League Baseball stadium as well as in all Minor League stadiums. A few years ago, PITCHf/x was replaced with something called TrackMan, a 3D Doppler radar system that is placed behind home plate and measures the location and trajectory of pitched and hit baseballs. TrackMan is used in conjunction with Statcast, a high-speed automated tool used for analyzing player movements on the field. Together these systems collect substantial amounts of data on each pitch and play, amassing large data sets for each baseball game played.

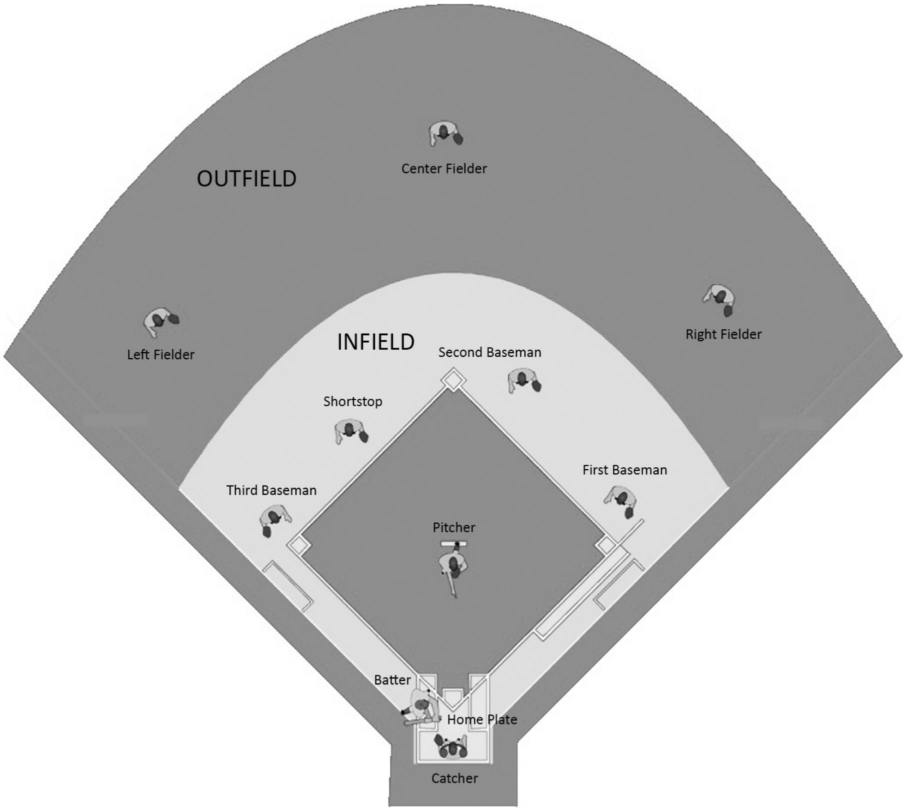


Figure 2: Fielding positions in baseball.

During a baseball game, a pitcher throws different types of pitches in an attempt to make it more difficult for the batter to effectively hit the ball. Sometimes a pitcher will grip the ball differently or release it in a different way to produce a particular pitch. These subtle changes are detected in the various pitch measurements that TrackMan collects. This system has a built-in algorithm to classify these pitch types, however it is often very inaccurate. For this reason, individuals are hired (called "TrackMan operators") to attend every home game of their assigned team and manually input the pitch types. While this process is somewhat subjective, it tends to be more accurate than the predictions from the automated algorithm.

Throughout the course of a game, one team has its nine players in defensive positioning on the field. These positions include the pitcher, the catcher, the infielders, and the outfielders, and they are positioned as displayed in Figure 2.[6] The catcher is centered behind home plate with the pitcher positioned on a mound in front of him. The remaining seven fielders are positioned around and behind the pitcher. A player from the opposing team stands at home plate to bat. The pitcher throws the ball across the plate to the catcher, and the batter attempts to hit it. For the next pitch, the pitcher may decide to change his grip or release point, thus throwing a new type of pitch than he did before. During a game, a pitcher usually throws three to five different types of pitches that vary in speed and movement. The different types of pitches that are thrown during a game are put into a list called his arsenal. The arsenal lists the pitch type and their associated velocities.

In baseball there are two basic types of pitchers: a starting pitcher and a relieving pitcher. The starter begins playing in the first inning and usually stays in the game until he has thrown 100 pitches, which usually occurs in the 5th or 6th inning, but can vary based on the game situation. A reliever comes into the game after the starter, and throws considerably less pitches than the starter. It is not uncommon for multiple relievers to be used in a game.

In the data set several pitch types are described. "Fastball" refers to a 4-seam fastball, which means the pitcher holds the ball with the seams of the ball horizontal to his grip instead of vertical. The 4-seam fastball typically has the highest velocity when compared to other pitch types. Slightly slower, the Sinker is another type of fastball, a 2-seam fastball; the pitcher has a slightly different grip that gives the 2-seam pitch more horizontal and vertical movement as it crosses home plate. Sliders and Curveballs have considerably lower velocity than a pitcher's fastball and are called "breaking balls," meaning they move suddenly as they cross home plate, often inducing a swing and miss outcome from the batter.

Unfortunately we are unable to ask a pitcher "what pitch was that supposed to be?" every time he throws the ball in a game. Therefore we should not take characteristics of known pitches and determine if new ones fit into those groups. Rather, we should create our own groups and see if they match the pitcher and coach impressions. This is where unsupervised learning would be useful; no labels are given to the learning algorithm as it attempts to find and create its own structure within the data. Cluster analysis is an unsupervised learning technique that aims to group a set of objects such that objects in the same group are more similar to each other than to those in other groups. These groups, called *clusters*, are based solely on information found in the data that describes the objects and their relationships to one another.

The objective of clustering is to maximize the similarity within groups and minimize the similarity between groups. It can be used as a means of exploratory analysis in order to identify natural structures that exist within the data. Clustering methods such as k-means and model-based offer sensible options for improving TrackMan's pitch classification algorithm. These methods are given no prior pitch type information and instead create groups based solely on the characteristics of the data at hand.

This paper will focus on the use of model-based clustering for pitch identification, with three primary goals:

1. To use k-means and model-based clustering to create an improved automated pitch classifier.
2. To compare and contrast the results of the clustering methods with those of the TrackMan operator.
3. To extend the methods used here to a broader context.

The sections are organized as follows: Section 1 introduces clustering procedures as methods of classification and density estimation. Section 2 applies k-means and model-based clustering methods to baseball data for pitch classification. Finally, Section 3 compares these methods and discusses other applications of model-based clustering and its relevance and importance in other contexts.

1 Cluster Analysis

There are several different clustering methods, such as k-means clustering and hierarchical clustering. K-means is an example of centroid-based clustering where clusters are represented by a central vector or centroid. K-means aims to minimize the mean distance within clusters. This method requires the analyst to specify the number of optimal clusters to extract (k). The basic steps of k-means clustering are as follows:

1. **Initialization** - k points are randomly chosen from the data set and become the initial centroids.
2. **Cluster Assignment** - for every point, the distance to each centroid is calculated and the points are assigned to the closest centroid, thus creating the clusters.
3. **Moving the Centroid** - now that the clusters are created, their centroids are recalculated as the mean of all the observations in that cluster.
4. **Iteration** - step 3 is repeated until the centroids stop changing after the recalculations.

K-means is a quick and efficient method as the complexity of one iteration is $k * d * n$ with $k =$ number of clusters, $d =$ time it takes to calculate the Euclidean distance between two points, and $n =$ number of observations.[7]

Hierarchical clustering aims to maximize the distance between clusters and uses a slightly different approach.[8] The steps are as follows:

1. **Initialization** - each observation is first treated as its own cluster.
2. **Merging** - the two clusters that are closest together are identified and merged.
3. **Iteration** - step 2 is repeated until only a "small" number of clusters remain.

K-means and hierarchical clustering are mainly heuristic and therefore not based on formal methods. They are also examples of "hard" clustering, which means the clusters do not overlap and observations must belong to one cluster only.[9]

Model-Based Clustering

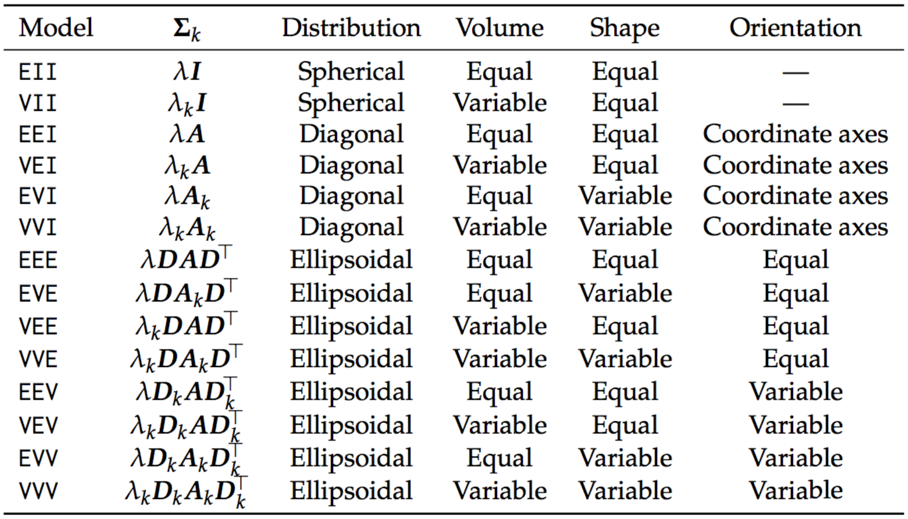
Model-based clustering is an example of "soft" clustering in which clusters are allowed to overlap and observations have a probability of belonging to each cluster. It is assumed that each cluster is generated by a multivariate normal distribution, so the data set is considered to be a mixture of different distributions. This method then applies maximum likelihood estimation and the Bayesian Information Criterion (BIC) to identify the best model and the number of clusters. The number of clusters is not required to be chosen beforehand; model-based clustering calculates the optimal number of clusters to be used. The steps of model-based clustering are as follows:

1. **Initialization** - points are randomly assigned to clusters.
2. **Calculation** - the means and covariances for each cluster are calculated.
3. **Cluster Assignment** - the probabilities of cluster membership for each point are calculated and points are assigned to the cluster with the highest probability.
4. **Iteration** - steps 2 and 3 are repeated until points no longer move between clusters.

Gaussian Mixture Modeling

Model-based clustering is based on finite Gaussian mixture modeling. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Mixture models generalize k-means clustering in a way to incorporate information about the covariance structure of the data in addition to the centers of the underlying Gaussians.

Suppose $x = \{x_1, x_2, ..., x_i, ..., x_n\}$ is a sample of n independent and identically distributed observations (each a vector of measured data). The distribution of each of these observations is determined



Model	$\boldsymbol{\Sigma}_k$	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable

Figure 3: Parametrizations of the within-group covariance matrix for multidimensional data available in the *mclust* package and the corresponding geometric characteristics.

by a probability density function through a finite mixture model of G components, which takes on the form

$$f(x_i; \Psi) = \sum_{k=1}^{G} \pi_k f_k(x_i; \theta_k)$$

where

1. $\Psi = \{\pi_1, ..., \pi_{G-1}, \theta_1, ..., \theta_G\}$ are the parameters of the mixture model.
2. $f_k(x_i; \theta_k)$ is the kth component density for observation x_i with parameter vector θ_k.
3. $(\pi_1, ..., \pi_{G-1})$ are the probabilities or mixing weights (such that $\pi_k > 0, \sum_{k=1}^{G} \pi_k = 1$).
4. G is the number of mixture components (clusters).

Assuming that G is fixed, the mixture model parameters Ψ are usually unknown and must be estimated, so we turn to the maximization of the log-likelihood of the equation above. Direct maximization of the log-likelihood is complex, so the maximum likelihood estimator (MLE) of a finite mixture model is usually obtained through the Expectation-Maximization (EM) Algorithm.[10]

In model-based clustering, each component of a finite mixture density is associated with a group or cluster. The Gaussian mixture model (GMM) is a popular model that assumes a multivariate Gaussian distribution for each component, i.e. $f_k(x; \theta_k) \sim N(\mu_k, \Sigma_k)$. Consequently the clusters are elliptical, centered at the mean vector μ_k, and with other geometric features such as shape, volume, and orientation, the clusters are determined by the covariance matrix Σ_k. Various parametrizations of the covariance matrices can be obtained through eigen-decomposition of the form $\Sigma_k = \lambda_k D_k A_k {D_k}^T$ where

1. λ_k is a scalar controlling the volume of the ellipsoid.
2. A_k is a diagonal matrix specifying the shape of the density contours with $det(A_k) = 1$.
3. D_k is an orthogonal matrix that determines the orientation of the corresponding ellipsoid.

The `mclust` package is used for model-based clustering, classification, and density estimation. The package allows for the modeling of data as a Gaussian finite mixture with different covariance structures and different numbers of mixture components. In the one-dimensional case, there are only

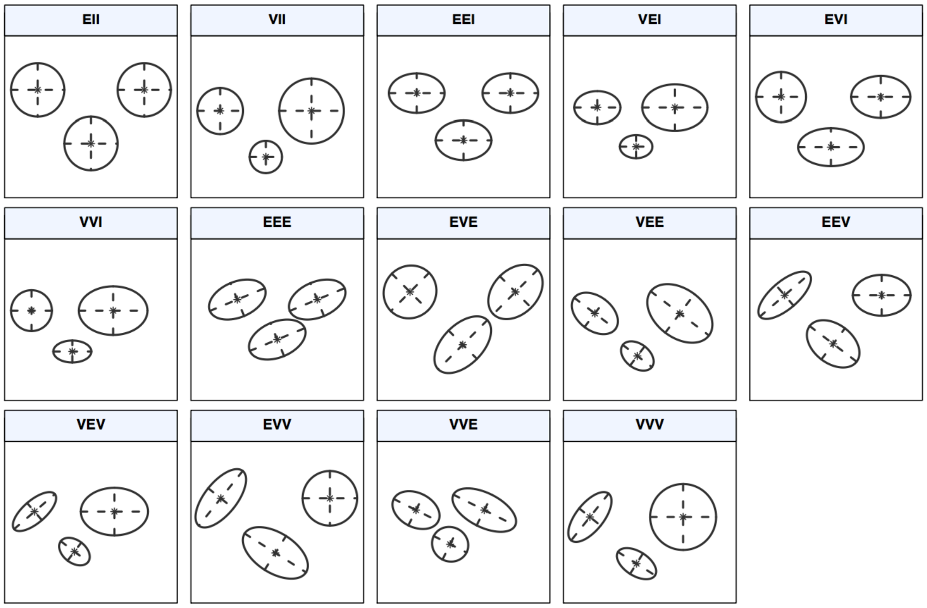


Figure 4: Ellipses of isodensity for each of the 14 Gaussian models obtained by eigen-decomposition in the case of three groups in two dimensions.

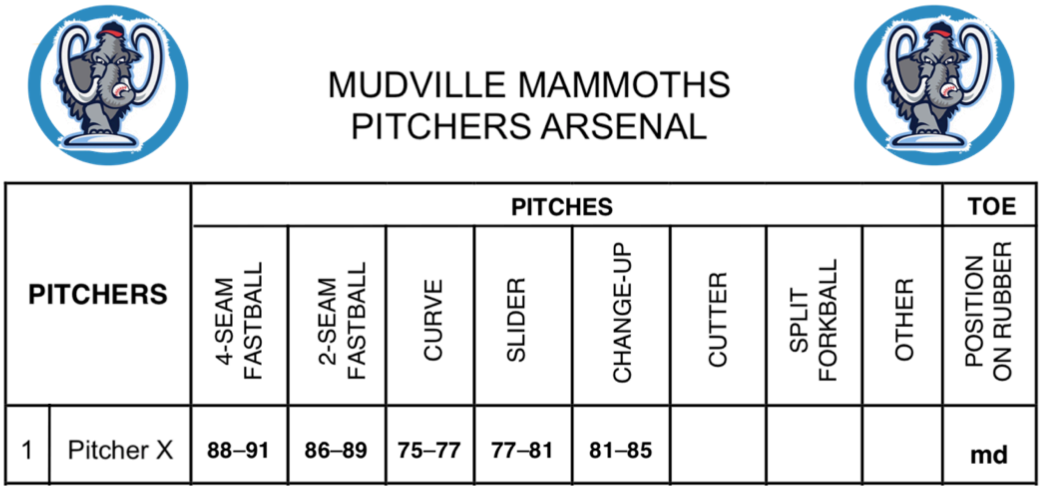
two models: E for equal variance and V for varying variance. However in the multivariate setting, the shape, volume, and orientation of the covariances can be constrained to be equal or variable across groups. This generates 14 possible models with different geometric characteristics. Figure 3 reports these models along with their corresponding distribution structure type, shape, volume, and orientation. Figure 4 graphically displays the geometric characteristics of these models.[11]

The `mclust` package provides a comprehensive strategy for clustering and can be used with a variety of data sets. For example, model-based clustering can assist with pitch classification in baseball and help determine a specific pitcher's arsenal, or a list of types of pitches that he throws.

2 Case Study

This anonymized data set is from a Class A Short Season minor league baseball team. It includes data from the team's home games throughout the course of one season that was collected by the on-site TrackMan operator. The only subjective variable associated with this data set is *TaggedPitchType*, and the main task of the TrackMan operator is to input the data for this variable. The operator sits behind home plate and watches every pitch of the game, storing the outcome of the pitch and the pitch type, which the operator must determine based on velocity, spin rate, and previous knowledge of the pitcher's arsenal (the types of pitches he throws and their associated velocities).

Although Statcast provides pitch type in the TrackMan data, their own proprietary algorithm is used to classify pitches, which is represented in the *AutoPitchType* column. This is the company's attempt to determine the intent of the pitch based on the measurements and characteristics of the moving ball. The algorithm utilizes the pitch's speed (relative to that pitcher's maximum speed), the spin rate, the spin axis, and the amount of break of the ball to suggest a probable pitch type.



MUDVILLE MAMMOTHS
PITCHERS ARSENAL

PITCHERS		PITCHES								TOE
		4-SEAM FASTBALL	2-SEAM FASTBALL	CURVE	SLIDER	CHANGE-UP	CUTTER	SPLIT FORKBALL	OTHER	POSITION ON RUBBER
1	Pitcher X	**88–91**	**86–89**	**75–77**	**77–81**	**81–85**				**md**

Figure 5: Pitching arsenal of Pitcher X.

Unfortunately this internal algorithm is often wrong and thus the subjective eye of a well-trained TrackMan operator is almost always trusted over the automated pitch type.

Clustering methods may be useful in this context in order to replicate and improve pitch classification. K-means clustering can be used to examine the accuracy of a pitcher's arsenal (the types of pitches that he supposedly throws). Since this approach requires a predetermined number of clusters prior to analysis, the quality of the clusters created can be a testament of the arsenal's accuracy.

Alternatively, model-based clustering can be applied in order to create a new arsenal entirely. This method is appropriate because an arbitrary number of clusters is not required to be chosen at the start of the analysis. Although pitchers have their predetermined arsenals, these are not assumptions that can be made in confidence. Model-based clustering is attractive as it reveals the optimal number of clusters, identifying potential discrepancies between these groupings and the pitcher's arsenal.

This method also allows for different means and variances for each cluster. This is useful because each pitch type is inherently different. For example, more variance in movement might be expected for a slider or a curveball as compared to a fastball. Since model-based clustering uses density estimation when creating clusters, it can reveal interesting information about the consistency of a pitcher across pitch types.

Clustering and Pitch Classification

To help illustrate the methods described above, we have chosen one pitcher to explore (the name of the pitcher has been anonymized per data use restrictions). Pitcher X is a starting pitcher who had 10 starts during this season. Figure 5 displays his pitching arsenal and Table 1 shows the distribution of pitch types and counts that Pitcher X threw during this season.

Overall we can see that Pitcher X throws a fastball, a sinker, a slider, a changeup, and a curveball. He also seems to throw a similar amount of each pitch type, except his curveball. Perhaps he does

Pitch Type	Average Velocity	Number of Pitches	% of All Pitches
4-Seam Fastball	87.78	208	31.1
2-Seam Fastball	87.03	184	27.5
Slider	79.86	135	20.2
ChangeUp	82.31	111	16.6
Curveball	76.61	30	4.6
Total		668	100

Table 1: Distribution of pitch types for Pitcher X.

not actually throw a curveball, and the pitch was simply misclassified during data collection.

We can also see that Pitcher X throws a disproportionately large amount of fastballs compared to the rest of his pitch types. This is common for many pitchers in baseball; the fastball is thrown most often because it puts minimal stress on a pitcher's arm compared to other types of pitches.

Notice that Pitcher X's fastballs are relatively split between the 4-seam fastball and the 2-seam fastball. This is interesting because the 2-seam is only a slight variation of the 4-seam, and pitchers usually choose one and stick with it. This may be another instance of misclassification.

Let's take a closer look at some of the attributes measured by the TrackMan system for different types of pitches. Table 2 displays examples of a 4-seam fastball, changeup, and slider thrown by Pitcher X.

Pitch Type	Release Speed	Vertical Release Angle	Spin Rate	Extension	Vertical Break
4-Seam Fastball	88.70	-2.52	1982.65	5.73	-11.28
ChangeUp	82.68	-0.38	1782.16	5.66	-31.44
Slider	80.08	0.75	2265.93	5.55	-43.15

Table 2: Example of pitches thrown by Pitcher X.

We can see how measurements like release speed, vertical release angle, and vertical break clearly vary by pitch type. On the other hand, changes in spin rate and extension can be more subtle at times.

Variable Selection

One limitation of cluster analysis is that there is no built-in method to identify relevant variables. Therefore the variables included in the analysis must be chosen carefully as the clusters can be very dependent on these factors. Variables relating to the velocity, location, and movement of a pitch, like those presented in Figure 6, may be helpful in classifying pitch types.[12]

The following variables are good indicators of how a pitch moves through the air as it crosses home plate and will be included in the clustering.

1. **RelSpeed** - Speed of the pitch, reported in miles per hour, when it leaves the pitcher's hand (mph).
2. **VertRelAngle** - Initial vertical (up-down) direction of the ball when it leaves the pitcher's hand. A positive number means the ball is released upward, while a negative number means the ball is released downward (degrees).

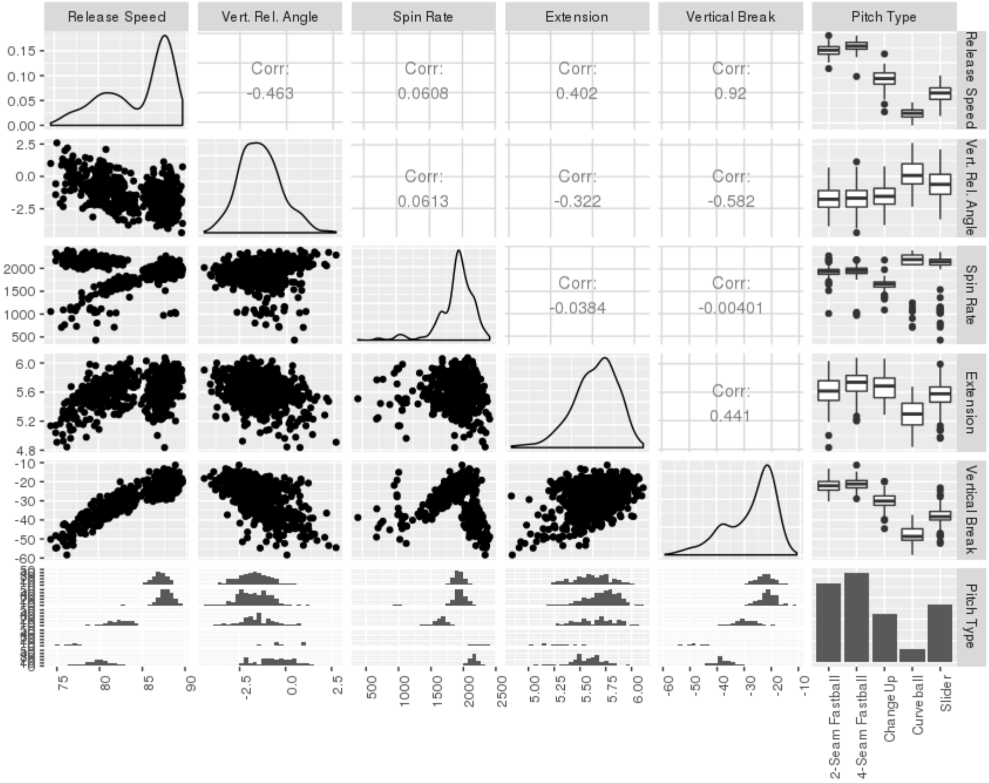


Figure 6: Scatterplot matrix for Pitcher X.

3. **HorzRelAngle** - Initial horizontal (left-right) direction of the ball when it leaves the pitcher's hand. A positive number means the ball is released to the right from the pitcher's perspective, while a negative number means the ball is released to the left from the pitcher's perspective (degrees).
4. **SpinRate** - How fast the ball is spinning as it leaves the pitcher's hand, reported in the number of times the pitched ball would spin per minute (revolutions per minute or rpm).
5. **RelHeight** - Height above home plate at which the pitcher releases the ball (feet).
6. **Extension** - The distance from which the pitcher releases the ball relative to the pitching rubber (feet).
7. **VertBreak** - Distance between where the pitch actually crosses the front of home plate height-wise, and where it would have crossed home plate height-wise if had it traveled in a perfectly straight line from release, completely unaffected by gravity (inches).
8. **HorzBreak** - Distance between where the pitch actually crosses the front of home plate side-wise, and where it would have crossed home plate side-wise if had it traveled in a perfectly straight line from release. A positive number means the break was to the right from the pitcher's perspective, while a negative number means the break was to the left from the pitcher's perspective (inches).[13]

K-Means Clustering

Now that we have chosen our variables, we can begin to utilize clustering methods. We begin with K-means clustering where the analyst must first choose the number of clusters, k, prior to analysis.

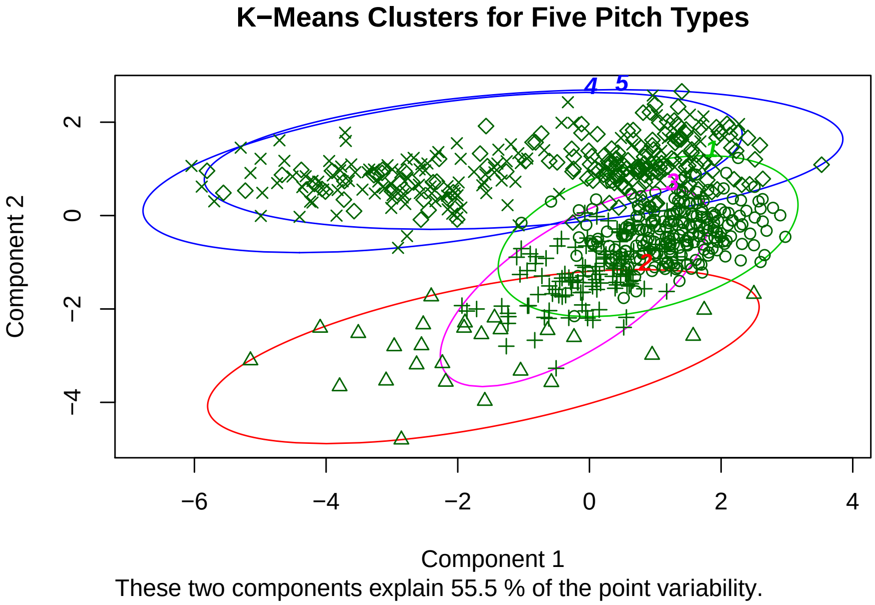


Figure 7: K-Means clustering attempt with five clusters.

This method then involves assigning each data point to a cluster, computing cluster centroids, reassigning points based on distance to centroids, and so on.[14]

Since we are required to choose the number of clusters, we can refer back to Pitcher X's arsenal in Figure 5. He is listed as throwing five different pitches, so let's begin clustering with $k = 5$. We can simply use the `kmeans()` function from the `stats` package to begin our analysis.[15]

K-Means Cluster	ChangeUp	Curveball	4-Seam Fastball	2-Seam Fastball	Slider
1	10	0	106	110	0
2	1	7	3	1	15
3	98	0	4	11	2
4	1	19	7	7	85
5	1	4	88	55	33

Table 3: Pitch type distributions of each k-means cluster for Pitcher X.

From Table 3 we can see that some of the clusters contain multiple pitch types. For example, cluster 5 has 88 4-seam fastballs, 55 2-seam fastballs, and 33 sliders. Fastballs and sliders are very different in terms of velocity and spin rate, so ideally we would not want this much overlap within one cluster.

K-Means Cluster	Average Release Speed	Average Spin Rate
1	87.18	1887.50
2	79.87	987.88
3	82.71	1635.61
4	80.40	2203.09
5	85.92	2026.38

Table 4: Attributes of pitches in k-means clusters.

Table 4 shows the average velocity and spin rate of pitches in each cluster. Some clusters have similar velocities and spin rates, so it is not completely clear which clusters represent which pitch types. To investigate further, let's examine a visual representation of these clusters.

In Figure 7 we can see that some of the clusters overlap; the two outlined in blue (Clusters 4 and 5) are essentially one cluster, while the other three overlap significantly. We want separate clusters with minimal overlap, and it seems that using $k = 5$ from the pitching arsenal for the number of clusters is not working well.

It seems that using the k-means clustering method is not producing the results we want. Instead, we can explore other methods to classify the pitches thrown by Pitcher X.[16]

Model-Based Clustering

Now let's approach the same question using a different method. To implement model-based clustering, we can use the `mclust` package.[17] Figure 8 shows that the best model chosen by the algorithm is *VVE*. Recall from Figure 3 that *VVE* represents an elliptical distribution with variable shape and volume and equal orientation. This makes sense because some types of pitches are more variable in their characteristics than others; the characteristics of a fastball vary much less than the characteristics of a slider, mostly due to the differences in spin rate and break. Additionally, the optimal number of clusters chosen by this model was 4. This differs from Pitcher X's arsenal (Figure 5), which lists 5 different pitches. This discrepancy may point to an error in the arsenal.

To represent these clusters visually, cluster grouping plots are very helpful and reveal more information about the clusters.

Figure 9 displays the different sizes, shapes, and locations of the clusters. We can see how some clusters are more circular in shape (cluster 1), suggesting equal mean and variance, whereas some are more elliptical (cluster 3), suggesting varying mean and variance. These different shapes help us think about the pitch types that each cluster represents.

When trying to compare the classifications made by the model to the manually tagged pitch types (tagged by the TrackMan operator in real time during the games), the accuracy can be represented in a table of counts.

Model-Based Cluster	ChangeUp	Curveball	4-Seam Fastball	2-Seam Fastball	Slider
1	2	0	204	182	0
2	105	1	1	1	2
3	2	23	0	0	118
4	2	6	3	1	15

Table 5: Comparison of model-based classifications and manually tagged pitch types.

Table 5 reveals a lot of information about the clusters. It can be inferred that Cluster 1 is "4-Seam/2-Seam Fastballs", Cluster 2 is "Changeups", Cluster 3 is "Sliders", and Cluster 4 is unclear (perhaps it is curveballs, with the uncertainty stemming from the small sample size). It is apparent that the model classification is grouping 4-seam and 2-seam fastballs together, and that it is also having trouble with Cluster 4.

We can also examine averages of certain attributes of each cluster to aid in matching the clusters to specific pitch types.[18] Table 6 displays the average velocity (speed) and spin rate of pitches in

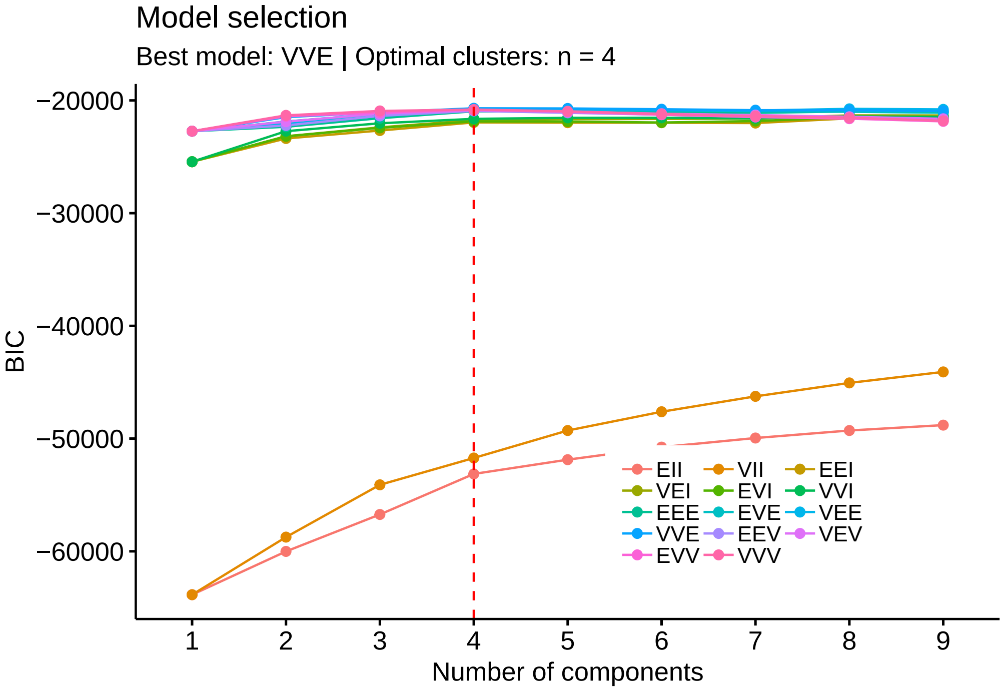


Figure 8: Model chosen and optimal number of clusters from model-based clustering.

each cluster. For insight into the range of velocities associated with different pitch types, we can refer back to Figure 5.

Model-Based Cluster	Average Velocity	Average Spin Rate
1	87.43	1941.00
2	82.23	1636.47
3	79.43	2170.42
4	80.07	992.44

Table 6: Average attributes of pitches across clusters.

The arsenal for Pitcher X includes an estimation of velocities for each pitch. The associated velocities are:

1. 4-Seam Fastball at 88-91 mph
2. 2-Seam Fastball at 86-89 mph
3. Changeup at 81-85 mph
4. Slider at 77-81 mph
5. Curveball at 75-77 mph.

Based on these velocities, we can be more confident in our initial assignments to Clusters 1, 2, and 3 (4-seam/2-seam Fastballs, Changeups, and Sliders, respectively).

The average spin rates of different pitch types can also be examined for reference. We can study the relationship between velocity and spin rate for all pitches in the data set. Figure 10 shows patterns hidden in this relationship across different pitch types. Several inferences can be made from this figure; changeups generally have the lowest spin rate, fastballs possess the greatest velocity, and

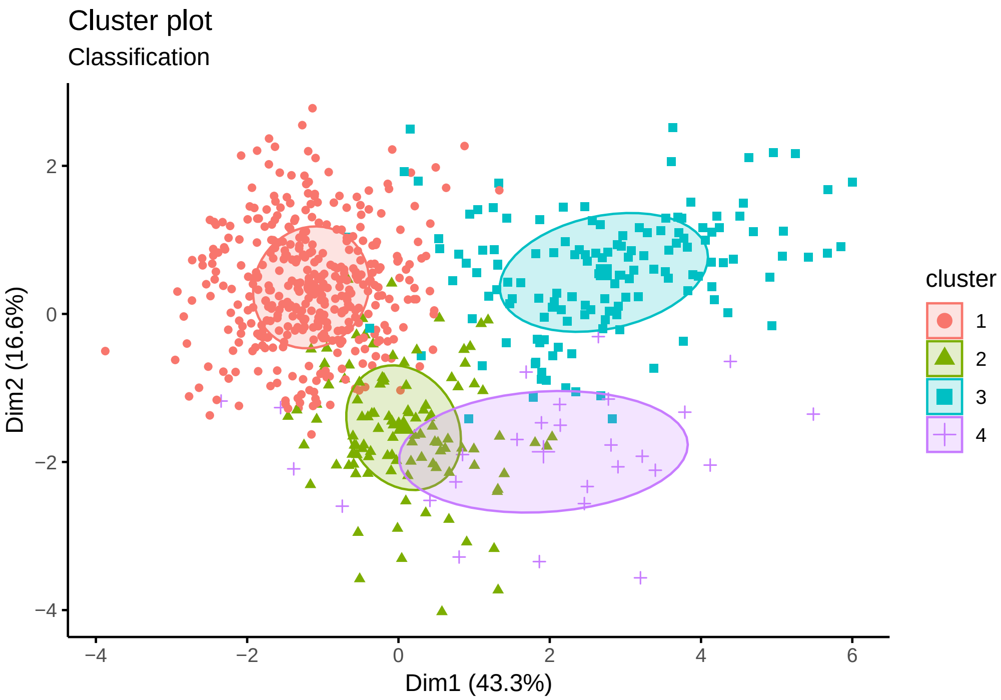


Figure 9: Distribution of first and second dimensions of clusters chosen by model for Pitcher X.

curveballs and sliders have the highest spin rates. This plot furthers our conjectures about the cluster assignments.

When following the attributes of Cluster 4 in the plot, we can see that there is no clear pitch type associated with that particular area. It is possible that Cluster 4 is simply a conglomeration of outliers that had peculiar velocities and spin rates, either due to a TrackMan system error or unusual pitcher mechanics.[19]

Evaluating Clustering Performance

While the k-means approach did not provide clear distinctions between pitch types, model-based clustering proved very useful in our quest to classify Pitcher X's different pitch types. We can be quite confident in our conclusions that Cluster 1 is 4-Seam/2-Seam Fastballs, Cluster 2 is Changeups, and Cluster 3 is Sliders. Due to a smaller number of pitches, Cluster 4 remains unclear.

With insight from the TrackMan operator, these classifications make sense. The operator informed us that while both Curveball and Slider are listed in the pitching arsenal, Pitcher X actually preferred to throw Sliders, although the coaches did encourage him to throw Curveballs in certain situations. These two pitch types are known as breaking balls, which means they both break or drop as they cross home plate. It is not uncommon for a pitcher at this minor league level to choose one or the other, instead of throwing both. At the start of our analysis we determined that Pitcher X did not throw many Curveballs (as seen in Table 1). Since Cluster 4 contains a small number of pitches, it is likely that this cluster does indeed represent Curveballs. The small sample size does not make the cluster as clear or defined as the others, but based on this outside information

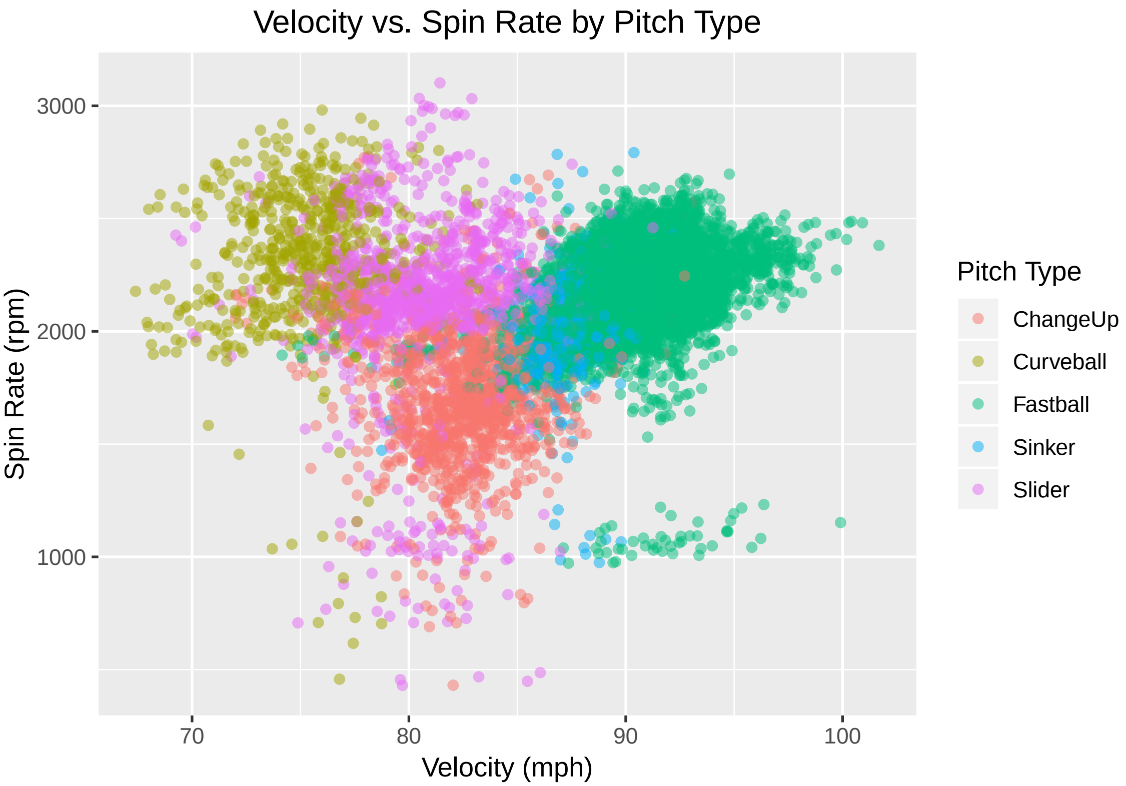


Figure 10: Relationship between velocity and spin rate for different pitch types.

it seems we can label Cluster 4 as Curveballs.

Additionally, 4-seam and 2-seam fastballs were grouped together in the same cluster. This could be a result of the similar characteristics of these two pitch types, but it is also unusual for a pitcher at this level to throw two different types of fastballs; usually they pick one or the other. The TrackMan operator pointed out that they were more likely to input a pitch as a 4-seam or a 2-seam at random simply because they were both listed in the pitching arsenal. The pitching arsenal is the only preliminary information provided to operators, so it is not surprising that they may rely on it for assistance at times of uncertainty. After conferring with Pitcher X, he informed us that he only threw 2-seam fastballs during this season. Thus it seems that the 4-seam fastballs in the data set were mislabeled, and that they were actually all 2-seam fastballs. The model-based clustering method correctly identified this discrepancy as it placed both fastball types inside one cluster.

It is clear that the pitching arsenal is an important resource that TrackMan operators use when collecting and inputting data. Thus it is important for these arsenals to be accurate and representative of the different pitch types thrown by each pitcher. This motivates the objective to implement clustering to correctly classify pitches that a pitcher actually throws. Overall we determined that model-based clustering was very useful in this situation as we were not required to choose the number of clusters beforehand; this method identifies hidden patterns in the data and chooses the optimal number of clusters itself. In Pitcher X's case, this clustering method identified discrepancies between his pitching arsenal and the pitch types he would actually throw during a game. From this analysis we would suggest that Pitcher X's arsenal be updated to only include 2-seam fastballs, changeups, sliders, and curveballs, with a note that he rarely throws Curveballs.

3 Future Work and Broader Applications

Model-based clustering can be utilized by Major League Baseball to improve the accuracy of TrackMan's automated pitch classifier and to create accurate arsenals for each pitcher. This would help decrease the frequency of mislabeled pitch types and increase the overall accuracy of pitch-by-pitch data. Other player statistics are often calculated from pitch data. For example, analysts could determine that 60% of the hits given up by Pitcher X occurred when he threw his Slider. This information would be misleading if the pitch data was not as accurate as it could be. Thus it is important to ensure that this baseline data is accurate and truly representative of Pitcher X.

While the use of model-based clustering is only demonstrated here in the context of pitch classification, there are several other scenarios in which this clustering method is implemented. Finite Gaussian mixture models can be applied to data where observations belong to certain groups and the group associations are unknown, or they can provide approximations for multi-modal distributions. For example, model-based clustering may be used when trying to identify earthquake epicenters.[20] The earthquake prediction problem is of great importance to the geosciences and society alike. Progress in this area is limited mainly because many of the variables associated with these occurrences are not accessible for direct observations. However, clustering can be implemented to identify and reveal more information about continental faults; most earthquakes should be clustered along these faults, and determining their location and magnitude could uncover valuable information. Model-based clustering is also useful in the realm of marketing. This machine learning technique can help marketers identify distinct groups in their customer bases by certain characteristics. They can then use this knowledge to develop targeted marketing initiatives to improve their marketing strategies.

Model-based clustering and the `mclust` package are gaining traction and recognition as an alternative clustering procedure comparable to k-means and hierarchical methods. The key advantage of model-based clustering is its model recommendation and its choice of the optimal number of clusters. This paper demonstrated how k-means is not always an appropriate or useful method, and in those cases model-based clustering may be implemented to achieve the desired result. Classification is an important statistical tool in many contexts, and model-based clustering is a useful resource for classifying various types of data, from pitch types in baseball to earthquake epicenters in the geosciences.

Bibliography

[1] Turkin, Hy, and Sherley C. Thompson. *The Official Encyclopedia of Baseball.* A.S. Barnes & Co., 1951.

[2] *The Baseball Encyclopedia: The Complete and Official Record of Major League Baseball.* The Macmillan Company and Information Concepts Inc., 1969.

[3] James, Bill. *The Bill James Historical Baseball Abstract.* Villard Books, 1985.

[4] Lahman, Sean. "Baseball in the Age of Big Data." SeanLahman.com, 4 Aug. 2013.

[5] "What Is PITCHf/x?" *FanGraphs Baseball.*

[6] "Cricket and Baseball Fielding Positions." Right Off the Bat, Oct. 2011.

[7] Wang, Haizhou, and Mingzhou Song. "Optimal k-Means Clustering in One Dimension by Dynamic Programming." *The R Journal*, Dec. 2011.

[8] Kodali, Teja. "Hierarchical Clustering in R." *R-Bloggers*, 22 Jan. 2016.

[9] "Cluster Analysis in R." *Girke Lab*, 11 May 2018.

[10] Leisch, Friedrich. *A General Framework for Finite Mixture Models and Latent Class Regression in R.* Journal of Statistical Software, 2004.

[11] Fop, Michael, et al. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal*, Aug. 2016.

[12] Sievert, Carson, and Brian Mills. "Using Publicly Available Baseball Data to Measure and Evaluate Pitching Performance." *Handbook of Statistical Methods and Analyses in Sport.*

[13] "Radar Measurement Glossary of Terms." *TrackMan*, Nov. 2018.

[14] Doukkali, Firdaouss. "Clustering Using K-Means Algorithm." *Towards Data Science*, 19 Dec. 2017.

[15] Jaiswal, Sejal. "K-Means Clustering in R." *DataCamp*, 14 Mar. 2018.

[16] Mills, Brian. "Pitch Classification with K-Means Clustering." *Exploring Baseball Data with R*, 23 Feb. 2015.

[17] Kassambara, Alboukadel. "Model Based Clustering Essentials." *Datanovia*, Oct. 2018.

[18] Cornish, Rosie. "Cluster Analysis." *Mathematics Learning*, 2007.

[19] Mills, Brian. "Pitch Classification with Mclust." *Exploring Baseball Data with R*, 12 Jan. 2015.

[20] Yuen, David, et al. *Earthquake Clusters Over Multi-Dimensional Space.* 2004.