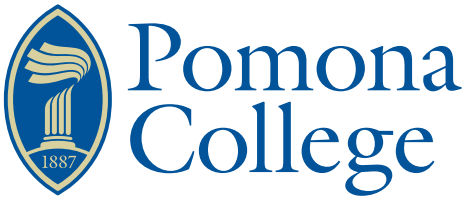

Senior Thesis in Mathematics

Modeling Missing Data in NCAA Division-III Baseball

Author:
Adam Hinthorne

Advisor:
Dr. Gabriel Chandler

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

May 24, 2018

Chapter 1

Introduction

1.1 Background

In 2002, the Oakland Athletics revolutionized the use of batting statistics in major league baseball. The 2003 book by Michael Lewis, *Moneyball: The Art of Winning and Unfair Game* chronicles how the Athletics overcame a limited payroll by building their team using players that were undervalued based on their conventional stats. The Athletics' method of using statistics revolutionized how Major League Baseball ("MLB") teams built their teams and new technologies that measure everything from spin rate on a pitchers curve ball to ball flight have even allowed teams to optimize in game strategy. But, the access to this technology is limited to only the upper echelons of baseball. In NCAA Division III ("D-III"), programs have neither the money to acquire the technology to provide accurate sabermetric data or the manpower to process the data, leaving D-III teams with significantly inferior data that limits their ability to execute in game strategy. The purpose of this paper is to use existing statistical models and MLB data to generate usable sabermetrics for D-III.

1.2 Data Introduction

One common visualization MLB teams utilize for in game strategy is a spray chart. A spray chart is a scatter plot of where hitters have hit the baseball in previous at bats. This information is useful because some batters will be consistent as to where they hit the ball and reach base successfully. With this data, managers can use this data to optimally position defensive players. For example, in figure 1.1, we have the spray chart for David Ortiz of the Boston Red Sox. From the spray chart we can see that Ortiz has no hits through the left side of the infield.

When a manager sees this kind of spray chart they may employ a defensive positioning called the "Lefty Shift" (See Figure 1.2), which is when managers move the majority of their players to one side of the field.

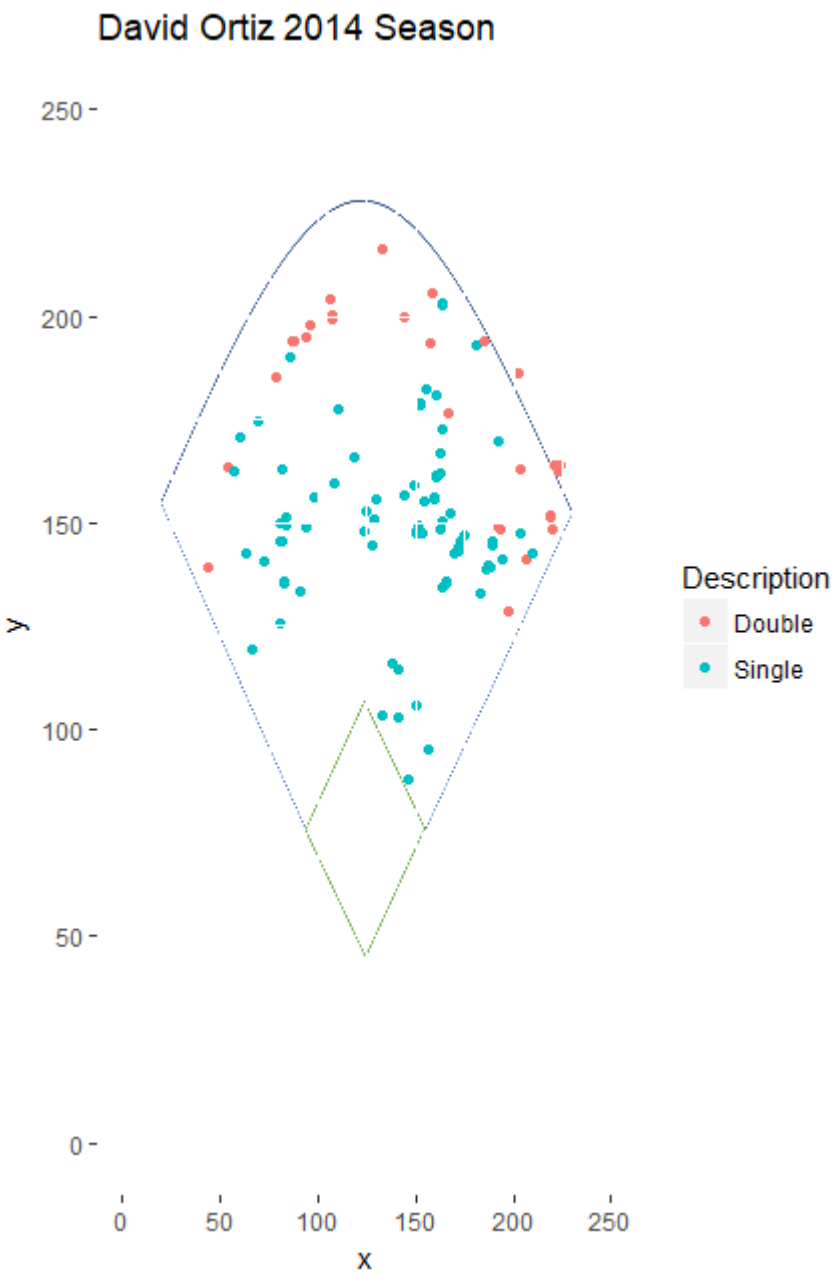

Figure 1.1: David Ortiz Hits Spray Chart 2014 Season

Figure 1.2: The New York Yankees shift three infielders to the right side of second base to defend David Ortiz of the Boston Red Sox

This kind of analysis is possible with MLB data because in every game, the location of every ball put into play is recorded and available for every team in the league. For D-III baseball, the data available on opposing teams is limited to play by plays pulled from team websites. Instead of location, which can be coded as x and y coordinates, D-III data is coded as follows:

1. A. Hinthorne singles to right field
2. R. Smith flies out to center field
3. J. Thomas doubles to right center field.

The problem with this play by play data is that it does not give enough information about the location of the data. A "double to left center field" could have been a hard hit that sneaks by the shortstop, a ball that is more likely to be a single rather than a double, or it could have been a ball hit off the wall, a ball more similar to a home run. Since each hit in a play by play data set can be very different, D-III teams cannot optimize strategy like MLB teams. This paper will try to use MLB data to give us a better understanding of D-III location data, so that it may be used as in game strategy.

1.3 Methodology and Hypothesis

The best way to relate the MLB location data to the D-III play by play data is by understanding that a given hitter is similar to some hitters and different to

others. For example, David Ortiz may be similar to another hitter who does not hit the ball through left side of the infield, but he is different from a hitter that only hits the ball through the left side of the infield. More generally, we can say that there are "types" of hitters, where hitters have similar hit location data to hitters of the same type, and different location data for hitters of a different type. However, types of hits may also differ between different types of hitters and we are unable to determine the difference in types of hits from D-III data. But, using the MLB data, we can get an idea for what a "double to left center field" looks like across all different types of hitters. From this information, we should then be able to classify where a D-III hitter may hit the ball like a MLB hitter, based on the MLB data. But, for there to be any impact of defensive positioning and in-game strategy in D-III, a "double to left center field" has to be different from one hitter to another. Thus, the hypothesis is for a given event, different hitter types will have different hit locations. For example, a "type 1" hitter and a "double to left center field" event suggests hit location will somewhere around where the shortstop is typically positioned, whereas a "type 2" hitter and a "double to left center field" event suggests that the hit location will be near the wall in left center field. To investigate this hypothesis, this paper will explore non-parametric density estimation, Kullback-Leibler divergence, and the Expectation Maximization algorithm and their applications to baseball sabermetrics. The MLB data for this paper is from the 2014 MLB regular season, which was collected by Mr. Danny Malter while creating a spray chart shiny app (https://github.com/danmalter). The D-III data from this paper is collected from the 2017 season from Pomona-Pitzer's website, using scrape code created by Professor Gabe Chandler at Pomona College (https://github.com/GabeChandler).

Chapter 2

Non-Parametric Density Estimation

2.1 Model Introduction

From the introduction, the MLB data can be coded as a scatter plot of the location of each hit for each hitter. While this is a good method to understand where a hitter has hit the ball in the past, it provides little information about where a hitter may hit the ball in the future, because the probability that a hitter will hit the ball to an exact location again is nearly zero. But, intuitively, a hitter that hits the ball into the same general area many times, will probably hit the ball there again in the future. Thus, when looking at a hitter's scatter plot, managers are intuitively thinking about a hitter's probability density, assigning a probability on if a hitter will hit the ball to an area and then positioning his players accordingly. For example, when a manager sees David Ortiz's scatter plot, he is more likely to be thinking about the representation in figure 2.1. So, to understand where a hitter may hit the ball in the future we want to convert a scatter plot to a contour plot representing a two dimensional probability density. Since, similar hitters have hit the ball to the same places in the past, they will probably hit the ball to the same place in the future probability densities can be used to classify the hitter types and understand the differences in types of hitters for each event. The following chapter will discuss the theory behind density estimation and how to optimize the density estimation.

2.2 Density Estimation

The goal is to estimate $\hat{f}_i$, which represents the probability density over a finite support (the baseball field) for player i, where player i is a collection of hit locations on the support. This first section will discuss model performance and introduce how to use of the bandwidth h to optimize the performance of the

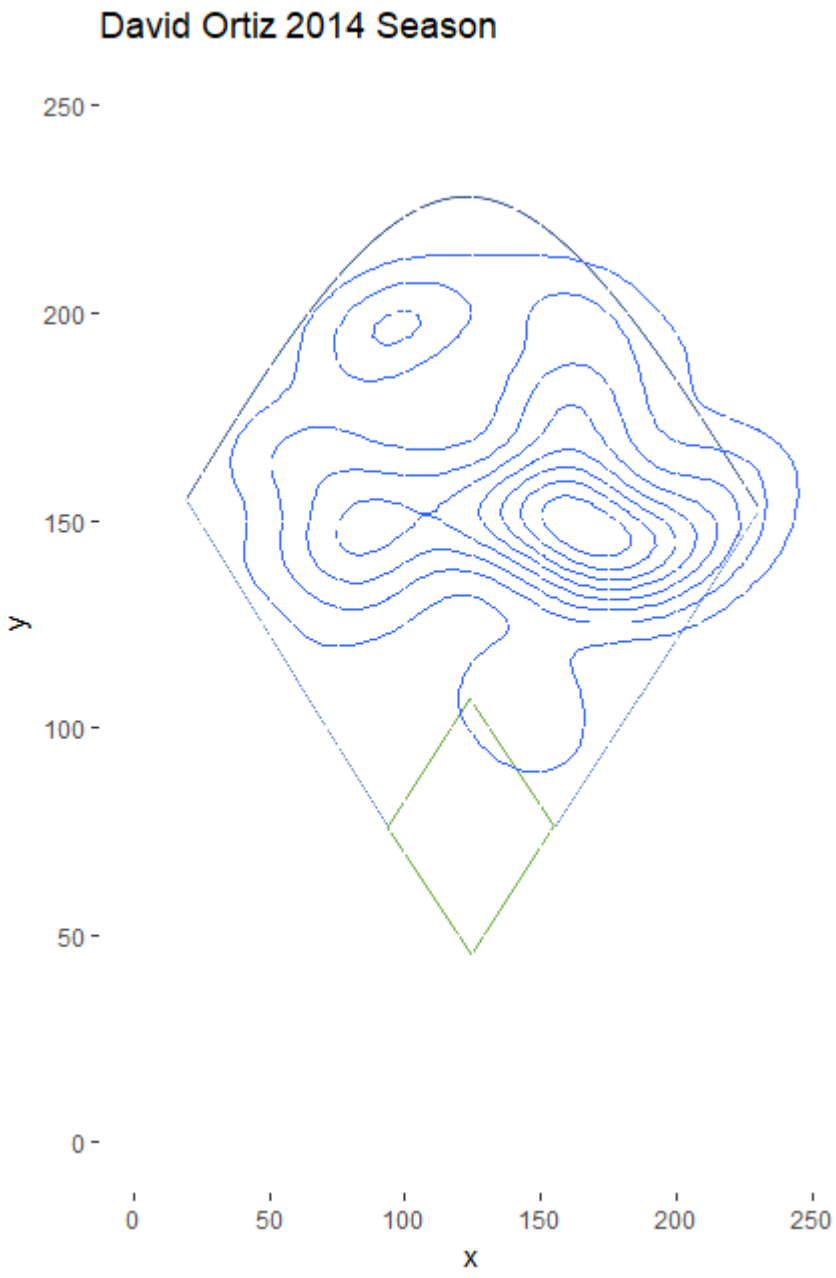

Figure 2.1: Contour plot representing the probability density describing David Ortiz's hits during the 2014 Season

estimator.

2.2.1 Performance

Nonparametric density estimators use local information to estimate a function $f(x)$ at a point x on the support. This function is assumed to be smooth on some level, but there are no specific assumptions made about the distribution of the function. A basic example of a nonparametric estimator is a histogram, where the map of a function is made by determining the frequency of data points within a bin, which is a local range that does not intersect with other bins and the union of all bins equals the support. A general kernel density estimator (in its discrete form) is given below where n is the number of observations, X_1, X_2,X_i are the observed data, K is a kernel function, and h is the bandwidth:

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

A kernel function assigns weights depending on the proximity between the data X_i and the estimation point x to the density estimator $\hat{f}_i$. Kernel functions tend to be positive definite and symmetric about zero [5]. The bandwidth, h, is a smoothing factor that determines how locally the estimator assigns probability around observed data. In practice, h can be thought of as the bin size, where a small h assigns probability locally around observed data, similar to a small bin, and a large h assigns probability to a large area around observed data, similar to a large bin.

The performance of the model is defined as how well the model represents the data. This is important because we want models to predict function given the observations. The typical measure of performance for density estimators is the mean integrated squared error (MISE), defined as:

$$MISE(h) = \int MSE_h(\hat{f}(x))dx$$

Where MSE is the mean squared error, defined as:

$$MSE_h(\hat{f}(x)) = E(\hat{f}(x) - f(x))^2 = var\hat{f}(x) + (bias\hat{f}(x))^2$$

The goal is to minimize this error, so as to create the most accurate estimation of the model. The key to minimizing the error is balancing the bias variance trade-off for the fixed number of observations n in our sample. We wish to choose an h so that the performance of the model is optimized. A low bandwidth will not smooth the data enough, resulting in a lot of variance. [5] A high bandwidth will smooth the data too much, resulting a lot of bias. An example of different choices of bandwidth are represented in figure 2.2.

Picking the optimal (or near optimal) bandwidth is crucial to picking an estimator with the best performance. It is also important for our purposes because if we pick a bad estimator to describe the hitters, we will not be able to rely on $\hat{f}_i$ to represent the observed data.

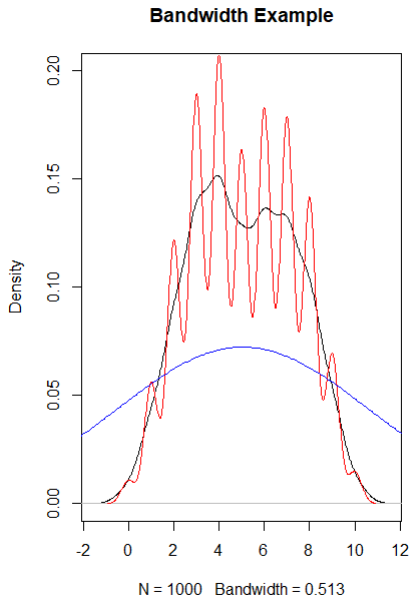

Figure 2.2: This plot is a density estimation of 1000 random draws from a binomial distribution. The red line represents a bandwidth of .3, which is too low, creating too much variance. The blue line represents a bandwidth of 5, which is too high, creating too much bias. The black line is a reasonable estimate of the optimal bandwidth of about .5, balancing the bias-variance trade off.

2.3 Bandwidth Selection

The optimal bandwidth can be derived from our definition of MISE. Assume that our kernel function K is a continuous probability density function with $\mu = 0$ and a variance $0 < \sigma_K^2 < \infty$. Let $R(g)$ be the roughness of a function g, which is defined as:

$$R(g) = \int g^2(z)dz$$

As $f(x)$ is assumed to have some level of smoothness, it thus has some level of roughness, and $R(K) < \infty$. Let $f(x)$ be the function we are trying to estimate. In order for the derivation to work we need to assume that f is sufficiently smooth by stating that it has two bounded, continuous derivatives. This is equivalent to defining the roughness of the second derivative to be less than infinity, i.e. $R(f'') < \infty$. Recall our definition of MISE and MSE:

$$MISE(h) = \int MSE_h(\hat{f}(x))\delta x = \int var(\hat{f}(x)) + \int (bias(\hat{f}(x)))^2\delta x$$

We want to analyze the bias-variance trade off as $h \to 0$, $n \to \infty$, and $nh \to \infty$. But, it is important to note that n needs to approach infinity faster than h approaches zero, as given by $nh \to \infty$. Let's analyze the variance and bias separately, starting with the bias. [5]

Using our general kernel density estimator, the expected value of $\hat{f}$ is as follows:

$$E(\hat{f}(x)) = \frac{1}{h}\int K(\frac{x-u}{h})(f(u))du$$

Using a change of variables we can manipulate the equation to be:

$$E(\hat{f}(x)) = \int K(t)(f(x-h(t))dt$$

To approximate this equation we replace $f(x-h(t))$ with its Taylor Series:

$$f(x-h(t)) = f(x) - htf'(x) + h^2t^2\frac{f''(x)}{2} + \mathcal{O}h^2$$

$\mathcal{O}h^2$ represents the rest of the Taylor series, which all have a shared factor h^2. After substituting the Taylor series expansion, we get the following expression:

$$E(\hat{f}(x)) = f(x)\int K(t)dt - hf'(x)\int K(t)tdt + h^2f''(x)\int K(t)t^2dt + \mathcal{O}(h^2)\int K(t)dt$$

From our assumptions that K is symmetric with $\mu = 0$ we see that the second term multiplied by the first derivative of f disappears, and the expression simplifies to:

$$E(\hat{f}(x)) = f(x) + h^2\sigma_K^2 t^2 f''(x)/2 + \mathcal{O}(h^2)$$

The absolute difference between the expected value of $\hat{f}$ and f is the bias. It is important to note that as $h \to 0$, $\mathcal{O}(h^2) \to 0$ faster than $h^2 \to 0$. Due to this fact we can simply ignore $\mathcal{O}(h^2)$ as $h \to 0$. Looking back to our original equation:

$$(bias(\hat{f}(x)))^2 = h^4\sigma_K^2[f''(x)]^2/4 + \mathcal{O}(h^2)$$

Integrating over the quantity as in our original equation gives us:

$$\int (bias(\hat{f}(x)))^2 = h^4\sigma_K^4 R(f'')^2/4 + \mathcal{O}(h^4)$$

The important points of this equality is that squared bias and thus the expected value of $\hat{f}$ moves on the order of h^4. This fits our intuition that as h increases and the density estimator becomes smooth, the estimator will exhibit more bias. Now we turn our attention to the variance.

$$var(\hat{f}(x)) = \frac{1}{n} var(\frac{1}{h}K(\frac{x - X_i}{h})$$

Using a change of variables and the definition of variance ($var = E(x^2) - (E(x))^2$), we can manipulate the expression to:

$$var(\hat{f}(x)) = \frac{1}{nh}\int K(t)^2[f(x) + \mathcal{O}1]\delta t - \frac{1}{n}[f(x) + \mathcal{O}(1)]^2$$

Which reduces to:

$$var(\hat{f}(x)) = \frac{1}{nh} f(x)R(K) + \mathcal{O}\frac{1}{nh}$$

Taking the integral of this expression, there is a rather clean result:

$$\int var(\hat{f}(x)) = \frac{R(K)}{nh} + \mathcal{O}\frac{1}{nh}$$

As we can see from this expression, the variance tends to decrease as $nh \to \infty$.[5] This fits with our intuition over a finite support, where we describe the variance as :$Var(\bar{X}) = \frac{\sigma^2}{n}$. This derivation suggests that the variance of our estimator $\hat{f}$ is $Var(\bar{X}) = \frac{\sigma^2}{nh}$, which makes sense from sense low bandwidth implies high variance and high bandwidth implies low variance. Now, returning to MISE, we can see the bias-variance trade off.

$$MISE(h) = \int MSE_h(\hat{f}(x))\delta x = \int var(\hat{f}(x)) + \int (bias(\hat{f}(x)))^2\delta x$$

$$MISE(h) = \frac{R(K)}{nh} + \frac{h^4\sigma_K^4 R(f'')}{4} + \mathcal{O}(\frac{1}{nh} + h^4)$$

The key terms that need to be balanced are the order of the variance $\frac{1}{nh}$ and the order of the squared bias h^4. We previously stated that $h \to 0$, $n \to \infty$, and $nh \to \infty$. This makes sense because as n increases and h decreases the error will decrease and we will have a perfect model. However, in reality there is never an infinite amount of observations, and we need to pick the h that instead balances the two terms. By minimizing the first two terms of our new MISE with respect to h we can find the optimal value for h.

$$h = (\frac{R(K)}{n\sigma_K^4 R(f'')})^{\frac{1}{5}}$$

The important part of the optimal bandwidth equation is the relation between h and n, which turns out to be $h = n^{-\frac{1}{5}}$. [5] Again, the intuition makes sense, as n increases the bandwidth should shrink to maintain the balance between the bias and the variance. This relationship between h and n doesn't let $\frac{1}{nh}$ dominate h^4 or vice versa. In practice, it is difficult to find a roughness function $R(K)$ or a function f that fits our data. Thus, it is near impossible to pick the optimal bandwidth, but there are ways to maintain the optimal relationship. One such way to maintain the optimal relationship is Silverman's rule. Silverman's rule suggests that you replace f with a normal distribution with a variance set to match the sample variance. This method derives a bandwidth of:

$$h = (\frac{4}{3n})^{1/5}\hat{\sigma}$$

We can see that the relationship of $h = n^{-\frac{1}{5}}$ is present in this rule, and Silverman proves that at a large n the bandwidth is close to its optimal value for a balance between the bias and the variance. The multivariate case follows the result for bandwidth, calculating h in each dimension x and y of the location data.

2.4 Data Application

One of the beneficial properties of density estimation is that two or more densities can be averaged to create a new density that encapsulates new data. For our purposes, this is how we will create densities for D-III players. Over the course of the MLB season, we will use data from all hitters to create "hit events", E. These hit events come from dividing the support into the areas that are described in the play by plays (e.g. "double to left center field"), then using the data from all hitters in that confined area to create a density. D-III players can then be described using a weighted average of these densities, depending on how many times an event occurred. Thus, we have a density $\hat{f}_i$ for a D-III player equal to the following:

$$\hat{f}_i = \frac{\sum \hat{f}_E w_{i,E}}{\sum w_{i,E}}$$

$\hat{f}_E$ represents the event density and $w_{i,E}$ represents the assigned weight for each player i and E calculated by dividing the hit event over the total number of hits. However, there is still unobserved data for D-III hitters: type. This same method of averaging densities can be applied to multiple MLB player densities and we can use those densities to run a clustering algorithm, to determine hitter types.

Chapter 3

The Expectation Maximization Algorithm

Now that we have the ability to calculate probability densities for hitters in both the MLB and D-III, we need a way to estimate the unobserved hitter "type". In baseball there is some intuition for hitter type, where a left handed singles hitter hits the ball to locations at a different frequency than a right handed home run hitter (resulting in a different density). But, the classification of hitter type should be chosen using the densities for hitters so as not to introduce other biases like size, speed, or a player on your favorite MLB team. For this purpose, we will explore the Expectation-Maximization ("EM") algorithm which will allow us to estimate hitter types from MLB densities. The density for a type of hitter should be the average of the densities of all hitters of that type.

3.1 EM Algorithm in Theory

The EM Algorithm is a data driven algorithm that estimates parameters from unobserved data. The algorithm uses an expectation ("E") step and a maximization ("M") step that by calculating the expected distribution of the data, and then given then maximizing the likelihood of the parameter with respect to the expected value. After repeating the E-step with the value calculated in the M-step, the algorithm eventually converges with a reasonable estimate of the parameters (the algorithm "converges" when the estimated parameter in step i and $i+1$ have an absolute difference smaller than a predetermined margin of error). We can articulate a simplified example of EM, as articulated by a number of coin flips [1]. Imagine you have two coins with different probabilities of landing heads. Let's label these parameters $\hat{\theta}_A$ and $\hat{\theta}_B$ where $\hat{\theta}$ represents the probability of heads. To estimate these parameters you flip each coin 30 times and calculate their Maximum Likelihood Estimator (MLE) using the following formula for a binomial: $\hat{\theta}_A = \frac{1}{n}\sum_1^n X_i^A$ and $\hat{\theta}_B = \frac{1}{n}\sum_1^n X_i^B$. This is reasonably

straight forward, but it becomes more complicated if the types $\hat{\theta}_A$ and $\hat{\theta}_B$ are unobserved. Essentially, you are left trying to estimate a parameter $\hat{\theta}_?$. The EM algorithm seeks to solve this problem.

When we know which coin corresponds to which series of flips, we are calculating the likelihood with respect to the complete data Y.

$$L(\theta|Y)$$

However, when we don't know which parameter we are trying to estimate, we can think of the complete data Y as a combination of two observations X and Z. X is the data we observe and Z is the missing or latent data. For our coin flips, X represents the data from the flips and Z the corresponding coin. The EM algorithm wants to maximize

$$L(\theta|X)$$

with respect to theta. The likelihood with respect to X is a marginalization of the total data likelihood, allowing us to estimate the total data likelihood. Due to multiple iterations, let $\theta^{(t)}$ denote the estimated parameter for iteration t for $t = 0, 1, 2, ...$. Define $Q(\theta|\theta^t)$ as the expectation of the joint log likelihood of Y, conditional on the observed data X.

$$Q(\theta|\theta^{(t)}) = E\{logL(\theta|Y)|x, \theta^{(t)}\}$$

Using this framework, how we mathematically think about the E-step and the M-step is as follows [5]:

- **E step**: Compute $Q(\theta|\theta^{(t)})$
- **M step**: Maximize $Q(\theta|\theta^{(t)})$ with respect to θ. This becomes $\theta^{(t+1)}$
- Return to E step until algorithm converges.

In our coin flip example, we can start at $t = 0$ by providing a guess for θ_A and θ_B. Then we use the data to calculate the distribution of the data using our initialized theta in the E step. In the M-step, maximizing the likelihood of theta is the MLE of the new distribution of data. This gives us $\theta^{(t+1)}$, which we plug back into the E step. This process is repeated until the estimates in E step for iteration n and $n + 1$ have a difference which is sufficiently small based on what makes sense for the given data.

Returning to our coin flip example [1], where $\hat{\theta}_A$, $\hat{\theta}_B$ represent the probability of heads for coin A and coin B respectively. First, with equal probability, we select a coin and flip that coin ten times. We then repeat this procedure five times, generating the following observations.

H	T	T	T	H	H	T	H	T	H	5H,5T
H	H	H	H	T	H	H	H	H	H	9H,1T
H	T	H	H	H	H	H	T	H	H	8H,2T
H	T	H	T	T	T	H	H	T	T	4H,6T
T	H	H	H	T	H	H	H	T	H	7H,3T

Our best guess for each coin is that $\hat{\theta}_A = .6$ and $\hat{\theta}_B = .5$. Now, for the E-step we calculate a probability distribution over possible completions using our best guess. So, for our first ten flips in the first row we calculate:

$$\frac{L(5|\theta_A)}{L(5|\theta_A) + L(5|\theta_B)} = .45$$

Thus we can attribute $.45 * 5 = 2.2$ heads in the first column to coin A, and 2.8 heads to coin B. Doing this for every ten flips gives us 21.3 heads and 8.6 tails attributed to coin A, and 11.7 heads and 8.4 tails attributed to coin B. This ends our E-step, and now we maximize the likelihood in the M step by calculating the MLE, $\theta_A = \frac{21.3}{21.3+8.6} = .7124$ and $\theta_B = \frac{11.7}{11.7+8.4} = .5821$. These new values become our new best guess and we then repeat the process until convergence. [1]

3.1.1 EM convergence

The proof of EM convergence was formalized in Dempster et al. 1977 and is briefly described here. We will start with the log likelihood of the distribution with respect to X represented as a function of the estimation of the joint likelihood of Y minus the conditional likelihood of Z with respect to X, which we will define as $H(\theta|\theta^{(t)})$

$$log f_X(x|\theta) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)})$$

The proof describes how $H(\theta|\theta^{(t)})$ is maximized when $\theta = \theta^{(t)}$. To see this we calculate the difference between $H(\theta|\theta^{(t)})$ and $H(\theta^{(t)}|\theta^{(t)})$.

$$H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)}) = \int -log \frac{f_{z|x}(z|x,\theta)}{f_{z|x}(z|x,\theta^{(t)})} f_{z|x}(z|x,\theta^{(t)}) dz$$

We also know that the ratio between the conditional probability with respect to θ and conditional probability with respect to $\theta^{(t)}$ is greater than zero and less than one, which means the following identity (see below) is true because log of the ratio will result in a value greater than one, and will be maximized when the ration is equal to one.[5]

$$\int -log \frac{f_{z|x}(z|x,\theta)}{f_{z|x}(z|x,\theta^{(t)})} f_{z|x}(z|x,\theta^{(t)}) dz \geq -log \int f_{z|x}(z|x,\theta)$$

Thus, $H(\theta|\theta^{(t)})$ is the maximized when $\theta = \theta^{(t)}$, and we know that the EM algorithm converges. Now, we have the framework to apply a variation of the algorithm to D-III baseball.

3.1.2 EM in practice: K-means Algorithm

The K-means algorithm is an algorithm that partitions a data set into K non-overlapping clusters by following the expectation and maximization steps we

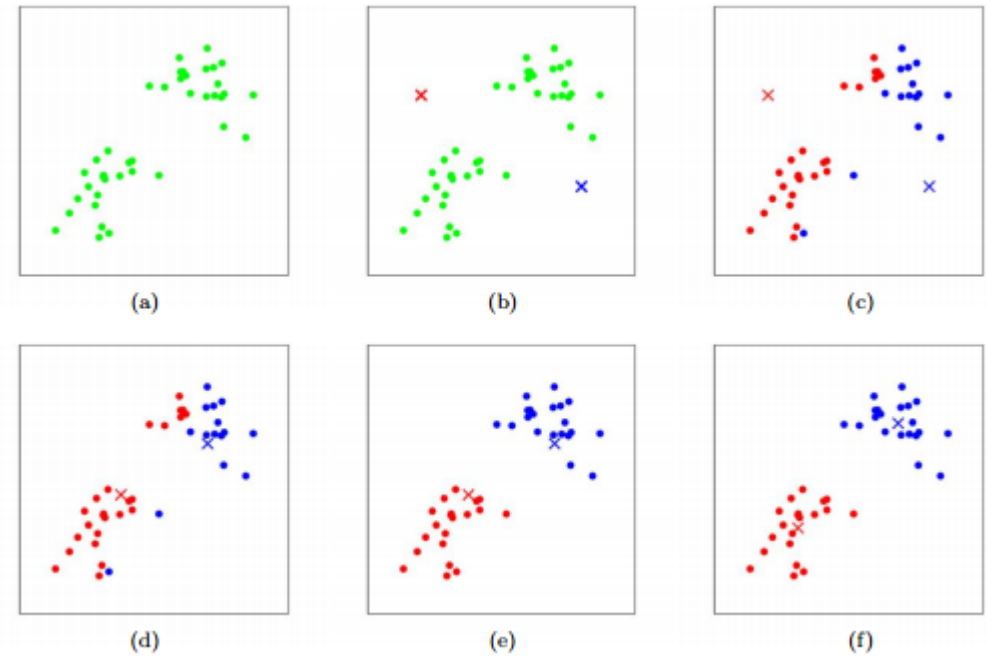

Figure 3.1: K-means algorithm using two clusters in two dimensional space. Step B is the random selection of centers. Steps C and E are E-steps. Steps D and F are M-steps. We can also see that in this instance the algorithm has converged quickly.

are familiar with in our EM algorithm. First, we decide on the number of clusters K and randomly select where the first cluster centroids will be. Then, for the E-Step we assign observations to the closest cluster center, assigning each observation to the closest expectation. Then in the maximization step, we calculate new cluster centers by finding the point that minimizes the sum of squared distances between the new cluster center and all the points in that cluster. Next, return to the E-step to repeat the process until there are no observations to switch clusters between iterations. Figure 3.1 shows a step by step rendition of this process.

One question still remaining about K-means is how to choose the correct number of clusters to fit the data. The key is to balance a trade off between the distance between cluster centers and the distance between points and the cluster center to which they are assigned.[4] Ideally, the distance between cluster centers should be large because we want there to be a difference between points in each cluster. This can be artificially created by choosing a small number of clusters. But, distance between points and their center should be small, suggesting that clustered observations are close to their cluster mean. This is achieved by choosing a large number of clusters. Generally, we choose the number of

clusters at the point when decreasing the number of clusters shows diminishing marginal increase in distance between centers and when increasing the number of clusters shows diminishing marginal decrease in the sum of distance within clusters.

3.2 K-means Data Application

From chapter two, probability densities can be averaged in an understandable way. Thus, the output of our clustering algorithm should be a probability density of hitter type j, where $\bar{f}$ represents the average of each density estimation for player i.

$$\bar{f}_j = \frac{\sum \hat{f_{i,j}} w_{i,j}}{\sum w_{i,j}}$$

Here, since each player is weighted equally, the $w_{i,j}$ will be equal to one for each player. From the clustering algorithm, these densities should be different from one another.

However, there is a problem when applying the K-means algorithm to the probability densities of MLB hitters. Unlike points in a Euclidean plane, we do not have an immediate intuition for how to measure the distance between probability densities. For the K-means algorithm to work, some densities need to be farther away from each other and some densities need to be closer together. But, densities often contain some information that overlaps, (i.e. two densities may well each have a probability assigned to some point x on the support), and some information that does not overlap. In the next chapter, we will exploit this fact to create a distance metric that can be used for K-means clustering.

Chapter 4

Kullback-Leibler Divergence

4.1 Kullback-Leibler Introduction

Now that we have a way to describe $\hat{f}$ for each player through density estimation and a way, in theory, to get from $\hat{f}_i$ (for MLB players) and $\hat{f}_{E,i}$ (for D-III players) to $\hat{f}_{i,j}$ for player i, type j, and event E through a weak Estimation Maximization algorithm, but we need a way to measure the difference, or specifically the distance between, these densities. The following chapter in this paper discusses how mathematicians have created a way to measure the difference between densities using Kullback-Leibler (KL) Divergence. KL divergence uses the information contained in probability densities to measure how different they are. From KL divergence, metrics have been developed to make KL divergence symmetric and follow the triangle inequality using Jensen Shannon divergence. At this point, since the metric is symmetric and satisfies the triangle inequality, we can say that it is a distance. From this distance metric, we will be able to run K-means clustering, and develop different hitter types.

4.2 Kullback Leibler Divergence

4.2.1 "On Information and Sufficiency"

In 1951 Solomon Kullback and Richard Leibler developed a method to measure the relative entropy between two probability densities. In their paper "On information and sufficiency, they describe the use of the method as,"...two populations differ more or less according as to how difficult it is to discriminate between them with the best test [6]. Two populations that share a lot of information are difficult to use, so Kullback-Leibler seeks to differentiate them by weighting information that two populations share less than information that is different in two populations. The Kullback-Leibler result is derived from

information theory and the relative entropy between points. The following subsections describe the information theory behind Kullback-Leibler and the theory itself.

4.2.2 Entropy

Entropy is the measure of uncertainty of a random variable. Given a random variable X, the entropy is defined by:

$$H(X) = -\sum p(x) \log_2 p(x)$$

This equation can also be defined as the expected value of the random variable $\frac{1}{\log p(x)}$, which is defined by the expression:

$$H(X) = E_p(\frac{1}{\log p(x)})$$

The definitions give an intuitive result: if a binomial random variable has a distribution where the probability of success is equal to either 0 or 1, the entropy of the function is zero, because the outcome of the random variable is certain.[2] But, if the random variable has a probability of $\frac{1}{2}$, there is an entropy because there is much more uncertainty of success. However, the definition of entropy can be expanded to comparing two different probability densities through joint entropy, conditional entropy, and mutual information.

Joint entropy is the total amount of entropy shared between two random variables. Similar to the union of two sets, it defines the entropy that is in either individually and both random variables. Joint entropy can be defined as:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

This expression can also be expressed as an expected value, similar to a single measurement of entropy:

$$H(X, Y) = -E(\log p(X, Y))$$

Joint entropy can be thought of as the combination of conditional entropy and mutual information. Conditional entropy is the entropy of one random variable given its relationship to another. So, the conditional entropy of the random variable X given Y is the amount of entropy that the information in Y contributes to the total entropy. As such, conditional entropy can be defined as:

$$H(X|Y) = \sum_{x \in X} p(x) H(Y|X = x)$$

Again, the expression can be represented as an expected value.

$$H(X|Y) = -E(\log p(Y|X))$$

One important result from joint entropy and conditional entropy is that joint entropy is the sum of the entropy of X and the conditional entropy of Y. The theorem is proved below [2]:

Theorem

$$H(X,Y) = H(X) + H(Y|X)$$

Proof

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y|x)$$

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

$$H(X,Y) = -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x)p(y|x)$$

$$H(X,Y) = H(X) + H(Y|X)$$

■

Now that we have defined conditional and joint entropy we can define mutual information. Mutual information is the entropy that is contributed by both random variables to the joint entropy. In other words, it is a measure of the amount of information that one random variable shares with another random variable and is the reduction in the uncertainty of one random variable due to the knowledge of the other. Following our other definitions, mutual information between random variables X and Y can be defined as:

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

This expression can be simplified to an expression of entropies that we know in the following steps:

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)}$$

$$I(X;Y) = -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(x,y)\log p(x|y)$$

$$I(X;Y) = -\sum_{x,y} p(x,y)\log p(x)\sum_{x,y} p(x,y)\log p(x|y)$$

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

To check if the final expression is accurate, it should follow that the mutual information plus the two conditional probabilities equals the total joint entropy of the random variables X and Y.

$$I(X;Y) + H(X|Y) + H(Y|X) = H(X,Y)$$

Applying our definition of mutual information, we get:

$$H(X) + H(Y) - H(X,Y) + H(X|Y) + H(Y|X) = H(X,Y)$$

Recalling that the joint entropy is equal to the entropy of one random variable plus the conditional entropy of the other we can see that this equation simplifies nicely, proving our result:

$$H(X) + H(Y|X) + H(Y) + H(X|Y) - H(X|Y) = H(X,Y)$$
$$H(X,Y) + H(X,Y) - H(X,Y) = H(X,Y)$$
$$H(X,Y) = H(X,Y)$$

The study of entropy and the foundations of information theory form the basis of Kullback-Leibler divergence, which we can think of as relative entropy. [2]

4.2.3 Relative Entropy and Kullback-Leibler Divergence

To follow the development of Kullback-Leibler Divergence it is useful to define entropy in a slightly different way. Entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. The information required to describe a random variable could be considered as the conditional entropy, or the entropy contributed only by a given random variable. Relative entropy compares the conditional entropy between two random variables. In this action it describes the difference in one relative to the other. The way Kullback Leibler separates the conditional entropies from the mutual information is through the logarithm of the likelihood ratio. So, given $p(x)$ and $q(x)$ are probability densities over the same support (which are effectively probability distributions), the logarithmic likelihood ratio is defined as:

$$\log \frac{p(x)}{q(x)}$$

This relationship is powerful because it separates the conditional entropy from the mutual information with respect to $p(x)$ [2]. As we can see if at a given x, $p(x) = q(x)$, the likelihood ratio returns a value of zero. This intuitively makes sense, because the mutual information of two random variables will not be able to tell us anything about how the two random variables are different on average. From the logarithmic likelihood ratio, we can define Kullback-Leibler as:

$$KL(p,q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

This identity can also be referred to as the relative entropy of $p(x)$ in relation to $q(x)$. As we can see from our earlier definitions of entropy, Kullback-Leibler is the combination of a distributions own entropy with the ratio of its conditional entropy to another distribution's conditional entropy. This result is powerful because we can measure the difference in information between one set of information and another to make a prediction about what the density group or type as we defined them. However, before designating the groups, Kullback-Leibler is not symmetric and does not satisfy the triangle inequality, which we need it to be in order to work with the K-means clustering algorithm.

4.3 Jensen Shannon Divergence

To solve the problem of symmetry and the triangle inequality in applied fields, mathematicians in fields such as quantum mechanics [7] have utilized Jensen-Shannon Divergence, an alteration of Kullback-Leibler. Let M be the average density of densities P and Q such that $M = \frac{1}{2}(P + Q)$. Then Jensen Shannon divergence is as follows:

$$JSD(P||Q) = \frac{1}{2}(KL(P|M) + KL(Q|M))$$

Unlike Kullback-Leibler, this identity is symmetric and the square root of this metric satisfies the triangle inequality. Thus, after taking the square root, we can use our Jensen Shannon distance to run the K-means clustering Algorithm.

4.4 Model Application

In this chapter we have shown that there exists a metric, Jensen-Shannon distance, that allows us to introduce a distance between probability densities. Thus, we can run our clustering algorithm on the probability densities for MLB players (i), $\hat{f}_i$ to determine hitter type (j) densities $\bar{f}_j$. The final step is to use Jensen Shannon distance to assign the probability density for a D-III player across all events E, $\hat{f}_{i,E}$ to the closest cluster, thus defining a D-III hitter's type. But, we have yet to show whether or not this will allow D-III managers to optimize in game strategy. The next chapter will explore

Chapter 5

Results and Conclusion

The data set is from the 2014 MLB baseball season, capturing 412 hitters resulting in $48,786$ hits. Due to computational constraints, only 100 hitters were selected at random from the sample amounting to a total of $12,146$ hits. From the play by play data, there are 14 total events, so the MLB data is divided into 14 distinct regions to created the event densities.

5.1 Density Results

One of the goals of density estimation was to use the individual event densities to estimate the density of a D-III hitter. Figures 5.1 and 5.2 are contour plots that show the result of the weighted average of event densities used to calculate $\bar{f}$ for two sample D-III players who played for the Pomona-Pitzer Sagehens in the Southern California Intercollegiate Athletic Conference. The densities to show similar characteristics of where other hitters hit the ball in the MLB. The first D-III player, Adam Hinthorne, hits mostly singles with few extra base hits, which is consistent with the plot in figure 5.1. The second D-III player, Tanner Nishioka, hits more doubles and a higher proportion of his hits to the outfield, consistent with the plot in figure 5.2.These plots show the usefulness of density estimation for D-III hitters when scatter plots are unavailable.

5.2 Clustering Results

For the K-means algorithm, we found that six clusters, and thus six types of hitters, as the optimal number of clusters for the algorithm. The primary reason for choosing six types of hitters is that adding more clusters resulted in some clusters being comprised of a single hitter, thus suggesting that there was no more to be gained by adding more clusters. The second reason is that for the analysis to be useful, there needs to be a difference between hitter types, and a fewer clusters would create little distinction between hitter types. The following list gives example hitters for the six different hitter types:

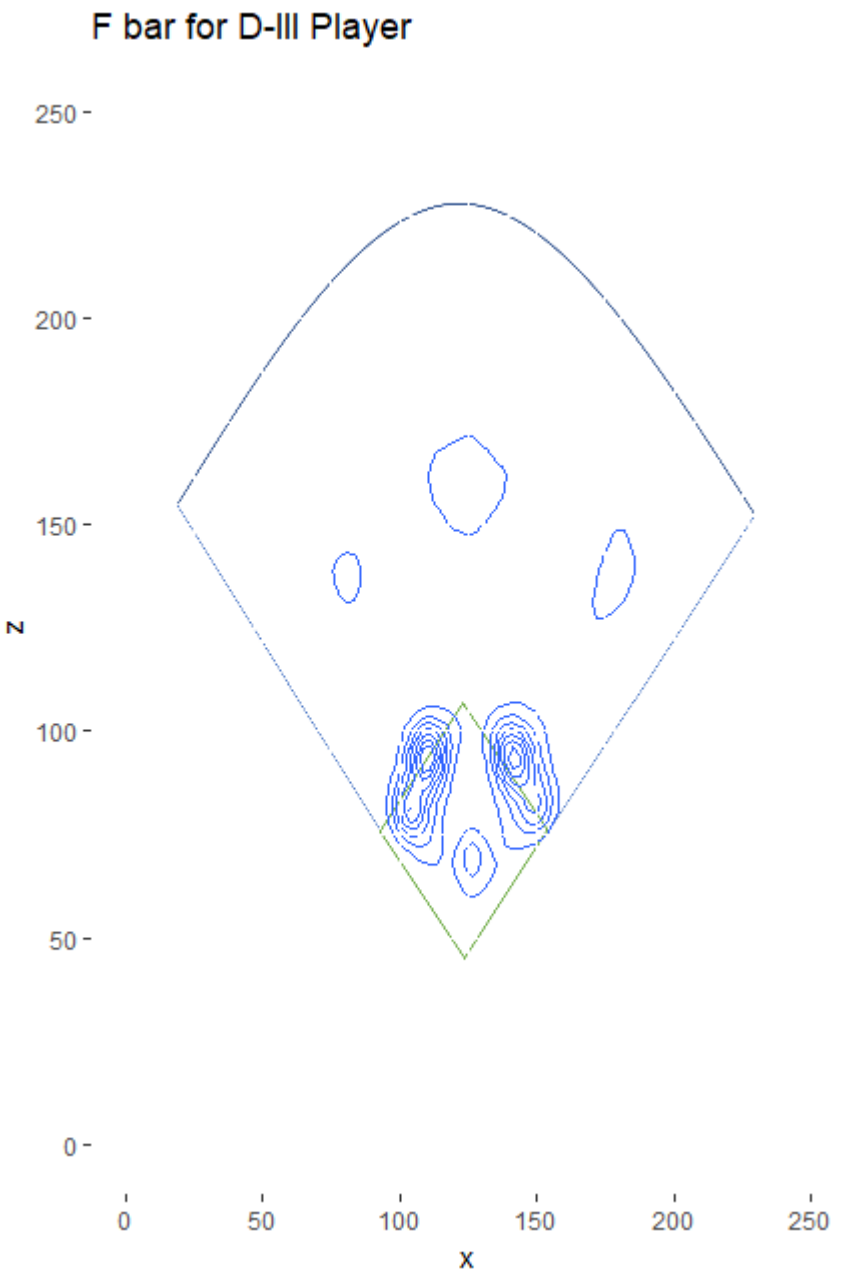

Figure 5.1: This figure shows $\bar{f}$ for Adam Hinthorne, a D-III player, using a weighted average of the observed events from the play by play data.This D-III hitter closely resembled type four hitters.

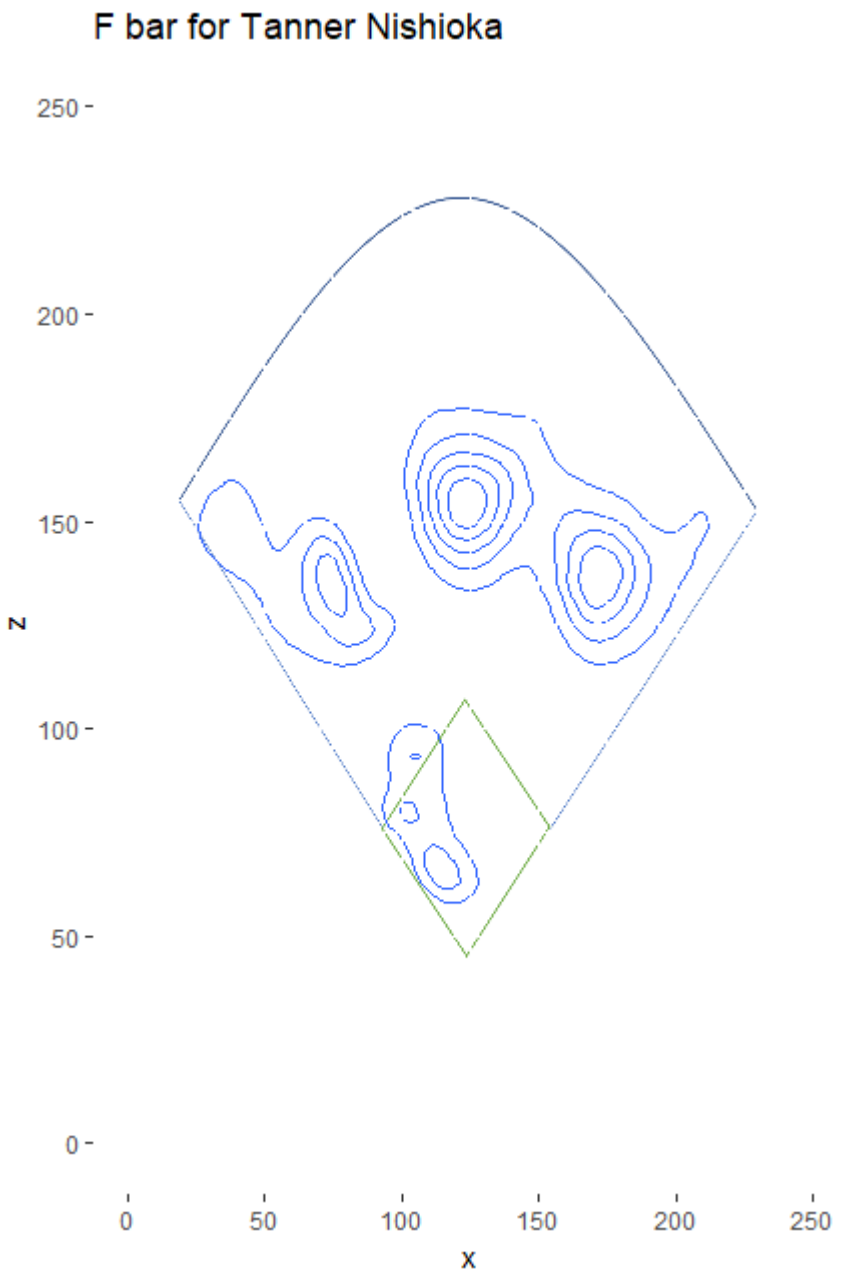

Figure 5.2: This figure shows $\bar{f}$ for another a D-III player, Tanner Nishioka, using a weighted average of the observed events from the play by play data. This D-III hitter closely resembled type three hitters.

- Type 1 Hitters: Ryan Howard, Kyle Seager, Kole Calhoun
- Type 2 Hitters: Bryce Harper, Matt Carpenter, Charlie Blackmon
- Type 3 Hitters: Howie Kendrik, Jason Hayward, Brad Miller
- Type 4 Hitters: Dee Gordon, Billy Hamilton, Leonys Martin
- Type 5 Hitters: Yasiel Puig, Albert Pujols, Mike Trout
- Type 6 Hitters: Yonder Alonso, Robinson Cano, Kolten Wong

At first glance, hitters within types seem to be relatively similar. For example, the hitters classified as type 1 seem to be left handed power hitters and type 5 seem to be right handed power hitters, whereas type 4 seems to be faster players that hit for average. We can also see this from analyzing figure 5.3. Type 1 and type 5 have the heavier part of their densities centered in the outfield, consistent with power hitting. While type 4 has a heavy concentration in the infield, suggesting a lot of infield hits.

Using our D-III hitter density, we can measure the distance between each of the aggregate densities for hitter type and the estimated D-III hitter density. According to our Jensen Shannon Divergence metric, the Adam Hinthorne's density is closes to the density of type 4 hitters. This makes sense looking at the contour plots because both the D-III player and the type 4 hitters seem to have a large concentration of infield hits. Whereas Tanner Nishioka's density resembles the density of type 3 hitters, which show a good mix of singles and doubles to right field and down the left field line. Nishoka's density was also close to type 5 hitters using Jensen Shannon divergence, but was slightly closer to the type 4 density, likely due to hits to the right side of the field.

5.3 Hypothesis Results

The key hypothesis for this paper was that for each type of hitter their event densities look differently. This is a key assumption relating to the usefulness of relating D-III hitters to MLB hitters since we create the D-III hitter densities from events. We want to be able to relate the density for an event and a hitter type to the D-III play by play information, then to shift the defense to where they get their hits.

To test whether event densities look differently we ran a permutation test for each event. The null hypothesis for this test is that each event density $\hat{f}_{j,E}$ that we calculated from MLB hitter data are not different (i.e. have a small distance between them). The alternate hypothesis is that these densities are different (i.e. a large distance) and we can use them to understand D-III data. To perform the permutation test, we calculated the density $\hat{f}_{j,E}$ for each hitter type j and each event E and using Jensen-Shannon distance, calculated the total distance between all densities. Then we randomly assigned clusters to each of the hundred players, took new densities $\hat{f}_{j,E,t}$ where t represents the iteration.

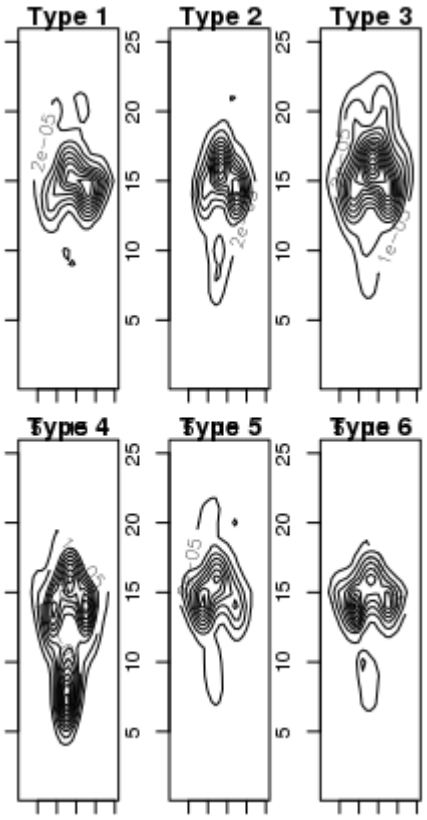

Figure 5.3: This figure shows the aggregate densities $\bar{f}_{i,j}$ for hitters i in hitter type j

Then we measured the total distance between the new densities, and recorded if that new distance was greater or smaller than the original distance between hitters of type j and event E. We repeated this process for 500 iterations and took the proportion of values that were greater than the original distance. The proportion of permutations with a greater total distance than the original distance represents the p-value. If the p-value is high it means that the event densities for hitter type are not different because a random cluster of densities creates aggregate densities with greater difference.

The chart below records the p-values for different event types. The p-values that are statistically significant at the 5 percent level are marked with an asterisk. The NA values in the table represent a lack of data for a specific event. From our knowledge of baseball, NA values makes sense for certain events. For example, a "double up the middle" would refer to a hitter that reaches second on a ball that lands between the second baseman and the shortstop, something that is nearly impossible.

Event Location	Singles P-Values	Doubles P-Values	Triples P-Values
Left Field Line	NA	.822	NA
Left Field	.024*	.066	NA
Left Center Field	.040*	.380	.604
Center Field	.750	.006*	.802
Right Center Field	.868	.026*	.948
Right Field	.662	.164	.940
Right Field Line	NA	.224	.826
Third Base	.736	NA	NA
Short Stop	.556	NA	NA
Up the Middle	.932	NA	NA
Second Base	.278	NA	NA
First Base	.964	NA	NA
Left Side	.690	NA	NA
Right Side	.800	NA	NA

5.4 Conclusion

From our permutation test, only four out of 42 events are statistically significant at a five percent level. But, the 12 types of hits that may make the most impact on defensive positioning are singles and doubles to the outfield. From out intuition, a single up the middle is not going to be that much different from hitter to hitter because "up the middle" is a relatively small area on the field compared to "double to left field". Therefore it is helpful that the four events which are statistically significant are hits to the outfield. These results suggest that positioning to defend doubles in the gaps and singles to left field may be the most important types of events to defend against (given that a batter hits the ball their with a high enough frequency), because those events are the most

dissimilar between types of hitters. Figure 5.3 shows the different densities of singles to left field, which suggests a type 4 hitter hits their singles shallow towards the left filed line, whereas a type 1 hitter hits those singles deeper towards the corner, which suggests that the left fielder should play shallower to type 4 hitters.

5.5 Acknowledgements

I would like to thank my advisor, Professor Chandler, for guiding me through the process of writing the thesis and the rest of the math department for their support while I worked through the major.

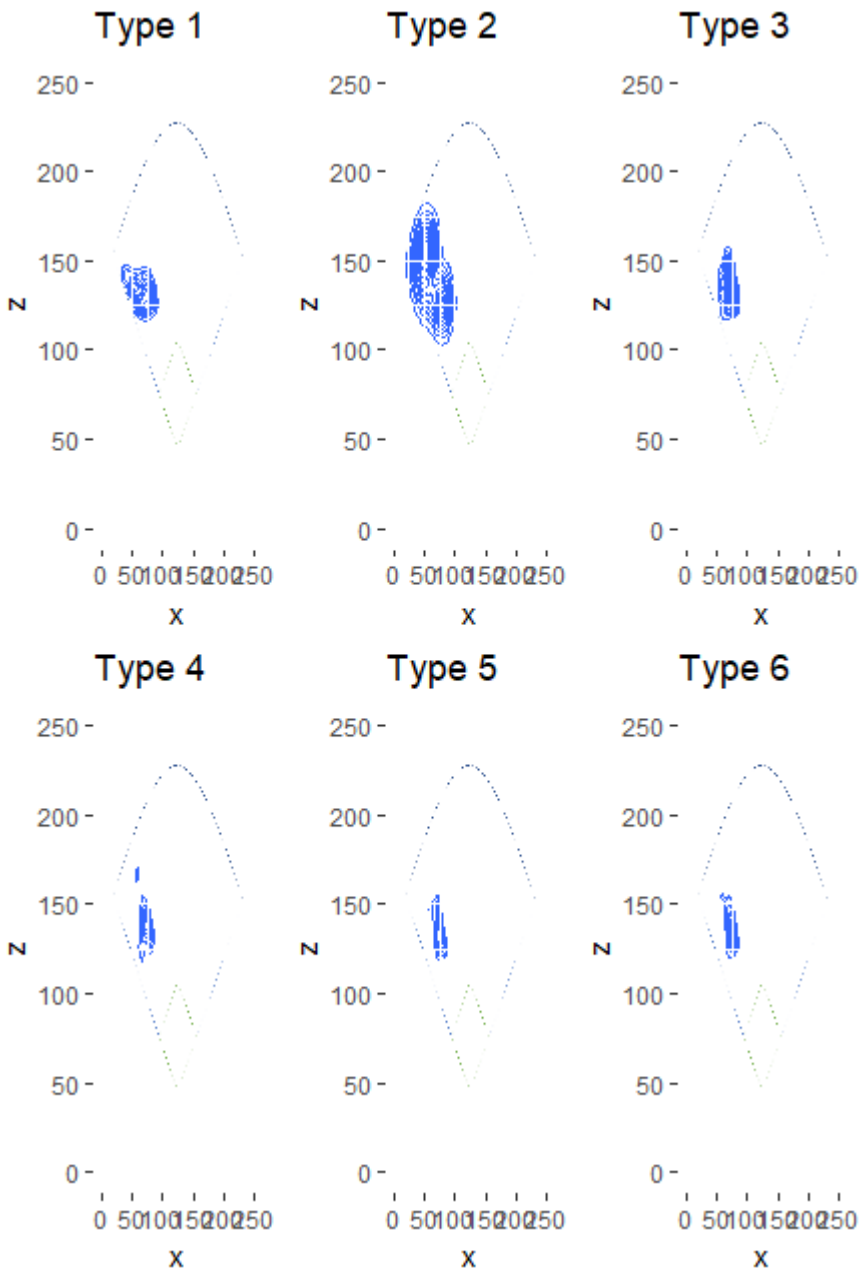

Figure 5.4: This figure shows the densities for the event "Single to Left Field", one of the few statistically significant events.

Bibliography

[1] Batzoglou, Serafim, Do, Chuong B; "What is the expectation maximization algorithm?", Nature Biotechnology 26, 897-899 (2008)

[2] Cover, Thomas. M, Thoms, Joy A.; "Elements of Information Theory", Second Edition, pp. 1-100

[3] Dempster, A; Laird, N.; Rubin.; 1977."Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Series B, 39(1): 1-38

[4] Gareth, James, Hastie, Trevor, Tibshirani, Robert, witten, Daniela: "An introduction to Statistical Learning", Springer Texts in Statistics, New York, NY. (2001), pp 386-387

[5] Givens, Geof H., Hoeting, Jennifer A; Computational Statistics, Wiley Series in Probability and Satistics, **277-284**, Hoboken, New Jersey, 2005

[6] Kullback, S.,Leibler, R.A.; "On Information and Sufficiency",Annals of Mathematical Statistics. (2005). 22 (1):79-86

[7] Majtey, A.P., Lamberti, P.W., Prato, D.P.; "Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states", Physical Review (2005)