Title

Using Multi-Camera and Radar Trajectory Data to Learn and Predict Performance in Baseball

Author

Zhao, Shiyuan

Publication Date

2021

UNIVERSITY OF CALIFORNIA,
IRVINE

Using Multi-Camera and Radar Trajectory Data to Learn and Predict Performance in Baseball

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Electrical Engineering

by

Shiyuan Zhao

Dissertation Committee:
Professor Glenn Healey, Chair
Professor Chen-Yu (Phillip) Sheu
Professor Michele Guindani

2021

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor, Professor Glenn Healey, who has been incredibly supportive and an integral part of my doctoral research. His involvement and guidance have fueled my progress in every step of the way, and I have learned a ton about what qualities a good researcher should have while working with him. Without his persistent help and patience, this dissertation would not have been possible.

I would like to thank Prof. Hal S. Stern, Phillip Sheu, Chin Lee, and Zhiying Wang, who have served on my candidacy exam committee. I would like to thank Prof. Phillip Sheu and Michele Guindani, who have served on my defense committee for their invaluable time, contributions, and helpful feedback to my research. I would like to thank Tom Tango and Mitchel Lichtman for their helpful comments. Specifically, I would like to thank Tom Tango for inspiring the research on the measurement space paritioning. I would also like to thank the anonymous reviewers for their very helpful feedback, which have helped improve my research.

I would like to thank Daniel Adler, Josh Kalk, and Heather Mac Murray for giving me the internship opportunity at Minnesota Twins and their invaluable advice, guidance and patience during my internship period.

I would like to thank all of the great friends who I have made during my stay at UCI, for their help and support, and also for making the life of graduate school less boring.

All pitch data used in this study was obtained from Brooks Baseball (www.brooksbaseball.net). The data used for computing strikeout rate was obtained from www.retrosheet.org. The batted ball data was obtained at baseballsavant.com.

VITA

Shiyuan Zhao

EDUCATION

Doctor of Philosophy in Electrical Engineering **2021**
University of California, Irvine *Irvine, California*

Master of Science in Electrical Engineering **2015**
University of California, Irvine *Irvine, California*

Bachelor of Engineering in Aeropace Engineering **2013**
Beihang University *Beijing, China*

RESEARCH EXPERIENCE

Graduate Student Researcher **2014–2021**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2014–2021**
University of California, Irvine *Irvine, California*

ABSTRACT OF THE DISSERTATION

Using Multi-Camera and Radar Trajectory Data to Learn and Predict Performance in Baseball

By

Shiyuan Zhao

Doctor of Philosophy in Electrical Engineering

University of California, Irvine, 2021

Professor Glenn Healey, Chair

Sensor systems that acquire large sets of data have been deployed to document sporting events at unprecedented levels of detail. Machine learning techniques have been applied to these sensor measurements to discover new skills, quantify known skills with greater accuracy, and understand biomechanical principles to improve performance and prevent injury. The use of learning methods to support the generation of predictive models has revolutionized decision making as teams search for an advantage in a highly competitive industry. Machine learning methods are particularly well suited for baseball due to the discrete structure of the sport.

We develop and apply learning methods to large sets of sensor data to address several of the most important and challenging problems in baseball analytics. We introduce a method for learning a function over distributions that generalizes nonparametric kernel regression by using the Wasserstein metric for distribution space. The technique is applied to the problem of learning the dependence of pitcher performance on multidimensional pitch distributions that are derived from sensor measurements which capture physical properties of each pitch. We also develop a method for estimation and prediction called measurement space partitioning. The method is applied to the problem of estimating batted-ball talent by using large sets of trajectory measurements acquired by in-game sensors to show that the predictive value of a batted ball depends on its physical properties. This knowledge is exploited to estimate batted-ball distributions defined over a multidimensional measurement space by using regression parameters that adapt to batted ball properties. This process is central to a new method for quantifying batted-ball skill. We present examples illustrating facets of the approach and use a set of experiments to show that the new methods generate predictions that are significantly more accurate than those generated using current methods.

Chapter 1

Introduction

An important use of machine learning techniques is the recovery of a model from observed data. The development of learning methods for the recovery of three-dimensional shape from image data, for example, has been a topic of recent interest in computer vision [7] [44]. The proliferation of sensor systems at sporting events has provided large data sets that support the generation of predictive models using machine learning algorithms. These models are playing an increasingly prominent role in the operational activities of professional sports teams. In an industry where the difference between success and failure is often small, models derived from sensor data can be used to gain an edge over the competition. In this work, we extend and apply learning techniques to multiple problems in baseball analytics.

In Chapter 2 we review the Wasserstein metric or Earth Mover's Distance (EMD) which can be used to compare distributions and has been applied to many problems in signal processing and machine learning [35]. We show how this metric can be configured to compare pitch and batted ball distributions derived from sensor measurements. The EMD is used in Chapter 3 to develop a method for measuring player similarity and we show how this method can be used for forecasting. In Chapter 4 we present a method for learning a function over distributions. This method is based on generalizing nonparametric kernel regression by using the EMD as a metric for distribution space. The technique is applied to the problem of learning the dependence of pitcher performance in baseball on multidimensional pitch distributions that are controlled by the pitcher. The distributions are derived from sensor measurements that capture the physical properties of each pitch. Finding this dependence allows the recovery of optimal pitch frequencies for individual pitchers. This application is amenable to the use of signatures to represent the distributions and a whitening step is employed to account for the correlations and variances of the pitch variables. Cross validation is used to optimize the kernel smoothing parameter. A set of experiments demonstrates that the new method accurately predicts changes in pitcher performance in response to changes in pitch distribution and also outperforms an existing technique for this application.

An important and challenging problem in the evaluation of baseball players is the quantification of batted-ball talent. This problem has traditionally been addressed using linear regression on the value of a statistic derived from a set of observations. In Chapter 5 we use large sets of trajectory measurements acquired by in-game sensors to show that the predictive value of a batted ball depends on its physical properties. This knowledge is exploited to estimate batted-ball distributions defined over a multidimensional measurement space from observed distributions by using regression parameters that adapt to batted ball properties. This process is central to a new method for estimating batted-ball talent. The domain of the batted-ball distributions is defined by a partition of measurement space that is selected to optimize the accuracy of the estimates. We present examples illustrating facets of the new approach and use a set of experiments to show that the new method generates estimates that are significantly more accurate than those generated using current methods. The new methodology supports the use of fine-grained contextual adjustments and we show that this process further improves the accuracy of the technique.

Chapter 2

The Earth Mover's Distance

2.1 Overview

The EMD is a standard method for computing the distance between distributions. The method utilizes a ground distance between individual points to determine the minimum amount of work that is required to transform one full distribution into the other. Small values of the EMD correspond to similar distributions while larger values correspond to less similar distributions.

For many applications [50], a distribution can be accurately represented as a signature S defined by a set of m clusters

$$S = \{(\mu_1, w_1), \ldots, (\mu_m, w_m)\} \tag{2.1}$$

where μ_i is the mean vector for cluster i and w_i is the fraction of the distribution represented by cluster i. Thus, the signature S approximates a distribution by a set of m point masses at the locations μ_i with the weights w_i where m depends on the distribution.

An established algorithm [50] for finding the EMD using signatures is based on the solution of the transportation problem [28] for finding the minimum cost to move product from a set of producers to a set of consumers with each having a known demand. For the transportation problem, the ground distance is the cost to move one unit of product from a given producer to a given consumer. The computation of the EMD can be formulated as a linear programming problem for which efficient solutions [27] and software [61] exist.

2.2 Sensor Data

A baseball game is defined by a set of one-on-one matchups between a pitcher and a batter. The pitcher throws a ball which the batter attempts to hit with a bat. Each throw is called a pitch and each matchup consists of one or more pitches. The pitcher's goal is to make it difficult for the batter to make solid contact with a pitch.

The PITCHf/x optical video and TrackMan(TM) phased-array Doppler radar sensors [22] capture data that is exploited to recover information about pitches and batted balls [14] [32]. Let s represent the initial speed of a pitch in three dimensions and let the pair (x, z) specify the pitch's movement as reported by Brooks Baseball (www.brooksbaseball.net). Let s_l represent the initial speed and v represent the launch angle of a batted ball hit by a batter as reported by Baseball Savant (baseballsavant.mlb.com). The parameter x is an estimate of the pitch horizontal movement between the release point and home plate relative to a theoretical pitch thrown at the same speed with no spin-induced movement and z is the corresponding estimate of vertical movement [46]. The coordinate system is arranged so that positive x is to the right from the catcher's perspective and positive z is up. The speed s is typically reported in miles per hour while x and z are reported in inches. The pitcher starts the process of throwing each pitch from a location that is 60.5 feet from home plate. By convention, Brooks Baseball reports s for $y = 55$ feet and (x, z) from $y = 40$ feet to home plate. The batted ball exit speed s_l is reported in miles per hour. Launch angle v represents the vertical angle at which the ball leaves a batter's bat in degrees relative to the ground plane. A ground ball has a v less than 10 degrees and a pop-up has a v greater than 50 degrees.

Major League Baseball Advanced Media (MLBAM) uses the GameDay application to distribute pitch information in real-time and also provides a classification label such as "four-seam fastball" or "slider" for each pitch. Brooks Baseball makes small adjustments to the calculations and uses manually-reviewed pitch classification results provided by Pitch Info (www.pitchinfo.com) to improve on the accuracy of the MLBAM reported data.

2.3 Signature Model

For the purpose of comparing players, the EMD has the advantage of allowing the comparison of all pitches thrown by a pair of pitchers regardless of pitch type, or all batted balls hit by a pair of batters regardless of the observed outcome. The EMD is also not sensitive to the vagaries of classification algorithms since clusters with similar properties will be seen as similar even if they have been assigned different labels.

Pitchers tend to throw a small number of distinct pitch types which allows the pitch distribution for a pitcher for a given year and one of the four possible platoon configurations (RHP vs. RHB, RHP vs. LHB, LHP vs. RHB, LHP vs. LHB) to be accurately modeled using the signature representation of Equation (2.1) where each pitch type corresponds to a cluster. The number of clusters m corresponds to the number of distinct pitch types as identified by the Pitch Info classifier where m can depend on both the specific pitcher and the platoon configuration. For each pitch type i, μ_i is the pitch parameter mean vector $(\overline{s}_i, \overline{x}_i, \overline{z}_i)$ and w_i is the fraction of pitches of that type for the pitcher and platoon configuration.

Batters are represented by distributions in the batted-ball parameter space. Separate distributions are used to capture information about batted ball initial speed s_l and vertical launch angle v against left-handed and right-handed pitchers.

2.4 Ground Distance

The computation of the EMD requires the specification of a ground distance between the μ_i mean vectors that define the point masses for each distribution. The use of a Euclidean distance between mean vectors is problematic because the component variables in the vectors can have different variances and these variables may also have significant correlations. We will illustrate the problems in Section 2.4.1 and Section 2.4.2.

We define the ground distance $G(i, j)$ between μ_i and μ_j as the Mahalanobis distance [11]

$$G(i, j) = \left[(\mu_i - \mu_j)\Sigma^{-1}(\mu_i - \mu_j)^T\right]^{\frac{1}{2}} \tag{2.2}$$

where the covariance matrix Σ for the population of mean vectors μ_i serves to correct for differences in the variances of the vector components and also for their correlation structure. This distance is equivalent to a Euclidean distance after a whitening transform [11] has been applied to transform the original variables to a new set of variables which are uncorrelated and have unit variance.

2.4.1 Ground Distance for Pitch Distributions

The signatures are used to compute the distance between distributions using the EMD as described in Section 2.1 with the whitened ground distance defined by Equation (2.2). As a two-dimensional example of this process, Figure 2.1 is a scatterplot of the mean $(\overline{s}_i, \overline{z}_i)$ values for each pitch cluster in a signature for the right-handed pitcher versus right-handed batter platoon configuration in 2016. We see that $\overline{s}_i$ and $\overline{z}_i$ have a large positive correlation so that a pitch thrown with a higher speed will tend to have a larger vertical movement. The variance of the $\overline{s}_i$ values is also larger than the variance of the $\overline{z}_i$ values. These effects are addressed by using the Mahalanobis ground distance defined by Equation (2.2).

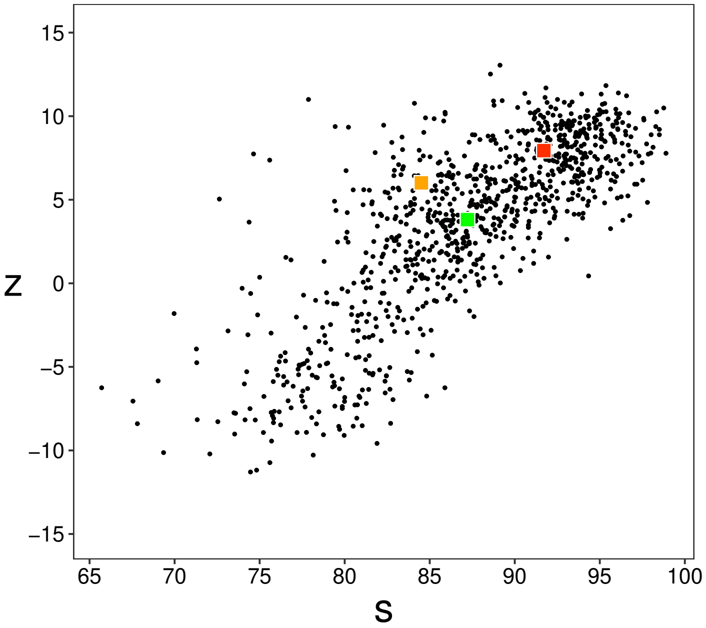

Figure 2.1: Cluster means $(\overline{s}_i, \overline{z}_i)$ for RHP versus RHB configuration, 2016

The impact of the correlation between the two variables can be seen by considering the orange, green, and red points in Figure 2.1 which correspond to the $(\overline{s}_i, \overline{z}_i)$ values for three specific pitch clusters in the figure as detailed in Table 2.1. The Euclidean distance of 6.10 between the green point (Latos cutter) and the red point (Chacin four-seam) is significantly larger than the Euclidean distance of 3.49 between the green point (Latos cutter) and the orange point (Kennedy changeup). Since the vector difference between the green point and the red point is aligned with the direction of correlation of the variables, however, a significant portion of the separation between these points is due to the correlation between s and z. On the other hand, the vector difference between the green point and the orange point is approximately orthogonal to the direction of correlation. If we compute the Mahalanobis distance using the s and z variables shown in Table 2.1, the distance of 0.81 between the green point and the red point is now significantly less than the distance of 1.32 between the green point and the orange point.

Table 2.1: Three $(\overline{s}_i, \overline{z}_i)$ pitch cluster means in Figure 2.1

Point color	Pitcher	Pitch type	$(\overline{s}_i, \overline{z}_i)$
Orange	Ian Kennedy	Changeup	(84.51, 6.01)
Green	Mat Latos	Cutter	(87.22, 3.81)
Red	Jhoulys Chacin	Four-seam	(91.71, 7.94)

2.4.2 Ground Distance for Batted Ball Distributions

As we mentioned in Section 2.4, the use of a standard Euclidean distance between vectors is problematic because the speed variable s_l corresponds to a different physical quantity that has a smaller variance than the launch angle variable v_l and also because the variables are correlated. Figure 2.2, for example, is a scatterplot of the (s_l, v) values for for the right-handed batter versus right-handed pitcher configuration in 2017. We see from the figure that s_l and v have different variances and a positive correlation. The covariance matrix Σ for the population of (s_l, v) vectors for a platoon configuration captures information about both the variance of the variables and their correlation. Using this information, the ground distance defined by Equation (2.2) can compensate for the correlation structure and the variance of the s_l and v variables.

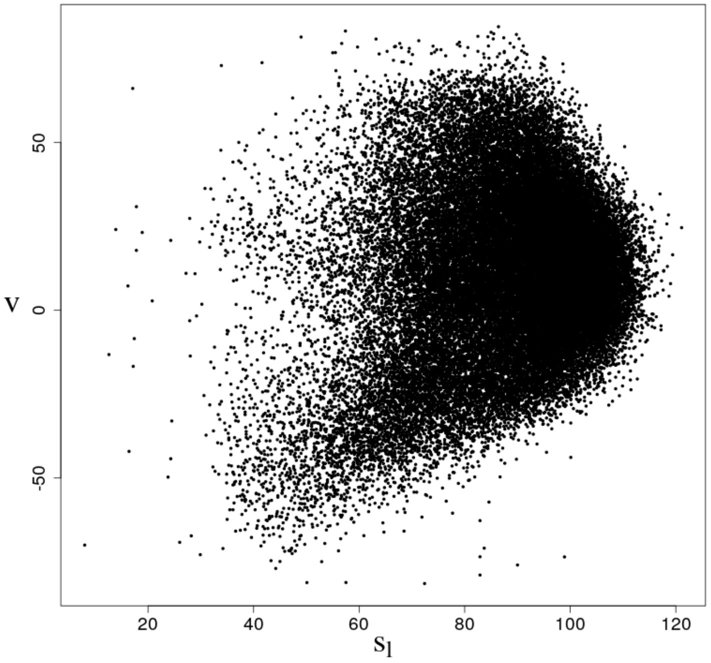

Figure 2.2: (s_l, v) for RHP versus RHB configuration, 2017

Chapter 3

Measuring Player Similarity

3.1 Measuring Pitcher Similarity

3.1.1 Overview

Sensors have recorded information about the speed and movement of pitches thrown in major league ballparks since 2006. This data can be used to develop pitcher similarity measures that are based on pitch physical properties. These measures are valuable not only for comparing major league pitchers but also for allowing the direct comparison of pitchers in other leagues (minor, amateur, and foreign) that deploy these sensors to their major league counterparts. The identification of groups of similar pitchers can be used by analysts to generate optimized projection models [55] or to generate larger samples for predicting the outcome of batter/pitcher matchups [18] [60]. A similarity measure can also be used to help quantify the relationship between pitch distributions and pitcher performance. In addition, such a measure allows individual pitchers to be monitored over time to detect possible changes in pitch characteristics, health, or throwing mechanics.

Previous methods for quantifying pitcher similarity have been limited to the comparison of pitches of the same type which makes these methods highly dependent on the outcome of pitch classification algorithms. Kalk [30] [31] developed a similarity measure that compared pitches of the same type using variables that included pitch frequency, speed, and movement. Loftus [39] [40] [41] improved on Kalk's approach by separating pitchers by handedness while using the Kolmogorov-Smirnov distance to compare distributions. Like Kalk's method, however, this approach only considers comparisons between pitches of the same type. A difficulty for these methods is that different pitch types for a single pitcher or across multiple pitchers can have similar properties. This causes the pitch frequency statistics used by similarity algorithms to depend heavily on the classification process and also prevents the comparison of similar pitches that are assigned different labels. Due to these complications, Loftus conceded [41] that his method is best suited for comparing individual pitches as opposed to comparing pitchers based on their entire arsenal. Gennaro [18] has proposed a more qualitative approach to measuring pitcher similarity by applying a cosine measure to a hand-selected vector of features and weightings. The features used by this method include a pitcher's two most common pitch types and his most common two-pitch sequence.

In this section we develop a pitcher similarity measure that is based on the comparison of multidimensional distributions that represent the collection of pitches thrown by a pitcher. The similarity measure separately considers the full pitch distributions used against right-handed and left-handed batters where each distribution captures information about pitch speed and movement. The distributions are modeled using signatures which enables the EMD [50] to efficiently measure the amount of work that is required to transform one distribution into another. A whitening transform [11] is used by the EMD to account for the variances and correlation structure of the pitch speed and movement variables when comparing individual elements of the distributions. The algorithm is structured to allow the incorporation of additional pitch descriptors as they become available. Since this approach compares full distributions instead of individual pitch types, the resulting similarity measure is relatively insensitive to the results of pitch classification. We demonstrate the similarity measure for several applications including the identification of similar and dissimilar pitchers, the identification of unique pitchers, the quantification of year-to-year pitcher stability, and the quantification of pitcher variation with batter handedness and the count. We also use non-metric multidimensional scaling (NMDS) [36] to visualize properties of the new measure.

3.1.2 Representing Pitch Distributions

Our pitcher similarity measure will consider the estimated s, x, and z parameters for each pitch, and we will represent pitchers by their distribution of (s, x, z) pitch vectors. We note that other factors such as pitch location [20], sequencing [25], and deception [42] also play a role in determining pitcher performance. Figure 3.1, for example, plots the distribution of pitches thrown by left-hander Jon Lester in 2016 and Figure 3.2 plots the distribution for left-hander Chris Sale. In each figure, different pitch types are labeled with different colors.

Given that pitchers typically throw a small number of different pitches as we mentioned in Section 2.1, their pitch distributions can be efficiently encoded as signatures defined by clusters of different pitch types. Suppose, for example, that a pitcher threw m different pitch types against RHB during 2016. Then his signature against RHB is given by the set of m clusters

$$P_{_R} = \{(\mu_1, w_1), \ldots, (\mu_m, w_m)\} \tag{3.1}$$

where μ_i is his mean vector $(\overline{s}_i, \overline{x}_i, \overline{z}_i)$ for (s, x, z) for pitch type i against RHB and w_i is his fraction of pitches of type i against RHB. Thus, the signature $P_{_R}$ approximates a pitcher's distribution of pitches against right-handed batters by a distribution defined by a set of point

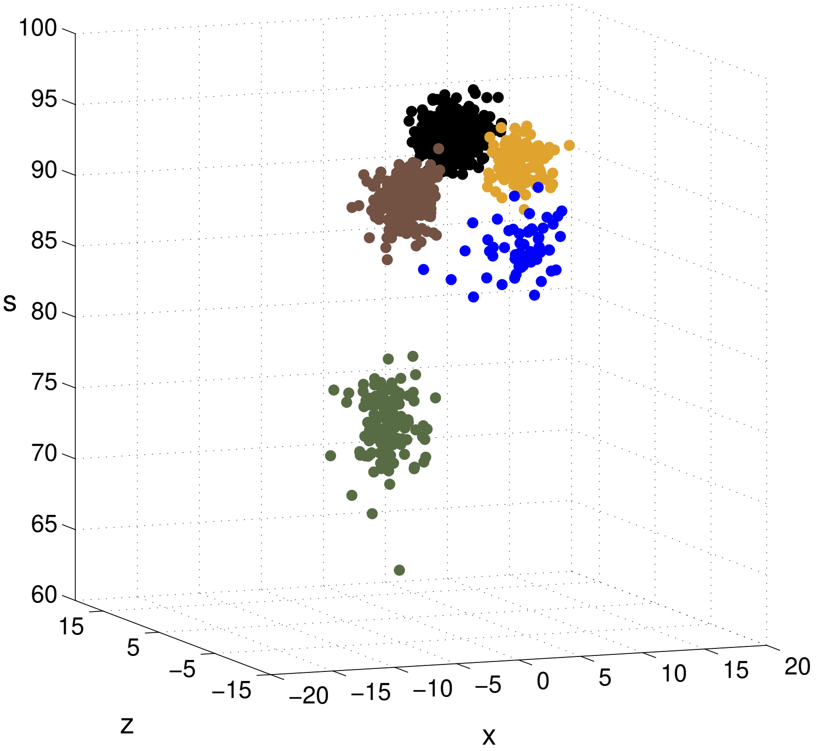

Figure 3.1: Jon Lester pitches in 2016

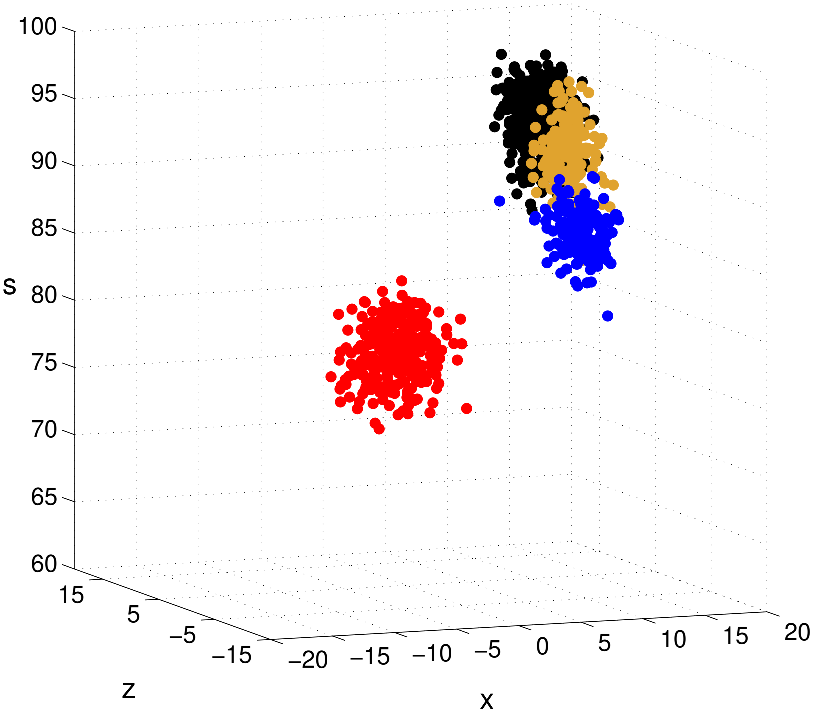

Figure 3.2: Chris Sale pitches in 2016

masses at the locations μ_i with the weights w_i. In a similar way, we can define his signature $P_{_L}$ against LHB. Note that the number of clusters m depends on both the specific pitcher and the batter handedness.

3.1.3 Pitcher Similarity Measure

We can assess the similarity of distributions that are represented by signatures using the Earth Mover's Distance [50]. Given the ground distance defined by Equation (2.2) and the signatures $P_{_R}$ for two right-handed pitchers A and B, we can compute the EMD $D_{_R}(A, B)$ to measure the similarity of the pitchers against right-handed batters. We can also use the $P_{_L}$ signatures to compute the EMD $D_{_L}(A, B)$ to measure their similarity against left-handed batters. The distances $D_{_R}(A, B)$ and $D_{_L}(A, B)$ can be combined into an overall measure of similarity using

$$D(A, B) = f_{_{RR}} D_{_R}(A, B) + f_{_{RL}} D_{_L}(A, B) \tag{3.2}$$

where $f_{_{RR}}$ and $f_{_{RL}}$ represent the league average fraction of pitches that right-handed pitchers throw to right-handed and left-handed batters respectively. We use the league average fractions so that $D(A, B)$ does not depend on the actual fraction of pitches that a particular pitcher threw to a given handedness of batter. In the same way, we can define the overall similarity score for a pair of left-handed pitchers Y and Z by

$$D(Y, Z) = f_{_{LR}} D_R(Y, Z) + f_{_{LL}} D_L(Y, Z) \tag{3.3}$$

where f_{LR} and f_{LL} are the league average fractions of pitches thrown by left-handed pitchers to right-handed and left-handed batters respectively. A small distance D between a pair of pitchers indicates a high degree of similarity while larger distances indicate that a pair of pitchers is less similar.

3.1.4 Data Analysis

All data analysis in Section 3.1 uses the Brooks Baseball adjustments to the PITCHf/x measurements and the Pitch Info classifications. Our 2016 data analysis considers the 196 right-handed pitchers and the 63 left-handed pitchers who threw at least 1000 pitches during the regular season.

Similar Pitchers

For each of these pitchers, Tables A.1 and A.2 in Appendix A present the most similar pitcher and the corresponding distance using the metric defined in Section 3.1.3. The most similar pair of right-handed pitchers in 2016 was Matt Harvey and Shelby Miller. Harvey and Miller each threw four-seam fastballs with similar (s, x, z) parameters at similar frequencies. In particular, each pitcher threw between 59 and 60 percent four-seamers to right-handed batters and between 56 and 57 percent four-seamers to left-handed batters with Harvey averaging 95.39 mph and Miller averaging 94.15 mph on these pitches. We also note that Harvey's slider $(\overline{s}, \overline{x}, \overline{z}) = (89.51, 0.90, 4.28)$ was similar to Miller's cutter $(\overline{s}, \overline{x}, \overline{z}) = (89.41, 1.17, 3.89)$ and each pitcher used this respective pitch between 25 and 26 percent of the time against right-handed batters. Similarity metrics that do not compare pitches of different type would be unaware of the similarity of these pitches.

The most similar pair of left-handed pitchers in 2016 was Jon Niese and Chris Rusin. The most frequent pitches for each left-hander against right-handed batters were their sinker and cutter which they threw at similar frequencies and with similar properties. For their sinkers against RHB, we had $(\overline{s}, \overline{x}, \overline{z}, w)$ vectors of (89.52, 9.63, 4.30, .272) for Niese and of (90.32, 9.74, 4.88, .244) for Rusin. For their cutters against RHB, we had vectors of (86.74, -0.30, 3.86, .272) for Niese and of (87.49, 1.62, 3.78, .299) for Rusin. Each pitcher's most frequent pitch to left-handed batters was their sinker which Niese threw 40.7 percent of the time and Rusin threw 38.8 percent of the time.

Dissimilar Pitchers

The most dissimilar pair of right-handed pitchers was Brad Ziegler and Marco Estrada with a distance of 5.688. The difference was largely due to an extreme difference in the vertical movement on their pitches. Ziegler threw 57.7 percent sinkers in 2016 with an average z of -6.72 while Estrada threw 50.1 percent 4-seam fastballs with an average z of 13.01. In 2016, Ziegler had the smallest average vertical movement $z = -5.33$ over all of his pitches while Estrada had the highest $z = 9.64$.

The most dissimilar pair of left-handed pitchers was Zach Britton and Tommy Milone with a distance of 4.238. Britton threw more than 90 percent sinkers in 2016 with an average s of over 97 mph and an average z of 3.70. Milone averaged only 88.19 mph on his hardest and most frequent pitch, a four-seam fastball, which he threw 45.5 percent of the time with an average vertical movement of 11.45.

Unique Pitchers

The similarity measure can also be used to find the most unique major league pitchers. Table 3.1 lists the right-handed pitchers with the greatest distance to their most similar match in 2016 and Table 3.2 lists the left-handed pitchers with the greatest distance to their most similar match in 2016. Hard-throwing left-hander Aroldis Chapman fell short of the 1000 pitch threshold in 2016, but would rank as the second most unique pitcher behind Britton in Table 3.2 with a distance of 1.5495 to the nearest left-handed pitcher Tony Cingrani.

Table 3.1: Most unique right-handed pitchers, 2016

Pitcher	Distance to nearest pitcher
Brad Ziegler	2.8651
Jered Weaver	1.7653
Chris Young	1.4429
Steve Cishek	1.3934
Marco Estrada	1.3896
Lance McCullers	1.3648
Fernando Rodney	1.2610
Tyler Clippard	1.2232
Aaron Nola	1.1660
Bryan Shaw	1.1258

Table 3.2: Most unique left-handed pitchers, 2016

Pitcher	Distance to nearest pitcher
Zach Britton	1.7251
Rich Hill	1.4946
Clayton Kershaw	1.4912
Zach Duke	1.4223
Andrew Miller	1.3264
Drew Pomeranz	1.2464
Tommy Milone	1.1309
Clayton Richard	1.0658
Julio Urias	0.9960
John Lamb	0.9782

Visualizing Similarity

The similarity structure for a group of pitchers can be visualized using non-metric multidimensional scaling (NMDS) [36] [37]. This technique maps a set of objects and the distances

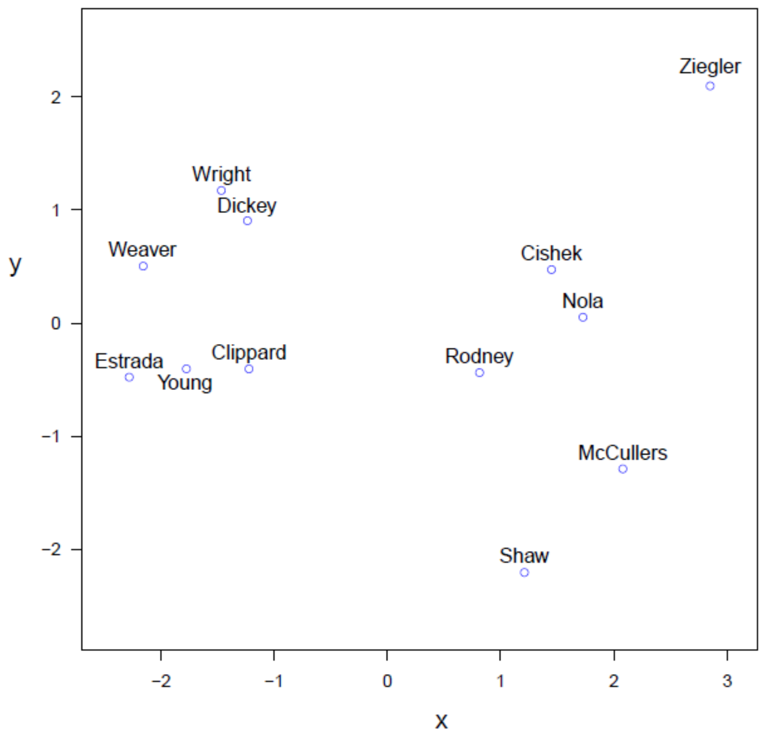

Figure 3.3: NMDS Result for Unique Right-handed Pitchers, 2016

between them to a low-dimensional space for visualization while attempting to preserve the rank ordering of the inter-object distances. We use NMDS to visualize properties of the similarity measure for the most unique right-handed and left-handed pitchers.

Figure 3.3 is the two-dimensional NMDS result for the ten most unique right-handed pitchers in Table 3.1 plus the two most prominent knuckleballers R.A. Dickey and Steven Wright. The most unique right-hander, Brad Ziegler, is located in the far upper right in Figure 3.3. Ziegler's uniqueness is largely due to throwing a large percentage ($w = 57.7\%$) of sinkers in 2016 with a low velocity ($\overline{s} = 84.74$) and heavy sink ($\overline{z} = -7.28$). The closest pitchers to Ziegler in the plot are Steve Cishek and Aaron Nola who each threw between 40 and 44 percent sinkers but at a higher velocity than Ziegler. The pitchers in the plot with the highest average velocity over their pitches (Rodney, McCullers, Shaw) are located in the lower right quadrant. In this group, Rodney appears closest to Cishek and Nola due to also throwing a high percentage of sinkers ($w = 39.1\%$), but the high vertical movement on his

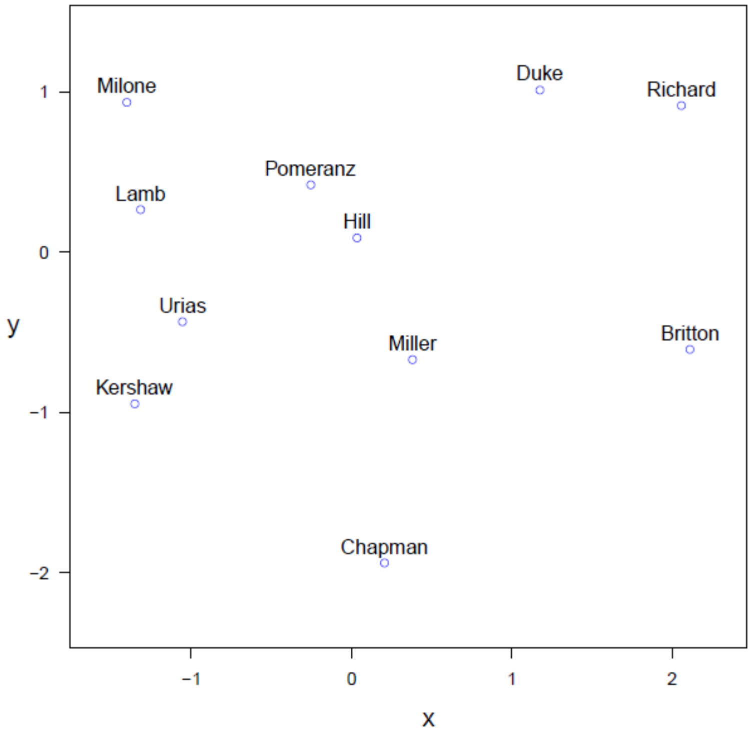

Figure 3.4: NMDS Result for Unique Left-handed Pitchers, 2016

pitches, particularly his four-seam fastball, pulls him to the left of these two. Bryan Shaw has the highest average velocity among pitchers in Figure 3.3 and appears at the lowest point in the plot. To the left of Rodney is a group of three pitchers (Estrada, Young, Clippard) who displayed the highest average vertical movement on their pitches among the pitchers in the figure. This high vertical movement was largely achieved by throwing between 45 and 51 percent four-seam fastballs. Above this group is Jered Weaver who also threw pitches with a high average vertical movement but who had the lowest average pitch velocity in the plot among the non-knuckleballers. Dickey and Wright appear together above Weaver and, as shown in Table A.1, the two knuckleballers are the best match for each other over the 196 right-handed pitchers in the 2016 data set. We see that the most dissimilar right-handed pitchers in the entire data set, Ziegler and Estrada, are also the most separated in Figure 3.3.

Figure 3.4 is the NMDS result for the ten most unique left-handed pitchers in Table 3.2

plus the hard-throwing Aroldis Chapman. The most unique left-hander, Zach Britton, is located on the far right edge of the plot. Britton achieved his uniqueness by throwing a high percentage ($w = 92.0\%$) of very hard ($\overline{s} = 97.44$) sinkers. The closest left-hander to Britton in the figure is Clayton Richard who also threw a high percentage of sinkers ($w = 65.0\%$) but at a lower velocity ($\overline{s} = 91.59$). To the left of Richard and farther removed from Britton is Zach Duke who also threw a large number of sinkers but at an even lower frequency ($w = 50.4\%$) and velocity ($\overline{s} = 90.13$). The second most unique left-hander in the group, Aroldis Chapman, who threw a high percentage ($w = 81.1\%$) of very hard ($\overline{s} = 101.32$) four-seam fastballs appears at the lowest point on the plot. On the left side of the figure are four left-handers (Milone, Lamb, Urias, Kershaw) who all favored the four-seam fastball with frequencies varying between 45.5 percent for Milone and 55.3 percent for Urias. The average four-seam velocity for the pitchers increases from top to bottom with values of 88.19, 90.49, 93.32, and 93.74 for Milone, Lamb, Urias, and Kershaw respectively. To the right of these four pitchers are Drew Pomeranz and Rich Hill who complemented their four-seam fastball with a large percentage of curves with sharp downward movement. Hill is the closest pitcher to Andrew Miller in the plot. Since Miller's four-seam fastball is harder than Hill's and Miller's most frequent off-speed pitch is a slider which is thrown substantially harder than's Hill's curve, Miller appears lower than Hill. We see that the most dissimilar left-handed pitchers in the full data set, Britton and Milone, are also the most separated in Figure 3.4. In Appendix B we use the pitcher similarity measure to examine several pitcher characteristics.

3.1.5 Summary

We have developed a new tool that analysts can exploit to study a range of application areas. The similarity measure allows the direct comparison of pitchers across various contexts including MLB, MiLB, amateur, and foreign leagues which can improve predictions for how a pitcher will perform in a new environment. The identification of similar pitchers increases the sample sizes that can be used to forecast the outcome of batter/pitcher matchups and supports regression to more appropriate population means by projection models. The measure can also be used to monitor pitchers over time and to develop improved models for the health risk and aging characteristics associated with different pitcher classes. For fans the new tool reveals similarities that we didn't know existed and shows us, once again, that there's more than one way to find success as a major league pitcher.

3.2 Measuring Batter Similarity

Traditional methods that are used to compare and forecast the performance of batters are based on observed outcomes. These measures are influenced by a number of variables that are beyond the control of the batter such as the defense, the ballpark, and the weather. Radar and optical sensors [22] have been installed in MLB stadiums that measure parameters of batted balls including their initial speed and direction. Previous work [21] [49] has used these measurements to build models for batted ball distributions and to derive mappings from batted ball parameters to intrinsic values which leads to more reliable batted-ball statistics for batters and pitchers. In this work, we use distributions defined over batted-ball parameters to develop a new approach for characterizing hitters using a similarity measure based on physical measurements. Similarity measures have been developed for pitchers using pitch trajectory parameters [18, 26, 30, 31, 39, 40, 41] but previous work has not considered the use of ball-tracking data to develop such metrics for batters. The new approach enables the generation of sets of similar and dissimilar batters as well as the identification of unique batters. The metric also allows the rendering of visualizations which illustrate batter characteristics.

The batter similarity measure supports several important applications. The measure can be used not only for comparing major league batters but also for comparing batters in other leagues (minor, amateur, and foreign) to their major league counterparts. The identification of groups of similar batters can be used to define optimized population models [55] for forecasting or to generate larger samples for predicting the outcome of batter/pitcher matchups [18] [60]. The metric can also be used to increase sample sizes for learning the strengths and weaknesses of sets of batters with similar swing characteristics as a function of pitch speed, location, and movement. In addition, a similarity measure allows individual batters to be monitored over time to detect possible changes in swing mechanics or health. As additional hit-tracking data becomes available, the measure can also be used to model the aging characteristics associated with different batter classes.

3.2.1 Representing Batted Ball Distributions

Batters are represented by distributions in the batted-ball parameter space. Separate distributions are used to capture information about batted ball initial speed s_l and vertical launch angle v against left-handed and right-handed pitchers. Figure 3.5, for example, plots the (s_l, v) distribution of batted balls for right-handed batters Aaron Judge and Ronald Torreyes against right-handed pitchers in 2017. We see that Judge has a higher average exit speed and vertical launch angle than Torreyes. For a pair of batters, the similarity of (s_l, v) distributions is evaluated against each pitcher handedness and the two values are combined into a single measure of similarity. The result is a metric that can be used to compare and group batters based on batted ball characteristics.

3.2.2 Batter Similarity Measure

Given the ground distance defined by Equation (2.2) and the (s_l, v) distributions for two right-handed batters A and B against right-handed pitchers, we can compute the EMD

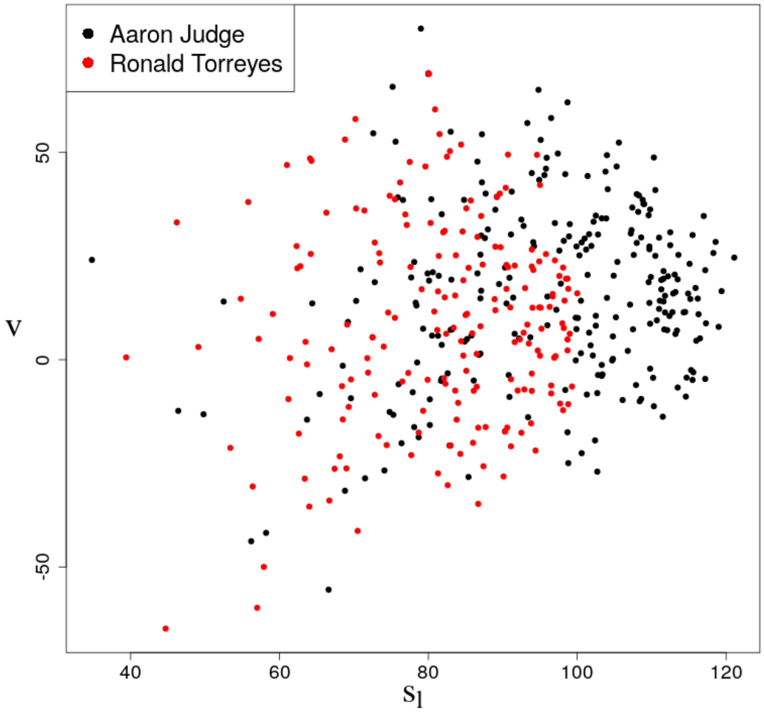

Figure 3.5: (s_l, v) for Aaron Judge and Ronald Torreyes vs. RHP, 2017

$D_R(A, B)$ to measure the similarity of the batters against right-handed pitchers. We can also use the (s_l, v) distributions against left-handed pitchers to compute the EMD $D_L(A, B)$ to measure their similarity against left-handed pitchers. The distances $D_R(A, B)$ and $D_L(A, B)$ can be combined into an overall measure of similarity using

$$D(A, B) = f_{RR} D_R(A, B) + f_{RL} D_L(A, B) \tag{3.4}$$

where f_{RR} and f_{RL} represent the league average fraction of batted balls from right-handed batters that occur against right-handed and left-handed pitchers respectively. We use the league average fractions so that $D(A, B)$ does not depend on the number of opportunities that a particular batter had against a given handedness of pitcher. Using the same approach, we can define an overall measure of similarity between pairs of left-handed batters and between pairs of switch-hitters. A small distance $D(A, B)$ between a pair of batters indicates a high degree of similarity while larger distances indicate that a pair of batters is less similar.

3.2.3 Data Analysis

We will demonstrate the similarity measure for several applications including the identification of similar and dissimilar batters, the identification of unique batters, and the quantification of year-to-year batter stability. Leaderboards and visualizations generated using non-metric multidimensional scaling are presented to illustrate the new approach. We will also use similarity groups derived from the metric to provide population models for batter classes that can be used for forecasting. All analysis uses Statcast batted ball data from Baseball Savant with bunts and foul balls removed. For 2017 we consider the 112 right-handed batters, the 71 left-handed batters, and the 29 switch-hitters who hit at least 250 batted balls during the regular season.

Similar Batters

For each of these batters, we used the method described in sections 3.2.1 and 3.2.2 to find the most similar batter and the corresponding distance. Smaller values of the distance correspond to more similar batters. The most similar pair of right-handed batters in 2017 was Andrew McCutchen and Nolan Arenado with a distance of 0.2099. The players had a similar average exit speed and launch angle with $(\overline{s_l}, \overline{v}) = (89.2, 14.0)$ for McCutchen and $(\overline{s_l}, \overline{v}) = (89.5, 15.0)$ for Arenado. The most similar pair of left-handed batters was Mitch Moreland and Shin-Soo Choo with a larger distance of 0.2359. The average exit speed and launch angle for these players was $(\overline{s_l}, \overline{v}) = (89.6, 10.3)$ for Moreland and $(\overline{s_l}, \overline{v}) = (88.8, 7.8)$ for Choo. The most similar pair of switch-hitters was Jose Ramirez and Francisco Lindor with a distance of 0.2166. The average exit speed and launch angle for these players was $(\overline{s_l}, \overline{v}) = (88.4, 13.6)$ for Ramirez and $(\overline{s_l}, \overline{v}) = (89.0, 13.4)$ for Lindor.

Dissimilar Batters

The most dissimilar pair of right-handed batters in 2017 was Aaron Judge and Ronald Torreyes with a distance of 1.0053. These players had large differences in average exit velocity and launch angle with $(\overline{s_l}, \overline{v}) = (95.6, 15.4)$ for Judge and $(\overline{s_l}, \overline{v}) = (81.9, 7.1)$ for Torreyes. The most dissimilar pair of left-handed batters was Dee Gordon and Matt Carpenter with a distance of 0.9581. These players also had large differences in average exit velocity and launch angle with $(\overline{s_l}, \overline{v}) = (81.8, 2.0)$ for Gordon and $(\overline{s_l}, \overline{v}) = (90.0, 22.6)$ for Carpenter. In particular, Gordon had the highest ground ball rate in MLB in 2017 at 57.6% while Carpenter had the lowest at 26.9%. The most dissimilar pair of switch-hitters was Billy Hamilton and Kendrys Morales with a distance of 0.8061. These players had a large difference in average exit velocity with Hamilton at 80.8 mph and Morales at 91.1 mph.

Unique Batters

The similarity measure can also be used to find the most unique major league batters. The right-handed batters with the greatest distance to their most similar match in 2017 are shown in Table 3.3. The most unique left-handed batters and switch-hitters are shown in Tables 3.4 and 3.5 respectively.

Visualizing Similarity

The similarity structure for a group of batters can be visualized using non-metric multidimensional scaling (NMDS) [36]. We use NMDS to visualize properties of the similarity

Unique RHB	Distance to nearest RHB
Aaron Judge	0.3286
Giancarlo Stanton	0.3282
Jose Iglesias	0.3038
Willson Contreras	0.3013
Hunter Renfroe	0.2987
Ronald Torreyes	0.2964
Paul DeJong	0.2915
Guillermo Heredia	0.2904
Jose Peraza	0.2904
Jorge Bonifacio	0.2871

Table 3.3: Most unique right-handed batters, 2017

Unique LHB	Distance to nearest LHB
Jarrod Dyson	0.3352
Didi Gregorius	0.3207
Norichika Aoki	0.3134
Lucas Duda	0.3057
Kyle Schwarber	0.3035
Ben Revere	0.3034
Dee Gordon	0.3034
Bryce Harper	0.3022
Ender Inciarte	0.2993
Carlos Gonzalez	0.2956

Table 3.4: Most unique left-handed batters, 2017

Unique Switch-Hitter	Distance to nearest Switch-Hitter
Kendrys Morales	0.3271
Billy Hamilton	0.3186
Erick Aybar	0.3186
Carlos Santana	0.3184
Cesar Hernandez	0.3055
Jose Reyes	0.2993
Yangervis Solarte	0.2964
Tucker Barnhart	0.2931
Melky Cabrera	0.2904
Matt Wieters	0.2835

Table 3.5: Most unique switch-hitters, 2017

measure for the unique batters described in section 3.2.3. In each plot, average exit speed s_l tends to increase as we move to the right while average launch angle v tends to increase as we move up.

The NMDS result for the ten most unique right-handed batters is shown in Figure 3.6. The most unique right-handed batter, Aaron Judge, appears on the far right side of the plot due to his high average exit velocity with Giancarlo Stanton located below and to the left due to a smaller average exit velocity and launch angle. On the far left are four players (Torreyes, Peraza, Iglesias, Heredia) with the smallest average exit velocity on the plot. Near the middle of Figure 3.6 are four players (DeJong, Bonifacio, Renfroe, Contreras) with intermediate average exit velocities and average launch angles that range from 7.0 degrees for Contreras near the bottom of the plot to 17.5 degrees for DeJong near the top of the plot. We see that the most dissimilar right-handed batters in the entire data set, Judge and Torreyes, are also the most separated in the plot.

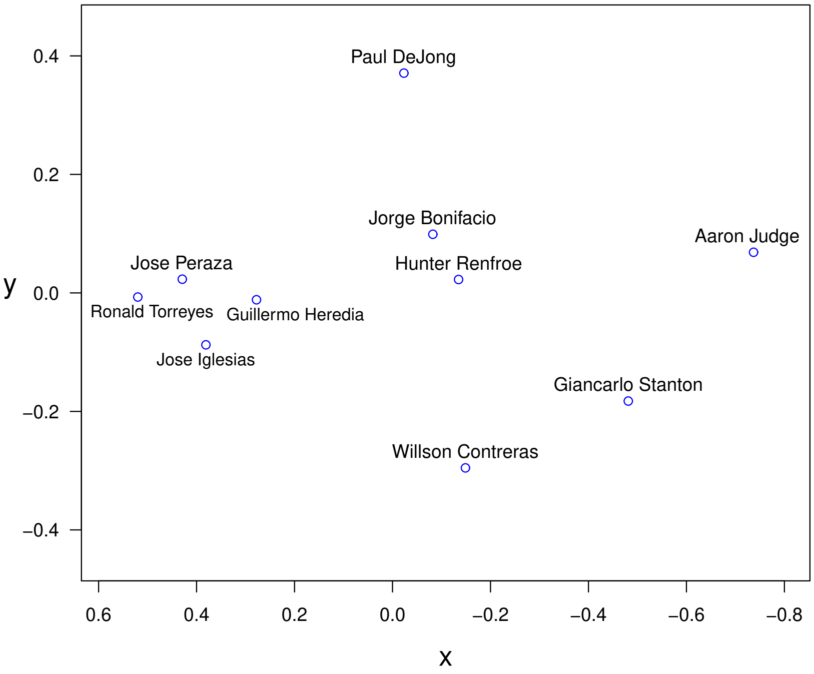

Figure 3.6: NMDS Result for Unique Right-handed Batters, 2017

The NMDS result for the ten most unique left-handed batters is shown in Figure 3.7. The three batters on the right side of the Figure (Schwarber, Duda, Harper) have the highest average exit velocity on the plot. The three players on the left side of the figure (Revere, Dyson, Gordon) have a low average exit velocity and a small average launch angle. To the right and below this group are Norichika Aoki and Carlos Gonzalez with a higher average exit velocity and small average launch angles while to the right and above this group are Ender Inciarte and Didi Gregorius with a higher average exit velocity and larger average launch angles.

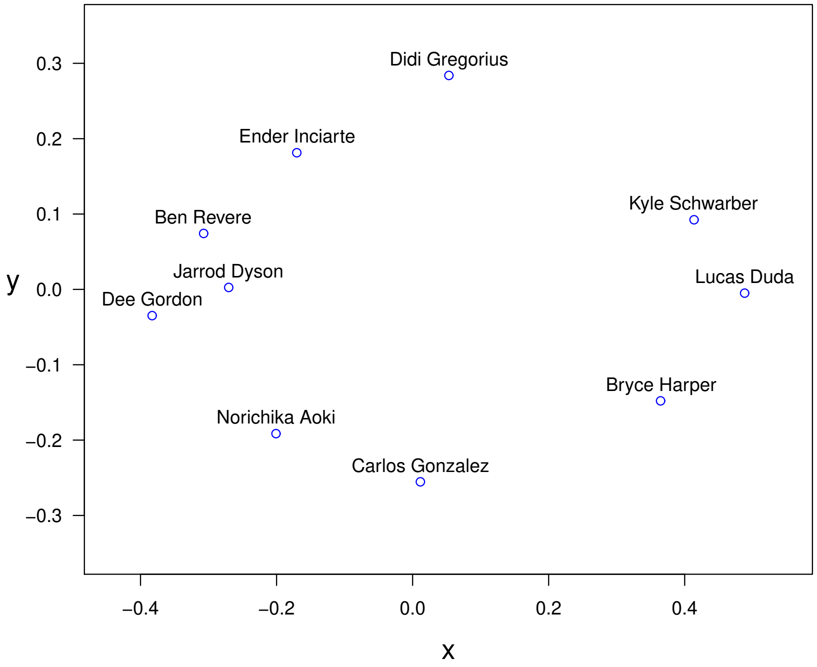

Figure 3.7: NMDS Result for Unique Left-handed Batters, 2017

The NMDS result for the ten most unique switch-hitters is shown in Figure 3.8. The most unique switch-hitter, Kendrys Morales, has the highest average exit velocity in the plot and appears to the far right. Carlos Santana appears above and to the left due to his lower average exit velocity and a higher average launch angle while Melky Cabrera appears below and to the left due to his lower average exit velocity and smaller average launch angle. Billy Hamilton appears to the far left due to the lowest average exit velocity in the plot. Below

Hamilton and to the right is Erick Aybar with the next lowest average exit velocity and a smaller average launch angle. Five players (Solarte, Reyes, Wieters, Barnhart, Hernandez) appear near the middle of the plot with intermediate average exit velocities and an average launch angle that is smallest for Hernandez at the bottom of the plot and largest for Reyes and Solarte near the top of the plot. We see that the most dissimilar switch hitters in the entire data set, Morales and Hamilton, are also the most separated in the plot. In Appendix C we use the batter similarity measure to examine several batter characteristics.

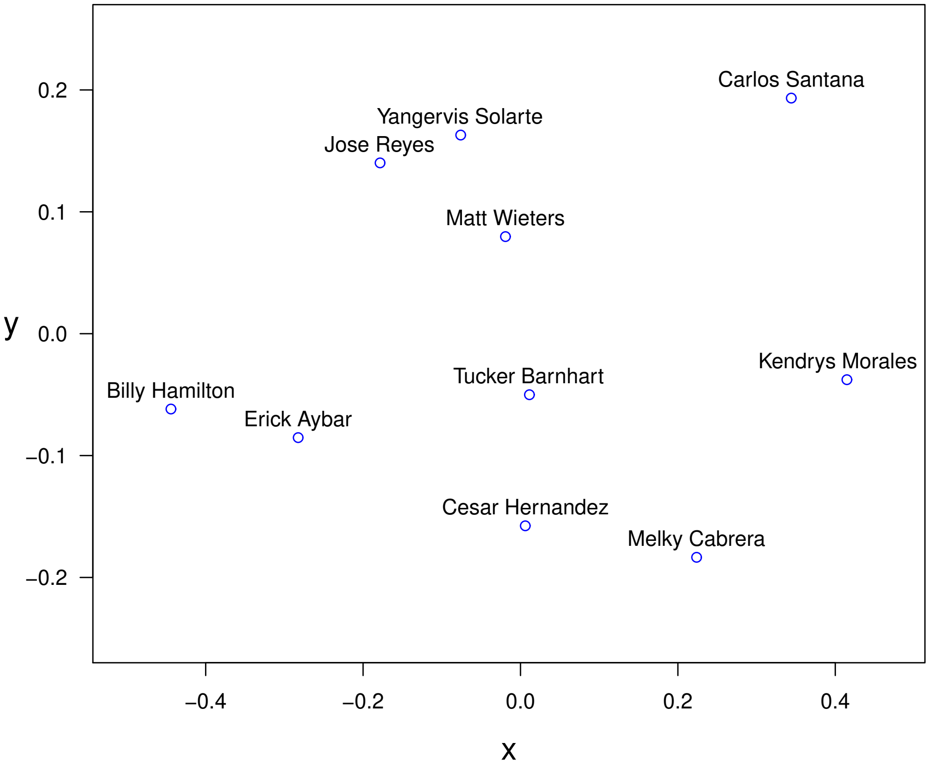

Figure 3.8: NMDS Result for Unique Switch-Hitters, 2017

3.2.4 Summary

We have developed a metric for comparing batters using hit-tracking data. The new metric can be exploited to advance a range of application areas. The measure allows the direct comparison of batter swing characteristics across various contexts including MLB, MiLB, amateur, and foreign leagues. The identification of similar batters increases the sample sizes that can be used to forecast the outcome of batter/pitcher matchups and supports regression to more appropriate population means by projection models. The measure can also be used to monitor batters over time and to develop improved models for the aging characteristics associated with different swing types. The new approach also allows teams to optimize pitch selection strategy according to batter strengths and weaknesses recovered by applying machine learning techniques [20] to groups of similar batters.

Chapter 4

Learning a Function over Distributions

4.1 Overview

Nonparametric methods [3] are a powerful tool for model recovery and continue to support a variety of applications [33] [63]. Nonparametric kernel regression can be used to estimate a function of unknown form and has been applied in a wide range of settings [34]. Generalizing this approach to learn a function over distributions requires a suitable metric for distribution space for which we use the Wasserstein metric or Earth Mover's Distance (EMD). The EMD uses a cost function called the ground distance to determine the minimum amount of work that is needed to transform one distribution into the other. The computational cost of finding the EMD can be expensive which leads to the use of signatures to approximate the distributions thereby enabling the use of efficient linear programming methods [50].

This methodology is used to learn a function over distributions to address the problem of optimizing pitch distributions in baseball. A nonparametric learning method is appropriate for this application because the effectiveness of a pitch distribution has a complicated dependence on the quality, frequency, and interaction of a pitcher's set of pitches. We represent a collection of pitches using a multidimensional distribution that is derived from sensor measurements that capture the physical properties of each pitch. These properties have been shown to have a strong effect on pitch value [23]. Pitchers typically use a small number of different pitch types which allows these distributions to be accurately encoded using signatures. A whitening transform [11] is used by the EMD ground distance to account for the variances and correlation structure of the component variables that define the distributions. A method that is similar to leave-one-out cross validation [54] is used to optimize the kernel smoothing parameter. After recovering the function over pitch distributions, an efficient low-dimensional search can be used to find the optimal frequencies for a pitcher's various pitch types. We show that the new model accurately predicts the dependence of pitcher performance on changes in pitch distribution and significantly outperforms an alternative approach based on game theory.

4.2 Method

We develop a method for learning a function over distributions when the underlying structure of the function is unknown. The method is based on generalizing nonparametric kernel regression using a whitened Earth Mover's Distance as the metric for distribution space. We will illustrate properties of the algorithm with a set of experiments in Section 4.3.

4.2.1 Nonparametric Kernel Regression

Let (x_i, y_i) for $i = 1, 2, \ldots, n$ be a set of observations where x is the explanatory variable and y is the response variable. The data can be modeled by

$$y = f(x) + \epsilon \tag{4.1}$$

where ϵ is an error term. Kernel regression [45] [64] is a nonparametric method that constructs an estimate for $f(x)$ using the weighted average

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} k(d_i) y_i}{\sum_{i=1}^{n} k(d_i)} \tag{4.2}$$

where $d_i = x - x_i$ and $k(\cdot)$ is a kernel probability density function that is typically maximum at zero and decreases with $|d_i|$ so that the largest weights $k(d_i)$ are given to the y_i associated with the x_i that are closest to x. A popular kernel function is the zero-mean Gaussian

$$k(d_i) = g(d_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{1}{2}(d_i/\sigma)^2} \tag{4.3}$$

which depends on the smoothing parameter σ.

Given a set of observations (X_i, y_i) where each X_i is a multidimensional distribution, we can generalize Equations (4.2) and (4.3) to approximate a function over distributions by replacing d_i with a distance D_i between the distributions X and X_i

$$\widehat{f}(X, \sigma) = \frac{\sum_{i=1}^{n} g(D_i, \sigma) y_i}{\sum_{i=1}^{n} g(D_i, \sigma)} \ . \tag{4.4}$$

4.2.2 Finding the Smoothing Parameter Using Cross Validation

The accuracy of kernel regression has a strong dependence on the smoothing parameter σ [11]. Let (X_i, y_i) for $i = 1, 2, \ldots, n$ be a set of observations that associate distributions X_i with responses y_i. For the distribution X_j we can use Equation (4.4) to compute

$$\widehat{f}(X = X_j, \sigma) = \frac{\displaystyle\sum_{\substack{1 \le i \le n \\ i \ne j}} g(D_{ij}, \sigma) y_i}{\displaystyle\sum_{\substack{1 \le i \le n \\ i \ne j}} g(D_{ij}, \sigma)} \tag{4.5}$$

where D_{ij} is the whitened EMD between X_i and X_j as described in Chapter 2 and the (X_j, y_j) observation is excluded from the sums. The error in the approximation is given by

$$E_j(\sigma) = y_j - \widehat{f}(X_j, \sigma). \tag{4.6}$$

We define the optimal smoothing parameter σ^* as the value of σ that minimizes the total absolute error in the approximation over the observations

$$\sigma^* = \arg\min_{\sigma} \sum_{j=1}^{n} |E_j(\sigma)| \,. \tag{4.7}$$

Note that if we include the (X_j, y_j) observation in the sums in Equation (4.5), then as σ approaches zero the approximation $\widehat{f}(X, \sigma)$ approaches a sum of Dirac delta functions centered at the observation points causing each $E_j(\sigma)$ and the sum in Equation (4.7) to approach zero. This yields a poor approximation to the underlying $f(X)$ function everywhere except at the observation points. The method described in this section for finding σ^* is similar to leave-one-out cross validation methods that are used for density estimation [54].

4.3 Experimental Results

4.3.1 Optimizing the Pitch Distribution

A pitcher's success is highly dependent on the characteristics of his pitch distribution. A larger speed s for an individual pitch reduces the batter's available reaction time while greater movement (x, z) makes it more difficult for the batter to determine the optimal contact point. In addition, the diversity of a pitcher's distribution of pitches affects the batter's ability to anticipate the speed and movement of the next pitch. A pitcher can benefit from having pitches with large differences in speed [19] or from having pitches with similar speed that move in different directions [42].

The best result of a matchup for a pitcher is a strikeout which means that the batter was unable to hit the ball successfully given multiple opportunities. A pitcher's strikeout rate is the fraction of his matchups that result in a strikeout. This rate is a repeatable pitcher skill [6] and is a strong determinant of a pitcher's success [15]. We can use the algorithm described in Section 4.2 to learn the dependence of pitcher strikeout rate on the pitch distribution defined over the s, x, and z variables. Since a given pitcher can throw several different pitch types, he can adjust his pitch distribution and expected strikeout rate by changing the frequency of each pitch type. Using the learned relationship between strikeout rate and pitch distribution, we can therefore find the pitch frequencies that optimize a pitcher's strikeout rate. We will evaluate this approach in the following sections.

Previous work on optimizing the pitch distribution has been based on game theory. Using this approach, Paine [47] has suggested that a pitcher's optimal pitch distribution occurs at Nash equilibrium where the pitcher's effectiveness is equal for each of his pitch types. This principle is used to derive the Nash score which is a measure of how close a pitch distribution is to Nash equilibrium. One difficulty with this method is that it requires the use of effectiveness values for each pitch type which are known to have a low reliability [1]. We will evaluate the use of the Nash score for assessing pitch distributions in Section 4.3.5.

4.3.2 Data Processing

We built the strikeout rate model described in Section 4.3.1 using 2016 sensor data for each MLB pitcher who threw at least 1500 pitches during the season. This threshold ensures the use of a reasonably large sample for generating the pitch distributions and strikeout rates and also removes pitchers who were used purely as relievers which often results in a different style of pitching. There were 108 right-handed pitchers and 41 left-handed pitchers who threw at least 1500 pitches in 2016.

The effectiveness of a given pitch depends on the handedness (left or right) of the batter and pitcher. Thus, we separately consider the dependence of strikeout rate on pitch distribution for each of the four possible platoon configurations (RHP vs. RHB, RHP vs. LHB, LHP vs. RHB, LHP vs. LHB). A pitcher's strikeout rate for a platoon configuration and year is defined as the ratio of strikeouts to the number of batters faced after removing all matchups with a pitcher as a batter and also removing all matchups that resulted in a bunt or an intentional walk. Using the 2016 constant of 4.262 batters per inning, the FIP equation [15] predicts that an increase of 0.03 in strikeout rate leads to 0.26 fewer runs allowed per game which is a significant improvement in pitcher performance.

The process of learning and applying a function over distributions can be summarized by the following steps. Training data is first partitioned by platoon configuration and each step is carried out separately for each configuration. The training data provides a set of pitch distributions specified by signatures S_i as defined in Section 2.1 and associated strikeout rates y_i. The covariance matrix Σ in Equation (2.2) is computed for the population of mean vectors specified by the S_i signatures. The smoothing parameter σ is found using cross validation as described in Section 4.2.2. The learned model can then be applied to a pitch distribution X described by a signature S to compute the expected strikeout rate by using Equation (4.4). This process is summarized by Figure 4.1 where application of the model will be described in more detail in Section 4.3.4.

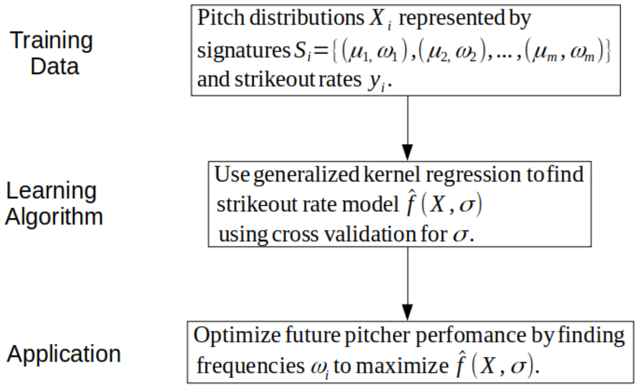
Training Data
Pitch distributions X_i represented by signatures $S_i = \{(\mu_1, \omega_1), (\mu_2, \omega_2), \dots, (\mu_m, \omega_m)\}$ and strikeout rates y_i.
Learning Algorithm
Use generalized kernel regression to find strikeout rate model $\hat{f}(X, \sigma)$ using cross validation for σ.
Application
Optimize future pitcher perfomance by finding frequencies ω_i to maximize $\hat{f}(X, \sigma)$.

Figure 4.1: Process of learning and applying a function over distributions

4.3.3 Cross Validation

The cross validation process described in Section 4.2.2 is used to find optimized values for the smoothing parameter σ for each platoon configuration using the total absolute error

$$E_T(\sigma) = \sum_{j=1}^{n} |E_j(\sigma)| \tag{4.8}$$

defined in Equation (4.7). In cases where $E_T(\sigma)$ is near its minimum value over a range of σ, we prefer smaller values of σ over the range since these yield more small values of $g(D_i, \sigma)$ in Equation (4.4) and therefore more terms in the sums that can be neglected without significantly affecting the approximation. Thus, we select the optimal value σ^* of the smoothing parameter as the smallest value of σ for which

$$E_T(\sigma) \leq 1.001 * \min\left[\mathrm{E_T}(\sigma)\right]. \tag{4.9}$$

The use of this equation to favor smaller values of σ has little effect on the accuracy of the model in Equation (4.4) but can improve the efficiency of the computation.

Figures 4.2 to 4.5 plot $E_T(\sigma)$ for each of the four platoon configurations. The resulting values of σ^* are shown in Table 4.1. For small values of σ, the $g(D_{ij}, \sigma)$ in Equation (4.5) are approximately Dirac delta functions and $\widehat{f}(X_j, \sigma)$ is approximately a sum of Dirac delta functions centered at the observations (X_i, y_i) for $i \neq j$. This results in a relatively large error $E_j(\sigma)$ for small σ in Equation (4.6) and a relatively large error in $E_T(\sigma)$ for small σ in Equation (4.8). As σ increases, the approximation in Equation (4.6) improves and the error decreases as shown in the figures.

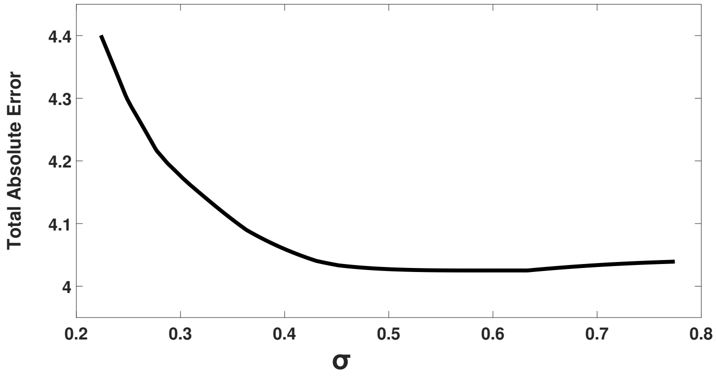

Figure 4.2: $E_T(\sigma)$ for RHP versus RHB configuration, 2016

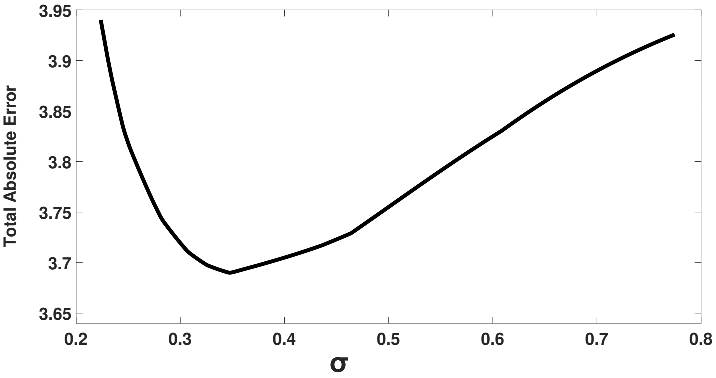

Figure 4.3: $E_T(\sigma)$ for RHP versus LHB configuration, 2016

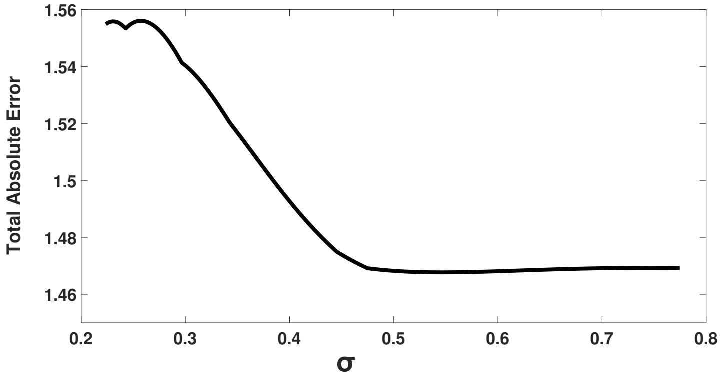

Figure 4.4: $E_T(\sigma)$ for LHP versus RHB configuration, 2016

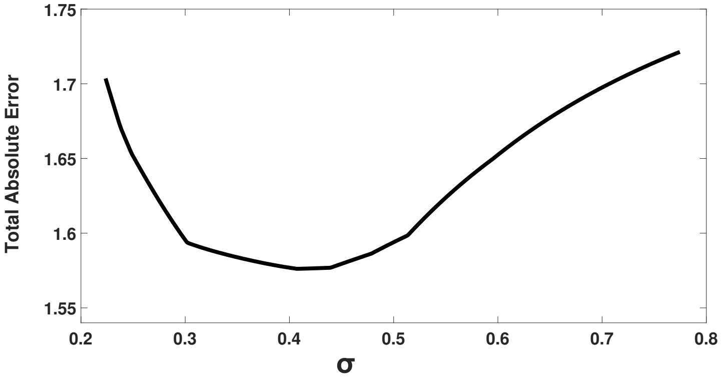

Figure 4.5: $E_T(\sigma)$ for LHP versus LHB configuration, 2016

Table 4.1: Optimized σ^* values found using cross validation

pitcher	batter	σ^*
RHP	RHB	0.48
RHP	LHB	0.34
LHP	RHB	0.48
LHP	LHB	0.39

4.3.4 Finding Optimized Pitch Frequencies

The goal for a pitcher is to maximize his future strikeout rate. This can be accomplished by using the estimated $\widehat{f}(X, \sigma^*)$ function which represents strikeout rate as a function of the pitch distribution X. Suppose that a pitcher has a pitch distribution X which is represented by a signature with m pitch types as in Equation (2.1). Each pitch type i has a pitch parameter vector $\mu_i = (\overline{s}_i, \overline{x}_i, \overline{z}_i)$ and a frequency w_i. For a given pitcher, the pitch parameter vector μ_i for each pitch type is characteristic of his ability and typically does not change. Each frequency w_i, however, can be easily changed by varying how often pitch type i is thrown. Thus, a pitcher can endeavor to maximize future strikeout rate by finding the values of w_i that maximize $\widehat{f}(X, \sigma^*)$ subject to the constraints $w_1 + w_2 + \cdots + w_m = 1$ and

$w_i \geq 0$. Since the number of pitch types m is typically small, the optimal w_i values can be found efficiently using an exhaustive search over combinations of the frequencies w_i.

We illustrate this process for left-handed pitcher Danny Duffy for the LHP vs. LHB platoon configuration using his 2016 signature as shown in Table 4.2. We note that the signature model S in Equation (2.1) is general and can accommodate any number of different pitch types. Individual pitchers, however, typically are not able to throw every pitch type effectively. As reported by Brooks Baseball, Danny Duffy only used the five pitch types listed in Table 4.2 during 2016. Other pitchers use other pitch types such as the cutter and the split which are represented in their signatures. Figure 4.6 is a visualization of $\widehat{f}(X, \sigma^*)$ for pitch distributions X formed by varying the frequency w_1 of his fourseam and w_2 of his slider. In order to limit the plot to two dimensions, the w_i for his two least frequent pitches are set to their 2016 values so that $w_4 = 0.0252$, $w_5 = 0.0069$, and w_3 is then constrained to $w_3 = 1-(w_1+w_2+w_4+w_5)$. The red point in the figure indicates the location of Duffy's 2016 signature and corresponds to an actual strikeout rate of 0.330 and an estimated strikeout rate using $\widehat{f}(X, \sigma^*)$ of 0.317. We see that the model predicts that the pitcher could improve his strikeout rate by increasing w_1 (fourseam frequency) and reducing w_2 (slider frequency). In 2017, Duffy's w_1 and w_2 frequencies for this configuration moved in the opposite direction to the point shown in black in the figure. This resulted in a reduced strikeout rate of 0.245 in 2017 which is consistent with a reduced strikeout rate model prediction as shown in Figure 4.6.

Table 4.2: Pitch signature for LHP Danny Duffy versus LHB for 2016

Pitch type	index	w	$\overline{s}$	$\overline{x}$	$\overline{z}$
Fourseam	1	0.6156	95.96	4.72	11.73
Slider	2	0.2357	84.43	-2.24	-0.85
Sinker	3	0.1167	95.39	8.02	9.21
Change	4	0.0252	86.21	9.79	8.08
Curve	5	0.0069	80.26	-4.26	-5.52

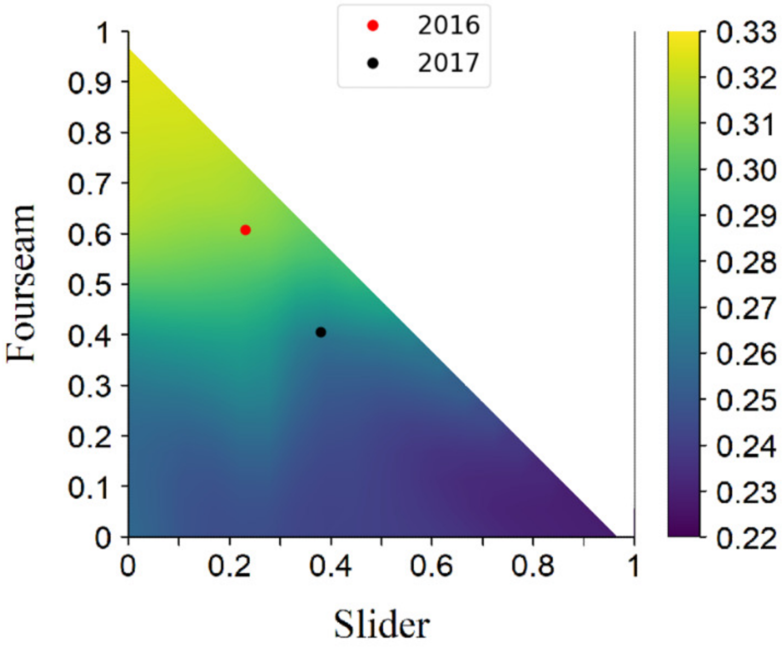

Figure 4.6: Danny Duffy $\widehat{f}(X, \sigma^*)$ for LHP versus LHB configuration, 2016

4.3.5 Predicting Strikeout Rate Changes

We can examine the ability of the $\widehat{f}(X, \sigma^*)$ model estimated from 2016 sensor data to predict pitcher strikeout rate changes as pitch distributions change from 2016 to out-of-sample data in 2017. For this purpose, we considered the 72 right-handed pitchers and 27 left-handed pitchers who threw at least 1500 pitches in both 2016 and 2017. We define a pitcher's actual change in strikeout rate Δ and his predicted change in strikeout rate $\widehat{\Delta}$ for a platoon configuration by

$$\Delta = (2017 \text{ rate}) - (2016 \text{ rate}) \tag{4.10}$$

$$\widehat{\Delta} = (2017 \text{ predicted rate}) - (2016 \text{ rate}) \tag{4.11}$$

where 2017 predicted strikeout rate is computed by evaluating $\widehat{f}(X, \sigma^*)$ using Equation (4.4) for the pitcher's 2017 pitch distribution with σ^* computed as described in Section 4.3.3. Figure 4.7 is a scatterplot with 198 points that represent $(\widehat{\Delta}, \Delta)$ for each of the 72 right-handed and 27 left-handed pitchers against each handedness of batter. We see that the points have a positive correlation. In particular, for the 25 points with strong positive predictions $\widehat{\Delta} > 0.03$ we have 21 points (84.0%) with a positive Δ in actual strikeout rate. For the 39 points with strong negative predictions $\widehat{\Delta} < -0.03$ we have 24 points (61.5%) with a negative Δ in actual strikeout rate. Thus, the model is useful for predicting the dependence of changes in strikeout rate on changes in pitch distribution.

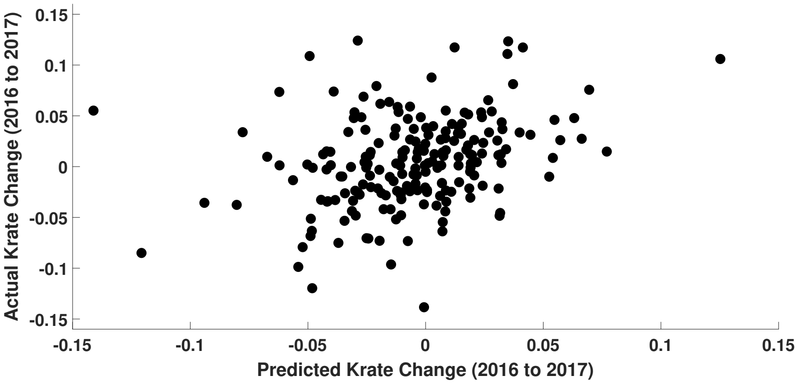

Figure 4.7: Predicting strikeout rate changes using $\widehat{f}(X, \sigma^*)$

For comparison, Figure 4.8 is a scatterplot of the actual change in strikeout rate from 2016 to 2017 for each of the 99 pitchers versus each pitcher's Nash score difference [47]

$$\Delta_N = 2016 \text{ Nash score} - 2017 \text{ Nash score} \tag{4.12}$$

As described in Section 4.3.1, a low Nash score indicates that a pitcher is close to Nash equilibrium while a higher Nash score indicates that a pitcher is farther from equilibrium. Thus, for Δ_N positive we would expect a pitcher to improve from 2016 to 2017 and for Δ_N

negative we would expect a pitcher to get worse from 2016 to 2017. In Figure 4.8, however, we see that the points in the scatterplot do not have an increasing trend and, in fact, the points have a small negative correlation. We believe that this is due to the low reliability for the pitch values [1] on which the Nash score is based.

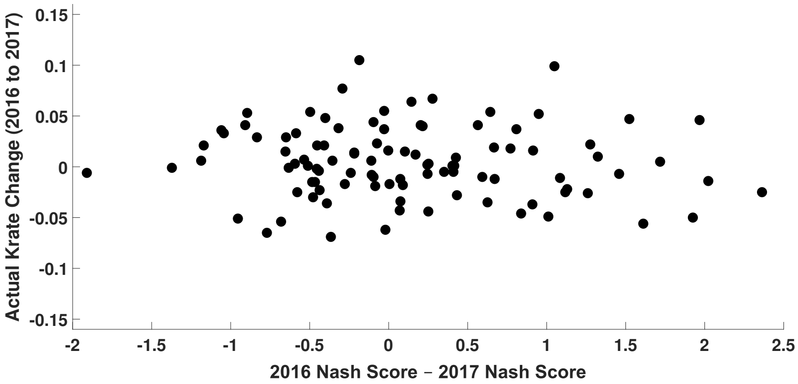

Figure 4.8: Predicting strikeout rate changes using Nash score changes

We can assess the statistical significance of the difference between the correlation coefficients of $r_1 = 0.320$ in Figure 4.7 and $r_2 = -0.081$ in Figure 4.8 using the Fisher z-transformation [16]. Even if we disregard the negative sign on r_2, this method yields a z_{observed} test statistic of 2.01 and a corresponding p-value of 0.044 which supports the conclusion that r_1 is significantly larger than r_2. Thus, the function $\widehat{f}(X, \sigma^*)$ has value for predicting future strikeout rate and can be used to find optimized pitch frequencies w_i using the approach described in Section 4.3.4.

4.4 Summary

We have developed and evaluated an algorithm for learning a function over distributions. The algorithm employs the earth mover's distance as a metric for distribution space within a nonparametric kernel regression scheme. We have demonstrated the algorithm for the task of learning a pitcher's strikeout rate as a function of a multidimensional pitch distribution that is generated from pitch trajectory measurements. The algorithm efficiently represents the pitch distributions using signatures and compensates for the correlation of the trajectory variables with a whitening step. The smoothing parameter for the regression kernel is learned using cross validation. We have assessed the algorithm for the prediction of strikeout rate from pitch distributions on out-of-sample data and have demonstrated that it performs better than an alternative algorithm based on game theory principles.

Chapter 5

Measurement Space Partitioning for Estimation and Prediction

5.1 Overview

Player talent level on batted balls is defined as the expected value of a statistic which can be estimated from a sample of observations. The utility of an estimate is often evaluated by its ability to predict player performance on unobserved data. An intuitively appealing estimate of talent level is simply the computed value of the statistic over a player's observed sample. But paradoxically this method is less accurate than estimators [12] [29] [57] that are defined by a weighted average of this computed value and the mean of the statistic over a group of players. An example of these estimators is linear regression (LR) for which the weighting depends on the correlation of the value of the statistic across samples. LR estimates have been used by several systems to predict player performance [55][60].

In recent years radar and optical sensors have been used in MLB stadiums to measure characteristics of batted balls such as speed, direction, and spin [22]. We use these sensor measurements to develop a new method for estimating talent level called measurement space partitioning (MSP). After constructing a discrete batted-ball distribution defined over a partition of the multidimensional measurement space for samples from a group of players, we use Cronbach's alpha to show that the expected correlation of distribution values across samples has a strong dependence on location in measurement space. This allows a player's underlying batted-ball distribution and the corresponding talent level to be estimated using regression parameters that adapt to his specific batted-ball distribution. The accuracy of the talent level estimate depends on the partition which leads to the derivation of a method for partition optimization. A set of experiments is used to show that the MSP method improves on the accuracy of linear regression for estimating batted-ball talent level.

Another advantage of the MSP approach is the ability to incorporate fine-grained contextual information into estimates. Contextual information includes a range of variables that can affect batted-ball value. The weather conditions and elevation, for example, will affect how far a batted ball will carry in the air [2]. Batted balls that follow similar trajectories can have different outcomes due to differences in outfield geometry from ballpark to ballpark [17]. A player's running speed [24] and variables that include the height of the infield grass and the composition of the infield surface [4] can affect the value of batted balls hit on the ground. The fate of batted balls also depends on the quality of the defenders in the field. Contextual factors are typically accounted for by a coarse adjustment that compensates for the average effect of the environment [48]. Since the MSP method computes talent level estimates from regressed batted ball distributions defined over physical parameters, contextual adjustments can be employed that depend on the characteristics of individual batted balls. A ball hit in the air at high speed, for example, can be adjusted differently from a ball hit softly on the ground. We will show that the use of fine-grained contextual adjustments improves the accuracy of predictions made by the MSP method.

5.2 Estimation and Prediction

5.2.1 Talent Level

Talent for a skill varies from player to player and can be represented by a statistic that is derived from a set of observations. The computed value of such a statistic equals talent level $T(j)$, which is the expected value of the statistic for player j, plus estimation error. In this work, we examine the problem of estimating player talent level on batted balls. Consider a dataset that contains information on $2N$ batted balls for each of P players where the data is arranged so that the first N batted balls for each player are observed and the second N batted balls for each player are unobserved. Let $R(i, j)$ represent the numerical value of batted ball i for player j and define the observed performance statistic for player j as the average over the first N batted balls

$$x(j) = \frac{1}{N} \sum_{i=1}^{N} R(i, j) \tag{5.1}$$

and define the unobserved performance for player j as the average over the second N batted balls

$$y(j) = \frac{1}{N} \sum_{i=N+1}^{2N} R(i, j). \tag{5.2}$$

We consider the task of using the observed batted ball data to estimate talent level $T(j)$ for the $x(j)$ statistic for each player j. The estimated $T(j)$ can be used to predict the unobserved performance $y(j)$.

5.2.2 Linear Regression

One estimate of $T(j)$ is the observed performance $x(j)$ for player j. However, the James-Stein paradox [29] [57] as illustrated by Effron and Morris [12] shows that a more accurate estimate of $T(j)$ is obtained by adjusting the $x(j)$ using an average of the observed $R(i, j)$ values over multiple players. Since an estimate for talent level can be assessed by its ability to predict the unobserved performance $y(j)$, we can define an estimate $\widehat{y}(j)$ for $T(j)$ by minimizing the sum of the square errors

$$E = \sum_{j=1}^{P} (y(j) - \widehat{y}(j))^2 \tag{5.3}$$

using the linear regression model

$$\widehat{y}(j) = a + bx(j). \tag{5.4}$$

The values of a and b that minimize E are

$$a = \mu_y - \frac{r\mu_x\sigma_y}{\sigma_x} \tag{5.5}$$

$$b = \frac{r\sigma_y}{\sigma_x} \tag{5.6}$$

where μ_x and σ_x are the mean and standard deviation for the $x(j)$, μ_y and σ_y are the mean and standard deviation for the $y(j)$, and r is the correlation coefficient for the set of P points $(x(j), y(j))$ [10].

Since the data used to generate the $y(j)$ are unobserved, the parameters μ_y, σ_y, and r in equations (5.5) and (5.6) cannot be computed directly. The $y(j)$, however, are generated in the same way for the same players as the $x(j)$ which allows us to use the approximations

$\mu_y = \mu_x$ and $\sigma_y = \sigma_x$. The remaining unknown parameter, the correlation coefficient r, can be approximated from the observed $R(i,j)$ values using Cronbach's alpha [9]

$$\alpha(N) = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_{R_i}^2}{\sigma_{R_T}^2}\right) \tag{5.7}$$

where $\sigma_{R_i}^2$ is the variance of the observed $R(i,j)$ values over players j for batted ball i and $\sigma_{R_T}^2$ is the variance of

$$R_T(j) = \sum_{i=1}^{N} R(i,j) \tag{5.8}$$

over players j. Using these approximations, equation (5.4) becomes

$$\widehat{y}(j) = \alpha(N)x(j) + (1 - \alpha(N))\mu_x \tag{5.9}$$

which can be computed using the observed data. $\widehat{y}(j)$ in equation (5.9) is consistent with the James-Stein result that an improved estimate for $T(j)$ can be obtained by adjusting $x(j)$ using the overall mean μ_x.

5.2.3 Varying Observed Sample Size

The $\alpha(N)$ that is used to compute the estimate $\widehat{y}(j)$ in equation (5.9) is derived using a dataset of N observed batted balls for each of P players using equation (5.7). The utility of the method is enhanced if we can use this dataset to compute the estimate $\widehat{y}(j)$ using a sample of N' batted balls for player j where $N' \neq N$. The value of $\alpha(N)$ tends to increase with N due to a decrease in the variance of the random error in the observed performance $x(j)$ [67]. The Spearman-Brown prophecy formula [5] [56] allows us to predict $\alpha(N')$ from the estimated

$\alpha(N)$ using

$$\alpha(N') = \frac{C\alpha(N)}{1 + (C-1)\alpha(N)} \tag{5.10}$$

where $C = N'/N$. This $\alpha(N')$ can be used in equation (5.9) to compute $\widehat{y}(j)$ using an observed performance $x(j)$ computed using any number of samples N'.

5.3 Exploiting Sensor Measurements

5.3.1 Partitioning the Measurement Space

Sensors allow batted balls to be represented by a point in a measurement space with dimensions defined by properties such as speed, direction, and spin. The measurement space can be partitioned into B disjoint subsets. For the dataset described in Sec. 5.2.1 let $M(i, j, k)$ be a binary-valued function which is one if batted ball i for player j is in subset k and zero otherwise. Define the observed batted ball distribution for player j over the subsets k by

$$p_x(j,k) = \frac{1}{N}\sum_{i=1}^{N} M(i,j,k) \tag{5.11}$$

and define the unobserved batted ball distribution for player j over the subsets k by

$$p_y(j,k) = \frac{1}{N}\sum_{i=N+1}^{2N} M(i,j,k). \tag{5.12}$$

We will show that an estimate for $p_y(j, k)$ can be used to generate an estimate for the talent level $T(j)$.

5.3.2 Estimating Measurement Space Distributions

For a given subset k we can use a linear regression model and approximations similar to those described in Sec. 5.2 to estimate $p_y(j,k)$ from the observed data according to

$$\widehat{p}_y(j,k) = \alpha(N,k)p_x(j,k) + (1 - \alpha(N,k))\mu(k) \tag{5.13}$$

where $\mu(k)$ is the average

$$\mu(k) = \frac{1}{P}\sum_{j=1}^{P} p_x(j,k) \tag{5.14}$$

and $\alpha(N,k)$ is the Cronbach approximation to the correlation coefficient for the set of P points $(p_x(j,k), p_y(j,k))$ for subset k. Specifically, $\alpha(N,k)$ is computed using

$$\alpha(N,k) = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{M_i}^2}{\sigma_{M_T}^2}\right) \tag{5.15}$$

where $\sigma_{M_i}^2$ is the variance of the observed $M(i,j,k)$ values over players j for batted ball i and subset k and $\sigma_{M_T}^2$ is the variance of

$$M_T(j) = \sum_{i=1}^{N} M(i,j,k) \tag{5.16}$$

over players j for subset k. $\alpha(N,k)$ can then be used in equation (5.13) to compute the regressed distribution $\widehat{p}_y(j,k)$ using only the observed data. We note that the calculation in equation (5.15) can yield $\alpha(N,k)$ values that are negative [67] and in these cases $\alpha(N,k)$ is set to zero for the calculation of $\widehat{p}_y(j,k)$.

5.3.3 Estimating Talent Using Measurement Space Partitioning

The batted ball distribution estimate $\widehat{p}_y(j,k)$ for player j can be used to estimate the player's talent level $T(j)$. If $\overline{R}(j,k)$ is an estimate of the expected value of batted balls for player j in subset k then $T(j)$ can be estimated by

$$\widehat{y}_s(j) = \sum_{k=1}^{B} \widehat{p}_y(j,k)\overline{R}(j,k). \tag{5.17}$$

For cases where we would like to estimate $\widehat{y}_s(j)$ using a sample of N' batted balls for player j, the values $\alpha(N',k)$ for each k in equation (5.13) can be computed using the Spearman-Brown formula as described in Sec. 5.2.3.

The $\widehat{y}_s(j)$ estimate in equation (5.17) is equivalent to the linear regression estimate $\widehat{y}(j)$ in equation (5.9) if $\alpha(N,k)$ has the same value $\alpha(N)$ for all subsets k and the average value of the observed batted balls in any subset k is the same for all players j. For this special case, if we let $\overline{R}(j,k)$ equal the overall mean of the observed $R(i,j)$ for subset k

$$\overline{R}(k) = \frac{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{P} M(i,j,k)R(i,j)}{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{P} M(i,j,k)} \tag{5.18}$$

then equation (5.17) can be written

$$\begin{aligned}\widehat{y}_s(j) &= \sum_{k=1}^{B} \left[\alpha(N)p_x(j,k) + (1-\alpha(N))\mu(k)\right]\overline{R}(k) \\ &= \left[\alpha(N)\sum_{k=1}^{B} p_x(j,k)\overline{R}(k)\right] + \left[(1-\alpha(N))\sum_{k=1}^{B}\mu(k)\overline{R}(k)\right]\end{aligned} \tag{5.19}$$

where the first sum in equation (5.19) equals $x(j)$ and the second sum equals μ_x which demonstrates the equivalence to equation (5.9). We will see that by allowing $\alpha(N, k)$ to vary over subsets k and by allowing $\overline{R}(j, k)$ to vary over players j, the model in equation (5.17) can generate estimates that are more accurate than the linear regression estimate in equation (5.9).

5.4 Experimental Results

5.4.1 Sensor Data

The TM radar has been used by MLB's Statcast system [22] since 2017 to track and characterize batted balls. The TM radar operates in the X-band at approximately 10.5 GHz and is positioned high behind home plate. The measured initial speed s_l and vertical launch angle v (Figure 5.1) for batted balls play an important role in determining batted ball value [21]. In particular, batters tend to achieve the best results for batted balls with an initial speed of greater than 90 miles per hour and a vertical launch angle between 10 and 30 degrees.

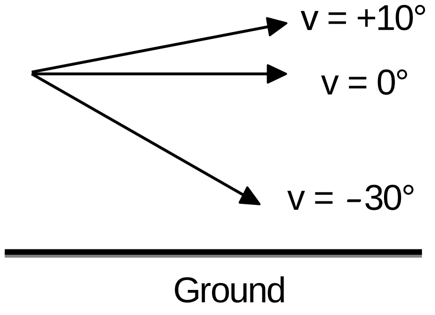

Figure 5.1: Vertical launch angle v

5.4.2 Representing Batted Ball Value

Many statistics [48] can be used to quantify a batter's performance on batted balls. Batting average, for example, is the fraction of batted balls that result in a hit but has the deficiency that all hits are given equal value. Slugging percentage allocates different weights to different kinds of hits, e.g. single or double, but has been shown to overweight doubles, triples, and home runs. Weighted on base average (wOBA) [60] uses weights for each batted ball outcome that are proportional to run value and, for this reason, we use wOBA to represent batted ball value $R(i, j)$.

5.4.3 Contextual Information

A batted ball with a given set of physical parameters such as s_l and v occurs in a context that can affect its value. Variation in the outfield geometry across stadiums [17] and variation in the ambient weather conditions [2] can affect the value of a ball hit in the air. The batter's running speed [24] plays a role in determining batted ball value especially for balls hit on the ground. The quality of defenders can also affect the value of a batted ball hit to a given region of the field. These factors cause the batted ball value $\overline{R}(j, k)$ for subset k to vary depending on the distribution of contextual variables for player j. We will show later in this section how contextual information can be combined with the batted ball distribution estimates $\widehat{p}_y(j, k)$ to improve the accuracy of the $\widehat{y}_s(j)$ predictions.

5.4.4 Assessing Prediction Accuracy

Statcast data from MLB games in 2019 was employed to evaluate methods for using observed data to predict player performance in unobserved data. After removing bunts from the dataset, each of the $P = 159$ players with at least 300 batted balls during the 2019 season was considered. Switch-hitters who bat both right-handed and left-handed were regarded as a different batter for each handedness. The first 300 batted balls for each player were divided into an observed set of $N = 150$ batted balls and an unobserved set of $N = 150$ batted balls. The odd batted balls in chronological order for each player defined the observed set and the even batted balls defined the unobserved set. The batted ball value $R(i, j)$ for batted ball i and player j was defined by the wOBA weight for the batted ball result as described in Sec. 5.4.2. For the 2019 MLB season the wOBA weights are out=0.000, single=0.870, double=1.217, triple=1.529, homerun=1.940, and batter reaches on error= 0.920 [65]. The observed batted ball data was used to generate predictions for the unobserved performance $y(j)$. The accuracy of a set of predictions $\widehat{y}(j)$ is evaluated using the sum of squared errors (SSE)

$$SSE = \sum_{j=1}^{P} (y(j) - \widehat{y}(j))^2 \tag{5.20}$$

between the unobserved performance and its prediction.

5.4.5 Linear Regression

The linear regression model defined by equation (5.9) was used to generate the $\widehat{y}(j)$ predictions for the data described in Sec. 5.4.4. The resulting model is

$$\widehat{y}(j) \;=\; 0.294x(j) + (1 - 0.294) \cdot 0.402 = 0.294x(j) + 0.284 \tag{5.21}$$

where the observed batted ball data was used to compute $\alpha(150) = 0.294$ and $\mu_x = 0.402$ as described in Sec. 5.2.2. This model gives an SSE of 0.647 using equation (5.20). Two boundary instances of the linear regression model are the naive prediction $\widehat{y}(j) = x(j)$ for

$\alpha(N) = 1$ and the baseline prediction $\widehat{y}(j) = \mu_x$ for $\alpha(N) = 0$. For this dataset, the naive prediction gives an SSE of 0.780 and the baseline prediction gives an SSE of 0.743 which are both larger than the SSE obtained using the linear regression model in equation (5.21). The $\widehat{y}(j)$ prediction lines for the linear regression model and the naive and baseline predictions are shown in Figure 5.2 along with the $(x(j), y(j))$ points for each of the 159 players.

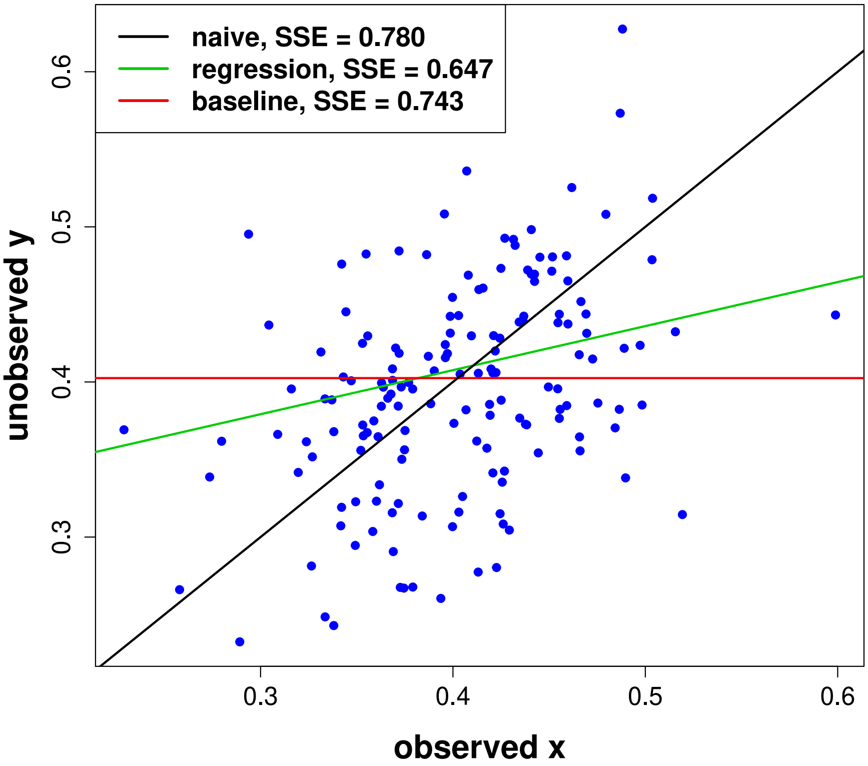

Figure 5.2: $(x(j), y(j))$ points with naive, regression, and baseline predictions

5.4.6 Measurement Space Partitioning

The measured initial speed and launch angle can be used to represent a batted ball as a point in a two-dimensional (s_l, v) measurement space. This space can be partitioned into B disjoint subsets as described in Sec. 5.3.1. In Appendix D, we show that the accuracy of the prediction in equation (5.17) depends on the partition. In this section we define different ways to partition the (s_l, v) measurement space and show how training data can be used to optimize partition selection.

Partition Definition

The (s_l, v) space can be divided into an internal region defined by

$$(s_{l,\text{min}} \leq s_l < s_{l,\text{max}}) \text{ and } (v_{\text{min}} \leq v < v_{\text{max}})$$

which includes the large majority of batted balls and four boundary regions B_1, B_2, B_3, B_4 defined by

$$\begin{aligned}
B_1 : &\quad s_l < s_{l,\text{min}} \\
B_2 : &\quad s_l \geq s_{l,\text{max}} \\
B_3 : &\quad (s_{l,\text{min}} \leq s_l < s_{l,\text{max}}) \text{ and } (v < v_{\text{min}}) \\
B_4 : &\quad (s_{l,\text{min}} \leq s_l < s_{l,\text{max}}) \text{ and } (v \geq v_{\text{max}}).
\end{aligned}$$

The internal region can be further divided into rectangular subregions $b(i, j)$ of dimension $s_{l,\text{width}} \times v_{\text{width}}$ which are defined by

$$\begin{aligned}
b(i, j) : &\quad (s_{l,\text{min}} + (i-1) * s_{l,\text{width}}) \leq s_l < (s_{l,\text{min}} + i * s_{l,\text{width}}) \text{ and} \\
&\quad (v_{\text{min}} + (j-1) * v_{\text{width}}) \leq v < (v_{\text{min}} + j * v_{\text{width}})
\end{aligned}$$

so that there are a total of

$$\frac{(s_{l,\max} - s_{l,\min})(v_{\max} - v_{\min})}{s_{l,\text{width}} * v_{\text{width}}}$$

$b(i,j)$ subregions.

We defined the internal and boundary regions for the 2019 data using $s_{l,\min} = 37.5$ mph, $s_{l,\max} = 117.5$ mph, $v_{\min} = -75°$, and $v_{\max} = 85°$ which yields an internal region that includes 99.5 percent of all batted balls. The internal region was partitioned into different configurations of fixed-size rectangular subregions $b(i,j)$ where the subregion widths were allowed to vary over the values

$$\begin{aligned} s_{l,\text{width}} &= 2.5, 5, 10, 20, 40, 80 \text{ mph} \\ v_{\text{width}} &= 2.5, 5, 10, 20, 40, 80, 160 \text{ degrees} \end{aligned}$$

By considering all combinations of the six $s_{l,\text{width}}$ values and the seven v_{width} values we can define 42 partitions with each denoted $\mathcal{P}_{s_{l,\text{width}}, v_{\text{width}}}$ where the boundary regions B_1, B_2, B_3, B_4 are the same for each partition. Figure 5.3, for example, depicts the $\mathcal{P}_{10,40}$ partition with the four boundary regions and thirty-two internal subregions $b(i,j)$ where $b(2,3)$ is explicitly labeled.

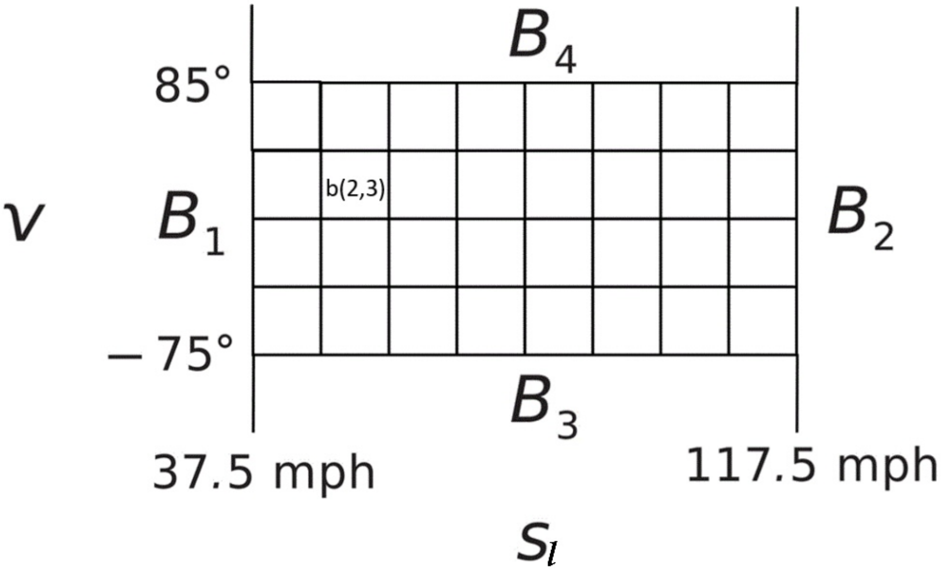

Figure 5.3: The $\mathcal{P}_{10,40}$ partition of measurement space

The prediction method described in Sec. 5.3.3 was used to process the observed and unobserved data described in Sec. 5.4.4 using each of the 42 partitions. For the finer partitions the observed data does not contain enough samples to reliably estimate $\overline{R}(j,k)$ for each (j,k). Therefore, the mean $\overline{R}(k)$ in equation (5.18) was used to approximate $\overline{R}(j,k)$ for each j. The smallest SSE of 0.532 was obtained for $\mathcal{P}_{2.5,40}$ while the largest SSE of 0.743 was obtained for $\mathcal{P}_{80,160}$. If we neglect the effect of the boundary regions, the use of $\mathcal{P}_{80,160}$ is equivalent to the baseline prediction $\widehat{y}(j) = \mu_x$ for which we also reported an SSE of 0.743 in Sec. 5.4.4.

Partition Selection

Partition selection is an important issue since there are large differences in the SSE for different partitions. To address this issue, we examine whether the analysis of previous year data can be used to optimize partition selection for current year data. To this end, we computed the SSE for each of the 42 partitions defined in Sec. 5.4.6 using 2018 batted ball data arranged as described in Sec. 5.4.4 for the 2019 data. There were $P = 158$ players with at least 300 batted balls in 2018 that were considered for analysis. Figure 5.4 plots the (SSE 2018, SSE 2019) point for each of the 42 partitions and we see that there is a strong correlation between the SSE values for the two years. In particular, the partitions that give the smallest SSE values in 2018 also give the smallest SSE values in 2019. This result suggests that we can use previous year data to select an optimized partition for current year data. The $\mathcal{P}_{5,10}$ partition gives the smallest SSE of .419 on 2018 data. Using this partition for the 2019 data gives an SSE of 0.546 which is close to the smallest value of 0.532 and significantly better than the linear regression SSE of 0.647 reported in Sec. 5.4.5.

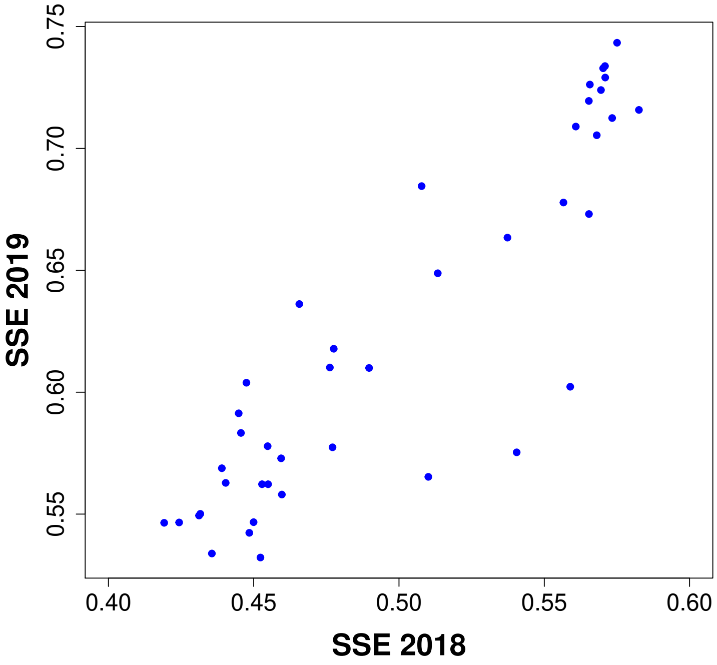

Figure 5.4: Prediction SSE in 2018 and 2019 for 42 partitions

Example

In this section we illustrate the mechanics of the MSP method using the 2019 batted ball data. The example considers the $\mathcal{P}_{5,10}$ partition defined in Sec. 5.4.6 that was selected using prior year data as described in Sec. 5.4.6. Figure 5.5 plots the $\alpha(150, k)$ function and Figure 5.6 plots the mean $\mu(k)$ function over the subregions k for this partition. The $\alpha(150, k)$ function is approximately in the shape of a rotated V with most of the larger values occurring for s_l greater than 95 mph. Figures 5.7 and 5.8 demonstrate properties of $\alpha(150, k)$ and $\mu(k)$ for specific subregions S_1 and S_2 of $\mathcal{P}_{5,10}$ defined by

$$S_1: \ (87.5 \text{ mph} \leq s_l < 92.5 \text{ mph}) \ \text{ and } \ (5^\circ \leq v < 15^\circ)$$
$$S_2: \ (107.5 \text{ mph} \leq s_l < 112.5 \text{ mph}) \ \text{ and } \ (15^\circ \leq v < 25^\circ)$$

which correspond respectively to $b(11, 9)$ and $b(15, 10)$ using the notation in Sec. 5.4.6. The observed data described in Sec. 5.4.4 gives values of

$$\alpha(150, S_1) = 0.01, \quad \mu(S_1) = 0.017,$$
$$\alpha(150, S_2) = 0.61, \quad \mu(S_2) = 0.011$$

which predict little correlation between the fraction of batted balls in the observed and unobserved data for S_1 and a larger correlation between the fraction of batted balls in the observed and unobserved data for S_2. Figure 5.7 plots the $P = 159$ points $(p_x(j, S_1), p_y(j, S_1))$ as defined by equations (5.11) and (5.12) along with the prediction line from equation (5.13) where each point in the figure has been moved by a small random amount to increase the visibility of the points. There is little correlation between the $p_x(j, S_1)$ and the $p_y(j, S_1)$ as predicted by the small estimated value of $\alpha(150, S_1)$. Figure 5.8 is the same plot for S_2 where the points have a larger positive correlation as predicted by $\alpha(150, S_2)$. In each figure the red prediction line agrees reasonably well with the structure of the data.

Figure 5.9 displays the full observed distribution $p_x(j, k)$ for player j = Jorge Polanco as left-handed batter using $\mathcal{P}_{5,10}$. Figure 5.10 is the corresponding regressed distribution $\widehat{p}_y(j,k)$ computed using equation (5.13). We see that the regressed distribution captures the overall structure of $p_x(j,k)$ but is substantially smoother. The regressed distribution results in a talent level estimate $\widehat{y}_s(j)$ in equation (5.17) of .397. This $\widehat{y}_s(j)$ is much closer to the unobserved performance $y(j)$ of 0.386 than the LR prediction $\widehat{y}(j) = .424$ or the naive prediction of $x(j) = .475$ which corresponds to the observed distribution shown in Figure 5.9.

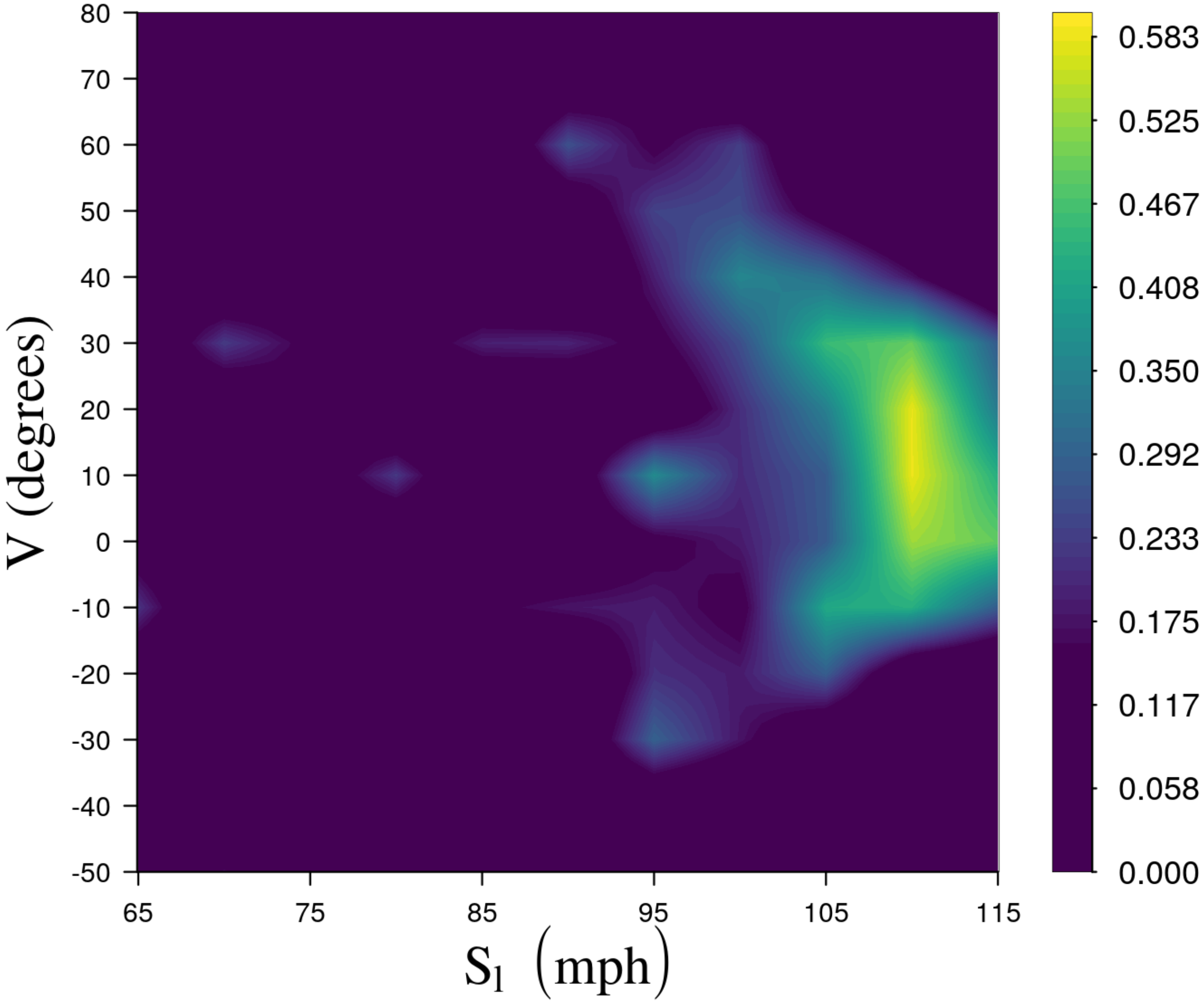

Figure 5.5: $\alpha(150,k)$ for $\mathcal{P}_{5,10}$ partition

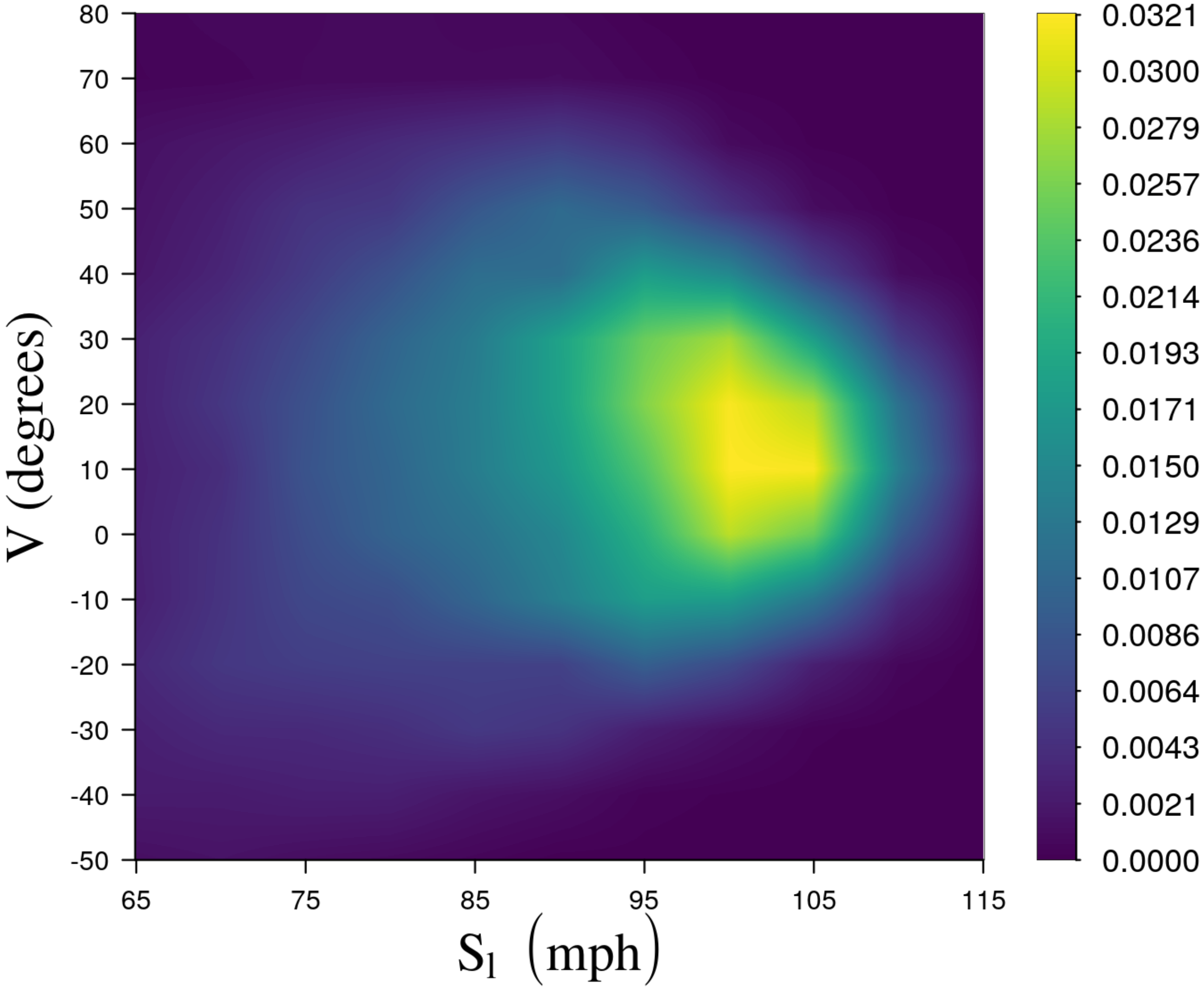

Figure 5.6: $\mu(k)$ for $\mathcal{P}_{5,10}$ partition

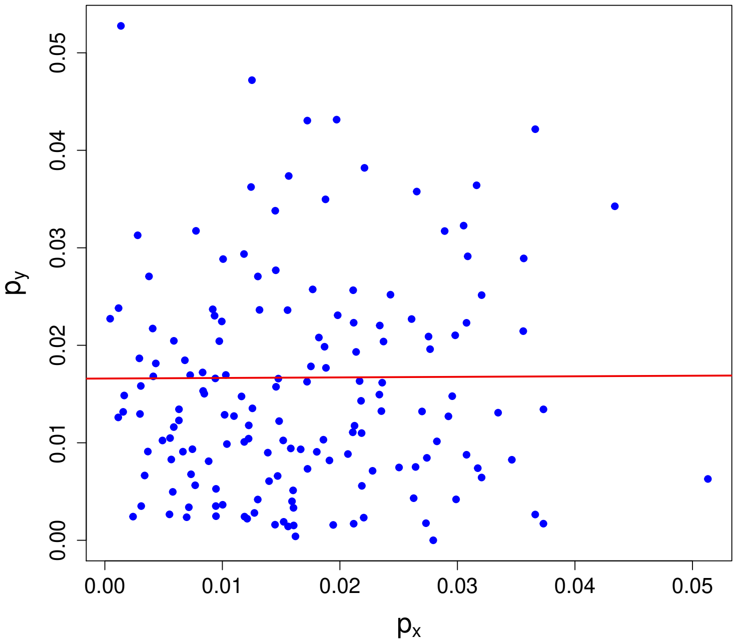

Figure 5.7: $p_y(j, S_1)$ versus $p_x(j, S_1)$ with $\alpha(150, S_1) = 0.01$

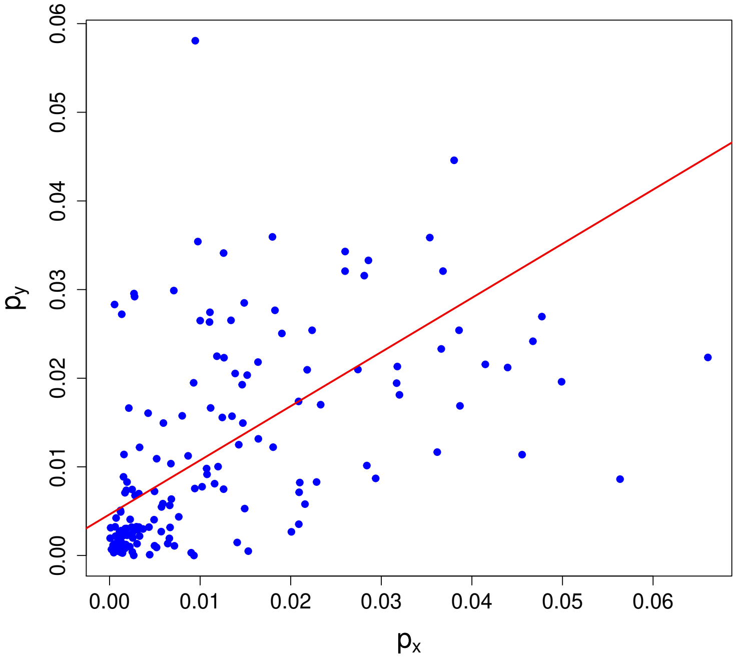

Figure 5.8: $p_y(j, S_2)$ versus $p_x(j, S_2)$ with $\alpha(150, S_2) = 0.61$

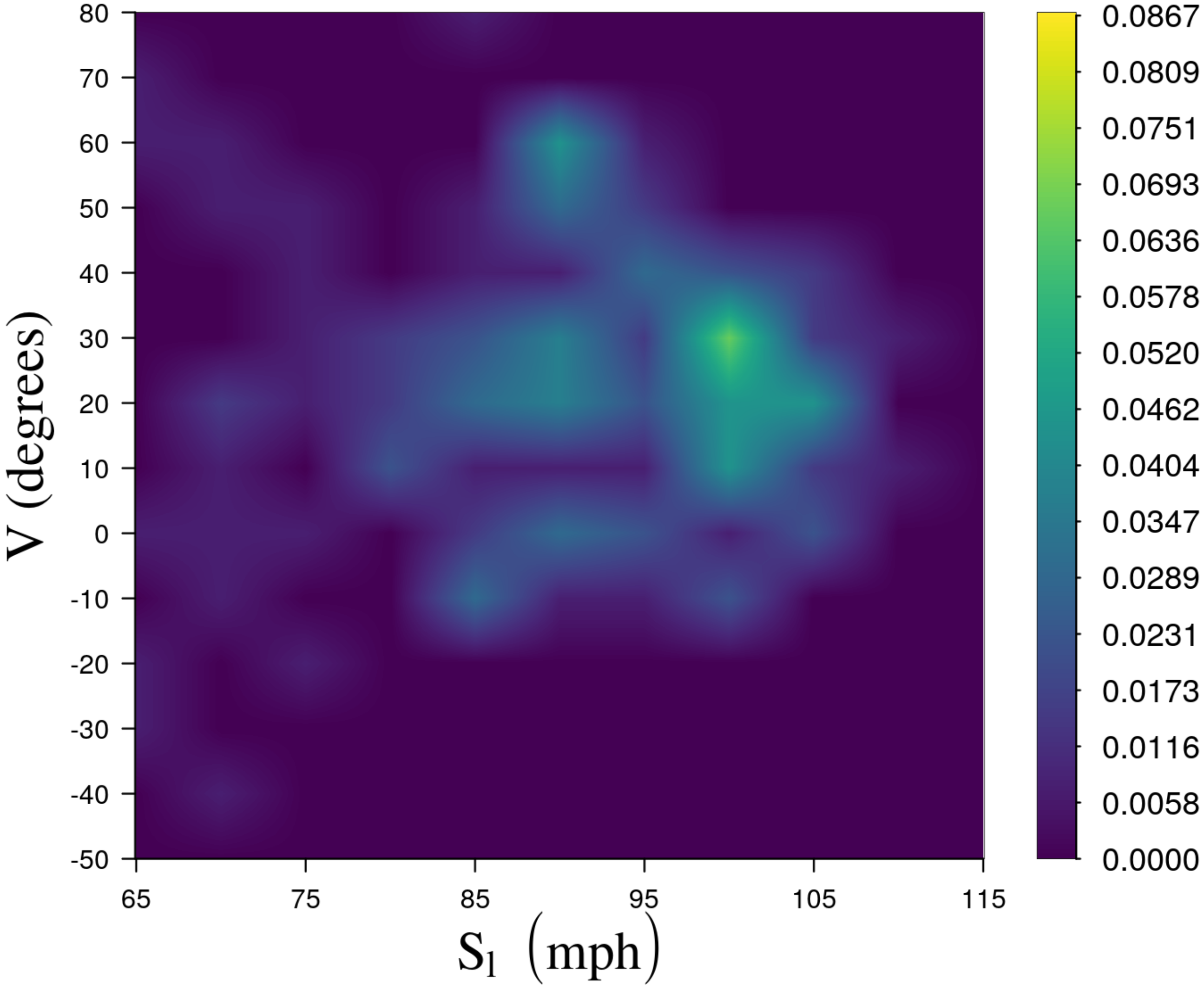

Figure 5.9: Observed distribution $p_x(j, k)$ for Jorge Polanco as left-handed batter

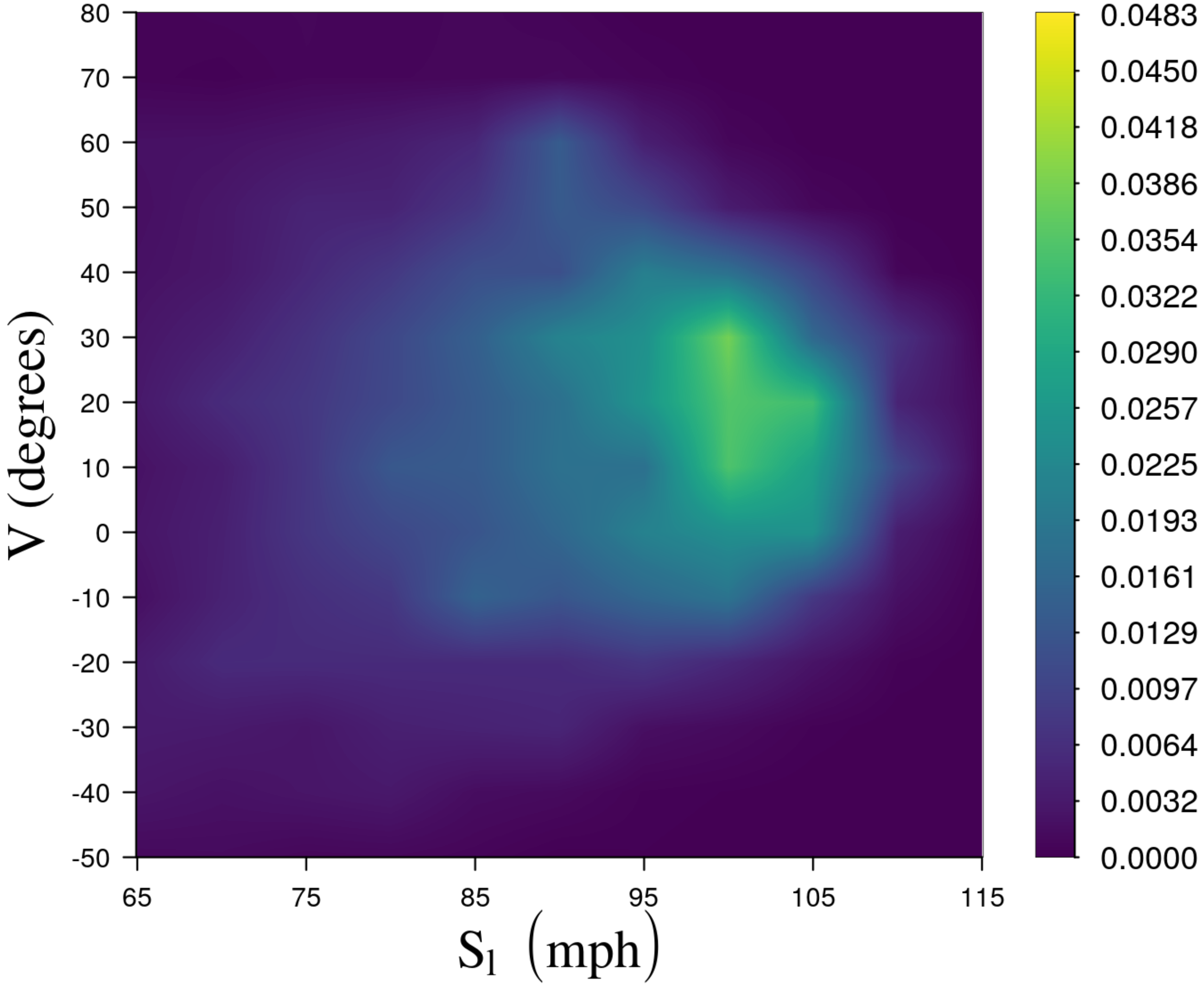

Figure 5.10: Regressed distribution $\widehat{p}_y(j,k)$ for Jorge Polanco as left-handed batter

Comparison with Linear Regression

In this section we compare properties of the LR and MSP predictions. For the data described in Sec. 5.4.4 the LR prediction is defined by the line (equation (5.21)) plotted in Figure 5.11. This figure also plots the 159 $\widehat{y}_s(j)$ predictions for the same data using the $\mathcal{P}_{5,10}$ partition. We see that players j_1 and j_2 with the same observed performance $x(j_1) = x(j_2)$ and therefore the same LR prediction $\widehat{y}(j_1) = \widehat{y}(j_2)$ can be assigned different MSP predictions $\widehat{y}_s(j_1) \neq \widehat{y}_s(j_2)$.

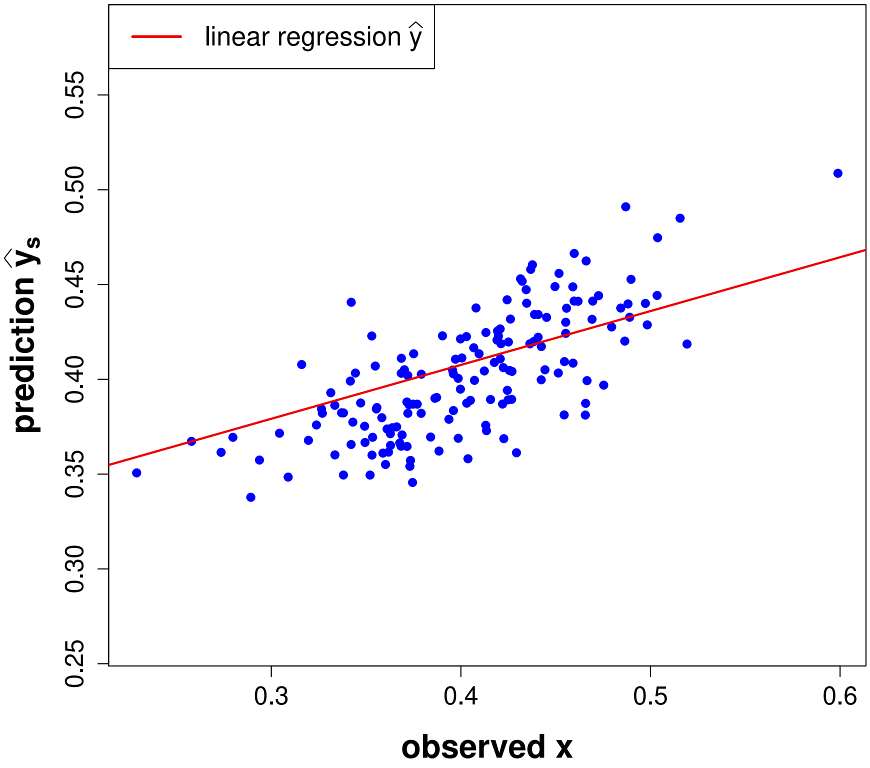

Figure 5.11: $\widehat{y}_s(j)$ predictions for 159 batters using $\mathcal{P}_{5,10}$ partition and LR line

In Sec. 5.3.3 we showed that an important difference between $\widehat{y}(j)$ and $\widehat{y}_s(j)$ is that the former is defined using a single $\alpha(N)$ while the latter employs a separate $\alpha(N, k)$ for each subset k. Players with an observed batted ball distribution $p_x(j, k)$ that includes a large fraction of batted balls in subsets k with large values of $\alpha(N, k)$ will have less regression to the mean in the calculation of $\widehat{y}_s(j)$ than players with a batted ball distribution that has smaller values of $\alpha(N,k)$. This allows the $\widehat{y}_s(j)$ prediction to adapt the amount of regression to a player's collection of batted balls. By comparing equations (5.13) and (5.17) with the LR model of equation (5.9) we see that the correlation-weighted expected wOBA

$$C(j) = \sum_{k=1}^{B} \alpha(N,k) p_x(j,k) \overline{R}(k) \tag{5.22}$$

should capture a large fraction of the variance in the difference $\widehat{y}_s(j) - \widehat{y}(j)$. Figure 5.12 is a scatterplot of $\widehat{y}_s(j) - \widehat{y}(j)$ versus $C(j)$ for the $\mathcal{P}_{5,10}$ partition of the 2019 data which shows that the variables have a strong relationship as expressed by a correlation coefficient of 0.87. Thus, $C(j)$ is a batter-controlled component of $\widehat{y}_s(j)$ that measures the combined value and α-correlation of a player's batted balls and is strongly related to the deviation of a player's $\widehat{y}_s(j)$ prediction from the LR prediction $\widehat{y}(j)$.

Table 5.1 considers four players with similar $\widehat{y}(j)$ LR predictions. The table also shows that several of the players have significant differences in correlation-weighted expected wOBA C. The players (Hernandez, DeJong) with below average values of C have negative $\widehat{y}_s - \widehat{y}$ differences while the players (Acuna, Donaldson) with above average values of C have positive $\widehat{y}_s - \widehat{y}$ differences as predicted by Figure 5.12. We see from the last two columns of the table that these differences benefit the MSP prediction as the LR prediction error $\widehat{y} - y$ is larger in absolute value than the MSP prediction error $\widehat{y}_s - y$ in each case.

Table 5.1: Players with similar $\widehat{y}$, 2019

Player	Hand	$\widehat{y}$	C	$\widehat{y}_s - \widehat{y}$	$\widehat{y} - y$	$\widehat{y}_s - y$
Cesar Hernandez	Left	.410	.054	-.049	.106	.057
Paul DeJong	Right	.410	.082	-.021	.067	.047
Ronald Acuna Jr.	Right	.411	.144	.040	-.077	-.036
Josh Donaldson	Right	.411	.146	.042	-.081	-.039

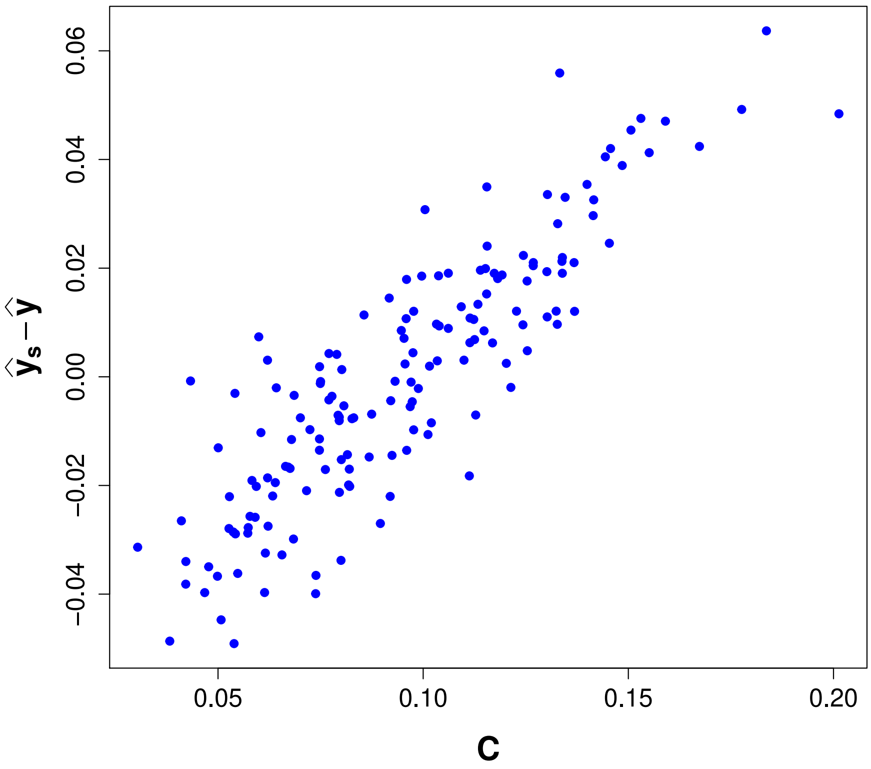

Figure 5.12: Prediction difference $\widehat{y}_s(j) - \widehat{y}(j)$ versus correlation-weighted expected wOBA C

Incorporating Contextual Information

In Sec. 5.4.3 we described several contextual factors that can affect the value of a batted ball with parameters (s_l, v). Accounting for each of these factors can improve the accuracy of the MSP predictions. In this section we describe a method that can be used to estimate $\overline{R}(j, k)$ in equation (5.17) to account for the effects of varying outfield geometry and atmospheric conditions across ballparks. Since a player j typically plays about half of his games in a single home park these effects can have a significant impact on $\overline{R}(j, k)$. As an example, Figure 5.13 plots the outfield boundaries for Fenway Park in Boston and Yankee Stadium in New York where the batter's location is at home plate in the lower left corner. A shorter distance from home plate to the outfield boundary typically improves the batter's likelihood of a home run for a batted ball hit in the air. In addition, the altitude of the ballpark affects the air density which plays an important role in determining how far a batted ball will carry [2]. The outfield geometry can affect players differently depending on whether they bat right-handed or left-handed since right-handed batters tend to hit most of their home runs to left field while left-handed batters tend to hit most of their home runs to right field.

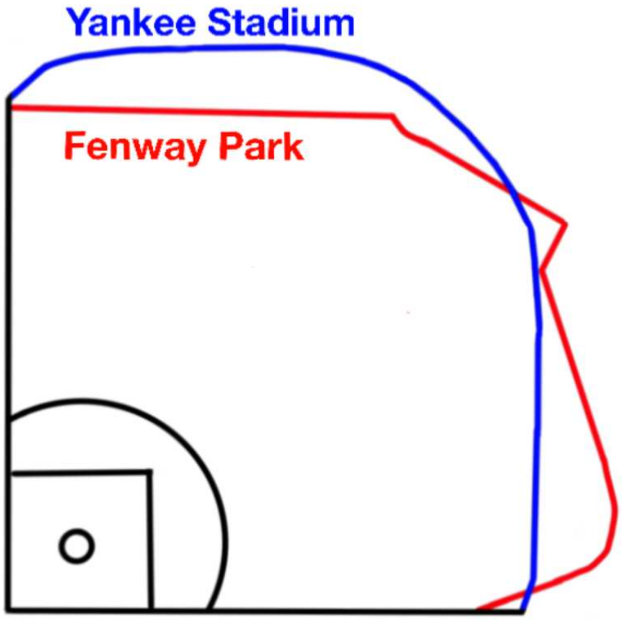

Figure 5.13: Outfield geometry for Fenway Park and Yankee Stadium

We will learn ballpark-dependent batted ball values from 2018 data and use these values to process the 2019 data described in Sec. 5.4.4. The value of batted balls in a subset k will depend on the quality of the fielders that defend against these batted balls. The home team defenders are on the field about half of the time for games played in park p which can cause bias in batted ball values for a given (k, p). Define $R_h(k, p)$ as the average wOBA value for batted balls hit by batters of hand h in subset k and park p with the visiting team on defense in 2018. Let $\overline{R}_h(k)$ be the average wOBA value for batted balls hit by all batters of hand h in subset k in all parks in 2018.

For (h, k, p) groups that correspond to vertical launch angles $v \geq 15°$ and include at least ten batted balls in the calculation of $R_h(k,p)$ we compute the factor

$$F_h(k,p) = \frac{R_h(k,p)}{\overline{R}_h(k)} \tag{5.23}$$

where otherwise $F_h(k,p)$ is set to 1. For a player j of hand h with home park p in 2019 we define

$$\overline{R}(j,k) = 0.5\left[\overline{R}(k) + \overline{R}(k)F_h(k,p)\right] \tag{5.24}$$

where $\overline{R}(k)$ is defined in equation (5.18) and the 0.5 accounts for the fact that a player plays approximately half of his games in the same home ballpark. The $\overline{R}(j,k)$ can be used to improve the accuracy of the prediction in equation (5.17).

To illustrate this process we consider the $b(13,12)$ subregion of the $\mathcal{P}_{5,10}$ partition which is defined by

$$(97.5 \text{ mph} \leq s_l < 102.5 \text{ mph}) \text{ and } (35^\circ \leq v < 45^\circ).$$

For this subregion we have the $\overline{R}(j,k)$ values shown in Table 5.2 which demonstrate that right-handed batters have an advantage in Fenway Park and left-handed batters have an advantage in Yankee Stadium. These observations are consistent with the outfield geometries shown in Fig. 5.13.

Table 5.2: $\overline{R}(j,k)$ for player j of hand h with home park p for $b(13,12)$

Hand (h)	Ballpark (p)	$\overline{R}(j,k)$
Right	Fenway Park	.557
Left	Fenway Park	.381
Right	Yankee Stadium	.378
Left	Yankee Stadium	.490

Let $\widehat{y}_{s1}(j)$ be the prediction of equation (5.17) using $\overline{R}(j,k) = \overline{R}(k)$ and let $\widehat{y}_{s2}(j)$ be the prediction using $\overline{R}(j,k)$ as defined by equation (5.24). As reported in Sec. 5.4.6, $\widehat{y}_{s1}(j)$ produces an SSE of 0.546 for partition $\mathcal{P}_{5,10}$ on the data described in Sec. 5.4.4. The use of $\widehat{y}_{s2}(j)$ reduces the SSE to 0.526.

Table 5.3: Players with largest $\widehat{y}_{s2} - \widehat{y}_{s1}$, 2019

Player	Hand	Home Ballpark	$\widehat{y}_{s2} - \widehat{y}_{s1}$	E_1	E_2
Trevor Story	Right	Coors Field	.020	-.080	-.061
Nolan Arenado	Right	Coors Field	.019	-.071	-.052
Ian Desmond	Right	Coors Field	.018	-.016	.001
Rhys Hoskins	Right	Citizens Bank Park	.017	-.060	-.042
Scott Kingery	Right	Citizens Bank Park	.016	-.029	-.013

Table 5.4: Players with smallest $\widehat{y}_{s2} - \widehat{y}_{s1}$, 2019

Player	Hand	Home Ballpark	$\widehat{y}_{s2} - \widehat{y}_{s1}$	E_1	E_2
Marcell Ozuna	Right	Busch Stadium	-.026	.088	.062
Paul Goldschmidt	Right	Busch Stadium	-.023	-.020	-.043
Paul DeJong	Right	Busch Stadium	-.021	.047	.025
Yadier Molina	Right	Busch Stadium	-.020	.080	.060
Brian Anderson	Right	Marlins Park	-.012	.045	.033

Table 5.3 presents the five players j with the largest differences $\widehat{y}_{s2}(j) - \widehat{y}_{s1}(j)$ and Table 5.4 presents the five players with the smallest differences $\widehat{y}_{s2}(j) - \widehat{y}_{s1}(j)$. Thus, the players in Table 5.3 are expected to benefit from their home ballpark while the players in Table 5.4 are expected to be hindered by their home ballpark. The parks represented in Table 5.3 are known to benefit batters. Coors Field in Denver has an altitude of 5197 feet which enables batted balls to carry longer distances and Citizens Bank Park in Philadelphia has an outfield geometry which is beneficial to right-handed batters. Similarly, both Busch Stadium in St. Louis and Marlins Park in Miami which appear in Table 5.4 have outfield geometries that are detrimental to right-handed batters. The last two columns in each table give the prediction errors $E_1(j) = \widehat{y}_{s1}(j) - y(j)$ and $E_2(j) = \widehat{y}_{s2}(j) - y(j)$ where $y(j)$ is the unobserved performance. $E_1(j)$ is negative for each of the players in Table 5.3 which is consistent with the expectation that these players should benefit from their home ballpark while $E_1(j)$ is

positive for four of the five players in Table 5.4 which is consistent with the expectation that these players should be hindered by their home ballpark. We see that for nine of the ten players in the two tables we have $|E_1| > |E_2|$ so that the use of home park information reduces the prediction error.

5.5 Summary

We have used ball-tracking radar data to show that the predictive value of a batted ball in baseball depends on its speed and vertical launch angle. This constraint enables a batted ball distribution to be estimated from a set of observations using a regression process that adapts to a player's particular collection of batted balls. We showed that these estimated distributions can be used to make improved predictions about unobserved data. The methodology can be adapted to include additional sensor measurements for properties such as spin and horizontal angle as they become available. Since the approach is based on estimating distributions defined over a partition of measurement space, fine-grained contextual adjustments can be included to improve the accuracy of the predictions. The measurement space partitioning process can be used for several applications in baseball including performance forecasting and defensive positioning as well as for a range of other estimation and prediction tasks involving large sets of multidimensional sensor data. In Appendix E we apply this method to 2021 MLB batted ball data.

Chapter 6

Conclusion

The ability to quantify player skill and team performance in sports has been revolutionized by the deployment of sensors that collect large amounts of data for athletic events including baseball [22], basketball [66], football [8], and golf [62]. This has led to the use of machine learning algorithms by teams to exploit this data to gain a competitive advantage. The assessment of player skills in baseball is increasingly dependent on data-driven models rather than subjective evaluation. The accuracy of these models is critical to a team's success as executives attempt to maximize performance while abiding by organizational financial constraints. While there are large disparities in the financial resources available to teams, the use of data-driven models has enabled small market franchises to compete successfully against their more affluent opponents [53].

We have applied learning methods to sensor data acquired at baseball games to develop new techniques for comparing players, optimizing pitch distributions, and quantifying batted ball talent. These techniques can be used for a number of applications in the areas of strategy [60], player development [38], and player evaluation [53] in baseball. By utilizing physical measurements, the methods have the advantage of allowing the direct comparison of players across environments. This enables, for example, predictions about how a college pitcher would perform in major league baseball after optimizing his pitch distribution. The use of physical models also enables isolation of the effect of contextual variables so that we can predict how a batter might perform in a different home ballpark. The framework can also be applied outside of the baseball domain. We could, for example, use a similar approach to build a model for the dependence of a football team's performance on the distribution of offensive play types, e.g. run or pass, that are used [43]. This model could then be utilized to determine the play distribution that a given offense should use to maximize success.

Bibliography

[1] D. Appelman. Pitch type linear weights. www.fangraphs.com/blogs/ pitch-type-linear-weights (May 20, 2009).

[2] A. Bahill, D. Baldwin, and J. Ramberg. Effects of altitude and atmospheric conditions on the flight of a baseball. *International Journal of Sports Science and Engineering*, 3(2):109–128, 2009.

[3] A. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis.* Clarendon Press, Oxford, 1997.

[4] J. Brosnan, A. McNitt, and T. Serensits. Effects of varying surface characteristics on the hardness and traction of baseball field playing surfaces. *International Turfgrass Society Research Journal*, 11:1053–1065, 2009.

[5] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, October 1910.

[6] R. Carleton. Should I worry about my favorite pitcher? www.baseball-prospectus.com/article.php?articleid=20516 (May 9, 2013).

[7] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, Long Beach, CA, 2019.

[8] K. Clark. The NFL's analytics revolution has arrived. www.theringer .com/nfl/2018/12/19/18148153/nfl-analytics-revolution (Dec. 19, 2018).

[9] L. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[10] N. Draper and H. Smith. *Applied Regression Analysis.* Wiley, New York, 3rd edition, 1998.

[11] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2001.

[12] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[13] A. Fagerstrom. (June 24, 2016). Trevor Bauer looks like a completely different pitcher [Online]. Available: www.fangraphs.com/blogs/trevor-bauer-looks-like-a-completely-different-pitcher.

[14] M. Fast. What the heck is PITCHf/x? In J. Distelheim, B. Tsao, J. Oshan, C. Bolado, and B. Jacobs, editors, *The Hardball Times Baseball Annual, 2010*, pages 153–158. The Hardball Times, 2010.

[15] Fielding Independent Pitching (FIP). www.fangraphs.com/library /pitching/fip/.

[16] R. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[17] D. Gartland. MLB outfield walls, ranked. Available: si.com/mlb/2021/03/24/mlb-outfield-walls-ranked-fenway-park-yankee-stadium (March 24, 2021).

[18] V. Gennaro. The Big Data approach to baseball analytics. In *SABR Analytics Conference*, Phoenix, AZ, March 2013.

[19] R. Gray. Behavior of college baseball players in a virtual batting task. *Journal of Experimental Psychology: Human perception and performance*, 28(5):1131–1148, 2002.

[20] G. Healey. The intrinsic value of a pitch. In *SABR Analytics Conference*, Phoenix, AZ, March 2017.

[21] G. Healey. Learning, visualizing, and assessing a model for the intrinsic value of a batted ball. *IEEE Access*, 5:13811–13822, 2017.

[22] G. Healey. The new moneyball: how ballpark sensors are changing baseball. *Proceedings of the IEEE*, 105(11):1999–2002, 2017.

[23] G. Healey. A Bayesian method for computing intrinsic pitch values using kernel density and nonparametric regression estimates. *Journal of Quantitative Analysis in Sports*, 15(1):59–74, March 2019.

[24] G. Healey. Combining radar and optical sensor data to measure player value in baseball. *Sensors*, 21(1):64, 2021.

[25] G. Healey and S. Zhao. Using PITCHf/x to model the dependence of strikeout rate on the predictability of pitch sequences. *Journal of Sports Analytics*, 2017.

[26] G. Healey, S. Zhao, and D. Brooks. Measuring pitcher similarity. www.baseball.prospectus.com/article.php?articleid=32199 (July 10, 2017).

[27] F. Hillier and G. Liberman. *Introduction to Mathematical Programming.* McGraw-Hill, 1990.

[28] F. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics*, 20:224–230, 1941.

[29] W. James and C. Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.

[30] J. Kalk. (Feb. 12, 2008). Pitcher similarity scores [Online]. Available: www.hardballtimes.com/pitcher-similarity-scores.

[31] J. Kalk. (Feb. 19, 2008). Pitcher similarity scores (part 2) [Online]. Available: www.hardballtimes.com/pitcher-similarity-scores-part-ii.

[32] J. Keri. (Mar. 4, 2014). Q&A: MLB Advanced Media's Bob Bowman discusses revolutionary new play-tracking system [Online]. Available: grantland.com/the-triangle/mlb-advanced-media-play-tracking-bob-bowman-interview.

[33] M. Khalily, T. Brown, and R. Tafazolli. Machine-learning based approach for diffraction loss variation prediction by the human body. *IEEE Antennas and Wireless Propagation Letters*, 18(11):2301–2305, November 2019.

[34] J. Kloke and J. McKean. *Nonparametric statistical methods using R.* Chapman and Hall/CRC, New York, 2014.

[35] S. Kolouri, S. Park, M. Thorpe, D. Slepcev, and G. Rohde. Optimal mass transport: Signal processing and machine learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[36] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[37] J. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.

[38] B. Lindbergh and T. Sawchik. *The MVP machine: How baseball's new nonconformists are using data to build better players.* Basic Books, New York, NY, 2019.

[39] S. Loftus. (Apr. 15, 2013). Pitcher similarity scores [Online]. Available: www.beyondtheboxscore.com/2013/4/15/4208426/pitcher-similarity-scores.

[40] S. Loftus. (Apr. 25, 2013). Testing and visualizing similarity scores [Online]. Available: www.beyondtheboxscore.com/2013/4/25/4260554/testing-and-visualizing-similarity-scores.

[41] S. Loftus. (Nov. 25, 2013). Pitcher similarity scores 2.0 [Online]. Available: www.beyondtheboxscore.com/2013/11/25/5133702/pitcher-similarity-scores-ervin-santana-sabermetrics.

[42] J. Long, J. Judge, and H. Pavlidis. Introducing pitch tunnels. www.baseball.prospectus.com/article.php?articleid=31030 (Jan. 24, 2017).

[43] E. McGough, C. Clemons, M. Ferrara, T. Norfolk, and G. Young. A game-theoretic approach to personnel decisions in american football. *Journal of Quantitative Analysis in Sports*, 6(4):1–15, October 2010.

[44] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, Long Beach, CA, 2019.

[45] E. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of probability and its applications*, 10(1):186–190, 1965.

[46] A. Nathan. Determining pitch movement from PITCHf/x data. baseball.physics.illinois.edu/Movement.pdf (Oct. 21, 2012).

[47] N. Paine. Game theory says R.A. Dickey should throw more knuckleballs. fivethirtyeight.com/features/game-theory-says-r-a-dickey-should-throw-more-knuckleballs (Aug. 13, 2015).

[48] L. Panas. *Beyond Batting Average.* Lulu Press, Morrisville, North Carolina, 2010.

[49] S. Powers. Toward a probability distribution over batted-ball trajectories. www.fangraphs.com/tht/toward-a-probability-distribution-over-batted-ball-trajectories (Aug. 19, 2016).

[50] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. *Int. J. Comp. Vision*, 40(2):99–121, 2000.

[51] E. Sarris. (June 9, 2016). James Paxton's new angle on life [Online]. Available: www.fangraphs.com/blogs/james-paxtons-new-angle-on-life.

[52] E. Sarris. (March 14, 2017). Yonder Alonso has changed his mind [Online]. Available: www.fangraphs.com/blogs/yonder-alonso-has-changed-his-mind.

[53] T. Sawchik. *Big data baseball.* Flatiron Books, New York, NY, 2016.

[54] S. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004.

[55] N. Silver. Why was Kevin Maas a bust? In J. Keri, editor, *Baseball between the numbers*, pages 253–271. Basic Books, New York, 2006.

[56] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, October 1910.

[57] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–208, 1956.

[58] J. Sullivan. Finding baseball's most improved hitter. www.fangraphs.com/blogs/finding-baseballs-most-improved-hitter (Sept. 21, 2017).

[59] J. Sullivan. Now Kelvin Herrera is almost impossible. www.fangraphs.com/blogs/now-kelvin-herrera-is-almost-impossible (April 13, 2016).

[60] T. Tango, M. Lichtman, and A. Dolphin. *The Book: Playing the Percentages in Baseball.* Potomac Books, Dulles, Virgina, 2007.

[61] S. Urbanek and Y. Rubner. Package ‘emdist’. Technical report, CRAN, February 19, 2015.

[62] S. Wang, Y. Xu, Y. Zheng, M. Zhu, H. Yao, and Z. Xiao. Tracking a golf ball with high-speed stereo vision system. *IEEE Transactions on Instrumentation and Measurement*, 68(8):2742–2754, August 2019.

[63] W. Wang, W. Han, X. Na, J. Gong, and J. Xi. A probabilistic approach to measuring driving behavior similarity with driving primitives. *IEEE Transactions on Intelligent Vehicles*, 5(1):127–138, March 2020.

[64] G. Watson. Smooth regression analysis. *Sankhyā: The Indian journal of statistics, series A*, 26(4):359–372, 1964.

[65] wOBA and FIP constants [Online]. Available: www.fangraphs.com/guts.aspx?type=cn.

[66] M. Woo. Artificial intelligence in NBA basketball. www.insidescience.org/news/artificial-intelligence-nba-basketball (Dec. 21, 2018).

[67] R. Zeller and E. Carmines. *Measurement in the Social Sciences: The Link Between Theory and Data.* Cambridge University Press, 1980.

Appendix A

Most Similar Match Tables

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Tim Adleman	Aaron Blair	0.7218
Cody Allen	Bud Norris	0.8042
Chase Anderson	James Shields	0.9830
Matt Andriese	Tom Koehler	0.7128
Chris Archer	Bud Norris	0.6269
Jake Arrieta	Anthony DeSclafani	0.7607
John Axford	Pedro Baez	0.7929
Pedro Baez	Carlos Estevez	0.6584
Matt Barnes	Joseph Biagini	0.8833
Kyle Barraclough	Luis Severino	0.7116
Trevor Bauer	Joseph Biagini	0.7694
Jose Berrios	Jacob deGrom	0.7022
Dellin Betances	Kyle Barraclough	0.9871
Chad Bettis	Jacob deGrom	0.6492
Joseph Biagini	Matt Harvey	0.6086
Aaron Blair	Tim Adleman	0.7218
Joe Blanton	Jason Hammel	0.7675
Matthew Bowman	Kendall Graveman	0.7262
Blaine Boyer	Eddie Butler	0.7772
Brad Brach	Stephen Strasburg	0.8760
Archie Bradley	Tom Koehler	0.6737
Clay Buchholz	Colin Rea	0.7716

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Eddie Butler	Anthony DeSclafani	0.5195
Trevor Cahill	Edinson Volquez	0.7771
Matt Cain	Ryan Dull	0.7166
Arquimedes Caminero	Hunter Strickland	0.7853
Carlos Carrasco	Michael Tonkin	0.6712
Andrew Cashner	Eddie Butler	0.5688
Luis Cessa	Brandon Maurer	0.6567
Jhoulys Chacin	Matt Wisler	0.7896
Tyler Chatwood	Andrew Cashner	0.7007
Jesse Chavez	Taijuan Walker	0.6910
Steve Cishek	Aaron Nola	1.3934
Paul Clemens	Michael Tonkin	1.0845
Tyler Clippard	Chase Anderson	1.2232
Gerrit Cole	Hansel Robles	0.6407
Bartolo Colon	Matt Shoemaker	0.9140
Jarred Cosart	Mark Melancon	0.9734
Nathan Eovaldi	Blake Wood	0.8319
Carlos Estevez	Pedro Baez	0.6584
Marco Estrada	Tyler Clippard	1.3896
Jeurys Familia	Luis Perdomo	0.7828
Scott Feldman	Adam Wainwright	0.6295
Michael Feliz	Roberto Osuna	0.7367
Jose Fernandez	Stephen Strasburg	0.7496
Mike Fiers	Ross Stripling	1.0287
Doug Fister	Zach Davies	0.9606
Mike Foltynewicz	Mike Wright	0.6340
Michael Fulmer	Jacob deGrom	0.7391
Yovani Gallardo	Anibal Sanchez	0.7852
Matt Garza	Colin Rea	0.6525
Kevin Gausman	Hansel Robles	0.7074
Dillon Gee	Rick Porcello	0.5729
Kyle Gibson	Johnny Cueto	0.6648
Ken Giles	Michael Pineda	0.7047
Mychal Givens	Hunter Strickland	0.8286
Zachary Godley	Jameson Taillon	0.8506
Jeanmar Gomez	Mike Pelfrey	0.6568
Miguel Gonzalez	James Shields	0.6537
Kendall Graveman	Chad Kuhl	0.7052
Jon Gray	Shelby Miller	0.5269

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Sonny Gray	Tyler Chatwood	0.7848
Zack Greinke	Matt Andriese	0.8323
A.J. Griffin	Colby Lewis	1.0014
Jason Grilli	David Hernandez	0.5254
Junior Guerra	Edwin Jackson	0.6731
Jason Hammel	Anthony DeSclafani	0.6111
Will Harris	David Robertson	0.6966
Matt Harvey	Shelby Miller	0.4067
Jeremy Hellickson	Zach Davies	0.9729
Kyle Hendricks	Rick Porcello	0.9840
Liam Hendriks	Tyler Chatwood	0.7218
David Hernandez	Jason Grilli	0.5254
Felix Hernandez	Tyler Duffey	0.9634
Kelvin Herrera	Carlos Estevez	0.8335
Daniel Hudson	Hansel Robles	0.8475
Raisel Iglesias	Blaine Boyer	0.8108
Hisashi Iwakuma	Dillon Gee	0.8792
Edwin Jackson	Junior Guerra	0.6731
Ubaldo Jimenez	Kyle Gibson	0.7786
Jim Johnson	Luis Perdomo	0.6994
Nate Karns	Archie Bradley	0.8307
Ian Kennedy	Justin Verlander	0.6537
Corey Kluber	Anthony DeSclafani	0.7820
Tom Koehler	Archie Bradley	0.6737
Chad Kuhl	Eddie Butler	0.6009
John Lackey	Jason Hammel	0.7547
Mat Latos	Anibal Sanchez	0.7750
Mike Leake	Kendall Graveman	0.9115
Colby Lewis	Carlos Villanueva	0.6744
Kenta Maeda	Jerad Eickhoff	0.7885
Carlos Martinez	Jeff Samardzija	1.0176
Brandon Maurer	Hansel Robles	0.6522
Lance McCullers	Mychal Givens	1.3648
Dustin McGowan	Blake Wood	0.8407
Collin McHugh	Adam Wainwright	0.7693
Mark Melancon	Jarred Cosart	0.9734
Daniel Mengden	Adam Warren	0.7971
Shelby Miller	Matt Harvey	0.4067
Jimmy Nelson	Marcus Stroman	0.8117

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Hector Neris	Carlos Carrasco	1.1101
Juan Nicasio	Shelby Miller	0.4577
Aaron Nola	Mike Leake	1.1660
Ricky Nolasco	Rick Porcello	0.8955
Bud Norris	Chris Archer	0.6269
Ivan Nova	Luis Perdomo	0.6706
Jake Odorizzi	Fernando Salas	0.6412
Seung-hwan Oh	Ryan Dull	0.4963
Ross Ohlendorf	Randall Delgado	0.5353
Roberto Osuna	Michael Feliz	0.7367
Jake Peavy	Ryan Vogelsong	0.8515
Mike Pelfrey	Jeanmar Gomez	0.6568
Wily Peralta	Mike Foltynewicz	0.6618
Luis Perdomo	Ivan Nova	0.6706
David Phelps	Shelby Miller	0.7011
Michael Pineda	Luis Cessa	0.6873
Rick Porcello	Dillon Gee	0.5729
Ryan Pressly	Chris Archer	0.7838
Kevin Quackenbush	Ross Stripling	0.8533
J.C. Ramirez	Hunter Strickland	0.7793
Erasmo Ramirez	Joel De La Cruz	0.7726
A.J. Ramos	Zack Greinke	0.8865
Colin Rea	Justin Verlander	0.6452
Addison Reed	Liam Hendriks	0.7261
Tanner Roark	Kyle Gibson	0.6736
David Robertson	Will Harris	0.6966
Hansel Robles	Gerrit Cole	0.6407
Fernando Rodney	Alfredo Simon	1.2610
Joe Ross	Matthew Bowman	0.7748
Fernando Salas	Jake Odorizzi	0.6412
Danny Salazar	Andrew Cashner	0.6537
Jeff Samardzija	Mike Foltynewicz	0.7275
Aaron Sanchez	Jim Johnson	0.7179
Anibal Sanchez	Fernando Salas	0.7685
Ervin Santana	Alex Wilson	0.6934
Max Scherzer	Chad Kuhl	0.6324
Luis Severino	Chris Archer	0.6592
Bryan Shaw	Carlos Torres	1.1258
James Shields	Miguel Gonzalez	0.6537

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Braden Shipley	Ian Kennedy	0.8771
Matt Shoemaker	Kyle Gibson	0.7481
Alfredo Simon	Tanner Roark	0.7760
Josh Smith	Matt Cain	0.7414
Joakim Soria	Taijuan Walker	0.7564
Dan Straily	Colby Lewis	0.7449
Stephen Strasburg	Matt Harvey	0.6034
Hunter Strickland	J.C. Ramirez	0.7793
Ross Stripling	Chris Tillman	0.8369
Marcus Stroman	Jimmy Nelson	0.8117
Albert Suarez	Aaron Blair	0.7603
Noah Syndergaard	Luis Severino	0.7567
Jameson Taillon	Zachary Godley	0.8506
Masahiro Tanaka	Dillon Gee	0.6355
Julio Teheran	Aaron Blair	0.7555
Tyler Thornburg	Archie Bradley	0.8752
Chris Tillman	Ross Stripling	0.8369
Josh Tomlin	Yovani Gallardo	0.8051
Michael Tonkin	Chad Kuhl	0.6546
Carlos Torres	Edwin Jackson	0.8924
Nick Tropeano	Randall Delgado	0.9390
Jose Urena	Jeff Samardzija	0.7994
Vincent Velasquez	Taijuan Walker	0.5854
Yordano Ventura	Jameson Taillon	0.9287
Justin Verlander	Colin Rea	0.6452
Logan Verrett	Ryan Vogelsong	0.7547
Carlos Villanueva	Colby Lewis	0.6744
Ryan Vogelsong	Logan Verrett	0.7547
Edinson Volquez	Jim Johnson	0.7348
Michael Wacha	Jake Odorizzi	0.8036
Adam Wainwright	Scott Feldman	0.6295
Taijuan Walker	Shelby Miller	0.5741
Adam Warren	Daniel Mengden	0.7971
Jered Weaver	A.J. Griffin	1.7653
Tyler Wilson	Ervin Santana	0.9275
Alex Wilson	Jordan Zimmermann	0.6897
Matt Wisler	David Hernandez	0.7196
Blake Wood	Nathan Eovaldi	0.8319
Vance Worley	Miguel Gonzalez	0.6543

Table A.1: Most similar match for each right-handed pitcher, 2016

Pitcher	Most Similar	Distance
Mike Wright	Mike Foltynewicz	0.6340
Steven Wright	R.A. Dickey	0.6293
Chris Young	Colby Lewis	1.4429
Brad Ziegler	Jeanmar Gomez	2.8651
Jordan Zimmermann	Ryan Dull	0.6001

Table A.2: Most similar match for each left-handed pitcher, 2016

Pitcher	Most Similar	Distance
Tyler Anderson	Scott Kazmir	0.7733
Antonio Bastardo	Travis Wood	0.7434
Matt Boyd	Adam Morgan	0.6427
Zach Britton	James Paxton	1.7251
Ryan Buchter	Travis Wood	0.7584
Madison Bumgarner	Jon Lester	0.6813
Wei-Yin Chen	Scott Kazmir	0.7659
Tony Cingrani	Robbie Ray	0.8608
Adam Conley	Justin Nicolino	0.7754
Patrick Corbin	Robbie Ray	0.7337
Pat Dean	Justin Nicolino	0.6367
Danny Duffy	Brandon Finnegan	0.8069
Zach Duke	Chris Sale	1.4223
Brandon Finnegan	Jose Quintana	0.6112
Christian Friedrich	Wade Miley	0.7675
Jaime Garcia	Martin Perez	0.8442
Gio Gonzalez	Mike Montgomery	0.8350
Cole Hamels	David Price	0.7686
Brad Hand	Dan Jennings	0.9119
J.A. Happ	Wei-Yin Chen	0.7908
Rich Hill	Christian Friedrich	1.4946
Derek Holland	Jeff Locke	0.7053
Dan Jennings	Brad Hand	0.9119
Scott Kazmir	Hector Santiago	0.6513
Clayton Kershaw	Travis Wood	1.4912
Dallas Keuchel	CC Sabathia	0.8611
John Lamb	Matt Boyd	0.9782
Jon Lester	Madison Bumgarner	0.6813

Table A.2: Most similar match for each left-handed pitcher, 2016

Pitcher	Most Similar	Distance
Francisco Liriano	Martin Perez	0.6337
Jeff Locke	Derek Holland	0.7053
Sean Manaea	Tony Watson	0.8821
Steven Matz	Mike Montgomery	0.7764
Wade Miley	Christian Friedrich	0.7675
Andrew Miller	Dan Jennings	1.3264
Tommy Milone	Drew Smyly	1.1309
Mike Montgomery	Steven Matz	0.7764
Matt Moore	Mike Montgomery	0.8197
Adam Morgan	Matt Boyd	0.6427
Justin Nicolino	Pat Dean	0.6367
Jon Niese	Chris Rusin	0.5540
Daniel Norris	Danny Duffy	0.9015
Brett Oberholtzer	Hector Santiago	0.6770
James Paxton	Felipe Rivero	0.9400
Martin Perez	Francisco Liriano	0.6337
Drew Pomeranz	Christian Friedrich	1.2464
David Price	Justin Nicolino	0.7498
Jose Quintana	Brandon Finnegan	0.6112
Robbie Ray	Eduardo Rodriguez	0.6191
Clayton Richard	Tony Watson	1.0658
Felipe Rivero	James Paxton	0.9400
Carlos Rodon	Eduardo Rodriguez	0.7011
Eduardo Rodriguez	Robbie Ray	0.6191
Jorge De La Rosa	Chris Rusin	0.7605
Chris Rusin	Jon Niese	0.5540
CC Sabathia	Jon Niese	0.8586
Chris Sale	Sean Manaea	0.9701
Hector Santiago	Scott Kazmir	0.6513
Kevin Siegrist	Hector Santiago	0.8364
Drew Smyly	Antonio Bastardo	0.9342
Blake Snell	Daniel Norris	0.9331
Julio Urias	Blake Snell	0.9960
Tony Watson	Sean Manaea	0.8821
Travis Wood	Antonio Bastardo	0.7434

Appendix B

Pitcher Characteristics Examined by the Pitcher Similarity Measure

B.1 Pitchers with Small Year-to-Year Variation

We can use the similarity measure defined in Section 3.1.3 to compare pitchers to themselves over time. For this purpose we computed the similarity measure between 2015 and 2016 for each pitcher who threw at least 1000 pitches in each regular season. The covariance matrix Σ for each platoon configuration in Equation (2.2) and the fractions $f_{RR}, f_{RL}, f_{LR}, f_{LL}$ in Equations (3.2) and (3.3) were computed using the combined data from both seasons.

Tables B.1 and B.2 list the right-handed and left-handed pitchers who changed the least between 2015 and 2016 along with their age on 30 June 2016. Many of the smallest changers are veterans with 13 of the 20 pitchers in the tables being at least 30 years old at midseason 2016 and with all pitchers except Carlos Rodon being at least 26. Two of the smallest changers are the knuckleballers R.A. Dickey and Steven Wright. Unsurprisingly, Bartolo Colon is also one of the least-changing right-handers.

Table B.1: Right-handed pitchers who changed the least between 2015 and 2016

Pitcher	Distance	Age
R.A. Dickey	0.1280	41
Fernando Salas	0.2584	31
Steven Wright	0.2654	31
Bartolo Colon	0.2801	43
Arquimedes Caminero	0.2881	29
Corey Kluber	0.2995	30
Adam Warren	0.3040	28
Jered Weaver	0.3062	33
Max Scherzer	0.3107	31
Scott Feldman	0.3215	33

Table B.2: Left-handed pitchers who changed the least between 2015 and 2016

Pitcher	Distance	Age
Jon Lester	0.2581	32
Carlos Rodon	0.3056	23
Jorge De La Rosa	0.3357	35
Francisco Liriano	0.3572	32
Drew Smyly	0.3922	27
Adam Conley	0.3963	26
Patrick Corbin	0.4007	26
Tony Watson	0.4147	31
Gio Gonzalez	0.4150	30
Chris Rusin	0.4169	29

B.2 Pitchers with Large Year-to-Year Variation

Tables B.3 and B.4 list the right-handed and left-handed pitchers who changed the most between 2015 and 2016 along with their age on 30 June 2016 and their ERA for the two seasons. We see that these pitchers are younger than their more stable counterparts with only 3 of the 20 pitchers being at least 30 years old at midseason 2016. Six of the ten right-handers in Table B.3 and eight of the ten left-handers in Table B.4 improved their ERA from 2015 to 2016. Several of the pitchers in these tables (Phelps, Chavez, Montgomery, Hand, Pomeranz) changed from starting in 2015 to relieving in 2016. Others near the top of the lists include Trevor Bauer and Kelvin Herrera who made significant changes to their pitch mix [13] [59] and James Paxton who made a significant change to his pitching mechanics [51].

Table B.3: Right-handed pitchers who changed the most between 2015 and 2016

Pitcher	Distance	Age	2015 ERA	2016 ERA
David Phelps	1.1081	29	4.50	2.28
Trevor Bauer	0.9869	25	4.55	4.26
Kelvin Herrera	0.9639	26	2.71	2.75
Jesse Chavez	0.9227	32	4.18	4.43
Matt Shoemaker	0.9156	29	4.46	3.88
Joe Blanton	0.9063	35	2.84	2.48
Will Harris	0.8785	31	1.90	2.25
Lance McCullers	0.8329	22	3.22	3.22
Noah Syndergaard	0.8240	23	3.24	2.60
Aaron Nola	0.8150	23	3.59	4.78

Table B.4: Left-handed pitchers who changed the most between 2015 and 2016

Pitcher	Distance	Age	2015 ERA	2016 ERA
James Paxton	1.4217	27	3.90	3.79
Mike Montgomery	1.0952	26	4.60	2.52
Brad Hand	1.0056	26	5.30	2.92
Matt Boyd	0.9570	25	7.53	4.53
Adam Morgan	0.9151	26	4.48	6.04
Daniel Norris	0.8312	23	3.75	3.38
Drew Pomeranz	0.8008	27	3.66	3.32
Danny Duffy	0.7765	27	4.08	3.51
Jeff Locke	0.7258	28	4.49	5.44
Chris Sale	0.6737	27	3.41	3.34

B.3 Pitchers with Small Platoon Distances

We can use the EMD to measure the difference between a pitcher's distribution of pitches against right-handed and left-handed batters. We considered all pitchers who threw at least 1000 pitches during the 2016 regular season. For this computation, the covariance matrix Σ for the ground distance in Equation (2.2) was generated for right-handed pitchers using all clusters of pitches thrown by right-handers to either right-handed or left-handed batters. Similarly, a covariance matrix Σ was computed for left-handed pitchers using all clusters of pitches thrown by left-handers to either right-handed or left-handed batters.

Tables B.5 and B.6 list the right-handed and left-handed pitchers with the smallest platoon distance along with each pitcher's 2016 wOBA allowed to right-handed and left-handed batters. A number of these pitchers relied heavily on a single pitch type. Reed ($w = 72.2\%$), Allen ($w = 63.3\%$), and Conley ($w = 65.5\%$) all threw a large fraction of four-seam fastballs in 2016. Dickey ($w = 87.6\%$) and Wright ($w = 83.1\%$) each threw a large fraction of knuckleballs while Harris ($w = 66.4\%$ cutter), Britton ($w = 92.0\%$ sinker), and Miller ($w = 60.7\%$ slider) also threw a large fraction of a single pitch type in 2016.

Throwing a similar distribution of pitches to right-handed and left-handed batters is a characteristic of a pitcher's approach, but is not necessarily indicative of his platoon results. While several of the pitchers (Reed, McCullers, Dickey, Happ) in Tables B.5 and B.6 had a very small wOBA platoon split, others (Young, DeSclafani) had large wOBA platoon splits.

B.4 Pitchers with Large Platoon Distances

Tables B.7 and B.8 list the right-handed and left-handed pitchers with the largest platoon distances in 2016 along with each pitcher's 2016 wOBA allowed to right-handed and left-

Table B.5: Right-handed pitchers with the smallest platoon distances in 2016

Pitcher	Distance	wOBA vs. R	wOBA vs. L
Addison Reed	0.0781	.229	.228
Cody Allen	0.0970	.222	.292
Will Harris	0.1592	.263	.229
Lance McCullers	0.1780	.324	.327
Chris Young	0.2242	.320	.476
Adam Warren	0.2338	.343	.258
Vance Worley	0.2352	.318	.333
R.A. Dickey	0.2400	.337	.339
Anthony DeSclafani	0.2486	.260	.353
Steven Wright	0.2517	.303	.271

Table B.6: Left-handed pitchers with the smallest platoon distances in 2016

Pitcher	Distance	wOBA vs. R	wOBA vs L
Adam Conley	0.2157	.316	.334
Dan Jennings	0.2538	.310	.290
Pat Dean	0.2632	.395	.356
J.A. Happ	0.2781	.292	.287
Madison Bumgarner	0.2919	.279	.223
Drew Smyly	0.3004	.328	.305
Steven Matz	0.3076	.296	.307
Tyler Anderson	0.3166	.333	.270
Zach Britton	0.3253	.180	.226
Andrew Miller	0.3339	.207	.220

handed batters. We see that by using very different distributions of pitches to right-handed and left-handed batters, several of these pitchers (Weaver, Milone, Pomeranz) had very small wOBA platoon splits while others (Iglesias, McGowan, Duffy) had large wOBA platoon splits.

None of the right-handers and only two of the left-handers (Rivero and Siegrist) in Tables B.7 and B.8 threw a single pitch type at least 60% of the time in 2016. Seven of the right-handers in Table B.7 (Ziegler, Weaver, Iglesias, McGowan, Herrera, Ramos, Chacin) contributed to their platoon variation by throwing a significantly higher fraction of sliders to right-handed batters and a significantly higher fraction of changeups to left-handed batters. For the purposes of this analysis, significantly refers to a fraction that is at least 0.10 higher. Similarly, four of the left-handers in Table B.8 (Rivero, Watson, Manaea, Corbin) threw a significantly higher fraction of sliders to left-handed batters and a significantly higher fraction of changeups to right-handed batters.

Another popular strategy which was used by six of the pitchers in Tables B.7 and B.8 (Weaver, McGowan, Hand, Duffy, Siegrist, Corbin) was to throw a significantly higher fraction of four-seam fastballs to same-side batters and a significantly higher fraction of sinkers to opposite-side batters. Right-hander Kyle Hendricks employed the opposite approach by throwing a significantly higher fraction of sinkers to right-handed batters and a significantly higher fraction of four-seam fastballs to left-handed batters. Left-handers Milone and Hill enhanced their platoon variation by throwing a significantly higher fraction of curveballs to left-handed batters.

Table B.7: Right-handed pitchers with the largest platoon distances in 2016

Pitcher	Distance	wOBA vs. R	wOBA vs. L
Brad Ziegler	1.8874	.278	.306
Jered Weaver	1.1993	.365	.365
Raisel Iglesias	1.0970	.224	.332
Dustin McGowan	1.0896	.212	.375
Kelvin Herrera	1.0802	.268	.246
Kyle Hendricks	0.9924	.243	.269
Matt Wisler	0.9723	.313	.334
A.J. Ramos	0.9458	.287	.262
Jhoulys Chacin	0.9152	.317	.327
Alfredo Simon	0.8719	.412	.454

Table B.8: Left-handed pitchers with the largest platoon distances in 2016

Pitcher	Distance	wOBA vs. R	wOBA vs. L
Brad Hand	1.1295	.297	.194
Felipe Rivero	1.0366	.272	.343
Tony Watson	0.9595	.302	.253
Tommy Milone	0.9032	.362	.357
Sean Manaea	0.8817	.322	.231
Danny Duffy	0.8480	.325	.201
Kevin Siegrist	0.8313	.269	.302
Rich Hill	0.8279	.244	.232
Patrick Corbin	0.7693	.363	.324
Drew Pomeranz	0.7610	.287	.284

B.5 Pitchers with Small Changes after Two Strikes

We can use the similarity measure defined by Equations (3.2) and (3.3) to measure how much a pitcher changes his distribution of pitches as the count changes. For each pitcher who threw at least 1000 pitches in 2016, we computed the distance described in Section 3.1.3 between the pitcher's distributions of pitches thrown before two strikes and his distributions of pitches thrown after two strikes. For each platoon configuration, the covariance matrix for the ground distance in Equation (2.2) was generated using all pitch clusters associated with pitches thrown before two strikes and all pitch clusters thrown after two strikes.

Tables B.9 and B.10 list the right-handed and left-handed pitchers who changed the least after reaching two strikes in 2016. The two right-handers who changed the least (Grilli 62.4% four-seam, Reed 72.2% four-seam) and the two left-handers who changed the least (Britton 92.0% sinker, Buchter 84.7% four-seam) each threw a large fraction of a single pitch type in 2016. In addition, several of the other pitchers in the two tables (Wright 83.1% knuckler, Quackenbush 63.2% four-seam, Oh 60.6% four-seam, Cingrani 87.4% four-seam, Bastardo 65.5% four-seam) each threw over 60% of a single pitch type in 2016.

Table B.9: Right-handed pitchers who changed the least with two strikes in 2016

Pitcher	Distance
Jason Grilli	0.2231
Addison Reed	0.2744
Chris Young	0.2877
Jered Weaver	0.2944
Fernando Salas	0.3022
Alex Wilson	0.3100
Steven Wright	0.3101
Kevin Quackenbush	0.3113
Seung-hwan Oh	0.3137
Jesse Chavez	0.3192

Table B.10: Left-handed pitchers who changed the least with two strikes in 2016

Pitcher	Distance
Zach Britton	0.2108
Ryan Buchter	0.2379
Brett Oberholtzer	0.3354
Tony Cingrani	0.3404
Chris Rusin	0.3549
Tyler Anderson	0.3730
Jeff Locke	0.3734
Eduardo Rodriguez	0.3756
Antonio Bastardo	0.3857
Steven Matz	0.4018

B.6 Pitchers with Large Changes after Two Strikes

Tables B.11 and B.12 list the right-handed and left-handed pitchers who changed the most after reaching two strikes in 2016. Each of these pitchers threw a significantly higher fraction of a particular breaking ball with two strikes. The pitch with the largest increase in frequency after two strikes over all batters faced is referred to as the Delta Pitch in the two tables. The Δw column in Tables B.11 and B.12 indicates how much more frequently a pitcher threw the Delta Pitch after two strikes as compared to before two strikes. Brad Ziegler, for example, threw his slider 10.16% of the time before two strikes and 40.45% of the time after two strikes for a Δw of 0.4045 - 0.1016 = 0.3029.

Among the pitchers in Tables B.11 and B.12 with smaller values of Δw for their Delta Pitch, Fiers (6 pitch types) and Darvish (7 pitch types) had a large set of possible pitch types with which to adjust frequencies and left-handers Kershaw and Snell used a higher fraction of sliders with two strikes in addition to a higher fraction of their Delta Pitch curve balls.

Table B.11: Right-handed pitchers who changed the most with two strikes in 2016

Pitcher	Distance	Delta Pitch	Δw
Brad Ziegler	2.4306	slider	0.3029
Dellin Betances	1.4501	curve	0.2086
Paul Clemens	1.3814	curve	0.2617
Carlos Martinez	1.2009	slider	0.2565
Jerad Eickhoff	1.1797	curve	0.2998
Mike Fiers	1.0923	curve	0.1828
Lance McCullers	1.0913	curve	0.3183
Raisel Iglesias	1.0753	slider	0.2601
Yu Darvish	1.0514	slider	0.1219
Aaron Nola	1.0365	curve	0.2094

Table B.12: Left-handed pitchers who changed the most with two strikes in 2016

Pitcher	Distance	Delta Pitch	Δw
Zach Duke	1.3792	curve	0.2997
Clayton Kershaw	1.3174	curve	0.1731
Jaime Garcia	1.1124	slider	0.3410
Brad Hand	1.0659	slider	0.2561
Carlos Rodon	1.0431	slider	0.2430
Chris Sale	1.0141	slider	0.2279
Patrick Corbin	0.9607	slider	0.3372
Gio Gonzalez	0.9532	curve	0.1923
Francisco Liriano	0.9263	slider	0.3101
Blake Snell	0.8758	curve	0.1426

Appendix C

Batter Characteristics Examined by the Batter Similarity Measure

C.1 Batters with Large Year-to-Year Variation

We can use the similarity measure to compare batters to themselves over time. For this purpose we computed the similarity measure between 2016 and 2017 for each of the 79 right-handed batters who hit at least 250 batted balls in each regular season. The right-handed batters who changed the most between 2016 and 2017 along with their age on 30 June 2017 and their wOBA for the two seasons are given in Table C.1. The left-handed batters who changed the most between 2016 and 2017 among the 53 left-handed batters with at least 250 batted balls in each season are given in Table C.2. The switch-hitters who changed the most between 2016 and 2017 among the 23 switch-hitters with at least 250 batted balls in each season are given in Table C.3.

Twelve players among the biggest changers in Tables C.1 to C.3 had a wOBA difference of

RHB	Distance	Age	2016 wOBA	2017 wOBA
Matt Holliday	0.4165	37	.335	.320
Giancarlo Stanton	0.4120	27	.344	.410
Hernan Perez	0.3926	26	.312	.298
Jonathan Lucroy	0.3903	31	.362	.311
Evan Longoria	0.3809	31	.350	.312
Mark Trumbo	0.3566	31	.358	.295
Mike Trout	0.3543	25	.418	.437
Jose Bautista	0.3533	36	.355	.295
Marcell Ozuna	0.3493	26	.330	.388
Hanley Ramirez	0.3444	33	.367	.318

Table C.1: Right-handed batters who changed the most between 2016 and 2017

LHB	Distance	Age	2016 wOBA	2017 wOBA
Yonder Alonso	0.4027	30	.299	.366
Logan Morrison	0.3751	29	.318	.363
Jackie Bradley Jr.	0.3596	27	.354	.313
Scooter Gennett	0.3569	27	.315	.367
Jake Lamb	0.3535	26	.352	.353
Jason Kipnis	0.3528	30	.347	.300
Dee Gordon	0.3504	29	.280	.312
Gerardo Parra	0.3456	30	.284	.337
Matt Carpenter	0.3442	31	.375	.361
Bryce Harper	0.3434	24	.343	.416

Table C.2: Left-handed batters who changed the most between 2016 and 2017

Switch-Hitter	Distance	Age	2016 wOBA	2017 wOBA
Erick Aybar	0.3853	33	.271	.282
Yasmani Grandal	0.3809	28	.350	.325
Jed Lowrie	0.3782	33	.282	.347
Yangervis Solarte	0.3704	29	.346	.311
Eduardo Escobar	0.3532	28	.269	.320
Carlos Beltran	0.3421	40	.358	.283
Francisco Lindor	0.3405	23	.340	.353
Kendrys Morales	0.3291	34	.339	.320
Matt Wieters	0.3253	31	.307	.273
Neil Walker	0.3238	31	.351	.346

Table C.3: Switch-Hitters who changed the most between 2016 and 2017

at least 50 points between the seasons with eight players (Stanton, Ozuna, Alonso, Gennett, Parra, Harper, Lowrie, Escobar) improving and four players (Lucroy, Trumbo, Bautista, Beltran) declining. The only player in the tables of biggest changers who had at least a 3 mph increase in average exit velocity between 2016 and 2017 was switch-hitter Jed Lowrie who also enjoyed a 65 point gain in wOBA between the two seasons. Lowrie's success was fueled in part by achieving the largest improvement in difference between in-zone swing rate and out-of-zone swing rate between the two seasons [58]. Seven of the players in the tables (Holliday, Perez, Longoria, Trumbo, Bautista, Kipnis, Grandal) had at least a 3 mph decline in average exit velocity between 2016 and 2017 with five of these players being at least 30 years old at midseason 2017. All seven of these players had a lower wOBA in 2017 than in 2016.

Seven of the players in the tables (Trout, Alonso, Morrison, Parra, Carpenter, Lowrie, Lindor) increased their average launch angle by at least four degrees and all but Carpenter, who finished with the highest average launch angle in the NL in 2017, increased their wOBA from 2016 and 2017. Three of these players (Alonso, Morrison, Lindor) set new career highs in home runs with at least 15 more than their previous best and Alsonso admitted to making a conscious effort to hit more balls in the air in 2017 [52]. Two of the players in the tables (Longoria, Lucroy) had at least a four degree decrease in average launch angle. These players are both over 30 and also experienced decreases in average exit velocity and substantial declines in wOBA across the two seasons.

C.2 Batters with Small Year-to-Year Variation

The right-handed batters who changed the least between 2016 and 2017 according to the similarity measure along with their age on 30 June 2017 and their wOBA for the two seasons are given in Table C.4. The left-handed batters who changed the least between 2016 and 2017 are given in Table C.5 and the switch-hitters who changed the least are given in Table C.6. While 12 players among the biggest changers in section C.1 had a wOBA change of at least 50 points between 2016 and 2017, only Ben Zobrist among the least-changers in Tables C.4 through C.6 had a wOBA difference of at least 50 points between the two seasons.

RHB	Distance	Age	2016 wOBA	2017 wOBA
Nolan Arenado	0.2268	26	.386	.395
Ian Kinsler	0.2291	35	.356	.313
Maikel Franco	0.2312	24	.311	.292
Brandon Phillips	0.2378	36	.315	.316
Josh Donaldson	0.2417	31	.403	.396
Jose Altuve	0.2477	27	.391	.405
Khris Davis	0.2478	29	.349	.361
Mookie Betts	0.2484	24	.379	.339
Brandon Drury	0.2522	24	.335	.325
Jose Abreu	0.2556	30	.349	.377

Table C.4: Right-handed batters who changed the least between 2016 and 2017

LHB	Distance	Age	2016 wOBA	2017 wOBA
Jay Bruce	0.2343	30	.340	.350
Corey Seager	0.2410	23	.372	.364
Daniel Murphy	0.2430	32	.408	.385
Joe Panik	0.2435	26	.300	.329
Curtis Granderson	0.2448	36	.339	.330
Freddie Freeman	0.2480	27	.402	.407
Joe Mauer	0.2487	34	.327	.349
Nick Markakis	0.2497	33	.321	.321
Brett Gardner	0.2511	33	.317	.336
Jason Heyward	0.2544	27	.282	.311

Table C.5: Left-handed batters who changed the least between 2016 and 2017

Switch-Hitter	Distance	Age	2016 wOBA	2017 wOBA
Freddy Galvis	0.2414	27	.284	.298
Chase Headley	0.2423	33	.311	.329
Tucker Barnhart	0.2451	26	.300	.317
Asdrubal Cabrera	0.2551	31	.345	.338
Ben Zobrist	0.2628	36	.360	.302
Melky Cabrera	0.2642	32	.342	.319
Cesar Hernandez	0.2762	27	.335	.346
Victor Martinez	0.2868	38	.351	.303
Jose Ramirez	0.2911	24	.355	.396
Carlos Santana	0.2957	31	.370	.350

Table C.6: Switch-Hitters who changed the least between 2016 and 2017

Appendix D

Dependence of Prediction Accuracy on the Partition

The error in a prediction generated using the MSP approach depends on the partition of measurement space. Using equation (5.17), we can write the unobserved performance for player j as

$$y(j) = \sum_{k=1}^{B} \left(\widehat{p}_y(j,k) + \epsilon_p(j,k)\right)\left(\overline{R}(j,k) + \epsilon_R(j,k)\right). \tag{D.1}$$

The error terms are defined by $\epsilon_p(j,k) = p_y(j,k) - \widehat{p}_y(j,k)$ and $\epsilon_R(j,k) = \overline{R}_y(j,k) - \overline{R}(j,k)$ where $\overline{R}_y(j,k)$ is the average value of the unobserved batted balls in subset k for player j.

The prediction error is given by

$$y(j) - \widehat{y}_s(j) = \sum_{k=1}^{B} \left[\widehat{p}_y(j,k)\epsilon_R(j,k) + \overline{R}(j,k)\epsilon_p(j,k) + \epsilon_p(j,k)\epsilon_R(j,k)\right] \tag{D.2}$$

where each term in the sum depends on the subset k.

The error terms have a complex dependence on the group of subsets that define the partition. Reducing the size of the $\epsilon_R(j,k)$ error depends on balancing the competing goals of using subsets k that include enough data to estimate $\overline{R}(j,k)$ accurately but which also allow a single $\overline{R}(j,k)$ to be representative of any particular sample within a subset that might occur in $y(j)$. The variance of the $\epsilon_p(j,k)$ error is given by [10]

$$\mathrm{VAR}\left[\epsilon_p(j,k)\right] = \sigma_p^2(k)\left(1 - \alpha^2(N,k)\right) \tag{D.3}$$

where $\sigma_p^2(k)$ is the variance of $p_x(j,k)$ over batters j for subset k. Thus, $\mathrm{VAR}\left[\epsilon_p(j,k)\right]$ depends on both the distribution of the $p_x(j,k)$ and the $\alpha(N,k)$. Since the error terms and the prediction error in equation (D.2) have a complex dependence on the interaction between the measurement space partition and the structure of the data we use a learning process for partition selection as described in Sec. 5.4.6.

Appendix E

Applying MSP to First-Half 2021 Data

In this Appendix we apply the MSP approach to batted ball data collected before the All-Star break during the 2021 MLB season. The study considers the 185 batters with at least 150 batted balls during this period. The $\mathcal{P}_{5,10}$ partition was used and the $\alpha(150, k)$ values were estimated using the first 150 batted balls for each of the 185 batters. The subset means $\overline{R}(k)$ were used to approximate $\overline{R}(j, k)$ for each player j. The full set of first half batted balls for each player was used to compute $\widehat{y}_s(j)$ where the $\alpha(150, k)$ values were adjusted to $\alpha(N'(j), k)$ using the Spearman-Brown formula to regress each distribution according to each individual player's batted ball distribution and number of batted balls $N'(j)$. The result is an estimate of true talent wOBA on contact (wOBAcon) that has removed all contextual information (ballpark, batter running speed, atmospheric conditions, defense, etc.). The accuracy of wOBAcon predictions can be improved by incorporating context into the $\overline{R}(j, k)$ for each player j as demonstrated in Sec. 5.4.6. Table E.1 presents the context-invariant true talent wOBAcon leaders based on first-half 2021 data.

Table E.1: Context-invariant $\widehat{y}_s(j)$ estimate of wOBAcon using first-half 2021 data

Player	True Talent wOBAcon
Shohei Ohtani	.556
Fernando Tatis Jr.	.518
Giancarlo Stanton	.515
Vladimir Guerrero Jr.	.506
Ronald Acuna Jr.	.506
Aaron Judge	.500
Kyle Schwarber	.493
Joey Gallo	.488
Tyler O'Neill	.480
Nelson Cruz	.474
Yordan Alvarez	.470
Bryce Harper	.470
Rafael Devers	.469
Pete Alonso	.460
Matt Olson	.460