IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2020, accepted September 12, 2020, date of publication September 18, 2020,
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024699

Learning and Applying a Function Over Distributions

GLENN HEALEY, (Fellow, IEEE), AND SHIYUAN ZHAO
Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92617, USA
Corresponding author: Glenn Healey (ghealey@uci.edu)

ABSTRACT We present a method for learning a function over distributions. The method is based on generalizing nonparametric kernel regression by using the earth mover's distance 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.

INDEX TERMS Baseball, earth mover's distance, function over distributions, kernel regression, machine learning, nonparametric, pitching, sensor data.

I. 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 [1], [2]. Nonparametric methods [3] are a powerful tool for model recovery and continue to support a variety of applications [4], [5]. In this work, we generalize nonparametric techniques that learn a function of multiple variables to the problem of learning a function over distributions.

The ability to quantify player skill and team performance in professional sports has been revolutionized by the deployment of sensors that collect large amounts of data during each game [6], [7] [8]. This has led to the use of machine learning algorithms by teams to exploit this data to gain a competitive advantage. Machine learning methods are particularly well suited for baseball due to the discrete structure of the sport [9]. We will apply the learning method derived in this article to one of the most challenging problems in baseball analytics.

Nonparametric kernel regression can be used to estimate a function of unknown form and has been applied in a wide range of settings [10]. Generalizing this approach to learn a function over distributions requires a suitable metric for distribution space. The Wasserstein metric or Earth Mover's Distance (EMD) can be used to compare distributions and has been applied to many problems in signal processing and machine learning [11]. 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 [12].

We develop an algorithm that learns a function over distributions by generalizing nonparametric kernel regression using the EMD as the distribution-space metric. The algorithm is applied to 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 [13]. Pitchers typically use a small number of different pitch types which allows these distributions to be accurately encoded using signatures. A whitening transform [14] is used

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Magno.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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 [15] 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.

II. LEARNING A FUNCTION OVER DISTRIBUTIONS

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 III.

A. 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

$$y = f(x) + \epsilon \tag{1}$$

where ϵ is an error term. Kernel regression [16], [17] is a nonparametric method that constructs an estimate for $f(x)$ using the weighted average

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} k(d_i) y_i}{\sum_{i=1}^{n} k(d_i)} \tag{2}$$

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

$$k(d_i) = g(d_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}(d_i/\sigma)^2} \tag{3}$$

which depends on the smoothing parameter σ.

B. EARTH MOVER's DISTANCE

Given a set of observations (X_i, y_i) where each X_i is a multidimensional distribution, we can generalize equations (2) and (3) to approximate a function over distributions by replacing d_i with a distance D_i between the distributions X and X_i

$$\widehat{f}(X, \sigma) = \frac{\sum_{i=1}^{n} g(D_i, \sigma) y_i}{\sum_{i=1}^{n} g(D_i, \sigma)}\,. \tag{4}$$

The Wasserstein metric which is also called the Earth Mover's Distance (EMD) is a standard method for computing the distance between distributions. The EMD 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.

For many applications [12], a distribution can be accurately represented as a signature S defined by a set of m clusters

$$S = \{(\mu_1, w_1), \ldots, (\mu_m, w_m)\} \tag{5}$$

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 [12] for finding the EMD using signatures is based on the solution of the transportation problem [18] 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 using signatures can be formulated as a linear programming problem for which efficient solutions [19] and software [20] exist.

C. 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 define the ground distance $G(i, j)$ between μ_i and μ_j as the Mahalanobis distance [14]

$$G(i, j) = \left[(\mu_i - \mu_j)\Sigma^{-1}(\mu_i - \mu_j)^T\right]^{\frac{1}{2}} \tag{6}$$

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 [14] has been applied to transform the original variables to a new set of variables which are uncorrelated and have unit variance.

D. FINDING THE SMOOTHING PARAMETER USING CROSS VALIDATION

The accuracy of kernel regression has a strong dependence on the smoothing parameter σ [14]. 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) to compute

$$\widehat{f}(X = X_j, \sigma) = \frac{\sum\limits_{\substack{1 \le i \le n \\ i \ne j}} g(D_{ij}, \sigma) y_i}{\sum\limits_{\substack{1 \le i \le n \\ i \ne j}} g(D_{ij}, \sigma)} \tag{7}$$

where D_{ij} is the whitened EMD between X_i and X_j as described in Sections II-B and II-C and the (X_j, y_j)

observation is excluded from the sums. The error in the approximation is given by

$$E_j(\sigma) = y_j - \widehat{f}(X_j, \sigma). \tag{8}$$

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

$$\sigma^* = \arg\min_{\sigma} \sum_{j=1}^{n} \left| E_j(\sigma) \right|. \tag{9}$$

Note that if we include the (X_j, y_j) observation in the sums in (7), 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 (9) 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 [15].

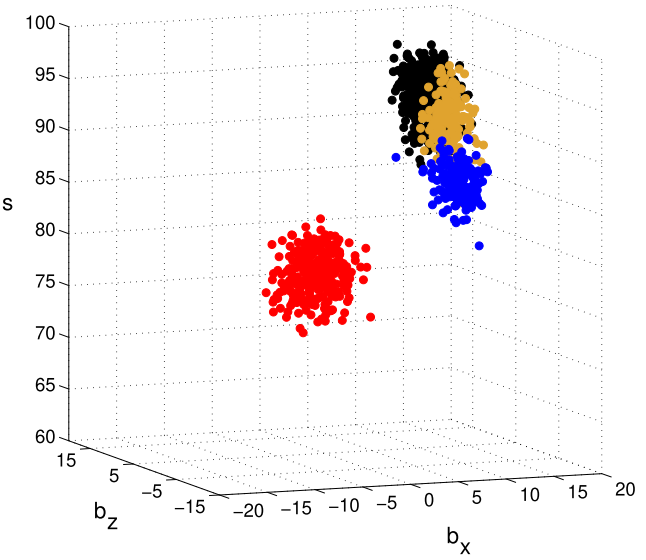

FIGURE 1. Chris Sale pitches in 2016.

III. EXPERIMENTAL RESULTS

A. 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 Doppler radar sensors [7] capture data during baseball games that can be exploited to recover information about pitches. Our analysis considers the estimated s, b_x, and b_z parameters for each pitch as reported by Brooks Baseball (www.brooksbaseball.net). The parameter s represents the speed of a pitch in three dimensions and the pair (b_x, b_z) specifies the pitch's horizontal and vertical movement relative to a theoretical pitch thrown at the same speed with no spin-induced movement [21].

The sensor coordinate system is arranged with the origin at home plate which is near the batter's location with positive z up, positive y parallel to the ground plane in the direction from the origin to the pitcher, and positive x chosen to complete a right-handed system. 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 (b_x, b_z) from $y = 40$ feet to home plate.

Pitchers typically throw different types of pitches to make the batter's task more difficult. A given pitch type has specific speed and movement characteristics. For example, a fourseam fastball from a right-handed major league baseball (MLB) pitcher will typically have a speed s above 90 miles per hour with a negative horizontal movement b_x and a positive vertical movement b_z due to backspin. A curveball from the same pitcher will typically have a speed s of less than 80 miles per hour with a positive b_x and a negative b_z due to topspin. For a left-handed pitcher, the sign of the horizontal movement b_x will reverse for these pitches. Major League Baseball Advanced Media (MLBAM) uses measured pitch parameters to classify the type of each pitch in real-time. After each game, Pitch Info (www.pitchinfo.com) uses a manual review process to improve on the accuracy of the MLBAM classifications. As an example, Figure 1 plots the distribution of pitches thrown by left-handed pitcher Chris Sale in 2016 for variables that represent pitch speed (s) in miles per hour and horizontal movement (b_x) and vertical movement (b_z) in inches. Different pitch types are shown in different colors in the figure.

B. 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 (b_x, b_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 [22] or from having pitches with similar speed that move in different directions [23].

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 [24] and is a strong determinant of a pitcher's success [25]. We can use the algorithm described in Section II to learn the dependence of pitcher strikeout rate on the pitch distribution defined over the s, b_x, and b_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 [26] 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 [27]. We will evaluate the use of the Nash score for assessing pitch distributions in Section III-C6.

C. DATA PROCESSING

1) OVERVIEW

We built the strikeout rate model described in Section III-B 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 [25] 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 II-B and associated strikeout rates y_i. The covariance matrix Σ in equation (6) 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 II-D. 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). This process is summarized by Figure 2 where application of the model will be described in more detail in Section III-C5.

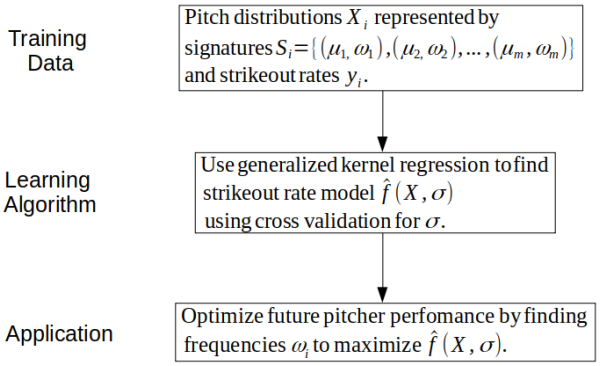

FIGURE 2. Process of learning and applying a function over distributions.

2) SIGNATURE MODEL

Pitchers tend to throw a small number of distinct pitch types which allows the pitch distribution for a pitcher for a given year and platoon configuration to be accurately modeled using the signature representation of equation (5) 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 $(\bar{s}_i, \bar{b}_{xi}, \bar{b}_{zi})$ and w_i is the fraction of pitches of that type for the pitcher and platoon configuration.

3) COMPUTING THE EMD

The signatures are used to compute the distance between distributions using the EMD as described in Section II-B with the whitened ground distance defined in Section II-C. As a two-dimensional example of this process, Figure 3 is a scatterplot of the mean $(\bar{s}_i, \bar{b}_{zi})$ values for each pitch cluster in a signature for the right-handed pitcher versus right-handed batter platoon configuration in 2016. We see that $\bar{s}_i$ and $\bar{b}_{zi}$ 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 $\bar{s}_i$ values is also larger than the variance of the $\bar{b}_{zi}$ values. These effects are addressed by using the Mahalanobis ground distance defined by equation (6).

The impact of the correlation between the two variables can be seen by considering the orange, green, and red points in Figure 3 which correspond to the $(\bar{s}_i, \bar{b}_{zi})$ values for three specific pitch clusters in the figure as detailed in Table 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 b_z. On the other hand, the vector difference between the green point and the orange point is approximately orthogonal

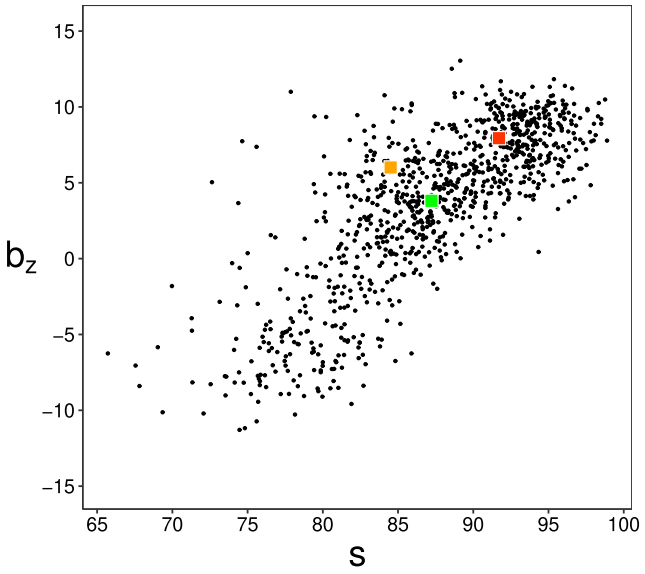

FIGURE 3. Cluster means $(\bar{s}_i, \bar{b}_{zi})$ for RHP versus RHB configuration, 2016.

TABLE 1. Three $(\bar{s}_i, \bar{b}_{zi})$ pitch cluster means in Figure 3.

Point color	Pitcher	Pitch type	$(\bar{s}_i, \bar{b}_{zi})$
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)

to the direction of correlation. If we compute the Mahalanobis distance using the s and b_z variables shown in Table 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.

4) CROSS VALIDATION

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

$$E_T(\sigma) = \sum_{j=1}^{n} \left| E_j(\sigma) \right| \tag{10}$$

defined in equation (9). 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) 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

$$E_T(\sigma) \leq 1.001 * min\left[E_T(\sigma)\right]. \tag{11}$$

The use of this equation to favor smaller values of σ has little effect on the accuracy of the model in equation (4) but can improve the efficiency of the computation.

Figures 4 to 7 plot $E_T(\sigma)$ for each of the four platoon configurations. The resulting values of σ^* are shown in Table 2.

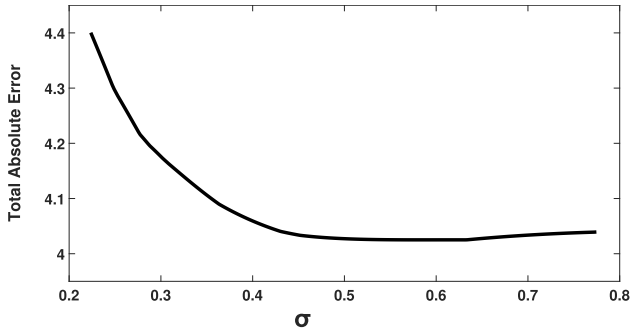

FIGURE 4. $E_T(\sigma)$ for RHP versus RHB configuration, 2016.

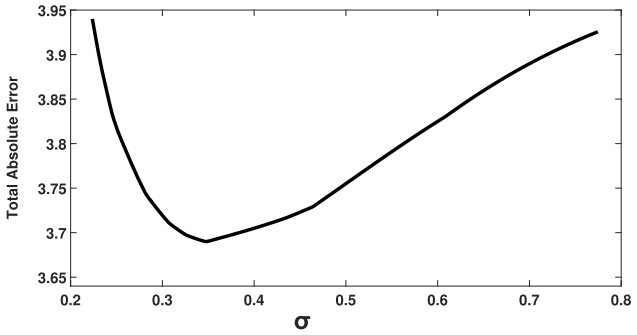

FIGURE 5. $E_T(\sigma)$ for RHP versus LHB configuration, 2016.

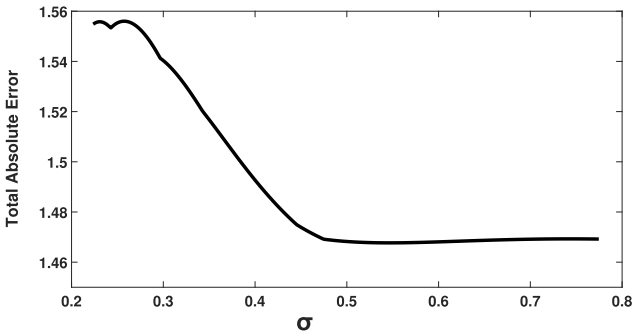

FIGURE 6. $E_T(\sigma)$ for LHP versus RHB configuration, 2016.

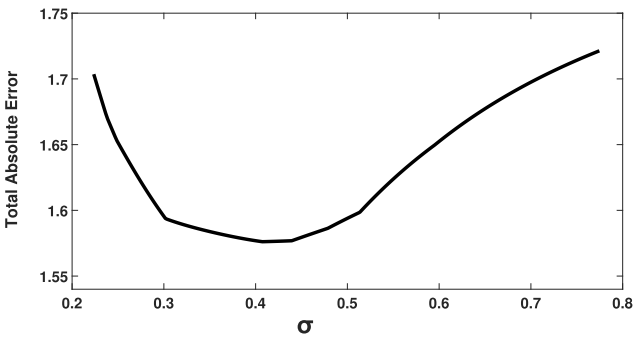

FIGURE 7. $E_T(\sigma)$ for LHP versus LHB configuration, 2016.

For small values of σ, the $g(D_{ij}, \sigma)$ in equation (7) are approximately Dirac delta functions and $\widehat{f}(X_j, \sigma)$ is approximately a sum of Dirac delta functions centered at the observations

TABLE 2. 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

(X_i, y_i) for $i \neq j$. This results in a relatively large error $E_j(\sigma)$ for small σ in equation (8) and a relatively large error in $E_T(\sigma)$ for small σ in equation (10). As σ increases, the approximation in equation (8) improves and the error decreases as shown in the figures.

5) 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 (5). Each pitch type i has a pitch parameter vector $\mu_i = (\bar{s}_i, \bar{b}_{xi}, \bar{b}_{zi})$ 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 3. We note that the signature model S in equation (5) 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 3 during 2016. Other pitchers use other pitch types such as the cutter and the split which are represented in their signatures. Figure 8 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 8.

TABLE 3. Pitch signature for LHP Danny Duffy versus LHB for 2016.

Pitch type	index	w	$\bar{s}$	$\bar{b}_x$	$\bar{b}_z$
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

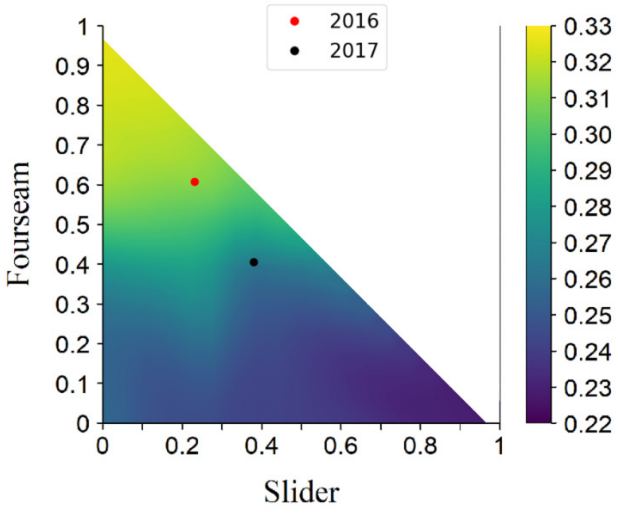

FIGURE 8. Danny Duffy $\widehat{f}(X, \sigma^*)$ for LHP versus LHB configuration, 2016.

6) 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

$$\Delta = (2017\ rate) - (2016\ rate) \tag{12}$$

$$\widehat{\Delta} = (2017\ predicted\ rate) - (2016\ rate) \tag{13}$$

where 2017 predicted strikeout rate is computed by evaluating $\widehat{f}(X, \sigma^*)$ using equation (4) for the pitcher's 2017 pitch distribution with σ^* computed as described in Section III-C4. Figure 9 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.

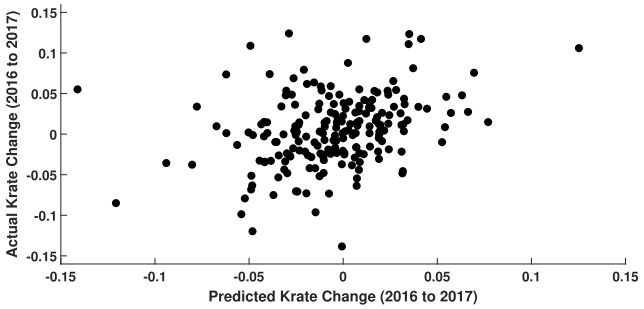

FIGURE 9. Predicting strikeout rate changes using $\widehat{f}(X, \sigma^*)$.

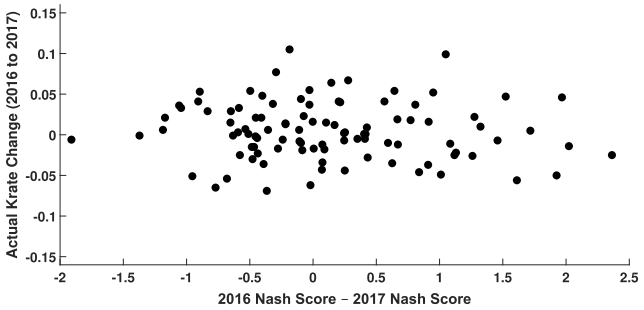

FIGURE 10. Predicting strikeout rate changes using Nash score changes.

For comparison, Figure 10 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 [26]

$$\Delta_N = 2016\ Nash\ score - 2017\ Nash\ score \qquad (14)$$

As described in Section III-B, 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 10, 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 [27] on which the Nash score is based.

We can assess the statistical significance of the difference between the correlation coefficients of $r_1 = 0.320$ in Figure 9 and $r_2 = -0.081$ in Figure 10 using the Fisher z-transformation [28]. 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 III-C5.

IV. CONCLUSION

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.

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.

The new technique can be used for a number of applications in the areas of strategy [29], player development [30], and player evaluation [31] in baseball as well as for play selection [32] in football. Given the physical characteristics of a pitcher's different pitch types, the function can be used to determine the frequencies for each pitch type that maximize strikeout rate. The method can also be used to evaluate the improvement in strikeout rate that is possible by adding a new pitch type to a pitcher's current collection of pitches. By utilizing physical measurements, the algorithm allows the direct comparison of pitchers across environments. This enables, for example, a prediction of how a college pitcher would perform in major league baseball after optimizing his pitch distribution. 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. This model could then be utilized to determine the play distribution that a given offense should use to maximize success.

ACKNOWLEDGMENT

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.

REFERENCES

[1] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5939–5948.
[2] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4460–4470.
[3] A. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*. Oxford, U.K.: Clarendon Press, 1997.
[4] M. Khalily, T. W. C. Brown, and R. Tafazolli, "Machine-learning-based approach for diffraction loss variation prediction by the human body," *IEEE Antennas Wireless Propag. Lett.*, vol. 18, no. 11, pp. 2301–2305, Nov. 2019.

[5] W. Wang, W. Han, X. Na, J. Gong, and J. Xi, "A probabilistic approach to measuring driving behavior similarity with driving primitives," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 1, pp. 127–138, Mar. 2020.
[6] K. Clark. (Dec. 19, 2018). *The NFL's Analytics Revolution Has Arrived*. [Online]. Available: https://www.theringer.com/nfl/2018/12/19/18148153/nfl-analytics-revolution
[7] G. Healey, "The new moneyball: How ballpark sensors are changing baseball," *Proc. IEEE*, vol. 105, no. 11, pp. 1999–2002, Nov. 2017.
[8] M. Woo. (Dec. 21, 2018). *Artificial Intelligence in NBA Basketball*. [Online]. Available: https://www.insidescience.org/news/artificial-intelligence-nba-basketball
[9] K. Koseler and M. Stephan, "Machine learning applications in baseball: A systematic literature review," *Appl. Artif. Intell.*, vol. 31, nos. 9–10, pp. 745–763, Nov. 2017.
[10] J. Kloke and J. McKean, *Nonparametric Statistical Methods Using R*. New York, NY, USA: Chapman & Hall/CRC, 2014.
[11] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 43–59, Jul. 2017.
[12] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[13] G. Healey, "A Bayesian method for computing intrinsic pitch values using kernel density and nonparametric regression estimates," *J. Quant. Anal. Sports*, vol. 15, no. 1, pp. 59–74, Mar. 2019.
[14] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley-Interscience, 2001.
[15] S. J. Sheather, "Density estimation," *Statist. Sci.*, vol. 19, no. 4, pp. 588–597, 2004.
[16] É. A. Nadaraya, "On non-parametric estimates of density functions and regression curves," *Theory Probab. Appl.*, vol. 10, no. 1, pp. 186–190, Jan. 1965.
[17] G. Watson, "Smooth regression analysis," *Sankhyā, Indian J. Statist., A*, vol. 26, no. 4, pp. 359–372, 1964.
[18] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *J. Math. Phys.*, vol. 20, nos. 1–4, pp. 224–230, Apr. 1941.
[19] F. Hillier and G. Liberman, *Introduction to Mathematical Programming*. New York, NY, USA: McGraw-Hill, 1990.
[20] S. Urbanek and Y. Rubner. (Feb. 19, 2015). *Package 'Emdist'*. [Online]. Available: http://cran.r-project.org/web/packages/emdist/emdist.pdf
[21] A. Nathan. (Oct. 21, 2012) *Determining Pitch Movement From PITCHf/x Data*. [Online]. Available: baseball.physics.illinois.edu/Movement.pdf
[22] R. Gray, "Behavior of college baseball players in a virtual batting task," *J. Experim. Psychol., Hum. Perception Perform.*, vol. 28, no. 5, pp. 1131–1148, 2002.
[23] J. Long, J. Judge, and H. Pavlidis. (Jan. 24, 2017). *Introducing Pitch Tunnels*. [Online]. Available: http://baseballprospectus.com/news/article/31030/prospectus-feature-introducing-pitch-tunnels/
[24] R. Carleton. (May 9, 2013). *Should I Worry About my Favorite Pitcher?* [Online]. Available: http://baseballprospectus.com/news/article/20516/baseball-therapy-should-i-worry-about-my-favorite-pitcher/
[25] S. Slowinski. (Feb. 15, 2010). *FIP*. [Online]. Available: http://library.fangraphs.com/pitching/fip/
[26] N. Paine. (Aug. 13, 2015). *Game Theory Says R.A. Dickey Should Throw More Knuckleballs*. [Online]. Available: fivethirtyeight.com/features/game-theory-says-r-a-dickey-should-throw-more-knuckleballs
[27] D. Appelman. (May 20, 2009). *Pitch Type Linear Weights*. [Online]. Available: https://www.fangraphs.com/blogs/ pitch-type-linear-weights
[28] R. Fisher, "On the probable error of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, no. 4, pp. 3–32, Sep. 1921.
[29] T. Tango, M. Lichtman, and A. Dolphin, *The Book: Playing the Percentages in Baseball*. Dulles, Virgina: Potomac Books, 2007.
[30] B. Lindbergh and T. Sawchik, *The MVP Machine: How Baseball's New Nonconformists are Using Data to Build Better Players*. New York, NY, USA: Basic Books, 2019.
[31] T. Sawchik, *Big Data Baseball*. New York, NY, USA: Flatiron Books, 2016.
[32] E. McGough, C. Clemons, M. Ferrara, T. Norfolk, and G. W. Young, "A game-theoretic approach to personnel decisions in American football," *J. Quant. Anal. Sports*, vol. 6, no. 4, pp. 1–15, Oct. 2010.

GLENN HEALEY (Fellow, IEEE) received the B.S.E. degree in computer engineering from the University of Michigan and the M.S. degree in computer science, the M.S. degree in mathematics, and the Ph.D. degree in computer science from Stanford University. He is currently a Professor of electrical engineering and computer science with the University of California at Irvine (UC Irvine). Before joining UC Irvine, he worked with IBM Research. He is the Director with the Computer Vision Laboratory, UC Irvine. His research interests include using physical models to develop algorithms that extract information from large sets of data. He has been elected a Fellow of SPIE and is a member of SABR. He has served on the editorial boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the *Journal of the Optical Society of America A*.

SHIYUAN ZHAO received the B.S.E. degree in aerospace engineering from Beihang University and the M.S. degree in electrical engineering from the University of California at Irvine where she is currently pursuing the Ph.D. degree in electrical engineering and computer science.