Learning a Function over Distributions

Glenn Healey and Shiyuan Zhao

Electrical Engineering and Computer Science

University of California, Irvine

ghealey@uci.edu,shiyuaz1@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 method accurately predicts changes in pitcher performance in response to changes in pitch distribution.

1 Introduction

An important application of machine learning is the recovery of a model from observed data. We consider the problem of learning a function over distributions with the subsequent goal of maximizing the function over low-dimensional subsets of the distribution space. Nonparametric kernel regression can be used to estimate a function of unknown form and has been applied in a wide range of settings [8]. 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 [9]. 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 [13].

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 which is one of the most challenging problems in baseball analytics. 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 [5]. Pitchers typically use a small number of different pitch types which allows these distributions to be accurately encoded using signatures. A whitening transform [1] 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 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.

2 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 regression using a whitened Earth Mover's Distance as the metric for distribution space. Cross-validation is used to optimize the smoothing parameter of the method. We will illustrate properties of the algorithm with a set of experiments in Section 3.

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

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

where ϵ is an error term. Kernel regression [11] [16] is a non-parametric 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.

2.2 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 [13], 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 [13] for finding the EMD using signatures is based on the solution of the transportation problem [7] 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 [6] and software [15] exist.

2.3 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 [1]

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

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

2.4 Finding the smoothing parameter using cross validation

The accuracy of kernel regression has a strong dependence on the smoothing parameter σ [1]. 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

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

where D_{ij} is the whitened EMD between X_i and X_j as described in Sections 2.2 and 2.3 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

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

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 (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 [14].

3 Experimental Results

3.1 Learning Strikeout Rate over Pitch Distributions

Strikeout rate is a strong determinant of a pitcher's success. We demonstrate the algorithm described in Section 2 for the problem of learning the dependence of pitcher strikeout rate on a multivariate pitch distribution defined over a vector of parameters derived from sensor measurements. Figure 1, for example, 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, e.g. sinker or slider, are shown in different colors in the figure. Pitchers tend to throw a small number of distinct pitch types which allows pitch distributions to be accurately modeled using the signature representation of (5) where each pitch type corresponds to a cluster. The ability to learn a function for predicting strikeout rate as a function of pitch distribution has several important applications. Given a pitcher's set of pitches, the function can be used to determine the frequencies for each pitch type that maximize strikeout rate. In addition, the function can be used to evaluate the potential value of new pitch types for improving strikeout rate. Thus, the new algorithm can be used to develop tools that guide pitchers in their quest to improve.

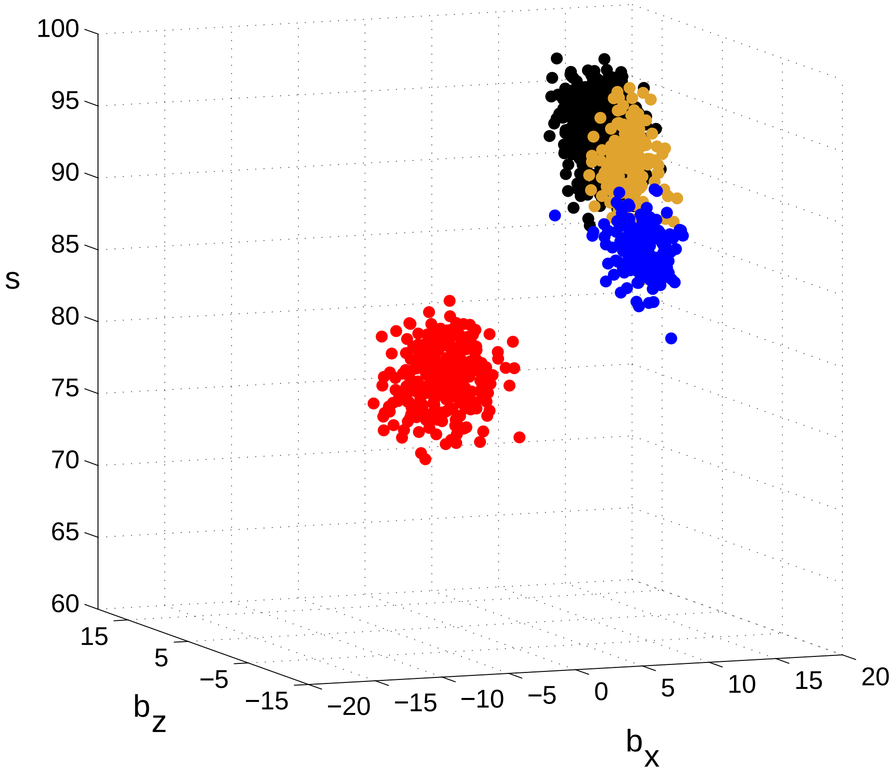

Figure 1: Chris Sale pitches in 2016

3.2 Sensor Data

The PITCHf/x optical video and TrackMan Doppler radar sensors [4] 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 [12]. The coordinate system origin is at home plate with positive x to the right from the catcher's perspective, positive z up, and positive y in the direction from home plate to second base. By convention, Brooks Baseball reports s for $y = 55$ feet and (b_x, b_z) from $y = 40$ feet to home plate. A pitcher's success is highly dependent on the speed and movement of his pitches. A larger speed s 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 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. 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. For a left-handed pitcher, the sign of the horizontal movement b_x will reverse for these pitches. A pitcher can benefit from having pitches with large differences in speed [3] or from having pitches with similar speed that move in different directions [10]. 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.

3.3 Data Processing

3.3.1 Overview

We built the strikeout rate model as described in Sections 3.1 and 3.2 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 [2] 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.

3.3.2 Signature Model

The pitch distribution for a pitcher for a given year and platoon configuration is represented using a signature as defined by (5). 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 $(\overline{s}_i, \overline{b}_{xi}, \overline{b}_{zi})$ and w_i is the fraction of pitches of that type for the pitcher and platoon configuration.

3.3.3 Computing the EMD

The signatures are used to compute the distance between distributions using the EMD as described in Section 2.2 with the whitened ground distance defined in Section 2.3. As a two-dimensional example of this process, Figure 2 is a scatterplot of the mean $(\overline{s}_i, \overline{b}_{zi})$ values for each pitch cluster in a signature for the right-handed pitcher versus right-handed batter platoon configuration in 2016. We see that $\overline{s}_i$ and $\overline{b}_{zi}$ 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 $\overline{s}_i$ values is also larger than the variance of the $\overline{b}_{zi}$ values. These effects are addressed by using the Mahalanobis ground distance defined by (6). As a specific example, a significant portion of the separation between the orange and red points in the figure is due to the correlation between the variables which results in a Euclidean distance between these points which is larger than the Euclidean distance between the orange and green points.

The Mahalanobis distance accounts for this correlation and results in a distance between the orange and red points which is less than the distance between the orange and green points.

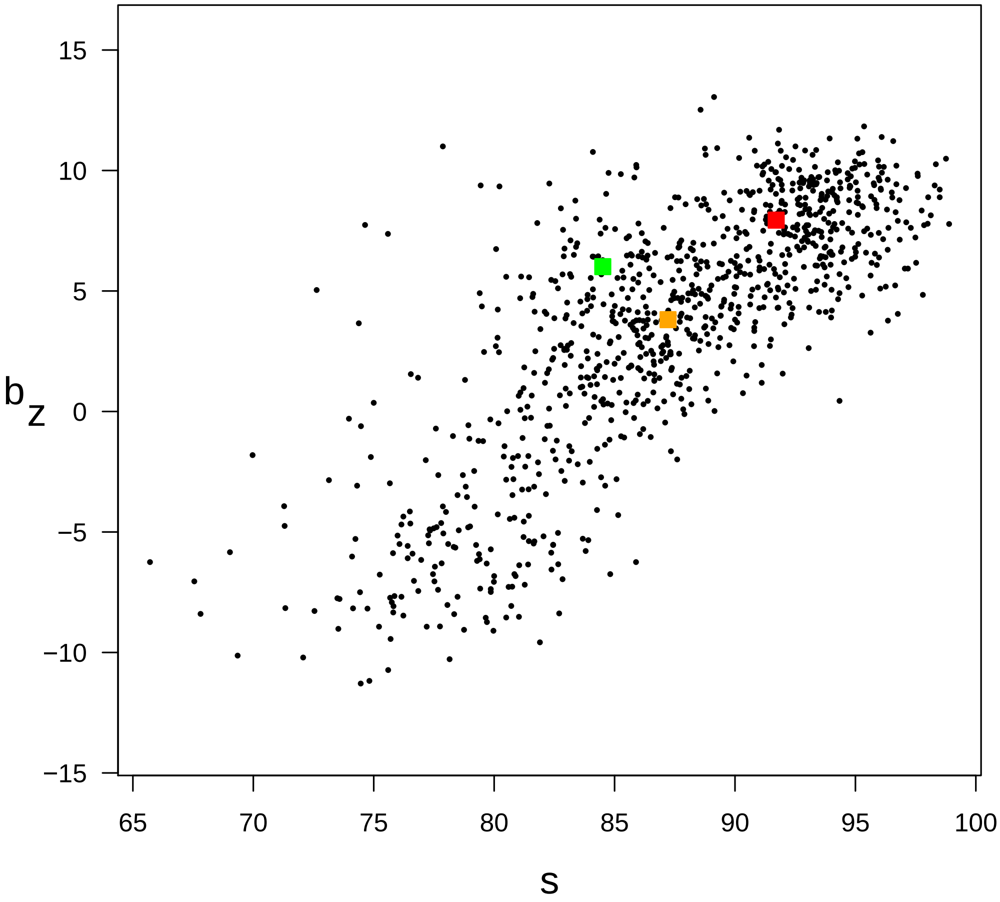

Figure 2: Cluster means $(\overline{s}_i, \overline{z}_i)$ for RHP versus RHB configuration, 2016

3.3.4 Cross validation

The cross validation process described in Section 2.4 is used to find optimized values for the smoothing parameter σ for each platoon configuration using the total absolute error

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

defined in (9). Figures 3 to 6 plot $E_T(\sigma)$ for each of the four platoon configurations. Since two of the curves decrease rapidly before remaining nearly flat for a significant range of σ, 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 resulting values of σ^* are shown in Table 1.

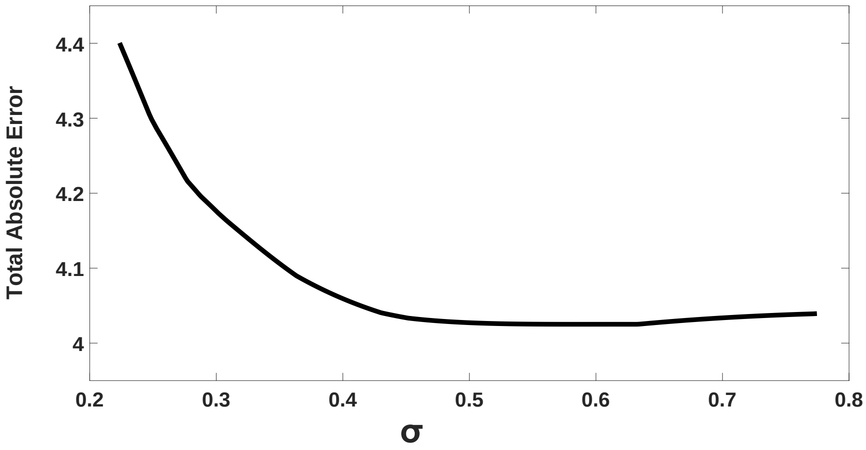

Figure 3: $E_T(\sigma)$ for RHP versus RHB configuration, 2016

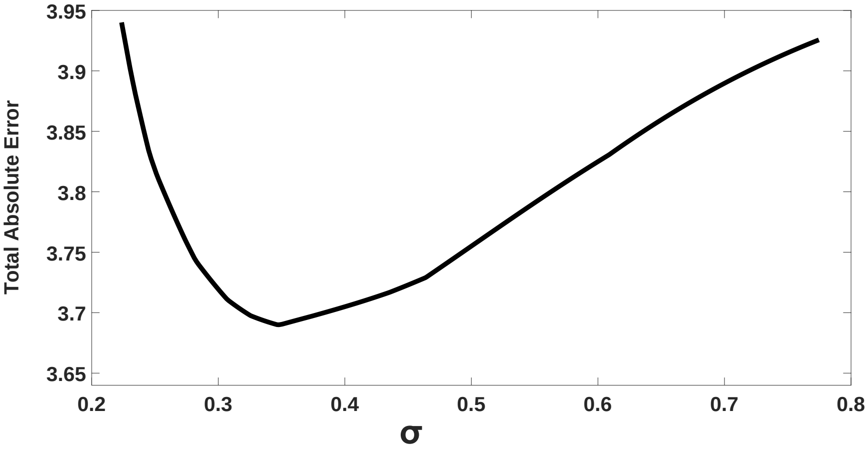

Figure 4: $E_T(\sigma)$ for RHP versus LHB configuration, 2016

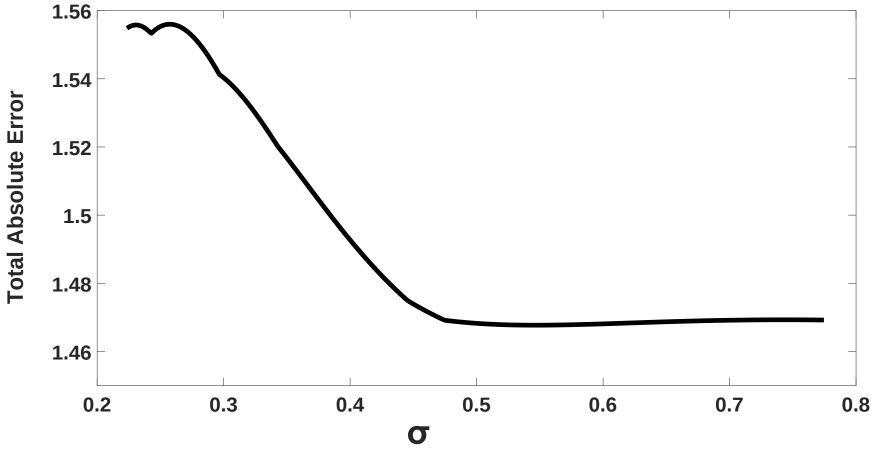

Figure 5: $E_T(\sigma)$ for LHP versus RHB configuration, 2016

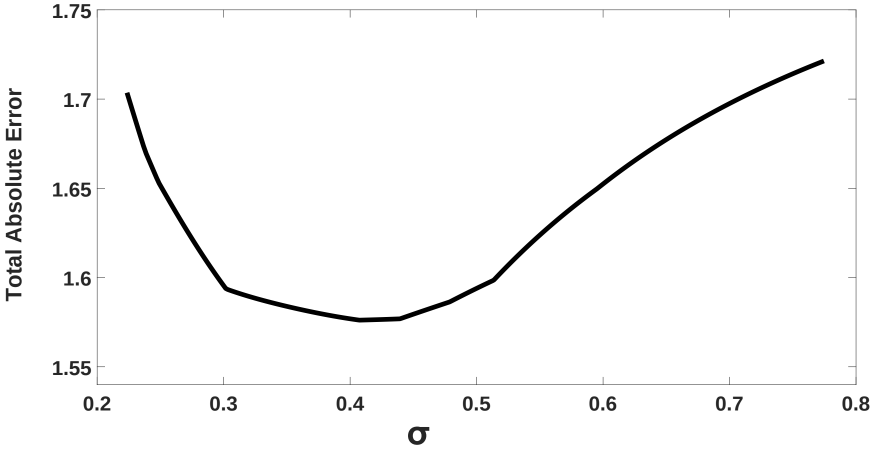

Figure 6: $E_T(\sigma)$ for LHP versus LHB configuration, 2016

Table 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

3.3.5 Finding optimized pitch frequencies

Given the estimated $\widehat{f}(X, \sigma^*)$ for representing strikeout rate as a function of the pitch distribution X, we can optimize the pitch frequencies for a pitcher with a given set of pitches. If X is represented by a signature as in (5), the optimization requires a search over the m weights w_i to maximize $\widehat{f}(X, \sigma^*)$ subject to the constraints $w_1 + w_2 + \ldots + w_m = 1$ and $w_i \geq 0$. The number of pitch types m is typically small which allows an exhaustive search to be performed efficiently.

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 2. Figure 7 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 7.

Table 2: Pitch signature for LHP Danny Duffy versus LHB for 2016

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

3.3.6 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

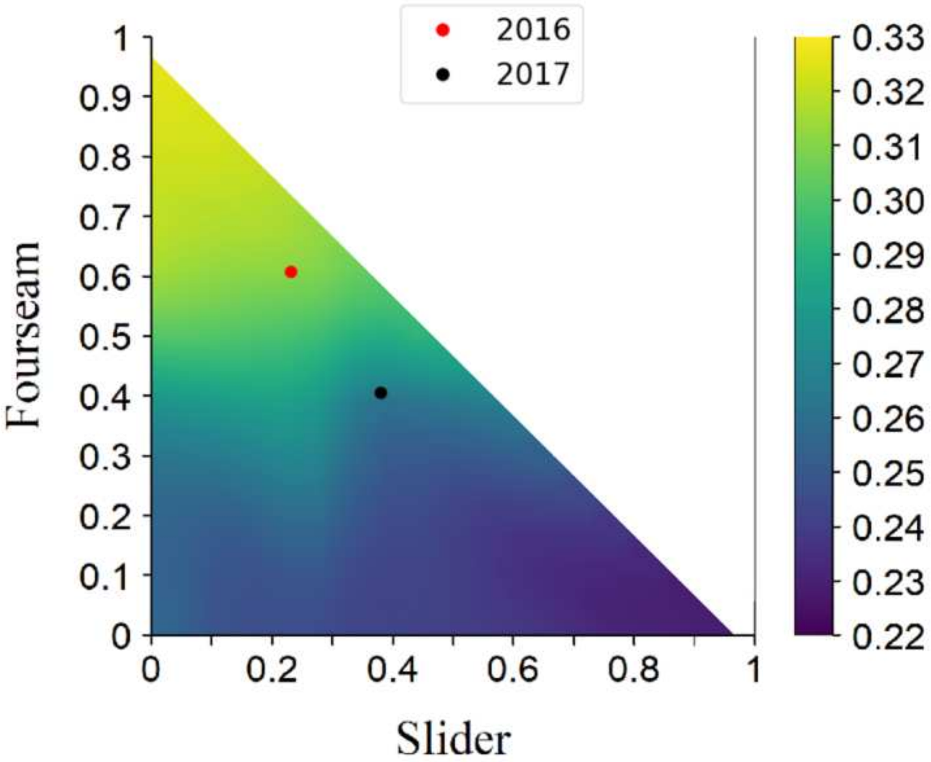

Figure 7: Danny Duffy $\widehat{f}(X, \sigma^*)$ for LHP versus LHB configuration, 2016

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

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

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

where 2017 predicted strikeout rate is computed by evaluating $\widehat{f}(X, \sigma^*)$ for the pitcher's 2017 pitch distribution. Figure 8 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.

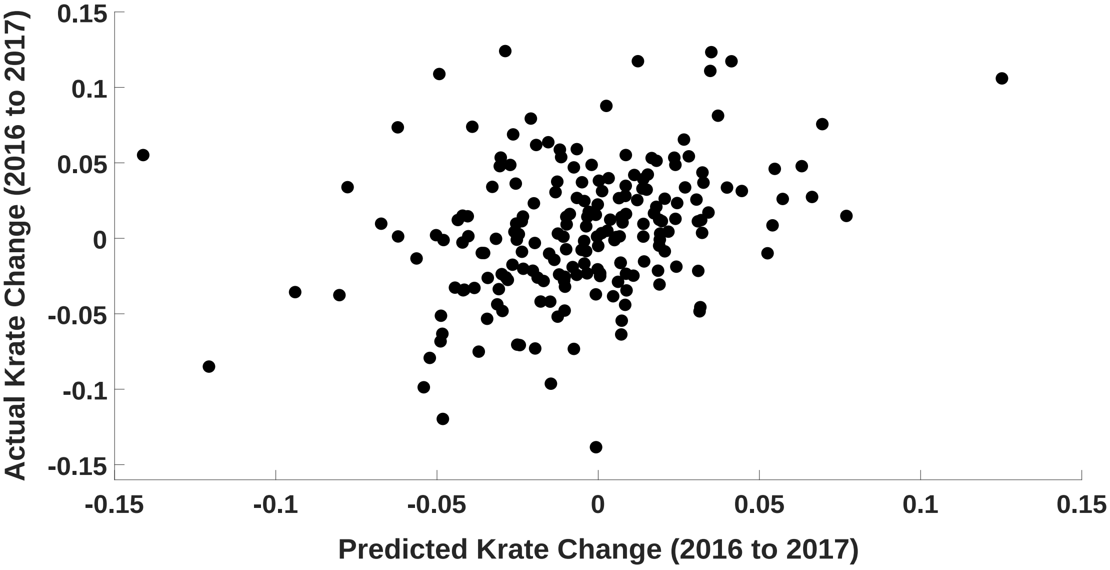

Figure 8: Predicting strikeout rate changes using $\widehat{f}(X, \sigma^*)$

4 Conclusion

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. The algorithm can be used by pitchers to find optimized pitch distributions or to evaluate the utility of adding a new pitch type. By utilizing physical measurements, the algorithm also allows the 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. We assessed the algorithm for the prediction of strikeout rate from pitch distributions on out-of-sample data. The method for learning a function over distributions can be easily adapted for other application areas.

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. We thank June Chen and Jutao Guo for their help with this manuscript.

References

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

[2] Fielding Independent Pitching (FIP) [Online]. Available: www.fangraphs.com/library/pitching/fip/.

[3] 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.

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

[5] 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.

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

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

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

[9] S. Kolouri, S.R. 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.

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

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

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

[13] 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.

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

[15] S. Urbanek and Y. Rubner. Package 'emdist'. Technical report, CRAN, February 19, 2015.

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