Measuring Pitcher Similarity: Technical Details

Glenn Healey, Shiyuan Zhao, Dan Brooks

Summary: Given the speed and movement for pitches thrown by a set of pitchers, we develop a measure of pitcher similarity.

1 Sensor Data

The PITCHf/x optical video and TrackMan Doppler radar sensors capture data that is exploited to recover information about pitches [1] [3]. Let s represent the initial speed of a pitch in three dimensions and let the pair (x, z) specify the pitch's movement. 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 [6]. 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. Our pitcher similarity measure will consider the estimated s, x, and z parameters for each pitch.

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 (www.brooksbaseball.net) 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 Representing Pitch Distributions

We will represent pitchers by their distribution of (s, x, z) pitch vectors. Given that pitchers typically throw a small number of different pitches, 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

$$P_{_R} = \{(\mu_1, w_1), \ldots, (\mu_m, w_m)\} \tag{1}$$

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 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 The Earth Mover's Distance

3.1 Overview

We can assess the similarity of distributions that are represented by signatures using the Earth Mover's Distance [7]. The Earth Mover's Distance (EMD) utilizes a ground distance between individual points in the distributions 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. The algorithm for finding the EMD is based on the solution of the transportation problem [5] 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 [4] and software [8] exist. For the purpose of comparing pitchers, the EMD has the advantage of allowing the comparison of all pitches thrown by a pair of pitchers regardless of pitch type. The EMD is also not sensitive to the vagaries of pitch classification algorithms since clusters with similar properties will be seen as similar even if they have been assigned different pitch labels.

3.2 Ground Distance

In order to apply the EMD to the measurement of pitcher similarity, we need to define a ground distance between the $\mu_i = (\overline{s}_i, \overline{x}_i, \overline{z}_i)$ mean vectors that define the pitcher signatures.

The use of a standard Euclidean distance between pitch vectors is problematic because the speed variable s corresponds to a different physical quantity that has a larger variance than the movement variables x and z and also because the three variables are strongly correlated. Figure 1, for example, 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 configuration in 2016.

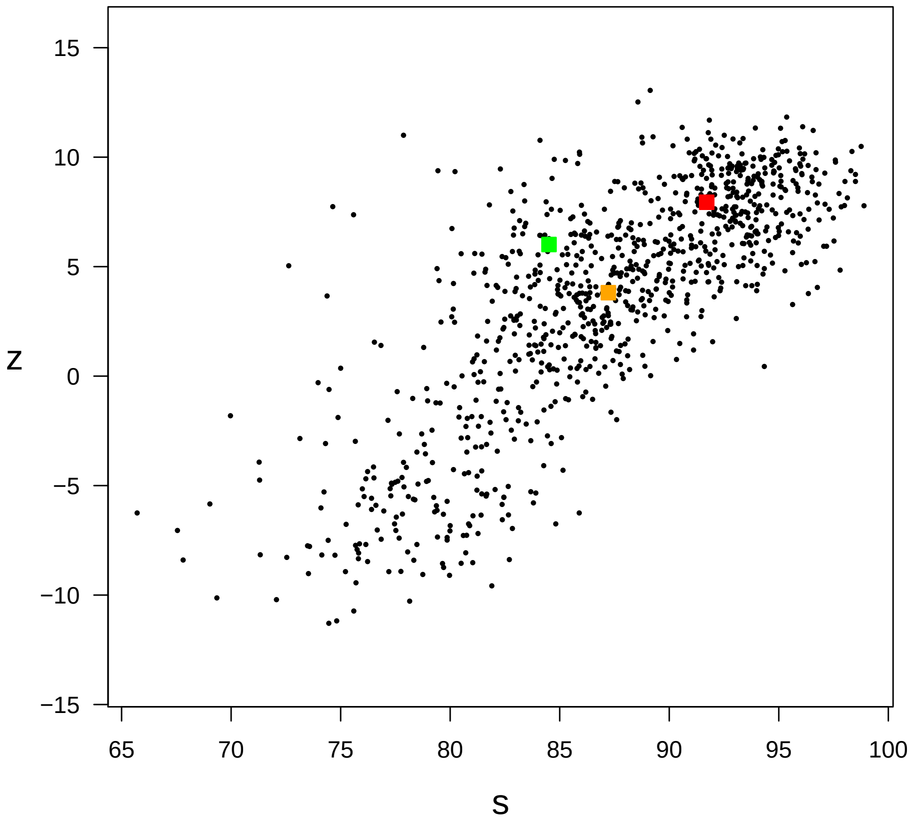

Figure 1: Cluster means $(\overline{s}_i, \overline{z}_i)$ for RHP versus RHB configuration, 2016

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

Table 1: Three points in figure 1

We see from the figure that s and z have a large positive correlation so that a pitch thrown with a higher speed s will tend to have a higher vertical movement z. The effect of this correlation can be seen by considering the green, orange, and red points in figure 1 which correspond respectively to the Ian Kennedy changeup, Mat Latos cutter, and Jhoulys Chacin four-seam fastball as detailed in Table 1. The Euclidean distance of 6.10 between the Latos cutter and the Chacin four-seam is significantly larger than the Euclidean distance of 3.49 between the Latos cutter and the Kennedy changeup. Since the vector difference between the Latos cutter and the Chacin four-seam 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 Latos cutter and the Kennedy changeup is approximately orthogonal to the direction of correlation. The variance of the s variable is also larger than the variance of z which tends to cause a larger separation between pitch clusters in s than in z. Ideally, the ground distance should compensate for the correlation structure and the variance of the s, x, and z variables.

The covariance matrix Σ for the population of mean vectors $\mu_i = (\overline{s}_i, \overline{x}_i, \overline{z}_i)$ for a platoon configuration captures information about both the variance of the individual components of μ_i and their correlations. Using this information, the Mahalanobis distance [2] can be used to compute a measure of separation between the mean vectors μ_i and μ_j for a pair of pitch clusters within a platoon configuration that corrects for differences in the variances of the s, x, and z components of the vectors and also for the correlation structure of the components. If we compute the Mahalanobis distance using the s and z variables shown in figure 1, the distance of 0.81 between the Latos cutter and the Chacin four-seam is now significantly less than the distance of 1.32 between the Latos cutter and the Kennedy changeup. We define the ground distance $G(i,j)$ between μ_i and μ_j as the Mahalanobis distance

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

This distance is equivalent to a Euclidean distance after a whitening transform [2] has been used to transform the original variables to a new set of variables which are uncorrelated and have unit variance.

4 Pitcher Similarity Measure

Given the ground distance defined by (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

$$D(A, B) = f_{RR} D_R(A, B) + f_{RL} D_L(A, B) \tag{3}$$

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

$$D(Y, Z) = f_{LR} D_R(Y, Z) + f_{LL} D_L(Y, Z) \tag{4}$$

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.

References

[1] Z. Day. (Dec. 5, 2013). Measuring pitching with Trackman [Online]. Available: www.baseball.prospectus.com/article.php?articleid=22362.

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

[3] 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.

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

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

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

[7] 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.

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