The Intrinsic Value of a Pitch

Glenn Healey (ghealey@uci.edu)
Electrical Engineering and Computer Science
University of California, Irvine, CA 92617
Research Paper for 2017 SABR Analytics Conference

1 Introduction

Traditional statistics that are employed to measure and predict the performance of pitchers are based on observed outcomes. These statistics are affected by a number of confounding variables that are beyond the control of the pitcher such as the defense, the ballpark, the umpire, and the catcher. Approaches that reduce the dependence of statistics on these variables include selectively removing plate appearances as in FIP [12] or attempting to adjust for the effect of variables as in DRA [25]. Both of these approaches have important limitations. Removing plate appearances restricts the scope of a statistic while the use of inexact adjustments introduces distortion in a statistic's computed value.

As baseball games have been recorded with increasing detail, analysts have developed methods for measuring pitcher performance at the level of an individual pitch. In his pioneering work, Burley [6] used a linear weights model [38] to define the observed value of a pitch as the change in run expectancy given the pitch outcome. Using PITCHf/x measurements, Walsh [39] extended the approach by using a pitch classification scheme to assign linear weight values to different pitch types. This inspired further research [1] [28] [29] and linear weight pitch values are readily accessible on the internet [2]. Since these pitch values depend on observed outcomes, however, they are affected by numerous variables that are independent of the quality of a pitch. Individual factors such as framing, for example, have been shown to have a large effect on observed pitch values [32]. Due to these confounding variables, observed pitch values have a low degree of repeatability [33] which limits their utility for prediction.

The deployment of sensors [11] [24] [26] that characterize the trajectory of pitches and batted balls in three dimensions provides the opportunity to assign an intrinsic value to a pitch that depends on its physical properties and not on its observed outcome. We exploit

this opportunity by utilizing a Bayesian framework to map five-dimensional PITCHf/x velocity, movement, and location vectors to pitch intrinsic values. HITf/x data is used by the model to obtain intrinsic quality-of-contact values [17] [18] for batted balls that are invariant to the defense, ballpark, and atmospheric conditions. Separate mappings are built to accommodate the effects of count and batter/pitcher handedness. A kernel method is used to generate nonparametric estimates for the component probability density functions in Bayes theorem while cross-validation enables the model to adapt to the size and structure of the data. The methodology does not suffer the loss of information that is inherent with schemes that rely on pitch classification and is sufficiently general to support the use of additional variables such as spin rate. The new model is efficient and supports the real-time dissemination of intrinsic pitch values during games.

We use the Cronbach's alpha [8] estimate of reliability [7] [40] to show that intrinsic pitch values have a significantly higher internal consistency than outcome-based pitch values which enables more accurate predictive models. We further develop a method to combine intrinsic values at the individual pitch level into a statistic that captures the value of a pitcher's collection of pitches over a period of time. We use this statistic to show that pitchers who outperform their intrinsic values during a season tend to perform worse the following year. Since intrinsic values are based on physical measurements, the new statistics can be used to predict how a pitcher's collection of pitches will translate from other levels (amateur, minors, foreign leagues) to the MLB environment. By directly relating the physical properties of a pitch to expected performance, this approach also promises to improve our understanding of how pitcher skill varies with age.

2 Computing Pitch Intrinsic Values

2.1 Sensor Data

PITCHf/x [11] is a system that uses two cameras to capture a set of images of a pitch. The PITCHf/x images can be used to estimate the three-dimensional path of a pitch and to derive information about its speed and movement. Our analysis of PITCHf/x data considers several of the reported attributes for each pitch. The pair (l_x, l_z) specifies the location of a pitch as it crosses home plate where l_x is the horizontal coordinate and l_z is the vertical

coordinate relative to an origin at the back vertex of home plate. The positive x-axis points to the right from the catcher's perspective, the positive y-axis points toward second base, and the positive z-axis points up. The coordinates l_x and l_z are typically reported in feet. The movement of a pitch (b_x, b_z) is defined as the difference between the pitch location (l_x, l_z) and the theoretical location of a pitch thrown at the same speed that does not deviate from a straight path due to spin [30]. The movement parameters b_x and b_z are typically reported in inches. The value s is an estimate of pitch speed in three dimensions near the release point in miles per hour. Since each batter has a unique strike zone in the z dimension, we transform l_z before processing into a coordinate system with a standard strike zone. The transform maps a pitch at the top or bottom of the batter's individual strike zone to the top or bottom of the standard strike zone. The batter's individual zone is stretched accordingly to map to the standard zone while pitches above or below the batter's individual zone by a vertical distance Δz are mapped to a transformed $z-$coordinate that is Δz above or below the standard zone.

The HITf/x system [24] uses the PITCHf/x images to estimate the initial speed and direction of batted balls in three dimensions. The direction is specified by two angles. The vertical launch angle is the angle that the batted ball's initial velocity vector makes with the plane of the playing field and the horizontal spray angle specifies the direction of the projection of the batted ball's initial velocity vector onto the plane of the playing field. The wOBA cube model [17] [18] is used to specify an intrinsic value for batted balls at contact using the measured exit speed, vertical angle, and horizontal angle as depicted in figure 1. This intrinsic value is independent of variables that include the defense, ballpark, and weather conditions.

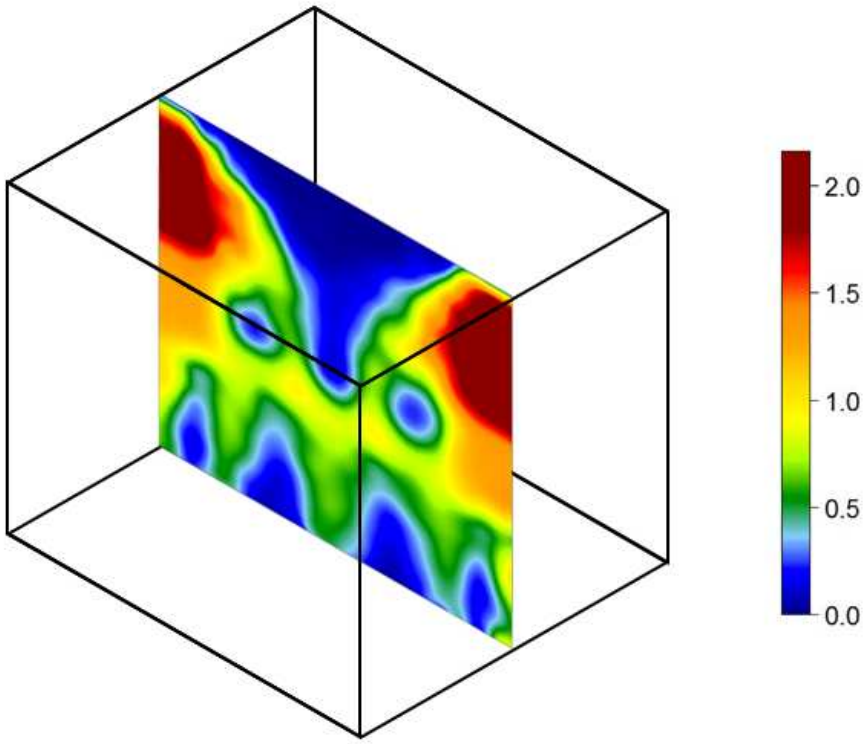

Figure 1: wOBA Cube

2.2 Bayesian Foundation

We develop a method for learning the dependence of a pitch's intrinsic value on its measured parameters. Using Bayes theorem, the posterior probability of an outcome R_j given a measured pitch vector $v = (s, b_x, b_z, l_x, l_z)$ is given by

$$P(R_j|v) = \frac{p(v|R_j)P(R_j)}{p(v)} \tag{1}$$

where $p(v|R_j)$ is the conditional probability density function for v given outcome R_j, $P(R_j)$ is the prior probability of outcome R_j, and $p(v)$ is the probability density function for v. We consider the six possible outcomes R_0 = ball in play, R_1 = called ball, R_2 = called strike, R_3 = swinging strike, R_4 = foul ball, and R_5 = batter hit-by-pitch where foul tips that are caught for strikeouts are classified as R_3 and not R_4. Our analysis will model the dependence of each of the factors in (1) on the count and the platoon configuration. We show in section 2.6 that a weighted sum of the $P(R_j|v)$ values over outcomes provides a measure of the intrinsic value of a pitch.

2.3 Kernel Density Estimation

The goal of density estimation for our application is to recover the conditional probability density functions $p(v|R_j)$ in equation (1) from a set of measured pitch vectors and their outcomes. Let v_i for $i = 1, 2, \ldots, n$ be a set of n five-dimensional measured pitch vectors with outcome R_j. Kernel methods [36] which are also known as Parzen-Rosenblatt [31] [35] window methods are widely used for nonparametric density estimation. A kernel density estimate for $p(v|R_j)$ is given by

$$\widehat{p}(v|R_j) = \frac{1}{n}\sum_{i=1}^{n} K(v - v_i) \tag{2}$$

where $K(\cdot)$ is a kernel probability density function that is typically unimodal and centered at zero. A standard kernel for approximating a $d-$dimensional density is the zero-mean Gaussian

$$K(v) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}v^T\Sigma^{-1}v\right] \tag{3}$$

where Σ is the $d \times d$ covariance matrix. For this kernel, $\widehat{p}(v|R_j)$ at any v is the average of a sum of Gaussians centered at the sample points v_i and the covariance matrix Σ determines the amount and orientation of the smoothing. Σ is often chosen to be the product of a scalar and an identity matrix which results in equal smoothing in every direction. To recover a more accurate approximation $\widehat{p}(v|R_j)$ the covariance matrix should allow different amounts of smoothing in different directions. We enable this goal while also reducing the number of unknown parameters by adopting a diagonal model for Σ with variance elements $(\sigma_s^2, \sigma_{b_x}^2, \sigma_{b_z}^2, \sigma_{l_x}^2, \sigma_{l_z}^2)$. For our five-dimensional data, this allows $K(v)$ to be written as a product of five one-dimensional Gaussians

$$K(v) = \frac{1}{(2\pi)^{5/2}\sigma_s\sigma_{b_x}\sigma_{b_z}\sigma_{l_x}\sigma_{l_z}} \exp\left[-\frac{1}{2}\left(\frac{s^2}{\sigma_s^2} + \frac{b_x^2}{\sigma_{b_x}^2} + \frac{b_z^2}{\sigma_{b_z}^2} + \frac{l_x^2}{\sigma_{l_x}^2} + \frac{l_z^2}{\sigma_{l_z}^2}\right)\right] \tag{4}$$

which depends on the five unknown bandwidth parameters $\sigma_s, \sigma_{b_x}, \sigma_{b_z}, \sigma_{l_x}$, and σ_{l_z}. Optimal bandwidth parameters are learned using the process described in the next section.

2.4 Bandwidth Selection for Kernel Density Estimation

The accuracy of the kernel density estimate $\widehat{p}(v|R_j)$ is highly dependent on the choice of the bandwidth vector $\sigma = (\sigma_s, \sigma_{b_x}, \sigma_{b_z}, \sigma_{l_x}, \sigma_{l_z})$ [9]. The recovered $\widehat{p}(v|R_j)$ will be spiky for small values of the parameters and, in the limit, will tend to a sum of Dirac delta functions centered at the v_i data points as the bandwidths approach zero. Large bandwidths, on the other hand, can induce excessive smoothing which causes the loss of important structure in the estimate of $p(v|R_j)$. A number of bandwidth selection techniques have been proposed and a recent survey of methods and software is given in [16]. Many of these techniques are based on maximum likelihood estimates for $p(v|R_j)$ which select σ so that $\widehat{p}(v|R_j)$ maximizes the likelihood of the observed v_i data samples. Applying these techniques to the full set of observed data, however, yields a maximum at $\sigma = (0, 0, 0)$ which corresponds to the sum of delta functions result. To avoid this difficulty, maximum likelihood methods for bandwidth selection have been developed that are based on leave-one-out cross-validation [36].

The computational demands of leave-one-out cross-validation techniques are excessive for our data set. Therefore, we have adopted a cross-validation method which requires less computation. From the set of n observed v_i vectors for outcome R_j, we generate M disjoint subsets S_k of fixed size n_v to be used for validation. For each validation set S_k, we construct the estimate $\widehat{p}(v|R_j)$ using the $n - n_v$ vectors that are not in S_k as a function of the bandwidth vector $\sigma = (\sigma_s, \sigma_{b_x}, \sigma_{b_z}, \sigma_{l_x}, \sigma_{l_z})$. The optimal bandwidth vector $\sigma_k^* = (\sigma_{sk}^*, \sigma_{b_xk}^*, \sigma_{b_zk}^*, \sigma_{l_xk}^*, \sigma_{l_zk}^*)$ for S_k is the choice that maximizes the pseudolikelihood [10] [16] according to

$$\sigma_k^* = \arg\max_{\sigma} \prod_{v_i \in S_k} \widehat{p}(v_i|R_j) \tag{5}$$

where the product is over the n_v vectors in the validation set S_k. The overall optimized bandwidth vector $\sigma^*(j) = (\sigma_s^*(j), \sigma_{b_x}^*(j), \sigma_{b_z}^*(j), \sigma_{l_x}^*(j), \sigma_{l_z}^*(j))$ for the $\widehat{p}(v|R_j)$ density estimate is obtained by averaging the M vectors σ_k^*.

2.5 Estimating the Posterior Probability

An estimate for $P(R_j|v)$ can be derived from estimates of the quantities on the right side of equation (1). The density estimate $\widehat{p}(v|R_j)$ for each $p(v|R_j)$ is obtained using the kernel

method defined by equations (2) and (4). Each prior probability $P(R_j)$ is estimated as the fraction $\widehat{P}(R_j)$ of the pitches in the full data set with outcome R_j. The density $p(v)$ is estimated using

$$\widehat{p}(v) = \sum_{j=0}^{5} \widehat{p}(v|R_j)\widehat{P}(R_j) \tag{6}$$

where the sum is over the six possible outcomes given in section 2.2. The estimate for $P(R_j|v)$ is then constructed by combining the estimates for $p(v|R_j), P(R_j)$, and $p(v)$ according to (1).

2.6 Intrinsic Values

In this section we present a method to compute the intrinsic value of a pitch. Define the context during a plate appearance (PA) as the platoon configuration and the number of balls b and strikes s on the batter. The context value is the league average wOBA [37] for PAs completed after the context. Let the pre-pitch value be the context value before a pitch. A pitch either completes a PA giving a post-pitch value defined by the wOBA coefficient for the PA result or the pitch causes a transition to a new context whose value defines the post-pitch value. The observed value O of a pitch is the difference between the post-pitch value and the pre-pitch value. This approach is used [2] to compute pitch values based on observed outcomes. Statistics that are based on observed pitch values depend on factors such as the catcher, the umpire, the defense, and the ballpark and have been shown to have a low degree of repeatability [33].

The posterior probabilities $P(R_j|v)$ can be used to define pitch intrinsic values. Let weight w_j denote the post-pitch value minus the pre-pitch value for a pitch with outcome R_j. The weights w_1, w_2, w_3, w_4, and w_5 depend on the count (b, s) and the platoon configuration. Since batted balls can take a range of values, the weight w_0 also depends on the vector v of pitch parameters. We describe a method for estimating the batted ball weight function $w_0(b, s, v)$ in section 2.7. We define the intrinsic value of a pitch for a platoon configuration as

$$I(b, s, v) = w_0(b, s, v)P(R_0|v) + \sum_{j=1}^{5} w_j(b, s)P(R_j|v) \tag{7}$$

which measures pitch value as a function of the physical pitch parameters, the count, and the platoon configuration. Positive values of $I(b, s, v)$ favor the batter while negative values favor the pitcher. $I(b, s, v)$ is the expected value of a pitch with parameter vector v on count (b, s) for a given platoon configuration and is not dependent on factors such as the catcher, umpire, defense, or ballpark associated with the pitch.

2.7 Estimating the Batted Ball Weight Function

The batted ball weight function $w_0(b, s, v)$ for a platoon configuration is estimated using nonparametric regression [5]. Let v_i for $i = 1, 2, \ldots, n_b$ be a set of n_b five-dimensional pitch vectors on count (b, s) for a platoon configuration that result in a batted ball (R_0) outcome. For each v_i, let y_i be the expected wOBA for the batted ball minus the pre-pitch value for the pitch that resulted in the batted ball. The expected wOBA is computed using the wOBA cube method [17] [18] from the batted ball exit speed, vertical angle, and horizontal angle as measured by the HITf/x system. The estimated $w_0(b, s, v)$ function is given by

$$\widehat{w}_0(b, s, v) = \frac{\sum_{i=1}^{n_b} K(v - v_i) y_i}{\sum_{i=1}^{n_b} K(v - v_i)} \tag{8}$$

where $K(\cdot)$ is a kernel function. For this work, we use the Gaussian kernel specified by (4) with cross-validation used for bandwidth selection. The n_b observed vectors v_i with a batted ball outcome and the associated y_i values are used to generate S_k validation sets. In this case, the optimal bandwidth vector σ_k^* is the choice that minimizes the sum of the absolute errors

$$\sigma_k^* = \arg\min_{\sigma} \sum_{v_i \in S_k} |y_i - \widehat{w}_0(b, s, v_i)| \tag{9}$$

where the sum is over the vectors in the validation set.

3 Visualizing Intrinsic Values

Data acquired by Sportvision's PITCHf/x and HITf/x systems during every regular-season MLB game in 2014 were used for this study. Figures 2 through 9 demonstrate properties and implications of pitch intrinsic values. Since I is a function of the count and the five

pitch parameters, we can examine the variation of intrinsic values along various dimensions while keeping other variables fixed. Figure 2 displays pitch intrinsic values for an 0-0 count on the (l_x, l_z) plane as viewed from the catcher's perspective for a pitch speed of $s = 90$ mph, horizontal movement $b_x = -3$ inches, and vertical movement $b_z = 6$ inches. Pitches with parameters near these values are typically classified as four-seam fastballs. We see that the locations with the smallest run value for these pitches are down-and-away within the strike zone. We also see that the locations with the largest run values are for pitches that are out of the strike zone which are often taken for balls.

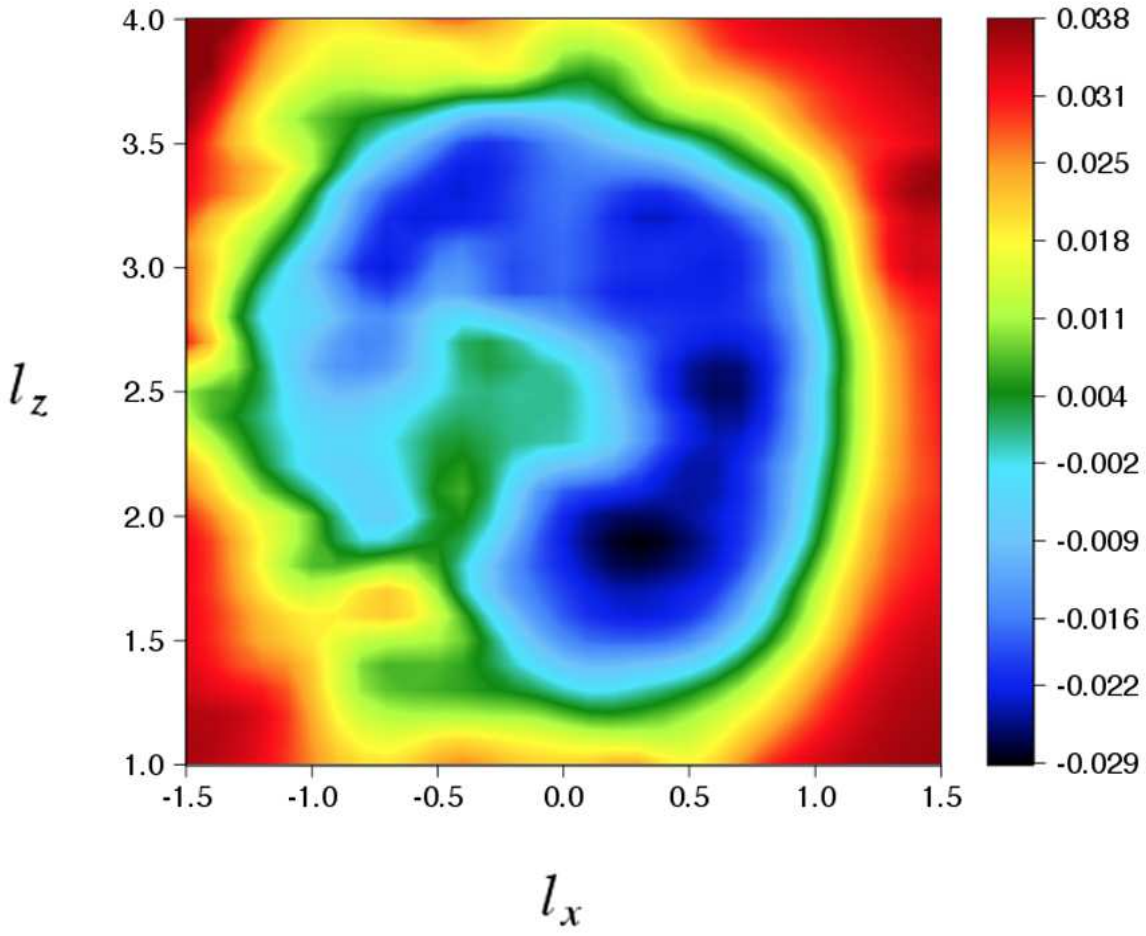

Figure 2: Pitch Intrinsic Value for 0-0 count, $s = 90, b_x = -3, b_z = 6$

Figures 3 through 6 illustrate how intrinsic values depend on the count. In these figures, we consider the pitch depicted in figure 2 for a vertical location near the center of the strike zone at $l_z = 2.5$. Figure 3 plots the probability of a swing as a function of the horizontal

location l_x for 0-0 and 0-1 counts. We see that the swing probability is higher for all values of l_x on 0-1 counts and that swing probability tends to be highest for pitches near the center of the strike zone for both counts. From equation (7), a pitch intrinsic value is the sum of components associated with the six R_j outcomes. Figure 4 plots the value of the ball in play $I_{bip} = w_0(b, s, v)P(R_0|v)$ component as a function of l_x for 0-0 and 0-1 counts. We see that I_{bip} is largest for pitches in the middle/inside part of the strike zone for pitches with these parameters. I_{bip} tends to be larger for an 0-1 count because there are more swings (figure 3) which increases $P(R_0|v)$ and also because $w_0(b, s, v)$ tends to be larger for balls in play on 0-1 since the pre-pitch value is smaller for an 0-1 count. Figure 5 plots the $I_{ball} = w_1(b, s)P(R_1|v)$ component as a function of l_x for 0-0 and 0-1 counts. We see that as we move away from the middle of the zone I_{ball} is larger for an 0-0 count because there are fewer swings (figure 3) on 0-0 which leads to larger values of the ball probability $P(R_1|v)$ and because the value of a ball ($w_1(0, 0) = .038$) on 0-0 is larger than the value of a ball ($w_1(0, 1) = .028$) on 0-1 for this platoon configuration.

Figure 6 plots pitch intrinsic value I for 0-0 and 0-1 counts. The shape of these curves is largely determined by the structure of the I_{bip} and I_{ball} functions plotted in figures 4 and 5. The pitch locations that minimize run value are near the edges of the strike zone whereas the pitch locations that maximize run value are near the middle/inside part of the strike zone or for pitches that are well outside the strike zone. The minima on both edges of the plate for the 0-1 count are more distant from the center of the zone than for the 0-0 count since the batter is more likely to swing at borderline pitches on an 0-1 count as shown in figure 3. Figure 7 plots the density of pitches thrown with the parameters considered in figure 2 as a function of l_x for 0-0 and 0-1 counts. Since the batter is more likely to swing at a given pitch on 0-1 and the cost of a ball is less on 0-1, pitchers throw fewer pitches near the center of the plate and more pitches out of the zone on 0-1.

Figure 8 illustrates the dependence of I on the pitch speed s for the previously considered movement parameters ($b_x = -3, b_z = 6$) with $l_z = 2.5$ on an 0-0 count. We see that for pitches in the strike zone the pitcher benefits from increased velocity. There is an inversion region where increased velocity benefits the batter near the inner edge of the strike zone which is due to umpires calling a smaller zone for higher velocity pitches with these parameters near the inner edge of the plate.

Figure 9 shows the dependence of I on pitch vertical movement b_z for $s = 90$ mph with $l_z = 2.5$ and $b_x = -3$ for an 0-0 count. We see that increasing vertical movement lowers the run value across the range of l_x.

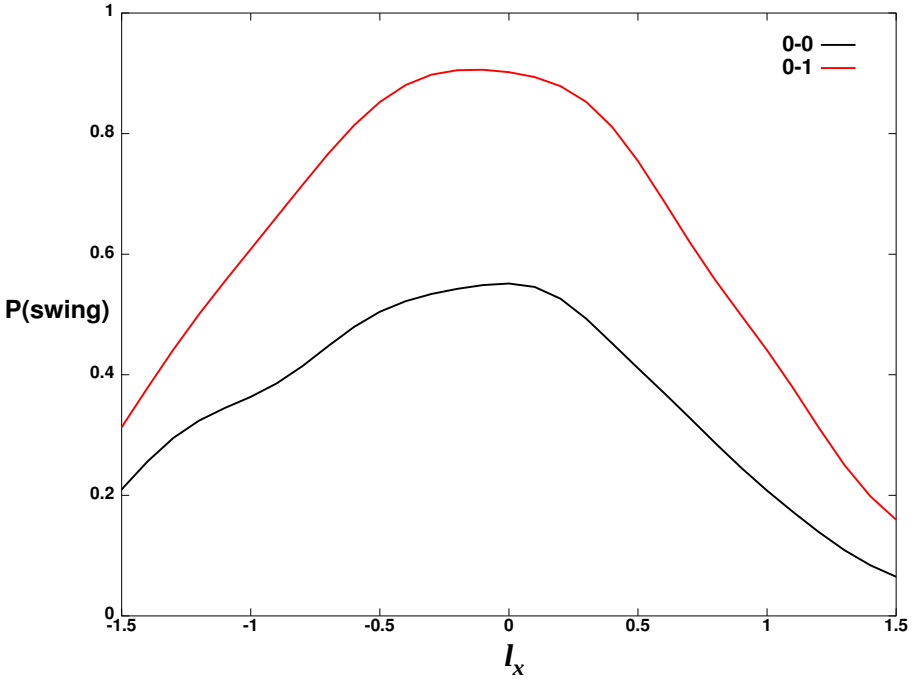

Figure 3: Dependence of swing probability on l_x for $s = 90, l_z = 2.5, b_x = -3, b_z = 6$

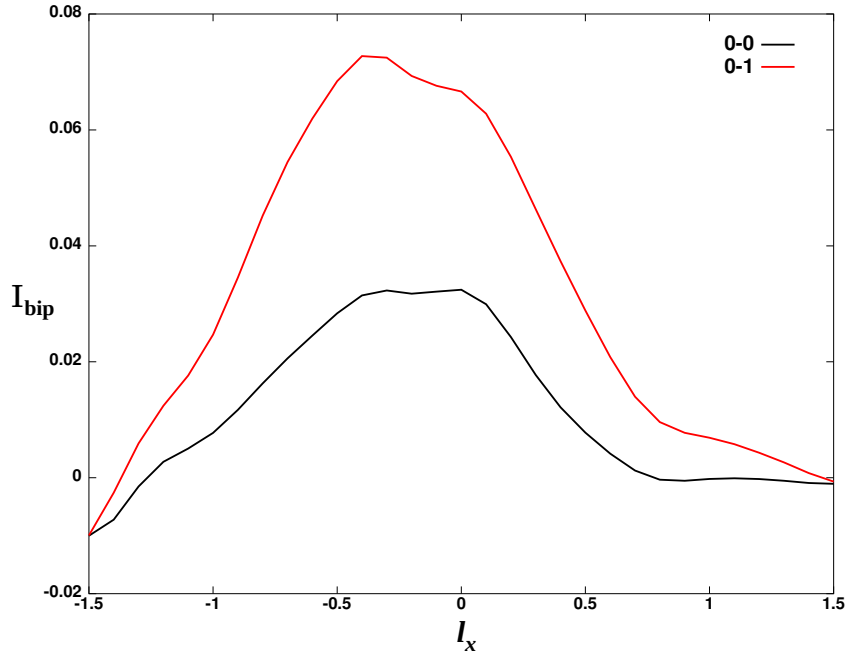

Figure 4: Dependence of BIP value component on l_x for $s = 90, l_z = 2.5, b_x = -3, b_z = 6$

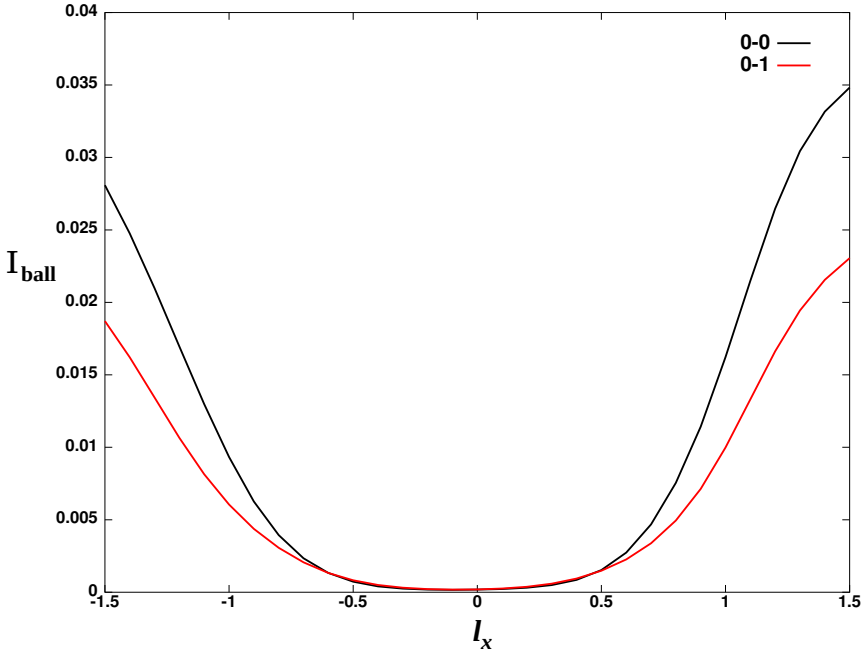

Figure 5: Dependence of ball value component on l_x for $s = 90, l_z = 2.5, b_x = -3, b_z = 6$

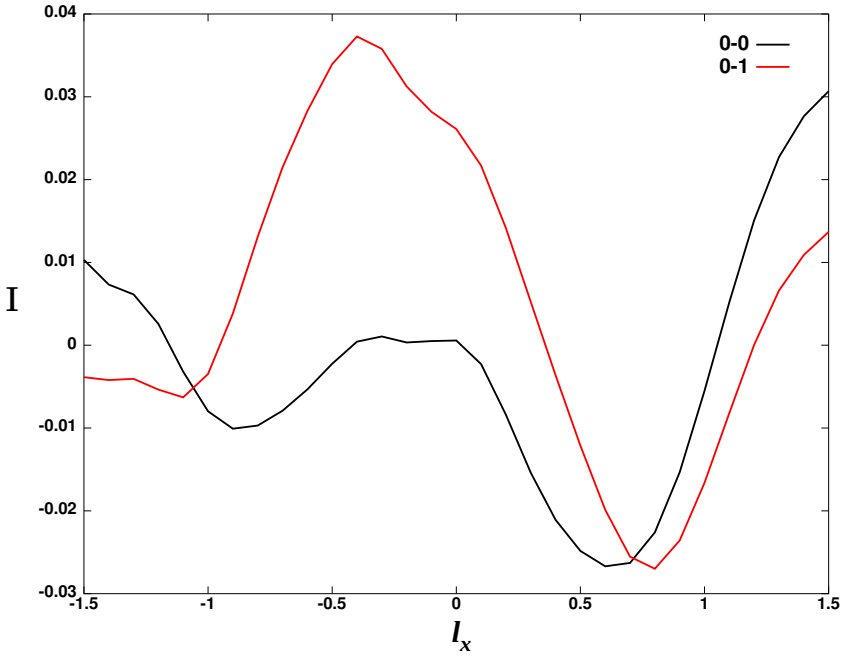

Figure 6: Dependence of pitch intrinsic value I on l_x for $s = 90, l_z = 2.5, b_x = -3, b_z = 6$

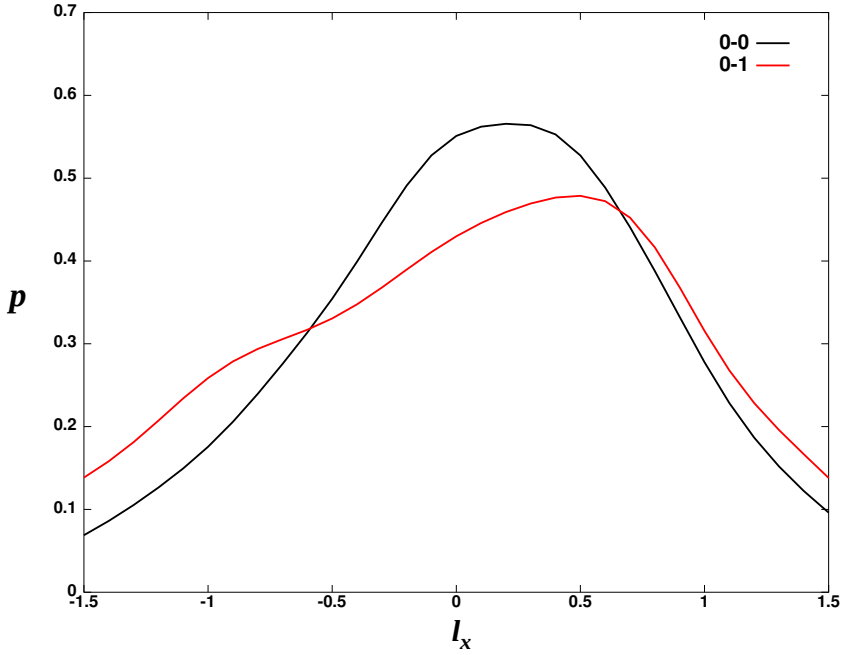

Figure 7: Dependence of pitch density on l_x for $s = 90, l_z = 2.5, b_x = -3, b_z = 6$

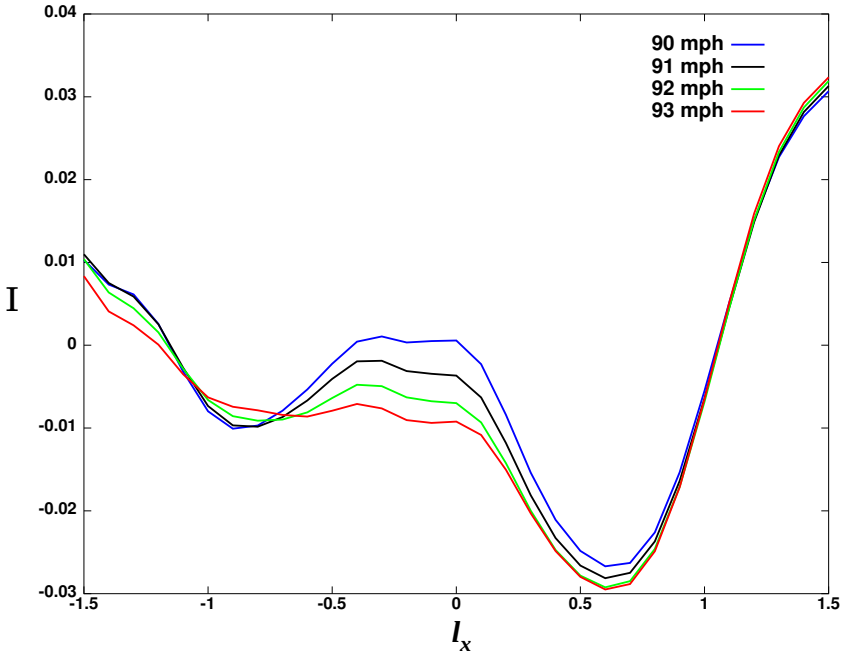

Figure 8: Dependence of I on s and l_x for 0-0 count, $l_z = 2.5, b_x = -3, b_z = 6$

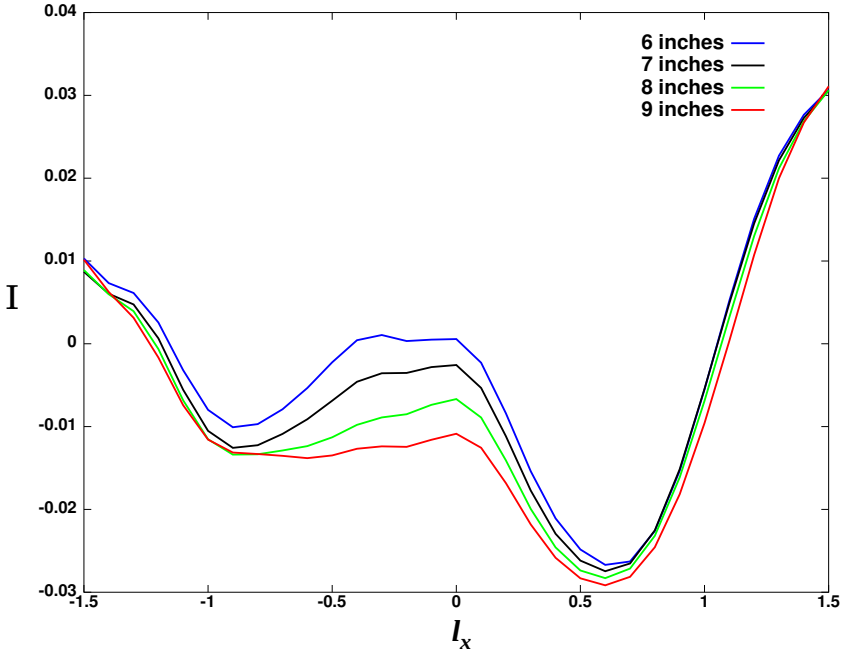

Figure 9: Dependence of I on b_z and l_x for 0-0 count, $s = 90, l_z = 2.5, b_x = -3$

4 Reliability

We use reliability estimates to demonstrate that the new intrinsic pitch values have a higher degree of repeatability than outcome-based pitch values. Reliability [40] is based on the premise that the measurement of an attribute is equal to true talent, which is the player's expected value for the measurement, plus random error. In the context of assessing a pitcher's skill level on pitches, the reliability of a measurement M over a sample of N pitches is defined by

$$R(N) = \frac{\sigma_t^2}{\sigma_o^2(N)} \tag{10}$$

where σ_t^2 is the variance of true talent across pitchers for M and $\sigma_o^2(N)$ is the variance of the observed values across pitchers for M as a function of the sample size. $R(N)$ quantifies the degree to which the measurement is repeatable and, therefore, is inversely related to the amount of random error in the measurement. In the context of forecasting, reliability determines how much the observed value of a measurement should be regressed in the direction of the mean to estimate true talent [18]. Measurements with a higher reliability require less regression and provide more accurate forecasts [19].

Split-half methods are a popular way to estimate reliability. These methods partition a data set into two halves and compute the correlation of the player measurements across the halves. A limitation of using split-half methods is that the estimated $R(N)$ can change depending on how the data is partitioned. An alternative approach is to compute Cronbach's alpha [8] which is an estimate of $R(N)$ that is an approximation to the average of all possible split-half correlations that would be computed from a full data set with $2N$ pitches for each player.

We used Cronbach's alpha ($\alpha(N)$) to estimate reliability for measurements defined by the average of the I pitch values and the average of the O pitch values over a set of N pitches. Figure 10 plots $\alpha(N)$ for these measurements for pitches thrown on an 0-0 count by a RHP to a RHB. The analysis considers the 116 pitchers who threw at least 200 tracked pitches in this configuration in 2014. For values of N ranging from 20 to 200 we computed $\alpha(N)$ for the I and O measurements using the first N of these pitches for each of the 116 pitchers. Figure 11 is the corresponding plot for the 112 RHP who threw at least 200 tracked

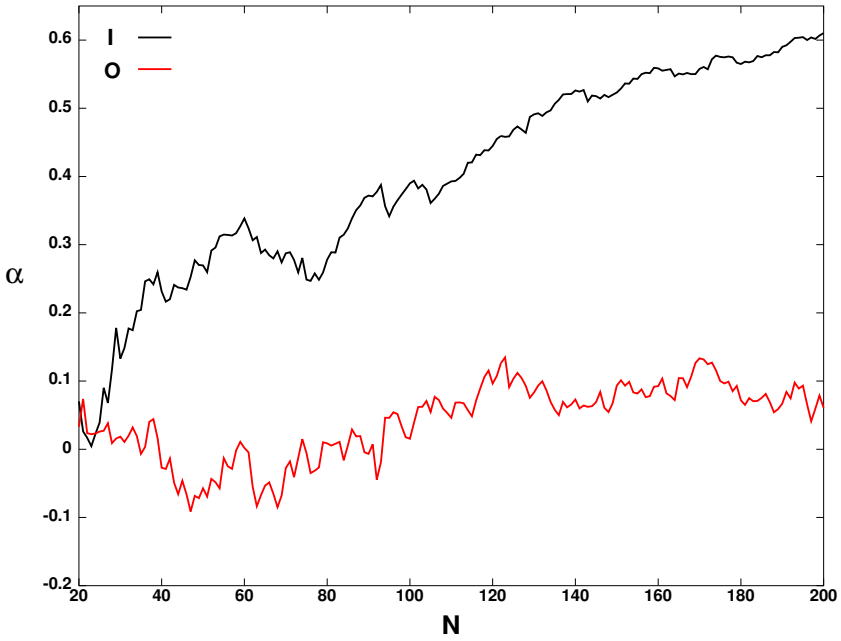

Figure 10: $\alpha(N)$ reliability estimate for RHP vs. RHB, 0-0 count

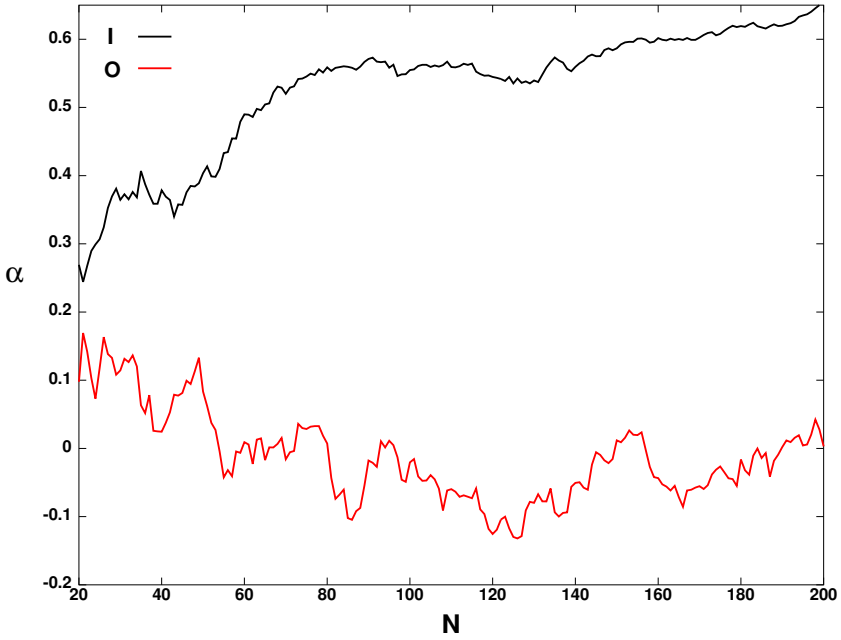

Figure 11: $\alpha(N)$ reliability estimate for RHP vs. LHB, 0-0 count

pitches on an 0-0 count to LHB. We see that the intrinsic pitch measurements have a significantly higher reliability than the observed pitch measurements. The intrinsic I measurement reaches 0.5 at 135 pitches for the configuration in figure 10 and at 65 pitches for the configuration in figure 11. The observed O measurement has relatively small values of $\alpha(N)$ and obtains negative values which can occur when $\alpha(N)$ is computed using small samples for measurements with low internal consistency.

Table 1 summarizes the reliability estimates for configurations involving right-handed pitchers for the second pitch of a plate appearance. For these cases, we considered pitchers who threw at least 100 pitches per count within a platoon configuration in 2014. The third column of the table indicates the number of pitchers who satisfied this criterion. The last two columns of the table give $\alpha(N)$ for a sample size of $N = 100$ pitches for the I and O measurements. We see that in each case, the estimated reliability for the I measurement is between 0.40 and 0.45 and is significantly larger than the corresponding reliability for the O measurement.

Table 1: $\alpha(N)$ for $N = 100$ for the I and O measurements

Configuration	count	# pitchers	$\alpha(100)$ for I	$\alpha(100)$ for O
RHP vs. RHB	0-1	117	0.45	0.18
RHP vs. RHB	1-0	82	0.42	0.20
RHP vs. LHB	0-1	108	0.40	-0.12
RHP vs. LHB	1-0	97	0.42	0.02

5 Intrinsic Pitch Statistics

In section 2.6 we described a method for computing the observed O and intrinsic I value of an individual pitch. In this section, we define statistics that summarize the observed and intrinsic value of the set of pitches thrown by a pitcher over a period of time. Consider a right-handed pitcher P who faces B_R right-handed batters and throws n_i pitches to the ith of these batters. Let $O_R(i, j)$ be the observed value of the jth pitch to the ith batter. For the plate appearance by the ith batter the sum of the observed pitch values

$$\sum_{j=1}^{n_i} O_R(i,j) \tag{11}$$

is equal to the wOBA coefficient for the outcome of the plate appearance minus the league average wOBA for the RHP vs RHB platoon configuration. We define pitcher P's observed pitch statistic O_R versus RHB as the sum of the $O_R(i,j)$ over all batters and pitches thrown within the platoon configuration divided by the number of batters faced

$$O_R = \frac{1}{B_R} \sum_{i=1}^{B_R} \sum_{j=1}^{n_i} O_R(i,j). \tag{12}$$

Thus, O_R is equal to the wOBA allowed by pitcher P against RHB minus the league average wOBA for RHP vs RHB. We will also find it convenient to write

$$O_R = \frac{1}{B_R} \sum_{c_i} N_R(c_i) \overline{O}_R(c_i) \tag{13}$$

where $N_R(c_i)$ is the number of pitches thrown by pitcher P in count c_i to RHB and $\overline{O}_R(c_i)$ is the average observed value of these pitches in count c_i. The sum is over the twelve possible counts c_i.

If we repeat the process for left-handed batters (LHB) to obtain O_L for pitcher P, we can define the overall observed pitch statistic O_P for pitcher P as

$$O_P = \frac{B_R O_R + B_L O_L}{B_R + B_L} \tag{14}$$

where B_L is the number of LHB faced by pitcher P. O_P can also be computed as the sum of the observed pitch values against all batters divided by the total number of batters faced.

In a similar way, we can use intrinsic pitch values to compute the intrinsic pitch statistic I_P for pitcher P. Unlike observed values, intrinsic values do not depend on factors such as the catcher, the umpire, the fielders, or the ballpark. Thus, intrinsic values should be more indicative of pitch quality than observed values. Let $\overline{I}_R(c_i)$ be the average intrinsic value of pitches thrown by P to RHB on count c_i. Instead of using the actual number $N_R(c_i)$ of pitches thrown in each count which depends on variables such as the catcher's framing ability, we will use the expected number of pitches thrown in each count. In sections 2.2, 2.3, and 2.5 we showed how to compute the probability $P(R_j|v)$ of each outcome R_j for a pitch with measured parameter vector v. Thus, for a given pitch, we can compute the

probability that the plate appearance ends on that pitch as well as the probability that the count transitions to a given new count or remains the same in the case of a two-strike foul ball. For each 0-0 pitch, for example, we can compute the probability that the plate appearance ends, the count moves to 1-0, or the count moves to 0-1. Considering all 0-0 pitches, we compute $N'_R(1,0)$ and $N'_R(0,1)$ which are the expected number of 1-0 and 0-1 pitches for the pitcher against RHB. Starting from $N'_R(1,0)$ and $N'_R(0,1)$ we continue the process to compute the expected number of pitches $N'_R(c_i)$ for each count c_i for pitcher P against RHB. Following (13), we then define the intrinsic pitch statistic for P against RHB as

$$I_R = \frac{1}{B_R} \sum_{c_i} N'_R(c_i) \overline{I}_R(c_i) \tag{15}$$

where $\overline{I}_R(c_i)$ is the average intrinsic value of the pitches thrown by pitcher P to RHB on count c_i. If we repeat the process for LHB to obtain I_L, we can define the overall intrinsic pitch statistic I_P for pitcher P as

$$I_P = \frac{B_R I_R + B_L I_L}{B_R + B_L}. \tag{16}$$

Given only the 2014 data set, the five-dimensional space of v vectors is too sparse to compute accurate kernel density estimates for the LHP vs LHB configuration and for deep counts involving right-handed pitchers. Thus, Table 2 presents the RHP with the lowest I_P for 2014 after restricting the analysis to the first two pitches of plate appearances ($c_i = \{(0,0),(0,1),(1,0)\}$). We see that Phil Hughes easily posted the lowest I_P which is not surprising given that he also enjoyed the highest strikeout-to-walk ratio ever recorded by a major league ERA qualifier.

Table 3 presents the five pitchers with the smallest $O_P - I_P$ differences. These pitchers significantly outperformed the intrinsic value of their pitches. We see that each of these pitchers had a much higher ERA the following year except for Adam Wainwright who was limited to 28 innings in 2015 after suffering an injury in April. In addition, the 2014 ERA for each of these pitchers is the best of their career through 2016 except for Wainwright's small-sample ERA in 2015.

Table 2: RHP with the lowest I_P over at least 400 batters faced, 2014

Pitcher	$I_P * 1000$
Phil Hughes	-19.1
John Lackey	-13.5
Zack Greinke	-12.6
Jordan Zimmermann	-12.5
Anibal Sanchez	-10.7
Jacob deGrom	-10.3
Colby Lewis	-10.3
Hisashi Iwakuma	-10.2
Bartolo Colon	-10.0
Adam Wainwright	-9.8

Table 3: RHP with the lowest $O_P - I_P$ over at least 400 batters faced, 2014

Pitcher	$(O_P - I_P) * 1000$	2014 ERA	2015 ERA
David Buchanan	-18.9	3.75	6.99
Carlos Carrasco	-15.3	2.55	3.63
Edinson Volquez	-13.3	3.04	3.55
Felix Hernandez	-13.3	2.14	3.53
Adam Wainwright	-11.5	2.38	1.61

6 Future Work

While we have shown that statistics based on pitch intrinsic values have a number of desirable properties, a pitcher's success depends on several additional factors. Pitch diversity affects performance as experiments have shown, for example, that contact rates degrade significantly as pitches are drawn from a wider range of speeds [14]. Other studies have shown that major league strikeout rates increase as a pitcher's number of distinct pitch types increases [3] and that pitchers who throw a high fraction of fastballs suffer a larger decline in performance when they face batters multiple times in a game [27]. Studies [4] [13] [15] [23] [34] have also shown that effective pitch sequencing can be used to obtain an advantage. Another important aspect of pitching is the use of a game plan that accounts for each batter's strengths/weaknesses and the computation of pitch intrinsic values is an important first step in the automated generation of matchup models [20] [21]. In addition to

a pitch's physical parameters, a pitcher's delivery can also affect results if, for example, he hides the ball well or inadvertently provides clues about the identity of an upcoming pitch. By accurately quantifying pitch intrinsic values, we have a framework that will enable the careful study of other factors that affect pitching success.

Acknowledgment

I am grateful to Sportvision and MLB Advanced Media for providing the HITf/x data which made this work possible. I also thank Tom Tango and Mitchel Lichtman for helpful comments on a previous draft of this manuscript. I am happy to acknowledge the assistance of Qi Shi in the preparation of this document.

References

[1] D. Allen. (Mar. 16, 2009). Run value by pitch location [Online]. Available: baseballanalysts.com/archives/2009/03/run_value_by_pi.php.

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

[3] R. Arthur. (Feb. 6, 2014). Entropy and the eephus [Online]. Available: www.baseball.prospectus.com/article.php?articleid=22758.

[4] P. Bonney. (Mar. 6, 2015). Defining the pitch sequencing question [Online]. Available: www.hardballtimes.com/defining-the-pitch-sequencing-question.

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

[6] C. Burley. (Oct. 15, 2004). The importance of strike one (and two, and three ...), part 2 [Online]. Available: www.hardballtimes.com/the-importance-of-strike-one-part-two.

[7] R. Carleton. (Apr. 20, 2011). 525,600 minutes: how do you measure a player in a year? [Online]. Available: www.fangraphs.com/blogs/525600-minutes-how-do-you-measure-a-player-in-a-year.

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

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

[10] R. Duin. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25(11):1175–1179, 1976.

[11] 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.

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

[13] C. Glaser. (Mar. 4, 2010). The influence of batters' expectations on pitch perception [Online]. Available: www.hardballtimes.com/tht-live/the-influence-of-batters-expectations-on-pitch-perception.

[14] 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.

[15] J. Greenhouse. (May 20, 2010). Lidge's pitches [Online]. Available: baseballanalysts.com/archives/2010/05/brad_lidges_out.php.

[16] A.C. Guidoum. Kernel estimator and bandwidth selection for density and its derivatives. The kedd package, version 1.03, October 2015.

[17] G. Healey. (Mar. 17, 2016). The intrinsic value of a batted ball [Online]. Available: www.hardballtimes.com/the-intrinsic-value-of-a-batted-ball.

[18] G. Healey. (Aug. 2, 2016). The reliability of intrinsic batted ball statistics [Online]. Available: www.hardballtimes.com/the-reliability-of-intrinsic-batted-ball-statistics.

[19] G. Healey. (July 26, 2016). The reliability of intrinsic batted ball statistics: Appendix [Online]. Available: vixra.org/pdf/1607.0497v1.pdf.

[20] G. Healey. Modeling the probability of a strikeout for a batter/pitcher matchup. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2415–2423, September 2015.

[21] G. Healey. Matchup models for the probability of a ground ball and a ground ball hit. *Journal of Sports Analytics*, September 2016.

[22] G. Healey. Appendix to The Intrinsic Value of a Pitch, 2017.

[23] 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.

[24] P. Jensen. (Jun. 30, 2009). Using HITf/x to measure skill [Online]. Available: www.hardballtimes.com/using-hitf-x-to-measure-skill.

[25] J. Judge, H. Pavlidis, and D. Turkenkopf. (Apr. 29, 2015). Introducing deserved run average DRA and all its friends [Online]. Available: www.baseballprospectus.com/article.php?articleid=26195.

[26] 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.

[27] M. Lichtman. (Nov. 15, 2013). Pitch types and the times through the order penalty [Online]. Available: www.baseball.prospectus.com/article.php?articleid=22235.

[28] M. Marchi. (Dec. 4, 2009). Pitch run value and count [Online]. Available: www.hardballtimes.com/pitch-run-value-and-count.

[29] D. Meyer. (May 6, 2015). Dynamic run value of throwing a strike (instead of a ball) [Online]. Available: www.hardballtimes.com/dynamic-run-value-of-throwing-a-strike-instead-of-a-ball.

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

[31] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[32] H. Pavlidis and D. Brooks. (Mar. 3, 2014). Framing and blocking pitches: a regressed probabilistic model [Online]. Available: www.baseballprospectus.com/arti-cle.php?articleid=22934.

[33] Pitch type linear weights [Online]. Available: www.fangraphs.com/library/pitching/ linear-weights.

[34] J. Roegele. (Nov. 24, 2014). The effects of pitch sequencing [Online]. Available: www.hardballtimes.com/the-effects-of-pitch-sequencing.

[35] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837, 1956.

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

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

[38] J. Thorn and P. Palmer. *The Hidden Game of Baseball*. Doubleday and Company, New York, 1984.

[39] J. Walsh. (Feb. 26, 2008). Searching for the game's best pitch [Online]. Available: www.hardballtimes.com/searching-for-the-games-best-pitch.

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

Biography

Glenn Healey is Professor of Electrical Engineering and Computer Science at the University of California, Irvine. Before joining UC Irvine, he worked at IBM Research. Dr. Healey received the B.S.E. in Computer Engineering from the University of Michigan and the M.S. in computer science, the M.S. in mathematics, and the Ph.D. in computer science from Stanford University. He is director of the Computer Vision Laboratory at UC Irvine. Dr. Healey has served on the editorial boards of *IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing,* and the *Journal of the Optical Society of America A*. He has been elected a Fellow of IEEE and SPIE. Dr. Healey has received several awards for outstanding teaching and research.