The Intrinsic Value of a Batted Ball
Technical Details

Glenn Healey, EECS Department

University of California, Irvine, CA 92617

Given a set of observed batted balls and their outcomes, we develop a method for learning the dependence of a batted ball's intrinsic value on its measured s, v, and h parameters.

1 HITf/x Data

The HITf/x data used for this study was provided by SportVision and includes measurements from every regular-season MLB game during 2014. We consider all balls in play with a horizontal angle in fair territory ($h \in [-45°, 45°]$) that were tracked by the system where bunts are excluded. This results in a set of 124364 batted balls and the distributions for s, v, and h are shown in figures 1 and 2. We see that the peak of the speed distribution is near 93 mph and that the peaks of the vertical and horizontal angle distributions are near zero. Since HITf/x tracks batted balls over a portion of their trajectory that occurs after the ball has slowed due to air drag and gravity, the estimated speeds are a few miles per hour less than the speeds recorded by other systems. Since this effect is systematic, these offsets will not have a significant impact on the batted ball statistics computed in this work.

2 Learning Algorithm

2.1 Bayesian Foundation

Using Bayes theorem, the probability of a batted ball outcome R_j given a measured vector $x = (s, v, h)$ is given by

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

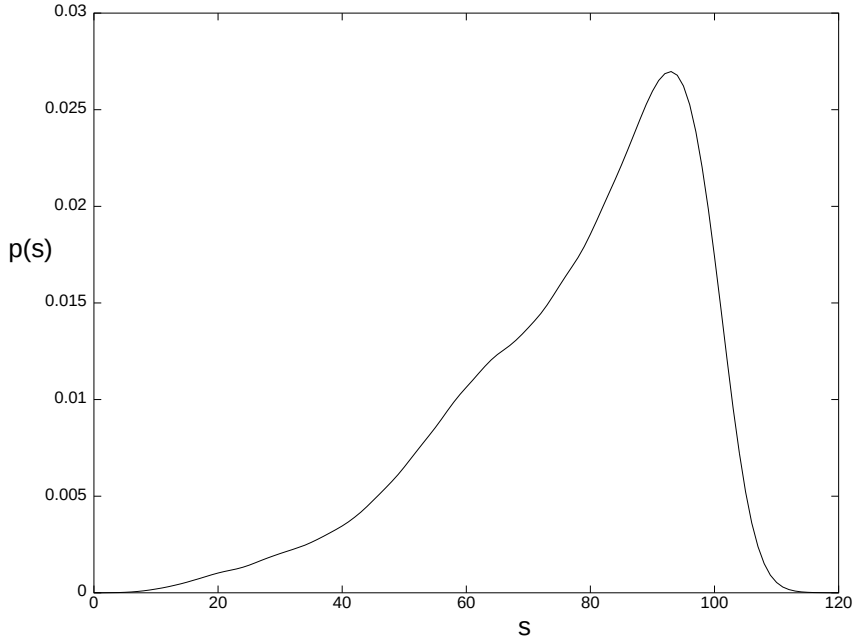

Figure 1: Distribution of initial speeds (mph) for batted balls in 2014

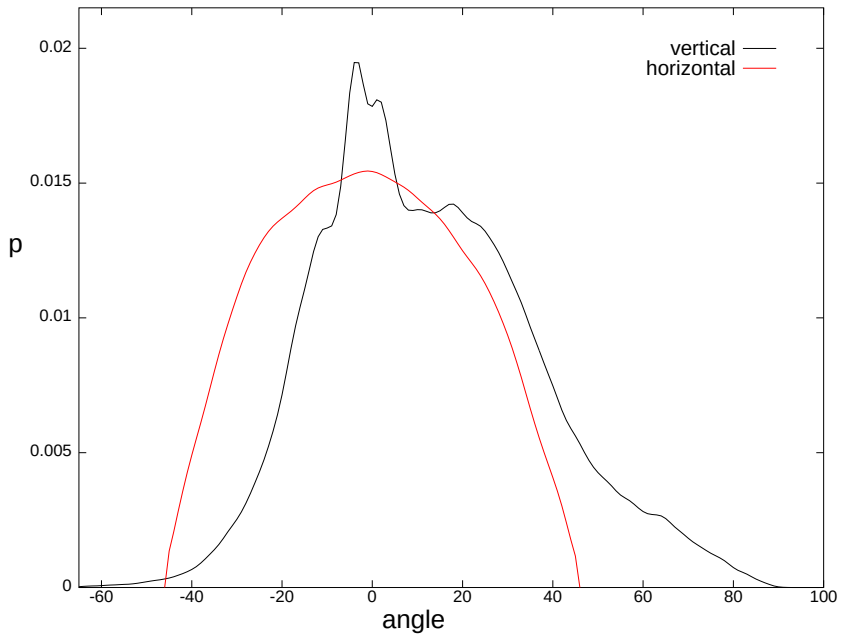

Figure 2: Distribution of vertical and horizontal angles (degrees) for batted balls in 2014

where $p(x|R_j)$ is the conditional probability density function for x given outcome R_j, $P(R_j)$ is the prior probability of outcome R_j, and $p(x)$ is the probability density function for x. Linear combinations of the $P(R_j|x)$ probabilities for different outcomes can be used to model the expected value of statistics such as batting average, wOBA, and slugging percentage as a function of the batted ball vector x. For a given batted ball, therefore, these statistics provide a measure of value that is separate from the batted ball's particular outcome.

2.2 Kernel Density Estimation

The goal of density estimation for this application is to recover the underlying probability density functions $p(x|R_j)$ and $p(x)$ in equation (1) from the set of observed batted ball vectors and their outcomes. Given the typical positioning of defenders on a baseball field and the various ways that an outcome such as a single can occur, we expect a conditional density $p(x|R_j)$ to have a complicated multimodal structure. Thus, we use a nonparametric technique for density estimation.

We first consider the task of estimating $p(x)$. Let $x_i = (s_i, v_i, h_i)$ for $i = 1, 2, \ldots, n$ be the set of n observed batted ball vectors. Kernel methods [6] which are also known as Parzen-Rosenblatt [4] [5] window methods are widely used for nonparametric density estimation. A kernel density estimate for $p(x)$ is given by

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i) \tag{2}$$

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

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

where Σ is the $d \times d$ covariance matrix. For this kernel, $\widehat{p}(x)$ at any x is the average of a sum of Gaussians centered at the sample points x_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. However, we see

from figures 1 and 2 that the distribution for v has detailed structure while the distributions for s and h are significantly smoother. Thus, to recover an accurate approximation $\widehat{p}(x)$ 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 $(\sigma_s^2, \sigma_v^2, \sigma_h^2)$. For our three-dimensional data, this allows $K(x)$ to be written as a product of three one-dimensional Gaussians

$$K(x) = \frac{1}{(2\pi)^{3/2}\sigma_s\sigma_v\sigma_h} \exp\left[-\frac{1}{2}\left(\frac{s^2}{\sigma_s^2} + \frac{v^2}{\sigma_v^2} + \frac{h^2}{\sigma_h^2}\right)\right] \tag{4}$$

which depends on the three unknown bandwidth parameters σ_s, σ_v, and σ_h.

2.3 Cross-Validation for Bandwidth Selection

The accuracy of the kernel density estimate $\widehat{p}(x)$ is highly dependent on the choice of the bandwidth vector $\sigma = (\sigma_s, \sigma_v, \sigma_h)$ [1]. The recovered $\widehat{p}(x)$ 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 x_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 $p(x)$. A number of bandwidth selection techniques have been proposed and a recent survey of methods and software is given in [3]. Many of these techniques are based on maximum likelihood estimates for $p(x)$ which select σ so that $\widehat{p}(x)$ maximizes the likelihood of the observed x_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 [6].

The computational demands of leave-one-out cross-validation techniques are excessive for our HITf/x data set. Therefore, we have adopted a cross-validation method which requires less computation. From the full set of n observed x_i vectors, we generate M disjoint subsets S_j of fixed size n_v to be used for validation. For each validation set S_j, we construct the estimate $\widehat{p}(x)$ using the $n - n_v$ vectors that are not in S_j as a function of the bandwidth vector $\sigma = (\sigma_s, \sigma_v, \sigma_h)$. The optimal bandwidth vector $\sigma_j^* = (\sigma_{sj}^*, \sigma_{vj}^*, \sigma_{hj}^*)$ for S_j

is the choice that maximizes the pseudolikelihood [2] [3] according to

$$\sigma_j^* = \arg\max_{\sigma} \prod_{x_i \in S_j} \widehat{p}(x_i) \tag{5}$$

where the product is over the n_v vectors in the validation set S_j. The overall optimized bandwidth vector σ^* is obtained by averaging the M vectors σ_j^*.

For our data set, we used five validation sets S_1, S_2, S_3, S_4, and S_5 to select the optimized bandwidth vector σ^* for the $p(x)$ estimate. Set S_i includes n_v batted balls that were hit on day $6i - 5$ of a calendar month. Set S_2, for example, includes only batted balls hit on the 7th day of a month. The size $n_v = 3820$ was taken to be the largest value so that each set S_i includes the same number of elements. The decision to use six days of separation for the validation sets was made with the goal of maximizing the independence of the sets. A regular-season series of consecutive games between the same pair of teams always lasts less than six days. In addition, major league teams in 2014 tended to use a rotation of starting pitchers that repeats every five days so that, if this tendency is followed, each starting pitcher will occur once per calendar month in each of the five validation sets.

For each validation set S_j, a three-dimensional search was conducted with a step size of 0.1 in σ_s, σ_v, and σ_h to find the optimized σ_j^* in equation (5). For each S_j and σ vector under consideration, we removed the twenty x_i batted ball vectors with the smallest value of $\widehat{p}(x_i)$ to prevent outliers from influencing the optimization. The vectors σ_j^* for each S_j are given in Table 1 and after averaging yielded an optimized $\sigma^* = (\sigma_s^*, \sigma_v^*, \sigma_h^*)$ of $(2.02, 1.50, 2.20)$. We see that vertical angle has the smallest smoothing parameter ($\sigma_v^* = 1.50$) which is consistent with the observation from figures 1 and 2 that vertical angle has more detailed structure in its density than batted ball speed or horizontal angle.

Table 1: Optimal bandwidths σ_j^* for validation sets S_j

S_1	S_2	S_3	S_4	S_5
(2.0,1.5,2.2)	(1.9,1.5,2.3)	(2.0,1.6,2.0)	(2.0,1.6,2.3)	(2.2,1.3,2.2)

2.4 Constructing the Estimate for $P(R_j|x)$

An estimate for $P(R_j|x)$ can be derived from estimates of the quantities on the right side of equation (1). The density estimate $\widehat{p}(x)$ for $p(x)$ is obtained using the kernel method defined by equations (2) and (4) with the optimized bandwidth vector σ^* learned using the process described in section 2.3. Each conditional probability density function $p(x|R_j)$ is estimated in the same way except that the training set is defined by the subset of the x_i vectors with outcome R_j. Since reduced sample sizes for specific outcomes R_j preclude the learning of individual bandwidth vectors for each $p(x|R_j)$, we use the σ^* derived for $p(x)$ for each case. This approach also has the desirable effect of providing the same smoothing to a batted ball vector in the numerator and denominator of (1) which prevents a probability $P(R_j|x)$ from exceeding one. Each prior probability $P(R_j)$ is estimated by the fraction of the n batted balls in the full training set with outcome R_j. The estimate for $P(R_j|x)$ is then constructed by combining the estimates for $p(x|R_j), P(R_j)$, and $p(x)$ according to Bayes theorem.

2.5 Batter Handedness

We repeated the process described in the previous sections to obtain separate densities for left-handed and right-handed batters. The $n = 124364$ batted balls were first partitioned into the 54948 for left-handed batters and 69416 for right-handed batters. The method described in section 2.3 was then used to build five validation sets for each case which resulted in a validation set size n_v of 1680 for left-handed batters and 2190 for right-handed batters. The optimal bandwidth vectors σ_j^* for each validation set and batter handedness are given in Table 2. After averaging, we arrive at an optimized $\sigma^* = (\sigma_s^*, \sigma_v^*, \sigma_h^*)$ of $(2.18, 1.72, 2.50)$ for left-handed batters and $(2.16, 1.56, 2.30)$ for right-handed batters. We note that, as seen in section 2.3, σ_v^* is the smallest for each case while σ_h^* is the largest. In addition, the bandwidth increases for each variable to provide more smoothing as the number of samples decreases.

Table 2: Optimal bandwidths σ_j^* for validation sets S_j by batter handedness

	S_1	S_2	S_3	S_4	S_5
L	(2.0,1.5,3.1)	(2.2,2.1,2.2)	(2.3,1.6,2.1)	(2.4,1.9,2.3)	(2.0,1.5,2.8)
R	(1.9,1.8,2.1)	(2.1,1.7,2.2)	(2.4,1.4,2.2)	(2.2,1.5,2.6)	(2.2,1.4,2.4)

References

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

[2] 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.

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

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

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

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