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

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

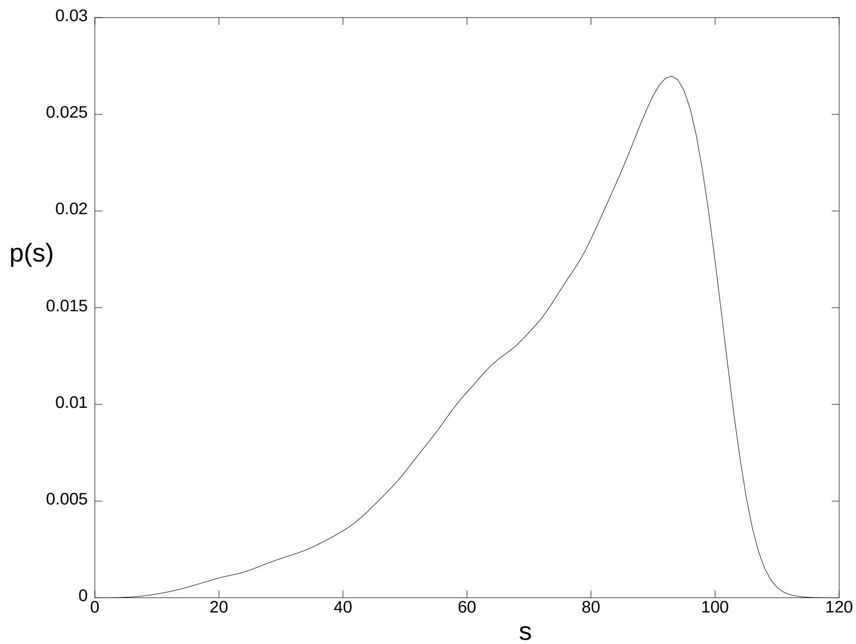

Figure 1: Distribution of initial speeds (mph) for batted balls in 2014

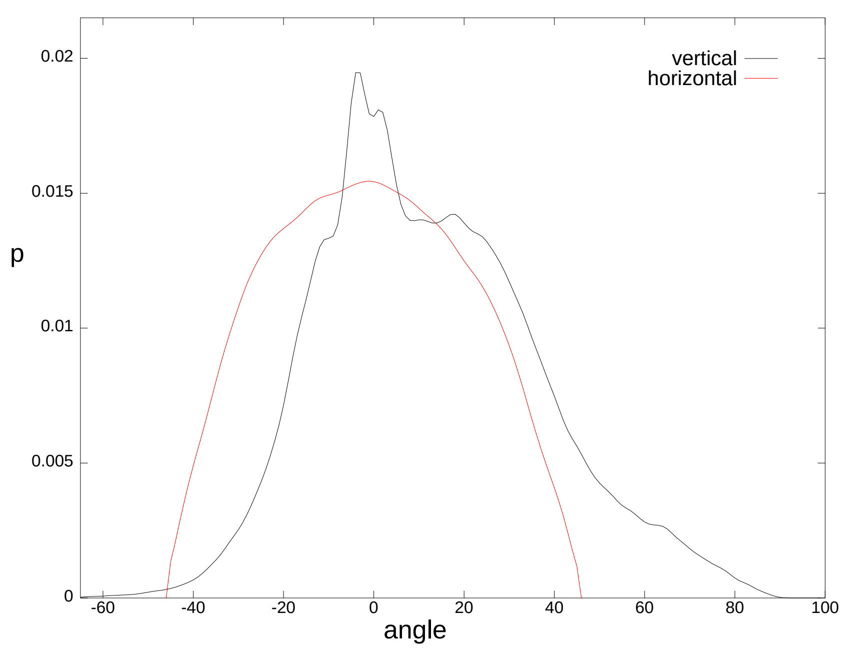

Figure 2: Distribution of vertical and horizontal angles (degrees) for batted balls in 2014

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

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

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.

Batter	I
Giancarlo Stanton	.526
Mike Trout	.498
Miguel Cabrera	.488
J. D. Martinez	.482
Matt Kemp	.477
Brandon Moss	.476
Jose Abreu	.469
Mike Morse	.468
Corey Dickerson	.465
Edwin Encarnacion	.461
Nelson Cruz	.459
Justin Upton	.454
Chris Carter	.454
Marlon Byrd	.446
Buster Posey	.443
David Ortiz	.442
Anthony Rizzo	.441
Marcell Ozuna	.439
Lucas Duda	.439
Jose Bautista	.438
Freddie Freeman	.436
Khris Davis	.426
Adrian Gonzalez	.426
Andrew McCutchen	.426
Ian Desmond	.426
Adam LaRoche	.425
Yan Gomes	.425
David Freese	.417
Jayson Werth	.416
Albert Pujols	.414
Victor Martinez	.413
Carlos Santana	.413
Starling Marte	.412
Todd Frazier	.411
Adam Jones	.410
Kyle Seager	.410
Matt Holliday	.410
Michael Brantley	.410
Carlos Gomez	.409
Matt Adams	.408
Starlin Castro	.407
Adrian Beltre	.406
Hanley Ramirez	.406
Billy Butler	.405
Ryan Howard	.404
Josh Donaldson	.404

Kole Calhoun	.400
Russell Martin	.400
Anthony Rendon	.400
Chase Headley	.399
Mark Teixeira	.399
Yoenis Cespedes	.398
Yasiel Puig	.397
Christian Yelich	.394
Dexter Fowler	.394
Nolan Arenado	.393
Joe Mauer	.392
Torii Hunter	.390
Garrett Jones	.389
David Wright	.389
Nick Castellanos	.388
Jhonny Peralta	.388
Justin Morneau	.387
Lonnie Chisenhall	.386
Seth Smith	.386
Jon Jay	.385
Luis Valbuena	.385
Chris Johnson	.385
Evan Longoria	.384
Robinson Cano	.383
Pablo Sandoval	.382
Yadier Molina	.382
Jacoby Ellsbury	.380
Ryan Braun	.379
Daniel Murphy	.379
Salvador Perez	.378
Alex Gordon	.377
Aramis Ramirez	.376
Jay Bruce	.376
Trevor Plouffe	.375
Alex Rios	.374
Howie Kendrick	.374
Jason Castro	.373
Martin Prado	.372
Curtis Granderson	.372
Jonathan Lucroy	.371
Josh Harrison	.371
Hunter Pence	.371
James Loney	.371
Brian Dozier	.371
Brett Gardner	.371
Asdrubal Cabrera	.369
Dioner Navarro	.368

Travis d'Arnaud	.368
Eric Hosmer	.367
Dayan Viciedo	.367
Wilin Rosario	.367
Neil Walker	.366
Melky Cabrera	.366
Nick Markakis	.365
Scooter Gennett	.365
Eduardo Escobar	.364
Brandon Phillips	.364
Carlos Beltran	.364
Miguel Montero	.363
Matt Carpenter	.361
Denard Span	.361
B. J. Upton	.361
Alejandro De Aza	.361
Austin Jackson	.359
J. J. Hardy	.359
Casey McGehee	.359
Jordy Mercer	.358
Charlie Blackmon	.358
Lorenzo Cain	.357
Matthew Joyce	.356
Chase Utley	.355
Jonathan Schoop	.355
Juan Lagares	.355
Angel Pagan	.354
Domonic Brown	.354
Desmond Jennings	.354
Gregor Blanco	.353
Brian McCann	.352
Kolten Wong	.352
Xander Bogaerts	.351
Brandon Crawford	.350
Matt Dominguez	.350
Aaron Hill	.350
Dustin Ackley	.349
Carlos Ruiz	.349
Shin-Soo Choo	.348
Jose Altuve	.348
Jimmy Rollins	.348
DJ LeMahieu	.347
Ian Kinsler	.345
Rajai Davis	.345
Ben Zobrist	.343
Leonys Martin	.343
Jed Lowrie	.341

Allen Craig	.340
Erick Aybar	.339
Dustin Pedroia	.339
Jose Reyes	.338
Conor Gillaspie	.338
Alexei Ramirez	.338
Yangervis Solarte	.336
Dee Gordon	.335
Brock Holt	.334
Jason Kipnis	.332
Gordon Beckham	.332
Jason Heyward	.331
Rougned Odor	.330
Gerardo Parra	.330
Michael Bourn	.330
Alcides Escobar	.329
Mike Moustakas	.328
Coco Crisp	.328
Adeiny Hechavarria	.328
Adam Eaton	.327
Nori Aoki	.325
Yunel Escobar	.322
Derek Jeter	.322
Ender Inciarte	.321
Alberto Callaspo	.319
Kurt Suzuki	.319
Omar Infante	.317
David Murphy	.314
Andrelton Simmons	.311
Elvis Andrus	.306
Alexi Amarista	.304
Ben Revere	.302
Jean Segura	.299
Billy Hamilton	.299
Zack Cozart	.285

Pitcher	I
Garrett Richards	.304
Anibal Sanchez	.309
Danny Duffy	.314
Chris Sale	.319
Matt Garza	.328
Dallas Keuchel	.329
Jarred Cosart	.329
Clayton Kershaw	.332
Alex Cobb	.336
Johnny Cueto	.337
Chris Archer	.339
Doug Fister	.340
Felix Hernandez	.341
Kyle Gibson	.342
Corey Kluber	.342
Tanner Roark	.345
Jake Arrieta	.346
Edinson Volquez	.347
Adam Wainwright	.347
Gio Gonzalez	.348
Lance Lynn	.348
Chris Tillman	.351
Carlos Carrasco	.351
Jacob deGrom	.352
Sonny Gray	.353
Vance Worley	.354
Rick Porcello	.355
Julio Teheran	.356
Francisco Liriano	.356
David Phelps	.356
Jon Lester	.356
Wily Peralta	.356
Jordan Zimmermann	.357
Charlie Morton	.357
John Danks	.357
Josh Collmenter	.358
Tyler Skaggs	.358
Masahiro Tanaka	.359
Andrew Cashner	.359
Alex Wood	.359
Yovani Gallardo	.359
R. A. Dickey	.360
Jose Quintana	.360
Roberto Hernandez	.360
James Shields	.360
Scott Kazmir	.360

Zack Greinke	.361
Hector Santiago	.361
David Price	.362
Jorge De La Rosa	.362
Kevin Correia	.362
Kyle Lohse	.363
Homer Bailey	.363
Trevor Bauer	.363
Max Scherzer	.363
Brad Hand	.364
Drew Hutchison	.364
David Buchanan	.364
Phil Hughes	.364
Jordan Lyles	.365
Hiroki Kuroda	.365
Chris Young	.365
Scott Feldman	.365
Tyson Ross	.365
Josh Beckett	.365
J. A. Happ	.367
Jeff Samardzija	.367
Shelby Miller	.368
Tom Koehler	.369
Hector Noesi	.369
Yu Darvish	.369
Yordano Ventura	.370
Rubby De La Rosa	.370
Justin Verlander	.370
Kevin Gausman	.370
Hisashi Iwakuma	.371
Zach Wheeler	.371
Mike Leake	.371
Jason Vargas	.372
Roenis Elias	.372
Nick Martinez	.372
Marco Estrada	.372
Collin McHugh	.373
Henderson Alvarez	.373
Cole Hamels	.373
Clay Buchholz	.374
Bartolo Colon	.375
Tyler Matzek	.376
Drew Smyly	.377
Jake Peavy	.377
C. J. Wilson	.377
Matt Shoemaker	.377
Jeff Locke	.377

Jered Weaver	.378
Alfredo Simon	.378
Jonathon Niese	.378
Jason Hammel	.378
Gerrit Cole	.379
T. J. House	.379
Hyun-jin Ryu	.379
Tim Hudson	.379
Nick Tepesch	.380
Jesse Chavez	.380
Jeremy Guthrie	.380
Brandon McCarthy	.381
Aaron Harang	.383
Bud Norris	.383
Nathan Eovaldi	.383
Dan Haren	.384
Jake Odorizzi	.384
Jerome Williams	.384
Dillon Gee	.384
Ervin Santana	.386
A. J. Burnett	.386
Miguel Gonzalez	.386
Madison Bumgarner	.387
John Lackey	.387
Ryan Vogelsong	.387
Justin Masterson	.389
Kyle Kendrick	.390
Eric Stults	.391
Ian Kennedy	.391
Wade Miley	.392
Vidal Nuno	.392
Brad Peacock	.393
Travis Wood	.394
Tim Lincecum	.395
Tommy Milone	.400
Wei-Yin Chen	.405
Trevor Cahill	.405
Danny Salazar	.408
Ubaldo Jimenez	.408
Mike Minor	.410
Stephen Strasburg	.411
Chase Anderson	.412
Colby Lewis	.414
Jacob Turner	.420
Ricky Nolasco	.420
Edwin Jackson	.427
Franklin Morales	.435