Spray chart distributions: a context rich approach to player evaluation

Charles Young, David Dalpiaz, Daniel J. Eck

April 28, 2020

Abstract

We develop a methodological and visual framework for assessing batter vs pitcher matchups in baseball. We first estimate the distribution of balls put into play by a batter facing a pitcher, called the spray chart distribution. This distribution is conditional on batter and pitcher characteristics that reflect the underlying talent and tendencies of the players under study. Many of these individual matchups have a sample size that is too small to be reliable. Synthetic versions of the batter and pitcher under consideration are constructed in order to alleviate these concerns. Weights governing how much influence these synthetic players have on the overall spray chart distribution are constructed to minimize expected mean square error. We then provide novel performance metrics that are calculated as expectations taken with respect to the spray chart distribution. These performance metrics provide a context rich approach to player evaluation. Our main contribution is a Shiny app that allows users to evaluate any batter-pitcher matchup that has occurred or could have occurred in the last five years. One can access this app here: (URL to Shiny app goes here). This interactive tool has utility for anyone interested in baseball ranging from casual fans to sportswriters to player to team executives.

1 Introduction

Baseball has had a rich statistical history dating back to the first box score created by Henry Chadwick in 1859. Fans, journalists, and teams have obsessed over baseball statistics and performance metrics ever since. This obsession about baseball statistics is best summarized by the existence of Schwarz [2004], a best selling book devoted entirely to the statistical history of baseball. Baseball data is analyzed in the classroom as well. Max Marchi, Jim Albert, and Benjamin S. Baumer have written a book that teaches R through baseball analysis [Marchi et al., 2019], and Jim Albert maintains an actively updated website Exploring Baseball Data with R that supplements this book. Quantification of players' skill has appeared in the Statistics literature, with articles devoted to hitting [Berry et al., 1999, Albert, 2008, Brown, 2008, Jensen et al., 2009a], pitching [Albert, 2006, Shinya et al., 2017], fielding [Jensen et al., 2009b, Piette and Jensen, 2012], and total value [Baumer et al., 2015].

Most baseball statistics used for player evaluations are obtained from raw box score totals. While box score totals are a enjoyable statistical summary for baseball fans and analysts, the information contained in them is not very substantive. They ignore rich contextual information. Most commonly used player evaluation metrics are functions of context-free box score totals. These include, and are far from limited to, adjusted earned run average (ERA+), adjusted on base plus slugging percentage (OPS+), weighted runs created plus (wRC+), and wins above replacement (WAR) [bre, 2020, fan, 2020]. More sophisticated techniques in Berry et al. [1999], Brown [2008], Jensen et al. [2009a], and Baumer et al. [2015] also constructed methodology grounded in raw box

score totals. While many of these tools account for some contextual information such as ball parks, position of a player, and a player's age, they ignore opponent strength. Eck [2020] showed that context-free metrics and the class of metrics that compares a player's accomplishments directly with that player's peers are ill-equipped for player comparisons across eras of baseball, although they may perform well over the course of a single season or a few consecutive seasons. That being said, these context-free metrics do not offer any guidance for how any particular batter will perform against a particular pitcher, the most interesting outcome in a baseball game. Worse yet is that the baseball outcomes have been assumed to be independent and identically distributed (iid) realizations in the literature [Brown, 2008, Jensen et al., 2009a]. The iid assumption of outcomes may be reasonable in the prediction contexts of Brown [2008] and Jensen et al. [2009a] that involve long time frames, but iid is not appropriate for small time frames when batter and pitcher variability can be very large.

In this article we develop spray chart distributions as a methodology for understanding batter-pitcher matchups visually and numerically. Informally, spray chart distributions are 2-dimensional contours that overlay spray charts [pet, 2009, Marchi et al., 2019]. We construct spray chart distributions for batter-pitcher matchups where separate batter spray chart distribution are constructed for each of the pitches that the pitcher throws. Rich pitch characteristic information is used to supplement labelled pitch type data since the velocity, trajectory, and movement characteristics and release points of a pitch exhibit large variation across pitcher. The reported spray chart distribution for the batter pitcher matchup is the aggregation of the spray chart distributions for each pitch that the pitcher throws, the aggregation is with respect to the percentage that the pitcher throws each pitch. These spray chart distributions are estimated nonparametrically using the `kde2d` function in the `Mass` R package [Ripley et al., 2019].

One concern with this approach is that batter-pitcher matchup data can be sparse. We alleviate this concern with the development of synthetic batters and pitchers with similar characteristics as the batter and pitcher under study. Our synthetic player creation methodology is inspired by the notion of similarity scores [James, 1994, Silver, 2003]. However, unlike the similarity scores presented in James [1994] and Silver [2003], we construct similarity scores using a nearest neighbor approach that is based on the underlying batter and pitcher characteristics of the players under study instead of observed statistics. The pitcher characteristics collected are averages of the velocity, trajectory, movement, and release point of a pitches thrown by pitchers. The batter characteristics collected are averages of launch angle, exit velocity, spray angle, and binned batted ball location information. These player characteristics are obtained from Statcast scraped using functionality in the `baseballr` R package [Petti et al., 2020], and they reflect the underlying talent and tendencies of baseball players. For each batter and pitcher matchup we estimate three spray chart densities, the first is the natural spray chart distribution corresponding to the players under study, the second is the spray chart of the synthetic pitcher versus the original batter, and the third is the spray chart of the original pitcher versus the synthetic batter. We report a synthetic spray chart distribution which is a weighted average of these spray chart distributions. The weights are chosen with the aim of minimizing mean squared error.

The main contribution of this work is a Shiny app which gives its users the ability to display the synthetic spray chart distribution for any batter pitcher matchup that has occurred or could have occurred in the last five years, the years that Statcast data exists. These synthetic spray charts are visualized over an image of a baseball field so that the batted ball distribution is displayed in its proper context. We also report performance metrics that are computed as expectations with respect to the synthetic spray chart distribution. The expected number of singles, doubles, triples, homeruns are reported, and the expected batting average on balls in play (xBABIP) and the expected total bases on contact (xBsCON) are also reported. These matchup dependent metrics

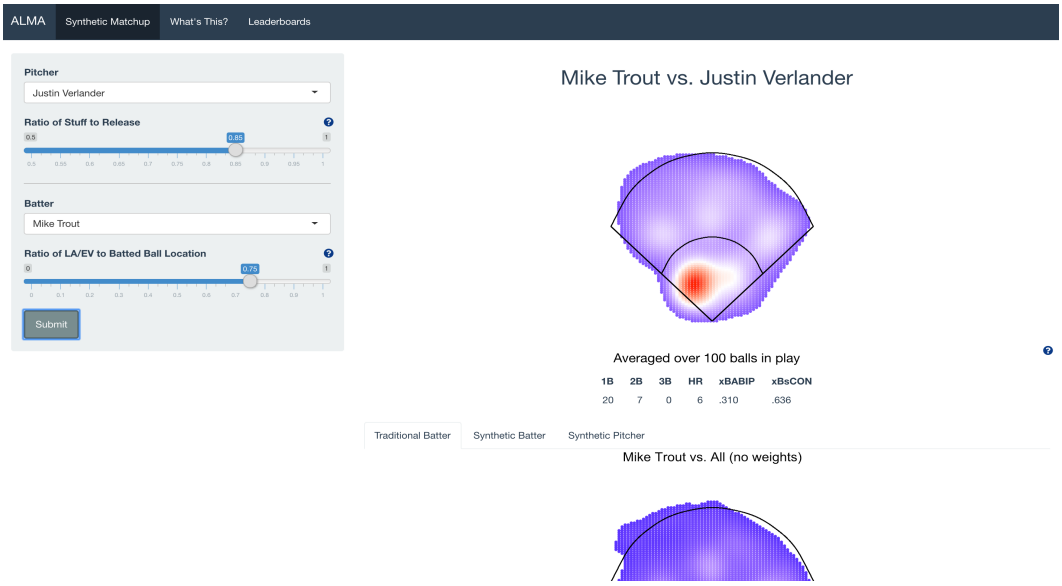

Figure 1: The layout of the application upon submission.

allow for any user to assess the expected performance of batters and pitchers when they face each other.

2 Motivating Example

In this section we present a snapshot of what our Shiny app offers users. The Shiny app is available at: (URL to Shiny app goes here). The default matchup in the application pairs the reigning American League Cy Young winning pitcher against the reigning American League MVP batter, in Justin Verlander vs. Mike Trout. The layout includes a sidebar with four filters: two dropdowns for pitcher/batter selection and two sliders for metric adjustment. A snapshot of the appearance of our visualization is depicted in Figures 1 and 2.

The slider displays allow users to determine the relative importance of stuff, a colloquial term for pitch quality, versus release information for pitchers and launch conditions versus batted ball locations for batters in forming the synthetic players. Stuff includes velocity, spin rate, and movement. Release includes release angles and release point. Launch conditions includes exit velocity and launch angle. Location includes pull%, middle%, oppo%. The default setting of the pitcher slider favors stuff over release information information. The logic for this is that quality of pitches is more representative of ability than release point. The default setting of the batter slider favors quality of contact over batted ball tendencies which appears to bias the synthetic batter's spray chart away from that of the batter under consideration. That being said, the batted ball tendencies are recorded as percentages of balls hit to six large grids on the baseball field, ignoring the quality, trajectory, and exact location of the batted ball. Thus, the quality of contact forms a more complete representation of a batter's skill than tendency.

As previously mentioned, these visualizations can help coaches position their fielders effectively. While a traditional spray chart may be useful in aggregate, building a custom spray chart to reflect a specific batter-pitcher matchup will yield more accurate results on a plate appearance by plate appearance level. This synthetically created spray chart will give the user an expected distribution of batted balls for the batter-pitcher matchup based on a combination the distribution

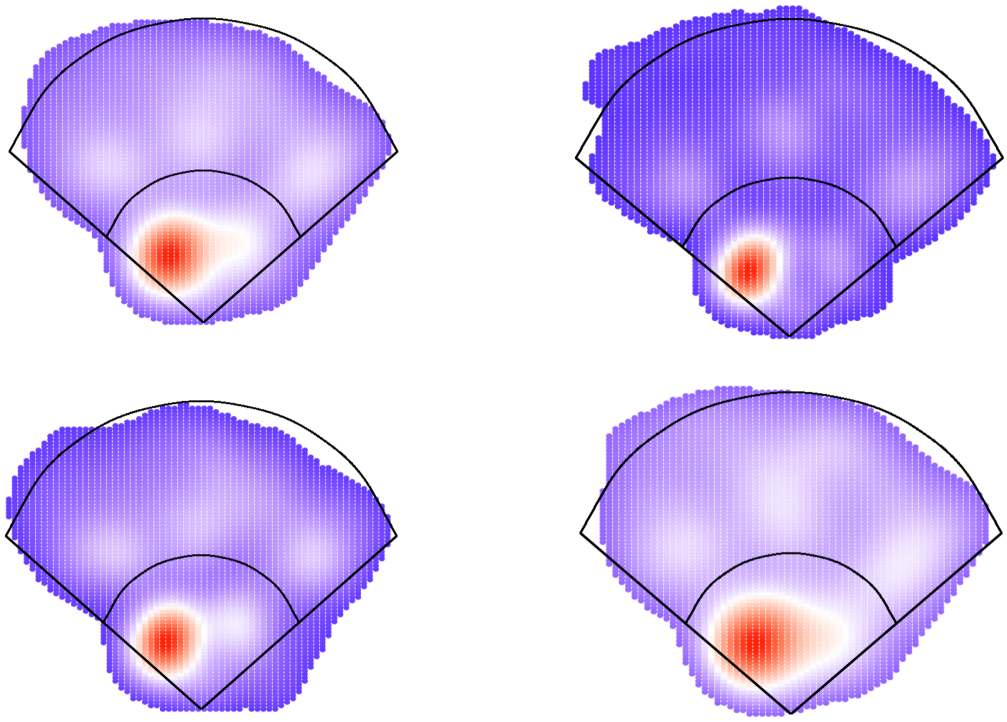

Figure 2: Spray chart distributions constructed by our app. This example corresponds to the spray chart distribution when batter Mike Trout faces pitcher Justin Verlander. The top-left panel is the complete synthetic spray chart for the batter-pitcher matchup. The top-right panel is the traditional batter-pitcher spray chart distribution, with no consideration of similar players. The bottom-left panel is the synthetic batter's spray chart distribution vs. the real pitcher. The bottom-right panel is the real batter's spray chart distribution vs. the syntheic batter.

Top 10 Similar Pitchers

Name	Season	Similarity	IP	ERA	K%	BB%	xFIP	WAR
Jake Arrieta	2018	0.33	172.20	3.96	19.1 %	7.9 %	4.08	1.90
Gerrit Cole	2019	0.32	212.10	2.50	39.9 %	5.9 %	2.48	7.40
Rick Porcello	2015	0.32	172.00	4.92	20.2 %	5.2 %	3.72	1.70
Jake Arrieta	2017	0.31	168.10	3.53	23.1 %	7.8 %	4.11	2.50
Rick Porcello	2017	0.30	203.10	4.65	20.5 %	5.4 %	4.43	2.00
Jeff Samardzija	2017	0.30	207.20	4.42	24.2 %	3.8 %	3.60	3.80
Jeff Samardzija	2019	0.30	181.10	3.52	18.9 %	6.6 %	5.02	1.50
Rick Porcello	2018	0.28	191.10	4.28	23.5 %	5.9 %	3.87	2.40
Mike Fiers	2019	0.28	184.20	3.90	16.7 %	7.0 %	5.19	1.70
Ivan Nova	2017	0.27	187.00	4.14	16.7 %	4.6 %	4.19	2.20

Figure 3: The most similar pitchers to Justin Verlander with an 85% stuff-to-release ratio

of similar batters against the pitcher, the distribution of similar pitchers against the batter, and any observations of pitcher vs batter since 2015. The app also displays two additional synthetic charts and a leaderboard displaying the most similar pitchers/batters. These include their overall similarity score and a variety of performance metrics. See Figure 3 for an example of the top 10 most similar pitchers to Justin Verlander.

3 Pitcher and batter characteristics

The data for our app was acquired via the `baseballr` R package [Petti et al., 2020]. This dataset contains every pitch thrown since 2015 that has been captured by Statcast. A few preprocessing steps are involved:

- Pitches classified with the following pitch types are removed: Eephus, Screwball.
- Pitches classified as Knuckle-Curve are renamed to Curveball.
- Pitches classified as Forkball are renamed to Splitter.
- Pitch launch angles are calculated using rudimentary kinematics.
- Batted ball locations are adjusted to reflect accurate baseball field coordinates [Petti, 2017].
- Spray angle is calculated from the x and y coordinates of the batted ball, and adjusted where a negative angle implies the ball was pulled.

For pitcher comparisons, pitchers are aggregated on a season and pitch type basis. These are the variables considered: velocity, spin rate, horizontal break, horizontal release angle, horizontal release point, vertical break, vertical release angle, vertical release point, and extension. Averages of these variables are taken across each pitch thrown by each pitcher. For a specific pitcher, the pool of pitchers must have thrown at least 10 pitches for at least $\lceil \frac{n_{\text{pitch\_types}}}{2} \rceil$ of the main pitcher's pitch

types. For example, for a main pitcher who throws three pitch types, to be eligible for comparing a pitcher must throw at least two of the same pitch types.

For batter comparisons, batter are aggregated on a season, handedness, and vs. pitch type basis. These are the variables considered: exit velocity, launch angle, pull%, middle%, and oppo%. Averages of these variables are taken across each pitch type faced across each batter.

4 Spray chart distributions

A spray chart distribution for a batter is a distribution F over a bounded subset $\mathcal{Y} \in \mathbb{R}^2$. The set $\mathcal{Y}$ contains plausible locations of batted balls from home plate. Let $(0,0) \in \mathcal{Y}$ denote the location of home plate where the batter stands. With this specification we can take $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{y}\| \leq 1000\}$ where values in $\mathcal{Y}$ are locations in feet and $\|\cdot\|$ is the Euclidean norm. This specification of $\mathcal{Y}$ guarantees that $F(\mathcal{Y}) = 1$ for all batters in history, no human in history has ever come close to hitting a ball 1000 feet.

Our main inferential goal will be to consider spray chart distributions that are conditional on several characteristics for pitchers $\mathbf{x}_p$ and batters $\mathbf{x}_b$, where $\mathbf{x} = (\mathbf{x}_p', \mathbf{x}_b')' \in \mathcal{X}$, and $\mathcal{X}$ is assumed to be bounded. The conditional spray chart density function will be denoted as $f(\cdot|\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. We will estimate $f(\cdot|\mathbf{x})$ with a multivariate kernel estimator

$$\hat{f}_{\mathbf{H}}(\mathbf{y}|\mathbf{x}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n_{\mathbf{x}}} K\left(\mathbf{H}^{-1}(\mathbf{y}_i - \mathbf{y})\right) \tag{1}$$

where K is a multivariate kernel function, $\mathbf{H}$ is a matrix of bandwidth parameters, and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_{\mathbf{x}}}$ is the $n_{\mathbf{x}}$ batted ball locations from home plate observed when the batter faced situation $\mathbf{x}$. The estimated spray chart density function (1) is a smoothed surface overlaying a spray chart. Our visualization of the spray chart distribution will be along n_g common grid points $g_1, \ldots, g_{n_g}$ for all batters and all conditional characteristics $\mathbf{x} \in \mathcal{X}$ under study. Commonality of grid of points across the batters and $\mathbf{x}$ allows for straightforward comparisons of spray chart distributions.

Our implementation will estimate $f(\cdot|\mathbf{x})$ using the `kde2d` and `kde2d.weighted` functions in R [Ripley et al., 2019, Hamilton, 2018]. These functions are chosen because of their presence in the `ggplot2` R package [Wickham, 2016] which will be employed for our visualizations. Therefore, we estimate $f(\cdot|\mathbf{x})$ using a bivariate nonparametric Gaussian kernel density estimator

$$\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) = \frac{1}{nh_{y_1}h_{y_2}} \sum_{i=1}^{n_{\mathbf{x}}} \phi\left(\frac{y_1 - y_{1i}}{h_{y_1}}\right) \phi\left(\frac{y_2 - y_{2i}}{h_{y_2}}\right) \tag{2}$$

where ϕ is a standard Gaussian density, $\mathbf{h} \in \mathbb{R}^2$ is a bandwidth parameter so that the matrix $\mathbf{H}$ in (1) is $\mathbf{H} = \text{diag}(\mathbf{h})$, and (y_{1i}, y_{2i}), $i = 1, \ldots, n$ are the observed batted ball locations.

4.1 Synthetic player construction

We develop a method for synthetically recreating baseball players in order to alleviate the small sample size concerns of individual batter-pitcher matchups. Matchup data involving these synthetic players will then be included in our analysis to estimate the spray chart density function for individual batter-pitcher matchups. Our synthetic player is similar in spirit to similarity scores [James, 1994, Silver, 2003].

Our similarity score of pitcher j to pitcher k is $s(\mathbf{x}_{p,j}, \mathbf{x}_{p,k}) = \exp(-\|\mathbf{x}_{p,j} - \mathbf{x}_{p,k}\|_{\mathbf{V}_p})$ where $\mathbf{x}_{p,j}$ and $\mathbf{x}_{p,k}$ are, respectively, the underlying pitch characteristics for pitcher j and k, $\|\mathbf{x}_{p,j} - \mathbf{x}_{p,k}\|_{\mathbf{V}_p} =$

$\sqrt{(\mathbf{x}_{p,j} - \mathbf{x}_{p,k})'\mathbf{V}_p(\mathbf{x}_{p,j} - \mathbf{x}_{p,k})}$, and $\mathbf{V}_p$ is a diagonal weight matrix that is chosen to scale the pitch characteristics and give preference to pitch characteristics that have higher influence on the spray chart distribution under study. The user of our shiny app has control of the entries of $\mathbf{V}_p$ by adjusting the pitcher slider. Similarity scores for batters are defined in a similar way. Implicit in this construction is the assumption that the underlying pitcher and batter characteristics that we collect represents the underlying talent of the players under study.

Our method for estimating spray chart densities for batter-pitcher matchups with synthetic players is as follows: first, without loss of generality, let $\mathbf{x}_p$ and $\mathbf{x}_b$ be the characteristics for the batter and pitcher under study so that $\mathbf{x} = (\mathbf{x}_p', \mathbf{x}_b')'$. There will be J pitchers and K batters available to form the pool of players that we compare to the pitcher and batter under study. Then line up the batter and pitcher characteristics for all of the available players, $\mathbf{x}_{b,j},\ j = 1, ..., J$ and $\mathbf{x}_{p,k},\ k = 1, ..., K$. Now obtain the similarity scores $s_{p,j} = s(\mathbf{x}_p, \mathbf{x}_{p,j}),\ j = 1, ..., J$ and $s_{b,k} = s(\mathbf{x}_b, \mathbf{x}_{b,k}),\ k = 1, ..., K$. Now convert the similarity scores to weights $w_{p,j} = s_{p,j} / \sum_{l=1}^{J} s_{p,l}$ and $w_{b,k} = s_{b,k} / \sum_{l=1}^{K} s_{b,l}$. The spray chart density for a batter facing the synthetic pitcher is

$$f_{\text{sp}}(\mathbf{y}|\mathbf{x}_b) = \sum_{j=1}^{J} w_{p,j} f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b). \tag{3}$$

The spray chart density for a pitcher facing the synthetic batter is

$$f_{\text{sb}}(\mathbf{y}|\mathbf{x}_p) = \sum_{k=1}^{K} w_{b,k} f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}). \tag{4}$$

It is clear that the above synthetic densities are biased in the population. We then estimate (3) and (4) with

$$\hat{f}_{\text{sp}}(\mathbf{y}|\mathbf{x}_b) = \sum_{j=1}^{J} w_{p,j} \hat{f}_{\mathbf{h}_{p,j}}(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b), \qquad \hat{f}_{\text{sb}}(\mathbf{y}|\mathbf{x}_p) = \sum_{k=1}^{K} w_{b,k} \hat{f}_{\mathbf{h}_{b,k}}(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}), \tag{5}$$

where we let $n_{p,j}$ denote the matchup sample size of pitcher j versus the batter under study, $n_{b,k}$ denote the matchup sample size of the pitcher under study versus batter k, and $\mathbf{h}_{p,j}$ and $\mathbf{h}_{b,k}$ are bandwidth parameters.

Our implementation estimates the densities in (5) with the `kde2d.weighted` function in the `ggtern` R package [Hamilton, 2018]. The estimators (5) are obviously biased estimators for $f(\mathbf{y}|\mathbf{x})$. However, they can lead to lower MSE in certain scenarios. One obvious case is when there exists weights $w_{p,j} \approx 1$, $w_{b,k} \approx 1$ and $n_{p,j} > n$, $n_{b,k} > n$. In these settings, the players under study are almost perfectly replicated by another player in the available pool and this player has a larger number of matchups with the batter or pitcher under study. Another obvious case is when the batter has never faced the pitcher before so that no data is available to estimate $f(\mathbf{y}|\mathbf{x})$ directly, although that does not guarantee that the estimators (5) are good estimators for $f(\mathbf{y}|\mathbf{x})$. Our implementation will estimate $f(\mathbf{y}|\mathbf{x})$ with

$$\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) = \lambda \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + \lambda_p \hat{f}_{\text{sp}}(\mathbf{y}|\mathbf{x}_p) + \lambda_b \hat{f}_{\text{sb}}(\mathbf{y}|\mathbf{x}_b) \tag{6}$$

where $\lambda, \lambda_p,\ \lambda_b$ form a convex combination. Note that these calculations are conditional on the pitch characteristics which implies that they are also conditional on $w_{p,j}$ and $w_{b,k}$ since the weights are a deterministic function of the pitch characteristics. Our implementation will estimate the elements of $\boldsymbol{\lambda}$ as

$$\lambda = \frac{\sqrt{n}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}}, \qquad \lambda_p = \frac{\sqrt{n_p}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}}, \qquad \lambda_b = \frac{\sqrt{n_b}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}},$$

where $n_p = \sum_{j=1}^{J} s_{p,j}^2 n_{p,j}$ and $n_b = \sum_{k=1}^{K} s_{b,k}^2 n_{b,k}$. Informally, these choices arise as a balance between the natural bias that exists in our synthetic player construction and the inherent estimation variation. In our application it is reasonable to take $n_{p,j} = O(n)$ and $n_{b,k} = O(n)$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$, it is also reasonable to assume that n will be too small to be of much use, hence the reason why n_p and n_b are aggregated with respect to similarity scores instead of weights that form a convex combination. However, in the event that n is large enough to provide reliable estimation of $f(\mathbf{y}|\mathbf{x})$ with $\hat{f}_h(\mathbf{y}|\mathbf{x})$, then n dominates n_p and n_b. Formal technical justification for selecting $\boldsymbol{\lambda}$ is given in the Appendix, in the Appendix we argue that our choices of $\boldsymbol{\lambda}$ lead to the estimator (6) having a lower MSE than the estimator (2).

4.2 Performance metrics

We develop novel performance metrics that are theoretically computed as expectations with respect to the synthetic spray chart distribution estimated in the last subsection. We estimate the expected number of singles, doubles, triples, homeruns that the batter hits versus the pitcher in a particular matchup. We also estimate the xBABIP and xBsCON as additional summary measures. The metric xBsCON is best interpreted as slugging percentage conditional on balls put into play. Our implementation will not estimate these expectations exactly, there is not enough historical batted ball data.

To theoretically estimate these quantities we first obtain five years of batted ball data. We then estimate the proportion of batted balls that were either an out (O), single (1B), double (2B), triple (3B), or homerun (HR) at locations $\mathbf{y}$ on the baseball field. Denote this vector of estimated proportions at $\mathbf{y}$ as $\widehat{\mathbf{P}}(\mathbf{y}) = (p_{\mathrm{O}}(\mathbf{y}), p_{\mathrm{1B}}(\mathbf{y}), p_{\mathrm{2B}}(\mathbf{y}), p_{\mathrm{3B}}(\mathbf{y}), p_{\mathrm{HR}}(\mathbf{y}))'$. Next, we obtain $\widehat{E}(\mathbf{x}) = \int \mathbf{P}(\mathbf{y}) \hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$, where $\widehat{E}(\mathbf{x}) = (\hat{e}_{\mathrm{O}}(\mathbf{x}), \hat{e}_{\mathrm{1B}}(\mathbf{x}), \hat{e}_{\mathrm{2B}}(\mathbf{x}), \hat{e}_{\mathrm{3B}}(\mathbf{x}), \hat{e}_{\mathrm{HR}}(\mathbf{x}))'$. Thus $\widehat{E}(\mathbf{x})$ is the estimated expected vector of outcomes where the expectation is taken with respect to the estimated spray chart distribution. Expected BABIP is then calculated as $\text{xBABIP} = \hat{e}_{\mathrm{1B}}(\mathbf{x}) + \hat{e}_{\mathrm{2B}}(\mathbf{x}) + \hat{e}_{\mathrm{3B}}(\mathbf{x})$ and xBsCON is calculated as $\text{xBsCON} = \hat{e}_{\mathrm{1B}}(\mathbf{x}) + 2\hat{e}_{\mathrm{2B}}(\mathbf{x}) + 3\hat{e}_{\mathrm{3B}}(\mathbf{x}) + 4\hat{e}_{\mathrm{HR}}(\mathbf{x})$. Our Shiny app also displays the floor of $100\hat{e}_{\mathrm{1B}}(\mathbf{x})$, $100\hat{e}_{\mathrm{2B}}(\mathbf{x})$, $100\hat{e}_{\mathrm{3B}}(\mathbf{x})$, and $100\hat{e}_{\mathrm{HR}}(\mathbf{x})$. The 100 multiplier is a normalization that is intended to easy interpretation.

The previous paragraph outlines how we would calculate our performance metrics if we could obtain $\widehat{\mathbf{P}}(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{Y}$. However, we do not have enough data available to achieve this task in reality. Therefore, we calculate discretized versions of these performance metrics over 10 feet by 10 feet bins.

5 Discussion

The primary contribution of this work is the development of synthetic spray chart distributions that are calculated under the hood of a Shiny app which provides users with visual and numeric summary measures of baseball matchups. This app will be of interest to baseball fans, analysts, players, and team executives alike. Our tools show users batter tendencies versus pitchers while providing summaries of their overall success or lack thereof. Our tools greatly improve upon the inferential power of spray charts [pet, 2009, Marchi et al., 2019] as a visualization of a batter's hitting tendencies. Spray charts may be uninformative for individual matchups due to a lack of data. Our synthetic player construction alleviates this problem.

We are not the first to incorporate additional players into an analysis via similarity scores with the understanding that doing so improves estimation performance. The PECOTA prediction methodology [Silver, 2003] tries to forecast the ability of players using aggregate estimates obtained from other similar players. To the best of our knowledge, we are the first to base similarity scores

exclusively on Statcast data which we believe provides a truer notion of similarity in the context of individual batter-pitcher matchups.

There needs to be a clear distinction made that clarifies the goal of this study. The goal of this synthetic spray chart approach is to provide a system to project where a batted ball will go given a certain batter-pitcher matchup - it is **not** to gauge true talent. On average, players who hit the ball harder with a more optimized launch angle will receive better projected stats, since these balls tend to produce more home runs (and thus take fielders out of the equation.) Tools like speed and eye at the plate, therefore, will will not be reflected in this application.

Appendix: Justification for our choice of $\boldsymbol{\lambda}$

We now motivate $\boldsymbol{\lambda}$ theoretically. We first assume some additional structure on the space of functions that $f(\cdot|\cdot)$ belongs to in order to facilitate our motivation. The best batters in baseball are good at hitting the ball with general intent but batted ball locations will still exhibit variation. Therefore we expect for spray chart densities to be smooth and lacking of sharp peaks. It is reasonable to assume that $f(\cdot|\cdot)$ belongs to a multivariate Hölder class of densities which we will denote by $H(\beta, L)$. The space $H(\beta, L)$ is the set of functions $f(\mathbf{y}|\mathbf{x})$ such that

$$|D^{\mathbf{s}}_{\mathbf{y}} f(\mathbf{y}|\mathbf{x}) - D^{\mathbf{s}}_{\mathbf{y}} f(\mathbf{y}'|\mathbf{x})| \leq L_{\mathbf{x}} \|\mathbf{y} - \mathbf{y}'\|^{\beta - |\mathbf{s}|},$$
$$|D^{\mathbf{t}}_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}) - D^{\mathbf{t}}_{\mathbf{x}} f(\mathbf{y}|\mathbf{x}')| \leq L_{\mathbf{y}} \|\mathbf{x} - \mathbf{x}'\|^{\beta - |\mathbf{t}|},$$

for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$, all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and all $\mathbf{s}$ such that $|\mathbf{s}| = \beta - 1$ where $D^{\mathbf{s}}_{\mathbf{y}} = \partial^{s_1 + s_2} / \partial y_1^{s_1} \partial y_2^{s_2}$, $D^{\mathbf{t}}_{\mathbf{x}} = \partial^{t_1 + \cdots + t_p} / \partial x_1^{t_1} \cdots \partial x_p^{t_p}$ and $L_{\mathbf{x}} \leq L$ for all $\mathbf{x} \in \mathcal{X}$ and $L_{\mathbf{y}} \leq L$ for all $\mathbf{y} \in \mathcal{Y}$. We will assume the following regularity conditions for our spray chart distributions and kernel functions:

A1. The density f is square integrable, twice continuously differentiable, and all the second order partial derivatives are square integrable. We will suppose that $\beta = 2$ in $H(\beta, L)$.

A2. The kernel K is a spherically symmetric and bounded pdf with finite second moment and square integrable.

A3. $\mathbf{H} = \mathbf{H}_n$ is a deterministic sequence of positive definite symmetric matrices such that, $n \det(\mathbf{H}) \to \infty$ when $n \to \infty$ and $\mathbf{H} \to 0$ elementwise.

Condition A2 holds for the multivariate Gaussian kernel function that we use in our implementation. We will let $\mathbf{H}$ be a matrix of bandwidth parameters that has diagonal elements $\mathbf{h}$, in our implementation $\mathbf{H} = \text{diag}(\mathbf{h})$. We will use the following notation: $R_{\mathbf{x}}(f) = \int f(\mathbf{y}|\mathbf{x})^2 d\mathbf{y}$, $\mu_2(K) = \int u^2 K(u) du$, and $\mathcal{H}_f(\mathbf{y}|\mathbf{x})$ is the Hessian matrix respect to $f(\mathbf{y}|\mathbf{x})$ where derivatives are taken with respect to $\mathbf{y}$. Assume that pitch outcomes are independent across at bats and that $n_{p,j} = O(n)$, $n_{b,k} = O(n)$ and $\mathbf{h}_{p,j} = O(\mathbf{h})$, $\mathbf{h}_{b,k} = O(\mathbf{h})$ for all $j = 1, \ldots, J$, $k = 1, \ldots, K$.

With the specification that $\beta = 2$ in Condition A1 we have that $f(\mathbf{y}|\mathbf{x}) - L\|\mathbf{x} - \mathbf{x}'\|^2 \leq f(\mathbf{y}|\mathbf{x}') \leq f(\mathbf{y}|\mathbf{x}) + L\|\mathbf{x} - \mathbf{x}'\|^2$. This result implies that

$$R_{\mathbf{x}'}(f) - R_{\mathbf{x}}(f) = \int (f(\mathbf{y}|\mathbf{x}')^2 - f(\mathbf{y}|\mathbf{x})^2) d\mathbf{y} = \int (f(\mathbf{y}|\mathbf{x}') - f(\mathbf{y}|\mathbf{x}))(f(\mathbf{y}|\mathbf{x}') + f(\mathbf{y}|\mathbf{x})) d\mathbf{y}$$
$$\leq L\|\mathbf{x}' - \mathbf{x}\|^2 \int (f(\mathbf{y}|\mathbf{x}') + f(\mathbf{y}|\mathbf{x})) d\mathbf{y} = 2L\|\mathbf{x}' - \mathbf{x}\|^2,$$

and $R_{\mathbf{x}}(f) - 2L\|\mathbf{x} - \mathbf{x}'\|^2 \leq R_{\mathbf{x}'}(f) \leq R_{\mathbf{x}}(f) + 2L\|\mathbf{x} - \mathbf{x}'\|^2$.

We will define $\tilde{\mathbf{x}}_{b,k} = (\mathbf{x}'_p, \mathbf{x}'_{b,k})'$ and $\tilde{\mathbf{x}}_{p,j} = (\mathbf{x}'_{p,j}, \mathbf{x}'_b)'$ for notational convenience, and will additionally assume the following regularity approximations:

A4. The quantities $\sum_{j=1}^{J} w_{p,j}^2 \|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^m$ and $\sum_{k=1}^{K} w_{b,k}^2 \|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^m$ are negligible, where and $m = 2, 4$.

A5. The quantities $\sum_{j=1}^{J} w_{p,j} \left(\mathbf{h}'_{p,j} \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b))\mathbf{h}_{p,j} - \mathbf{h}' \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}\right)$ and $\sum_{k=1}^{K} w_{b,k} \left(\mathbf{h}'_{b,k} \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}))\mathbf{h}_{b,k} - \mathbf{h}' \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}\right)$ are negligible.

Approximation A4 is reasonable in our baseball application where there are many players similar enough to the players under study so that $\sum_{j=1}^{J} s_{p,j} > 1$ and $\sum_{k=1}^{K} s_{b,k} > 1$ and $s_{p,j}\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^m, s_{b,k}\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^m \to 0$ as $\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|, \|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\| \to \infty$ for all integers m. Approximation A5 is reasonable by similar logic. Specification of $\beta = 2$ implies that $\|\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k})) - \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\| \leq \sqrt{d_p}L$ and $\|\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b)) - \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\| \leq \sqrt{d_b}L$ where d_p and d_b are, respectively, the dimension of $\mathbf{x}_p$ and $\mathbf{x}_b$.

We now have enough structure to estimate the MSE of (2) and (6). Standard results from nonparametric estimation theory give

$$\text{E}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})) - f(\mathbf{y}|\mathbf{x}) = \frac{\mu_2(K)\mathbf{h}' \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2),$$

and

$$\text{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})) = \frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n \det(\mathbf{H})} + O\left(\frac{1}{n}\right).$$

Our multivariate Hölder class specifications yield,

$$\begin{aligned}
\text{E}(\hat{g}_{\boldsymbol{\lambda}}(y|x)) &= \lambda \,\text{E}\, \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + \lambda_p \,\text{E}\, \hat{f}_{\text{sp}}(\mathbf{y}|\mathbf{x}_b) + \lambda_b \,\text{E}\, \hat{f}_{\text{sb}}(\mathbf{y}|\mathbf{x}_p) \\
&= \lambda f(\mathbf{y}|\mathbf{x}) + \lambda \frac{\mu_2(K)\mathbf{h}' \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + \lambda_p \sum_{j=1}^{J} w_{p,j} \,\text{E}\, \hat{f}_{\mathbf{h}_{p,j}}(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b) \\
&\quad + \lambda_b \sum_{k=1}^{K} w_{b,k} \,\text{E}\, \hat{f}_{\mathbf{h}_{b,k}}(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}) + o(\|\mathbf{h}\|^2) \\
&= \lambda f(\mathbf{y}|\mathbf{x}) + \lambda \frac{\mu_2(K)\mathbf{h}' \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) \\
&\quad + \lambda_p \sum_{j=1}^{J} w_{p,j} f_{\mathbf{h}_{p,j}}(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b) + \lambda_p \sum_{j=1}^{J} w_{p,j} \frac{\mu_2(K)\mathbf{h}'_{p,j} \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b))\mathbf{h}_{p,j}}{2} \\
&\quad + \lambda_b \sum_{k=1}^{K} w_{b,k} f_{\mathbf{h}_{b,k}}(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}) + \lambda_b \sum_{k=1}^{K} w_{b,k} \frac{\mu_2(K)\mathbf{h}'_{b,k} \text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}))\mathbf{h}_{b,k}}{2},
\end{aligned}$$

and

$$\begin{aligned}
&\text{Var}(\hat{g}_{\boldsymbol{\lambda}}(y|x)) = \text{Var}\left(\lambda \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + \lambda_p \hat{f}_{\text{sp}}(\mathbf{y}|\mathbf{x}_b) + \lambda_b \hat{f}_{\text{sb}}(\mathbf{y}|\mathbf{x}_p)\right) \\
&= \lambda^2 \frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n \det(\mathbf{H})} + O\left(\frac{1}{n}\right) + \lambda_p^2 \sum_{j=1}^{J} w_{p,j}^2 \,\text{Var}\, \hat{f}_{\mathbf{h}_{p,j}}(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b) + \lambda_b^2 \sum_{k=1}^{K} w_{b,k}^2 \,\text{Var}\, \hat{f}_{\mathbf{h}_{b,k}}(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}) \\
&= \lambda^2 \frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n \det(\mathbf{H})} + O\left(\frac{1}{n}\right) + \lambda_p^2 \sum_{j=1}^{J} w_{p,j}^2 \frac{R_{\tilde{\mathbf{x}}_{p,j}}(f)f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b)}{n_{p,j} \det(\mathbf{H}_{p,j})} + \lambda_b^2 \sum_{k=1}^{K} w_{b,k}^2 \frac{R_{\tilde{\mathbf{x}}_{b,k}}(f)f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k})}{n_{b,k} \det(\mathbf{H}_{b,k})},
\end{aligned}$$

Let $\theta_{p,j} = n\det(\mathbf{H})/n_{p,j}\det(\mathbf{H}_{p,j})$ and $\theta_{b,k} = n\det(\mathbf{H})/n_{b,k}\det(\mathbf{H}_{b,k})$. With these specifications, we have that

$$
\begin{aligned}
&\mathrm{Var}(\hat{g}_{\boldsymbol{\lambda}}(y|x)) - \mathrm{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})) + O\left(\frac{1}{n}\right) \\
&= (\lambda^2 - 1)\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n\det(\mathbf{H})} + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 \frac{R_{\tilde{\mathbf{x}}_{p,j}}(f) f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b)}{n\det(\mathbf{H})} + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 \frac{R_{\tilde{\mathbf{x}}_{b,k}}(f) f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k})}{n\det(\mathbf{H})} \\
&\leq \left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 - 1\right) \frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n\det(\mathbf{H})} \\
&\qquad + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 \left(\frac{R_{\mathbf{x}}(f)\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2 + 2Lf(\mathbf{y}|\mathbf{x})\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2 + 2L^2\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^4}{n\det(\mathbf{H})}\right) \\
&\qquad + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 \left(\frac{R_{\mathbf{x}}(f)\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^2 + 2Lf(\mathbf{y}|\mathbf{x})\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^2 + 2L^2\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^4}{n\det(\mathbf{H})}\right).
\end{aligned}
$$

Our assumption on the $w_{b,k}^2\|\mathbf{x} - \tilde{\mathbf{x}}_{,k}\|^m$ and $w_{p,j}^2\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^m$, for $m = 2, 4$, and an identical lower bound argument implies that

$$
\mathrm{Var}(\hat{g}_{\boldsymbol{\lambda}}(y|x)) - \mathrm{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})) \approx \left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 - 1\right) \frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n\det(\mathbf{H})} + O\left(\frac{1}{n}\right).
$$

We also have

$$
\mathrm{Bias}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}), f(\mathbf{y}|\mathbf{x}))^2 = \left(\frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2)\right)^2,
$$

and regularity approximations A4 and A5 yield

$$
\begin{aligned}
\mathrm{Bias}(\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}), f(\mathbf{y}|\mathbf{x}))^2 &= \Bigg(\lambda f(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x}) + \lambda\frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) \\
&\quad + \lambda_p \sum_{j=1}^{J} w_{p,j} f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b) + \lambda_p \sum_{j=1}^{J} w_{p,j} \frac{\mu_2(K)\mathbf{h}_{p,j}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j}, \mathbf{x}_b))\mathbf{h}_{p,j}}{2} \\
&\quad + \lambda_b \sum_{k=1}^{K} w_{b,k} f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}) + \lambda_b \sum_{k=1}^{K} w_{b,k} \frac{\mu_2(K)\mathbf{h}_{b,k}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_p, \mathbf{x}_{b,k}))\mathbf{h}_{b,k}}{2}\Bigg)^2 \\
&\leq \Bigg(\lambda f(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x}) + \lambda\frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) \\
&\quad + \lambda_p \sum_{j=1}^{J} w_{p,j}(f(\mathbf{y}|\mathbf{x}) + L(-1)^t\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2) + \lambda_p \sum_{j=1}^{J} w_{p,j} \frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} \\
&\quad + \lambda_b \sum_{k=1}^{K} w_{b,k}(f(\mathbf{y}|\mathbf{x}) + L(-1)^t\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^2) + \lambda_b \sum_{k=1}^{K} w_{b,k} \frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2}\Bigg)^2 \\
&\approx \Bigg(\lambda_p \sum_{j=1}^{J} (-1)^t L w_{p,j}\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2 + \lambda_b \sum_{k=1}^{K} (-1)^t L w_{b,k}\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^2
\end{aligned}
$$

$$+ \frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2)\Bigg)^2,$$

where $t \in \{0, 1\}$ is chosen to satisfy the above inequality. Putting these variance and bias results together without the lower order terms yields

$$\begin{aligned}
&MSE(\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}), f(\mathbf{y}|\mathbf{x})) - MSE(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}), f(\mathbf{y}|\mathbf{x})) \\
&\quad \leq \left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 - 1\right) \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n \det(\mathbf{H})} \\
&\quad + \left(\lambda_p \sum_{j=1}^{J} (-1)^t L w_{p,j} \|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2 + \lambda_b \sum_{k=1}^{K} (-1)^t L w_{b,k} \|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|^2 + \frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2}\right)^2 \\
&\quad - \left(\frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2}\right)^2.
\end{aligned}$$

This motivates the following choice of $\boldsymbol{\lambda}$,

$$\lambda = \frac{\sqrt{n}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}}, \qquad \lambda_p = \frac{\sqrt{n_p}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}}, \qquad \lambda_b = \frac{\sqrt{n_b}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}},$$

where $n_p = \sum_{j=1}^{J} s_{p,j}^2 n_{p,j}$ and $n_b = \sum_{k=1}^{K} s_{b,k}^2 n_{b,k}$. We will now develop intuition for these choices. First, notice that $\lambda_p, \lambda_b \to 0$ as $\min_j(\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|), \min_k(\|\mathbf{x} - \tilde{\mathbf{x}}_{b,k}\|) \to \infty$. These cases correspond, to there being no similar pitchers or batters to the players under study. We turn attention to the bias terms, notice that

$$\lambda_p \sum_{j=1}^{J} (-1)^t L w_{p,j} \|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2 = \frac{\sqrt{\sum_{j=1}^{J} s_{p,j}^2 n_{p,j}} \sum_{j=1}^{J} (-1)^t L w_{p,j} \|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|^2}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}} \longrightarrow 0,$$

when there exists some j' such that $\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\| \to 0$ or $\min_j(\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|) \to \infty$. These cases correspond, respectively, to there being a few highly similar pitchers or there being no similar pitchers to the pitcher under study. Thus, the discrepancy in bias vanishes in the extreme cases. The same argument holds for batters. Now notice that

$$\lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 = \frac{\sum_{j=1}^{J} s_{p,j}^2 n_{p,j} \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2}{(\sqrt{n} + \sqrt{n_p} + \sqrt{n_b})^2} \longrightarrow \begin{cases} 0, & \min_j(\|\mathbf{x} - \tilde{\mathbf{x}}_{p,j}\|) \to \infty; \\ \frac{n}{(\sqrt{n} + \sqrt{n_{p,j'}} + \sqrt{n_b})^2}, & w_{p,j'} \to 1. \end{cases}$$

The same argument holds fo batters. Therefore, when there is a pitcher j' and batter k' so that $w_{p,j'}, w_{b,k'} \to 1$, we have that

$$\left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j} w_{p,j}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k} w_{b,k}^2 - 1\right) \longrightarrow \frac{3n}{(\sqrt{n} + \sqrt{n_{p,j'}} + \sqrt{n_{b,k'}})^2} - 1.$$

The above is not always less than 0 for all configurations. However, it will be less than 0 when n is comparable to $n_{p,j'}$ and $n_{b,k'}$, a setting that we will guard against in our implementation by specifying a minimal sample size to enter into available player pool.

References

The interactive spray chart tool. `https://billpetti.shinyapps.io/shiny_spraychart/`, 2009. Accessed: 2020-04-22.

Baseball reference. `https://www.baseball-reference.com`, 2020. Accessed: 2020-04-22.

Fangraphs. `https://www.fangraphs.com`, 2020. Accessed: 2020-04-22.

J. Albert. Pitching statistics, talent and luck, and the best strikeout seasons of all-time. *Journal of Quantitative Analysis in Sports*, 2(1), 2006.

J. Albert. Streaky hitting in baseball. *Journal of Quantitative Analysis in Sports*, 4(1), 2008.

B. S. Baumer, S. T. Jensen, and G. J. Matthews. openwar: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84, 2015.

S. M. Berry, C. S. Reese, and P. D. Larkey. Bridging different eras in sports. *Journal of the American Statistical Association*, 94(447):661–676, 1999.

L. .D. Brown. In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, pages 113–152, 2008.

D. J. Eck. Challenging nostalgia and performance metrics in baseball. *Chance*, 33(1):16–25, 2020.

N. Hamilton. *ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams*, 2018.

B. James. *The politics of glory: how baseball's Hall of Fame really works*. Macmillan, 1994.

S. T. Jensen, B. B. McShane, and A. J. Wyner. Hierarchical bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, 4(4):631–652, 2009a.

S. T. Jensen, K. E. Shirley, and A. J. Wyner. Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, 3(2):491–520, 2009b.

M. Marchi, J. Albert, and B. S. Baumer. *Analyzing baseball data with R 2nd Edition*. CRC Press, 2019.

B. Petti. Research notebook: New format for statcast data export at baseball savant. *The Hardball Times*, 2017.

B. Petti, B. Baumer, and B. Dilday. *baseballr: Functions for acquiring and analyzing baseball data*, 2020.

J. Piette and S. T. Jensen. Estimating fielding ability in baseball players over time. *Journal of Quantitative Analysis in Sports*, 8(3), 2012.

B. Ripley, B. Venables, D. Bates, K. Hornik, A. Gebhardt, and D. Firth. *MASS: R package*, 2019.

A. Schwarz. *The numbers game: Baseball's lifelong fascination with statistics*. Macmillan, 2004.

Masahiro Shinya, Shinji Tsuchiya, Yousuke Yamada, Kimitaka Nakazawa, Kazutoshi Kudo, and Shingo Oda. Pitching form determines probabilistic structure of errors in pitch location. *Journal of sports sciences*, 35(21):2142–2147, 2017.

N. Silver. Introducing pecota. *Baseball Prospectus*, pages 507–514, 2003.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL `https://ggplot2.tidyverse.org`.