SEAM methodology for context-rich player matchup evaluations in baseball

Julia Wapner*, Charlie Young†, David Dalpiaz‡, and Daniel J. Eck§

Department of Statistics, University of Illinois Urbana-Champaign

Abstract

We develop the SEAM (synthetic estimated average matchup) method for describing batter versus pitcher matchups in baseball. We first estimate the distribution of balls put into play by a batter facing a pitcher, called the empirical spray chart distribution. Many individual matchups have a sample size that is too small to be reliable for use in predicting future outcomes. 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 provide a Shiny app that allows users to visualize and evaluate any batter-pitcher matchup that has occurred or could have occurred in the last five years. This provides a tool that could be used to determine defensive alignments, lineup construction, or pitcher selection through estimation of spray densities based on any input matchup. One can access this app at

https://seam.stat.illinois.edu/app/

The computational speed with which the method calculates the spray densities allows the app to display the visualizations for any input almost instantly. Therefore, SEAM offers distributional interpretations of dependent matchup data which is computationally fast.

Keywords: Nonparametric density estimation; Similarity scores; Model averaging; Reproducible research; Sabermetrics; Big data applications and visualization

1 Introduction

Baseball has a rich statistical history dating back to the first box score created by Henry Chadwick in 1859. Fans of the game, academics, and professionals alike have throughout the

*jwapner2@illinois.edu
†cdy2@illinois.edu
‡dalpiaz2@illinois.edu
§dje13@illinois.edu

years made use of baseball statistics as a means of quantifying and comparing the ability of baseball players (Berry et al., 1999; Schwarz, 2004; Albert, 2006; Brown, 2008; Jensen et al., 2009,?; Piette and Jensen, 2012; Baumer et al., 2015; Marchi et al., 2019). Most baseball statistics arrive from the box score, a thorough account of many discrete events that occurred throughout the course of a baseball game. Many additional "advanced" baseball statistics combine these simple box score statistics and account for some contextual information including stadium and league effects (Marchi et al., 2019). One can view the glossary on MLB.com (`https://www.mlb.com/glossary`) or explore comprehensive websites (Baseball-Reference, 2022; Fangraphs, 2022; Baseball-Prospectus, 2022) for descriptions on a wide array of such statistics. This vast effort to quantify the ability of baseball players through statistics has been increasingly embraced by baseball teams due to a successful track record for statistics to explain how players' abilities translate to winning games (Lewis, 2004).

The basic box score statistics which serve as the foundation for most baseball statistics do not represent the fundamental unit for which outcomes are realized. Baseball is a game with a pitcher on the mound, with a defense behind him, who first throws a ball to a batter. From here the ball is put into play or it is not, and at this moment box score statistics can then be tabulated. When we look closer at this chain of events we see that the pitcher who throws the ball possesses an arsenal of pitches whose movement, velocity, and release point are unique to that pitcher. The batter also has a unique set of traits including how hard they hit the ball, the angle in which they hit the ball in the air and spray the ball. When a ball is put into play it has a final location that is relevant to, but not accounted in, the box score. As recently as 2015 this type of high-resolution data has been collected and made publicly available (Baseball-Savant, 2014). Proprietary sources maintain similar high-resolution data dating back to several years prior to 2015 (info solutions, 2022).

This type of high resolution data has ushered in a new sophisticated class of techniques for finding and exploiting players' strengths and weaknesses which has had a major impact on players' statistics when put into practice. As an example, Jensen et al. (2009) showed how this data could be used to better quantify defensive ability using Baseball Information Solutions data through 2007. One of the most prominent example of the success of this type of high-resolution data has been the implementation of defensive shifting based on the spray chart, a plot of locations for which a batter has hit the ball in the past when the ball has been put into play. Through the early goings of the 2022 MLB season, MLB had seen infield shifts on 37% of pitches (Baccellieri, 2022).

2 Spray chart distributions and densities

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. 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.

Let f be the spray chart density function corresponding to the spray chart distribution F for the matchup between a particular pitcher and a particular batter. Let $(y_{1i}, y_{2i}) \in \mathcal{Y}$,

$i = 1, \ldots, n$, be the observed batted-ball locations for this matchup. We will estimate f with a multivariate kernel density estimator

$$\hat{f}_{\mathbf{H}}(\mathbf{y}) = \frac{1}{n|\mathbf{H}|} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1}(\mathbf{y}_i - \mathbf{y})\right), \qquad \mathbf{y} \in \mathcal{Y}, \tag{1}$$

where K is a multivariate kernel function and $\mathbf{H}$ is a matrix of bandwidth parameters. Our implementation will estimate f 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 visualization. Therefore, we estimate f using a bivariate nonparametric Gaussian kernel density estimator

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

where $\mathbf{y} = (y_1, y_2) \in \mathcal{Y}$, 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 $\mathbf{H}$ is chosen according to the default bandwidth selection procedures within the `kde2d` and `kde2d.weighted` functions. The estimated spray chart density function $\hat{f}_{\mathbf{h}}$ 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 matchups under study. Commonality of grid points allows for straightforward comparisons of spray chart distributions in practice.

We extend this framework to 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. Denote the conditional spray chart distribution as $F(\mathbf{y}|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. The conditional spray chart density function corresponding to $F(\mathbf{y}|\mathbf{x})$ will be denoted as $f(\mathbf{y}|\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. Thus, $f(\mathbf{y}|\mathbf{x})$ is a nonparametric regression model that we will again estimate with a bivariate nonparametric Gaussian kernel density estimator

$$\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) = \frac{1}{nh_{y_1}h_{y_2}} \sum_{i=1}^{n} \phi\left(\frac{y_1 - y_{1i}}{h_{y_1}}\right) \phi\left(\frac{y_2 - y_{2i}}{h_{y_2}}\right), \tag{3}$$

where the sample of batted-ball locations $(y_{1i}, y_{2i}) \in \mathcal{Y}$, $i = 1, \ldots, n$ are now conditional on $\mathbf{x} \in \mathcal{X}$.

2.1 Synthetic player construction

We develop a method for synthetically recreating baseball players in order to alleviate the small sample size concerns inherent in the estimation of f for any batter-pitcher matchup. Matchup data involving these synthetic players will then be included in our analysis to estimate f. We first develop the similarity scores used in this methodology. We will suppose that there are J pitchers and K batters available in our donor pool. We will suppose that the pitcher in the matchup under study throws n_{type} different types of pitches. We will let $\mathbf{x}_{p,t}$ be the pitcher covariates for pitch type $t = 1, \ldots, n_{\text{type}}$. Similarly, let $\mathbf{x}_{b,t}$ be the batter covariates when facing pitch type $t = 1, \ldots, n_{\text{type}}$. The covariates in $\mathbf{x}_{p,t}$ and $\mathbf{x}_{b,t}$ are averages

across the pitch-by-pitch realizations. We will denote d_p and d_b as the dimensions of $\mathbf{x}_{p,t}$ and $\mathbf{x}_{b,t}$ respectively. For a pitch type t, the similarity score of pitcher j_1 to pitcher j_2 is defined as $s(\mathbf{x}_{p,j_1,t}, \mathbf{x}_{p,j_2,t}) = \exp(-\|\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t}\|_{\mathbf{V}_{p,t}})$ where $\mathbf{x}_{p,j_1,t}$ and $\mathbf{x}_{p,j_2,t}$ are, respectively, the underlying pitch characteristics for pitcher j_1 and j_2,

$$\|\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t}\|_{\mathbf{V}_{p,t}} = \left((\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t})'\mathbf{V}_{p,t}(\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t})\right)^{1/d_p}, \tag{4}$$

and $\mathbf{V}_{p,t}$ is a diagonal weight matrix that is chosen to scale the pitch characteristics and give preference to pitch characteristics that are chosen to have higher influence on the spray chart distribution under study. Similarity scores of the form $s(\mathbf{x}_{p,j_1,t}, \mathbf{x}_{p,j_2,t})$ have desirable theoretical properties that are explained in the Appendix and, in practice, they guard against downplaying the effect of the players under study. Users of our Shiny app has some control of the entries of $\mathbf{V}_{p,t}$ by adjusting the pitcher slider. Similarity scores between batters k_1 and k_2 are defined in a similar manner and are noted as $s(\mathbf{x}_{b,k_1,t}, \mathbf{x}_{b,k_2,t})$.

Implicit in this construction is the assumption that the collected pitcher and batter characteristics are an exhaustive set of inputs to properly estimate the spray chart distribution. Therefore, we are assuming that f is conditional on $\mathbf{x}_{b,t}, \mathbf{x}_{p,t}, \rho_t$, for $t = 1, \ldots, n_{\text{type}}$, where ρ_t be the proportion of time that the pitcher in the matchup under study throws pitch type t. We therefore represent $f(y)$ as $\sum_t \rho_t f(y|\mathbf{x}_t)$, where $\mathbf{x}_t = (\mathbf{x}'_{p,t}, \mathbf{x}'_{b,t})'$.

We describe the synthetic spray chart density for the batter under study facing the synthetic version of the pitcher under study. Without loss of generality, let $\mathbf{x}_{p,t}$ be the characteristics for pitch type t thrown by the pitcher under study, let $\mathbf{x}_{b,t}$ be the characteristics for the batter under study. We then line up the pitcher characteristics for all of the pitchers in the donor pool, $\mathbf{x}_{p,j,t}$, $j = 1, ..., J$. Now obtain the similarity scores $s_{p,j,t} = s(\mathbf{x}_{p,t}, \mathbf{x}_{p,j,t})$ and then construct the weights $w_{p,j,t} = s_{p,j,t} / \sum_{l=1}^{J} s_{p,l,t}$, for $j = 1, ..., J$. For pitch type t, the spray chart density for a batter facing the synthetic pitcher is

$$f_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t} f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}). \tag{5}$$

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

$$f_{\text{sp}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{type}}} \rho_t f_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}). \tag{6}$$

The conditioning on $\mathbf{x}_{b,t}, \mathbf{x}_{p,j,t}, \rho_t$, for $t = 1, \ldots, n_{\text{type}}$ and $j = 1, \ldots, J$ is suppressed in the density $f_{\text{sp}}(\mathbf{y})$.

Similarly, we describe the synthetic spray chart density for the synthetic batter facing the pitcher under study. For pitch type t, we line up the batter characteristics for all of the available batters that faced pitch type t thrown by the pitcher under study, $\mathbf{x}_{b,k,t}$, $k = 1, ..., K$. We obtain the similarity scores $s_{b,k,t} = s(\mathbf{x}_{b,t}, \mathbf{x}_{b,k,t})$ and then construct the weights $w_{b,k,t} = s_{b,k,t} / \sum_{l=1}^{K} s_{b,l,t}$, for $k = 1, ..., K$. For pitch type t, the spray chart density for a pitcher facing the synthetic batter is

$$f_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}) = \sum_{k=1}^{K} w_{b,k,t} f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}). \tag{7}$$

The spray chart density for the synthetic batter facing the pitcher under study is then

$$f_{\text{sb}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{type}}} \rho_t f_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t)}. \tag{8}$$

The conditioning on $\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}, \rho_t$, for $t = 1, \ldots, n_{\text{type}}$ and $k = 1, \ldots, K$ is suppressed in the density $f_{\text{sb}}(\mathbf{y})$.

We then estimate (5) and (7) with

$$\hat{f}_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t} \hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}), \qquad \hat{f}_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}) = \sum_{k=1}^{K} w_{b,k,t} \hat{f}_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}), \tag{9}$$

where, for pitch type t, we let $n_{p,j,t}$ denote the matchup sample size of pitcher j versus the batter under study, $n_{b,k,t}$ denote the matchup sample size of the pitcher under study versus batter k, and $\mathbf{h}_{p,j,t}$ and $\mathbf{h}_{b,k,t}$ are bandwidth parameters. We estimate the densities in (6) and (8) with,

$$\hat{f}_{\text{sp}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{type}}} \rho_t \hat{f}_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}), \qquad \hat{f}_{\text{sb}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{type}}} \rho_t \hat{f}_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}). \tag{10}$$

The estimators (10) are obviously biased estimators for f. However, they have the potential to reduce MSE. One obvious case is when, for all $t = 1, \ldots, n_{\text{type}}$, there exists weights $w_{p,j,t}, w_{b,k} \approx 1$ and $n_{p,j,t}, n_{b,k,t} > n$. In such settings, $f_{\text{sp}}(\mathbf{y})$ and $f_{\text{sb}}(\mathbf{y})$ have minimal bias when estimating f and can be more efficient than $\hat{f}_h$. Another obvious case is when the batter has never faced the pitcher so that no data is available to estimate f directly, although that does not guarantee that the estimators (10) are good estimators for f. Our implementation will estimate f with

$$\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}) = \lambda \hat{f}_{\mathbf{h}}(\mathbf{y}) + \lambda_p \hat{f}_{\text{sp}}(\mathbf{y}) + \lambda_b \hat{f}_{\text{sb}}(\mathbf{y}) \tag{11}$$

where $\lambda, \lambda_p, \lambda_b$ form a convex combination. The conditioning on $\mathbf{x}_{p,j,t}, \mathbf{x}_{b,k,t}, \rho_t$, for $t = 1, \ldots, n_{\text{type}}$ and $k = 1, \ldots, K$ is suppressed in the density $\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y})$. 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 $n_p = \sum_t \rho_t \sum_{j=1}^{J} s_{p,j,t}^2 n_{p,j,t}$ and $n_b = \sum_t \rho_t \sum_{k=1}^{K} s_{b,k,t}^2 n_{b,k,t}$. These choices arise as a balance between the natural bias that exists in our synthetic player construction and the inherent estimation variation. It is reasonable to assume that $n_{p,j,t} = O(n)$ and $n_{b,k,t} = O(n)$ for all $j = 1, \ldots, J$, all $k = 1, \ldots, K$, and all $t = 1, \ldots, n_{\text{type}}$. 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 (11) having a lower MSE than the estimator (2).

3 Data considerations

Our methodology will consider the following variables comprising $\mathbf{x}_{p,t}$: 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 pitcher-pitch type combination. Our methodology will consider the following variables comprising $\mathbf{x}_{b,t}$: exit velocity, launch angle, pull%, middle%, and oppo%. Averages of these variables are taken across each batter-pitch type combination. One should note that these variables will not allow us to measure the complete talent profile of baseball players. Tools such as speed and eye at the plate will not be fully captured by our methodology.

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 as Eephus, Knuckleball, and Screwball are removed since these pitch types are rare.
- 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:
  - $launch_h = \arctan(\frac{vx_r}{vy_r})$
  - $launch_v = \arctan\left(\frac{vz_r}{\sqrt{vx_r^2+vy_r^2}}\right)$

  where vx_r, vy_r, vz_r are, respectively, the x, y, z components of release velocity.
- 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.

Pitchers are aggregated on a season and pitch type basis and batters are aggregated on a season, handedness, and pitch type basis. To be eligible for comparison, a pitcher must share at least $\lceil \frac{n_{\text{type}}}{2} \rceil$ pitches with the pitcher under study.

4 Shiny app

In this section we present a snapshot of what our Shiny app implementing SEAM methodology offers users. The Shiny app is available at `https://seam.stat.illinois.edu/app/`. The default matchup in the application pairs the 2019 American League Cy Young winner Justin Verlander against the 2019 American League MVP Mike Trout. The layout includes a sidebar with five filters: two dropdowns for batter/pitcher selection, two sliders for metric

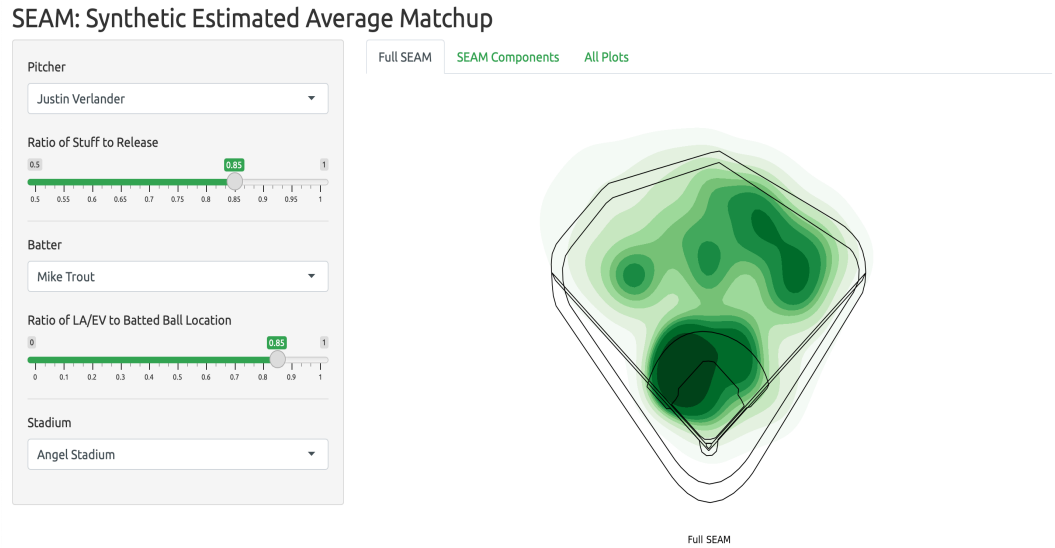

Figure 1: The layout of the application upon submission.

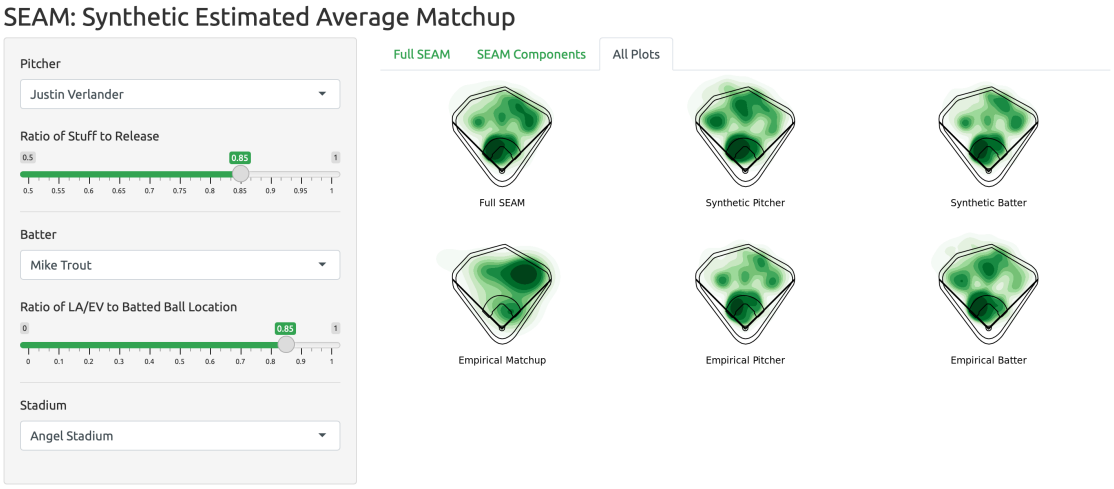

Figure 2: Spray chart distributions constructed by our app. This example corresponds to the spray chart distribution when batter Mike Trout faces pitcher Justin Verlander. The top-left panel is the complete synthetic spray chart for the batter-pitcher matchup. The top-center panel is the synthetic pitcher's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

adjustment, and a dropdown to select the stadium appearance. A snapshot of the appearance of our visualization is depicted in Figures 1 and 2.

The pitcher slider allows users to determine the relative importance of "stuff", a colloquial term for pitch quality, versus release information. Stuff includes velocity, spin rate, and movement. Release includes release angles and release point. The batter slider allows users to determine the relative importance of launch conditions versus batted ball locations. Launch conditions includes exit velocity and launch angle. Location includes pull%, middle%, oppo% (the percentage of batted balls place into the corresponding thirds of a baseball field). The default setting of the pitcher slider favors stuff over release information. The logic for this is quality of pitches being 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.

Note that the same slider weights are applied to each pitch type so that $V_{p,t} = V_p$ in (4), the same holds for batter characteristics. Additionally, our implementation calculates ρ_t as the marginal proportion of pitch types computed across all pitches thrown by the pitcher under study. This design construction is appropriate for the descriptive nature of the analyses presented. For more prescriptive uses, one could adapt more flexible choices of ρ_t to incorporate contextual information known about a batter-pitcher matchup, possibly in real time. Further note that pitch proportions are not considered in our similarity score constructions, therefore we are not accounting for the batter-pitcher meta game.

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 of the distribution of similar batters against the pitcher, the distribution of similar pitchers against the batter, and the distribution of any observations of the pitcher vs batter since 2015.

This matchup presents a good example of how to interpret the resulting spray charts. Trout seems to be a pull-heavy hitter in general according to his traditional chart. When facing pitchers similar to Verlander, he seems to push the ball the opposite way. This may be explained by Verlander's high velocity fastball. In general, batters have a hard time "getting around" (pulling) an upper-90's fastball, so they end up hitting the ball to the opposite field. Given this spray chart distribution, a coach may position the shortstop more towards third base, the second baseman more up the middle, and the first baseman more towards second base. This will protect against Trout's usual habit of pulling the ball, and also put the first baseman in a position to cover the opposite field soft ground ball. If this decision were made just by Trout's traditional chart, the first baseman might not have been moved to cover ground balls through the right side.

5 Discussion

The primary contribution of this work is the development of SEAM methodology in which a synthetic spray chart density function $\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y})$ is estimated. In our context of batter-pitcher matchups, this estimated density function is a weighted average of $\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})$, $\hat{f}_{\text{sp}}(\mathbf{y})$, and $\hat{f}_{\text{sb}}(\mathbf{y})$, where these weights are chosen to minimize MSE under an assumed smooth function space. The synthetic players are constructed to best mimic the players under study. Our method of synthetic player construction is generalizable to other settings in baseball as well as other sports.

We also developed a Shiny app which implements SEAM methodology. This app provides users with visual measures of batter-pitcher matchups and it will be of interest to the broad sports community. Our application shows users batter tendencies versus pitchers and greatly improves upon the inferential power of spray charts (Petti, 2009; Marchi et al., 2019) as a visualization of a batter's talent and 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 talent similarity.

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 spray chart densities to be smooth and lacking 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 assume that $\mathbf{h} = \mathbf{h}_t$, the bandwidth parameters for the batter-pitcher matchup are the same across pitch types. 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 $n_{p,j,t} = O(n)$, $n_{b,k,t} = O(n)$ and $\mathbf{h}_{p,j,t} = O(\mathbf{h})$, $\mathbf{h}_{b,k,t} = O(\mathbf{h})$ for all $j = 1, \ldots, J$, $k = 1, \ldots, K$, $t = 1, \ldots, n_{\text{type}}$. 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

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

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

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

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 now have enough structure to estimate the MSE of (2) and (11). Our multivariate Hölder class specifications yield,

$$\begin{aligned} \text{E}(\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y})) &= \lambda \,\text{E}\, \hat{f}_{\mathbf{h}}(\mathbf{y}) + \lambda_p \,\text{E}\, \hat{f}_{\text{sp}}(\mathbf{y}) + \lambda_b \,\text{E}\, \hat{f}_{\text{sb}}(\mathbf{y}) \\ &= \lambda f(\mathbf{y}) + \lambda \sum_t \rho_t \frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}}{2} + \lambda_p \sum_t \rho_t \sum_{j=1}^{J} w_{p,j,t} \,\text{E}\, \hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}) \\ &\quad + \lambda_b \sum_t \rho_t \sum_{k=1}^{K} w_{b,k,t} \,\text{E}\, \hat{f}_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}) + o(\|\mathbf{h}\|^2) \\ &= \lambda f(\mathbf{y}) + \lambda \sum_t \rho_t \frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) \\ &\quad + \lambda_p \sum_t \rho_t \sum_{j=1}^{J} w_{p,j,t} f_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}) \\ &\quad + \lambda_p \sum_t \rho_t \sum_{j=1}^{J} w_{p,j,t} \frac{\mu_2(K)\mathbf{h}'_{p,j,t}\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}))\mathbf{h}_{p,j,t}}{2} \end{aligned}$$

$$+ \lambda_b \sum_t \rho_t \sum_{k=1}^{K} w_{b,k,t} f_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t})$$
$$+ \lambda_b \sum_t \rho_t \sum_{k=1}^{K} w_{b,k,t} \frac{\mu_2(K)\mathbf{h}'_{b,k,t}\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}))\mathbf{h}_{b,k,t}}{2},$$

and

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

We will define $\tilde{\mathbf{x}}_{b,k,t} = (\mathbf{x}'_{p,t}, \mathbf{x}'_{b,k,t})'$ and $\tilde{\mathbf{x}}_{p,j,t} = (\mathbf{x}'_{p,j,t}, \mathbf{x}'_{b,t})'$ for notational convenience, and will additionally assume the following regularity approximations:

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

A5. The quantities $\sum_{j=1}^{J} w_{p,j,t} \left(\mathbf{h}'_{p,j,t}\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}))\mathbf{h}_{p,j,t} - \mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}\right)$ and $\sum_{k=1}^{K} w_{b,k,t} \left(\mathbf{h}'_{b,k,t}\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}))\mathbf{h}_{b,k,t} - \mathbf{h}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}\right)$ are negligible.

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

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

$$\leq \sum_t \rho_t \left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j,t} w_{p,j,t}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k,t} w_{b,k,t}^2 - 1 \right) \frac{R_{\mathbf{x}_t}(f) f(\mathbf{y}|\mathbf{x}_t)}{n \det(\mathbf{H})}$$
$$+ \lambda_p^2 \sum_t \rho_t \sum_{j=1}^{J} \theta_{p,j,t} w_{p,j,t}^2 \left(\frac{R_{\mathbf{x}_t}(f) \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^2 + 2L f(\mathbf{y}|\mathbf{x}_t) \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^2 + 2L^2 \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^4}{n \det(\mathbf{H})} \right)$$
$$+ \lambda_b^2 \sum_t \rho_t \sum_{k=1}^{K} \theta_{b,k,t} w_{b,k,t}^2 \left(\frac{R_{\mathbf{x}_t}(f) \|\mathbf{x}_t - \tilde{\mathbf{x}}_{b,k,t}\|^2 + 2L f(\mathbf{y}|\mathbf{x}_t) \|\mathbf{x}_t - \tilde{\mathbf{x}}_{b,k,t}\|^2 + 2L^2 \|\mathbf{x}_t - \tilde{\mathbf{x}}_{b,k,t}\|^4}{n \det(\mathbf{H})} \right).$$

Assumption A4 and an identical lower bound argument implies that $\mathrm{Var}(\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y})) - \mathrm{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y}))$ is approximately bounded above by

$$\sum_t \rho_t \left(\lambda^2 + \lambda_p^2 \sum_{j=1}^{J} \theta_{p,j,t} w_{p,j,t}^2 + \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k,t} w_{b,k,t}^2 - 1 \right) \frac{R_{\mathbf{x}_t}(f) f(\mathbf{y}|\mathbf{x}_t)}{n \det(\mathbf{H})} + O\left(\frac{1}{n}\right).$$

We also have

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

and regularity approximations A4 and A5 yield

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

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

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

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 $n_p = \sum_t \rho_t \sum_{j=1}^{J} s_{p,j,t}^2 n_{p,j,t}$ and $n_b = \sum_t \rho_t \sum_{k=1}^{K} s_{b,k,t}^2 n_{b,k,t}$. We will now develop intuition for these choices. First, notice that $\lambda_p, \lambda_b \to 0$ as $\min_j(\|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|), \min_k(\|\mathbf{x}_t - \tilde{\mathbf{x}}_{b,k,f}\|) \to \infty$ for all $t = 1, \ldots, n_{\text{type}}$. 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

$$
\begin{aligned}
&\lambda_p \sum_t \rho_t \sum_{j=1}^{J} (-1)^z L w_{p,j,t} \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^2 \\
&\quad = \frac{\sqrt{\sum_t \rho_t \sum_{j=1}^{J} s_{p,j,t}^2 n_{p,j,t}} \left(\sum_t \rho_t \sum_{j=1}^{J} (-1)^z L w_{p,j,t} \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^2 \right)}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}} \longrightarrow 0,
\end{aligned}
$$

when there exists some j' such that $\|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j',t}\| \to 0$ or $\min_j(\|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|) \to \infty$ for each pitch type $t = 1, \ldots, n_{\text{type}}$. These cases correspond, respectively, to there being a few highly similar pitchers or there being no similar pitchers for each pitch thrown by the pitcher under study. Thus, the discrepancy in bias vanishes in the extreme cases. The same argument holds for batters. Now notice that

$$
\begin{aligned}
\lambda_p^2 \sum_t \rho_t \sum_{j=1}^{J} \theta_{p,j,t} w_{p,j,t}^2 &= \frac{\left(\sum_t \rho_t \sum_{j=1}^{J} s_{p,j,t}^2 n_{p,j,t} \right) \left(\sum_t \rho_t \sum_{j=1}^{J} \theta_{p,j,t} w_{p,j,t}^2 \right)}{(\sqrt{n} + \sqrt{n_p} + \sqrt{n_b})^2} \\
&\longrightarrow \begin{cases} 0, & \min_{j,t}(\|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|) \to \infty; \\ \frac{\sum_t \rho_t n_{p,j_t,t}}{(\sqrt{n} + \sqrt{\sum_t \rho_t n_{p,j_t,t}} + \sqrt{n_b})^2}, & w_{p,j_t,t} \to 1, \text{ for all } t = 1, \ldots, n_{\text{type}}, \end{cases}
\end{aligned}
$$

under the specifications that $\theta_{p,j_t,t} = 1$. The same argument holds for batters under the specifications that $\theta_{b,k_t,t} = 1$. Therefore, when there is a pitcher j_t and batter k_t so that

$w_{p,j_t,t}, w_{b,k_t,t} \to 1$ for each pitch type $t = 1, \ldots, n_{\text{type}}$, we have that

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

Our choices of the elements of $\boldsymbol{\lambda}$ will work well in the presence or absence of pitchers and batters that recover the traits of the players under study. Less is known about middle ground cases, especially when sample sizes are are small.

References

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

Baccellieri, E. (2022). Infield shifts are increasing in the final season before mlb restricts them. https://www.si.com/mlb/2022/04/26/shifts-increasing-the-opener. Accessed: 2022-06-13.

Baseball-Prospectus (2022). https://www.baseballprospectus.com/. Accessed: 2022-06-13.

Baseball-Reference (2022). https://www.baseball-reference.com. Accessed: 2022-06-13.

Baseball-Savant (2014). Statcast. https://baseballsavant.mlb.com/. Accessed: 2022-06-13.

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

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

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

Fangraphs (2022). https://www.fangraphs.com. Accessed: 2022-06-13.

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

info solutions, B. (2022). www.baseballinfosolutions.comm. Accessed: 2022-06-13.

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

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

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

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

Petti, B. (2009). The interactive spray chart tool. https://billpetti.shinyapps.io/shiny_spraychart/. Accessed: 2020-04-29.

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

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

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

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

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

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.