SEAM methodology for context-rich player matchup evaluations in baseball

Julia Wapner[1]*, David Dalpiaz[2]†, and Daniel J. Eck[2]‡

1. Baseball Analytics, Baltimore Orioles
2. 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 estimated spray chart distribution are constructed to minimize expected mean square error. We provide a Shiny web application that allows users to visualize and evaluate any batter-pitcher matchup that has occurred or could have occurred during the Statcast era (specifically 2017-present). This methodology and web application 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 web application at

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

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

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

1 Introduction

Baseball has a rich statistical history dating back to the first box score created by Henry Chadwick in 1859. Throughout the years, fans of the game, academics, and professionals alike have 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., 2009a,b, Piette and Jensen, 2012, Baumer et al., 2015, Marchi et al., 2019]. Most statistics arrive from the box score, a simple summary 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 organizations 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 (and 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 are 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 (often referred to as exit velocity), and the horizontal and vertical angles in which they hit the ball (often referred to as spray and launch angles respectively). When a ball is put into play it has a final location that is relevant to, but not included 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 several years prior to 2015 [info solutions, 2022].

This high resolution data has ushered in a new sophisticated class of techniques for finding and exploiting players' strengths and weaknesses. Use of these techniques have affected player performance and outcomes, which is reflected but noted in the boxscore. As an example, Jensen et al. [2009b] showed how this data could be used to better quantify defensive ability using Baseball Information Solutions data through 2007. One of the most prominent examples of the success of this high-resolution data has been the implementation of defensive shifting based on the spray chart [Petti, 2009, Marchi et al., 2019], a plot of locations a batter has hit the ball in the past. Spray charts are used extensively by MLB organizations for fielder positioning [Marchi et al., 2019, page 271]. Through the early part of the 2022 MLB season, MLB had seen infield shifts on 37% of pitches [Baccellieri, 2022]. This rate of deployment demonstrates the effectiveness of the strategy.

Even though the shift has a demonstrated record of success, the strategy of the shift can be justified by batters' empirical spray chart or pitchers' empirical spray chart allowed. These charts are compiled marginally across all matchups and they therefore do not account for the interaction effect when a particular batter faces a particular pitcher. However, spray charts for specific batter-pitcher matchups are not often able to be reliably constructed due to sparse or non-existent matchup data.

In this article, we develop SEAM methodology which alleviates the challenge posed by sparse matchup data through a principled incorporation of batted-ball outcomes of players that are similar to the batter and pitcher under study. The SEAM method poses each batter-pitcher matchup as having a spray chart distribution which are 2-dimensional spatial distributions representing the potential batted-ball locations when a particular batter faces off against a particular pitcher. Estimation of the specific batter-pitcher matchup involves batted-ball locations from three sources: 1) those from the batter under study vs the pitcher under study; 2) those from the batter under study vs a synthetic version of the pitcher under study; 3) those from a synthetic version of the batter under study vs the pitcher under study.

Here the batted-ball locations corresponding to synthetic versions of the players under study are taken across all other MLB players, the locations are weighted to reflect similarity to the player under study where similarity ranges from 0 (not considered at all) to 1 (a clone of the player under study up to the information used to compute the similarity score). These three components represent separate spray chart distributions and they are combined through a convex combination with weights reflecting overall similarity of the synthetic players to those under study where these weights collapse on the specific batter-pitcher empirical spray distribution when enough data is available to reliably estimate this distribution.

Mathematical justification for the combination of these estimated spray distributions is provided. We also demonstrate that our SEAM method outperforms the empirical batter or pitcher spray chart distributions through a detailed validation study in which we hold out data from the 2021 season and determine how well $(1 - \alpha) \times 100\%$ highest density regions include the batted-ball locations from the held out data. SEAM exhibits nominal marginal coverage, but is no better than the existing batter or pitcher spray charts. SEAM outperforms batter or pitcher spray charts when we evaluate conditional coverage, indicating that our approach is better suited for predicting batted-ball locations. That being said, the SEAM confidence region is larger than those constructed from empirical batter or pitcher spray chart distributions. We then consider fixed-size confidence regions as a compromise between size and coverage. We find that the SEAM confidence region has higher average conditional coverage than its competitors when we compare fixed-size confidence regions. The size of these fixed-size regions is chosen to reflect the proportion of a baseball field that fielders could cover. Therefore, the SEAM method exhibits an important advantage over its competitors.

We argue that the advantage that SEAM offers translates to an additional win over the course of the season. MLB organizations have repeatedly demonstrated a willingness to spend millions of dollars to obtain a comparable advantage via free agency [**?**]. The importance of winning one additional game per season can not be overstated. Had the 2011

Atlanta Braves won one additional game, they would have forced a one-game tiebreaker between them and the St. Louis Cardinals where the winner of this tiebreaker goes to the playoffs and the loser goes home. Instead, the Braves missed the playoffs altogether and the Cardinals won the World Series in historic fashion[1].

2   Setting

Suppose that left-handed batter Joey Votto is about to face right-handed pitcher Justin Verlander, and further suppose that the fielding team desires to employ a Votto-specific shift. Both players are unique talents and they seldom face each other, this matchup has only produced two balls in play. Therefore, past matchup data is too sparse to be of much use. The fielding team could instead select fielder locations based on an estimated density constructed from Votto's spray chart or Verlander's spray chart allowed. These approaches suggest different fielder positioning schemes, see Figure 1. In this article, we propose the SEAM (synthetic estimated average matchup) method as a solution to these problems. The SEAM method provides a batter-pitcher specific spray chart density that can be used for more reliable fielder positioning. We will demonstrate that SEAM is superior to positioning strategies that are formed from exclusive use of either the batter or pitcher data. See Figure 1 for the estimated SEAM spray chart density corresponding to the Joey Votto vs. Justin Verlander matchup.

3   SEAM method

3.1   Spray chart distribution preliminaries

In this article, we seek to provide a batter-pitcher specific spray chart density which is more accurate than spray chart densities estimated exclusively from either batter or pitcher

---

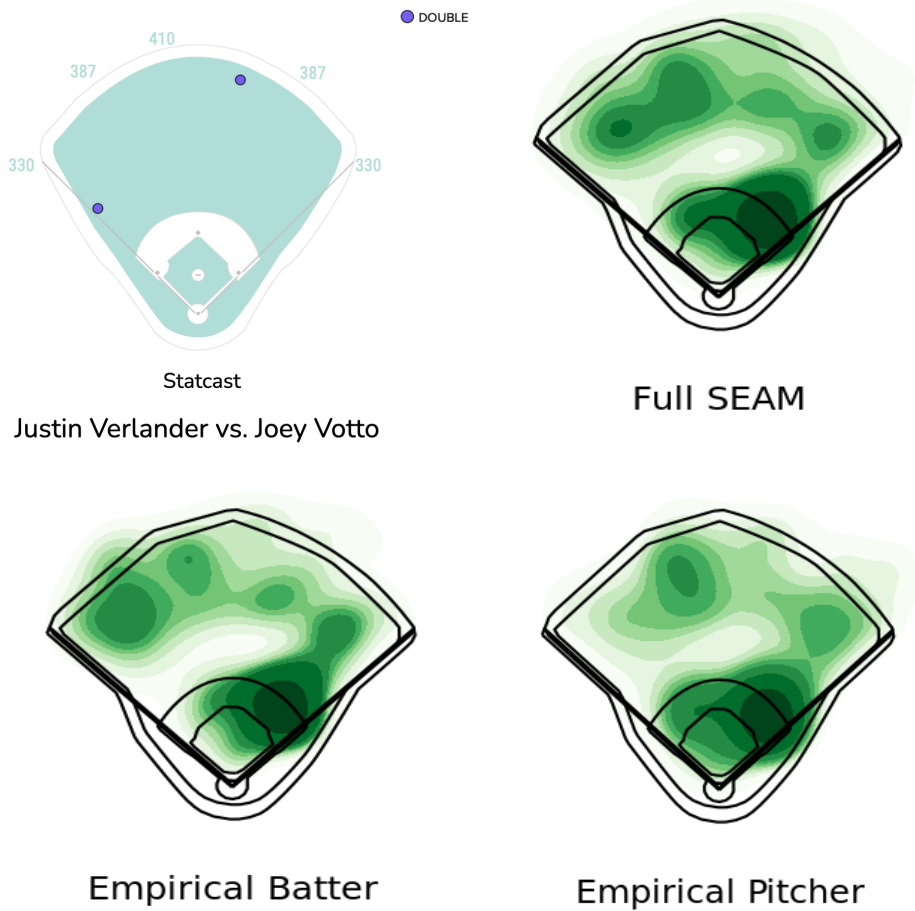[1] https://www.youtube.com/watch?v=46QITnnPINY&ab_channel=SecretBase

Figure 1: Visualization of potential batted-ball locations corresponding to the Joey Votto vs. Justin Verlander matchup according to past data (top-left), our SEAM method (top-right), Joey Votto's estimated spray chart density against all right-handed pitchers (bottom-left), and Justin Verlander's estimated spray chart allowed density against left-handed batters (bottom-right).

data marginally. The densities that we estimate correspond to what we call a spray chart distribution which is a spatial distribution over a bounded subset $\mathcal{Y} \in \mathbb{R}^2$. The set $\mathcal{Y}$ contains plausible locations of batted balls from home plate. In our application we will let $(0, 0) \in \mathcal{Y}$ denote the location of home plate. A spray chart distribution for a batter is a distribution F over a bounded subset $\mathcal{Y} \in \mathbb{R}^2$.

3.2 Nonparametric estimation

A spray chart distribution F is not likely to have a known parametric form so we will consider nonparametric estimation techniques to estimate spray chart densities f. The particular nonparametric estimation methods that we use are chosen for their computational speed and theoretical properties. Let $(y_{1i}, y_{2i}) \in \mathcal{Y}$, $i = 1, \ldots, n$, be the observed batted-ball locations, then 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 variants of the `kde2d` and `kde2d.weighted` functions in R [Ripley et al., 2019, Hamilton, 2018]. Therefore, we estimate f using a bivariate nonparametric Gaussian kernel density estimator

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

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 $\mathbf{x} \in \mathcal{X}$.

3.3   Similarity scores

The estimator (3) is often not feasible for $f(\cdot|\mathbf{x})$ in practice since there is not enough individual matchup data. To address this challenge we will incorporate batted-ball data from other matchups involving one of the batter or pitcher under study where the batted-ball locations from such matchups are weighted by their similarity to the players in the matchup under study. Similarity will be assessed via similarity scores which are motivated by James [1994] and Silver [2003].

We now describe our procedure for determining similarity scores. We will suppose that there are J pitchers who allowed a ball in play against the batter under study and K batters who put a ball in play against the pitcher under study. 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,

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

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 web application have some control of the entries of $\mathbf{V}_{p,t}$ in (4) by adjusting the pitcher slider. Similarity scores between batters k_1 and k_2 are defined in a similar manner and are denoted 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 is the proportion of time that a ball in play yielded by the pitcher in the matchup under study corresponds to pitch type t. We therefore represent $f(y)$ as $\sum_t \rho_t f(y|\mathbf{x}_t)$, where $\mathbf{x}_t = (\mathbf{x}'_{p,t}, \mathbf{x}'_{b,t})'$.

3.4   The SEAM density function

We now describe the construction and estimation of the SEAM density function. The SEAM density function is a linear combination of three density functions, the spray chart density $f(\cdot|\mathbf{x})$, and two "synthetic" player density functions, the synthetic pitcher density function $f_{\text{sp}}(\cdot|\mathbf{x})$ and the synthetic batter density function $f_{\text{sb}}(\cdot|\mathbf{x})$. Both $f_{\text{sp}}(\cdot|\mathbf{x})$ and $f_{\text{sb}}(\cdot|\mathbf{x})$ are weighted averages of spray chart densities corresponding to matchups in which the batter (pitcher) under study faces every other pitcher (batter). Attention is restricted to matchups that yielded a ball in play. We now describe these synthetic densities in more detail, describe their estimation, and present an estimator of the SEAM density function.

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

10

density $f_{\mathrm{sb}}(\mathbf{y})$.

We then estimate (5) and (7) with

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

(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,

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

(10)

The estimators (10) are biased estimators for f. However, they have the potential to reduce MSE. One obvious case is when, for all $t = 1,\ldots,n_{\mathrm{type}}$, 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, $f_{\mathrm{sp}}(\mathbf{y})$ and $f_{\mathrm{sb}}(\mathbf{y})$ 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

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

(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 $t = 1,\ldots,n_{\mathrm{type}}$ and $k = 1,\ldots,K$ is suppressed in the density $\hat{f}_{\boldsymbol{\lambda}}(\mathbf{y})$. 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).

4   Validation

We evaluate the performance of our SEAM method through a validation approach in which we train the SEAM method on data through the 2020 season and then test the coverage of highest density regions with 2021 data. We define coverage as the proportion of batted balls that land in a location on the field that is within a level α highest density region. We consider both marginal and conditional coverage. Marginal coverage is assessed across all 2021 batted balls. Conditional coverage is assessed for 2021 batter-pitcher matchups that yielded 10 or more balls in play. We compute the proportion of individual batter-pitcher conditional coverages that is at or above the nominal level, which we call conditional coverage success rate.

We compare the coverage properties of our SEAM method to empirical batter spray charts and empirical pitcher spray charts which are estimated marginally across all pitchers and batters respectively using similar nonparametric estimation techniques. The results are included in Table 1. We see that the highest density regions for each method achieve nominal marginal coverage. This is especially encouraging for SEAM since it is expected

	Marginal coverage			Conditional coverage success rate		
level	0.5	0.75	0.9	0.5	0.75	0.90
SEAM	0.559	0.811	0.953	0.579 (0.035)	0.641 (0.034)	0.779 (0.030)
batter	0.547	0.793	0.938	0.513 (0.036)	0.574 (0.035)	0.662 (0.034)
pitcher	0.541	0.788	0.935	0.549 (0.036)	0.595 (0.035)	0.713 (0.032)

Table 1: Marginal and conditional coverage properties for SEAM, the empirical batter spray distribution, and the empirical pitcher spray distribution (standard errors in parentheses). The conditional coverage success rate is the proportion of individual batter-pitcher conditional coverages that is at or above the nominal level.

that the empirical batter and pitcher spray charts should obtain marginal coverage by their construction, provided that the players under study did not dramatically change in a material way. We also see that the conditional coverage of the SEAM method is noticeably and usefully higher than the other methods.

However, the highest density region for SEAM is slightly larger than its competitors. As an example, the level $\alpha = 0.10$ highest density region for SEAM includes 34% of grids forming the rectangle that contains the playing field while similar highest density regions corresponding to the empirical batter and pitcher spray distributions include, respectively, 33.4% and 33% of such grids.

We will now consider a different type of confidence region which will balance these conflicting conclusions with size and coverage. This region is constructed using the n_g grids containing the highest probable batted ball locations as estimated by each method. Calculating average conditional coverages for fixed-size regions is an important consideration for fielding placement due to the physical constraint on how much of the baseball field defenders can reasonably cover. We considered [1500, 2500] as a range for the number of grids. This range corresponds to fixed-size regions that cover approximately [50%, 80%] of fair territory of a baseball field that is 115,000 square feet.

We then assess the average conditional coverage of these fixed-size regions. What we find is that the SEAM fixed-size region has higher average conditional coverage than its competitors which ignore matchup considerations. This is despite the fact that the cumulative

probability mass of the SEAM fixed-size region is lower than its competitors. The results are in Figure 2. Note that were 129937 baseballs put into play in 2021. Thus, an increase in coverage by 1% roughly corresponds to an increase of fielding 40 outs per team, and a difference of 40 outs corresponds to about one and half baseball games. It is reasonable to expect that any baseball team that implements our SEAM method could win an additional game over the course of the season. As stated in the Introduction, MLB organizations have repeatedly demonstrated a willingness to spend millions of dollars to obtain a comparable advantage via free agency [?], and an additional win could be all that is needed to have a chance at competing for a championship.

We now explore statistical significant differences between the methods depicted in Figure 2. Individual coverages for particular matchups are quite variable. Therefore, point-wise differences between methods are not statistically significant. However, a linear regression model with coverage as the response, and method and the number of grids as predictors, reveals that the SEAM main effect is statistically significantly greater than those corresponding to competing methods. But classical tests of regression coefficients are dubious because diagnostic plots reveal modest violations of modeling assumptions. Violations of modeling assumptions were expected since the response variable lives in $[0, 1]$. A nonparametric bootstrap of ordinary least squares (OLS) estimates was then performed with bootstrap sample size set at 10000. This analysis agreed with the original regression analysis, the proportion of bootstrap iterates for which the estimated coefficient corresponding to the SEAM method was exceeded by estimated coefficient corresponding to the empirical batter and empirical pitcher spray chart distributions was, respectively, 0 and 0.0147. More comprehensive results are depicted in Figure 3. This analysis suggests that the advantage offered by the SEAM method is indeed statistically significant. Note that a bootstrap analysis revealed that the regression model presented here was preferred over a larger model which included a separate slope coefficient for the empirical pitching method.
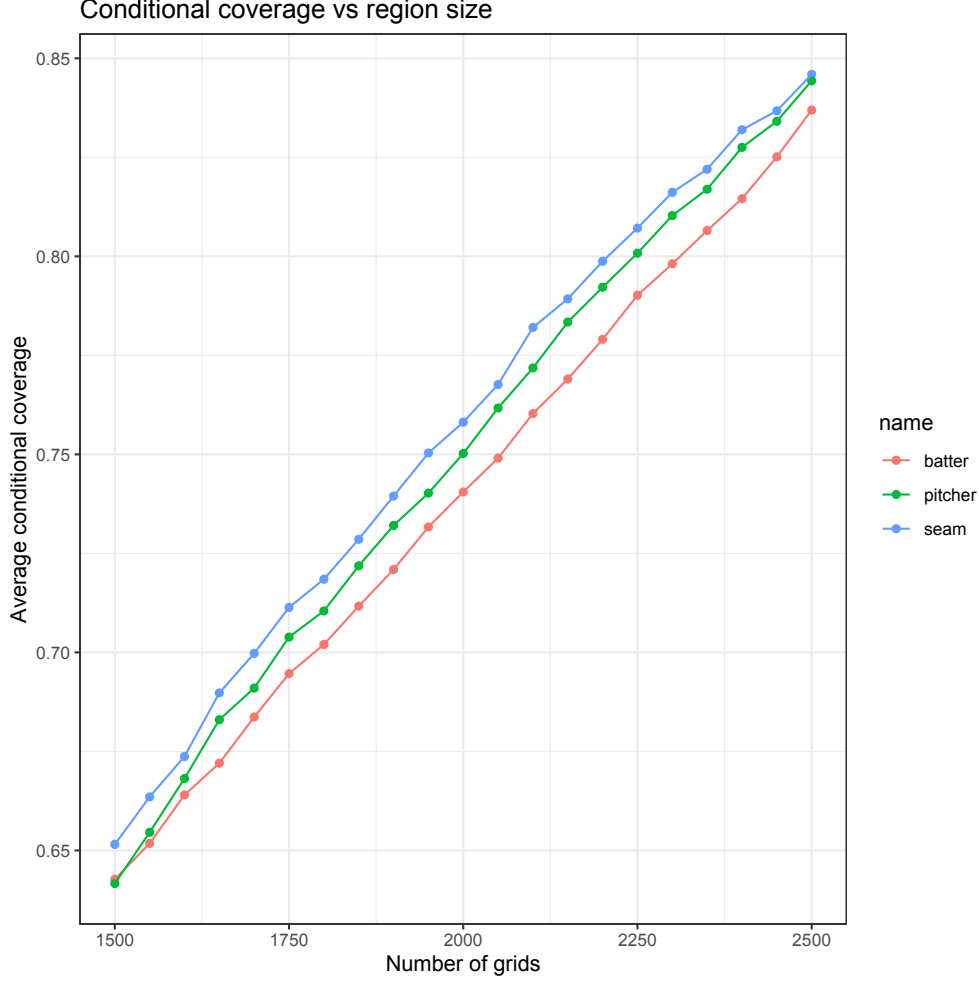
Figure 2: Average conditional coverage of fixed-size confidence regions for SEAM, the empirical batter spray distribution, and the empirical pitcher spray distribution. The fixed-size regions are constructed from the n_g most probable grids according to each method. The proportion of iterates that are less than 0 is 0 (left panel) and 0.0147 (right panel).

5 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-year combination. Our methodology will consider the following variables comprising $\mathbf{x}_{b,t}$: exit velocity, launch angle, left%, center%, and right%. Averages of these variables are taken across each batter-pitch type-year combination. We separate for
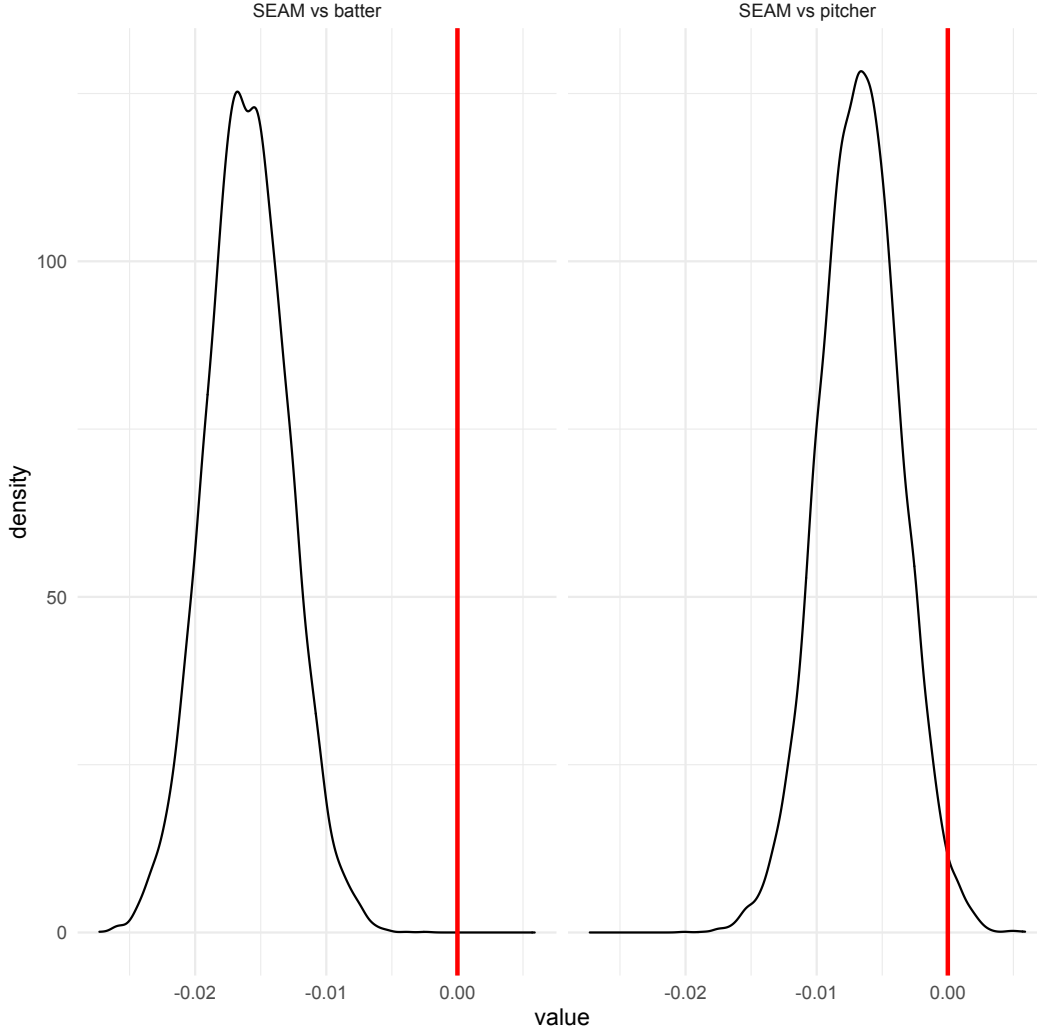
15

Figure 3: Nonparametric bootstrap investigation of the statistical significance of the increased coverage offered by our SEAM method. The distributions depicted are differences between OLS estimated coefficients between the empirical batter (left panel) and pitcher (right panel) spray chart distributions and the SEAM method. Negative values indicate preference for the SEAM method.

year to account for subtle season-to-season shifts in pitcher and batter characteristics. One should note that these variables will not allow us to measure the complete talent profile of baseball players. Skills such as speed and eye at the plate will not be fully captured by our methodology.

Data for our web application was acquired via Statcast [Baseball-Savant, 2014]. Our data set contains every pitch that was thrown since the start of the 2017 season. We ignored

Statcast data pre-2017 due to data integrity issues. 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 make home plate the origin $(0,0)$ [Petti, 2017].

- Spray angle [Petti, 2017] is calculated from the x and y coordinates of the batted ball locations (what we previously called y_1 and y_2).

- Data was limited to regular season batter-pitcher matchups.

- Sacrifice hits and sacrifice flies are removed from consideration.

  Pitchers and batters are aggregated on a season, handedness, and pitch type basis.

6   Interactive web application

In this section we present a snapshot of what our Shiny web application implementing SEAM methodology offers users. The web application is available at
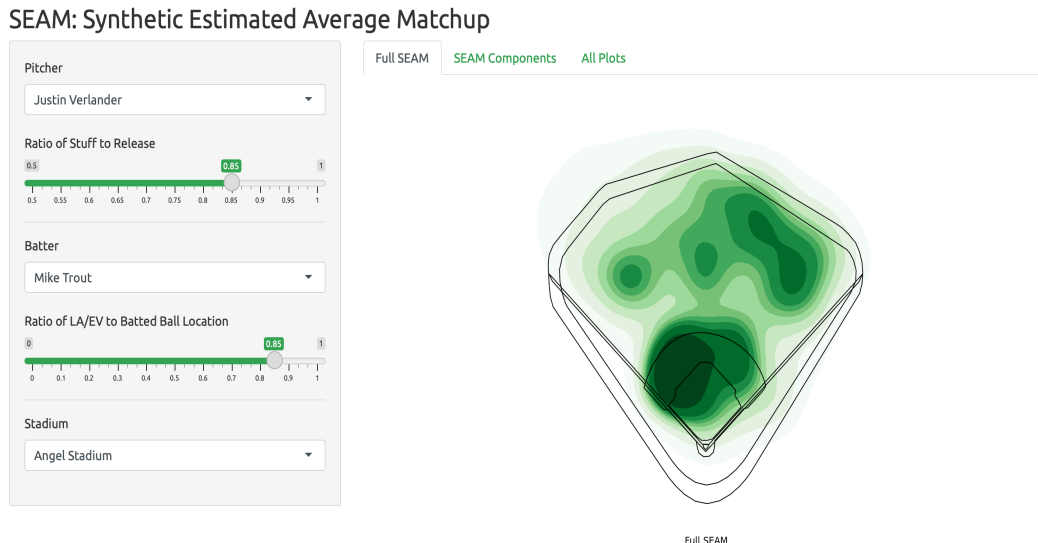
https://seam.stat.illinois.edu/

Figure 4: The layout of the web application upon submission.

The default matchup in the web application pairs the 2019 American League Cy Young winner Justin Verlander against the 2019 American League MVP Mike Trout, two of the greatest players of all-time [Yan et al., 2022]. Note that our matchup will assess the outcomes currently, rather than during their award winning seasons. The layout includes a sidebar with five filters: two dropdowns for batter/pitcher selection, two sliders for metric adjustment, and a dropdown to select the stadium appearance. A snapshot of the appearance of our visualization is depicted in Figures 4 and 5.

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
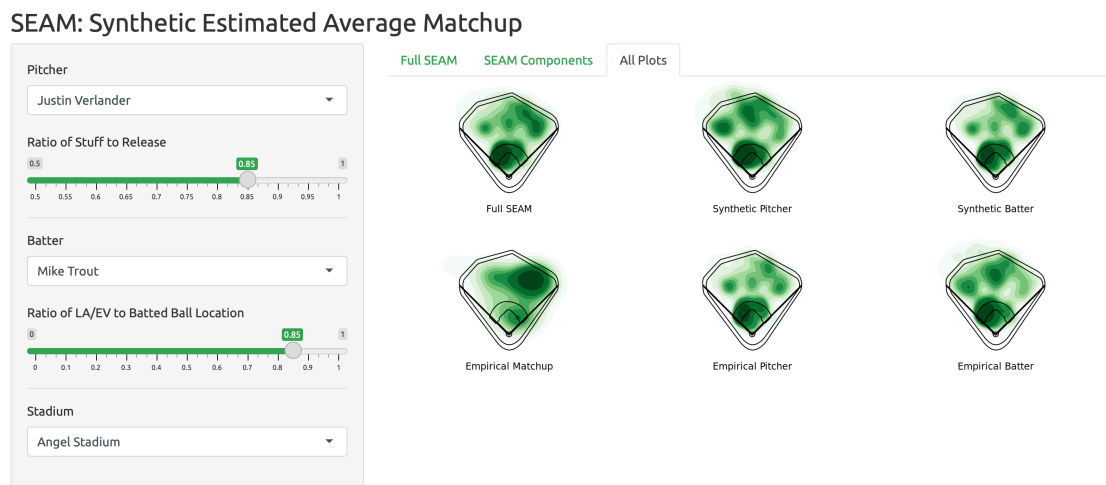
Figure 5: Spray chart distributions constructed by our web application. This example corresponds to the spray chart distribution when batter Mike Trout faces pitcher Justin Verlander. The top-left panel is the complete synthetic spray chart for the batter-pitcher matchup. The top-center panel is the 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.

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 that yielded a ball in play. 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.

7 Discussion

The primary contribution of this work is the development of SEAM methodology in which a synthetic spray chart density function $f_{\boldsymbol{\lambda}}(\mathbf{y})$ is estimated. In our presented context of batter-pitcher matchups, this estimated density function is a weighted average of $\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})$, $\hat{f}_{\mathrm{sp}}(\mathbf{y})$, and $\hat{f}_{\mathrm{sb}}(\mathbf{y})$, 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. It was demonstrated that our SEAM method provided more accurate predictions of future batted-ball locations than spray chart distributions computed from either the batter or pitcher in isolation. Our method of synthetic player construction is generalizable to other settings in baseball as well as other sports. 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.

We also developed an interactive web application which implements and provides visualizations of our SEAM methodology. This web application is fast, providing users with visual measures of batter-pitcher matchups nearly instantaneously. Our methodology displays an improvement upon empirical batter or pitcher spray charts which are computed marginally [Petti, 2009, Marchi et al., 2019]. Additionally, matchup-specific spray charts are often not informative due to a lack of data. Our synthetic player construction alleviates this problem.

As previously mentioned, the visualizations provided by our web application 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 2017.

The default matchup between Justin Verlander and Mike Trout provides a good example of how to interpret our SEAM spray chart distribution estiamtes. Trout seems to be a pull-heavy hitter in general according to his traditional marginal spray 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.

References

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

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

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

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

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

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.

Ben Clemens. https://blogs.fangraphs.com/what-are-teams-paying-per-war-in-free-agency/, 2021. Accessed: 2022-08-29.

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

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

Baseball info solutions. www.baseballinfosolutions.comm, 2022. Accessed: 2022-06-13.

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.

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

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

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

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

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.

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

Shen Yan, Adrian Burgos Jr, Christopher Kinson, Brandon Niedert, and Daniel J Eck. Comparing baseball players across eras via the novel full house model. *arXiv preprint arXiv:2207.11332*, 2022.

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

$$|D_{\mathbf{y}}^{\mathbf{s}} f(\mathbf{y}|\mathbf{x}) - D_{\mathbf{y}}^{\mathbf{s}} f(\mathbf{y}'|\mathbf{x})| \leq L_{\mathbf{x}} \|\mathbf{y} - \mathbf{y}'\|^{\beta - |\mathbf{s}|},$$

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

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 $D_{\mathbf{y}}^{\mathbf{s}} = \partial^{s_1 + s_2} / \partial y_1^{s_1} \partial y_2^{s_2}$, $D_{\mathbf{x}}^{\mathbf{t}} = \partial^{t_1 + \cdots + t_p} / \partial x_1^{t_1} \cdots \partial x_p^{t_p}$ 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 $\mathbf{H} = \mathrm{diag}(\mathbf{h})$. 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

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

and

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

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

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

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

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

$$\mathrm{E}(\hat{f}_{\boldsymbol{\lambda}}(\mathbf{y})) = \lambda \mathrm{E} \hat{f}_{\mathbf{h}}(\mathbf{y}) + \lambda_p \mathrm{E} \hat{f}_{\mathrm{sp}}(\mathbf{y}) + \lambda_b \mathrm{E} \hat{f}_{\mathrm{sb}}(\mathbf{y})$$

$$= \lambda f(\mathbf{y}) + \lambda \sum_t \rho_t \frac{\mu_2(K)\mathbf{h}'\mathrm{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}}{2} + \lambda_p \sum_t \rho_t \sum_{j=1}^J w_{p,j,t}\,\mathrm{E}\,\hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t},\mathbf{x}_{b,t})$$

$$+ \lambda_b \sum_t \rho_t \sum_{k=1}^K w_{b,k,t}\,\mathrm{E}\,\hat{f}_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t},\mathbf{x}_{b,k,t}) + o(\|\mathbf{h}\|^2)$$

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

$$+ \lambda_p \sum_t \rho_t \sum_{j=1}^J w_{p,j,t} f_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t},\mathbf{x}_{b,t})$$

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

$$+ \lambda_b \sum_t \rho_t \sum_{k=1}^K w_{b,k,t} f_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t},\mathbf{x}_{b,k,t})$$

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

and

$$\mathrm{Var}(\hat{f}_{\boldsymbol{\lambda}}(\mathbf{y})) = \mathrm{Var}\left(\lambda\hat{f}_{\mathbf{h}}(\mathbf{y}) + \lambda_p\hat{f}_{\mathrm{sp}}(\mathbf{y}) + \lambda_b\hat{f}_{\mathrm{sb}}(\mathbf{y})\right)$$

$$= \lambda^2 \sum_t \rho_t \frac{R_{\mathbf{x}_t}(f)f(\mathbf{y}|\mathbf{x}_t)}{n\det(\mathbf{H})} + O\left(\frac{1}{n}\right) + \lambda_p^2 \sum_t \rho_t \sum_{j=1}^J w_{p,j,t}^2\,\mathrm{Var}\,\hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t},\mathbf{x}_{b,t})$$

$$+ \lambda_b^2 \sum_t \rho_t \sum_{k=1}^K w_{b,k,t}^2\,\mathrm{Var}\,\hat{f}_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t},\mathbf{x}_{b,k,t})$$

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

$$+ \lambda_b^2 \sum_t \rho_t \sum_{k=1}^K w_{b,k,t}^2 \frac{R_{\tilde{\mathbf{x}}_{b,k,t}}(f)f(\mathbf{y}|\mathbf{x}_{p,t},\mathbf{x}_{b,k,t})}{n_{b,k,t}\det(\mathbf{H}_{b,k,t})},$$

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 $\sum_{j=1}^{J} w_{p,j,t} \left(\mathbf{h}_{p,j,t}' \operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t})) \mathbf{h}_{p,j,t} - \mathbf{h}' \operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t)) \mathbf{h} \right)$ and

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

Approximation A4 is reasonable in our baseball application where there are many play-ers 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 $\beta = 2$ im-plies that $\|\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t})) - \operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\| \leq \sqrt{d_p} L$ and $\|\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t})) - \operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\| \leq \sqrt{d_b} L$ 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

$$\operatorname{Var}(\hat{f}_{\boldsymbol{\lambda}}(\mathbf{y})) - \operatorname{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y})) + O\left(\frac{1}{n}\right)$$

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

$$+ \lambda_b^2 \sum_{k=1}^{K} \theta_{b,k,t} w_{b,k,t}^2 \frac{R_{\tilde{\mathbf{x}}_{b,k,t}}(f) f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t})}{n \det(\mathbf{H})}$$

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

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

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

Assumption A4 and an identical lower bound argument implies that $\operatorname{Var}(\hat{f}_{\boldsymbol{\lambda}}(\mathbf{y})) - \operatorname{Var}(\hat{f}_{\mathbf{h}}(\mathbf{y}))$ 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

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

and regularity approximations A4 and A5 yield

$$
\begin{aligned}
\text{Bias}(\hat{f}_{\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}'\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{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}\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}))\mathbf{h}_{p,j,t}}{2} \\
&\quad + \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}\text{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}'\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{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}'\text{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}}{2} \\
&\quad + \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}'\text{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 \\
&\quad + \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) \Bigg)^2,
\end{aligned}
$$

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

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

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

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

$$\lambda_p \sum_t \rho_t \sum_{j=1}^J (-1)^z L w_{p,j,t} \|\mathbf{x}_t - \tilde{\mathbf{x}}_{p,j,t}\|^2$$

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

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

$$\lambda_p^2 \sum_t \rho_t \sum_{j=1}^J \theta_{p,j,t} w_{p,j,t}^2 = \frac{\left(\sum_t \rho_t \sum_{j=1}^J s_{p,j,t}^2 n_{p,j,t} \right) \left(\sum_t \rho_t \sum_{j=1}^J \theta_{p,j,t} w_{p,j,t}^2 \right)}{(\sqrt{n} + \sqrt{n_p} + \sqrt{n_b})^2}$$

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

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.