SEAM methodology for context-rich player matchup evaluations in baseball

Abstract

We develop SEAM (synthetic estimated average matchup) methodology which can be used to evaluate batter versus pitcher matchups, both numerically and visually. We first estimate the distribution of balls put into play by a batter facing a pitcher, called the spray chart distribution. This distribution is conditional on batter and pitcher characteristics. These characteristics are a better expression of talent than any conventional statistics. Many individual matchups have a sample size that is too small to be reliable. Synthetic versions of the batter and pitcher under consideration are constructed in order to alleviate these concerns. Weights governing how much influence these synthetic players have on the overall spray chart distribution are constructed to minimize expected mean square error. We provide novel performance metrics that are calculated as expectations taken with respect to the spray chart distribution. These performance metrics provide a context rich approach to player evaluation. We also 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. One can access this app at `https://seam.shinyapps.io/seam/`. Our methodology and interactive tool has utility for anyone interested in baseball as well as team executives and players.

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

Most baseball statistics used for player evaluations are obtained from raw box score totals. While box score totals are an enjoyable statistical summary for baseball fans and analysts, the information contained in them is not very substantive. They ignore rich contextual information such as era, opposing team strength, and individual level matchup effects. Most commonly used player evaluation metrics are functions of context-free box score totals. These include, and are far from limited to, adjusted earned run average (ERA+), adjusted on base plus slugging percentage (OPS+), weighted runs created plus (wRC+), and wins above replacement (WAR) (Baseball-

Reference, 2020; Fangraphs, 2020). More sophisticated techniques in Berry et al. (1999), Brown (2008), Jensen et al. (2009a), and Baumer et al. (2015) also constructed methodology grounded in raw box score totals. While many of these tools account for some contextual information such as ball parks, position of a player, and a player's age, they ignore opponent strength. Eck (2020) showed that context-free metrics and the class of metrics that compares a player's accomplishments directly with that player's peers are ill-equipped for player comparisons across eras of baseball, although they may perform well over the course of a single season or a few consecutive seasons. That being said, these context-free metrics do not offer any guidance for how any particular batter will perform against a particular pitcher, the most important and relevant outcome in baseball. Furthermore, baseball outcomes have been assumed to be independent and identically distributed (iid) realizations in the literature (Brown, 2008; Jensen et al., 2009a). The iid assumption of outcomes may be reasonable in the prediction contexts of Brown (2008) and Jensen et al. (2009a) that involve long time frames, but this assumption is not appropriate for small time frames when the variability in quality of batter and pitcher characteristics can be very large.

In this article we develop SEAM methodology that can be used to evaluate batter-pitcher matchups visually and numerically. This methodology is built upon spray chart distributions which are 2-dimensional spatial distributions representing the potential batted-ball locations when a particular batter faces off against a particular pitcher. Spray chart distributions provide contours that overlay traditional spray charts (Petti, 2009; Marchi et al., 2019). We construct spray chart distributions for batter-pitcher matchups where separate batter spray chart distributions are constructed for each of the pitches that the pitcher throws. Rich pitch characteristic information is used to supplement labelled pitch type data since the velocity, trajectory, movement characteristics, and release points of a pitch exhibit large variation across pitchers. The reported spray chart distribu-

tion for the batter pitcher matchup is the aggregation of the spray chart distributions for each pitch that the pitcher throws. The aggregation is with respect to the percentage that the pitcher throws each pitch. The density functions corresponding to these spray chart distributions are estimated nonparametrically using the `kde2d` function in the `MASS` R package (Ripley et al., 2019).

One concern with the use of spray chart distributions is the potential sparsity of batter-pitcher matchup data. We alleviate this concern through the development and aggregation of synthetic batters and pitchers with similar characteristics as the batter and pitcher under study. Our synthetic player creation methodology is inspired by the notion of similarity scores (James, 1994; Silver, 2003). However, unlike the similarity scores presented in James (1994) and Silver (2003), we construct similarity scores using a nearest neighbor approach that is based on the underlying batter and pitcher characteristics of the players under study instead of observed statistics. The pitcher characteristics that we consider are averages of the velocity, trajectory, movement, and release point of pitches thrown. The batter characteristics that we consider are averages of launch angle, exit velocity, spray angle, and binned batted ball location information. These player characteristics are obtained from Statcast (Baseball, 2014) scraped using functionality in the `baseballr` R package (Petti et al., 2020), and reflect the underlying talent and tendencies of players. For each batter-pitcher matchup we estimate three spray chart densities, the first is the natural spray chart density corresponding to the players under study, the second is the spray chart of the synthetic pitcher versus the original batter, and the third is the spray chart of the original pitcher versus the synthetic batter. We report a synthetic spray chart density which is a weighted average of these spray chart densities where the weights are chosen with the aim of minimizing mean squared error.

We also provide a Shiny app which gives users the ability to display the synthetic spray chart distribution for any batter pitcher matchup that has occurred or could have occurred in the last

five years, the years that Statcast data exists. These synthetic spray charts are visualized over an image of a representative baseball field so that this spray chart distribution is displayed to give proper context. We also report performance metrics that are computed as expectations with respect to the synthetic spray chart distribution. The expected number of singles, doubles, triples, home runs are reported. The expected batting average on balls in play (xBABIP) and the expected bases on contact (xBsCON) are also reported. These matchup dependent metrics allow for any user to assess the expected performance of batters and pitchers when they face each other.

2 Spray chart distributions and densities

A spray chart distribution for a batter is a distribution F over a bounded subset $\mathscr{Y} \in \mathbb{R}^2$. The set $\mathscr{Y}$ contains plausible locations of batted balls from home plate. Let $(0,0) \in \mathscr{Y}$ denote the location of home plate. With this specification we can take $\mathscr{Y} = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{y}\| \leq 1000\}$ where values in $\mathscr{Y}$ are locations in feet and $\|\cdot\|$ is the Euclidean norm. This specification of $\mathscr{Y}$ guarantees that $F(\mathscr{Y}) = 1$ 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 $(y_{1i}, y_{2i}) \in \mathscr{Y}$, $i = 1, \ldots, n$, be the observed batted-ball locations for this matchup. We will estimate f with a multivariate kernel density estimator

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

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 $\mathbf{y} = (y_1, y_2) \in \mathscr{Y}$, 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 $\mathbf{x} = (\mathbf{x}_p', \mathbf{x}_b')' \in \mathscr{X}$, and $\mathscr{X}$ is assumed to be bounded. Denote the conditional spray chart distribution as $F(\mathbf{y}|\mathbf{x})$ for $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{Y}$. The conditional spray chart density function corresponding to $F(\mathbf{y}|\mathbf{x})$ will be denoted as $f(\mathbf{y}|\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}$ and $\mathbf{y} \in \mathscr{Y}$. 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 $(y_{1i}, y_{2i}) \in \mathscr{Y}$, $i = 1, \ldots, n$ are now conditional on $\mathbf{x} \in \mathscr{X}$.

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 $n_{\text{pitch types}}$ different types of pitches. We will let $\mathbf{x}_{p,t}$ be the pitcher covariates for pitch type $t = 1, \ldots, n_{\text{pitch types}}$. We will let $\mathbf{x}_{b,t}$ be the batter covariates when facing pitch type $t = 1, \ldots, n_{\text{pitch types}}$. 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,

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

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. The user of our Shiny app has some control of the entries of $\mathbf{V}_{p,t}$ by adjusting the pitcher slider. Similarity scores for batters $s(\mathbf{x}_{b,k_1,t}, \mathbf{x}_{b,k_2,t})$, $1 \leq k_1, k_2 \leq K$ are defined in a similar manner.

Implicit in this construction is the assumption that the underlying pitcher and batter characteris-

tics that we collect 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 $t = 1, \ldots, n_{\text{pitch types}}$, 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

$$f_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t} f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}). \tag{4}$$

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

$$f_{\text{sp}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{pitch types}}} \rho_t f_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}). \tag{5}$$

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

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 $k = 1, ..., K^t$. For pitch type t, the spray chart density for a pitcher facing the synthetic batter is

$$f_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}) = \sum_{k=1}^{K} w_{b,k,t} f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}). \tag{6}$$

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

$$f_{\text{sb}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{pitch types}}} \rho_t f_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t)}. \tag{7}$$

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

We then estimate (4) and (6) with

$$\hat{f}_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t} \hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}), \qquad \hat{f}_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}) = \sum_{k=1}^{K} w_{b,k,t} \hat{f}_{\mathbf{h}_{b,k,t}}(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}), \tag{8}$$

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 (8) with,

$$\hat{f}_{\text{sp}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{pitch types}}} \rho_t \hat{f}_{\text{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}), \qquad \hat{f}_{\text{sb}}(\mathbf{y}) = \sum_{t=1}^{n_{\text{pitch types}}} \rho_t \hat{f}_{\text{sb},t}(\mathbf{y}|\mathbf{x}_{p,t}). \tag{9}$$

Our implementation estimates the densities in (9) with the `kde2d.weighted` function. The estimators (9) are obviously biased estimators for f. However, they have the potential to reduce MSE. One obvious case is when, for all $t = 1, \ldots, n_{\text{pitch types}}$, 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 (9) are good estimators for f. Our implementation will estimate f with

$$\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}) = \lambda \hat{f}_{\mathbf{h}}(\mathbf{y}) + \lambda_p \hat{f}_{\text{sp}}(\mathbf{y}) + \lambda_b \hat{f}_{\text{sb}}(\mathbf{y}) \tag{10}$$

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 $t = 1, \ldots, n_{\text{pitch types}}$ 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 $t = 1,\ldots,n_{\text{pitch types}}$. 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 (10) having a lower MSE than the estimator (2).

2.2 Performance metrics

We develop novel performance metrics that are theoretically computed as expectations with respect to $\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x})$. We estimate the expected number of singles, doubles, triples, and home runs that the batter hits versus the pitcher in a particular matchup. We also estimate the xBABIP and xBsCON as additional summary measures. The metric xBsCON is best interpreted as slugging percentage conditional on balls put into play. Our implementation will not estimate these expectations exactly as there is not enough historical batted ball data.

To theoretically estimate these quantities we first obtain five years of batted ball data. We then estimate the proportion of batted balls that were either an out (O), single (1B), double (2B), triple (3B), or home run (HR) at locations $\mathbf{y}$ on the baseball field. Denote this vector of estimated propor-

tions at $\mathbf{y}$ as $\widehat{\mathbf{P}}(\mathbf{y}) = (\hat{p}_{\mathrm{O}}(\mathbf{y}), \hat{p}_{1\mathrm{B}}(\mathbf{y}), \hat{p}_{2\mathrm{B}}(\mathbf{y}), \hat{p}_{3\mathrm{B}}(\mathbf{y}), \hat{p}_{\mathrm{HR}}(\mathbf{y}))'$. Next, we obtain $\widehat{E}(\mathbf{x}) = \int \mathbf{P}(\mathbf{y}) \hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$, where $\widehat{E}(\mathbf{x}) = (\hat{e}_{\mathrm{O}}(\mathbf{x}), \hat{e}_{1\mathrm{B}}(\mathbf{x}), \hat{e}_{2\mathrm{B}}(\mathbf{x}), \hat{e}_{3\mathrm{B}}(\mathbf{x}), \hat{e}_{\mathrm{HR}}(\mathbf{x}))'$. Thus $\widehat{E}(\mathbf{x})$ is the estimated expected vector of outcomes where the expectation is taken with respect to the estimated spray chart distribution. Expected BABIP is then calculated as $\text{xBABIP} = \hat{e}_{1\mathrm{B}}(\mathbf{x}) + \hat{e}_{2\mathrm{B}}(\mathbf{x}) + \hat{e}_{3\mathrm{B}}(\mathbf{x})$ and xBsCON is calculated as $\text{xBsCON} = \hat{e}_{1\mathrm{B}}(\mathbf{x}) + 2\hat{e}_{2\mathrm{B}}(\mathbf{x}) + 3\hat{e}_{3\mathrm{B}}(\mathbf{x}) + 4\hat{e}_{\mathrm{HR}}(\mathbf{x})$. Our Shiny app also displays the floor of $100\hat{e}_{1\mathrm{B}}(\mathbf{x})$, $100\hat{e}_{2\mathrm{B}}(\mathbf{x})$, $100\hat{e}_{3\mathrm{B}}(\mathbf{x})$, and $100\hat{e}_{\mathrm{HR}}(\mathbf{x})$. The 100 multiplier is a normalization that is intended for ease of interpretation. Note that there is not enough data available to estimate $\widehat{\mathbf{P}}(\mathbf{y})$ for every $\mathbf{y} \in \mathscr{Y}$. Therefore, we calculate discretized versions of these performance metrics over 10 feet by 10 feet bins.

3 Data considerations

Our methodology will consider the following variables comprising $\mathbf{x}_p$ and $\mathbf{x}_{p,t}$: 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 $\mathbf{x}_b$ and $\mathbf{x}_{b,t}$: 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.
- Player characteristic data are standardized to have mean zero and standard deviation 1. The player characteristic data are physically dimensionless after standardization.

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 $\lceil \frac{n_{\text{pitch types}}}{2} \rceil$ 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.shinyapps.io/seam/`. The default

[Figure 1 about here]

Figure 1: The layout of the application upon submission.

[Figure 2 about here]

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

matchup in the application pairs the reigning American League Cy Young winner Justin Verlander against the reigning American League MVP Mike Trout. The layout includes a sidebar with four filters: two dropdowns for batter/pitcher selection and two sliders for metric adjustment. 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

[Figure 3 about here]

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

chart away from that of the batter under consideration. That being said, the batted ball tendencies are recorded as percentages of balls hit to six large grids on the baseball field, ignoring the quality, trajectory, and exact location of the batted ball. Thus, the quality of contact forms a more complete representation of a batter's skill than tendency.

As previously mentioned, these visualizations can help coaches position their fielders effectively. While a traditional spray chart may be useful in aggregate, building a custom spray chart to reflect a specific batter-pitcher matchup will yield more accurate results on a plate appearance by plate appearance level. This synthetically created spray chart will give the user an expected distribution of batted balls for the batter-pitcher matchup based on a combination 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. The app also displays two additional synthetic charts and a leaderboard displaying the most similar pitchers/batters. These include their overall similarity score and performance metrics. The overall similarity score for batter k and pitcher j are given, respectively, by $\sum_t \rho_t s(\mathbf{x}_{b,t}, \mathbf{x}_{b,k,t})$ and $\sum_t \rho_t s(\mathbf{x}_{p,t}, \mathbf{x}_{p,j,t})$.

See Figure 3 for an example of the top 10 most similar pitchers to Justin Verlander.

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 $f_{\mathbf{h}}(\mathbf{y}|\mathbf{x})$, $\hat{f}_{\text{sp}}(\mathbf{y})$, and $\hat{f}_{\text{sb}}(\mathbf{y})$. The 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 and numeric summary measures of batter-pitcher matchups and it will be of interest to baseball fans, analysts, players, and team executives alike. Our application shows users batter tendencies versus pitchers while providing summaries of their overall success or lack thereof. Our application 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 $\mathbf{y}, \mathbf{y}' \in \mathscr{Y}$, all $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$, 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 $\mathbf{x} \in \mathscr{X}$ and $L_{\mathbf{y}} \leq L$ for all $\mathbf{y} \in \mathscr{Y}$. 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 $\mathscr{H}_f(\mathbf{y}|\mathbf{x})$ 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 $j = 1,\ldots,J$, $k = 1,\ldots,K$, $t = 1,\ldots,n_{\text{pitch types}}$. Standard results from nonparametric estimation theory give

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

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 (10). Our multivariate Hölder class specifications yield,

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

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

and

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

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 $\sum_{j=1}^{J} w_{p,j,t} \left(\mathbf{h}'_{p,j,t}\mathrm{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}))\mathbf{h}_{p,j,t} - \mathbf{h}'\mathrm{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}\right)$ and $\sum_{k=1}^{K} w_{b,k,t} \left(\mathbf{h}'_{b,k,t}\mathrm{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t}))\mathbf{h}_{b,k,t} - \mathbf{h}'\mathrm{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}\right)$ 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 $\|\text{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_{p,t}, \mathbf{x}_{b,k,t})) - \text{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_t))\| \leq \sqrt{d_p}L$ and $\|\text{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t})) - \text{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_t))\| \leq \sqrt{d_b}L$ 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

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

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

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

and regularity approximations A4 and A5 yield

$$
\text{Bias}(\hat{g}_{\boldsymbol{\lambda}}(\mathbf{y}), f(\mathbf{y}))^2 = \left((\lambda - 1)\sum_t \rho_t f(\mathbf{y}|\mathbf{x}_t) + \lambda \sum_t \rho_t \frac{\mu_2(K)\mathbf{h}'\text{diag}(\mathscr{H}_f(\mathbf{y}|\mathbf{x}_t))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) \right.
$$

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

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

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

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

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

where $n_p=\sum_{j=1}^{J}s_{p,j}^2n_{p,j}$ and $n_b=\sum_{k=1}^{K}s_{b,k}^2n_{b,k}$. We will now develop intuition for these choices.

First, notice that $\lambda_p, \lambda_b \to 0$ as $\min_j(\|\mathbf{x}-\tilde{\mathbf{x}}_{p,j}\|), \min_k(\|\mathbf{x}-\tilde{\mathbf{x}}_{b,k}\|) \to \infty$. These cases correspond, to there being no similar pitchers or batters to the players under study. We turn attention to the bias terms, notice that

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

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

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

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

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

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

References

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

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

Major League Baseball. Statcast. `https://baseballsavant.mlb.com/`, 2014. Accessed: 2020-04-29.

Baseball-Reference. `https://www.baseball-reference.com`, 2020. Accessed: 2020-04-29.

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

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

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

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

Fangraphs. `https://www.fangraphs.com`, 2020. Accessed: 2020-04-29.

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

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

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

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

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

B. Petti. 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.

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

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

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

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

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

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

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