Spray chart distributions: a context rich approach to player evaluation

Charlie Young, David Dalpiaz, Daniel J. Eck

April 5, 2020

Abstract

1 Introduction

Baseball has had a rich statistical history dating back to the first box score created by Henry Chadwick in 1859. Fans, journalists, and baseball teams have been enamored and obsessed with statistics in baseball ever since. This obsession about baseball statistics is best summarized by the existence of Schwarz [2004], a best selling book devoted 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, Jensen et al., 2009a], pitching (need articles), fielding [Jensen et al., 2009b, Piette and Jensen, 2012], and total value [Baumer et al., 2015].

Most baseball statistics used for player evaluations are obtained from raw box score totals. While box score totals are a enjoyable statistical summary for baseball fans, the information contained in them is not very substantive, they ignore rich contextual information. Most commonly used player evaluation metrics are functions of context-free box score totals. These include, and are far from limited to, adjusted earned run average (ERA+), adjusted on base plus slugging percentage (OPS+), and wins above replacement (WAR). The more sophisticated techniques in Berry et al. [1999], Jensen et al. [2009a], and Baumer et al. [2015] are also constructed from raw box score totals. These metrics all ignore which pitchers a batter faced and the game situations which complement the outcomes that are recorded. Eck [2020] showed that these 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 any of the particular batter pitcher matchups that occur throughout the season outside of park effects.

Discuss how PECOTA, Jensen et al. [2009a], [Jensen et al., 2009b, Section 2.5] pool players

In this article we develop spray chart distributions as a methodology for understanding batter pitcher matchups visually and numerically. Informally, spray chart distributions are 2-dimensional contours that overlay spray charts [Marchi et al., 2019, Section 12]. We construct spray chart distributions for batter pitcher matchups where separate batter spray chart distribution are constructed for each of the pitches that the pitcher throws. Rich pitch characteristic information is used to

supplement labelled pitch type data since the velocity, trajectory, and other characteristics of a pitch exhibit large variation across pitcher. The reported spray chart distribution for the batter pitcher matchup is the aggregation of the spray chart distributions for each pitch that the pitcher throws, the aggregation is with respect to the percentage that the pitcher throws each pitch. **Need to describe what our developments bring to the table, both visually and numerically.**

One concern with this approach is that batter pitcher matchup data can be sparse. We alleviate this concern with the development of a synthetic pitcher with similar characteristics as the pitcher under study. The synthetic pitcher is constructed in a similar fashion as how synthetic controls are created via synthetic control methodology (SCM) [Abadie et al., 2010] in the policy evaluation with observational data literature (need citation). The final spray chart distribution reported is a convex combination of that for the pitcher under study and the synthetic pitcher. The parameter which controls the amount of influence that the synthetic pitcher's spray chart distribution has on the final spray chart distribution is chosen with the goal of minimizing mean squared error (MSE).

2 Motivating Example

In this section we present a snapshot of what our proposed visualization and methodology can provide users. Perhaps the spray chart distribution of Mike Trout vs. Justin Verlander should go here.

3 Pitcher and batter characteristics

We need to explain the pitch data and all of the preprocessing employed.

4 Spray chart distributions

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

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

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

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

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

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

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

5 Synthetic player construction

We develop a method for synthetically recreating baseball players in order to alleviate the small sample size concerns of individual batter pitcher matchups. Matchup data involving these synthetic players will then be included in our analysis to estimate the spray chart density function for individual batter pitcher mathcups. Our synthetic player creation method is inspired by the notion of similarity scores [James, 1994, Silver, 2003]. Unlike these notions of similarity scores we base similarity on the underlying pitch characteristics and not observed statistics. Our similarity score of pitcher j to pitcher k is $s(\mathbf{x}_{p,j}, \mathbf{x}_{p,k}) = \exp(-\|\mathbf{x}_{p,j} - \mathbf{x}_{p,k}\|_{\mathbf{V}_p})$ where $\mathbf{x}_{p,j}$ and $\mathbf{x}_{p,k}$ are, respectively, the underlying pitch characteristics for pitcher j and k, $\|\mathbf{x}_{p,j} - \mathbf{x}_{p,k}\|_{\mathbf{V}_p} = \sqrt{(\mathbf{x}_{p,j} - \mathbf{x}_{p,k})'\mathbf{V}_p(\mathbf{x}_{p,j} - \mathbf{x}_{p,k})}$, and $\mathbf{V}_p$ is a weight matrix that is chosen to scale the pitch characteristics and give preference to pitch characteristics that have higher influence on the spray chart distribution under study. Similarity scores for batters are defined in the same way. Implicit in this construction is the assumption that the underlying pitcher and batter characteristics that we collect represents the true talent of the players under study.

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

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

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

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

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

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

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

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

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

where $\lambda, \lambda_p, \lambda_b$ form a convex combination. The choices of the elements of $\boldsymbol{\lambda}$ will be discussed in the next Section where we calculate and compare the MSE of (2) and (6). Note that these calculations are conditional on the pitch characteristics which implies that they are also conditional on $w_{p,j}$ and $w_{b,k}$ since the weights are a deterministic function of the pitch characteristics.

Our implementation will estimate the elements of $\boldsymbol{\lambda}$ as

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

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

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

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

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

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 $\mathbf{H} = \text{diag}(\mathbf{h})$. We will use the following notation: $R_{\mathbf{x}}(f) = \int f(\mathbf{y}|\mathbf{x})^2 d\mathbf{y}$, $\mu_2(K) = \int u^2 K(u) du$, and $\mathcal{H}_f(\mathbf{y}|\mathbf{x})$ is the Hessian matrix respect to $f(\mathbf{y}|\mathbf{x})$ where derivatives are taken with respect to $\mathbf{y}$. Assume that pitch outcomes are independent across at bats and that $n_{p,j} = O(n)$, $n_{b,k} = O(n)$ and $\mathbf{h}_{p,j} = O(\mathbf{h})$, $\mathbf{h}_{b,k} = O(\mathbf{h})$ for all $j = 1, \ldots, J$, $k = 1, \ldots, K$.

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

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

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

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

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

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

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

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

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

and

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

Our multivariate Hölder class specifications yield,

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

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

and

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

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

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

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

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

We also have

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

and regularity approximations A4 and A5 yield

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

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

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

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

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

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

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

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

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

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

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

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

References

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

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

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

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

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

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

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

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

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

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

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.

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