Spray chart distributions: a context rich approach to player evaluation

Charlie Young, David Dalpiaz, Daniel J. Eck

March 26, 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 pitcher, batter, defense, ballpark, and other characteristics belonging to a space $\mathcal{X}$. The conditional spray 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

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

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

Our implementation will estimate $f(\cdot|\mathbf{x})$ using the `kde2d` 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

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

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}), $i = 1, \ldots, n_x$ 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. 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 method for building spray chart distributions for batter pitcher matchups with an additional synthetic pitcher is as follows: we isolate the pitch repertoire of the pitcher of interest. We obtain the types of pitches that the pitcher throws and then compute the mean of all pitch characteristics for each of these pitches. This will form the basis of our comparison. **Finish setup. Get similarity score construction**.

We then construct the spray chart distribution against the synthetic pitcher. For each pitch thrown by the pitcher under study, we take a weighted average of the hitter's spray chart. The estimate of the synthetic pitcher $\tilde{f}(\mathbf{y}|\mathbf{x})$ is

$$\tilde{f}(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{J} w_j^* \hat{f}_{\mathbf{h}_j}(\mathbf{y}|\mathbf{x}_j) \tag{3}$$

where $\hat{f}_{\mathbf{h}_j}(\mathbf{y}|\mathbf{x}_j)$ are estimated spray chart distributions from the batter's matchups against the jth pitcher with pitch characteristics $\mathbf{x}_j$. Our implementation estimates (3) with the `kde2d.weighted` function in the `ggtern` R package [Hamilton, 2018]. The estimator (3) is obviously a biased estimator for $f(\mathbf{y}|\mathbf{x})$. However, it can lead to lower MSE in certain scenarios. One obvious case is when $w_j^* = 1 - \varepsilon$ where $\varepsilon > 0$ is small and $n_{\mathbf{x}_j} > n_{\mathbf{x}}$. In this setting, pitcher j possess very similar pitch characteristics as the pitcher that the batter is facing, and the batter has had more encounters with pitcher j than the current pitcher. 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 $\tilde{f}(\mathbf{y}|\mathbf{x})$ is a good estimator for $f(\mathbf{y}|\mathbf{x})$. Our implementation will estimate $f(\mathbf{y}|\mathbf{x})$ with

$$\hat{g}_\lambda(\mathbf{y}|\mathbf{x}) = \lambda \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + (1 - \lambda)\tilde{f}(\mathbf{y}|\mathbf{x}), \tag{4}$$

where the choice of $0 \leq \lambda \leq 1$ will be discussed in the next Section where we calculate and compare the MSE of (2) and (4). Note that these calculations are conditional on the pitch characteristics which implies that they are also conditional on $\mathbf{W}^*$ since $\mathbf{W}^*$ is a deterministic function of the pitch characteristics.

5.1 Choice of λ

We motivate a choice of λ in (4) by comparing the MSE of (2) and (4). Let $n_{\mathbf{W}^*} = \sum_{j=1}^J w_j^* n_{\mathbf{x}_j}$ and define $\text{logit}(x) = 1/(1+\exp(-x))$. Our implementation will use

$$\lambda^* = 1 - \text{logit}\left(-\frac{n_{\mathbf{W}^*} - n_{\mathbf{x}}}{n_{\mathbf{x}}}\right)\left(\max_j(w_j^*) - \frac{1}{J}\right)\left(\frac{1}{n_{\mathbf{x}}}1(n_{\mathbf{x}} \geq M) + 1(n_{\mathbf{x}} < M)\right)$$

as our choice for λ where $\text{logit}(x) = 1/(1+\exp(-x))$ and M is a user specified input that is meant to void out the influence of the synthetic pitcher if $n_{\mathbf{x}}$ is thought to be large enough. Intuitively, λ^* favors the synthetic pitcher when there exists a j such that $n_{\mathbf{x}_j} > n$ and the weight w_j^* is large, and it protects against scenarios where no w_j^* is large but $n_{\mathbf{x}_j} > n_{\mathbf{x}}$ for several $1 \leq j \leq J$. Technical justification for selecting λ^* is given in the Appendix.

Appendix: Justification for our choice of λ

We now motivate λ^* 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

$$|D_{\mathbf{y}}^{\mathbf{s}} f(\mathbf{y}_1|\mathbf{x}) - D_{\mathbf{y}}^{\mathbf{s}} f(\mathbf{y}_2|\mathbf{x})| \leq L_{\mathbf{x}} \|\mathbf{y}_1 - \mathbf{y}_2\|^{\beta - |\mathbf{s}|},$$
$$|D_{\mathbf{x}}^{\mathbf{t}} f(\mathbf{y}|\mathbf{x}_1) - D_{\mathbf{x}}^{\mathbf{t}} f(\mathbf{y}|\mathbf{x}_2)| \leq L_{\mathbf{y}} \|\mathbf{x}_1 - \mathbf{x}_2\|^{\beta - |\mathbf{t}|},$$

for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, 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.

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, $n_x \det(\mathbf{H}) \to \infty$ when $n_{\mathbf{x}} \to \infty$ 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 $n_{\mathbf{x}_j} = O(n_{\mathbf{x}})$ and $\mathbf{h}_j = O(\mathbf{h})$ for all $j = 1, \ldots, J$.

From Hansen's notes:

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

and

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

With these properties we have

$$\begin{aligned}
\operatorname{MSE}(\hat{g}_\lambda(\mathbf{y}|\mathbf{x})) &= \operatorname{Var}\left(\hat{g}_\lambda(\mathbf{y}|\mathbf{x})\right) + \left(\operatorname{E}\hat{g}_\lambda(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x})\right)^2 \\
&= \operatorname{Var}\left(\lambda \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + (1-\lambda)\tilde{f}(\mathbf{y}|\mathbf{x})\right) + \left(\lambda \operatorname{E}\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + (1-\lambda)\operatorname{E}\tilde{f}(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x})\right)^2 \\
&= \lambda^2 \operatorname{Var}\left(\hat{f}_h(\mathbf{y}|\mathbf{x})\right) + (1-\lambda)^2 \operatorname{Var}\left(\tilde{f}(\mathbf{y}|\mathbf{x})\right) \\
&\quad + \left(\lambda \operatorname{E}\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + (1-\lambda)\operatorname{E}\tilde{f}(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x})\right)^2 \\
&= \lambda^2 \operatorname{Var}\left(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})\right) + (1-\lambda)^2 \operatorname{Var}\left(\sum_{j=1}^{J} w_j^* \hat{f}(\mathbf{y}|\mathbf{x}_j)\right) \\
&\quad + \left(\lambda \operatorname{E}\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + (1-\lambda)\operatorname{E}\sum_{j=1}^{J} w_j^* \hat{f}(\mathbf{y}|\mathbf{x}_j) - f(\mathbf{y}|\mathbf{x})\right)^2 \\
&= \lambda^2 \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + O\left(\frac{1}{n_{\mathbf{x}}}\right) + (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} \\
&\quad + \left(\lambda \left[f(\mathbf{y}|\mathbf{x}) + \frac{\mu_2(K)\mathbf{h}'\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2)\right] - f(\mathbf{y}|\mathbf{x})\right. \\
&\quad \left. + (1-\lambda)\sum_{j=1}^{J} w_j^* \left[f(\mathbf{y}|\mathbf{x}_j) + \frac{\mu_2(K)\mathbf{h}_j'\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_j))\mathbf{h}_j}{2} + o(\|\mathbf{h}\|^2)\right]\right)^2 \\
&= \lambda^2 \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + O\left(\frac{1}{n_{\mathbf{x}}}\right) + (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} \\
&\quad + \left(\lambda \frac{\mu_2(K)\mathbf{h}'\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h}}{2} + o(\|\mathbf{h}\|^2) - (1-\lambda) f(\mathbf{y}|\mathbf{x})\right. \\
&\quad \left. + (1-\lambda)\sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) + (1-\lambda)\sum_{j=1}^{J} \frac{\mu_2(K)\mathbf{h}_j'\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}_j))\mathbf{h}_j}{2}\right)^2 .
\end{aligned}$$

In our application $\max_{\mathbf{y}\in\mathcal{Y},\mathbf{x}\in\mathcal{X}} \|\mathcal{H}_f(\mathbf{y}|\mathbf{x})\|$ is thought be small, so that $\mathbf{h}'\operatorname{diag}(\mathcal{H}_f(\mathbf{y}|\mathbf{x}))\mathbf{h} = O(\|\mathbf{h}\|^2)$. With this, we have

$$\begin{aligned}
\operatorname{MSE}(\hat{g}_\lambda(\mathbf{y}|\mathbf{x})) &= \lambda^2 \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + O\left(\frac{1}{n_{\mathbf{x}}}\right) \\
&\quad + (1-\lambda)^2 \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) + O(\|\mathbf{h}\|^2)\right)^2 \\
&= \lambda^2 \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + O\left(\frac{1}{n_{\mathbf{x}}}\right) + O(\|\mathbf{h}\|^2)
\end{aligned}$$

$$+ (1-\lambda)^2 \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 .$$

Similar reasoning yields

$$\begin{aligned}\mathrm{MSE}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})) &= \mathrm{Var}\left(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})\right) + \left(\mathrm{E}\,\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) - f(\mathbf{y}|\mathbf{x})\right)^2 \\ &= \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + O\left(\frac{1}{n_{\mathbf{x}}}\right) + O(\|\mathbf{h}\|^2).\end{aligned}$$

Ignoring the lower order terms, we want to pick λ so that $\mathrm{MSE}(\hat{g}_\lambda(\mathbf{y}|\mathbf{x})) \leq \mathrm{MSE}(\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}))$. Therefore, we want to pick λ so that

$$\begin{aligned}
&\lambda^2 \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + (1-\lambda)^2 \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 \\
&\qquad \leq \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} \\
\implies\ & (1-\lambda)^2 \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + (1-\lambda)^2 \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 \\
&\qquad \leq \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} (1-\lambda^2) \\
\implies\ & (1-\lambda) \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + (1-\lambda) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 \\
&\qquad \leq \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} (1+\lambda) \\
\implies\ & \sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} - \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 \\
&\qquad \leq \lambda \left(\sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2 \right) \\
\implies\ & \lambda \geq \frac{\sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} - \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2}{\sum_{j=1}^{J} {w_j^*}^2 \frac{R_{\mathbf{x}_j}(f) f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j} \det(\mathbf{H}_j)} + \frac{R_{\mathbf{x}}(f) f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}} \det(\mathbf{H})} + \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j) \right)^2}.
\end{aligned}$$

We will now choose the bandwidth parameters such that $\det(\mathbf{H}_j) = \det(\mathbf{H}) n_{\mathbf{x}} / n_{\mathbf{x}_j}$ in order to simplify our derivation for λ. We will also assume that we can specify $\beta = 2$ for the Hölder class $H(\beta, L)$ that governs the smoothness of $f(\cdot|\cdot)$. The specification of $\beta = 2$ implies that

$$\begin{aligned} f(\mathbf{y}|\mathbf{x}) - L\|\mathbf{x} - \mathbf{x}_j\|^2 \leq & f(\mathbf{y}|\mathbf{x}_j) \leq f(\mathbf{y}|\mathbf{x}) + L\|\mathbf{x} - \mathbf{x}_j\|^2, \\ R_{\mathbf{x}}(f) - L\|\mathbf{x} - \mathbf{x}_j\|^2 \leq & R_{\mathbf{x}_j}(f) \leq R_{\mathbf{x}}(f) + L\|\mathbf{x} - \mathbf{x}_j\|^2. \end{aligned}$$

We will suppose that ${w_j^*}^2\|\mathbf{x}-\mathbf{x}_j\|^2$ and ${w_j^*}^2\|\mathbf{x}-\mathbf{x}_j\|^4$ are negligible. The intuition for this approximation is that an appreciable value ${w_j^*}^2$ of implies that $\|\mathbf{x}-\mathbf{x}_j\|^k$, for $k=2,4$ is negligible and vice-versa.

The specifications of $\det(\mathbf{H}_j)=\det(\mathbf{H})n_{\mathbf{x}}/n_{\mathbf{x}_j}$ and $\beta=2$ in the Hölder class $H(\beta,L)$, and the negligibility of ${w_j^*}^2\|\mathbf{x}-\mathbf{x}_j\|^2$ and ${w_j^*}^2\|\mathbf{x}-\mathbf{x}_j\|^4$ combine to yield

$$\begin{aligned}
\sum_{j=1}^J {w_j^*}^2\frac{R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j}\det(\mathbf{H}_j)}-\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}}\det(\mathbf{H})}&=\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)-R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})\right)\\
&\le\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2(R_{\mathbf{x}}(f)+L\|\mathbf{x}-\mathbf{x}_j\|^2)(f(\mathbf{y}|\mathbf{x})+L\|\mathbf{x}-\mathbf{x}_j\|^2)-R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})\right)\\
&=\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2\left(R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})+R_{\mathbf{x}}(f)L\|\mathbf{x}-\mathbf{x}_j\|^2+L\|\mathbf{x}-\mathbf{x}_j\|^2f(\mathbf{y}|\mathbf{x})+L^2\|\mathbf{x}-\mathbf{x}_j\|^4\right)\right.\\
&\quad\left.-\,R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})\right)\\
&\approx\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})-R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})\right)\\
&=\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2-1\right)f(\mathbf{y}|\mathbf{x})R_{\mathbf{x}}(f).
\end{aligned}$$

A similar argument gives

$$\begin{aligned}
&\sum_{j=1}^J {w_j^*}^2\frac{R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j}\det(\mathbf{H}_j)}-\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}}\det(\mathbf{H})}\\
&\qquad\ge\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2(R_{\mathbf{x}}(f)-L\|\mathbf{x}-\mathbf{x}_j\|^2)(f(\mathbf{y}|\mathbf{x})-L\|\mathbf{x}-\mathbf{x}_j\|^2)-R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})\right)\\
&\qquad\approx\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2-1\right)f(\mathbf{y}|\mathbf{x})R_{\mathbf{x}}(f),
\end{aligned}$$

and

$$\sum_{j=1}^J {w_j^*}^2\frac{R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j}\det(\mathbf{H}_j)}+\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}}\det(\mathbf{H})}\approx\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2+1\right)f(\mathbf{y}|\mathbf{x})R_{\mathbf{x}}(f).$$

Putting all of this together yields

$$\begin{aligned}
&\frac{\sum_{j=1}^J {w_j^*}^2\frac{R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j}\det(\mathbf{H}_j)}-\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}}\det(\mathbf{H})}+\left(f(\mathbf{y}|\mathbf{x})-\sum_{j=1}^J w_j^*f(\mathbf{y}|\mathbf{x}_j)\right)^2}{\sum_{j=1}^J {w_j^*}^2\frac{R_{\mathbf{x}_j}(f)f(\mathbf{y}|\mathbf{x}_j)}{n_{\mathbf{x}_j}\det(\mathbf{H}_j)}+\frac{R_{\mathbf{x}}(f)f(\mathbf{y}|\mathbf{x})}{n_{\mathbf{x}}\det(\mathbf{H})}+\left(f(\mathbf{y}|\mathbf{x})-\sum_{j=1}^J w_j^*f(\mathbf{y}|\mathbf{x}_j)\right)^2}\\
&\qquad\approx\frac{\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2-1\right)f(\mathbf{y}|\mathbf{x})R_{\mathbf{x}}(f)+\left(f(\mathbf{y}|\mathbf{x})-\sum_{j=1}^J w_j^*f(\mathbf{y}|\mathbf{x}_j)\right)^2}{\frac{1}{n_{\mathbf{x}}\det(\mathbf{H})}\left(\sum_{j=1}^J {w_j^*}^2+1\right)f(\mathbf{y}|\mathbf{x})R_{\mathbf{x}}(f)+\left(f(\mathbf{y}|\mathbf{x})-\sum_{j=1}^J w_j^*f(\mathbf{y}|\mathbf{x}_j)\right)^2}
\end{aligned}$$

$$= \frac{\left(\sum_{j=1}^{J} {w_j^*}^2 - 1\right) f(\mathbf{y}|\mathbf{x}) R_{\mathbf{x}}(f) + n_{\mathbf{x}} \det(\mathbf{H}) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j)\right)^2}{\left(\sum_{j=1}^{J} {w_j^*}^2 + 1\right) f(\mathbf{y}|\mathbf{x}) R_{\mathbf{x}}(f) + n_{\mathbf{x}} \det(\mathbf{H}) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j)\right)^2}.$$

We want to pick λ so that

$$\lambda \geq \frac{\left(\sum_{j=1}^{J} {w_j^*}^2 - 1\right) f(\mathbf{y}|\mathbf{x}) R_{\mathbf{x}}(f) + n_{\mathbf{x}} \det(\mathbf{H}) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j)\right)^2}{\left(\sum_{j=1}^{J} {w_j^*}^2 + 1\right) f(\mathbf{y}|\mathbf{x}) R_{\mathbf{x}}(f) + n_{\mathbf{x}} \det(\mathbf{H}) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j)\right)^2}. \tag{5}$$

We will now motivate choices for λ based on scenarios of $\mathbf{W}^*$ and $n_{\mathbf{x}}$ and $n_{\mathbf{x}_j}$. Note that our approximations which led to (5) are based upon specifying the bandwidths as $\det(\mathbf{H}_j) = \det(\mathbf{H}) n_{\mathbf{x}} / n_{\mathbf{x}_j}$ for $j = 1, \ldots, J$.

Case 1: $w_j^* \approx 1$. It is easy to see that the existence of a j such that $w_j^* \approx 1$ implies that λ can be essentially any value between 0 and 1. When this is the case we will choose $\lambda \leq 1/2$ when $n_{\mathbf{x}_j} \geq n_{\mathbf{x}}$ and $\lambda > 1/2$ when $n_{\mathbf{x}_j} < n_{\mathbf{x}}$. In this setting we have $\lambda \to 0$ and $n_{\mathbf{x}} \to 0$ or $n_{\mathbf{x}_j} \to \infty$.

Case 2: w_j^* is not large or small, and $n_{\mathbf{x}}$ is large. If there exists no such j for case 1 to hold and $n_{\mathbf{x}}$ is large then the right hand side (RHS) of (5) is close to 1. Thus we specify that $\lambda \to 1$ as $n_{\mathbf{x}} \to \infty$.

Case 3: $\max_j(w_j^*) \approx 1/J$. In this setting, the synthetic pitcher poorly approximates the pitcher under study. The smoothness of the space $H(L, \beta = 2)$ and lack of pronounced modes implies that the $n_{\mathbf{x}} \det(\mathbf{H}) \left(f(\mathbf{y}|\mathbf{x}) - \sum_{j=1}^{J} w_j^* f(\mathbf{y}|\mathbf{x}_j)\right)^2$ terms dominate the other terms provided that $n_{\mathbf{x}}$ is large enough. Thus we specify that $\lambda \to 1$ as $\max_j(w_j^*) \to 1/J$. When $n_{\mathbf{x}}$ is prohibitively small then we recommend not making inferences using spray chart distributions in this setting.

Case 4: w_j^* is not large or small, and $n_{\mathbf{x}}$ is not large or small. If there exists no such j for case 1 to hold and $n_{\mathbf{x}}$ is not large enough for case 2, then the RHS of (5) depends on $\mathbf{W}^*$, $n_{\mathbf{x}}$, and $n_{\mathbf{x}_j}$ for all $j = 1, \ldots, J$. Also note that the supposition that ${w_j^*}^2 \|\mathbf{x} - \mathbf{x}_j\|^2$ and ${w_j^*}^2 \|\mathbf{x} - \mathbf{x}_j\|^4$ are negligible may be questionable in this setting.

These four cases lead to the following choice for λ:

$$\lambda^* = 1 - \text{logit}\left(-\frac{n_{\mathbf{W}^*} - n_{\mathbf{x}}}{n_{\mathbf{x}}}\right) \left(\max_j(w_j^*) - \frac{1}{J}\right) \left(\frac{1}{n_{\mathbf{x}}} 1(n_{\mathbf{x}} \geq M) + 1(n_{\mathbf{x}} < M)\right)$$

where $n_{\mathbf{W}^*} = \sum_{j=1}^{J} w_j^* n_{\mathbf{x}_j}$, $\text{logit}(x) = 1/(1 + \exp(-x))$, and M is a user specified input that is meant to void out the influence of the synthetic pitcher if $n_{\mathbf{x}}$ is thought to be large enough. The choice λ^* satisfies Cases 1-3 and we hope that it is a sufficient balance of all the design parameters to be reliable when in Case 4.

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`.