Comparing baseball players across eras via the novel Full House Model

Shen Yan, Daniel J. Eck

Jan 2021

Abstract

We motivate a new framework that is appropriate for making cross-contextual comparisons of different components in an evolving system at different times. The systems that we consider involve rich data collection on both the quality of components in the system as well as understanding how components enter the system. Our methodology is a crystallization of the conceptual ideas put forward by Stephen Jay Gould. We name this methodology the Full House Model in his honor. The Full House Model works by balancing the performance of components within the system at any given set time and the number of components that are eligible to be included within the system at that time. We assume that each component has a latent ability which can be computed from this balancing act provided the system inclusion mechanism for components of the system is known. The distribution of components can be estimated from nonparametric probability distribution without any assumptions on the distribution of the system components. We demonstrate the utility of the Full House Model in an application of comparing baseball players' statistics across eras. We show that our approach yields defensible era-adjusted baseball statistics which properly balance how players performed against their peers and how many people were eligible to play in Major League Baseball (MLB). We compare our approach with existing approaches which we argue are not properly calibrated for the task of cross-era comparisons of baseball players. Our results reveal a radical reranking of baseball's greatest players that is consistent with what one would expect under a common-sense uniform talent generation assumption. Most importantly, we found that the greatest African American and Latino players now sit atop the greatest all-time lists of historical baseball players.

1 Introduction

This paper deals with the following problem: Suppose we have a known number of components which comprise a system where these components are chosen from a population of eligible components also of known size. In this setting we allow the components and the number of eligible components of the system to change over time. The aim is then to make cross-contextual comparisons of components of different systems. We achieve this aim through balancing the performance of specific components as judged against components within the same system and the number of eligible components available to be included in the system. Our methodology is a crystallization of the conceptual ideas put forward by Stephen Jay Gould. We name this methodology the Full

House Model, this name is a reference to Gould's book book Full House: The Spread of Excellence from Plato to Darwin [Gould, 2011].

Gould [2011] reached a paradoxical conclusion about baseball's greatest players from an evolutionary biological perspective. At time, it was widely thought that the players of past were the game's greatest players because these players' statistical dominance had not been equaled by modern players. Gould [2011] showed that the statistical nature of baseball is a balance between pitching and batting, and this balance between pitchers and batters has been preserved. He also showed that the statistical distribution of batting average, a measure of success for batters, has become more concentrated as time has gone on. Gould [2011] argued that these changes reflected the evolution of baseball; rules changed, bad strategies and training methods got weeded out, and the eligible population has expanded as time continues. The Full House Model which we develop in this paper can be explicitly tailored to accommodate this evolving nature, and, more importantly, disentangle it to yield era-neutral metrics for baseball players which allow for proper cross-contextual comparisons to be made.

We are not the first authors to use statistical thinking in attempt to compare players' statistics across eras [Berry et al., 1999, Petersen et al., 2011, Schell, 2013, 2016, Petersen and Penner, 2020]. There also exists a wide class of publicly available "vs your peers" metrics that compare players to their contemporaries after accounting for park effects, league differences, and overall run environment. Examples of "vs your peers" statistics include Wins Above Replacement (WAR) which is a holistic one-number summary of a players overall value translated to how many wins that player is expected to add if they are substituted for a hypothetical replacement level player [Baseball-Reference, FanGraphs]. All of these approaches do not properly compare players across eras or serve as an era-adjustment in any meaningful sense of the word because they make no attempt to account for changing dynamics of the inclusion mechanism which governs who enters the MLB [Eck, 2020]. We further show that the testing procedure developed in Eck [2020] reveals that the Full House Model is the only technique among these various approaches which yields era-neutral rankings of players which are consistent with what one would expect under the assumption that the underlying talent is evenly distributed across time.

This paper is organized in the following manner. In Section 2 we motivate the Full House Model using parametric and nonparametric probability distributions to estimate the performance of the components in the system. We also provide some theoretical properties of our model. In Section 3 we demonstrate our method on historical batting averages (BA) by using parametric probability distribution and home runs (HR), earned run allowed (ERA), Strikeout (SO) and Win Above Replacement (WAR) by using nonparametric probability distribution in baseball. In Section 4 we investigate comparisons of players across eras obtained from conventional approaches as well as other more sophisticated era-adjustment approaches. We wrap up with a discussion of results, extension of the Full House Model, philosophies of era-adjustment techniques, and the social implications of our findings.

2 Model Setting

We now motivate the structure of the Full House Model. Supposed that all individuals in the population N_i have different underlying aptitudes and denote this aptitude by $X_{i,1}, \ldots, X_{i,N_i} \overset{\text{iid}}{\sim} F_{X_i}$. $X_{i,(j)}$ is the jth ordered aptitudes in the population N_i. For each individual in the population

N_i, it can be included in the ith system and $g_i\left(X_{i,j}, \mathbf{X}_{i,-j}\right)$ is the system inclusion function:

$$g_i\left(X_{i,j}, \mathbf{X}_{i,-j}\right) = 1\left(X_{i,j} \in \{l_{i,1}(X_{i,j}, \mathbf{X}_{i,-j}), l_{i,2}(X_{i,j}, \mathbf{X}_{i,-j}), \ldots, l_{i,n_i}(X_{i,j}, \mathbf{X}_{i,-j})\}\right)$$

where $l_{i,k}(X_{i,j}, \mathbf{X}_{i,-j}) = X_{i,(r(N_i,n_i,k))}, k = 1, 2, \ldots, n_i$, n_i is the number of individuals included in the ith system.

$g_i\left(X_{i,j}, \mathbf{X}_{i,-j}\right) = 1$ indicates that components j is included in the ith system and $g_i\left(X_{i,j}, \mathbf{X}_{i,-j}\right) = 0$ indicates that component j is not included in the ith system, where $\mathbf{X}_{i,-j}$ is the vector of all individual aptitudes not including the component j.

For example, suppose $r(N_i, n_i, k) = N_i - n_i + k$, $k = 1, 2, \ldots, n_i$, then inclusion function g_i will select the top n_i components from the population N_i and the top n_i components are included in the ith system.

Then the n_i individuals arising from the population N_i show their aptitudes by competing with each other in the ith system and one or several statistics will be observed. In our setting, we assume only one statistic is observed.

Now we have n_i observable components $Y_{i,1}, \ldots, Y_{i,n_i} \sim F_{Y_i}$ representing the one statistic we observe, where F_{Y_i} could be known or unknown, and $Y_{i,(j)}$ is the jth ordered components from the ith system. Each observable component $Y_{i,j}$ in the ith system corresponds to the underlying aptitude $X_{i,(r(N_i,n_i,j))}$ in the population N_i. Therefore we can combine the components and underlying aptitudes as pairs, $(Y_{i,(j)}, X_{i,(r(N_i,n_i,j))})$, where $X_{i,(r(N_i,n_i,j))}$ is the $r(N_i, n_i, j)$th ordered underlying aptitude in the population N_i from the ith system and corresponds to the jth ordered components from the ith system, which is $Y_{i,(j)}$. For example, suppose $r(N_i, n_i, k) = N_i - n_i + k$, then $l_{i,k}(X_{i,j}, \mathbf{X}_{i,-j}) = X_{i,(N_i-n_i+k)}$, $k = 1, 2, \ldots, n_i$, and the ith system will include the top n_i underlying aptitudes from the population N_i .

The difference between the approaches that just look at observable components $Y_{i,j}$ and our methodology is population is taken into account for underlying aptitudes, which is an important part in Gould's conjecture [Gould, 2011].

2.1 Parametric distribution measuring the components

Consider the pair $(Y_{i,(j)}, X_{i,(r(N_i,n_i,j))})$ and when the distribution corresponding to $Y_{i,j}$ from the ith system is known to belong to a continuous parametric family indexed by unknown parameter θ_i, and let $F_Y(\cdot \mid \theta_i)$ be a parametric CDF with parameters $\theta_i \in \mathbb{R}^{p_Y}$. We can estimate θ_i with $\hat{\theta}_i$ and plug the estimator into the CDF $F_Y(\cdot \mid \hat{\theta}_i)$. The distribution function $F_{Y_{i,(j)}}(y \mid \hat{\theta}_i)$ is

$$F_{Y_{i,(j)}}(y \mid \hat{\theta}_i) = \sum_{k=j}^{n_i} \binom{n_i}{k} \left(F_{Y_i}(y \mid \hat{\theta}_i)\right)^k \left(1 - F_{Y_i}(y \mid \hat{\theta}_i)\right)^{n_i-k}$$

We will make use of the following classical order statistics properties,

$$F_{Y_i}\left(Y_{i,(j)} \mid \theta_i\right) \sim U_{i,(j)}, \quad F_{Y_i}\left(Y_{i,(j)} \mid \hat{\theta}_i\right) \approx U_{i,(j)}$$
$$F_{Y_{i,(j)}}\left(Y_{i,(j)} \mid \theta_i\right) \sim U_{i,j}, \quad F_{Y_{i,(j)}}\left(Y_{i,(j)} \mid \hat{\theta}_i\right) \approx U_{i,j}$$

where $U_{i,j} \sim U(0, 1)$ and $U_{i,(j)} \sim \text{Beta}(j, n_i + 1 - j)$ and the approximation in the right hand side depends upon the estimator $\hat{\theta}_i$ and the sample size.

We now connect the order statistics to the underlying aptitude distribution that comes from a population with $N_i \geq n_i$ observations when F_{X_i} is known. This connection is established with the relation

$$F^{-1}_{X_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(Y_{i,(j)} \mid \theta_i\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(r(N_i,n_i,j))}$$

We estimate the above with

$$F^{-1}_{X_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(Y_{i,(j)} \mid \hat{\theta}_i\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(r(N_i,n_i,j))}$$

Consider $r(N_i, n_i, j) = N_i - n_i + j$, then the relation becomes

$$F^{-1}_{X_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(Y_{i,(j)} \mid \theta_i\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(N_i-n_i+j)}$$

We estimate the above with

$$F^{-1}_{X_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(Y_{i,(j)} \mid \hat{\theta}_i\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(N_i-n_i+j)}$$

2.2 Non-parametric distribution measuring the components

2.2.1 Past methods and challenges of non-parametric approach

There are three kinds of non-parametric method that are widely used in estimating the cumulative distribution function, such as piecewise linear function estimation[Leenaerts and Bokhoven, 1998], Kaczynski et al. [2012], kernel estimation[Silverman, 1986] and semi-parametric conjugated estimation [Scholz, 1995]. One of the important issues of non-parametric method is extrapolation problem: how to generate samples below the first order statistics and above the last order statistics. In other word, how to estimate the boundary of the samples using non-parametric method.

The methodology of Kaczynski et al. [2012] is to stretch and shift the original data values so that the mean and variance of the piecewise-linear cumulative density function model matches the mean and variance of the sample values. It partially solves extrapolation problem mentioned above. The kernel estimation from Silverman [1986] does not have special treatment of the tail of the distribution. Semi-parametric conjugated estimation is widely used in dealing with tail behavior of the distribution. Scholz [1995] extends the scope of these nonparametric confidence bounds by introducing an adaptive type of QQ-plot, which plots the sample extremes against corresponding transformed probability using extreme value distribution. Stein [2020] uses parametric families of generalized Pareto distribution that have flexible behavior in both tails, which works well for estimating all quantiles when both tails of a distribution are heavy tailed. Both extreme value distribution and generalized Pareto distribution fail to provide reasonable estimate for the maximum values and computation becomes fairly tedious as the sample size increases.

All of these methods would work fine within the general Full House Model for the first specific applications what they motivate these particular methods. But in the application on baseball data, where the range of the distribution are naturally constrained, and the prevailing thought is that the outlying aptitudes for the influential points of the components is rare, these methods fail to capture the differences between the influential points and reminder of the components, which leads the results that are nonsensical.

2.2.2 New interpolated and extrapolated approach

In the nonparametric setting we motivate an another interpolated empirical CDF as an estimator of the system components distribution F_Y. The classical empirical CDF estimator $\widehat{F}_Y$ fails because it places cumulative probability 1 at the observation $Y_{i,(n_i)}$. Also the classical empirical CDF estimator places much higher cumulative probability at the tier values. We therefore consider surrogate sample points to construct an interpolated version of the empirical CDF $\tilde{F}_Y$ to alleviate this problem. [Kaczynski et al., 2012] provides a very similar interpolated CDF that has been widely used and it also use surrogated sample points to construct empirical CDF given same mean and variance as raw data, but it fail to show the distance between empirical CDF is decreasing as the sample size increases. We construct $\widetilde{F}_Y$ in the following manner: We first construct surrogate sample points $\widetilde{Y}_{i,(1)}, \ldots, \widetilde{Y}_{i,(n_i+1)}$ as,

$$
\begin{aligned}
\tilde{Y}_{i,(1)} &= Y_{i,(1)} - Y_i^* \\
\widetilde{Y}_{i,(j)} &= \left(Y_{i,(j)} + Y_{i,(j-1)}\right)/2, \quad j = 2, \ldots, n \\
\widetilde{Y}_{i,(n_i+1)} &= Y_{i,(n_i)} + Y_i^{**},
\end{aligned}
$$

where Y_i^* is the value to construct the lower bound and Y_i^{**} is the value to construct the upper bound. Technically Y_i^* and Y_i^{**} can be many different real numbers in different applications. With this construction, we build $\widetilde{F}_Y$ as

$$
\widetilde{F}_Y(t) = \sum_{j=1}^{n_i} \left(\frac{j-1}{n_i} + \frac{t - \tilde{Y}_{i,(j)}}{n_i \left(\tilde{Y}_{i,(j+1)} - \tilde{Y}_{i,(j)} \right)} \right) 1\left(\widetilde{Y}_{i,(j)} \leq t < \tilde{Y}_{i,(j+1)} \right) + 1\left(t \geq \tilde{Y}_{i,(n_i+1)} \right) \tag{1}
$$

Now we have

$$
\widetilde{F}_Y(Y_{i,(1)}) = \frac{1}{n_i} \frac{Y_{i,(1)} - \tilde{Y}_{i,(1)}}{\widetilde{Y}_{i,(2)} - \tilde{Y}_{i,(1)}} = \frac{1}{n_i} \frac{Y_i^*}{\frac{Y_{i,(2)} - Y_{i,(n_1)}}{2} + Y_i^*}
$$

$$
\widetilde{F}_Y(Y_{i,(n_i)}) = \frac{n_i - 1}{n_i} + \frac{1}{n_i} \frac{Y_{i,(n_i)} - \tilde{Y}_{i,(n_i)}}{\widetilde{Y}_{i,(n_i+1)} - \tilde{Y}_{i,(n_i)}} = \frac{n_i - 1}{n_i} + \frac{1}{n_i} \frac{\frac{Y_{i,(n_i)} - Y_{i,(n_i-1)}}{2}}{\frac{Y_{i,(n_i)} - Y_{i,(n_i-1)}}{2} + Y_i^{**}}
$$

2.2.3 Choosing the Y_i^* and Y_i^{**}

The estimator $\widetilde{F}_Y$ is desirable for two reasons. First, it does not assume that the observed minimum and observed maximum constitute the actual boundaries of the support of Y. Furthermore, $\widetilde{F}_Y\left(Y_{i,(1)}\right)$ and $\widetilde{F}_Y\left(Y_{i,(n_i)}\right)$ that are showed above provide reasonable estimates for the cumulative probability at $Y_{i,(1)}$ and $Y_{i,(n_i)}$ by considering their respective value of Y_i^* and Y_i^{**}. $\widetilde{F}_Y\left(Y_{i,(1)}\right)$ and $\widetilde{F}_Y\left(Y_{i,(n_i)}\right)$ are two functions of the observed data and, Y_i^* and Y_i^{**}. If we can choose the Y_i^* and Y_i^{**} sensibly in a way that captures the discrepancy of $Y_{i,(1)}$ and $Y_{i,(n_i)}$ from the remainder of the components, then we can perfectly control the cumulative probability of Y_i^* and Y_i^{**}.

Since the Y_i^* and Y_i^{**} are two important part of the tail behavior of the interpolated distribution, we are fairly careful of choosing the value of Y_i^* and Y_i^{**}. Note that Y_i^* determines the lower tail behavior of talent distribution and in fact most normal or low talents would concentrate in the similar scale or size [Newman, 2005]. Therefore Y_i^* should be a small positive value and we choose Y_i^* is equal to $Y_{i,(2)} - Y_{i,(1)}$. We would expect a more similar and reasonable tail behavior of talent distribution around $Y_{i,(1)}$.

Choice of Y_i^{**} is harder than the Y_i^*. We try out extreme value distribution and generalized Pareto distribution to fit the upper tail behavior of the interpolated distribution, but both methods fail to show the fact that when some components stand out from their peers, it is unlikely to observe the other components that better than the outstanding components from the upper tail of the distribution. This also indicates the larger the discrepancy between the outstanding components and their peers, the less likely we observe the other components that better than the outstanding components from the upper tail of the distribution.

To solve it, we consider a minimization problem that if a component $Y_{i,j}$ is transported to a new system k, the individual aptitude of $Y_{i,j}$ in system i should be similar to the individual aptitude of $Y_{i,j}$ in system k. However, it is fairly difficult to define a primary Y^{**} and compute the rest of them.

Scholz [1995] develops a new non-parametric tail extrapolation method and by using one term Taylor expansion on extreme quantile of distribution, the generalized least square model is well-established. Let $Y_{i,1}, \ldots, Y_{i,n_i}$ be a random sample from the ith system with continuous cumulative distribution function $F_Y(y) = P\left(Y_{i,j} \leq y\right)$, and denote by $Y_{i,(n_i)} \geq Y_{i,(n_i-1)} \geq \ldots \geq Y_{i,(1)}$ the ordered sample, in order from largest to smallest. The p-quantile y_p of F is defined as the smallest value for which $F\left(y_p\right) = p$, i.e. $y_p = \inf\{y : F(y) \geq p\}$. Hence $P\left(Y_{i,j} \leq y_p\right) = p$.

Suppose p_{i,j,γ,n_i} is the value that satisfies $\gamma = P\left(Y_{i,j} \geq y_p\right)$, then for $\gamma = 0.5$, an excellent approximation [Hoaglin et al., 1983] of p_{i,j,γ,n_i} is given by

$$p_{i,j,.5,n_i} \approx \frac{i - \frac{1}{3}}{n_i + \frac{1}{3}}.$$

Then we make an appeal to linear extrapolation with widely used transformed percentile scales as in Scholz [1995]. The specific transformations that we use are functions which are linear in its arguments, linear or quadratic in the logistic function of the arguments. More formally, let $g(\cdot)$ denote one of these transformations. Then we consider the regression fit on points $(g(p_{j,i,.5,n_i}), Y_{i,(j)})$, $j = n_i - k + 1, \ldots, n_i$. To pick Y_i^{**} we solve the following optimization problem

$$Y_i^{**} = \operatorname{argmin}_y \left| g^{-1}(Y_{i,(n_1)}) - \widetilde{F}_{Y_i}(Y_{i,(n_1)}; y) \right|,$$

where $\widetilde{F}_{Y_i}(\cdot; y)$ is $\widetilde{F}_{Y_i}$ defined with respect to a value y replacing Y_i^{**} in its construction. Notice that $\widetilde{F}_Y(t)$ is explicitly constructed to be close to $\widehat{F}_Y(t)$. We formalize this statement below.

Proposition 2.1. *Let $\widetilde{F}_Y(t)$ be defined as in (2) and let $\widehat{F}_Y(t)$ be the empirical distribution function. Then,*

$$\sup_{t \in \mathbb{R}} \left| \tilde{F}_Y(t) - \widehat{F}_Y(t) \right| \leq \frac{1}{n}$$

This leads to a Glivenko-Cantelli result which is appropriate for $\widetilde{F}_Y$.

Corollary 2.1.1. *Let $\widetilde{F}_Y(t)$ be defined as in (1) and let $\widehat{F}_Y(t)$ be the empirical distribution function. Then,*

$$\sup_{t \in \mathbb{R}} \left| \widetilde{F}_Y(t) - F_Y(t) \right| \xrightarrow{a.s} 0$$

More properties of $\widehat{F}_Y$ are provided in the Appendix. Corollary 2.1.1 shows that the interpolated empirical distribution function is a serviceable estimator for F_Y. We will make use of the following approximations to facilitate our methodology,

$$\widetilde{F}_Y\left(Y_{i,(j)}\right) \approx U_{i,(j)}, \quad \tilde{F}_{Y_{(j)}}\left(Y_{i,(j)}\right) \approx U_{i,j}$$

where $U_{i,j} \sim U(0,1)$ and $U_{i,(j)} \sim \text{Beta}(j, n_i + 1 - j)$ and the quality of the approximation in the right hand side depends upon the sample size and the shape of F_Y. We now connect the order statistics to the underlying distribution that comes from a population with $N_i \geq n_i$ observations when F_X is known. We estimate the hidden trait values by with

$$F^{-1}_{X_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(y_{i,(j)}\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(r(N_i,n_i,j))}$$

We estimate the above with

$$F^{-1}_{X_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(\widetilde{F}_{Y_i}\left(y_{i,(j)}\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(r(N_i,n_i,j))}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(r(N_i,n_i,j))}$$

Consider $r(N_i, n_i, j) = (N_i - n_i + j)$, then the relation becomes

$$F^{-1}_{X_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(F_{Y_i}\left(y_{i,(j)}\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(N_i-n_i+j)}$$

We estimate the above with

$$F^{-1}_{X_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(\widetilde{F}_{Y_i}\left(y_{i,(j)}\right)\right)\right) \sim F^{-1}_{X^{-1}_{i,(N_i-n_i+j)}}\left(F_{U_{i,(j)}}\left(U_{i,(j)}\right)\right) \sim X_{i,(N_i-n_i+j)}$$

2.3 Estimate how components will perform in another systems

We now take the hidden trait values and reverse the process to extract the components if these components were from a new system, where the distribution for aptitude X is known. More formally, the hypothetical components arise from a new system k is computed as:

Parametric distribution:

$$Y_{k,r^{-1}(N_k,n_k,j)} = F^{-1}_{Y_k}\left(F^{-1}_{U_{k,(r^{-1}(N_k,n_k,j))}}\left(F_{X_{k,(r(N_k,n_k,j))}}\left(X_{i,(r(N_i,n_i,j))}|\hat{\theta}_i\right)\right)\right)$$

Non-Parametric distribution:

$$Y_{k,r^{-1}(N_k,n_k,j)} = \widetilde{F}^{-1}_{Y_k}\left(F^{-1}_{U_{k,(r^{-1}(N_k,n_k,j))}}\left(F_{X_{k,(r(N_k,n_k,j))}}\left(X_{i,(r(N_i,n_i,j))}\right)\right)\right)$$

where $X_{i,r(N_i,n_i,j)}$ is the hidden trait from the ith system.

3 Full House Model on Baseball Data

3.0.1 Background and eligible baseball population

Comparing the achievements of baseball players across eras has resulted in endless debates among scientists [Kvam, 2011], participants in social media platforms, network personalities, family members and friends. Of all the possible across-era comparisons to be made, the comparison of baseball players' statistics has a lively discussion in the scientific literature[Berry et al., 1999, Schell, 2016, Eck, 2020, Gould, 2011, Petersen and Penner, 2020, Petersen et al., 2011, Schmidt and Berri, 2005]. The methodology of Petersen et al. [2011] and Petersen and Penner [2020](PPS) is to detrend the statistics and account for changes in the components of the systems resulting from both exogenous and endogenous. But as we point out in Eck [2020], PPS misunderstands the effect of talent dilution from expansion and ignores reality. The talent pool was more diluted in the earlier eras of baseball than now because of a small relative eligible population size and the exclusion of entire populations of people on racial grounds.

The methodology of Berry et al. [1999] is to use hierarchal model to estimate the innate ability of players, the effects of aging on performance, and the relative difficulty of each year within a sport. Compared with our method, Berry et al. [1999] ignores segregation, increases in the MLB eligible population relative to available roster spots, and increases in the average overall talent of that population. Second, the Bayesian method they use fails to compute the talent in some special eras. For example, about 80% of people MLB eligible population are removed from the major league due to the World War II and overall talent in the MLB should got worse. But Berry et al. [1999] can not capture this change and only cares about distribution of the statistics conditioned on the MLB players. Therefore, their methodology does not fully address the characteristics of a changing talent pool. Third, in Berry et al. [1999], they assumes the getting a hit and hitting a home run follows binomial distribution. This assumption makes sense but has limitation that we can only perform this method on the statistics that we know the distribution of it. For the distribution of Win Above Replacement, it does not follow some common distributions and it has heavy tail. Schell [2016] makes a similar assumption as our method that after adjusting for ballpark effects, a pth percentile player in one year is equal in ability to a pth percentile player in another year for each basic offensive event, but he fails to consider the baseball eligible population effect.

In this application, the Full House Model is used to construct an era-neutral environment which allows for comparisons of the different statistics of baseball players from fundamentally different eras. In our setting, we consider the statistics in the ith season is the ith system in our Full House Model and suppose the system selects the highest aptitudes, so that $r(N_i, n_i, k) = (N_i - n_i + k)$. We motivate this methodological approach through the goal of inferring the values of $X_{i,N_i-n_i+j}, j = 1, \ldots, n_i$ from the observed values of $Y_{i,1}, \ldots, Y_{i,n_i}$.

The MLB eligible population is not well-defined and we use the definition from Eck [2020]. MLB eligible population is the decennial count of males aged 20-29 that are living in the Aruba, Australia, Bahamas, Brazil, Canada, Colombia, Cuba, Curaçao, Dominican Republic, Jamaica, Japan, Mexico, Nicaragua, Panama, Puerto Rico, South Korea, Taiwan, the United States, the United States Virgin Islands, and Venezuela. Latin American countries' populations will often be added four years before their first MLB player reached the MLB.

. We estimate the MLB eligible population factoring in the changing levels of interest in the game and will reference the Baseball Reference [1], the US Census [2], Statistics Canada [3], Gallup survey data [4] and Wikipedia [5]. We also closely follow the calculation made in Eck [2020].

[1] https://www.baseball-reference.com/

[2] https://www.census.gov/

[3] https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

[4] https://news.gallup.com/poll/4735/sports.aspx

[5] https://en.wikipedia.org/wiki/Main_Page

In the Table 1, The cumulative population proportion means that at each era, the population of the previous eras is also included. As an example of how to interpret this dataset, consider the year 1950. There were 3.41 million eligible males aged 20-29. The proportion of the historical MLB eligible population that existed at or before 1950 is 0.141.

year	population	cumulative population proportion
1870	0.39	0.004
1880	0.56	0.011
1890	0.67	0.019
1900	0.79	0.028
1910	1.27	0.042
1920	1.05	0.054
1930	1.36	0.070
1940	2.82	0.102
1950	3.41	0.141
1960	5.62	0.205
1970	7.80	0.294
1980	9.30	0.400
1990	8.18	0.494
2000	14.14	0.655
2010	14.50	0.820
2020	15.73	1.000

Table 1: Eligible MLB population throughout the years. The 1st column indicates the year; 2nd column indicates the estimated eligible population size(in millions) and the 3rd column indicates the proportion of the MLB eligible population in row x that was eligible at or before row x.

We also suppose the underlying talent follows Pareto(α) distribution, which is motivated by Berri and Schmidt [2010] and α we select is 1.16. They mentioned that the superstars are really important for the teams to win and state that about 80% of wins appear to be produced by top 20% of the players. We will additionally assume that talent distributions are independent and identically distributed across years.

We analyze data the from Chadwick Baseball Bureau database [cha], Baseball-Reference [Baseball-Reference] and FanGraphs [FanGraphs], which have batting and pitching data for over 17,000 players from 1871 season to 2021 season.

3.1 Batting statistics

In this section, we explore the talents of four batting statistics, such as BA, HR, bWAR and fWAR for batters and era-adjusted career statistics using our Full House Model. WAR is a primary example of one number summary to measure the player's total values and contributions to wins. We use the batting data from Chadwick Baseball Bureau database, FanGraphs and Baseball Reference from 1871 season to 2021 season and make some modifications, such as combining the statistics when players played in the different leagues in a single season.

Park factor adjustment [6] is also considered in our model and we apply the adjusted park index from [Schell, 2016] to all ballpark from 1871 season to 2021 season. BA and HR are two statistics that can be affected by the ballpark and we apply the park factor adjustment to these two statistics.

We will use the parametric distribution to measure the BA since it is widely recognized that the BA follows a normal distribution [Gould, 2011]. Also we perform Shapiro-Wilk tests [Shapiro and Wilk, 1965] of normality on the BA of each season and the p values of BA in 121 seasons out of total 151 seasons are greater than 0.05.

We will use nonparametric methods to measure the HR, bWAR and fWAR since they do not follow any common distribution we know. We also use HR per at bats (AB), bWAR per game and fWAR per game as the components in the system to compute the talent score corresponding to HR, bWAR and fWAR.

We restrict attention to full-time players. We define the full-time hitter cutoff as the median plate appearances (PA) after screening out hitters who batted fewer than 75 PAs. Then the full-time hitters are the hitters batted greater than PA cutoff and we include the full-time hitters in the system.

Once the underlying talents for the four batting statistics in all seasons are calculated, we could map back the underlying talents and estimate the BA, HR, bWAR and fWAR of baseball players in some old or modern seasons even they never played. Before we estimate the four statistics in some old or modern seasons even they never played, we notice Walks (BB), Hit-By-Pitch (HBP), Sacrifice Bunt (SH) and Sacrifice Fly (SF) are important in estimating the number of ABs, and the BB changes in the different eras. It is widely acknowledged that HBP, SH and SF do not change over time. Then we calculate the era-adjusted BB and calculate the era-adjusted AB as

$$\text{adjusted-AB} = \text{mapped-PA - adjusted-BB - HBP - SH - SF},$$

where the mapped-PA is calculated by apply quantile mapping for the full-time hitters and non-full-time hitters. Then the era-adjusted home run total is obtained by multiplying estimated HR per AB with adjusted-AB.

Instead of using the raw games in the dataset, we calculate the mapped games by applying quantile mapping for the full-time hitters and non-full-time hitters. Then the era-adjusted bWAR and fWAR are obtained by multiply estimated bWAR per Game and fWAR per Game with mapped game.

We now extend our model to compute the hypothetical careers in which we suppose that every player who start their career in 2021 and compare their four statistics at the same span. We take the talents scores $X_{i,j}$ and reverse the process to extract the four predicted statistics of the players if these seasons were to take place in 2021 to present. More formally, the hypothetical 2021 statistics for player j in year i of BA is computed as

$$Y_{2021,j} = F^{-1}_{Y_{2021}}\left(F^{-1}_{U_{2021,(j)}}\left(F_{X_{2021,(N_{2021}-n_{2021}+j)}}\left(X_{i,(N_i-n_i+j)}|\hat{\theta}_i\right)\right)\right) \tag{2}$$

More formally, the hypothetical 2021 statistics for player j in year i of HR, bWAR and fWAR is computed as

$$Y_{2021,j} = \widetilde{F}^{-1}_{Y_{2021}}\left(F^{-1}_{U_{2021,(j)}}\left(F_{X_{2021,(N_{2021}-n_{2021}+j)}}\left(X_{i,(N_i-n_i+j)}\right)\right)\right) \tag{3}$$

[6] https://en.wikipedia.org/wiki/Batting_park_factor

We only consider the careers of batters who have 4000 career at bats. Then we rank MLB players by their era-adjusted hypothetical career BA, HR, bWAR and fWAR under the scenario that every player began their career in 2021. The players who began their career before 1950 are highlighted in the Table 7, Table 8, Table 9 and Table 10 in the Appendix.

Figure 1 illustrate the yearly effect for BA from 1871 to 2021. It shows that the difficulty of getting a base hit for a batter has decreased since the early 1920s. It also shows that the talent of batters has decreased from 1940s to 1950s. This is probably due the WWII and a large portion of people are removed from the eligible baseball population.

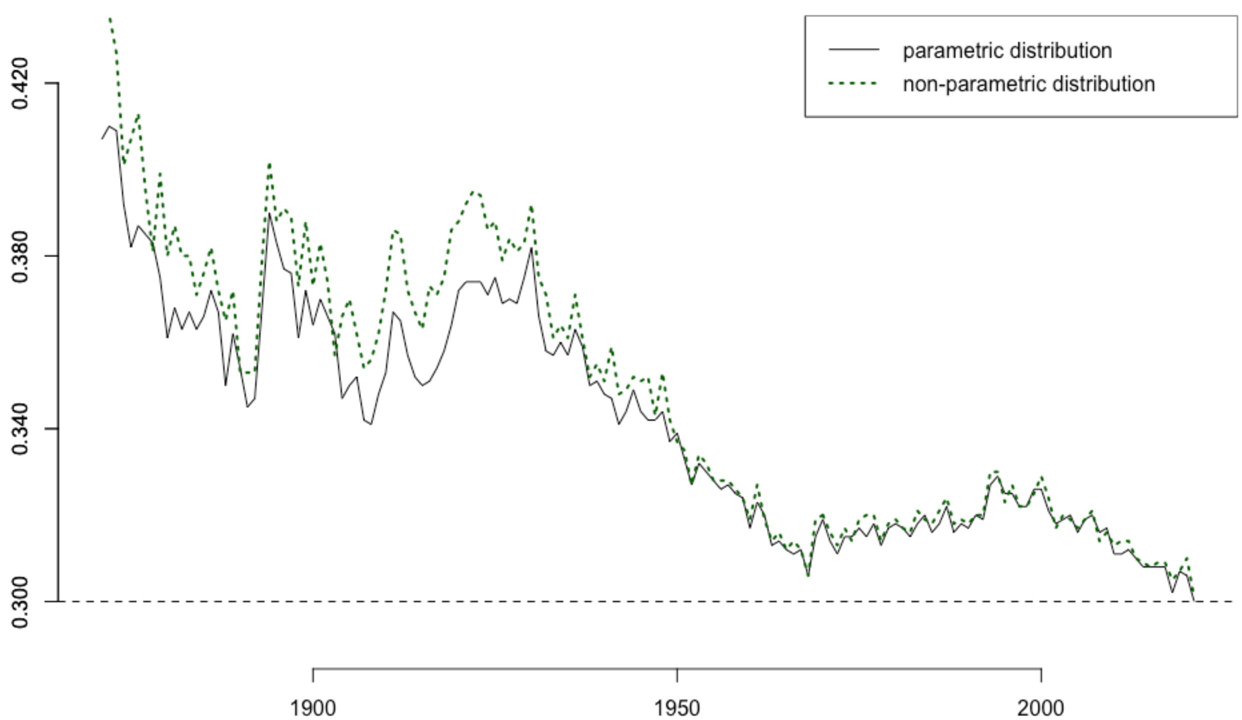

Figure 1: The year effect for the batting average study from Full House Model using parametric and non-parametric distribution measuring the components. The batting average plot shows the estimated batting average for a player who is a .300-batter in 2021.

The bottom plot in Figure 1 is the year effect for the batting averages study from Full House Model using nonparametric distribution measuring the components, and it is fairly similar with the first plot using parametric distribution. We can conclude that there is no significant difference between the models using parametric distribution and nonparametric distribution.

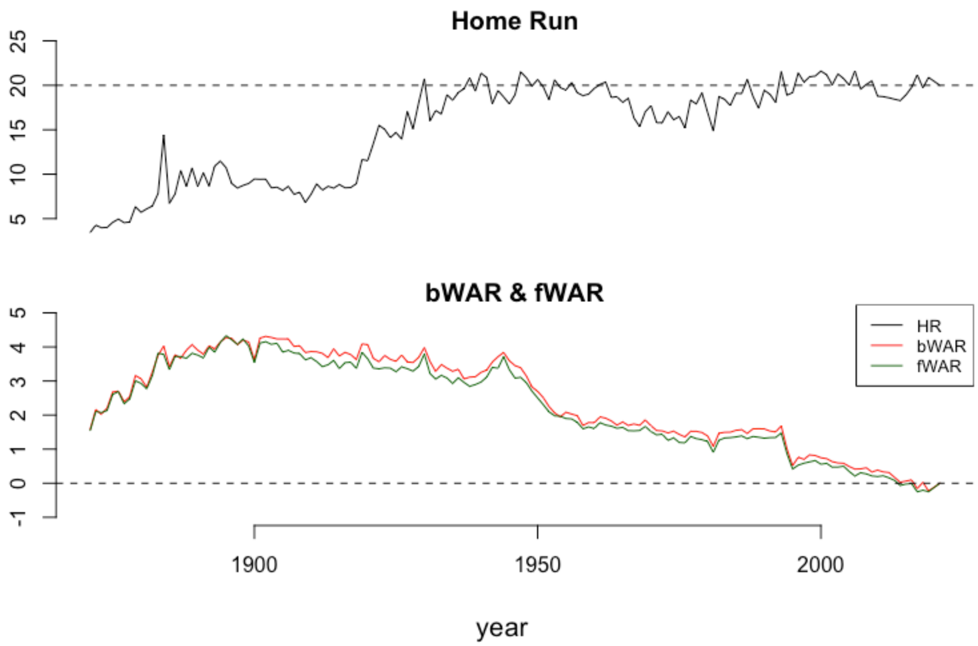

Figure 2: The year effect for the HR, bWAR and fWAR for batters study from Full House Model. The three plots show the smoothed estimated HR, bWAR and fWAR for batters for a player who achieve 20 home runs, 0 bWAR and 0 fWAR in 2021.

Figure 2 illustrate the yearly effect for home runs, bWAR and fWAR for batters from 1871 to 2021. It shows that after 1920, when the dead-ball era ended, the difficulty of hitting home runs has not changed a great deal. A 20-home run hitter in 2021 is estimated to have hit about 25 in the mid-1920s. It also shows that the bWAR and fWAR talent of batters has decreased from 1940s to 1950s. This is probably due the WWII and a large portion of people are removed from the eligible baseball population.

3.2 Pitching statistics

In this section, we explore the talents of four pitching statistics, such as earned run average (ERA), Strikeouts (SO), bWAR and fWAR for pitchers and era-adjusted career statistics using our Full House Model. We use the pitching data from Chadwick Baseball Bureau database, FanGraphs and Baseball Reference from 1871 season to 2021 season and make some modifications, such as combining the statistics when players played in the different leagues in a single season.

Park factor adjustment [7] is also considered in our model and we apply the adjusted park index from [Schell, 2016] to all ballpark from 1871 season to 2021 season.

We will use the nonparametric distribution to measure the ERA, SO, bWAR and fWAR since they do not follow any common distribution we know. We typically use the negative ERA, SO per 9 innings pitched (IP), bWAR per IP and fWAR per IP as the components in the system to compute the talent score for them since we would expect a high talent score for smaller ERA value and they are more reasonable than the raw statistics.

[7] https://en.wikipedia.org/wiki/Batting_park_factor

To define the full-time pitchers, we compute the number of average starting pitchers by measuring the average rotation size for each team and multiplying it with the numbers of team in each season. Then the full-time pitchers are the pitchers who are most innings pitched and the number of them is the same as the number of average starting pitchers. The computing the number of average starting pitchers is motivated by the rotation size changes significantly in different eras.

Once the underlying ERA, SO, bWAR and fWAR talents in all seasons are calculated, we could map back the underlying talents and estimate the ERA, SO, bWAR and fWAR of baseball players in some old or modern seasons even they never played. Before mapping the talents, the talent for old-era pitchers need to be adjusted since their rotation size was smaller than the rotation size in nowadays, and it is not fair to compare with the peers in a modern setting with a larger rotation size. For the pitchers who are not included in the pitching rotation in some old seasons when the rotation size is about 3 or 4, they may be included in the rotation in the nowadays, but their statistics and talent could be smaller due to the size of the rotation. The details of how to adjust the talents based on the rotation size are in the Appendix.

We now extend our model to compute the hypothetical careers in which we suppose that every player who start their career in 2021 and compare their ERA, SO, bWAR and fWAR at the same span. We take the talents scores $X_{i,j}$ and reverse the process to extract predicted ERA, SO, bWAR and fWAR of the players if these seasons were to take place in 2021 to present. More formally, the hypothetical 2021 ERA, SO, bWAR and fWAR for player j in year i is computed as

$$Y_{2021,j} = \widetilde{F}^{-1}_{Y_{2021}} \left(F^{-1}_{U_{2021,(j)}} \left(F_{X_{2021,(N_{2021}-n_{2021}+j)}} \left(X_{i,(N_i - n_i + j)} \right) \right) \right) \tag{4}$$

We only consider the careers of pitchers who have over 1000 career Innings Pitched. Then we rank MLB players by their era-adjusted hypothetical career ERA, SO, bWAR and fWAR under the scenario that every player began their career in 2021.The players who began their career before 1950 are highlighted in the Table 11, Table 12, Table 13 and Table 14.

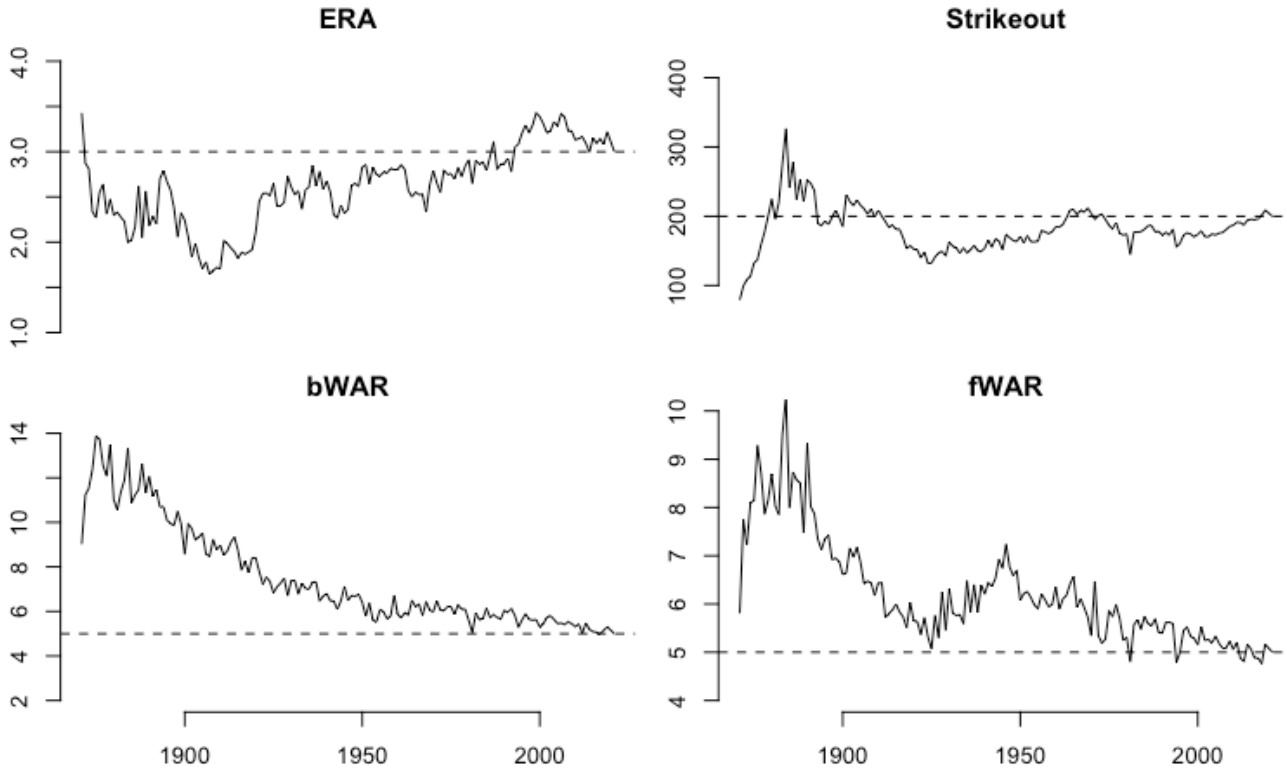

Figure 3: The year effect for the ERA, SO, bWAR and fWAR for pitchers study from the Full House Model. The three plots show the smoothed estimated ERA, SO, bWAR and fWAR for pitchers if a player who achieve 3 ERA, 200 SO, 5 bWAR and 5 fWAR in 2021.

Figure 3 illustrate the yearly effect for ERA, SO, bWAR and fWAR for pitchers from 1871 to 2021.

	name	ebWAR	name	efWAR
1	Barry Bonds	131.03	Roger Clemens	131.96
2	Roger Clemens	125.77	Barry Bonds	131.44
3	Willie Mays	121.71	Willie Mays	117.73
4	Henry Aaron	115.79	Henry Aaron	114.02
5	**Babe Ruth**	109.81	Greg Maddux	108.43
6	Alex Rodriguez	102.65	**Babe Ruth**	105.14
7	Greg Maddux	98.14	Alex Rodriguez	98.25
8	Randy Johnson	93.82	Randy Johnson	97.81
9	Rickey Henderson	92.69	Mike Schmidt	92.31
10	Mike Schmidt	92.18	Nolan Ryan	90.42
11	**Stan Musial**	89.96	Rickey Henderson	89.58
12	Albert Pujols	89.30	**Ted Williams**	88.43
13	Frank Robinson	87.63	**Stan Musial**	85.63
14	Tom Seaver	86.47	Frank Robinson	84.02
15	Adrian Beltre	85.88	Steve Carlton	82.53
16	**Ted Williams**	83.94	Bert Blyleven	82.15
17	**Walter Johnson**	82.47	Cal Ripken Jr.	81.97
18	Cal Ripken Jr.	82.35	Gaylord Perry	81.26
19	**Cy Young**	80.59	**Cy Young**	80.13
20	**Ty Cobb**	80.12	**Ty Cobb**	80.10
21	**Lefty Grove**	79.82	Mickey Mantle	79.88
22	Chipper Jones	79.21	**Walter Johnson**	78.90
23	Bert Blyleven	79.14	Albert Pujols	78.37
24	Eddie Mathews	76.67	**Lefty Grove**	78.05
25	Mike Mussina	76.18	Adrian Beltre	77.80

Table 2: Top 25 bWAR and fWAR leaders for MLB players with era-adjusted hypothetical career in 2021. 1st and 3rd columns are the name of the players; 2nd column is the estimated bWAR using the Full House Model; 4th column is the estimated fWAR using full house model.

Table 2 shows the top 25 MLB leaders in bWAR and fWAR using the Full House Model.

4 Compare rankings and era-adjusted method

4.1 Compare rankings from different sources

The table below displays baseball's all-time greatest players according to five sources. The first source is Ranker which is the overall rankings by baseball fans. The second and third sources are, respectively, bWAR and fWAR. The fourth source is from ESPN, and it is a proxy measure for the overall rankings among sports journalists. The fifth source is from the Hall of Stats, which removed all 235 inductees and replaced them with the top 235 eligible players in history, according to a mathematical formula. Players who started their career in 1950 or before appear in bold text.

rank	Ranker	bWAR	fWAR	ESPN	Hall of Stats
1	**Babe Ruth**	**Cy Young**	**Babe Ruth**	**Babe Ruth**	**Babe Ruth**
2	Willie Mays	**Babe Ruth**	Barry Bonds	Willie Mays	Barry Bonds
3	**Lou Gehrig**	Barry Bonds	Willie Mays	Barry Bonds	**Walter Johnson**
4	**Ty Cobb**	Willie Mays	**Ty Cobb**	**Ted Williams**	Willie Mays
5	**Ted Williams**	**Walter Johnson**	**Honus Wagner**	Hank Aaron	**Cy Young**
6	Hank Aaron	**Ty Cobb**	Hank Aaron	**Ty Cobb**	**Ty Cobb**
7	**Cy Young**	Hank Aaron	Roger Clemens	Roger Clemens	Hank Aaron
8	**Walter Johsnon**	Roger Clemens	**Cy Young**	**Stan Musial**	Roger Clemens
9	**Rogers Hornsby**	**Tris Speaker**	**Tris Speaker**	Mickey Mantle	**Rogers Hornsby**
10	**Honus Wagner**	**Honus Wagner**	**Ted Williams**	**Honus Wagner**	**Honus Wagner**
11	Mickey Mantle	**Stan Musial**	**Rogers Hornsby**	**Lou Gehrig**	**Tris Speaker**
12	**Joe Dimaggio**	**Rogers Hornsby**	**Stan Musial**	**Walter Johnson**	**Ted Williams**
13	**Stan Musial**	**Eddie Collins**	**Eddie Collins**	Greg Maddux	**Stan Musial**
14	**Joe Jackson**	**Ted Williams**	**Walter Johsnon**	Rickey Henderson	**Eddie Collins**
15	**Jimmie Foxx**	Alex Rodriguez	Greg Maddux	**Rogers Hornsby**	**Pete Alexander**
16	**Christy Mathewson**	**Kid Nichols**	**Lou Gehrig**	Mike Schmidt	**Lou Gehrig**
17	Roberto Clemente	**Pete Alexander**	Alex Rodriguez	**Cy Young**	Mickey Mantle
18	**Jackie Robinson**	**Lou Gehrig**	Mickey Mantle	Joe Morgan	**Lefty Grove**
19	Johnny Bench	**Lefty Grove**	**Mel Ott**	**Joe Dimaggio**	**Mel Ott**
20	**Warren Spahn**	Rickey Henderson	Randy Johnson	Frank Robinson	Rickey Henderson
21	Ernie Banks	**Mel Ott**	Nolan Ryan	Randy Johnson	**Kid Nichols**
22	**Satchel Paige**	Mickey Mantle	Mike Schmidt	Tom Seaver	Mike Schmidt
23	**Yogi Berra**	Frank Robinson	Rickey Henderson	Alex Rodriguez	**Nap Lajoie**
24	Ken Griffey Jr	**Nap Lajoie**	Frank Robinson	**Tris Speaker**	**Christy Mathewson**
25	Bob Gibson	Mike Schmidt	Bert Blyleven	Steve Carlton	Greg Maddux
pre-1950 in top 25:	17/25	16/25	12/25	11/25	17/25

Table 3: Lists of the top 25 greatest baseball players to ever play in the MLB according to Ranker.com (1st column), bWAR (2nd column), fWAR (3rd column), and ESPN (4th column), Hall of Stats (5th column). Players that started their career before 1950 are indicated in bold text. The last row counts the number of players that started their careers before 1950 in top 25 lists.

Given the assumptions in Eck [2020], we could calculate the chance of extreme event in top 25 list using the Binomial distribution. The results provided in Table 2 present overwhelming evidence that players who started their careers before 1950 are overrepresented in top 25 list from the perspectives of fans, analytic assessment of performance, and experts' rankings.

Ranking list	chance of extreme event in top 25 list
Ranker	1 in 7544203
bWAR	1 in 922605
fWAR	1 in 992
ESPN	1 in 102
Hall of Stats	1 in 7544203
Full House with fWAR	1 in 3
Full House with bWAR	1 in 3

Table 4: The change of each extreme event calculation corresponding to the seven lists in Table 2 and Table 3

As an example of how to interpret the results of Table 3 with Ranker's top 25 list, the Table 3 shows that the probability of observing 17 or more players that started their career at or before 1950 of the top 25 all time players, based on the population dynamics, is about 1 in 7544203. The same interpretation applies to remainder of Table 3. We see that all approaches have drastically over-included players from the pre-1950s time period in their all-time rankings.

4.2 Compare with Era Bridging method

In the Section 3 we point out three disadvantages of era-bridging method compared with our Full House Model and we compare our results with the results from era-bridging method. Berry et al. [1999] provides two tables of top 25 peak players for the BA study (Table 9) and HR study (Table 10) on every player who has batted in MLB in the modern era (1901-1996) by accounting for the benchmark year of 1996. We also apply our Full House Model to the same dataset and same reference year with Berry et al. [1999].

	Full House			Era Bridging		
1	Tony Gwynn	1987	0.391	**Ty Cobb**	1886	0.368
2	**Nap Lajoie**	1904	0.387	Tony Gwynn	1960	0.363
3	Rod Carew	1977	0.384	**Ted Williams**	1918	0.353
4	**Honus Wagner**	1908	0.376	Wade Boggs	1958	0.353
5	Willie McGee	1985	0.376	Rod Carew	1945	0.351
6	**Tris Speaker**	1916	0.372	**Joe Jackson**	1889	0.347
7	Henry Aaron	1959	0.372	**Nap Lajoie**	1874	0.345
8	**Ty Cobb**	1912	0.372	**Stan Musial**	1920	0.345
9	Norm Cash	1961	0.372	Frank Thomas	1968	0.344
10	Wade Boggs	1985	0.372	**Ed Delahanty**	1867	0.340
11	Joe Torre	1971	0.371	**Tris Speaker**	1888	0.339
12	Cecil Cooper	1980	0.369	**Rogers Hornsby**	1896	0.338
13	Kirby Puckett	1988	0.368	Hank Aaron	1934	0.336
14	**Rogers Hornsby**	1924	0.367	Álex Rodríguez	1975	0.336
15	Robin Yount	1982	0.366	Pete Rose	1941	0.335
16	Alan Trammell	1987	0.365	**Honus Wagner**	1874	0.333
17	Alex Rodriguez	1996	0.365	Roberto Clemente	1934	0.332
18	**George Sisler**	1922	0.365	George Brett	1953	0.331
19	Pete Rose	1973	0.364	Don Mattingly	1961	0.330
20	Mickey Mantle	1956	0.364	Kirby Puckett	1961	0.330
21	Frank Thomas	1996	0.363	Mike Piazza	1968	0.330
22	Ralph Garr	1974	0.363	**Eddie Collins**	1887	0.329
23	**Cy Seymour**	1905	0.362	Edgar Martinez	1963	0.328
24	Mike Piazza	1996	0.362	Paul Molitor	1956	0.328
25	Willie Mays	1960	0.361	Willie Mays	1931	0.328

Table 5: Top 25 peak players for the BA study on every player who has batted in MLB in the modern era (1901-1996) by accounting for the benchmark year of 1996 using Full House Model and Era bridging conditioned on at least 500 at bats. Players that started their career before 1950 are indicated in bold text.

From the Table 5, 7 out of 25 players started their career before 1950 from Full House Model and the pre-150s cumulative population proportion during the modern era (1901-1996) is 0.257, which is calculated from the Table 1. Then the probability of observing 7 or more players that started their career at or before 1950 of the top 25 all time players, based on the population dynamics, is about 1 in 2.12, which is 0.47.

Compared with Full House Model, 10 out of 25 players started their career before 1950 from the era-bridging list and the probability of observing 10 or more players that started their career at or before 1950 of the top 25 all time players is 1 in 11.95, which is 0.08.

Therefore, the *p*-value from the era-bridging method is fairly small compared to Full House Model and it is unlikely to see 10 out of 25 players started their career before 1950 from the era-bridging list. In addition, the top two BA leaders in 1996 after taking into account the hits park factor are Alex Rodriguez and Frank Thomas and their BA in 1996 are 0.365 and 0.363, which are same to the result on the list. For the Era-bridging list, only two players' estimated BAs in 1996 are greater than 0.361, which goes against the fact that Jim Eisenreich does not appear on this list, who achieve 0.361 BA in 1996.

	Full House			Era Bridging		
	name	yearID	HR	name	Born	θ
1	Jose Canseco	1988	60	Mark MeGwire	1963	0.104
2	George Foster	1977	60	Juan Gonzalez	1969	0.098
3	Jim Wynn	1967	60	**Babe Ruth**	1895	0.094
4	Eddie Mathews	1953	58	Dave Kingman	1948	0.093
5	Willie Mays	1965	57	Mike Schmidt	1949	0.092
6	Albert Belle	1995	56	Harmon Killebrew	1936	0.090
7	Frank Howard	1968	56	Frank Thomas	1968	0.089
8	Willie Stargell	1971	56	Jose Canseco	1964	0.088
9	Johnny Bench	1972	55	Ron Kittle	1958	0.086
10	John Mayberry	1975	54	Willie Stargell	1940	0.084
11	Mark McGwire	1987	54	Willie McCovey	1938	0.084
12	Jim Rice	1983	54	Darryl Strawberry	1962	0.084
13	Gorman Thomas	1982	54	Bo Jackson	1962	0.083
14	Cecil Fielder	1990	53	**Ted Williams**	1918	0.083
15	Mickey Mantle	1956	53	**Ralph Kiner**	1922	0.083
16	Kevin Mitchell	1989	53	**Pat Seerey**	1923	0.081
17	Reggie Jackson	1969	52	Reggie Jackson	1946	0.081
18	Mike Schmidt	1980	52	Ken Griffey	1969	0.080
19	Henry Aaron	1957	51	Albert Belle	1966	0.080
20	**Babe Ruth**	1927	51	Dick Allen	1942	0.080
21	Darryl Strawberry	1988	51	Barry Bonds	1964	0.079
22	George Bell	1987	50	Dean Palmer	1968	0.079
23	Barry Bonds	1993	50	Hank Aaron	1934	0.078
24	Nate Colbert	1972	50	**Jimmie Foxx**	1907	0.078
25	Tony Armas	1983	49	Mike Piazza	1968	0.078

Table 6: Top 25 peak players for the HR study on every player who has batted in the MLB in the modern era (1901-1996) by accounting for the benchmark year of 1996 using Full House Model and Era bridging model conditioned on at least 500 at bats.

From the Table 6, 1 out of 25 players started their career before 1950 from Full House Model and the pre-1950s cumulative population proportion during the modern era is 0.257, which is calculated from the Table 1. Then the probability of observing 5 or more players that started their career at or before 1950 of the top 25 all time players, based on the population dynamics, is about 1 in 1.01, which is 0.99.

Compared with Full House Model, 5 out of 25 players started their career before 1950 and the probability of observing 5 or more players that started their career at or before 1950 of the top 25 all time players is 1 in 1.24, which is 0.81.

Therefore, it is possible that the pre-1950s time period could have produced 1 or 5 historically great baseball players during the modern era based on home runs and both model work well on estimating the leaders in home runs. However, θ in the era-bridging model does not help on estimating the adjusted home runs. The estimated home runs could be over 200 by applying the formulas in Baumer et al. [2015].

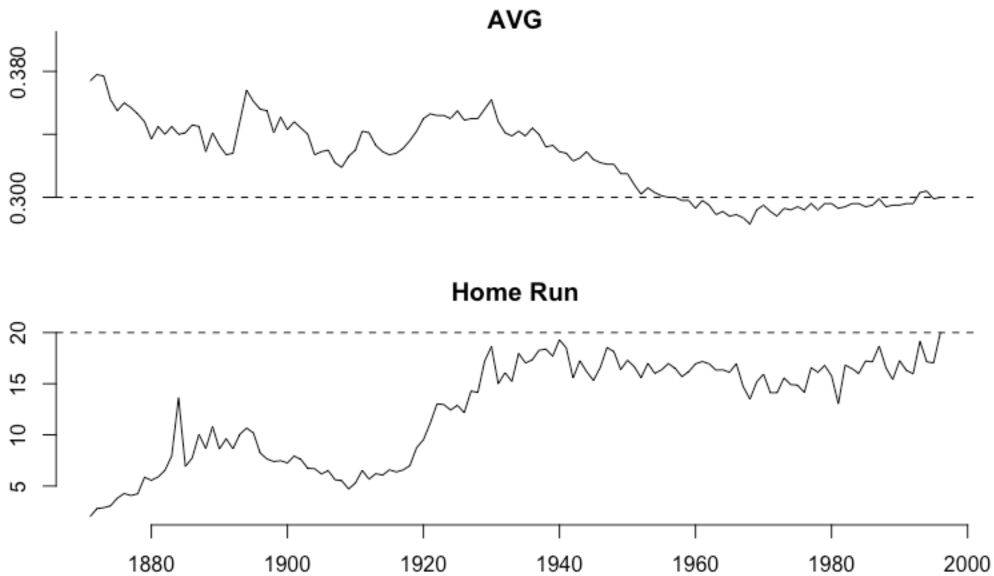

Figure 4: The year effect for BA (a) and HR (b) study from the Full House Model. The batting average plot shows how a .300 batter in 1996 would perform in other seasons. The home run plot shows how a 20 home run batter in 1996 would perform in other seasons.

Figure 4 illustrate the yearly effects for batting average and home runs. Compared with result in Berry et al. [1999], most of the seasons between 1900 to 1920 season are below 0.300 from era-bridging model, and this indicates the 300 batters from 1900 to 1920 season are more talented than the 300 batters in 1996, which is hard to believe. Compared with result in Berry et al. [1999], the batters in the early 1900s would get about 20 home runs, which is contradicted with the fact that in the dead-ball era, batters relied much more on plays such as the stolen bases and hit-and-run than on home runs. [Okrent et al., 2000].

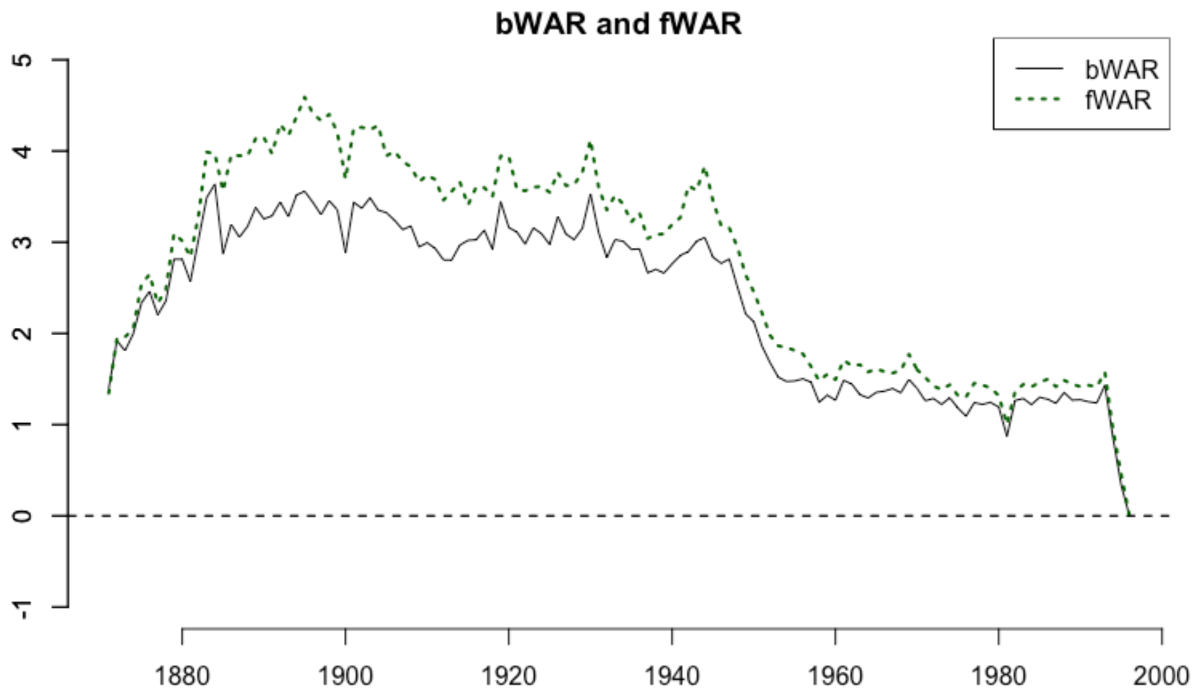

Figure 5: The year effect for the bWAR and fWAR study from the Full House Model. The bWAR and fWAR plot shows how a replacement batter (0 WAR batter) in 1996 would perform in other seasons.

Figure 5 illustrate the yearly effects for WAR and show that WAR has slowly decreased over the years, which is consistent with fact that the population of players in baseball is continually improving. We can clearly see the sudden decline in talent between 1940s and 1950s because of world War II, when more than 80% of the eligible players joined the army. This phenomena is quite important in baseball history but can not be captured by era-adjusted method.

5 Summary and Discussion

In this article we have developed a model motivated by Stephen J. Gould's book Full House: The Spread of Excellence from Plato to Darwin [Gould, 2011] for making statistical inference on cross-system components. Then we apply this model to several important statistics in baseball and obtain fairly reasonable results with era-adjusted hypothetical career. These results challenge the nostalgia from media and fans, and the MLB players from the early eras of baseball are overrepresented in rankings of the greatest players.

In Stephen Jay Gould's video [8], he mentioned that Wade Boggs would hit 0.420 or more if he played in 1900s since he was sitting near the limit of excellent. We apply our Full House Model to Wade Boggs assuming he starts his baseball career in 1901. The batting average Wade Boggs

[8] https://www.youtube.com/watch?v=BNM6ait4LOc

achieve in 1901 and 1902 are both above 0.400. Although the estimated BAs do not reach 0.420, Wade Boggs is still fairly excellent when he was playing in 1900s. The possible reason that Stephen Jay Gould failed to make the correct prediction is he did not take park factor into account. After applying the Full House Model to the baseball dataset without considering park factor, we find the BA of Wade Boggs successfully reaches to 0.420.

real season	projeced season	estimated park-factored BA	estimated raw BA	adj AB
1982	1901	0.404	0.433	283
1983	1902	0.394	0.417	577
1984	1903	0.361	0.382	600
1985	1904	0.371	0.389	625
1986	1905	0.372	0.383	570
1987	1906	0.373	0.385	548
1988	1907	0.360	0.376	584
1989	1908	0.330	0.345	610
1990	1909	0.314	0.330	606
1991	1910	0.329	0.343	544
1992	1911	0.322	0.329	502
1993	1912	0.354	0.346	535
1994	1913	0.360	0.354	459
1995	1914	0.336	0.338	487
1996	1915	0.320	0.322	465
1997	1916	0.316	0.315	315
1998	1917	0.313	0.313	372
1999	1918	0.329	0.326	273
career AVG		0.349	0.359	8955

Table 7: Wade Boggs' hypothetical career that started in 1901

In the future, we could extend our model to multivariate Full House Model using multivariate order statistics and multivariate empirical distribution, and make statistical inference on cross-system multi-dimensional components. It would be helpful to compare batter's talent by using several batting statistics together instead of using BA or Hits separately.

In this model we assume the components in different systems are mutually exclusive and independent and this assumption may fail in some scenarios. The extension of this work is accounting for the time-variation between the systems [Spearing et al., 2021], which would be helpful on predicting the seasons have not yet opened.

References

Baseball-Reference. Baseball-reference archive. URL `https://www.baseball-reference.com/`.

Benjamin S Baumer, Shane T Jensen, and Gregory 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.

David Berri and Martin Schmidt. *Stumbling on Wins (Bonus Content Edition): Two Economists Expose the Pitfalls on the Road to Victory in Professional Sports, Portable Documents*. Pearson Education, 2010.

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

Enrique Castillo. *Extreme value theory in engineering*. Elsevier, 2012.

Arnold LM Dekkers, John HJ Einmahl, and Laurens De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

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

FanGraphs. Fangraphs baseball archive. URL `https://www.fangraphs.com/`.

Stephen Jay Gould. *Full house*. Harvard University Press, 2011.

David C Hoaglin et al. Letter values: A set of selected order statistics. *Understanding robust and exploratory data analysis*, pages 33–57, 1983.

William Kaczynski, L Leemis, N Loehr, and J McQueston. Nonparametric random variate generation using a piecewise-linear cumulative distribution function. *Communications in Statistics-Simulation and Computation*, 41(4):449–468, 2012.

Paul H Kvam. Comparing hall of fame baseball players using most valuable player ranks. *Journal of Quantitative Analysis in Sports*, 7(3), 2011.

Domine MW Leenaerts and Wim M Van Bokhoven. *Piecewise Linear Modeling and Analysis*. Kluwer Academic Publishers, 1998.

Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5): 323–351, 2005.

Daniel Okrent, Harris Lewine, and David Nemec. *The Ultimate Baseball Book: The Classic Illustrated History of the World's Greatest Game*. Houghton Mifflin Harcourt, 2000.

Alexander M Petersen and Orion Penner. Renormalizing individual performance metrics for cultural heritage management of sports records. *Chaos, Solitons & Fractals*, 136:109821, 2020.

Alexander M Petersen, Orion Penner, and H Eugene Stanley. Methods for detrending success metrics to account for inflationary and deflationary factors. *The European Physical Journal B*, 79(1):67–78, 2011.

Michael J Schell. *Baseball's all-time best hitters*. Princeton University Press, 2013.

Michael J Schell. *Baseball's all-time best sluggers*. Princeton University Press, 2016.

Martin B Schmidt and David J Berri. Concentration of playing talent: evolution in major league baseball. *Journal of Sports Economics*, 6(4):412–419, 2005.

FW Scholz. Nonparametric tail extrapolation. *White paper from Boeing Information & Support Services) http://www. stat. washington. edu/fritz/Reports/ISSTECH-95-014. pdf*, 1995.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Harry Spearing, Jonathan Tawn, David Irons, Tim Paulden, and Grace Bennett. Ranking, and other properties, of elite swimmers using extreme value theory. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):368–395, 2021.

Michael L Stein. A parametric model for distributions with flexible behavior in both tails. *Environmetrics*, page e2658, 2020.

6 Appendix

6.1 Some results from Scholz [1995]

Based on the extreme value theory, see Castillo [2012] and induction from Scholz [1995], we would expect the $Y_1, \ldots, Y_k$ to show a approximately linear pattern with the log odds ratio of $p_1, \ldots, p_k$ when the extreme-value index c is equal to 0, and the $Y_1, \ldots, Y_k$ to show a approximately linear pattern with $f_c(p_1), \ldots, f_c(p_k)$ when c is not equal to 0, where k is the data values in the tail of the distribution to use in the linear approximation, $p_i = p_{i,\gamma,n}$, and

$$f_c(p_i) = \frac{(-n \ln(p_i))^{-c} - 1}{c}$$

The extreme-value index c can be estimated from the data directly by using the moment estimate proposed by Dekkers et al. [1989], which is

$$\widehat{c}_k = M_{1,k} + 1 - .5\left(1 - \frac{M_{1,k}^2}{M_{2,k}}\right)^{-1}$$

with

$$M_{1,k} = \frac{1}{k-1}\sum_{i=1}^{k-1} \log\left(\tilde{Y}_i/\tilde{Y}_k\right) \quad \text{and} \quad M_{2,k} = \frac{1}{k-1}\sum_{i=1}^{k-1}\left[\log\left(\tilde{Y}_i/\tilde{Y}_k\right)\right]^2,$$

where $\tilde{Y}_i = Y_i - \text{median}(X_1, \ldots, X_n)$

An issue, that has not yet been addressed, is the number k of data values in the tail of the distribution to use in the linear approximation step. Scholz [1995] shows that the k can be found in $[K_1, K_2]$ and satisfies $T_k \in \left[\kappa_k t_{k-2,1/\kappa}(.25), \quad \kappa_k t_{k-2,1/\kappa}(.75)\right]$, where $K_1 = \max(6, \lfloor 1.3\sqrt{n} \rfloor)$, $K_2 = 2\lfloor \log_{10}(n)\sqrt{n} \rfloor$, $T_k = \kappa_k t_{k-2,1/\kappa_k}$, and κ_k is the standard deviation of the slope parameter in the generalized linear model. Then we choose the value k so that the plotted points become most linear using linear model and linear quadratic model, which based on the goodness of fit to pick up the best value of k. Also we would expect the larger k so that we can minimize the influences of the influential points.

Consider the one term Taylor expansion on the extreme p-quantile of distribution F and induction from Scholz [1995], the generalized least square model is well-established.

6.2 Some theoretical properties

Proposition 6.1. *Let $\widetilde{F}_Y(t)$ be defined as in (1) and let $\widehat{F}_Y(t)$ be the empirical distribution function. Then,*

$$\sup_{t\in\mathbb{R}} \left| \tilde{F}_Y(t) - \widehat{F}_Y(t) \right| \leq \frac{1}{n}$$

Proof. We will prove this result in cases. First, when $t \leq \widetilde{Y}_{i,(1)}$ or $t \geq \widetilde{Y}_{i,(n+1)}$ we have that $|\ \widetilde{F}_Y(t) - \widehat{F}_Y(t)\ | = 0$. For any $j = 1, \ldots, n$ and $\tilde{Y}_{i,(j)} \leq t < Y_{i,(j)}$, we have

$$\left| \widehat{F}_Y(t) - \widetilde{F}_Y(t) \right| = \left| \frac{j-1}{n} - \frac{j-1+\left(t-\widetilde{Y}_{i,(j)}\right) / \left(\tilde{Y}_{i,(j+1)} - \tilde{Y}_{i,(j)}\right)}{n} \right| \leq \frac{1}{n}$$

For any $j = 1, \ldots, n$ and $Y_{i,(j)} < t < \tilde{Y}_{i,(j+1)}$, we have

$$\left| \widehat{F}_Y(t) - \widetilde{F}_Y(t) \right| = \left| \frac{j}{n} - \frac{j-1+\left(t-\tilde{Y}_{i,(j)}\right) / \left(\tilde{Y}_{i,(j+1)} - \tilde{Y}_{i,(j)}\right)}{n} \right| \leq \frac{1}{n}$$

Our conclusion follows. □

Corollary 6.1.1. *Let $\widetilde{F}_Y(t)$ be defined as in (1) and let $\widehat{F}_Y(t)$ be the empirical distribution function. Then,*

$$\sup_{t\in\mathbb{R}} \left| \widetilde{F}_Y(t) - F_Y(t) \right| \xrightarrow{a.s} 0$$

Proof. We have, $\sup_{t\in\mathbb{R}} \left| \widetilde{F}_Y(t) - F_Y(t) \right| \leq \sup_{t\in\mathbb{R}} \left| \widetilde{F}_Y(t) - \widehat{F}_Y(t) \right| + \sup_{t\in\mathbb{R}} \left| \widehat{F}_Y(t) - F_Y(t) \right|$. The conclusion follows from the Glivenko-Cantelli Theorem and Proposition 2.1 . □

6.3 Era-Adjusted Hypothetical Career Statistics Based on Full House Model

	Name	rookie year	BA
1	Tony Gwynn	1982	0.325
2	Rod Carew	1967	0.309
3	Ichiro Suzuki	2001	0.306
4	**Ty Cobb**	1906	0.303
5	Jose Altuve	2011	0.302
6	Mike Trout	2011	0.300
7	Miguel Cabrera	2003	0.297
8	Buster Posey	2009	0.297
9	Vladimir Guerrero	1996	0.295
10	Wade Boggs	1982	0.295
11	Robinson Cano	2005	0.294
12	Mike Piazza	1992	0.293
13	Henry Aaron	1954	0.293
14	**Shoeless Joe Jackson**	1909	0.293
15	Roberto Clemente	1955	0.292
16	Matty Alou	1961	0.291
17	Tony Oliva	1962	0.291
18	Daniel Murphy	2008	0.291
19	Derek Jeter	1995	0.291
20	Joe Mauer	2004	0.290
21	Christian Yelich	2013	0.290
22	Pete Rose	1963	0.290
23	Joey Votto	2007	0.290
24	Willie Mays	1951	0.289
25	José Abreu	2014	0.289

Table 8: Top 25 MLB batters with era-adjusted hypothetical career BA. The second column is the name of the players. The third column is the year when they start their career. The fourth column is the cumulative estimated BA.

	Name	rookie year	Home Runs
1	Henry Aaron	1954	796
2	Albert Pujols	2001	793
3	Barry Bonds	1986	780
4	**Babe Ruth**	1914	715
5	Alex Rodriguez	1994	691
6	Willie Mays	1951	680
7	Mike Schmidt	1972	655
8	Frank Robinson	1956	630
9	Ken Griffey Jr.	1989	618
10	David Ortiz	1997	615
11	Eddie Mathews	1952	598
12	Willie Stargell	1963	593
13	Miguel Cabrera	2003	589
14	Jim Thome	1991	585
15	Manny Ramirez	1993	583
16	Mickey Mantle	1951	582
17	Reggie Jackson	1967	570
18	Mark McGwire	1986	558
19	Frank Thomas	1990	553
20	Rafael Palmeiro	1987	551
21	Sammy Sosa	1992	540
22	Harmon Killebrew	1956	534
23	Nelson Cruz	2005	526
24	Adrian Beltre	1998	522
25	Fred McGriff	1986	518

Table 9: Top 25 MLB batters with era-adjusted hypothetical career HR. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career home runs.

	name	rookie year	bWAR
1	Barry Bonds	1986	131.03
2	Willie Mays	1951	121.71
3	Henry Aaron	1954	115.79
4	Alex Rodriguez	1994	102.65
5	**Babe Ruth**	1914	97.80
6	Rickey Henderson	1980	92.69
7	Mike Schmidt	1972	92.18
8	**Stan Musial**	1941	89.96
9	Albert Pujols	2001	89.30
10	Frank Robinson	1956	87.63
11	Adrian Beltre	1998	85.88
12	**Ted Williams**	1939	83.94
13	Cal Ripken Jr.	1981	82.35
14	**Ty Cobb**	1906	80.12
15	Chipper Jones	1993	79.21
16	Eddie Mathews	1952	76.67
17	Joe Morgan	1963	75.82
18	Wade Boggs	1982	75.73
19	Roberto Clemente	1955	75.10
20	Mickey Mantle	1951	75.10
21	**Tris Speaker**	1907	73.86
22	**Lou Gehrig**	1923	72.63
23	**Rogers Hornsby**	1915	72.63
24	Jeff Bagwell	1991	71.82
25	Ken Griffey Jr.	1989	71.31

Table 10: Top 25 MLB batters with era-adjusted hypothetical career bWAR. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career bWAR.

	name	rookie year	fWAR
1	Barry Bonds	1986	131.44
2	Willie Mays	1951	117.73
3	Henry Aaron	1954	114.02
4	Alex Rodriguez	1994	98.25
5	**Babe Ruth**	1914	96.97
6	Mike Schmidt	1972	92.31
7	Rickey Henderson	1980	89.58
8	**Ted Williams**	1939	88.43
9	**Stan Musial**	1941	85.63
10	Frank Robinson	1956	84.02
11	Cal Ripken Jr.	1981	81.97
12	**Ty Cobb**	1906	80.10
13	Mickey Mantle	1951	79.88
14	Albert Pujols	2001	78.37
15	Adrian Beltre	1998	77.80
16	Chipper Jones	1993	76.52
17	Eddie Mathews	1952	76.27
18	Joe Morgan	1963	75.39
19	Wade Boggs	1982	73.54
20	**Lou Gehrig**	1923	72.40
21	Jeff Bagwell	1991	71.49
22	**Rogers Hornsby**	1915	71.41
23	**Honus Wagner**	1898	70.74
24	Derek Jeter	1995	70.10
25	**Tris Speaker**	1907	70.05

Table 11: Top 25 MLB batters with era-adjusted hypothetical career fWAR. The second column is the name of the players. The third column is the year when they start their career. The fourth column is the career fWAR.

	Name	rookie year	ERA
1	Mariano Rivera	1995	1.676
2	Clayton Kershaw	2008	2.349
3	Pedro Martinez	1992	2.392
4	Jacob deGrom	2014	2.467
5	Brandon Webb	2003	2.509
6	Chris Sale	2010	2.844
7	Greg Maddux	1986	2.909
8	Hoyt Wilhelm	1952	2.921
9	Roy Halladay	1998	2.922
10	Corey Kluber	2012	2.925
11	Johan Santana	2000	2.958
12	**Al Spalding**	1871	2.988
13	Roger Clemens	1984	3.002
14	Max Scherzer	2008	3.036
15	**Tommy Bond**	1874	3.039
16	Sandy Koufax	1955	3.046
17	Roy Oswalt	2001	3.059
18	Justin Verlander	2005	3.062
19	Stephen Strasburg	2010	3.078
20	Gerrit Cole	2013	3.121
21	Randy Johnson	1988	3.123
22	**Old Hoss Radbourn**	1881	3.147
23	Cole Hamels	2006	3.176
24	**Jim McCormick**	1878	3.181
25	Juan Marichal	1960	3.184

Table 12: Top 25 MLB pitchers with era-adjusted hypothetical career ERA. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career ERA.

	Name	rookie year	SO
1	Nolan Ryan	1968	5218
2	Roger Clemens	1984	4831
3	Randy Johnson	1988	4671
4	Steve Carlton	1965	4157
5	Greg Maddux	1986	3872
6	Bert Blyleven	1970	3789
7	Tom Seaver	1967	3704
8	**Walter Johnson**	1907	3582
9	Don Sutton	1966	3537
10	Gaylord Perry	1962	3489
11	Phil Niekro	1964	3253
12	Fergie Jenkins	1965	3225
13	CC Sabathia	2001	3200
14	John Smoltz	1988	3175
15	Max Scherzer	2008	3104
16	Curt Schilling	1989	3092
17	Justin Verlander	2005	3086
18	Pedro Martinez	1992	3053
19	Zack Greinke	2004	2916
20	Bob Gibson	1959	2909
21	Mike Mussina	1991	2888
22	Chuck Finley	1986	2788
23	Clayton Kershaw	2008	2735
24	David Cone	1986	2714
25	Jim Bunning	1955	2712

Table 13: Top 25 MLB pitchers with era-adjusted hypothetical career SO. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career strikeouts.

	name	rookie year	bWAR
1	Roger Clemens	1984	125.77
2	Greg Maddux	1986	98.14
3	Randy Johnson	1988	93.82
4	Tom Seaver	1967	86.47
5	**Walter Johnson**	1907	82.47
6	**Cy Young**	1890	80.59
7	**Lefty Grove**	1925	79.82
8	Bert Blyleven	1970	79.14
9	Mike Mussina	1991	76.18
10	Pedro Martinez	1992	75.78
11	Justin Verlander	2005	74.88
12	Phil Niekro	1964	74.37
13	Clayton Kershaw	2008	71.66
14	Curt Schilling	1989	70.25
15	Gaylord Perry	1962	69.39
16	Zack Greinke	2004	69.19
17	Tom Glavine	1987	68.52
18	Max Scherzer	2008	68.15
19	Roy Halladay	1998	67.44
20	Steve Carlton	1965	67.06
21	Bob Gibson	1959	65.90
22	**Warren Spahn**	1942	65.46
23	CC Sabathia	2001	61.53
24	Nolan Ryan	1968	61.52
25	Fergie Jenkins	1965	61.46

Table 14: Top 25 MLB pitchers with era-adjusted hypothetical career bWAR. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career bWAR.

	name	rookie year	fWAR
1	Roger Clemens	1984	131.96
2	Greg Maddux	1986	108.43
3	Randy Johnson	1988	97.81
4	Nolan Ryan	1968	90.42
5	Steve Carlton	1965	82.53
6	Bert Blyleven	1970	82.15
7	Gaylord Perry	1962	81.26
8	**Cy Young**	1890	80.13
9	**Walter Johnson**	1907	78.90
10	**Lefty Grove**	1925	78.05
11	Pedro Martinez	1992	76.58
12	Clayton Kershaw	2008	75.30
13	Justin Verlander	2005	74.85
14	Mike Mussina	1991	72.54
15	Tom Seaver	1967	71.48
16	Curt Schilling	1989	68.53
17	Bob Gibson	1959	68.05
18	John Smoltz	1988	66.57
19	Max Scherzer	2008	66.11
20	Zack Greinke	2004	65.60
21	Kevin Brown	1986	65.32
22	Don Sutton	1966	63.36
23	CC Sabathia	2001	61.94
24	Roy Halladay	1998	61.33
25	Andy Pettitte	1995	60.53

Table 15: Top 25 MLB pitchers with era-adjusted hypothetical career fWAR. The second column is the name of the players. The third column is the year when they start their career. The fourth column is career fWAR.

We also rank MLB players, for both batters and pitchers, by their era-adjusted hypothetical career bWAR and fWAR under the scenario that every player only played in 2021.

6.4 Talent Adjustment Based on Rotation Size

For example, we assume that a pitcher played in 1894 season and the rotation size in that season is 4, which indicates that the average number of starting pitchers in each team is 4. Since we want to map this player's statistics to 2021 season, which has average 5 starting pitchers in each team, more pitchers will be taken into account for the starting pitchers. The pitchers that are chosen as new starting pitchers in the 2021 season performs bad in 1894 season since they are not in a rotation size. Then we would add some talents to them. Also compared with the pitchers mentioned above, the pitchers who are already in a rotation size in 1894 season are also credited with some talents since they are all in the same distribution and compete with each others. The reasonable talent that credited to these pitchers is the talent of the pitchers who are in the 5th rotation in 2021.

6.5 Latent Distribution Sensitivity Analysis

We consider two possible latent distribution of the underlying traits, folded normal distribution and standard normal distribution, and apply our Full House Model to BA data. The Table 16 is the top 25 MLB players with era-adjusted hypothetical career BA using folded normal distribution and standard normal distribution as latent distribution of the underlying traits

	folded normal			normal		
	name	rookie year	BA	name	rookie year	BA
1	Tony Gwynn	1982	0.326	Tony Gwynn	1982	0.318
2	Rod Carew	1967	0.307	Rod Carew	1967	0.296
3	Ichiro Suzuki	2001	0.306	Ichiro Suzuki	2001	0.293
4	Ty Cobb	1905	0.303	Ty Cobb	1905	0.292
5	Jose Altuve	2011	0.302	Jose Altuve	2011	0.291
6	Buster Posey	2009	0.299	Buster Posey	2009	0.287
7	Mike Trout	2011	0.297	Mike Trout	2011	0.285
8	Miguel Cabrera	2003	0.296	Wade Boggs	1982	0.284
9	Vladimir Guerrero	1996	0.296	Vladimir Guerrero	1996	0.284
10	Wade Boggs	1982	0.295	Miguel Cabrera	2003	0.283
11	Edgar Martinez	1987	0.294	Roberto Clemente	1955	0.282
12	Mike Piazza	1992	0.294	Edgar Martinez	1987	0.282
13	Robinson Cano	2005	0.293	Matty Alou	1960	0.281
14	Derek Jeter	1995	0.292	Robinson Cano	2005	0.281
15	Roberto Clemente	1955	0.292	Derek Jeter	1995	0.279
16	Matty Alou	1960	0.292	Mike Piazza	1992	0.279
17	Shoeless Joe Jackson	1908	0.292	Henry Aaron	1954	0.278
18	Joe Mauer	2004	0.291	Willie Mays	1951	0.278
19	Willie Mays	1951	0.291	Joe Mauer	2004	0.277
20	Manny Mota	1962	0.290	Shoeless Joe Jackson	1908	0.277
21	Christian Yelich	2013	0.290	Manny Mota	1962	0.277
22	Henry Aaron	1954	0.290	Christian Yelich	2013	0.276
23	Joey Votto	2007	0.290	Joey Votto	2007	0.276
24	José Abreu	2014	0.289	José Abreu	2014	0.276
25	David Wright	2004	0.289	David Wright	2004	0.276

Table 16: Top 25 MLB hitters with era-adjusted hypothetical career BA using folded normal distribution and standard normal distribution as latent distribution of the underlying traits.

Compared the result using Pareto distribution as latent distribution of the underlying traits from Table 7, the rankings on the list have changed slightly but the hitters on the list keep the same. Also, the estimated BA is almost the same. Therefore, the result is not sensitive to the latent distribution.