Full House Methodology for estimating components of describable systems by accounting for context

Daniel J. Eck

June 14, 2020

Abstract

We motivate a new viewpoint on statistical inference on components of systems when placed in the broader context of the population that these components arise from. Particular emphasis is placed on a class of techniques for inferring the values of hidden inputs in systems with estimable components. Each estimable component in the system is thought to be a function of hidden traits which are themselves realizations from a parametric probability distribution. With this structure we can estimate the values of the hidden traits after estimating the distribution of the system components. Estimation of the distribution of system components can be done with no parametric assumptions when necessary. This framework allows for one to infer how individuals will perform in different states of a describable system that remains stable but has a changing hidden inputs distribution. Our methodology is named in homage to Stephen J. Gould's book Full House: The Spread of Excellence from Plato to Darwin [Gould, 1996]. We demonstrate our method on historical batting averages in baseball, a compelling example of how seemingly paradoxical conclusions arise when the full house of system variability is taken into account.

1 Setting

We will suppose that we have a sample of n observable components $Y_{i_1}, \ldots, Y_{i_n}$ arising from a population with $n_{\text{pop}} \geq n$ members. We will additionally suppose that each $Y_j = f_j(X_j)$, $j = 1, \ldots, n_{\text{pop}}$, where X_j is some underlying hidden trait arising from a distribution F_X, f_j is an indvidual specific function connecting the hidden trait and the outcome, and $g_j(X_j)$ is the system inclusion mechanism, $g_j(X_j) = 1$ indicates that subject j is included in the system and $g_j(X_j) = 0$ indicates that subject j is not included in the system. The inclusion mechanism can be probabilistic or deterministic. This framework encapsulates simple random sampling where $X_j = Y_j$, f_j is the identity function, and $g_j(X_j) = 1$ with probability n/n_{pop}.

2 The spread of excellence model

Let n_{pop} be the size of the population and let $n \leq n_{\text{pop}}$ be the number of components in the system. Suppose that we have a sample of outputs $Y_1, \ldots, Y_n \overset{iid}{\sim} F_Y$ measured on the components in the system. Suppose that all individuals in the population have an underlying aptitude to participate in the system. Denote this aptitude by $X_1, \ldots, X_{n_{\text{pop}}} \overset{iid}{\sim} F_X$. Suppose that the system selects the highest aptitudes, so that $g(X_j) = 1(X_j \geq X_{n_{\text{pop}}-n+1})$, where we observe $Y_{(j)} = f_j(X_{(n_{\text{pop}}-n+j)})$. We motivate this methodological approach through the goal of inferring the values of $X_{(n_{\text{pop}}-n+j)}$, $j = 1, \ldots, n$ from the observed values of $Y_1, \ldots, Y_n$.

In this setup, we will assume that F_X is completely known up to p_X unknown parameters $\theta \in \mathbb{R}^{p_X}$ and that F_Y is known, known up to p_Y unknown parameters $\psi \in \mathbb{R}^{p_Y}$, or unknown and is estimated empirically.

2.1 Parametric case

Let $F_Y(\cdot|\theta)$ be a parametric CDF with parameters $\theta \in \mathbb{R}^{p_Y}$. We can estimate θ with $\hat{\theta}$ and plug the estimator into the CDF $F_Y(\cdot|\hat{\theta})$. The distribution function $F_{Y_{(j)}}(y|\hat{\theta})$ is

$$F_{Y_{(j)}}(y|\hat{\theta}) = \sum_{k=j}^{n} \binom{n}{k} \left(F_Y(y|\hat{\theta})\right)^k \left(1 - F_Y(y|\hat{\theta})\right)^{n-k}.$$

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

$$\begin{aligned} F_Y(Y_{(j)}|\theta) &\sim U_{(j)}, & F_Y(Y_{(j)}|\hat{\theta}) &\approx U_{(j)}, \\ F_{Y_{(j)}}(Y_{(j)}|\theta) &\sim U_j, & F_{Y_{(j)}}(Y_{(j)}|\hat{\theta}) &\approx U_j, \end{aligned}$$

where $U_j \sim U(0,1)$ and $U_{(j)} \sim \text{Beta}(j, n+1-j)$ and the approximation in the right hand side depends upon the estimator $\hat{\theta}$ and the sample size. We now connect the order statistics to the underlying distribution that comes from a population with $n_{\text{pop}} \geq n$ observations when F_X is known. This connection is established with the relation

$$F^{-1}_{X_{(n_{\text{pop}}-n+j)}}\left(F_{Y_{(j)}}(Y_{(j)}|\theta)\right) \sim F^{-1}_{X_{(n_{\text{pop}}-n+j)}}(U_j) \sim X_{(n_{\text{pop}}-n+j)}.$$

We estimate the above with

$$F^{-1}_{X_{(n_{\text{pop}}-n+j)}}\left(F_{Y_{(j)}}(Y_{(j)}|\hat{\theta})\right) \approx F^{-1}_{X_{(n_{\text{pop}}-n+j)}}(U_j) \sim X_{(n_{\text{pop}}-n+j)}.$$

2.2 Nonarametric case

[This section needs work; the interpolated CDF motivated here does not have desirable empirical properties. The interpolated CDF does well when the goal is to extract X scores, but it does not do well when mapping people from different time periods into a common time period.]

In the nonparametric setting we motivate an 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 $Y_{(n)}$. We therefore consider an interpolated version of the empirical CDF $\widetilde{F}_Y$ to alleviate this problem. We construct $\widetilde{F}_Y$ in the following manner: We first construct surrogate sample points $\widetilde{Y}_{(1)}, \ldots, \widetilde{Y}_{(n+1)}$ as,

$$\begin{aligned} \widetilde{Y}_{(1)} &= Y_{(1)} - 1/(Y_{(2)} - Y_{(1)}), \\ \widetilde{Y}_{(j)} &= (Y_{(j)} + Y_{(j-1)})/2, \qquad j = 2, \ldots, n, \\ \widetilde{Y}_{(n+1)} &= Y_{(n)} + 1/(Y_{(n)} - Y_{(n-1)}). \end{aligned}$$

With this construction, we build $\widetilde{F}_Y$ as

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

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, $\widetilde{F}_Y(Y_{(1)})$ and $\widetilde{F}_Y(Y_{(n)})$ provide reasonable estimates for the cumulative probability at $Y_{(1)}$ and $Y_{(n)}$ by considering their respective discrepancy from $Y_{(2)}$ and $Y_{(n-1)}$. Notice that $\widetilde{F}_Y(t)$ is close to $\widehat{F}_Y(t)$. We formalize this statement below.

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

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

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

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

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

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

Our conclusion follows. □

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

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

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

Proof. We have, $\sup_{t\in\mathbb{R}} |\widetilde{F}_Y(t) - F_Y(t)| \leq \sup_{t\in\mathbb{R}} |\widetilde{F}_Y(t) - \widehat{F}_Y(t)| + \sup_{t\in\mathbb{R}} |\widehat{F}_Y(t) - F_Y(t)|$. The conclusion follows from the Glivenko-Cantelli Theorem and Proposition 1. □

More properties of $\widehat{F}_Y$ are provided in the Appendix. Corollary 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,

$$\widetilde{F}_Y(Y_{(j)}) \approx U_{(j)}, \qquad \widetilde{F}_{Y_{(j)}}(Y_{(j)}) \approx U_j,$$

where $U_j \sim U(0,1)$ and $U_{(j)} \sim \text{Beta}(j, n+1-j)$ 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 $n_{\text{pop}} \geq n$ observations when F_X is known. We estimate the hidden trait value by with

$$F^{-1}_{X_{(n_{\text{pop}}-n+j)}} \left(\widetilde{F}_{Y_{(j)}}(Y_{(j)}) \right) \approx F^{-1}_{X_{(n_{\text{pop}}-n+j)}} (U_j) \sim X_{(n_{\text{pop}}-n+j)}.$$

3 Connection to the aster models

Aster models were originally developed as a statistically valid model for Darwinian fitness by accounting for the fitness components (life cycle) of the system under study and by employing appropriate probability models for each fitness component [Geyer et al., 2007, Shaw et al., 2008]. In the absence of aster models the individual fitness components are estimated separately, and therefore the interplay of fitness components in their contributions to lifetime fitness and how selection operates over the entire life cycle cannot be quantified. Aster models belong to the full house methodology framework. In an aster analysis we can specify lifetime fitness as $Y_j = f(X_j, \beta)$ for all $j = 1, \ldots, n_{\text{pop}}$, where $n = n_{\text{pop}}$, X_j is a vector of observable traits and fitness components, and β is a vector of regression parameters linking the observable traits to fitness and modeling parameters associated with the probability models for each fitness component. Expected Darwinian fitness is estimated for every individual through invariance of maximum likelihood estimation for β.

Eck et al. [2015] provided an aster analysis of total egg counts (fitness) for a field population of *Manduca sexta*. One conclusion from this analysis was that estimated fitness surfaces revealed strong and significant directional selection favoring both larger adult size (via effects on egg counts) and more rapid rates of early larval development (via effects on larval survival). The incorporation of timing of reproduction and its influence on population growth rate resulted in larger values for size in early larval development at which fitness is maximized. The original analysis of this field population comes from Kingsolver et al. [2012] who estimated selection via survival, reproduction, and other components separately. In the original analysis, total egg counts is estimated with multiple regression after conditioning on survival. In this setting there are $n < n_{\text{pop}}$ females with $Y'_{i_j} = f'(X'_{i_j}, \beta')$ total eggs. Females had to survive to their reproduction stage for inclusion in this regression, therefore $g_j(X_j) = 1$ if female j survived to reproduction and $g_j(X_j) = 0$ otherwise. This analysis cannot quantify how selection on survival effected reproduction. By accounting for the life cycles of the full population into the analysis, this aster analysis of *M. sexta* could then illustrate how the interplay of different components of fitness can influence selection on size and development time.

4 Connection to causal inference

Discuss selection bias, and read Imbens and Menzel [2018].

5 Examples

5.1 Era adjustment for batting averages in baseball

[see R script]

Comparing the achievements of baseball players across eras has resulted in endless debates among family members, friends, participants in social media platforms, network personalities, and scientists [Gould, 1996, Berry et al., 1999, Schell, 2005, Petersen et al., 2011, Eck, 2020]. Of all the possible across-era comparisons to be made, the comparison of baseball players' batting averages has a lively discussion in the scientific literature [Gould, 1996, Berry et al., 1999, Schell, 2005]. In this example, the spread of excellence model is used to construct an era-neutral environment which allows for comparisons of the batting averages of baseball players from fundamentally different eras.

We compare the results and philosophies of approach to those of Gould [1996], Berry et al. [1999], Schell [2005] and Petersen et al. [2011].

The spread of excellence model is philosophically rooted in Part 3 of Gould [1996] and Eck [2020]. Gould made the paradoxical observation that the diminishing rate of extraordinary individual batting averages signalled an overall increase in the hitting ability of the typical major league baseball player. The rationale for this finding is fourfold: 1) the distribution of annual batting averages among full time players forms a stable system that historically follows a normal distribution with mean roughly equal to .260; 2) the available talent pool of eligible major league players becomes deeper and richer as time continues; 3) baseball players have gotten bigger; 4) records in other sports with absolute standards have historically improved. Schell [2005] makes a similar observation. For our comparisons we will assume that all players grow up in the same era-neutral environment. Under this hypothetical only the stability of the annual batting average distribution and the size and richness of the underlying talent pool are relevant.

Batting averages in baseball have historically followed a normal distribution. We suppose that underlying talent follows a Pareto(α) distribution. In this example, we can take $f_j = \Phi^{-1} \circ F_{X_{(n_{\text{pop}}-n+j)}}$ where Φ is the CDF of the normal distribution and $F_{X_{(n_{\text{pop}}-n+j)}}$ is the CDF of the order statistics of the latent distribution.

Appendix

Properties of $\widetilde{F}_Y$

In this section we study mathematical properties of the interpolated empirical distribution function $\widetilde{F}_Y$. First, we expand on some classical empirical results for $\widetilde{F}_Y$. In particular, $\widetilde{F}_Y$ possess a Komlós-Major-Tusnády (KMT) embedding [Komlós et al., 1975] and a Dvoretzky-Kiefer–Wolfowitz (DFW) inequality bound of the tail probability [Dvoretzky et al., 1956, Massart, 1990].

Proposition 2. *Let F_Y be a distribution function, $\widetilde{F}_Y(t)$ be defined as in* (1)*, and $\widehat{F}_Y(t)$ be the empirical distribution function. Let $G_{F,n} = B_n(F_Y(t))$ be a Gaussian process where $\{B_n(t), 0 \le t \le 1\}$ is a sequence of Brownian bridges. Then,*

$$\limsup_{n\to\infty} \frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}(\widetilde{F}_Y - F_Y) - G_{F,n} \right\|_\infty < \infty, \qquad a.s.$$

Proof. The result follows from a simple derivation,

$$\begin{aligned}
\frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}(\widetilde{F}_Y - F_Y) - G_{F,n} \right\|_\infty &= \frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}((\widetilde{F}_Y - \widehat{F}_Y) - (\widehat{F}_Y - F_Y) - G_{F,n}) \right\|_\infty \\
&\le \frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}(\widehat{F}_Y - F_Y) - G_{F,n} \right\|_\infty + \frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}(\widetilde{F}_Y - \widehat{F}_Y) \right\|_\infty \\
&\le \frac{\sqrt{n}}{\log(n)} \left\| \sqrt{n}(\widehat{F}_Y - F_Y) - G_{F,n} \right\|_\infty + \frac{1}{\log(n)}
\end{aligned}$$

where the last line follows from Proposition 1. Our conclusion follows from Komlós et al. [1975]. □

Proposition 3. *Let F_Y be a distribution function, $\widetilde{F}_Y(t)$ be defined as in* (1)*, and $\widehat{F}_Y(t)$ be the empirical distribution function. Then, for any $C > 0$,*

$$\mathbb{P}\left(\sqrt{n} \|\widetilde{F}_Y - F_Y\|_\infty > \sqrt{C \log(n)/2} \right) = O\left(n^{-C}\right).$$

Proof. Proposition 1 gives

$$\mathbb{P}\left(\sqrt{n}\|\widetilde{F}_Y - F_Y\|_\infty > \sqrt{C\log(n)/2}\right) \leq \mathbb{P}\left(\sqrt{n}\|\widehat{F}_Y - F_Y\|_\infty > \sqrt{C\log(n)/2} - 1/\log(n)\right).$$

Massart [1990] gives

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}\|\widehat{F}_Y - F_Y\|_\infty > \sqrt{C\log(n)/2} - 1/\log(n)\right) &\leq 2\exp\left(-2(\sqrt{C\log(n)/2} - 1/\log(n))^2\right) \\
&= 2\exp\left(-C\log(n) + \frac{4\sqrt{C/2}}{\sqrt{\log(n)}} - \frac{2}{\log(n)^2}\right) \\
&= 2n^{-C}O\left(1 + + \frac{4\sqrt{C/2}}{\sqrt{\log(n)}} - \frac{2}{\log(n)^2}\right) \\
&= O\left(n^{-C}\right).
\end{aligned}$$

Our conclusion follows. □

References

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.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

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

Daniel J Eck, Ruth G Shaw, Charles J Geyer, and Joel G Kingsolver. An integrated analysis of phenotypic selection on insect body size and development time. *Evolution*, 69(9):2525–2532, 2015.

Charles J Geyer, Stuart Wagenius, and Ruth G Shaw. Aster models for life history analysis. *Biometrika*, 94(2):415–426, 2007.

Stephen Jay Gould. *Full house: The Spread of Excellence from Plato to Darwin*. Harvard University Press, 1996.

Guido Imbens and Konrad Menzel. A causal bootstrap. Technical report, National Bureau of Economic Research, 2018.

Joel G Kingsolver, Sarah E Diamond, Sarah A Seiter, and Jessica K Higgins. Direct and indirect phenotypic selection on developmental trajectories in manduca sexta. *Functional Ecology*, 26(3): 598–607, 2012.

János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32 (1-2):111–131, 1975.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

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: How statistics can level the playing field*. Princeton University Press, 2005.

Ruth G Shaw, Charles J Geyer, Stuart Wagenius, Helen H Hangelbroek, and Julie R Etterson. Unifying life-history analyses for inference of fitness and population growth. *The American Naturalist*, 172(1):E35–E47, 2008.