Shrinkage Estimators

Vasco Villas-Boas

September 5, 2022

Shrinkage Estimators

Introduction with a Paradox

Suppose you're interested in simultaneously estimating the proportion of attendees at Wimbledon that wear a hat each day (μ_{hat}), the fraction of red cars driving around Columbus Circle in Manhattan every day (μ_{car}), and the 2022 batting average of Aaron Judge of the New York Yankees (μ_{ba}). You will report your estimates $\hat{\mu}_{\text{hat}}$, $\hat{\mu}_{\text{car}}$, and $\hat{\mu}_{\text{ba}}$.

You observe N iid people at Wimbledon: $\{H_i : 1 \leq i \leq N\}$, where $H_i := 1_{\text{person i wears a hat}}$, you observe N iid cars driving through Columbus Circle: $\{C_i : 1 \leq i \leq N\}$, where $C_i := 1_{\text{car i is red}}$, and you observe N iid at-bats of Aaron Judge: $\{B_i : 1 \leq i \leq N\}$, where $B_i := 1_{\text{at-bat i is a hit}}$. For simplicity, suppose that $\text{Var}(H_i) = \text{Var}(C_i) = \text{Var}(B_i) = \sigma^2$ for all i [1]. Also define $h := \frac{\sum_{i=1}^N H_i}{N}$, $c := \frac{\sum_{i=1}^N H_i}{N}$, and $b := \frac{\sum_{i=1}^N A_i}{N}$.

The most intuitive report would be to give

$$(\hat{\mu}_{\text{hat}}, \hat{\mu}_{\text{car}}, \hat{\mu}_{\text{ba}}) = (h, c, b)$$

That is, estimate the sample frequency for each problem. This estimate is unbiased, meaning that in expectation across samples, we are estimating the true parameter for each problem. Can we provide a "better" estimate? The answer to that question depends on our definition of "better"- we want an estimate that's "probably closer" to the three values we're estimating. Mathematically, if we're interested in reducing the *mean squared error* of our parameter estimates, we can do better than predicting the sample frequencies (Stein 1956). In fact, the James-Stein estimator[2] estimates,

$$(\hat{\mu}_{\text{hat}}, \hat{\mu}_{\text{car}}, \hat{\mu}_{\text{ba}}) = (h(1 - \frac{\sigma^2/N}{h^2 + c^2 + b^2}), c(1 - \frac{\sigma^2/N}{h^2 + c^2 + b^2}), b(1 - \frac{\sigma^2/N}{h^2 + c^2 + b^2}))$$

have lower mean squared error than the intuitive estimates. The alarming part of this result is that Wimbledon hat attendance has seemingly become useful in predicting the batting average of Aaron Judge and the number of red cars driving around Columbus Circle. Thinking carefully, the Wimbledon hat attendance aids in estimating these two other quantities because our objective is to pick estimators to minimize the mean squared error for all three quantities simultaneously, not to pick estimators that minimize each estimator's mean squared error individually. The paradox lies in the fact that to reduce the mean squared error of the joint problem, you can do better than estimating the sample frequencies for each problem (Stein 1956). A sanity-check question: when would we ever want to SIMULTANEOUSLY estimate the proportion of attendees at Wimbledon that wear a hat each day, the fraction of red cars driving around Columbus Circle in Manhattan every day, and the 2022 batting average of Aaron Judge of the New York Yankees? Probably never! Thus, the paradox doesn't make much sense in real-life but it's something to be aware of. I'll provide a more formal introduction to the James-Stein Estimator later in this article along with a simplified proof.

Bias-Variance Tradeoff of Estimators

Suppose we're trying to estimate a parameter x and give estimate $\hat{x}$ based on data D. We want $\hat{x}$ to be "close" to x. To formulate that, we typically pick $\hat{x}$ so as to minimize the *mean squared error*. Mathematically, we pick $\hat{x}$ to minimize

[1] In our initial analysis of the paradox, it's not necessary that N be the same for all observations nor that the random variables have identical variance; it just simplifies the analysis.

[2] The reason that this estimator is called a shrinkage estimator is because of the shrinkage factor $1 - \frac{\sigma^2/N}{h^2+c^2+b^2}$. If we're worried that the shrinkage factor could take on a negative value, we could modify the shrinkage factor to $\max(1$ - $\sigma^2/N \frac{}{h^2+c^2+b^2}$, 0).

$MSE(\hat{x}) := \mathbb{E}_D[(\hat{x} - x)^2]$. The D subscript indicates that the expectation is taken over the sampling outcome D[3]- I omit the subscript from here on out for brevity. I assume the reader is familiar with the decomposition of error into bias and variance though I'll provide the derivation here:

$$\begin{aligned}
MSE(\hat{x}) &= \mathbb{E}[(\hat{x} - x)^2] \\
&= \mathbb{E}[((\hat{x} - \mathbb{E}[\hat{x}]) - (\mathbb{E}[\hat{x}] - x))^2] \\
&= \mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2 + (\mathbb{E}[\hat{x}] - x)^2 - 2(\hat{x} - \mathbb{E}[\hat{x}])(\mathbb{E}[\hat{x}] - x)] \\
&= \mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2] + \mathbb{E}[(\mathbb{E}[\hat{x}] - x)^2] - 2\mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])(\mathbb{E}[\hat{x}] - x)] \\
&= \mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2] + \mathbb{E}[(\mathbb{E}[\hat{x}] - x)^2] - 2(\mathbb{E}[\hat{x}] - x)\cancelto{0}{(\mathbb{E}[\hat{x}] - \mathbb{E}[\hat{x}])} \\
&= \mathbb{E}[(\hat{x} - \mathbb{E}[\hat{x}])^2] + (\mathbb{E}[\hat{x}] - x)^2 \\
&= \text{Var}(\hat{x}) + (\text{bias}(\hat{x}))^2
\end{aligned}$$

To explain the nontrivial steps, in step 5, I use the fact the term $(\mathbb{E}[\hat{x}] - x)$ is independent of the outer expectation on D and so we can pull it out. Then, I simplify the other part of that last term: $\mathbb{E}[\hat{x} - \mathbb{E}[\hat{x}]] = \cancelto{0}{\mathbb{E}[\hat{x}] - \mathbb{E}[\hat{x}]}$ since $\mathbb{E}[\hat{x}]$ is a constant and hence can be pulled out of the outer expectation by linearity. In the last step, notice that the first term is simply the definition for Var$(\hat{x})$ and in the second term, we can interpret $\mathbb{E}[\hat{x}] - x$ as the bias (ie., how far our estimate is from the true parameter on average).

The result, $MSE(\hat{x}) = \text{Var}(\hat{x}) + (\text{bias}(\hat{x}))^2$, has many nice interpretations and implications. Our error from predicting $\hat{x}$ can be decomposed into a variance and bias component[4]. The variance component is the part of the estimation error that arises because our sample is a random. The bias component is the estimation error if our sample were the entire population.

In estimating $\hat{x}$, we equally penalize error due to variance and bias^2, and we penalize extremes of each more due to the squared nature of each term. It is arbitrary that we pick $\hat{x}$ so as to minimize Var$(\hat{x})$ + $(\text{bias}(\hat{x}))^2$. We could also insert $\lambda \in [0, 1]$ and pick $\hat{x}$ so as to minimize $\lambda * \text{Var}(\hat{x}) + (1 - \lambda) * (\text{bias}(\hat{x}))^2$ or raise each of the terms to different powers. If an estimator is worse than another in both variance and bias, it's clear it's the worse estimator, but it's typically the case that one's better in one component and worse in the other and we must decide how much we care about each (ie., λ). The success of an estimator $\hat{x}$ also depends on what the true value of x ends up being. If one estimator is worse than another for all values of x, it's easy to say that it's the worse estimator; however, if the estimator is better for some values of x and worse for others, the statistician again must make a choice.

When we use an unbiased estimator, $0 = \text{bias}(\hat{x}) = \mathbb{E}[\hat{x}] - x$. That means, our estimation error $MSE(\hat{x}) = \text{Var}(\hat{x})$ comes entirely from the variance term and depends only on the sample being random. In the case of estimating the frequency of hat attendance at Wimbledon, red cars at Columbus Circle, and Aaron Judge batting average in section , the unbiased estimator was predicting the sample frequencies for each of the values. It's often the case that an unbiased estimator has the lowest mean squared error of any estimator but that's not always the case as you will see clearly in the next example.

The Zero Estimator

Suppose we observe a single draw X_1 from a distribution X with mean x and variance σ^2. Consider two different estimators $\hat{x}_1 = X_1$ and $\hat{x}_2 = 0$. In the case of $\hat{x}_1$, we estimate x as our single sample. In $\hat{x}_2$, we estimate x as 0 regardless of what we observe with X_1- I call $\hat{x}_2$ the zero estimator. Is it possible that $\hat{x}_2$ is better than $\hat{x}_1$ in terms of mean-squared error?

[3] This detail is important and glanced over by those learning and critical to analyzing this error function.

[4] Rasmusen (2021) explains nicely that these aren't necessarily the only types of errors. For example, you might have some strange preference that $\hat{x}$ not lie in the interval $[1, 3]$. That is a situational type of error and not applicable here- I'll ignore this in my future discussion.

$$\begin{aligned} MSE(\hat{x}_1) &= \text{Var}(\hat{x}_1) + (\text{bias}(\hat{x}_1))^2 \\ &= \mathbb{E}[(\hat{x}_1 - \mathbb{E}[\hat{x}_1])^2] + (\mathbb{E}[\hat{x}_1] - x)^2 \\ &= \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] + \cancelto{0}{(\mathbb{E}[X_1] - x)^2} \\ &= \sigma^2 \end{aligned}$$

$$\begin{aligned} MSE(\hat{x}_2) &= \text{Var}(\hat{x}_2) + (\text{bias}(\hat{x}_2))^2 \\ &= \mathbb{E}[(\hat{x}_2 - \mathbb{E}[\hat{x}_2])^2] + (\mathbb{E}[\hat{x}_2] - x)^2 \\ &= \cancelto{0}{\mathbb{E}[(0 - 0)^2]} + (0 - x)^2 \\ &= x^2 \end{aligned}$$

As you can see from the derivation, $\hat{x}_1$ is unbiased- the bias component of $MSE(\hat{x}_1)$ equals 0 since $\mathbb{E}[\hat{x}_1] = \mathbb{E}[X_1] = x$ and thus the $MSE(\hat{x}_1)$ depends solely on it's variance component. The estimator $\hat{x}_2$ has zero variance since we always predict zero and so the $MSE(\hat{x}_2)$ depends solely on the bias component. To decide which estimator is better (according to MSE), we need to see if $MSE(\hat{x}_1) < MSE(\hat{x}_2) \iff \sigma^2 < x^2$ or vice versa. In other words, depending on the values of σ and x, an estimator that ignores the sample and always predicts zero could be "better" than the unbiased estimator. To give intuition for this result, if the variance of X is very large relative to it's mean, we don't learn very much from the sample so we might as well predict something arbitrary, say 0.

We have an issue though, we don't actually know the values of x and σ^2 when we're trying to estimate. So it's hard to say which of $\hat{x}_1$ or $\hat{x}_2$ is better apriori. If we did know x, the best estimator of x would be $\hat{x} = x$- we ignore the sample completely and the thought experiment between $\hat{x}_1$ and $\hat{x}_2$ makes no sense.

To say that one estimator is *completely superior* to another (according to the MSE criterion), we want the former estimator to have lower MSE than the latter regardless of what value x, the estimand, takes on. In the case of the zero estimator and the unbiased estimator here, we don't have a a conclusive result either way. The power of the James-Stein (1956) estimator is that it is indeed *completely superior* to the unbiased estimator of sample frequencies. Before we get to the James-Stein estimator though, there is a more simple class of shrinkage estimators to look- the Oracle estimator. Even before that, I'll summarize why shrinkage estimators are useful when looking at this tennis data.

Shrinkage Estimators Definition

Consistent with our earlier discussion, suppose we're trying to estimate the mean x of some distribution X with known finite variance σ^2 given N iid draws $D = [X_1, ..., X_N]$ from the distribution. A shrinkage estimator $\hat{x}_{\text{shrink}}$ for estimating x takes the form:

$$\hat{x}_{\text{shrink}} := (1 - B)\bar{x} + Bx_0 \tag{1}$$

where $\bar{x} := \sum_{i=1}^{N} \frac{X_i}{N}$, x_0 is some value we want to "shrink" our estimate towards (typically due to prior information), and $B : (X_1, ..., X_N, X, N) \mapsto [0, 1]$ specifies how much we want to shrink our estimate towards x_0. First, notice that $\bar{x}$ is the unbiased estimator of x and so for any value of $B \neq 0$, $\hat{x}_{\text{shrink}}$ is a biased estimator. Another remark, we want B to be a nonincreasing function of N since if our sample is larger, our sample is more representative of the overall population and thus we want to place at least as much trust in $\bar{x}$. I encourage the reader to look at the subsequently presented shrinkage estimators through this form.

x_0-*Oracle Estimator*

Let's write the x_0-Oracle slightly differently from the structure above, as in Rasmusen (2021).

$$\hat{x}_{\text{Oracle}(x_0)} := \bar{x} - B(\bar{x} - x_0) \tag{2}$$

We're interested in computing the MSE of $\hat{x}_{\text{Oracle}(x_0)}$ to compare it to the MSE of $\bar{x}$. Again, the expectation is taken over the data D.

$$\begin{aligned}
MSE(\hat{x}_{\text{Oracle}(x_0)}) &= \mathbb{E}[((\bar{x} - B(\bar{x} - x_0)) - x)^2] \\
&= \mathbb{E}[(\bar{x} - x) - B(\bar{x} - x_0))^2] \\
&= \mathbb{E}[(\bar{x} - x)^2 + B^2(\bar{x} - x_0)^2 - 2B(\bar{x} - x)(\bar{x} - x_0)] \\
&= \frac{\sigma^2}{N} + B^2\mathbb{E}[\bar{x}^2 + x_0^2 - 2\bar{x}x_0] - 2B\mathbb{E}[\bar{x}^2 - \bar{x}x - \bar{x}x_0 + xx_0] \\
&= \frac{\sigma^2}{N} + B^2((\frac{\sigma^2}{N} + x^2) + x_0^2 - 2xx_0) - 2B((\frac{\sigma^2}{N} + x^2) - x^2 - \cancelto{0}{xx_0 + xx_0)} \\
&= \frac{\sigma^2}{N} + B^2(\frac{\sigma^2}{N} + (x - x_0)^2) - 2B\frac{\sigma^2}{N}
\end{aligned}$$

Now, we differentiate $MSE(\hat{x}_{\text{Oracle}(x_0)})$ with respect to B and set equal to 0 to find the minimizing B. It's clear the second order condition is satisfied for a global minimum since we're a differentiating a degree two polynomial on B with positive coefficient on squared term (ie., $\frac{\sigma^2}{N} + (x - x_0)^2 > 0$).

$$0 = \frac{\mathrm{d}MSE(\hat{x}_{\text{Oracle}(x_0)})}{\mathrm{d}B} = 2B(\frac{\sigma^2}{N} + (x - x_0)^2) - 2\frac{\sigma^2}{N}$$

Solving for B gives:

$$B = \frac{\sigma^2/N}{\sigma^2/N + (x - x_0)^2}$$

and consequently

$$\hat{x}_{\text{Oracle}(x_0)} = \bar{x} - \frac{\sigma^2/N}{\sigma^2/N + (x - x_0)^2}(\bar{x} - x_0)$$

and the mean squared error at the optimum B:

$$\begin{aligned}
MSE(\hat{x}_{\text{Oracle}(x_0)}) &= \frac{\sigma^2}{N} + (\frac{\sigma^2/N}{\sigma^2/N + (x - x_0)^2})^2(\frac{\sigma^2}{N} + (x - x_0)^2) - 2(\frac{\sigma^2/N}{\sigma^2/N + (x - x_0)^2})\frac{\sigma^2}{N} \\
&= \frac{\sigma^2}{N}(1 - \frac{\sigma^2/N}{\sigma^2/N + (x - x_0)^2})
\end{aligned}$$

First, notice that the mean squared error for the x_0-Oracle estimator is lower than the mean squared error of the unbiased estimator ($\frac{\sigma^2}{N}$) regardless of what is the true value of x. This result is surprising since no matter how close x_0 is to x, the x_0-Oracle estimator has lower MSE than the unbiased estimator. We can have terrible prior information x_0 and still benefit. Though, if x_0 is close to x, we benefit more. Also notice that B is an increasing function of σ^2- the intuition is that if our draws from X have larger variance, they're typically further from x and so we should trust these draws less when predicting x. Also, notice that B is a decreasing function of N, which agrees with our sanity check in the shrinkage estimators definition section (section). As a remark, we do not know the value of x and so we must estimate this parameter when employing this estimator.

The James-Stein Estimator, Shrinking to 0

Now it's time to mathematically justify the paradoxical claim I made in introducing shrinkage estimators in section , that the James-Stein shrunken estimators have lower total mean squared error than the sample frequencies when we're trying to jointly estimate the hat attendance percentage at Wimbledon, the percentage of red cars in Columbus Circle, and the batting average of Aaron Judge. In other words, I want to justify the paradox that Wimbledon hat attendance is ”useful” in estimating the fraction of red cars in Columbus Circle and the batting average of Aaron Judge.

For ease of notation, assume we have k Bernoulli distributions $X_1, ..., X_k$ and we're trying to estimate their means $x_1, ..., x_k$, respectively. Each distribution has identical variance σ^2 and we observe N iid draws [5] $(X_{11}, ..., X_{1N}, X_{21}, ..., X_{2N}, ..., X_{k1}, ..., X_{kN})$ from each distribution. A reminder that in my notation, $\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$ and hence $\bar{x}_i \sim \mathcal{N}(x_i, \frac{\sigma^2}{N})$ approximately.

We define the James-Stein Estimator, shrinking to 0, for x_1 with analogous expressions for $x_2, ..., x_k$:

$$\hat{x}_{1\text{JS}} := \bar{x}_1 - (k-2)\frac{\sigma^2/N}{\sum_{i=1}^{k}\bar{x}_i^2}\bar{x}_1 \tag{3}$$

To aid in demonstrating that the James-Stein estimator has lower total MSE than the unbiased sample averages, as in Rasmusen (2021), I define:

$$g(\bar{x}_1) := (k-2)\frac{\sigma^2/N}{\sum_{i=1}^{k}\bar{x}_i^2}\bar{x}_1$$

with derivative

$$\frac{\mathrm{d}g(\bar{x})}{\mathrm{d}\bar{x}} := (k-2)\frac{\sigma^2}{N}(\frac{1}{\sum_{i=1}^{k}\bar{x_i}^2} - \frac{2\bar{x_1}^2}{(\sum_{i=1}^{k}\bar{x_i}^2)^2})$$

There's one other useful fact to know in our derivation of the MSE: $\mathbb{E}[(\bar{x}_1 - x_1)g(\bar{x}_1)] = \frac{\sigma^2}{N}\mathbb{E}[\frac{\mathrm{d}g(\bar{x}_1)}{\mathrm{d}\bar{x}_1}]$. As Rasmusen (2021) explains, this comes as an implication of Stein's Lemma from Stein (1981) for rotationally symmetric densities like the normal distribution. Now onto our simplification of the $MSE(\hat{x}_{1\text{JS}})$.

$$\begin{aligned}
MSE(\hat{x}_{1\text{JS}}) &= \mathbb{E}[(\hat{x}_{1\text{JS}} - x_1)^2] \\
&= \mathbb{E}[((\bar{x}_1 - g(\bar{x}_1)) - x_1)^2] \\
&= \mathbb{E}[((\bar{x}_1 - x_1) - g(\bar{x}_1))^2] \\
&= \mathbb{E}[(\bar{x}_1 - x_1)^2] + \mathbb{E}[g(\bar{x}_1)^2] - 2\mathbb{E}[(\bar{x}_1 - x_1)g(\bar{x}_1)] \\
&= \frac{\sigma^2}{N} + \mathbb{E}[(k-2)^2\frac{(\sigma^2/N)^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}\bar{x}_1^2] - 2\frac{\sigma^2}{N}\mathbb{E}[(k-2)\frac{\sigma^2}{N}(\frac{1}{\sum_{i=1}^{k}\bar{x}_i^2} - \frac{2\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2})] \\
&= \frac{\sigma^2}{N} + (k-2)^2(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] - 2(k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\sum_{i\neq 1}^{k}\bar{x}_i^2 - \bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] \\
&= \frac{\sigma^2}{N} + (k-2)^2(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] + 2(k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] - 2(k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\sum_{i\neq 1}^{k}\bar{x}_i^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] \\
&= \frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}]((k-2)+2) - 2(k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{\sum_{i\neq 1}^{k}\bar{x}_i^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}] \\
&= \frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2(k\mathbb{E}[\frac{\bar{x}_1^2}{(\sum_{i=1}^{k}\bar{x_i}^2)^2}] - 2\mathbb{E}[\frac{\sum_{i\neq 1}^{k}\bar{x}_i^2}{(\sum_{i=1}^{k}\bar{x}_i^2)^2}]) \qquad (4)
\end{aligned}$$

At this point, it's difficult to tell whether $MSE(\hat{x_1}_{\text{JS}})$ is larger or smaller than $MSE(\bar{x_1}) = \frac{\sigma^2}{N}$. Let's look at the sum of the MSE's across all k dimensions.

[5] The assumptions of identical variance, same number of samples from each distribution, and Bernoulli distributions can be relaxed but the math becomes algebraicly more complicated and is out of scope of this article. In other words, the James-Stein Estimator is indeed generalizable for different sample sizes, different variances, and different distributions.

$$
\begin{aligned}
\sum_{j=1}^{k} MSE(\hat{x}_{j\text{JS}}) &= \sum_{j=1}^{k}(\frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2(k\mathbb{E}[\frac{\bar{x}_j^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}] - 2\mathbb{E}[\frac{\sum_{i\neq j}^k \bar{x}_i^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}])) \\
&= k\frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2\sum_{j=1}^{k}(k\mathbb{E}[\frac{\bar{x}_j^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}] - 2\mathbb{E}[\frac{\sum_{i\neq j}^k \bar{x}_i^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}]) \\
&= k\frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{k\sum_{j=1}^k \bar{x}_j^2 - 2(k-1)\sum_{j=1}^k x_j^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}] \\
&= k\frac{\sigma^2}{N} + (k-2)(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{(k-2(k-1))\sum_{j=1}^k \bar{x}_j^2}{(\sum_{i=1}^k \bar{x}_i^2)^2}] \\
&= k\frac{\sigma^2}{N} - (k-2)^2(\frac{\sigma^2}{N})^2\mathbb{E}[\frac{1}{\sum_{i=1}^k \bar{x}_i^2}] \\
&< k\frac{\sigma^2}{N} = \sum_{j=1}^{k} MSE(\bar{x}_j) \text{ if } k > 2
\end{aligned}
$$

We have shown that the total mean squared error when using the James-Stein estimator is less than that of using the arithmetic average for each of the components if $k > 2$. If $k = 1$, the James-Stein estimator loses, and if $k = 2$ the James-Stein estimator ties. Even if $k > 2$, it's not clear that each component will have a smaller mean squared error than the arithmetic average, as we see in equation 4- our only guarantee is that the sum of the mean squared errors of all components will be smaller with the James-Stein estimator.

Understanding the James-Stein Estimator

To gain some intuition for what's going on, one useful move is to compare the oracle estimator for $x_0 = 0$ and James-Stein estimator shrinking to 0:

$$
\begin{aligned}
\hat{x}_{\text{Oracle}(x_0)} &:= \bar{x} - \frac{\sigma^2/N}{\sigma^2/N + x^2}\bar{x} \\
\hat{x_1}_{\text{JS}} &:= \bar{x_1} - (k-2)\frac{\sigma^2/N}{\sum_{i=1}^k \bar{x_i}^2}\bar{x_1}
\end{aligned}
$$

First, notice that $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 = \sigma^2/N + x^2$ and so we can write $\hat{x}_{\text{Oracle}(x_0)} = \bar{x} - \frac{\sigma^2/N}{\mathbb{E}[X^2]}\bar{x}$. In the JS estimator, for intuition-reasons, lets replace[6] the $k-2$ with k and sweep it into the denominator so that $\hat{x}_{1\text{JS}} \approx \bar{x}_1 - \frac{\sigma^2/N}{\frac{1}{k}\sum_{i=1}^k \bar{x}_i^2}\bar{x}_1$. The expressions for the two now estimators now look very similar except for the fact that the denominator of the Oracle shrinkage multiplier is the expectation of X^2 and the denominator of the James-Stein Estimator is the arithmetic average of the $\bar{x}_i^2$- it's a sample estimate of $\bar{x}_1^2$ in its own right if we "pretend" that all of the $\bar{x}_i$'s come from the same distribution.

Given our construction[7], in the James-Stein estimator, we're forced to have one shrinkage multiplier for all of the estimands and hence this "pretending" is the best we can do- we did prove that it's still better than the unbiased estimates! In the extreme where all distributions are the same, the denominator of the JS estimator becomes a proper unbiased of $\bar{x}_1^2$ and our estimators align even more with the corresponding Oracle estimators. In these cases where $X_1, ..., X_k$ are very similar, since our JS estimators are very similar to the analogous Oracle estimators, we're guaranteed[8] that each component will have less mean squared error than the unbiased estimates. Note this guarantee is stronger than having having the the sum of mean squared errors smaller than the sum of mean squared errors for unbiased estimates.

[6] The $k-2$ (as opposed to k) comes from the fact that fact that for samples that give unrealistically large $\bar{x_1}$, we want to shrink these samples more but the correspondingly large $\bar{x_1}^2$ in the denominator results in a smaller shrinkage than desired. In other words, large sample averages are correlated with small shrinkage factors and so we replace k with $k-2$ to correct for this effect.

[7] We here required that each distribution have identical variance and that we took the same number of draws from each distribution.

[8] It's an interesting and loaded question, how similar do these distributions need to be to properly guarantee that the mean squared error of each component of the James-Stein Estimator is less than the mean squared error for the unbiased estimates? The answer to this question is out of scope of this article. Though, Rasmusen (2021) does provide a nice simple algebraic proof of a positive result to this question when all estimands are equal.

James-Stein Estimator Applied to Tennis Analysis

In the space of tennis, let's focus on estimating the serve win rate for various players in the year. We have k Bernoulli distributions of serve win rates: $S_1, ..., S_k$ and we're trying to estimate their means $s_1, ..., s_k$, respectively. s_i should be interpreted as the true probability that player i wins a serve against a "representative" opponent and S_i the fraction of serves player i won in the past year. We say that $\frac{\sigma_i^2}{N_i} := \text{Var}(S_i)$ where N_i is the number of serves by player i in the past year and σ_i^2 is the variance of the probability of player i winning a single serve.

Let's add some additional assumptions in the lens of tennis- I want concrete-ize the notion that the serve win rates of all of the players must somehow be related. Let's say that s_i are selected iid from the following distribution: $s_i \sim \mathcal{N}(s, d)$ where s and d are the mean and variance of the distribution, respectively. Then, let's say that each S_i is selected from the distribution: $S_i \sim N(s_i, \frac{\sigma_i^2}{N_i})$ so that each S_i is also conditionally independent given the s_is. Then, algebra and some calculus implies that:

$$s_i | S_i \sim \mathcal{N}((1 - B_i)S_i + B_i s, \frac{\sigma_i^2}{N_i}(1 - B_i)) \tag{5}$$

where $B_i := \frac{\sigma_i^2/N_i}{\sigma_i^2/N_i + d}$. What's going on here? We're taking a two stage approach to generate the S_is. We're assuming that player i has its true serve win probability, s_i selected iid from a normal distribution. Then, we assume that player i wins each of its N_i serves iid with probability s_i so that S_i is approximately distributed normally (by the central limit theorem) with mean s_i and variance $\frac{\sigma_i^2}{N_i}$. We observe S_i and we want to estimate s_i and hence the derived conditional distribution in equation 5. If we want to estimate $s_i | S_i$, one intuitive approach, and that's what I'll do, is to estimate s_i as its conditional expectation. In other words, we estimate:

$$\hat{s}_{i(\text{tennis})} := \mathbb{E}[s_i | S_i] = (1 - B_i)S_i + B_i s \tag{6}$$

There's also nice intuition in the expression for $\hat{s}_{i(\text{tennis})}$. This estimator is a weighted average of the presumably known quantities S_i and s. We weight s more when B_i is larger- that happens when $\frac{\sigma_i^2}{N_i}$ is relatively larger than d. In other words, that means when the sample average (S_i) comes with more uncertainty, we weight the globally known variable s more. When the sample average comes with less uncertainty ($\frac{\sigma_i^2}{N_i}$ is relatively smaller than d), we weight the sample average more. It's nice to link this definition of $\hat{s}_{i(\text{tennis})}$ back to the original definition of a shrinkage estimator, x_{shrink}, in equation 1- the expressions basically say the same thing! When we're more confident in our sample average, we weight that quantity by more in our estimator; when we're less confident in our sample average, we weight our prior information more[9].

[9] Coincidentally, the solved expression for $\hat{s}_{i(\text{tennis})}$ was discovered analogously in the setting of an insurance company selecting a rate to cover the (potential) claim filed in period $n + 1$ ($X_{i(n+1)}$) for individual i after observing n prior claims (X_{ij} for $1 \leq j \leq n$) for k individuals (Buhlmann and Straub (1970)). The insurance company assumes a risk parameter $\theta_i \sim \Pi_\theta$ and claims distributed iid for individual i: $X_{ij} \sim f_{X|\theta}(X|\theta_i)$. There are no assumptions on these distributions beyond the fact that they have a finite mean and variance. The insurance company then solves for the best linear predictor (according to mean squared error) of $X_{i(n+1)}$ given $X_1, ..., X_n$. In their solution, using B_i from our notation, they equivalently write that $\hat{x}_{i(n+1)} := (1 - B_i)\bar{X}_i + B_i x$ where $1 - B_i = \frac{n}{n + \frac{\mathbb{E}[\text{Var}(X|\theta)]}{\text{Var}(\mathbb{E}[X|\theta])}}$, $\bar{X}_i := \frac{1}{n}\sum_{j=1}^n X_{ij}$, and $x := \frac{1}{kn}\sum_{i=1}^k \sum_{j=1}^n X_{ij}$. In their intuition, holding everything else fixed, if the θ_i are more homogeneous, then we expect $\text{Var}(\mathbb{E}[X|\theta])$ to be small and so $1 - B_i$ small and therefore we place more weight on the global mean x when trying to predict the next claim. If the θ_i are more heterogeneous, vice-versa.

REFERENCES

Introduction to Buhlmann credibility, February 2, 2010. URL: `https://mathmodelsblog.wordpress.com/2010/02/02/introduction-to-buhlmann-credibility/`.

Eric Rasmusen. Understanding Shrinkage Estimators: From Zero to Oracle to James-Stein. 2021. URL: `http://www.rasmusen.org/papers/shrinkage-rasmusen.pdf`.