Understanding Shrinkage Estimators: From Zero to Oracle to James-Stein

June 29, 2015

Abstract

The standard estimator of the population mean is the sample mean ($\hat{\mu_y} = \overline{y}$), which is unbiased. Constructing an estimator by shrinking the sample mean results in a biased estimator, with an expected value less than the population mean. On the other hand, shrinkage always reduces the estimator's variance and can reduce its mean squared error. This paper tries to explain how that works. I start with estimating a single mean using the zero estimator (a neologism, $\hat{\mu_y} = 0$) and the oracle estimator ($\hat{\mu_y} = \left(\frac{\mu_y^2}{\mu_y^2+\sigma^2}\right)\overline{y}$), and continue with the unrelated-average estimator (another neologism, $\hat{\mu_y} = \frac{\overline{w}+\overline{y}+\overline{z}}{3}$). Thus prepared, it is easier to understand the James-Stein estimator in its simple form with known homogeneous variance ($\hat{\mu_y} = \left(1 - \frac{(k-2)\sigma^2}{\overline{w}^2+\overline{y}^2+\overline{z}^2}\right)\overline{y}$)) and in extensions. The James-Stein estimator combines the oracle estimate's coefficient shrinking with the unrelated-average estimator's cancelling out of overestimates and underestimates.

Eric Rasmusen: John M. Olin Faculty Fellow, Olin Center, Harvard Law School; Visiting Professor, Economics Dept., Harvard University, Cambridge, Massachusetts (till

Economics and Public Policy, Kelley School of Business, Indiana University.

Structure

1. Biased estimators can be "better".
2. The zero estimator.
3. The seventeen estimator.
4. The oracle estimator.
5. The unrelated-average estimator.
6. The James-Stein estimator with equal and known variances.
7. The positive-part James-Stein estimator.
8. The James-Stein estimator with shrinkage towards the unequal-average.
9. Understanding the James-Stein estimator.
10. The James-Stein estimator with unequal but known variances.
11. The James-Stein estimator with unequal and unknown variances.

The James-Stein Estimator

W, Y, and Z are normally distributed with unknown means μ_w, μ_y, and μ_z and known identical variances σ^2. We have one observation on each variable, w, y, z. The sample means are $\hat{\mu}_w(\overline{w}) = w, \hat{\mu}_y(\overline{y}) = y$, and $\hat{\mu}_z(\overline{z}) = y$. But for *any* values that μ_w, μ_y, and μ_z might happen to have, an estimator with lower total mean squared error is the James-Stein estimator which for y is this and for w and z is similar:

$$\hat{\mu}_{JS,w} = \overline{w} - \frac{(k-2)\sigma^2}{w^2+y^2+z^2}\overline{w}, \tag{1}$$

Some questions to think about

1. Why $k-2$ instead of k?
2. Why not shrink towards the unrelated-average mean instead of towards zero?
3. Why not shrink all three towards $\overline{y}$ instead of towards zero?
4. Why does it not work if σ^2 is different for W, Y, Z and needs to be estimated?
5. Why not use just Y and Z to calculate W's shrinkage percentage?

The Sequence of Thought

1. Hypothesize a value μ^r for the true parameter, μ.
2. Pick an estimator of μ as a function of the observed sample: $\hat{\mu}(y)$.
3. Compare μ and $\hat{\mu}(y)$ for the various possible samples we might have give that $\mu = \mu^r$. Usually we'll condense this to the mean, variance, and mean squared error of the estimator: $E\hat{\mu}(y), E(\hat{\mu}(y) - E\hat{\mu}(y))^2$, and $E(\hat{\mu}(y) - \mu^r)^2$.
4. Go back to (1) and try out how the estimator does for another hypothetical value of μ. Keep looping till you've covered all possible value of μ.

The Zero Estimator

The sample mean is $\hat{\mu}_{\overline{y}} = y$

Our new estimator, "the zero estimator" is $\hat{\mu}_{zero} = 0$.

$$MSE(\hat{\mu}) \; = \; E(\hat{\mu} - \mu)^2 \tag{2}$$

After some algebra,

$$\begin{aligned} MSE(\hat{\mu}) &= E[\hat{\mu} - E\hat{\mu}]^2 + E[\hat{\mu} - \mu]^2 \\ \boxed{MSE(\hat{\mu}) &= E(Sampling\ Error)^2 + Bias^2} \end{aligned} \tag{3}$$

The sampling error is the distance between $\hat{\mu}$ and μ that you get because the sample is randomly drawn, different every time you draw it.

The bias is the distance between $\hat{\mu}$ and μ that you'd get if your sample was the entire population, so there was no sampling error.

Often one estimator will be better in sampling error and another one in bias. Or, it might be that which estimator is better depends on the true value of μ.

Mean squared error weights sampling error and bias equally, but extremes of either of them get more than proportional weight. This will be important.

Mean Squared Errors

How do our two estimators do in terms of mean square error? The population variance is σ^2.

$$MSE(\hat{\mu}_{\overline{y}}) = E[\overline{y} - E\overline{y}]^2 + E[E\overline{y} - \mu]^2 \tag{4}$$

$$\boxed{MSE(\hat{\mu}_{\overline{y}}) = \sigma^2}$$

and

$$MSE(\hat{\mu}_{zero}) = E[0 - E(0)]^2 + E[E(0) - \mu]^2 \tag{5}$$

$$\boxed{MSE(\hat{\mu}_{zero}) = \mu^2}$$

Thus, $\overline{y}$ is better than the zero estimator if and only if $\sigma < \mu$. That makes sense. The zero estimator's bias is μ, but its variance is zero. By ignoring the data, it escapes sampling error. I

If the population variance is high, it is better to give up on using the sample for estimation and just guess zero.

The Seventeen Estimator

Let me emphasize that the key to the superiority of the zero estimator over $\overline{y}$ is that variance is high so sampling error is high. The key is *not* that 0 is a low estimate. The intuition is that there is a tradeoff between bias and sampling error, and so a biased estimator might be best.

The "seventeen estimator" is like the zero estimator, except it is defined as $\hat{\mu}_{17} = 17$.

$$MSE(\hat{\mu}_{seventeen}) = E[17 - E(17)]^2 + E[E(17) - \mu]^2 \tag{6}$$

$$\boxed{MSE(\hat{\mu}_{seventeen}) = (17 - \mu)^2}$$

The seventeen estimator is better than $\overline{y}$ if $\sigma > 17 - \mu$. Thus, it is a good estimator if the variance is big, and a good estimator if the true mean is big and positive. It is not shrinking the estimate from $\overline{y}$ towards 0 that helps when variance is big: it is making the estimate depend less on the data.

III. The Oracle Estimator

Let's next think about shrinkage estimators generally, of which $\overline{y}$ and the zero estimator are the extreme limits.

How about an "expansion estimator", e.g. $\hat{\mu} = 1.4\overline{y}$? That estimator is biased, plus it depends *more* on the data, not less, so it will have even bigger sampling error than $\overline{y}$. Hence, we can restrict attention to shrinkage estimators.

The "oracle estimator" is the best possible (not proved here). It is:

$$\boxed{\hat{\mu}_{oracle} \equiv \overline{y} - \left(\frac{\sigma^2}{\sigma^2+\mu^2}\right)\overline{y}} \tag{7}$$

Equation (7) says that if μ is small, we should shrink a bigger percentage. If σ^2 is big, we should shrink a lot. The James-Stein estimator will use that idea.

IV. The Unrelated-Average Estimator

Suppose we have $k = 3$ independent estimands, W, Y, and Z. We can still use the sample means, of course— that is to say, use the observed values w, y, and z as our estimator. Or we could use the zero estimator, (0,0,0). But consider "the unrelated average estimator" : the average of the three independent estimands,

$$\boxed{\hat{\mu}_{UAE,w} = \hat{\mu}_{UAE,y} = \hat{\mu}_{UAE,z} \equiv \frac{w+y+z}{3}} \tag{8}$$

After lots of algebra,

$$\boxed{MSE_{UAE} = \sigma^2 + \frac{2}{3}\Big((\mu_w^2 + \mu_y^2 + \mu_z^2) - (\mu_w\mu_z + \mu_w\mu_y + \mu_y\mu_z)\Big)} \tag{9}$$

Not bad! In this context,

$$MSE_{wbar,\overline{y},zbar} = 3\sigma^2 \tag{10}$$

The unrelated-average estimator cuts the sampling error back by 2/3, though at a cost of adding bias equal to $\frac{2}{3}\Big((\mu_w^2 + \mu_y^2 + \mu_z^2) - (\mu_w\mu_z + \mu_w\mu_y + \mu_y\mu_z)\Big)$. So if the variances are high and the means aren't too big, we have an improvement over the unbiased estimator.

The Unrelated-Average Estimator with Coincidentally Close Estimands

. Notice what happens if $\mu_w = \mu_y = \mu_z = \mu$. Then $MSE_{UAE} = \sigma^2 + \frac{2}{3}\Big((\mu^2 + \mu^2 + \mu^2) - (\mu \cdot \mu + \mu \cdot \mu + \mu \cdot \mu\Big) = \sigma^2$, better than the standard estimator no matter how low the variance is! (unless, of course, $\sigma_2 = 0$, in which case the two estimators perform equally well). The closer the three estimands are to each other, the better the unrelated-average estimator works. If they're even slightly unequal, though, the negative terms in the second part of (10) are outweighed by the positive terms.

If $\mu_w = 3, \mu_y = 3, \mu_z = 10$, for example, the last part of the MSE is $\frac{2}{3}\Big((9+9+100) - (30 + 9 + 30)\Big) = \frac{2}{3}\Big(39\Big)$, and if the variance were only $\sigma^2 = 4$ then $MSE_{UAE} = 17$ and $MSE_{wbar,\overline{y},zbar} = 12$.

Return to the case of $\mu_w = \mu_y = \mu_z$, and suppose we know this in advance of getting the data. We have one observation on each of three different independent variables to estimate the population mean when that mean is the same for all three. But that is a problem identical ("isomorphic", because it maps one to one) to the problem of having three independent observations on one variable.

Close Estimands and Measurement Error

One variable with three observations, it's like having observations with measurement error where some of the observations' measurement errors don't have zero means. It's as if we have observations w and y without error, but observation z has measurement error. We would then have the decision of whether to use z in our estimation. If we knew the measurement error was -1, we'd use z, but if the measurement error is the $+7$ in the example, we'd do better leaving out z. (If we know the exact measurement error, we can use that fact in the estimation, of course, but think of this as knowing z has a little measurement error bias vs. a lot without knowing specifics.)

What's going on is regression to the mean. We're shrinking the biggest overestimate from 3 samples means and inflating the biggest underestimate, roughly speaking. When $k = 1$, just one estimand, it's either an overestimate or an underestimate, with equal probability. When $k = 2$, there is an equal chance of (a) one overestimate and one underestimate, cancelling each other nicely, or (b) an imbalance of two underestimates or two overestimates that don't cancel. When $k \geq 3$, we can expect cancellation on average.

Fama Portfolios

I never understood before why in finance studies they start by putting stocks into "portfolios" before doing their regressions, as in the famous paper Fama & MacBeth (1973) . Finance economists say they do this to reduce variance, but it looked to me like they were doing this by throwing away information and it must be a misleading trick. After all, the underlying stock price movements are extremely noisy, even if the portfolios aren't, and the aim is to find out something about stock prices. Why not do a regression with bigger n by making the corporation the individual observation instead of the portfolio?

Here, I think, we may have the answer. Fama probably should have made a correction to his results for the fact that he was using portfolios, not individual stocks, since he wanted to apply his estimates to individual stocks in the end. But what he was doing was using the unequal-average estimator. The portfolio average over 20 stocks is really the unequal-average estimator for *each* stock. It is biased, because each stock is different, but it does cut down the variance a lot. And so for estimating something about 100's of stocks, where only the total error matters and we don't care about individual stocks, he did the right thing.[1]

Two Ideas

1. Shrink if variance is high relative to the mean, to reduce mean squared error.

2. Combine info from three unrelated estimands because regression to the mean will help us— their errors will "cancel out".

V. The James-Stein Estimator for k Means, Variances Identical and Known

1. "Stein's Paradox", from Stein (1956), is that there exists an estimator with lower mean squared error than $\overline{y}$ if $k \geq 3$ whatever values μ might take.
2. The "James-Stein estimator" of James & Stein (1961) describes a particular estimator.
3. "Stein's Lemma" from Stein (1974, 1981) makes it easier to show that the James-Stein estimator has lower MSE than $\overline{y}$.

For $k = 3$ and $n = 1$ and known homogeneous variance σ^2,

$$\boxed{\hat{\mu}_{y,JS} \equiv y - \left(\frac{\sigma^2}{w^2+y^2+z^2}\right)y} \tag{11}$$

$$\boxed{MSE(JS, total) = 3\sigma^2 - (k-2)^2\sigma^4\left[E\frac{1}{w^2+y^2+z^2}\right]} \tag{12}$$

The James-Stein Estimator: What's Really Going On?

Compare the JS shrinkage with the oracle estimator:

$$\hat{\mu}_{JS,y} = \Big(1 - \frac{(k-2)\sigma^2}{w^2 + y^2 + z^2}\Big)y \tag{13}$$

$$\hat{\mu}_{oracle,y} = \Big(1 - \frac{\sigma^2}{\sigma^2 + \mu_y^2}\Big)y \tag{14}$$

It happens that

$$Ey^2 \;=\; \mu_y^2 + \sigma^2 + 0 \tag{15}$$

Thus, another way to write the optimal oracle estimator for the $k = 1$ case is

$$\hat{\mu}_{oracle,y} = \Big(1 - \frac{(k-2)\sigma^2}{Ey^2}\Big)y \tag{16}$$

The analog of the oracle estimator is

$$\hat{\mu}_{oracle,y;w,z} = \Big(1 - \frac{(k-2)\sigma^2}{E(w^2 + y^2 + z^2)}\Big)y \tag{17}$$

Why the $k-2$ Correction?

We need the $(k-2)$ correction because of the bias in the shrinkage being correlated with the bias in $\overline{y}$. Think of there being k variances combined in the denominator. So if $\overline{y}$ is combined with 2 other parameters, we need to multiply the shrinkage amount by 1/3. If with 4, by 1/2. If with 5, by 3/5. If with 6, by 2/3. If by 20, by 9/10. If with 1, then by 0/2.

Regression to the Mean

Really, JS is just using regression to the mean. Suppose we knew that $\mu_w = \mu_y = \mu_z$ Then we'd have just 1 value to estimate. We could use the mean instead of 0 as the level to which ot shrink, and that would work better— would be optimal in fact (we can find the first-best here because we're in effect back to $k = 1$). But let's stick with shrinking to zero. Then, the biggest variable's estimate won't be shrunk down from its observation enough. All three variables are shrunk the same percentage. The smallest shouldn't be shrunk at all, but it is. Since it's smallest, though, and its percentage shrinkage is the same, its absolute shrinkage is the smallest. So what we've got is an estimator that shrinks the small observations less and the big observations more— just what we want.

Equal Estimands

Of course, when $\mu_w = \mu_y = \mu_z$ we will end up with an overall improvement, since that's true of the James-Stein estimator even when the true means aren't equal. But it works out even better. The mean squared error back in equation (**??**) for just Y was

$$\begin{aligned} MSE_{JS,y} &= \sigma^2 + (k-2)\sigma^4\left(kE\frac{y^2}{(w^2+y^2+z^2)^2} - 2E\frac{w^2+z^2}{(w^2+y^2+z^2)^2}\right) \\ &= \sigma^2 + \sigma^4\left(3E\frac{y^2}{(w^2+y^2+z^2)^2} - 2E\frac{w^2}{(w^2+y^2+z^2)^2} - 2E\frac{z^2}{(w^2+y^2+z^2)^2}\right) \end{aligned} \tag{18}$$

As I said then, we can't tell if (18) is bigger than σ^2 or not, even though when we add it to the mean squared errors for W and Z we can tell the sum is less than $3\sigma^2$. But suppose $\mu_w = \mu_y = \mu_z$. Then

$$\begin{aligned} MSE_{JS,y} &= \sigma^2 + \sigma^4\left(3E\frac{y^2}{(w^2+y^2+z^2)^2} - 2E\frac{y^2}{(w^2+y^2+z^2)^2} - 2E\frac{y^2}{(w^2+y^2+z^2)^2}\right) \\ &= \sigma^2 - \sigma^4 E\frac{y^2}{(w^2+y^2+z^2)^2} \end{aligned} \tag{19}$$

Equation (19) tells us that the mean squared error for *each* estimand is lower with James-Stein than with *ybar* if the true population means are equal. And that means that it will be lower for each estimand if the true population means are fairly close to

VI. The Full James-Stein Estimator for k Means, Variances Not Identical, But Known

This turns out not to be as hard a case as you might think. There's a trick we can use. Suppose we have three estimands, each with a separate known variance, $\sigma_w^2, \sigma_y^2, \sigma_z^2$. Before we start the estimation, transform the variables so they have identical variances, all equal to one. We can do that by using $\frac{y_i}{\sigma_y}$ instead of y_i. Now, all the variances are equal, so we can use the plain old James-Stein estimator. At the end, untransform the estimator so we have a number we can multiply by the original, untransformed, data.

The three transformed variables will each have a different mean but all will have the same variance, $\sigma_2 = 1$. Thus we're back to our old case of equal variances. Because we've used this trick, we are still shrinking each estimator the same amount, even though in this case it would seem to make sense to shrink $\overline{y}$ more than $\overline{z}$ if $\sigma_y^2 > \sigma_z^2$. Maybe the transformation process does that somehow, though. I do see that if $\sigma_z^2 = 0$, the transformation breaks down because it requires dividing by zero.

VII. The Full James-Stein Estimator for k Means, Variances Identical or Not, and Also Needing To Be Estimated

The James-Stein estimator (with $k = 3$ in this case) will turn out to be, for the case of equal unknown variances and equal sample sizes,

$$\boxed{\mu_{y,JS} \equiv \overline{y} - \left(\frac{n-1}{n+1}\right)(k-2)\left(\frac{\hat{\sigma}^2}{\overline{y}^2+\overline{w}^2+\overline{z}^2}\right)\overline{y},} \tag{20}$$

$$MSE_{JS,y} \;=\; \sigma_y^2 + \gamma\sigma_y^4\left(\gamma + \frac{2\gamma}{n_y-1} + 2\right)E\frac{\overline{y}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2} - 2\gamma\sigma_y^4 E\frac{\overline{w}^2+\overline{z}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2} \tag{21}$$

This shows why we need more than one estimand to get the James-Stein estimator to work. It would be nice if we could find a value for γ that would make the second term of this MSE negative. We can't, though— there is no way to pick γ so that $(\gamma + \frac{2\gamma}{n_y-1} + 2) < 0$. On the other hand, there's that third term, which we get by having $\overline{z}$ and $\overline{w}$ in the problem. It's negative, so we can hope it would outweigh the first two terms.

Full MSE with equal unknown variances

$$
\begin{aligned}
MSE(JS, total) \;=\; & \left(\sigma_y^2 + \sigma_z^2 + \sigma_w^2\right) \\
& + \gamma\sigma_y^4\left[\left(\gamma + \tfrac{2\gamma}{n_y - 1}\right) E\tfrac{\overline{y}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2} + 2E\tfrac{\overline{y}^2-\overline{w}^2-\overline{z}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2}\right] + \\
& + \gamma\sigma_z^4\left[\left(\gamma + \tfrac{2\gamma}{n_z - 1}\right) E\tfrac{\overline{z}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2} + 2E\tfrac{\overline{z}^2-\overline{w}^2-\overline{y}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2}\right] + \\
& + \gamma\sigma_w^4\left[\left(\gamma + \tfrac{2\gamma}{n_w - 1}\right) E\tfrac{\overline{w}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2} + 2E\tfrac{\overline{w}^2-\overline{y}^2-\overline{z}^2}{(\overline{y}^2+\overline{w}^2+\overline{z}^2)^2}\right]
\end{aligned}
\tag{22}
$$

This expression looks hopeful. We have a lot of negative numbers in the "+2E(jkjkljl)" terms— more negatives than positives in each numerator. And positive terms like $\frac{2\gamma}{n_z-1}$ will get small as our sample size rises above $n = 2$. But there's a fatal problem. We can't cancel out across the $\overline{y}$, $\overline{z}$, and $\overline{w}$ expressions, because $\sigma_w^4 \neq \sigma_z^4 \neq \sigma_z^4$.

More correctly, those variances *might* not be equal, so we can't count on that. I wish we had a symbol for "is not necessarily equal to but it might happen to be equal to."

A Special Case

Think about what happens if σ_w^2, σ_z^2, μ_w, and μ_z are very small, and $n_y = 2$ so that $\frac{2\gamma}{n_y - 1}$ is big. The third and fourth lines of (22) are now small, and

$$\begin{aligned} MSE(JS, total) &\approx \sigma_y^2 + \gamma\sigma_y^4 \left[\left(\gamma + \tfrac{2\gamma}{2-1} \right) E\frac{\overline{y}^2}{\overline{y}^2} + 2E\frac{\overline{y}^2}{(\overline{y}^2)^2} \right] \\ &\approx \sigma_y^2 + \gamma\sigma_y^4 \left[3\gamma + 2E\frac{1}{\overline{y}^2} \right] \end{aligned} \tag{23}$$

There is no γ that can make this MSE smaller than $\sigma_y^2 + \sigma_z^2 + \sigma_w^2$. Since only $\overline{y}$ is important, we can't trade off likely errors in one estimand against likely errors in another. Thus, we do need the assumption of equal variances if the variances are unknown. Without it, we're effectively back in the $k = 1$ case.

Shrinking towards the Unrelated Average

Let's try, for $k = 3$ and $n = 1$ and known homogeneous variance σ^2, the more general James-Stein estimator:

$$\boxed{\hat{\mu}_y \equiv y + (k-2)\sigma^2\left(\frac{1}{w^2+y^2+z^2}\right)(A-y)} \tag{24}$$

where we will consider two possibilities, $A = w$ and $A = \frac{w+y+z}{3}$.

Define

$$g(y) \equiv (k-2)\sigma^2\left(\frac{1}{w^2+y^2+z^2}\right)(A-y) \tag{25}$$

with derivative

$$\frac{dg}{dy} = (k-2)\sigma^2\left(\frac{\frac{dA}{dy}-1}{w^2+y^2+z^2} - \frac{2y(A-y)}{(w^2+y^2+z^2)^2}\right) \tag{26}$$

The mean squared error is

$$\begin{aligned} MSE_y &= E\Big(y + g(y) - \mu_y\Big)^2 \\ &= E\Big((y-\mu_y) + g(y)\Big)^2 \end{aligned} \tag{27}$$

Stein's Lemma implies that for Y distributed $N(\mu_y, \sigma^2)$,

$$E\Big(g(y)(y-\mu_y)\Big) = \sigma^2 E\frac{dg}{dy}. \tag{28}$$

so we get

$$\begin{aligned} MSE_y &= \sigma^2 + E(k-2)^2\sigma^4\left(\tfrac{A-y}{w^2+y^2+z^2}\right)^2 + 2\sigma^2 E\left[(k-2)\sigma^2\left(\tfrac{\frac{dA}{dy}-1}{w^2+y^2+z^2} - \tfrac{2y(A-y)}{(w^2+y^2+z^2)^2}\right)\right] \\ &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{A^2+y^2-2Ay}{(w^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{(\frac{dA}{dy}-1)(w^2+y^2+z^2)-2Ay+2y^2}{(w^2+y^2+z^2)^2} \end{aligned} \tag{29}$$

Now let's introduce $A = w$, so we get

$$\begin{aligned} MSE_y &= \sigma^2 + (k-2)^2\sigma^4 E\tfrac{w^2+y^2-2wy}{(w^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\tfrac{(0-1)(w^2+y^2+z^2)-2wy+2y^2}{(w^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)^2\sigma^4\left(E\tfrac{w^2+y^2-2wy}{(w^2+y^2+z^2)^2} + E\tfrac{-2w^2-2y^2-2z^2-4wy+4y^2}{(w^2+y^2+z^2)^2}\right) \\ &= \sigma^2 + (k-2)^2\sigma^4 E\left(\tfrac{w^2+y^2-2wy-2w^2-2y^2-2z^2-4wy+4y^2}{(w^2+y^2+z^2)^2}\right) \\ &= \sigma^2 + (k-2)^2\sigma^4 E\left(\tfrac{-w^2+3y^2-6wy-z^2}{(w^2+y^2+z^2)^2}\right) \end{aligned} \tag{30}$$

adding up the three we get

$$\begin{aligned} MSE(total) &= \sigma^2 \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{-w^2+3y^2-6wy-z^2}{(w^2+y^2+z^2)^2}\right) \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{-w^2+3z^2-6wz-y^2}{(w^2+y^2+z^2)^2}\right) \qquad (31) \\ &= 3\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{2y^2+2z^2-2w^2-6wy-6wz}{(w^2+y^2+z^2)^2}\right) \end{aligned}$$

Not much use.

Now let's introduce $A = \frac{w+y+z}{3}$, so we get

$$\begin{aligned} MSE_y &= \sigma^2 + (k-2)^2\sigma^4 E\frac{A^2+y^2-2Ay}{(w^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\frac{(\frac{dA}{dy}-1)(w^2+y^2+z^2)-2Ay+2y^2}{(w^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)^2\sigma^4 E\frac{A^2+y^2-2Ay}{(w^2+y^2+z^2)^2} + 2(k-2)\sigma^4 E\frac{(\frac{1}{3}-1)(w^2+y^2+z^2)-2Ay+2y^2}{(w^2+y^2+z^2)^2} \\ &= \sigma^2 + (k-2)^2\sigma^4\left(E\frac{A^2+y^2-2Ay}{(w^2+y^2+z^2)^2} + E\frac{-\frac{4}{3}(w^2+y^2+z^2)-4Ay+4y^2}{(w^2+y^2+z^2)^2}\right) \\ &= \sigma^2 + (k-2)^2\sigma^4 E\left(\frac{A^2-6Ay+\frac{13}{3}y^2-\frac{4}{3}w^2-\frac{4}{3}z^2}{(w^2+y^2+z^2)^2}\right) \end{aligned}$$

Adding up the three estimands MSE's, we get

$$\begin{aligned} MSE(total) &= \sigma^2 + (k-2)^2\sigma^4 E\left(\frac{A^2-6Aw+\frac{13}{3}w^2-\frac{4}{3}y^2-\frac{4}{3}z^2}{(w^2+y^2+z^2)^2}\right) \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{A^2-6Ay+\frac{13}{3}y^2-\frac{4}{3}w^2-\frac{4}{3}z^2}{(w^2+y^2+z^2)^2}\right) \\ &\quad +\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{A^2-6Az+\frac{13}{3}z^2-\frac{4}{3}w^2-\frac{4}{3}y^2}{(w^2+y^2+z^2)^2}\right) \end{aligned}$$

$$\begin{aligned} &= 3\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{3A^2-6A(w+y+z)+\frac{5}{3}w^2+\frac{5}{3}y^2+\frac{5}{3}z^2}{(w^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{3A^2-6A(w+y+z)+\frac{5}{3}w^2+\frac{5}{3}y^2+\frac{5}{3}z^2}{(w^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{\frac{(w+y+z)^2}{3}-2(w+y+z)^2+\frac{5}{3}w^2+\frac{5}{3}y^2+\frac{5}{3}z^2}{(w^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 + (k-2)^2\sigma^4 E\left(\frac{-\frac{5}{3}(w+y+z)^2+\frac{5}{3}(w^2+y^2+z^2)}{(w^2+y^2+z^2)^2}\right) \\ &= 3\sigma^2 - (k-2)^2\sigma^4 E\left(\frac{\frac{5}{3}(wy+yz+wz)}{(w^2+y^2+z^2)^2}\right) \end{aligned} \tag{33}$$

This is better MSE than the sample means, to be sure.

This comapres with

$$MSE(total, JS) = 3\sigma^2 - (k-2)^2\sigma^4 E\left(\frac{w^2+y^2+z^2}{(w^2+y^2+z^2)^2}\right) \tag{34}$$

The JS MSE looks like it would usually but now always be lower. It would be higher if, for example W, Y, Z and we are in the $k = 1, n = 3$ non-independent draws case.

$k = 1, n = 3$, independent draw case, which doesn't look likely here— so what is going on? Ah—not enough care to how far to shrink, maybe.

How about stretching each estimand towards the average of the other two?