1. Optimality of the AHL Estimator

Consider the problem of estimating the mean of a random variable $X \sim F_X(\mu; \sigma^2)$ using two observations X_1 and X_2. With probability π, X_1 is drawn from the true distribution F_X and X_2 is noise drawn from the distribution $F_Y(\kappa, \omega^2)$. With probability $1 - \pi$, X_2 is drawn from the correct distribution and X_1 is noise. Under this specification, exactly one of X_1 or X_2 is drawn from the distribution of interest at all times.

Observe that if π is known, we can construct an unbiased estimator using only X_1,

$$\hat{\mu}_1 = \frac{X_1}{\pi} - \frac{1-\pi}{\pi}\kappa \tag{1}$$

Similarly, we can construct an unbiased estimator using only X_2,

$$\hat{\mu_2} = \frac{X_2}{1-\pi} - \frac{\pi}{1-\pi}\kappa \tag{2}$$

Compare these to an estimator that uses both X_1 and X_2,

$$\hat{\mu} = a_1 X_1 + a_2 X_2 - a_3 \kappa \tag{3}$$

which has the following expectation,

$$E[\hat{\mu}] = (a_1\pi + a_2(1-\pi))\mu + (a_1(1-\pi) + a_2\pi - a_3)\kappa$$

so that unbiased, requires

$$a_1\pi + a_2(1-\pi) = 1 \implies a_2(a_1) = \frac{1}{1-\pi} - \frac{a_1\pi}{1-\pi} \tag{4}$$

$$a_1(1-\pi) + a_2\pi = a_3 \implies a_3(a_1) = \frac{\pi}{1-\pi} + \frac{a_1 - 2a_1\pi}{1-\pi} \tag{5}$$

Hence we can write $\hat{\mu}$ as a function of a_1,

$$\hat{\mu}(a_1) = a_1 X_1 + \left(\frac{1}{1-\pi} - \frac{a_1\pi}{1-\pi}\right) X_2 - \left(\frac{\pi}{1-\pi} + \frac{a_1 - 2a_1\pi}{1-\pi}\right)\kappa$$

When $a_1 = \frac{1}{\pi}$, then $\hat{\mu} = \hat{\mu}_1$; and if $a_1 = 0$ then $\hat{\mu} = \hat{\mu_2}$.

We can write:

$$\hat{\mu} = (a_1\pi)\hat{\mu}_1 + (1 - a_1\pi)\hat{\mu}_2 + a_1(1-\pi)\kappa + \frac{\pi}{1-\pi}(1-a_1)\kappa - \left(\frac{\pi}{1-\pi} + \frac{a_1 - 2a_1\pi}{1-\pi}\right)\kappa$$
$$= (a_1\pi)\hat{\mu}_1 + (1 - a_1\pi)\hat{\mu}_2 - (a_1\pi)\kappa$$

Hence any unbiased estimator $\hat{\mu}$ that uses X_1 and X_2 can be written as a linear combination of estimators using only X_1 or X_2. The problem of finding the minimum variance, unbiased estimator $\hat{\mu}$ reduces to finding d^* that solves

$$\min_d \ \mathrm{Var}\left(d\hat{\mu}_1 + (1-d)\hat{\mu}_2\right)$$

which is solved by $d^* = 0$ or $d^* = 1$ depending on whether $\mathrm{Var}(\hat{\mu}_1)$ or $\mathrm{Var}(\hat{\mu}_2)$ is smaller.

I now show that $\mathrm{Var}(\hat{\mu}_1) > \mathrm{Var}(\hat{\mu}_2)$, without loss of generality, except when $\pi = 0.5$, or $\sigma^2 = \omega^2 = (\mu - \kappa)^2$. Observe that,

$$\mathrm{Var}(\hat{\mu}_1) = \frac{\mathrm{Var}(X_1)}{\pi^2} = \frac{1}{\pi^2}\left(\pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right)$$
$$\mathrm{Var}(\hat{\mu}_2) = \frac{\mathrm{Var}(X_2)}{(1-\pi)^2} = \frac{1}{(1-\pi)^2}\left((1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right)$$

This follows from the law of total variance, with the random variable $D = 1$ if X_1 is drawn from the correct distribution (and X_2 is drawn from the incorrect distribution), and $D = 0$ otherwise.

$$\mathrm{Var}(X_1) = E[\mathrm{Var}(X_1|D)] + \mathrm{Var}(E[X_1|D])$$
$$= P(D=1)\sigma^2 + P(D=0)\omega^2 + \mathrm{Var}(\mu D + \kappa(1-D))$$
$$= \pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2$$

Similarly,

$$\mathrm{Var}(X_2) = (1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu-\kappa)^2$$

Thus, $\text{Var}(\hat{\mu}_1)$ and $\text{Var}(\hat{\mu}_2)$ can be written as functions of σ^2, ω^2, and $(\mu-\kappa)^2$,

$$g(\sigma^2, \omega^2, (\mu-\kappa)^2, x) \equiv \frac{1}{x^2}\left(x\sigma^2 + (1-x)\omega^2 + x(1-x)(\mu-\kappa)^2\right)$$

$$\text{Var}(\hat{\mu}_1) = g(\sigma^2, \omega^2, (\mu-\kappa)^2, \pi)$$

$$\text{Var}(\hat{\mu}_2) = g(\sigma^2, \omega^2, (\mu-\kappa)^2, 1-\pi)$$

Importantly,

$$\frac{\partial g(\sigma^2, \omega^2, (\mu-\kappa)^2, x)}{\partial x} = \frac{\omega^2(x-2) - x(\sigma^2 + (\mu-\kappa)^2)}{x^3} < 0, \;\; x \in (0,1)$$

and so $\text{Var}(\hat{\mu}_\ell)$ is strictly decreasing in the probability that the observation ℓ is drawn from the correct distribution, and $\text{Var}(\hat{\mu}_1) \neq \text{Var}(\hat{\mu}_2)$ unless $\pi = 0.5$. Thus, the minimum variance unbiased estimator is equal to $\hat{\mu}_\ell$ for the observation ℓ that has the highest probability of being correct.

The above result holds also for L observations $X_1, \ldots, X_L$ with corresponding probabilities $\pi_1, \ldots, \pi_L$; that is, the minimum variance unbiased estimator $\hat{\mu}$ will use only X_ℓ with the highest π_ℓ and apply inverse probability weighting.

Now consider a sample of N sets of i.i.d. observations, $\left\{\{X_{i\ell}\}_{\ell=1}^{L_i}\right\}_{i=1}^{N}$. If the values

$$\pi_{i\ell} = \Pr(X_{i\ell} \text{ is drawn from the correct distribution}), \sum_{\ell=1}^{L_i} \pi_{i\ell} = 1$$

are known for all i, then the optimal estimator is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i = \frac{1}{N}\sum_{i=1}^{N} \frac{X_{i\ell_i}}{\pi_{i\ell_i}} - \frac{1-\pi_{i\ell_i}}{\pi_{i\ell_i}}\kappa$$

where $\hat{\mu}_i$ is the minimum variance estimator constructed for observations $\{X_{i\ell}\}_{\ell=1}^{L_i}$.

Since $\hat{\mu}$ is an inverse probability weighting estimator, small values of $\pi_{i\ell}$ may be detrimental for its finite sample performance. Rather than dropping observations whose maximal $\pi_{i\ell}$ is small, however, it is possible to give these observations equal weights for all $\{X_{i\ell}\}$, as in the AHL estimator. I hypothesize that if $\pi_{i\ell} < 0.5$ (or another threshold) for all ℓ, then it is better to give equal weights to all $X_{i\ell}$ associated with i, even when the $\pi_{i\ell}$ are known.

In practice, $\pi_{i\ell}$ needs to be estimated. If $\hat{\pi}_{i\ell}$ is imprecise, the bias may potentially be very large. This can be seen via simulation. By contrast, the AHL estimator does not require knowledge about $\pi_{i\ell}$, and has (possibly optimal) worst case performance in terms of MSE.