Stein's Paradox

We have seen that the MLE is asymptotically unbiased, consistent, and asymptotically efficient, in that it reached the Cramer-Rao lower bound as the amount of observed data grows.

The Cramer-Rao bound is somehow fundamental, but does it tell the whole story? It provides a unbeatable lower-bound on mean-square error for any unbiased estimator. Is it possible, however, that being biased can actually help us? Are there biased estimators that perform uniformly better that unbiased ones?

As you can probably guess from the fact that we are even posing these questions that the answers are: no, yes, yes.

A strikingly simple example of this was provided by Charles Stein and W. James in the late 1950s. Their problem scenario is completely benign. We observe a Gaussian random vector of length D whose entries are independent, but have different means:

$$X \sim \text{Normal}(\boldsymbol{\theta}_\star, \mathbf{I}), \quad \boldsymbol{\theta}_\star = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_D \end{bmatrix}.$$

Given a single observation $X = \boldsymbol{x}$, what is the best estimate for $\boldsymbol{\theta}_\star$?

The obvious answer is the MLE, which in this case is also the sample mean (though there is only one sample):

$$\hat{\boldsymbol{\Theta}}_{\text{mle}} = X. \tag{1}$$

The log likelihood function when $X = \boldsymbol{x}$ in this case is

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \text{Constant} - \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{x}\|_2^2$$

which is maximized at $\boldsymbol{x}$. It is also clear that the error is itself normally distributed,

$$\hat{\boldsymbol{\Theta}}_{\text{mle}} - \boldsymbol{\theta}_\star \sim \text{Normal}(\mathbf{0}, \mathbf{I}),$$

and so the estimator is unbiased, $\mathrm{E}[\hat{\boldsymbol{\theta}}_{\text{mle}}] = \boldsymbol{\theta}_\star$, with MSE

$$\mathrm{E}[\|\hat{\boldsymbol{\Theta}}_{\text{mle}} - \boldsymbol{\theta}_\star\|_2^2]] = \text{trace}(\mathbf{I}) = D.$$

A quick calculation also shows that the Fisher information matrix is also $\boldsymbol{J}(\boldsymbol{\theta}_\star) = \mathbf{I}$, and so $\hat{\boldsymbol{\Theta}}_{\text{mle}}$ is indeed the best unbiased estimator.

Stein rocked the world of statistics when he showed that there was an estimator that was better than $\boldsymbol{\Theta}_{\text{mle}}$ *uniformly* over all $\boldsymbol{\theta}_\star$ when $D \geq 3$, then along with James he formulated exactly what such an estimator looks like. Here is what they came up with:

$$\hat{\boldsymbol{\Theta}}_{\text{js}} = \left(1 - \frac{D-2}{\|X\|^2}\right) X. \tag{2}$$

It is clear that this estimator is biased; in fact, it is a *shrinkage* operator that pulls X back towards the origin (you can see immediately that the multiplier in front will usually have magnitude less than 1).

Let's calculate the MSE of (2). We have

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{\Theta}_{\text{js}} - \boldsymbol{\theta}_\star\|_2^2] &= \left\|X - \boldsymbol{\theta}_\star - \frac{(D-2)X}{\|X\|_2^2}\right\|_2^2 \\ &= D - 2(D-2)\sum_{d=1}^{D} \mathrm{E}\left[\frac{X_d(X_d - \theta_d)}{\|X\|_2^2}\right] + (D-2)^2\, \mathrm{E}\left[\frac{1}{\|X\|_2^2}\right]\end{aligned}$$

In the Technical Details section below, we show that

$$\sum_{d=1}^{D} \mathrm{E}\left[\frac{X_d(X_d-\theta_d)}{\|X\|_2^4}\right] = (D-2)\,\mathrm{E}\left[\frac{1}{\|X\|_2^2}\right],$$

and so

$$\mathrm{E}[\|\boldsymbol{\Theta}_{\mathrm{js}} - \boldsymbol{\theta}_\star\|_2^2] = D - (D-2)^2\,\mathrm{E}\left[\frac{1}{\|X\|_2^2}\right].$$

Thus for $D \geq 3$, the term on the right is strictly positive, and

$$\mathrm{E}[\|\boldsymbol{\Theta}_{\mathrm{js}} - \boldsymbol{\theta}_\star\|_2^2] \;<\; \mathrm{E}[\|\boldsymbol{\Theta}_{\mathrm{mle}} - \boldsymbol{\theta}_\star\|_2^2]$$

A small modification to the above results in an even better estimator; we simply make sure the multiplier is positive,

$$\hat{\boldsymbol{\Theta}}_{\mathrm{js+}} = \left(1 - \frac{D-2}{\|X\|^2}\right)_+ X \tag{3}$$

where $(q)_+ = \max(q,0)$ is the "positive part" of q. We will not prove this here, but it has been shown that

$$\mathrm{E}[\|\boldsymbol{\Theta}_{\mathrm{js+}} - \boldsymbol{\theta}_\star\|_2^2] \;\leq\; \mathrm{E}[\|\boldsymbol{\Theta}_{\mathrm{js}} - \boldsymbol{\theta}_\star\|_2^2],$$

with the inequality being strict for at least one $\boldsymbol{\theta}_\star \in \mathbb{R}^D$.

Notes:

- Notice that even though each of the components is independent, (2) provides a joint estimation through the $\|X\|^2$ term. It is a little counter-intuitive that we can do better by coupling the estimates, even when the entries are independent, but this is not the only place that something like this occurs in statistics.
- We have shown that (2) is better than (1) for every $\boldsymbol{\theta}_\star$. But this does not mean that (2) is optimal; indeed, (3) uniformly outperforms (2) (we didn't show this, but it is true). It is also know that this (3) estimator can again be outperformed uniformly.
- There are two ideas in this work that were extremely influential in latter 20th century statistics. The first is that introducing bias can actually help you. The other is that "shriking" the observations in one way or another often times introduces this bias at a much lower rate than it decreases the variance.
- Two great resources for the material in this section, along with very good qualitative commentary and historical perspective, can be found in [EM77],[Sam12]. These papers will be posted on the course website.

Technical Details

Lemma. Let $X \sim \mathrm{Normal}(\boldsymbol{\theta}, \mathbf{I})$ be a Gaussian random vector in $\mathbb{R}^D$ with entries $X_1, \ldots, X_D$. Then

$$\mathrm{E}\left[\frac{X_i(X_i-\theta_i)}{\|X\|_2^2}\right] = \mathrm{E}\left[\frac{\|X\|_2^2 - 2X_i^2}{\|X\|_2^4}\right],$$

and so

$$\sum_{d=1}^{D} \mathrm{E}\left[\frac{X_i(X_i-\theta_i)}{\|X\|_2^4}\right] = (D-2)\,\mathrm{E}\left[\frac{1}{\|X\|_2^2}\right].$$

Proof: Start with $i = 1$, and write out the integral to compute the expectation

$$\mathrm{E}\left[\frac{X_1(X_1-\theta_i)}{\|X\|_2^2}\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{x_1(x_1-\theta_1)}{\|\boldsymbol{x}\|_2^2}\,(2\pi)^{-D/2} e^{-\|\boldsymbol{x}-\boldsymbol{\theta}\|_2^2/2}\ \mathrm{d}x_1\cdots\mathrm{d}x_D,$$

and then compute the inner integral by parts:

$$\begin{aligned}&\int_{-\infty}^{\infty} \frac{x_1(x_1-\theta_1)}{\|\boldsymbol{x}\|_2^2}\,(2\pi)^{-D/2} e^{-\|\boldsymbol{x}-\boldsymbol{\theta}\|_2^2/2}\ \mathrm{d}x_1 \\ &= (2\pi)^{-D/2} e^{-\sum_{d=2}^{D}(x_d-\theta_d)^2/2} \int_{-\infty}^{\infty} \frac{x_1(x_1-\theta_1)}{\|\boldsymbol{x}\|_2^2} e^{-(x_1-\theta_1)^2/2}\,\mathrm{d}x_1,\end{aligned}$$

With $c = \sum_{d=2}^{D} x_d^2$, take

$$u(x_1) = \frac{x_1}{x_1^2+c}, \Rightarrow u'(x_1) = \frac{c - x_1^2}{(c+x_1^2)^2} = \frac{\|\boldsymbol{x}\|_2^2 - 2x_1^2}{\|\boldsymbol{x}\|_2^4}$$

$$v'(x_1) = (x_1-\theta_1)e^{-(x_1-\theta_1)^2/2} \Rightarrow v(x_1) = -e^{-(x_1-\theta_1)^2/2}.$$

Then the integral above is

$$\int_{-\infty}^{\infty} u(x)v'(x)\,\mathrm{d}x = [u(x_1)v(x_1)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} u'(x_1)v(x_1)\ \mathrm{d}x_1$$
$$= \int_{-\infty}^{\infty} \frac{\|\boldsymbol{x}\|_2^2 - 2x_1^2}{\|\boldsymbol{x}\|_2^4} e^{-(x_1-\theta_1)^2/2}\ \mathrm{d}x_1$$

where we have used the fact that

$$-\frac{x_1}{x_1^2+c}e^{-(x_1-\theta_1)^2/2} \to 0, \quad \text{as} \ \ x_1 \to \pm\infty.$$

Thus

$$\mathrm{E}\left[\frac{X_1(X_1-\theta_i)}{\|X\|_2^2}\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\|\boldsymbol{x}\|_2^2 - 2x_1^2}{\|\boldsymbol{x}\|_2^4}\,(2\pi)^{-D/2}e^{-\|\boldsymbol{x}-\boldsymbol{\theta}\|_2^2/2}\ \mathrm{d}x_1\cdots\mathrm{d}x_D$$
$$= \mathrm{E}\left[\frac{\|X\|^2 - 2X_1^2}{\|X\|_2^4}\right].$$

The lemma follows from doing the same integration by parts for each of the $i = 2, \ldots, D$ in turn. ■

References

[EM77] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.

[Sam12] R. Samworth. Stein's paradox. *Eureka*, 62:38–41, 2012.