Bayesians: Pay No Attention to Neyman-Scott

William H. Press

March 14, 2016

Maximum likelihood estimates (MLEs) are not necessarily *consistent*, meaning that they do not necessarily converge to the value of the underlying parameter being estimated, even in the limit of an infinite amount of data. A particularly clear example is the so-called Neyman-Scott Paradox (J. Neyman and E.L. Scott, *Econometrica*, **16**, 1, 1948).

Consider pairs of i.i.d. random deviates (x_i, y_i), $i = 1, \ldots, N$,

$$x_i \sim N(\mu_i, \sigma^2), \qquad y_i \sim N(\mu_i, \sigma^2) \tag{1}$$

That is, each pair (x_i, y_i) has its own mean μ_i, but all the pairs share a common variance σ^2. The plethora of μ_i's are to be viewed as uninteresting nuisance parameters. Our goal is to estimate the common σ^2 using all the data, that is, all N pairs of values.

The likelihood function is

$$\mathcal{L} = \frac{1}{(2\pi)^N \sigma^{2N}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i \left[(x_i - \mu_i)^2 + (y_i - \mu_i)^2 \right] \right\} \tag{2}$$

The log-likelihood is thus

$$\ln \mathcal{L} = -N \ln(2\pi) - N \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_i \left[(x_i - \mu_i)^2 + (y_i - \mu_i)^2 \right] \tag{3}$$

We can obtain the MLE by setting the derivatives of the log-likelihood with respect σ^2 and all of the μ_i's to zero, yielding the equations

$$0 = \frac{\partial \ln \mathcal{L}}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \frac{1}{2\sigma^4} \sum_i \left[(x_i - \mu_i)^2 + (y_i - \mu_i)^2 \right] \tag{4}$$

$$0 = \frac{\partial \ln \mathcal{L}}{\partial \mu_i} = \frac{x_i + y_i - 2\mu_i}{\sigma^2} \tag{5}$$

whose simultaneous solution yields the MLE estimators

$$\widehat{\sigma^2} = \frac{1}{4}\left\langle (X-Y)^2 \right\rangle, \qquad \widehat{\mu_i} = \tfrac{1}{2}(x_i + y_i) \tag{6}$$

where we have defined in the obvious way

$$\left\langle (X-Y)^2 \right\rangle \equiv \frac{1}{N}\sum_i (x_i - y_i)^2 \tag{7}$$

The problem with the first equation in (6) is that it is "wrong" by a factor of 2 no matter how large N is,

$$E[\widehat{\sigma^2}] = \frac{1}{4}E\left[\frac{1}{N}\sum_i (x_i - y_i)^2\right] = \frac{1}{4}(2\sigma^2) = \tfrac{1}{2}\sigma^2 \tag{8}$$

Thus, the MLE is not consistent. The problem is that the number of nuisance parameters μ_i grows with the data size N. The small-sample bias in the MLE for σ^2 from a single pair (x_i, y_i) is replicated in every such pair, rather than being averaged away asymptotically.

So much for MLE. How does a Bayesian approach the same problem? Maximum a posteriori (MAP) is the Bayesian's close analog of frequentist MLE. The analog of the likelihood, equation (2), is the posterior probability,

$$\begin{aligned} P(\{\mu_i\}, \sigma | \{x_i, y_i\}) \propto{}& \frac{1}{(2\pi)^N \sigma^{2N}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i \left[(x_i - \mu_i)^2 + (y_i - \mu_i)^2\right]\right\} \\ & \times P(\{\mu_i\}, \sigma^2) \end{aligned} \tag{9}$$

where $P(\{\mu_i\}, \sigma^2)$ is the prior on the parameters. Let's assume a uniform (i.e., non-informative) prior on all the μ_i's, and (for now) any desired prior $P(\sigma^2)$ on σ^2. Then we can marginalize over the nuisance parameters μ_i by

$$\begin{aligned} P(\sigma | \{x_i, y_i\}) &\propto \int\int \cdots \int P(\{\mu_i\}, \sigma | \{x_i, y_i\}) d\mu_1 d\mu_2 \cdots d\mu_N \\ &\propto \prod_i \int \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{2\sigma^2}\right] d\mu_i \times P(\sigma^2) \\ &\propto \prod_i \frac{1}{2\sigma\sqrt{\pi}} \exp\left[-\frac{(x_i - y_i)^2}{4\sigma^2}\right] \times P(\sigma^2) \end{aligned} \tag{10}$$

where the integral over each μ_i has been done analytically.

To see what is going on in equation (10), take its logarithm,

$$\ln P(\sigma|\{x_i, y_i\}) = \text{const} - N \ln \sigma - \frac{N}{4\sigma^2} \left\langle (X - Y)^2 \right\rangle + \ln P(\sigma^2) \tag{11}$$

As N becomes large, the assumed prior on σ^2 becomes negligible as compared with the the terms contining σ that scale as N, so that the prior becomes immaterial (as it should when there is a lot of informative data). The argmax of equation (11) is the MAP estimator,

$$\sigma^2 \equiv \sigma^2_{\text{MAP}} = \frac{1}{2} \left\langle (X - Y)^2 \right\rangle \tag{12}$$

Instead of equation (8), we have

$$E[\sigma^2_{\text{MAP}}] = \frac{1}{2} E[\left\langle (X - Y)^2 \right\rangle] = \frac{1}{2}(2\sigma^2) = \sigma^2 \tag{13}$$

so the estimator is consistent.

The point is simply that marginalizing to get a MAP estimator gives the "right" (i.e., consistent) answer, while maximizing the functionally identical likelihood for MLE gives the "wrong" (i.e., inconsistent) answer. The unbounded number of nuisance parameters μ_i is not a problem for Bayes. They are regularized by their priors—in this case even when the non-informative prior is improper. Were I to summarize as, "As always, Bayes is better," I might attract some angry responses, so, I won't say that.

If we were to put a nontrivially different prior on the μ_i's, either jointly or independently, we would get a different answer for σ^2_{MAP}. We should. Such a prior, by adding information about the values of the μ_i's, would also add information about the value of σ^2. This would be consistently reflected in the MAP estimator.