Anomalous Mean Scatter Memo

Michael Wilensky

January 2025

Suppose we have some data, d, that are i.i.d. Gaussian distributed with true mean, μ, and true variance σ^2_{true}. Suppose that we are interested in the unknown mean of this distribution, but for whatever the reason are convinced that the variance is exactly σ^2_{assume} not necessarily equal to σ^2_{true}. Suppose then we adopt a prior on the mean parameter that is Gaussian with some mean, μ_{prior}, and variance σ^2_{prior}. These assumptions imply that our posterior distribution for the mean is Normal:

$$\left(\mu | d, \sigma^2 = \sigma^2_{\text{assume}}\right) \sim \mathcal{N}\left(\mu_{\text{post}}, \sigma^2_{\text{post}}\right) \tag{1}$$

where N is the number of data, $\mathbb{1}$ is a vector of ones,

$$\sigma^2_{\text{post}} = \left(\frac{N}{\sigma^2_{\text{assume}}} + \frac{1}{\sigma^2_{\text{prior}}}\right)^{-1} \tag{2}$$

and

$$\mu_{\text{post}} = \sigma^2_{\text{post}} \left(\frac{\mathbb{1}^T d}{\sigma^2_{\text{assume}}} + \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}}\right) \tag{3}$$

If one conducted many such experiments, an interesting quantity would be, what is the scatter in the posterior means, and how does that compare to the width of the posterior distribution of any of the experiments. The latter of these two quantities is independent of the data draw, but the former is determined by the data in a simple way. Mathematically, this can be represented in terms of squared errors:

$$u^2 \equiv E\left[(\mu_{\text{post}} - \mu)^2\right]_{d|\mu, \sigma^2 = \sigma^2_{\text{true}}} \tag{4}$$

where the expectation is taken over the *true* data generating distribution.

Expanding Equation 4, we have

$$u^2 = E\left[\mu^2_{\text{post}}\right]_{d|\mu, \sigma^2 = \sigma^2_{\text{true}}} - 2\mu E\left[\mu_{\text{post}}\right]_{d|\mu, \sigma^2 = \sigma^2_{\text{true}}} + \mu^2 \tag{5}$$

The second term is easy, noting that $E[d]_{d|\mu, \sigma^2 = \sigma^2_{\text{true}}} = \mu \mathbb{1}$, and $\mathbb{1}^T \mathbb{1} = N$

$$E\left[\mu_{\text{post}}\right]_{d|\mu, \sigma^2 = \sigma^2_{\text{true}}} = \sigma^2_{\text{post}} \left(\frac{N\mu}{\sigma^2_{\text{assume}}} + \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}}\right). \tag{6}$$

The first term is a bit more complicated, but expanding in terms of the definition of $\mu_{\rm post}$, we have

$$\begin{aligned} E\left[\mu_{\rm post}^2\right]_{d|\mu,\sigma^2=\sigma^2_{\rm true}} = (\sigma^2_{\rm post})^2 \Bigg(& E\left[\frac{\mathbb{1}^T d d^T \mathbb{1}}{(\sigma^2_{\rm assume})^2}\right]_{d|\mu,\sigma^2=\sigma^2_{\rm true}} \\ & + 2\frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}} E\left[\frac{\mathbb{1}^T d}{\sigma^2_{\rm assume}}\right]_{d|\mu,\sigma^2=\sigma^2_{\rm true}} \\ & + \left(\frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}}\right)^2 \Bigg) \end{aligned} \tag{7}$$

We've already dealt with the second of these terms. To do the first term, note

$$E\left[dd^T\right]_{d|\mu,\sigma^2=\sigma^2_{\rm true}} = \sigma^2_{\rm true} I_N + \mu^2 \mathbb{1}\mathbb{1}^T, \tag{8}$$

where I_N is the identity matrix in N dimensions. Combining all this we get

$$u^2 = (\sigma^2_{\rm post})^2 \left(\frac{N\sigma^2_{\rm true}}{(\sigma^2_{\rm assume})^2} + \frac{N^2\mu^2}{(\sigma^2_{\rm assume})^2} + 2N\frac{\mu\mu_{\rm prior}}{\sigma^2_{\rm prior}\sigma^2_{\rm assume}} + \left(\frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}}\right)^2 \right) - 2\mu\sigma^2_{\rm post}\left(\frac{N\mu}{\sigma^2_{\rm assume}} + \frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}}\right) + \mu^2 \tag{9}$$

We can collect some terms here to clean things up. We can rewrite this as

$$u^2 = \left(\frac{\sigma^2_{\rm post}}{\sigma^2_{\rm assume}}\right)^2 N\sigma^2_{\rm true} + (\sigma^2_{\rm post})^2\left(\frac{N\mu}{\sigma^2_{\rm assume}} + \frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}}\right)^2 - 2\mu\sigma^2_{\rm post}\left(\frac{N\mu}{\sigma^2_{\rm assume}} + \frac{\mu_{\rm prior}}{\sigma^2_{\rm prior}}\right) + \mu^2, \tag{10}$$

and we can further rewrite it as

$$u^2 = \left(\frac{\sigma^2_{\rm post}}{\sigma^2_{\rm assume}}\right)^2 N\sigma^2_{\rm true} + \left(E\left[\mu_{\rm post}\right]_{d|\mu,\sigma^2=\sigma^2_{\rm true}} - \mu\right)^2. \tag{11}$$

Let's play with this expression a bit to make sense of it. First, note that it has two terms. The first term depends on the true noise variance, the second term does not (cf. Equation 11). This second term results from biases in the prior. One can make this bias disappear in two ways: by guessing $\mu_{\rm prior} = \mu$ (or enforcing it for the sake of theoretical study!), or by taking the limit of a flat prior, $\sigma^2_{\rm prior} \to \infty$.

In the flat prior limit, the first term reduces to

$$\lim_{\sigma^2_{\rm prior}\to\infty} \left(\frac{\sigma^2_{\rm post}}{\sigma^2_{\rm assume}}\right)^2 N\sigma^2_{\rm true} = \frac{\sigma^2_{\rm true}}{N}. \tag{12}$$

In particular, in the noiseless case, the mean will be achieved exactly (which makes sense, because inspecting the data vector would let you read off the mean). If $\mu_{\rm post}$ is being estimated via monte carlo (e.g. post-processing an MCMC chain), then there may be some sample variance associated with the monte carlo error, and one may see some scatter, but at some drastically reduced level based on how many independent monte carlo samples are being used.