Appendix B: Nonlinearity of Shrinkage Operations

Rachael Meager

November 11, 2020

1 The Set-up

Consider K parallel experiments each containing N participants randomized 50/50 into "treatment" and "control" groups (indicated by a binary variable $T_{ik} = 1$ if individual i in trial k is treated, else 0). Within a single experiment, the average treatment effect (ATE) is the difference of the means in the treatment and control groups, which is also the mean of the difference between the two groups due to the linearity of the expectation operator. Denote the ATE in experiment k by τ_k, estimated using the sample counterpart $\bar{\tau}_k = E[Y_{1ki} - Y_{0ki}] = E[Y_{1ki}] - [Y_{0ki}] = \bar{y_{1k}} - \bar{y_{0k}}$. This result relies only on the linearity of the expectations operator.

However, once the analyst places a hierarchical model on the data and jointly analyses all K experiments together, the updated expectations of these objects cease to obey this linear relationship. This is because shrinkage on any of these objects is a nonlinear operation in the unknown parameters, with particular issue caused by the unknown hypervariances. This appendix contains an illustration and proof of the problem, and concludes that analysts need to choose the object of interest and shrink directly on that object.

Consider the following Gaussian hierarchical model in which all the random parameters are independent to simplify the exposition.

$$
\begin{aligned}
y_{0k} &\sim N(y_0, \sigma_0^2) \\
\tau_k &\sim N(\tau, \sigma_\tau^2) \\
y_{1k} &\equiv y_{0k} + \tau_k \\
\therefore y_{1k} &\sim N(y_0 + \tau, \sigma_0^2 + \sigma_\tau^2) \\
\bar{y_{0k}} &\sim N(y_{0k}, \hat{se}_{y0}^2) \\
\bar{\tau_k} &\sim N(\tau_k, \hat{se}_{\tau k}^2) \\
\therefore \bar{y_{1k}} &\sim N(y_{0k} + \tau_k, \hat{se}_{0k}^2 + \hat{se}_{\tau k}^2)
\end{aligned}
$$

2 The Nonlinearity Result

This model generates new estimates of the parameters in each site k updated given the information in the other sites – that is, the model performs shrinkage.[1] Per Gelman et al (2004), if one knew the hyperparameters that govern $\{\tau_k\}_{k=1}^K$, i.e. (τ, σ_τ^2), one could manually compute the shrinkage on the observed $\bar{\tau_k}$ and thus the new posterior ATE $\tilde{\tau_k}$ for a given site k as follows:

$$\tilde{\tau_k} = \frac{\frac{1}{\hat{se}_{\tau k}^2}\bar{\tau_k} + \frac{1}{\sigma_\tau^2}\tau}{\frac{1}{\hat{se}_{\tau k}^2} + \frac{1}{\sigma_\tau^2}}.$$

Analogous objects exist for the y_{0k} and the y_{1k} if shrinkage is performed on them:

$$\tilde{y_{0k}} = \frac{\frac{1}{\hat{se}_{y_{0k}}^2}\bar{y_{0k}} + \frac{1}{\sigma_{y_0}^2}y_0}{\frac{1}{\hat{se}_{y_{0k}}^2} + \frac{1}{\sigma_{y_0}^2}}, \quad \tilde{y_{1k}} = \frac{\frac{1}{\hat{se}_{y_{1k}}^2}\bar{y_{1k}} + \frac{1}{\sigma_{y_1}^2}y_1}{\frac{1}{\hat{se}_{y_{1k}}^2} + \frac{1}{\sigma_{y_1}^2}}.$$

However, despite the fact that $\bar{\tau_k} = \bar{y_{1k}} - \bar{y_{0k}}$ and that $\tau = y_1 - y_0$, the following result holds:

Theorem 2.1. *Given the model above, $\tilde{\tau_k} \neq \tilde{y_{1k}} - \tilde{y_{0k}}$ and hence shrinkage is not a linear operation.*

[1] Notice that at most two of the triple (y_{0k}, y_{1k}, τ_k) can be independently distributed because the third object is constructed from the other two, and this third object will always be dependent on the components from which it is constructed. Here I have chosen y_{0k} and τ_k, which amounts to taking linear regression seriously as a model.

Proof. To see this, we begin with $\tilde{y_{1k}}$ and substitute in its components:

$$\begin{aligned}\tilde{y_{1k}} &= \frac{\frac{1}{\hat{se}^2_{y_{1k}}}(\bar{y_{0k}} + \bar{\tau_k}) + \frac{1}{\sigma^2_{y_1}}(y_0 + \tau)}{\frac{1}{\hat{se}^2_{y_{1k}}} + \frac{1}{\sigma^2_{y_1}}} \\ &= \frac{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}}(\bar{y_{0k}} + \bar{\tau_k}) + \frac{1}{\sigma^2_{y_0} + \sigma^2_\tau}(y_0 + \tau)}{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_{y_0} + \sigma^2_\tau}}\end{aligned}$$

Focusing only on the term that contains τ, we can see that it is being given the wrong weight in this expression:

$$\frac{\frac{1}{\sigma^2_{y_0} + \sigma^2_\tau}\tau}{\frac{1}{\hat{se}^2_{y_{0k}} + \hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_{y_0} + \sigma^2_\tau}} \neq \frac{\frac{1}{\sigma^2_\tau}\tau}{\frac{1}{\hat{se}^2_{\tau k}} + \frac{1}{\sigma^2_\tau}}$$

Since there is no other instance of τ in the formula for $\tilde{y_{1k}}$, this issue cannot be rectified by the presence of some other term. □

Notice that the only time the shrinkage will be "correct" on τ_k is when both the sampling error and the cross-site heterogeneity in y_{0k} is exactly zero. This turns out to be the key to understanding why the two kinds of shrinkage do not coincide: it is because shrinking the composite object y_{1k} allows the noise from the control mean to corrupt the shrinkage on the treatment effects (and vice versa).

To make this clear, consider an example. Suppose that $y_0 = 50$, $\tau = 30$ and thus $y_1 = 80$. Suppose that $\sigma_{y0} = 50$ and $\sigma_\tau = 1$, so that $\sigma_{y1} = \sqrt{50^2 + 1^2} = 50.01$. Then suppose that $\bar{y_{0k}} = 10$, $\hat{se}_{y_{0k}} = 10$, $\bar{\tau_k} = 20$, $\hat{se}_{\tau k} = 8$, and that therefore $\bar{y_{1k}} = 30$ and $\hat{se}_{y_{1k}} = \sqrt{10^2 + 8^2} = 12.80625$. Applying the formulae above, we get:

$$\begin{aligned}\tilde{y_{0k}} &= 11.53846 \\ \tilde{y_{1k}} &= 32.38005 \\ \tilde{\tau_k} &= 29.61538 \neq \tilde{y_{1k}} - \tilde{y_{0k}} = 21.53846.\end{aligned}$$

In this case the discrepancy is about 30% of the underlying parameter's true magnitude. This occurs it is because the control means y_{0k} vary greatly across the K sites, but the treatment effects τ_k vary little across the K sites. Faced with the task of shrinking on the composite object y_{1k}, the treatment group's mean, the model compromises between the two patterns – but the huge variance in y_{0k} across the K sites dominates the weight, so we still see essentially zero shrinkage on y_{1k}. However, when we direct the hierarchical model to shrink directly on τ_k, we isolate it from the noise on y_{0k} and the model can therefore detect that the τ_k component

of y_{1k} is very similar across the K sites and should be shrunk strongly towards the common mean of 30.