Chapter 4. Bayesian Analysis (C4)

♠ Hierarchical Bayes Analysis

Consider two stage priors.

Assume $\boldsymbol{x} \sim f(\boldsymbol{x}|\boldsymbol{\theta})$.

(1) 1st stage: $\pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda})$; and

(2) 2nd stage: $\pi_2(\boldsymbol{\lambda})$ (proper/noninformative).

We further assume $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2)$. Then

$$\pi_2(\boldsymbol{\lambda}) = \pi_{2,1}(\boldsymbol{\lambda}^1|\boldsymbol{\lambda}^2)\pi_{2,2}(\boldsymbol{\lambda}^2).$$

We have the following result:

Result: *Supposing all densities below exist and are nonzero,*

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \int_\Lambda \pi_1(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\lambda})\pi_{2,1}(\boldsymbol{\lambda}^1|\boldsymbol{x}, \boldsymbol{\lambda}^2)\pi_{2,2}(\boldsymbol{\lambda}^2|\boldsymbol{x})d\boldsymbol{\lambda}.$$

Here

$$\pi_1(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda})}{m_1(\boldsymbol{x}|\boldsymbol{\lambda})},$$

where $m_1(\boldsymbol{x}|\boldsymbol{\lambda}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}$*,*

$$\pi_{2,1}(\boldsymbol{\lambda}^1|\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{m_1(\boldsymbol{x}|\boldsymbol{\lambda})\pi_{2,1}(\boldsymbol{\lambda}^1|\boldsymbol{\lambda}^2)}{m_2(\boldsymbol{x}|\boldsymbol{\lambda}^2)},$$

where $m_2(\boldsymbol{x}|\boldsymbol{\lambda}^2) = \int m_1(\boldsymbol{x}|\boldsymbol{\lambda})\pi_{2,1}(\boldsymbol{\lambda}^1|\boldsymbol{\lambda}^2)d\boldsymbol{\lambda}^1$, *and*

$$\pi_{2,2}(\boldsymbol{\lambda}^2|\boldsymbol{x}) = \frac{m_2(\boldsymbol{x}|\boldsymbol{\lambda}^2)\pi_{2,2}(\boldsymbol{\lambda}^2)}{m(\boldsymbol{x})},$$

where $m(\boldsymbol{x}) = \int m_2(\boldsymbol{x}|\boldsymbol{\lambda}^2)\pi_{2,2}(\boldsymbol{\lambda}^2)d\boldsymbol{\lambda}^2$.

Example 1 (Example 17 of Berger's book): Suppose that seven independent IQ test scores $X_i \sim N(\theta_i, 100)$, and assume the θ_i are independently from a common distribution $N(\mu_\pi, \sigma_\pi^2)$ distribution. Thus, $\boldsymbol{X} = (X_1, \ldots, X_7)' \sim N_7(\boldsymbol{\theta}, 100I_7)$ (Defining $f(\boldsymbol{x}|\boldsymbol{\theta})$) and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_7)' \sim N_7((\mu_\pi, \ldots, \mu_\pi)', \sigma_\pi^2 I_7)$ (defining $\pi_1(\boldsymbol{\theta}|\boldsymbol{\lambda})$, $\boldsymbol{\lambda} = (\mu_\pi, \sigma_\pi^2)$). The seven IQ scores are 105, 127, 115, 130, 110, 135, and 115.

Rather than estimating $\boldsymbol{\lambda}$ as in empirical Bayes analysis, we can put a second stage hyperprior on $\boldsymbol{\lambda}$. It is natural to give μ_π a $N(100, 225)$ prior distribution, this being the overall population distribution of IQs; denote this density by $\pi_{2,1}(\mu_\pi)$. Our knowledge about σ_π^2 might be very vague, so that an (improper) constant density $\pi_{2,2}(\sigma_\pi^2) = 1$ would seem appropriate, Thus, the second stage hyperprior on $\boldsymbol{\lambda}$ would be (assuming independence of μ_π and σ_π^2),

$$\pi_2(\boldsymbol{\lambda}) = \pi_{2,1}(\mu_\pi)\pi_{2,2}(\sigma_\pi^2).$$

It can be shown that $\pi_1(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\lambda})$ is

$$N_7\left(\left[\boldsymbol{x} - \frac{100}{(100+\sigma_\pi^2)}(\boldsymbol{x} - \mu_\pi \mathbf{1})\right], \frac{100\sigma_\pi^2}{(100+\sigma_\pi^2)} I_7\right)$$

and $m_1(\boldsymbol{x}|\boldsymbol{\lambda})$ is $N_7(\mu_\pi \mathbf{1}, (100+\sigma_\pi^2)I_7)$. Since m_1 is normal and $\pi_{2,1}(\mu_\pi)$ is $N(100, 225)$, it can also be shown that $\pi_{2,1}(\mu_\pi|\boldsymbol{x}, \sigma_\pi^2)$ is

$$N\left(\left[\bar{x} - \frac{(100+\sigma_\pi^2)}{(1675+\sigma_\pi^2)}(\bar{x} - 100)\right], \frac{(100+\sigma_\pi^2)(225)}{(1675+\sigma_\pi^2)}\right)$$

and $m_2(\boldsymbol{x}|\sigma_\pi^2)$ is $N_7(100\mathbf{1}, (100+\sigma_\pi^2)I_7 + 225\mathbf{1}\mathbf{1}')$.
Finally,

$$\pi_{2,2}(\sigma^2|\boldsymbol{x})$$
$$= K \frac{\exp\left\{-\frac{1}{2}\left[\frac{s^2}{(100+\sigma_\pi^2)} + \frac{7(\bar{x}-100)^2}{(7(225)+100+\sigma_\pi^2)}\right]\right\}}{(100+\sigma_\pi^2)^3(7(225)+100+\sigma_\pi^2)^{1/2}225^{-1/2}} \pi_{2,2}(\sigma_\pi^2),$$

where $s^2 = \sum_{i=1}^7 (x_i - \bar{x})^2$ and K is the appropriate normalizing constant.

With $\pi_{2,2}(\sigma_\pi^2) = 1$, we obtain that the posterior mean and variance of θ_7 are 118.61 and 30, respectively. An approximate 95% credible set is

$$118.61 \pm (1.96)\sqrt{30} = (107.87, 129.35),$$

which is shorter than the one obtained by the empirical Bayes method.

Note: A comment is in order concerning the choice $\pi_{2,2}(\sigma_\pi^2) = 1$ in the example. Knowledge about hyperparameters is often quite vague so that π_2 is frequently chosen to be at least partially noninformative. In the example, it would have been tempting to use the standard noninformative prior, $\pi_{2,2}(\sigma_\pi^2) = 1/\sigma_\pi^2$ for σ_π^2, but care must be taken. Indeed, if $\pi_{2,2}(\sigma_\pi^2) = 1/\sigma_\pi^2$, then

$$\pi_{2,2}(\sigma^2|\boldsymbol{x})$$
$$= K \frac{\exp\left\{-\frac{1}{2}\left[\frac{s^2}{(100+\sigma_\pi^2)} + \frac{7(\bar{x}-100)^2}{(7(225)+100+\sigma_\pi^2)}\right]\right\}}{(100+\sigma_\pi^2)^3(7(225)+100+\sigma_\pi^2)^{1/2}225^{-1/2}} / \sigma_\pi^2.$$

Thus, $\pi_{2,2}(\sigma^2|\boldsymbol{x})$ is not integrable because of nonintegrability as $\sigma_\pi^2 \to 0$.

Because of such potential problems in hierarchical Bayes analysis, it is indeed often best to simply choose constant noninformative priors on hyperparameters. (Perhaps Laplace was right, in a practical sense, to simply pretend that unknown parameters had constant priors.)

♠ The Relationship Between the Power Prior and Hierarchical Models

We consider the case of one historical dataset. Consider the normal hierarchical model

$$y_i = \theta + \epsilon_i,\ i = 1, 2, \ldots, n,$$

$$y_{0i} = \theta + \epsilon_{0i}, i = 1, 2, \ldots, n_0,$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ denote the current data with the sample size of n and $\boldsymbol{y}_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})$ denote the historical data with the sample size of n_0. We further assume that the ϵ_i are i.i.d. $N(0, \sigma^2)$ and the ϵ_{0i} are i.i.d. $N(0, \sigma_0^2)$ and independent of the ϵ_i's, where σ^2 and σ_0^2 are known.

Now the hierarchical model is completed by independently taking

$$\theta_0 \mid \mu, \tau^2 \sim N(\mu, \tau^2), \qquad \theta \mid \mu, \tau^2 \sim N(\mu, \tau^2),$$

and then taking

$$\mu \sim N(\alpha, \nu^2),$$

where α, ν^2, and τ^2 are all fixed hyperparameters.

Within the development of this hierarchical model, our goal is to make inferences about the current study through the marginal posterior distribution of $[\theta|\boldsymbol{y}, \boldsymbol{y}_0]$. Here, θ is the parameter of interest for the current study, such as a treatment effect, and θ_0 denotes the corresponding parameter based on the historical study. The following theorem gives the form of the marginal posterior distribution of $[\theta|\boldsymbol{y}, \boldsymbol{y}_0]$.

Theorem 1: *Letting* $\bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_{0i}$ *and* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, *we have*

$$(\theta \mid \boldsymbol{y}, \boldsymbol{y}_0) \sim N(\mu_h, \sigma_h^2),$$

where

$$\begin{aligned}\mu_h =& \sigma_h^2 \Bigg[\frac{n\bar{y}}{\sigma^2} + \frac{\alpha}{\tau^2\nu^2\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)} \\ &+ \frac{\frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}\left(\frac{n_0\bar{y}_0}{\sigma_0^2} + \frac{\alpha}{\tau^2\nu^2\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}\right)}{\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}} \Bigg]\end{aligned}$$

and

$$\begin{aligned}\sigma_h^2 = \Bigg[& \frac{n}{\sigma^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)} \\ & - \frac{1}{\left(\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)\right)^2\left(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}\right)} \Bigg]^{-1}.\end{aligned}$$

Now we consider a power prior formulation of the model. To do this, we set $\theta_0 = \theta$, and the resulting model becomes

$$y_i = \theta + \epsilon_i, \text{ and } y_{0i} = \theta + \epsilon_{0i}.$$

Thus in the power prior formulation, the ϵ_i are i.i.d. $N(0, \sigma^2)$, and the ϵ_{0i} are i.i.d. $N(0, \sigma_0^2)$ and independent of the ϵ_i's, where σ^2 and σ_0^2 are known. Under this model, the power prior based on the historical data $\boldsymbol{y}_0$ using the initial prior $\pi_0(\theta) \propto 1$, is given by

$$\pi(\theta|\boldsymbol{y}_0) \propto \exp\left\{-\frac{a_0}{2\sigma_0^2}\sum_{i=1}^{n_0}(y_{0i} - \theta)^2\right\}.$$

Straightforward calculations show that

$$\theta|\boldsymbol{y}, \boldsymbol{y}_0 \sim N(\mu_p, \sigma_p^2),$$

where

$$\mu_p = \frac{\frac{n\bar{y}}{\sigma^2} + a_0\frac{n_0\bar{y}_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + a_0\frac{n_0}{\sigma_0^2}},$$

and

$$\sigma_p^2 = \frac{1}{\frac{n}{\sigma^2} + a_0\frac{n_0}{\sigma_0^2}}.$$

We now examine the relationship between these two different formulations. To do this, we need to find an explicit relationship between μ_h and μ_p as well as a relationship between σ_h^2 and σ_p^2. We are led to the following theorem which characterizes this relationship.

Theorem 2: *The two posteriors match, i.e., $\mu_h = \mu_p$ and $\sigma_h^2 = \sigma_p^2$ if and only if $\alpha = 0$ and $\nu^2 \to \infty$, and*

$$a_0 = \frac{1}{\frac{2\tau^2 n_0}{\sigma_0^2} + 1}.$$

Proof: It is easy to show that μ_h matches to μ_p if and only if $\alpha = 0$. When $\alpha = 0$, μ_h reduces to

$$\mu_h = \sigma_h^2 \left[\frac{n\bar{y}}{\sigma^2} + \frac{\frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)} \cdot \frac{n_0\bar{y}_0}{\sigma_0^2}}{\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}} \right].$$

To match μ_h to μ_p, we have to set

$$a_0 = \frac{1}{\tau^4(\frac{1}{\nu^2} + \frac{2}{\tau^2})(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4(\frac{1}{\nu^2}+\frac{2}{\tau^2})})}$$

and

$$\frac{n}{\sigma^2} + a_0 \frac{n_0}{\sigma_0^2} = \sigma_h^{-2}.$$

Note that the above equation directly yields that σ_h^2 is equal to σ_p^2. Now, using the above expression of a_0, the equation,

$$\frac{n}{\sigma^2} + a_0 \frac{n_0}{\sigma_0^2} = {\sigma_h}^{-2}.$$

reduces to

$$\begin{aligned} &\frac{\frac{n_0}{\sigma_0^2}}{\tau^4(\frac{1}{\nu^2} + \frac{2}{\tau^2})(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4(\frac{1}{\nu^2}+\frac{2}{\tau^2})})} \\ = \ & \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2} + \frac{2}{\tau^2}\right)} \\ & - \frac{1}{\left(\tau^4\left(\frac{1}{\nu^2} + \frac{2}{\tau^2}\right)\right)^2 \left(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}\right)}. \end{aligned}$$

After some algebra, we obtain

$$\begin{aligned} &\frac{\frac{n_0}{\sigma_0^2}}{\tau^4(\frac{1}{\nu^2} + \frac{2}{\tau^2})(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4(\frac{1}{\nu^2}+\frac{2}{\tau^2})})} \\ = \ & \frac{1}{\tau^4\left(\frac{1}{\nu^2} + \frac{2}{\tau^2}\right)} \\ & - \frac{\frac{\tau^2}{\nu^2} + 1}{\left(\tau^4\left(\frac{1}{\nu^2} + \frac{2}{\tau^2}\right)\right)^2 \left(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4\left(\frac{1}{\nu^2}+\frac{2}{\tau^2}\right)}\right)}. \end{aligned}$$

Thus, the above equality holds if and only if $\nu^2 \to \infty$.

When $\nu^2 \to \infty$,

$$\begin{aligned} a_0 =& \frac{1}{\tau^4(\frac{1}{\nu^2} + \frac{2}{\tau^2})(\frac{n_0}{\sigma_0^2} + \frac{1}{\tau^2} - \frac{1}{\tau^4(\frac{1}{\nu^2}+\frac{2}{\tau^2})})} \\ \to& \frac{1}{\frac{2\tau^2 n_0}{\sigma_0^2} + 1}, \end{aligned}$$

which completes the proof. □

Corollary 1: *The choice of* a_0 *given in Theorem 2 satisfies* $0 < a_0 < 1$.

Corollary 2: *The result as in Theorem 2 can be alternatively obtained by taking a uniform improper prior for* μ *at the outset, i.e.,* $\pi(\mu) \propto 1$.

Theorem 2 gives us a useful characterization of the explicit relationship between the power prior and the hierarchical model, and we see from this theorem that the two models are equivalent if a_0 is chosen as

$$a_0 = \frac{1}{\frac{2\tau^2 n_0}{\sigma_0^2} + 1}.$$

We also see that a_0 is a monotonic function of τ and that if $\tau^2 \to 0$, then $a_0 \to 1$. This implies that if $\theta = \theta_0$ with probability 1, the historical and current data should be weighted equally. Also, the larger the sample size for the historical data, the less the weight given to the historical data. This is a desirable property since, in general, we would never want the historical data to dominate the posterior distribution of θ by simply increasing n_0.

Note: The result given in Theorem 2 establishes a formal relationship between the power prior and the usual normal Bayesian hierarchical model for incorporating historical data. This result is critical since it obtains an analytic relationship between the power parameter and the variance components in the hierarchical model, thereby motivating and justifying the use of the power prior as an informative prior for incorporating historical data.

Another implication of the result given Theorem 2 is that instead of the use of power prior, one can simply build a usual Bayesian hierarchical model for both historical and current datasets and then one makes inference based on the marginal posterior distribution of the model parameters associated with the current data.