

## Hierarchical Models

A *hierarchy* is a classification system having different levels.

A *hierarchical model* has levels as well.

Simple example:

Observed datum is  $X$ .

- $X|\theta \sim N(\theta, 1)$
- First stage prior:  $\theta|\tau^2 \sim N(0, \tau^2)$
- Second stage prior:  $\tau^2 \sim \pi$ , where  $\pi$  is completely specified.

In some instances we would specify  $\tau^2$  according to our prior beliefs about  $\theta$ . Then we would have only the first stage prior. (Or, we could think of  $\pi$  as a distribution degenerate at  $\tau^2$ .)

If we can't easily quantify our uncertainty about  $\theta$ , then the second stage prior could be used.

Of course, we could carry the process further. We might say

- Second stage prior:  $\tau^2 | \alpha \sim \text{gamma}(\alpha, 1)$
- Third stage prior:  $\alpha$  has exponential density with mean 1.

## *Connection of hierarchical and empirical Bayes methods*

### Example 21

Suppose that, conditional on  $\theta_1, \dots, \theta_n$ ,  $X_1, \dots, X_n$  are independent with

$$X_i | \theta_i \sim N(\theta_i, 1)$$

and  $\theta_1, \dots, \theta_n$  are i.i.d.  $N(\mu, \sigma^2)$ , where  $\mu$  and  $\sigma^2$  are unknown.

In empirical Bayes,  $\mu$  and  $\sigma^2$  are estimated from the observations  $X_1, \dots, X_n$ , and then the estimated “prior” for  $\theta$  is employed in usual Bayesian fashion to obtain estimates of  $\theta_1, \dots, \theta_n$ .

Let's see how this works.

The joint density of  $X_1, \dots, X_n$  and  $\theta_1, \dots, \theta_n$  is

$$\begin{aligned}
 f(x_1, \dots, x_n, \theta_1, \dots, \theta_n) &= \\
 f(x_1, \dots, x_n | \theta_1, \dots, \theta_n) \sigma^{-n} \prod_{i=1}^n \phi\left(\frac{\theta_i - \mu}{\sigma}\right) &= \\
 \prod_{i=1}^n \phi(x_i - \theta_i) \sigma^{-n} \prod_{i=1}^n \phi\left(\frac{\theta_i - \mu}{\sigma}\right) &=
 \end{aligned}$$

Now let's find the marginal of  $X_1, \dots, X_n$ . We need to integrate out  $\theta_1, \dots, \theta_n$ , which can be done separately for each  $\theta_i$ .

$$\begin{aligned}
 \int_{-\infty}^{\infty} \phi(x_i - \theta_i) \phi\left(\frac{\theta_i - \mu}{\sigma}\right) d\theta_i &= \\
 \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left(x_i^2 + \frac{\mu^2}{\sigma^2}\right)\right] \times \\
 \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\theta_i^2(1 + \sigma^{-2}) - 2\theta_i\left(x_i + \frac{\mu}{\sigma^2}\right)\right]\right\} d\theta_i.
 \end{aligned}$$

Using our cherished device of completing the square, the previous quantity can be shown to equal

$$\frac{1}{\sqrt{2\pi}} \frac{\sigma}{\sqrt{\sigma^2 + 1}} \phi \left( \frac{x_i - \mu}{\sqrt{\sigma^2 + 1}} \right).$$

It follows that, marginally,  $X_1, \dots, X_n$  are i.i.d.  $N(\mu, \sigma^2 + 1)$ .

In the empirical Bayes approach, frequentist methods are often used to estimate parameters of the prior.

The MLEs of  $\mu$  and  $\sigma^2 + 1$  are  $\bar{X}$  and  $S^2$ , respectively. The MLE of  $\sigma^2$  is  $\hat{\sigma}^2 = \max(0, S^2 - 1)$ .

If  $\mu$  and  $\sigma^2$  were known, the Bayes estimate of  $\theta_i$  would be

$$\hat{\theta}_i = wx_i + (1 - w)\mu, \quad (*)$$

where  $w = \sigma^2 / (\sigma^2 + 1)$ .

The empirical Bayes estimate of  $\theta_i$  plugs estimates of  $\mu$  and  $\sigma^2$  into (\*):

$$\hat{\theta}_{i,EB} = \hat{w}x_i + (1 - \hat{w})\bar{X},$$

where  $\hat{w} = \hat{\sigma}^2 / (\hat{\sigma}^2 + 1)$ .

The estimate  $\hat{\theta}_{i,EB}$  has an interesting interpretation. When  $\sigma^2$  (and hence  $\hat{\sigma}^2$ ) is large, most of the weight is put on  $x_i$ , and when  $\sigma^2$  is small, most of the weight is on  $\bar{X}$ .

Note that

$$\sigma^2 = 0 \implies \text{all } \theta_i\text{s equal } \mu,$$

in which case we *should* estimate each  $\theta_i$  by  $\bar{X}$ .

When  $\sigma^2$  is big relative to 1, information from other  $\theta_j$ s doesn't help in estimating  $\theta_i$ .

Empirical Bayes (EB) uses the hierarchy idea:

- First stage:  $X_i|\theta_i \sim N(\theta_i, 1)$
- Second stage:  $\theta_i|(\mu, \sigma^2) \sim N(\mu, \sigma^2)$

Bayesian “purists” don't like EB since it uses the data to estimate the prior.

A hierarchical Bayes model in the setting of Example 21 would use a second level prior for  $(\mu, \sigma)$ . Call this prior  $\pi_2$ . If only one extra level is used,  $\pi_2$  would be completely specified. Then a proper Bayesian analysis could be done in which a posterior for  $(\theta_1, \dots, \theta_n, \mu, \sigma)$  is obtained.

The posterior is

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mu, \sigma^2 | \mathbf{x}) &\propto f(\mathbf{x} | \boldsymbol{\theta}, \mu, \sigma) \pi(\boldsymbol{\theta} | \mu, \sigma) \pi_2(\mu, \sigma) \\ &= \prod_{i=1}^n \phi(x_i - \theta_i) \\ &\times \prod_{i=1}^n \frac{1}{\sigma} \phi\left(\frac{\theta_i - \mu}{\sigma}\right) \pi_2(\mu, \sigma). \end{aligned}$$

Note:  $\pi_2(\mu, \sigma) = \pi_2(\mu|\sigma)p(\sigma)$ . For the setting on the previous page, It's ok to use

$$\pi_2(\mu|\sigma) \propto 1,$$

but  $p(\sigma) \propto \sigma^{-1}$  results in an improper posterior. Using  $p(\sigma) \propto 1$  yields a *proper* posterior.

A discussion of more general normal hierarchical models is given in Chapter 5 of GCSR.