Lecture 19: March 16

Lecturer: Donlapark Pornnopparath

19.1 Hierarchical models

Let us recall a simple Bayesian model $x \sim N(\theta, \sigma^2)$ with a prior $\theta \sim N(\mu, \tau^2)$. In this case, we need to figure out appropriate values of μ and τ. If we were using an empirical Bayes approach, then we could pick the MLE (which is the posterior mean) of μ. Alternatively, we could have gone full Bayesian approach and impose another prior on τ, and keep going with a chain of priors until we obtain infinite hierarchy of hyperparameters. To find a right number of layers in the hierarchy and a right choice of priors, one would need to run sensitivity analyses.

Let us see a situation in which hierarchical model cam come into play.

Example 19.1. Consider the one-way normal random effect ANOVA:

$$y_{ij} = \theta_j + \epsilon_{ij},$$

of J groups $j = 1, \dots, J$. θ_j is the expected effect from $j-$th group and θ_{ij} are the deviations from the expectation of the sample in the $j-$th group. We can treat θ_j as latent variables of the following hierarchical model:

$$\begin{aligned} y_{ij}|\theta_j, \sigma^2 &\sim N(\theta_j, \sigma^2) \\ \theta_j|\mu, \tau &\sim N(\mu, \tau^2) \end{aligned}$$

with the prior

$$\pi(\mu, \tau^2) \propto \pi(\tau),$$

where σ^2 is known. We could estimate θ_j with the simple sufficient statistics $\bar{y}_j = \frac{1}{n_j}\sum_i y_{ij} \sim N(\theta_j, \sigma_j^2)$ where $\sigma_j^2 = \sigma^2/n_j$. On the contrary, using the hierarchical model, we allow information from other θ_k's to estimate θ_j; the posterior of θ_j is given by

$$\theta_j|\mu, \tau, \boldsymbol{y} \sim N(\hat{\theta}_j, \hat{\sigma}_j),$$

where

$$\hat{\theta}_j = \frac{\tau^2}{\sigma_j^2 + \tau^2}\bar{y}_j + \frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\mu \quad \text{and} \quad \hat{\sigma}_j = \frac{\sigma_j^2\tau^2}{\sigma_j^2 + \tau^2}. \tag{19.1}$$

From this, we can compute the posterior distribution for μ.

$$\begin{aligned} p(\mu|\tau, \boldsymbol{y}) &\propto p(\tau)\prod_{j=1}^{J} p(\bar{y}_j|\mu, \tau) \\ &= p(\tau)\prod_{j=1}^{J}\int p(\bar{y}_j|\theta_j)p(\theta_j|\mu, \tau)\, d\theta_j \\ &\propto p(\tau)\prod_{j=1}^{J} N(\mu|\bar{y}_j, \sigma_j^2 + \tau^2), \end{aligned}$$

which leads to

$$\mu|\tau, \boldsymbol{y} \sim N(\hat{\mu}, V_\mu),$$

where

$$\hat{\mu} = \frac{\sum_j \bar{y}_j(\sigma_j^2 + \tau^2)^{-1}}{\sum_j(\sigma_j^2 + \tau^2)^{-1}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_j(\sigma_j^2 + \tau^2)^{-1}}. \tag{19.2}$$

We can see from (19.1) and (19.2) that τ controls how much information are shared across the groups. To pick a suitable prior for τ, we might want to try different priors and do some sensitivity analysis. ◇

Example 19.2. Gelman et al. 2003 A study was performed on students in different schools to measure the effect of a standardized test preparation course. After collecting the test scores, linear regression analysis was done to measure the median difference in the scores between the treatment and the control groups The data is shown in Table 19.1

School	Difference	Standard error
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Table 19.1: Difference in standized test scores of students from various schools.

If we were to do a frequentist analysis then we have to decided between two hypotheses: either all schools have different effects or they have the same effects. The former gives somewhat undesirable conclusion; since the data are assumed to be normally distributed, 50% of students from school A actually gained by more than 28 points from the prep course. This might be unrealistic compared to the data from the other schools. On the other hand, assuming that the effects are the same is also questionable since most students from school A would have improved their scores by more than the overall mean (which is 8.75).

Suppose that we use the hierarchical Bayes with a uniform prior on τ. Then the posterior distribution is shown in Figure 19.1. As a result, the model *shrinks* the median score of school A (according to (19.1)) from 28 to 19. ◇

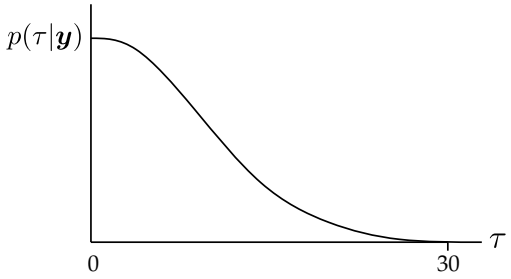

Figure 19.1: The posterior distribution of τ.

19.1.1 Inference in hierarchical models

Gibbs sampling

We have first introduced MCMC as a tool to calculate the gradient in each gradient descent step in MLE of an MRF. In Bayesian model, we can apply MCMC directly to infer posterior distributions. In particular, Gibbs sampling is pretty straightforward for the hierarchical Gaussian example above; assuming that $\tau^2 \sim \text{IG}(\alpha, \beta)$ where IG is the inverse-gamma, we compute the marginal distributions

$$\theta_j | \mu, \tau, \boldsymbol{y} \sim N(\hat{\theta}_j, \hat{\sigma}_j).$$

For μ, we take advantage of the conditional independence.

$$\begin{aligned} \mu | \boldsymbol{\theta}, \tau, \boldsymbol{y} &= \mu | \boldsymbol{\theta}, \tau \\ &= \mu | \bar{\theta}, \tau \\ &\sim N\left(\bar{\theta}, \frac{\tau^2}{\sqrt{J}}\right), \end{aligned}$$

where $\bar{\theta} = \frac{1}{J}\sum_j \theta_j$. For τ^2, we have to take the prior into account

$$\begin{aligned} \tau^2 | \boldsymbol{\theta}, \mu, \boldsymbol{y} &= \tau^2 | \boldsymbol{\theta}, \mu \\ &= \tau^2 | \bar{\theta}, \mu \\ &\sim \text{IG}\left(\alpha + \frac{J}{2} + 1, \hat{\beta}\right), \end{aligned}$$

where $\hat{\beta} = \beta + \sum_{j=1}^{J} (\mu - \theta_j)^2$. Note that this just an approximation since the support is finite from the uniform prior.

In summary, the Gibbs sampling for hierarchical model for random effects ANOVA is as follows:

1. Initialize $\theta_1^0, \theta_2^0, \ldots, \theta_J^0, \mu^0$ and τ^0.
2. For $t = 0, 1, \ldots, T$ sample

$$\begin{aligned} &\theta_j^{t+1} | \mu^t, \tau^t, \boldsymbol{y} \sim N\left(\tfrac{(\tau^t)^2}{\sigma_j^2 + (\tau^t)^2}\bar{y}_j + \tfrac{\sigma_j^2}{\sigma_j^2 + (\tau^t)^2}\mu^t, \tfrac{\sigma_j^2 (\tau^t)^2}{\sigma_j^2 + (\tau^t)^2}.\right) \text{ for } j = 1, \ldots, J. \\ &\mu^{t+1} | \boldsymbol{\theta}^{t+1}, \tau^t, \boldsymbol{y} \sim N\left(\tfrac{1}{J}\sum_{j=1}^{J}, \tfrac{(\tau^t)^2}{\sqrt{J}}\right). \\ &\tau^{t+1} | \boldsymbol{\theta}^{t+1}, \mu^{t+1}, \boldsymbol{y} \sim \text{IG}\left(\tfrac{J}{2}, \textstyle\sum_{j=1}^{J} (\mu^{t+1} - \theta_j^{t+1})^2\right). \end{aligned}$$

Variational inference

In this case, it is easier to study the precision $\gamma^2 = 1/\tau^2$ and $\delta^2 = 1/\sigma^2$ with a gamma prior $\gamma^2 \sim \Gamma(\alpha, \beta)$. Recall the mean-field approximation:

$$p(\boldsymbol{\theta}, \mu, \gamma | \boldsymbol{y}) \approx q_\mu(\mu) q_\gamma(\gamma) \prod_{j=1}^{J} q_j(\theta_j).$$

Define the unnormalized joint probability

$$\tilde{p}(\boldsymbol{\theta}, \mu, \gamma | \boldsymbol{y}) = \pi(\tau) \prod_{j=1}^{J} p(\theta_j | \mu, \gamma) p(\bar{y}_j | \theta_j, \sigma).$$

Then, to find each optimal distribution $q(x)$, we can just look at the Markov blanket of x in $\log \tilde{p}(\boldsymbol{\theta}, \mu, \gamma)$ and treat the rest as constants.

$$\begin{aligned} q_j(\theta_j) &\propto \exp\left\{\mathbb{E}_{\mu,\tau} \log \tilde{p}(\boldsymbol{\theta}, \mu, \gamma | \boldsymbol{y})\right\} \\ &\propto \exp\left\{\mathbb{E}_{\mu,\tau} \log p(\theta_j|\mu,\gamma) p(\bar{y}_j|\theta_j,\sigma)\right\} \\ &\propto \exp\left\{\mathbb{E}_{\mu,\tau}\left[-\frac{\gamma^2(\theta_j-\mu)^2}{2} - \frac{n_j\delta^2(\bar{y}_j-\theta_j)^2}{2}\right]\right\} \\ &\propto N\left(\theta_j \,\middle|\, \frac{\mathbb{E}\gamma^2\mathbb{E}\mu + n_j\delta^2\bar{y}_j}{\mathbb{E}\gamma^2 + n_j\delta^2},\ \frac{1}{\mathbb{E}\gamma^2 + n_j\delta^2}\right). \end{aligned}$$

Similarly,

$$\begin{aligned} q_\mu(\mu) &\propto \exp\left\{\mathbb{E}_{\theta_1,\ldots,\theta_J,\gamma} \log \prod_{j=1}^{J} p(\theta_j|\mu,\gamma)\right\} \\ &\propto N\left(\mu \,\middle|\, \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\theta_j,\ n\gamma^2\right), \end{aligned}$$

and

$$\begin{aligned} q_\gamma(\gamma) &\propto \exp\left\{\mathbb{E}_{\theta_1,\ldots,\theta_J,\mu} \log \pi(\gamma) \prod_{j=1}^{J} p(\theta_j|\mu,\gamma)\right\} \\ &\propto \Gamma\left(\gamma^2 \,\middle|\, \frac{J}{2},\ \sum_{j=1}^{J}(\mathbb{E}\mu^2 - 2\mathbb{E}\mu\mathbb{E}\theta_j^2 + \mathbb{E}\theta_j^2)\right). \end{aligned}$$

In practice, the second moment of each variable is written in terms of its mean and variance. We can see that the choice of the prior helps in arriving at this closed form.