STAT 525 Lecture 18

October 31, 2017

1 Hierarchical Model

$$
\begin{aligned}
y \mid \theta &\sim p(y \mid \theta) \\
\theta \mid \varphi &\sim p(\theta \mid \varphi)
\end{aligned}
$$

Possible approaches:

1. Fix q based on prior knowledge or diffuse (Jeffrey's)
2. Fully Baues φ unknown, so we use prior and integrating it out

   $$\varphi \sim p(\varphi)$$

3. Empirical Bayes

   $$
   \begin{aligned}
   p(y \mid \varphi) &= \int p(y, \theta \mid \varphi)\, d\theta \\
   &= \int \underbrace{p(y \mid \theta, \varphi)}_{p(y|\theta)} p(\theta \mid \varphi)\, d\theta \\
   &= \int p(y \mid \theta)\, p(\theta \mid \varphi)\, d\theta
   \end{aligned}
   $$

   where

   $$\hat{\varphi} = \arg\max_{\varphi} p(y \mid \varphi)$$

   which is the marginalized max likelihood estimator (MMLE).

1.1 Eight School Example

$$
\begin{aligned}
y_{ij} &\sim N\left(\theta_j, \sigma_y^2\right) \\
&\rightleftarrows \\
\bar{y}_{\cdot j} &\sim N\left(\theta_j, \sigma_j^2\right) \\
\bar{y}_{\cdot j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \\
\sigma_j^2 &= \frac{\sigma_y^2}{n_j}
\end{aligned}
$$

where $\theta_j \sim N\left(\mu, \sigma_\theta^2\right)$ and marginally is

$$\bar{y}_{\cdot j} \mid \mu, \left\{\sigma_j^2\right\}, \sigma_\theta^2 \sim N\left(\mu, \sigma_j^2 + \sigma_\theta^2\right)$$

where

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\cdot\cdot} \\
\hat{\theta}_j &= \mathbb{E}\left[\theta_j \mid y, \mu = \hat{\mu}, \sigma_j^2, \sigma_\theta^2\right] \\
&= \left(1 - \kappa_{\theta_j}\right)\bar{y}_{\cdot j} + \kappa_{\theta_j}\bar{y}_{\cdot\cdot} \\
\kappa_{\theta_j} &= \frac{\sigma_j^2}{\sigma_j^2 + \sigma_\theta^2}
\end{aligned}$$

Comments on Empirical Bayes

1. Use optim() on MLE tools for $\hat{\varphi}$
2. More completed models. Use EM algorithm (θ_j latent).
3. $\hat{\theta}_j$ is an estimator with freq properties to study.
4. However, use data to estimate prior is nonsense to Bayesian.
5. Underestimated the uncertainty in the post distribution.

1.2 Ridge Regression

Mixed effect model :

$$\begin{aligned}
y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \\
\epsilon_i &\overset{\text{iid}}{\sim} N\left(0, \sigma_\epsilon^2\right) \\
\beta_j &\sim N\left(0, \sigma_\beta^2\right)
\end{aligned}$$

Idea: Use (RE)ML o estimate $(\sigma_\epsilon, \sigma_\beta)$ as in a standard mixed model.

$$\begin{aligned}
y_i &\sim N\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma_\epsilon^2\right) \\
\boldsymbol{y} &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma_\epsilon^2 \mathbb{I}\right)
\end{aligned}$$

1. Ridge $\beta_j \sim N\left(0, \sigma_\beta^2\right)$
2. g-prior $\boldsymbol{\beta} \sim N\left(0, \sigma_\epsilon^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$

Bayesian Factors

Goal: compute two models/ hypothesis H_0, H_1

$$\begin{aligned}
H_0 &: \theta \in \Theta_0 \\
H_1 &: \theta \in \Theta_1
\end{aligned}$$

where Θ_0 and Θ_1 partition the parameter space.

Prior odds ratios

$$\frac{\pi_0^{\text{prior}}}{\pi_1^{\text{prior}}} = \frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)} = \frac{\int_{\Theta_0} p(\theta)\, d\theta}{\int_{\Theta_1} p(\theta)\, d\theta}$$

The observed chosen data $\boldsymbol{y}$

$$\begin{aligned}
\frac{\pi_0^{\text{post}}}{\pi_1^{\text{post}}} &= \frac{P(\theta \in \Theta_0 \mid \boldsymbol{y})}{P(\theta \in \Theta_1 \mid \boldsymbol{y})} \\
&= \frac{\int_{\Theta_0} p(\theta \mid \boldsymbol{y})\, d\theta}{\int_{\Theta_1} p(\theta \mid \boldsymbol{y})\, d\theta} \\
&= \frac{\int_{\Theta_0} p(\boldsymbol{y} \mid \theta)\, p(\theta)\, d\theta}{\int_{\Theta_1} p(\boldsymbol{y} \mid \theta)\, p(\theta)\, d\theta}
\end{aligned}$$

Bayes Factor (BF): ratios of post odds to the prior odds

$$\begin{aligned}
BF(H_0, H_1) &= \frac{\text{post odds}}{\text{prior odds}} = \frac{\pi_0^{\text{prior}}/\pi_1^{\text{prior}}}{\pi_0^{\text{post}}/\pi_1^{\text{post}}} \\
P(H_1 \mid \boldsymbol{y}) &= 1 - P(H_0 \mid \boldsymbol{y}) \\
\pi_0^{\text{post}} &= p(H_0 \mid \boldsymbol{y}) \\
&= \frac{\pi_0^{\text{prior}}\text{BF}}{\pi_0^{\text{prior}}\text{BF} + \left(1 - \pi_0^{\text{prior}}\right)}
\end{aligned}$$

More generally for H_0 and H_1

$$\begin{aligned}
\frac{P(H_0 \mid \boldsymbol{y})}{P(H_1 \mid \boldsymbol{y})} &= \underbrace{\frac{P(\boldsymbol{y} \mid H_0)}{P(\boldsymbol{y} \mid H_1)}}_{BF(H_0,H_1)} \times \frac{P(H_0)}{P(H_1)} \\
BF(H_0, H_1) &= \frac{P(\boldsymbol{y} \mid H_0)}{P(\boldsymbol{y} \mid H_1)} \\
&= \frac{\int p(\theta_0 \mid H_0)\, p(\boldsymbol{y} \mid \theta_0, H_0)\, d\theta_0}{\int p(\theta_1 \mid H_1)\, p(\boldsymbol{y} \mid \theta_1, H_1)\, d\theta_1}
\end{aligned}$$

- General hypothesis testing
- Evidence in favor of the null
- model averaging

$$\begin{aligned}
\hat{\theta}_{\text{avg}} &= \sum_l \underbrace{\hat{\theta}_h}_{\text{estimate from model } h} \pi_h^{\text{post}} \\
&= \sum_h \hat{\theta}_h P(H_h \mid \boldsymbol{y})
\end{aligned}$$

- Good examples: same model different values
  - $H_0 : \theta = \theta_0,\ H_1 : \theta = \theta_1$
    * $\text{BF}(H_0, H_1) = \frac{P(\boldsymbol{y}|H_0)}{P(\boldsymbol{y}|H_1)} \times \frac{P(H_0)}{P(H_1)} = \frac{P(\boldsymbol{y}|H_0)}{P(\boldsymbol{y}|H_1)} = \text{Likelihood ratio}$

- One side test of normal mean
  - $y_i \sim N\left(\theta, \sigma^2\right)$
  - $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$
  - Estimate

$$\hat{P}\left(\theta \leq \theta_0 \mid \boldsymbol{y}\right)$$

- Bad examples:
  - Point null
    - $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
    - $P\left(H_0\right)$ versus $P\left(H_1\right)$?
    - Examples: Eight school no pooling vs complete pooling?