Empirical Bayes Appoaches

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

03/15/2021

The Empirical Bayes Approach

- An empirical Bayes (EB) method is one in which the prior is estimated from the data.
- Useful when borrowing strength across multiple related parameters $\theta = (\theta_1, \ldots, \theta_k)$
- Model $\theta_i' s$ as iid from shared (estimated) prior $\hat{p}_\theta(\cdot)$.
  - *Parametric EB*: Estimate p_θ from a parametric family with hyperparameters η: $p(\cdot \mid \eta)$
  - *Nonparametric EB*:Assume no parametric form for p_θ

Parametric empirical Bayes

- For the parametric EB approach we consider prior $p_\theta(\cdot \mid \eta)$ and estimate η.
- A natural estimate is the maximizer of the marginal density

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\ p(\mathbf{y} \mid \eta) = \underset{\eta}{\operatorname{argmax}} \int p(\mathbf{y} \mid \theta) p(\theta \mid \eta) d\theta$$

  - Used in the previous regression example, with $\theta := \beta$ and $\eta := \tau^2$.

- Alternatively, could estimate η by matching moments or some other approach.

Example: Liver disease (cont.)

- Recall: mice from 50 genetic strains
  - n_i is the total number of mice from strain i
  - y_i is the number of mice with liver disease from strain i.
  - θ_i is probability a mouse from strain i develops liver disease:
$$y_i \sim \text{Binomial}(n_i, \theta_i).$$
  - $\theta_i's$ are iid from Beta(a, b) distribution.
- Would like to estimate a, b empirically.

Example: Liver disease (cont.)

- Consider reparameterization $\mu = \frac{a}{a+b}$ and $M = a + b$.
- The mean and variance of a beta distributed variable θ are

$$E(\theta) = \mu \quad \text{and} \quad \mathrm{Var}(\theta) = \frac{\mu(1-\mu)}{M+1}$$

- To compute mean and variance of Y in terms of μ and M, use
  - Law of total expectation:

$$E_y Y = E_\theta[E_{Y|\theta} Y]$$

  - Law of total variance:

$$\mathrm{Var}_y Y = E_\theta[\mathrm{Var}_{Y|\theta} Y] + \mathrm{Var}_\theta[E_{Y|\theta} Y]$$

Example: Liver disease (cont.)

- Let $\hat{\theta}_i = y_i/n_i$. Then,

$$E(\hat{\theta}_i) = \mu,$$
$$\text{Var}(\hat{\theta}_i) = \frac{\mu(1-\mu)}{n_i}\left(1 + \frac{n_i - 1}{M+1}\right).$$

$$E_y(\hat{\theta}) = E_\theta E_{y|\theta} \hat{\theta} = E_\theta \theta = \mu$$

$$V_x \hat{\theta} = E_\theta V_{x|\theta} \hat{\theta} + V_\theta E_{y|\theta} \hat{\theta}$$

$$= E_\theta \frac{\theta(1-\theta)}{n} + V_\theta \theta$$

$$= \frac{1}{n}(E_\theta \theta - E_\theta \theta^2) + V_\theta \theta$$

$$= \frac{1}{n}\left(\mu - \left((E_\theta \theta)^2 + V_\theta \theta\right)\right) + V_\theta \theta$$

$$= \frac{1}{n}(\mu - \mu^2) + \frac{n-1}{n} V_\theta \theta$$

$$= \frac{\mu(1-\mu)}{n} + \frac{n-1}{n} \frac{\mu(1-\mu)}{m+1}$$

$$= \frac{\mu(1-\mu)}{n}\left(1 + \frac{n-1}{m+1}\right)$$

Example: Liver disease (cont.)

- Consider "iterated method of moments" estimator based on $y_1, \ldots, y_m$

$$E(\hat{\theta}_i) \approx \hat{\mu} = \frac{1}{m} \sum_{j=1}^{m} \frac{y_j}{n_j}$$

Plugging in $\hat{\mu}$, note that

$$s^2 = \frac{1}{m} \sum_{i=1}^{m} \text{Var}(\hat{\theta}_i) = \frac{1}{m} \sum_{i=1}^{m} \frac{\hat{\mu}(1-\hat{\mu})}{n_i} \left(1 + \frac{n_i - 1}{M+1}\right) \qquad (1)$$

Equating (1) with the sample variance of the $\hat{\theta}_i'$s, s^2, gives

$$\hat{M} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{\cancel{N}\; n} \sum_{i=1}^{\cancel{N}\; n} 1/n_i}.$$

Example: Liver disease (cont.)

- Can solve for $\hat{a}$ and $\hat{b}$:

$$\hat{a} = \hat{\mu}\hat{M}$$

$$\hat{b} = \frac{\hat{a}(1-\hat{\mu})}{\hat{\mu}}$$

$\theta_i \sim Beta(\hat{a}, \hat{b})$

- Note that $\tilde{\theta}_i = E(\theta_i \mid y_i, n_i, \hat{M}, \hat{\mu}) =$

$\theta_i | y_i, \ldots \sim Beta(\hat{a}+y_i, \hat{b}+n_i-y_i)$

$E(\theta_i | y_i, \ldots) = \frac{\hat{a}+y_i}{\hat{a}+y_i+\hat{b}+n_i-y_i}$

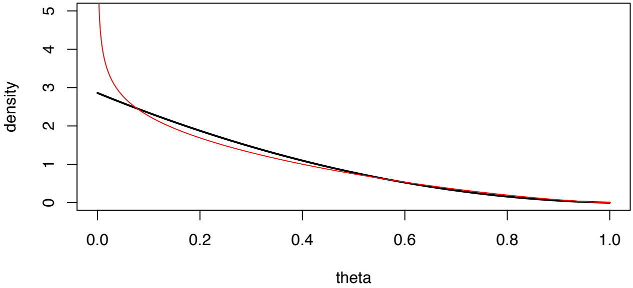
$= \ldots = \frac{\hat{\mu}\hat{M}+y_i}{\hat{M}+n_i}$

- Can write $\tilde{\theta}_i = \hat{B}_i\hat{\mu} + (1-\hat{B}_i)\hat{\theta}_i$

   - The "shrinkage factor" is $\hat{B}_i = \frac{\hat{M}}{\hat{M}+n_i}$

Example: Liver disease (cont.)

- For mouse liver data $\hat{a} = 0.81, \hat{b} = 2.35$
- Plot of empirical prior Beta$(0.81, 2.35)$ red, with formerly obtained posterior predictive for θ_{51} using a discrete uniform prior for $a = 1, 2, 3$ and $b = 1, 2, 3$:



- - http://www.ericfrazerlock.com/Emperical_Bayes_Approaches_Rcode1.r

Non-parametric empirical Bayes

- A non-parametric empirical Bayes method does not assume a parametric form for p_θ.
- Base inference directly on marginal distribution

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \theta) p(\theta) d\theta$$

- Example: Robbins' method.
  - If $Y_i \mid \theta_i \sim \text{Poisson}(\theta_i)$, and θ_i have unspecified prior. Estimate

$y_i | \theta_i$

$$\hat{\theta}_i = (y_i + 1) \frac{\#ys \text{ equal to } y_i + 1}{\#ys \text{ equal to } y_i}$$

$\theta_i \sim p_\theta$ for $i = 1, \ldots, n$

$$P(y_i|\theta_i) = \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!}, \quad P(\theta_i|y_i) = \frac{\frac{P(y_i|\theta_i)\cdot P(\theta_i)}{\int P(y_i|\theta_i) P(\theta_i)\, d\theta_i}}{P(y_i)}$$

$$E(\theta_i|y_i) = \int \theta_i P(\theta_i|y_i)\, d\theta_i$$

$$= \int \frac{\theta_i \left(\frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!}\right)\cdot P(\theta_i)}{P(y_i)}\, d\theta_i \cdot \left(\frac{y_i+1}{y_i+1}\right)$$

$$= (y_i+1) \frac{\int \frac{\theta_i^{y_i+1} e^{-\theta_i}}{(y_i+1)!} P(\theta_i)\, d\theta_i}{P(y_i)} = \frac{(y_i+1) P(y_i+1)}{P(y_i)}$$

Let $\hat{m}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i = y\}$

$$\hat{\theta}_i = (y_i + 1) \cdot \frac{\hat{m}(y_i + 1)}{\hat{m}(y_i)}$$