Bayesian Statistics
Note 4
Hierarchical and Empirical Bayes Analysis

Keunbaik Lee

Sungkyunkwan University

Exchangeability I

- Let $\mathcal{K} = \{k_1, k_2, \cdots, k_n\}$ be a permutation of $\mathcal{L} = \{1, 2, \cdots, n\}$.
- Example: $\mathcal{L} = \{1, 2, 3\}$. Permutations: $\mathcal{K} = \{2, 1, 3\}$, $\mathcal{K} = \{2, 3, 1\}$, $\mathcal{K} = \{3, 2, 1\}$, $\cdots$.
- $X_1, X_2, \cdots, X_n$ are *exchangeable* stochastic variables if $X_{k_1}, X_{k_2}, \cdots, X_{k_n}$ has the same joint distributions for all $n!$ permutations of $\{k_1, k_2, \cdots, k_n\}$.
- Example: $n = 2$, X_1, X_2 are exchangeable if

$$P(X_1 = a, X_2 = b) = P(X_2 = a, X_1 = b).$$

- *iid* $=>$ Exchangeability, but the opposite does not always hold. Exchangeability is less restrictive than *iid*.

Exchangeability II

- Example: Urn with m marbles: r white and $m - r$ black. Draw $n \leq m$ marbles without replacement

$$X_i = \begin{cases} 1, & \text{if } i\text{th draw gives black marble;} \\ 0, & \text{if } i\text{th draw gives white marble.} \end{cases}$$

for $i = 1, \cdots, n$. Then $X_1, \cdots, X_n$ are exchangeable, but not *iid*.

Exchangeability III

- de Finetti's Theorem (1930):
  $X_1, X_2, \cdots, X_n$ are exchangeable stochastic variables. Then there is guaranteed to exist parameter θ, its probability density function $\pi(\theta)$, and conditional probability density function $P(x_i|\theta)$ such that

$$P(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} P(x_i|\theta)\pi(\theta)d\theta.$$

  Note: When $X_1, X_2, \cdots, X_n$ are exchangeable, there exists conditional i.i.d. distribution $\prod_{i=1}^{n} P(x_i|\theta)$ and the prior distribution $\pi(\theta)$ for θ.

Hierarchical Models I

Multicenter Clinical Trial

A cardiac treatment is used for patients in several hospitals.
Let $\theta_i =$probability of survival of a patient for the ith hospital
$y_i =$sample proportion of patients surviving in the ith hospital.
Suppose there are n_i patients who are given the cardiac treatment in the ith hospital

$$n_i y_i | \theta_i \sim^{ind} B(n_i, \theta_i).$$

It is natural to expect that the estimates of the θ_i's should be related to each other.

Hierarchical Models II

General Theme

There are multiple parameter which are related or connected in a certain way. If we can use a prior for the parameters which builds the dependence, then that serves the purpose.

Further Examples

1. Proportions of defectives in a series of lots of parts from the same supplier.
2. Mean bushels of corn per acre for a random selection of farms from a given county.
3. Mean worker accident rates from a sample of similar companies in a given industry.

Hierarchical Models III

4. Bone loss rates of a random sample of individuals in a osteoporosis high risk group.
5. Estimating simultaneously the proportion of poor children in the age-group 5-17 for all the counties in Florida.

Hierarchical Models IV

A Mathematical Formulation

Likelihood: $y_1, \cdots, y_n \mid \theta_1, \cdots, \theta_n, \lambda \ \sim^{ind} P(y_i|\theta_i)$

Prior (Stage 1): $\theta_1, \cdots, \theta_n \mid \lambda \ \sim^{ind} P(\theta_i|\lambda)$

Prior (Stage 2): λ has prior $P(\lambda)$

Note that the joint prior of $\theta_1, \cdots, \theta_n$ is

$$P(\theta_1, \cdots, \theta_n) = \int \prod_{i=1}^{n} P(\theta_i|\lambda)P(\lambda)d\lambda.$$

They are no longer independent

Hierarchical Models V

Lindley and Smith (JRSS-B, 1972)

$$y_i|\theta_i,\mu \sim^{ind} N(\theta_i,\sigma^2), \quad \sigma^2(>0) \text{ known}$$
$$\theta_i|\mu \sim^{iid} N(\mu,\tau^2), \quad \tau^2(>0) \text{ known}$$
$$\mu \sim \text{uniform}(-\infty,\infty).$$

Goal: Find the joint posterior of $\theta_1,\cdots,\theta_n$ given $y_1,\cdots,y_n$.

$$P(\theta_1,\cdots,\theta_n,\mu|y_1,\cdots,y_n) \propto e^{-\frac{1}{2\sigma^2}\sum_i(y_i-\theta_i)^2} e^{-\frac{1}{2\tau^2}\sum_i(\theta_i-\mu)^2}$$

Hierarchical Models VI

Note that

$$\begin{aligned}\sum_i (\theta_i - \mu)^2 &= n\mu^2 - 2n\mu\bar{\theta} + \sum_i \theta_i^2 \\ &= n(\mu - \bar{\theta})^2 + \sum_i (\theta_i - \bar{\theta})^2.\end{aligned}$$

Integrating out μ,

$$P(\theta_1, \cdots, \theta_n | y_1, \cdots, y_n) \propto e^{-\frac{1}{2\sigma^2}\sum_i (\theta_i - y_i)^2} e^{-\frac{1}{2\tau^2}\sum_i (\theta_i - \bar{\theta})^2}.$$

Write $\sum_i (\theta_i - y_i)^2 = \theta^T\theta - 2\theta^T y + y^T y$ where

$$\theta^T = (\theta_1, \cdots, \theta_n), \quad y^T = (y_1, \cdots, y_n).$$

Hierarchical Models VII

$$\sum_i (\theta_i - \bar{\theta})^2 = \theta^T \left(I_n - \frac{1}{n} J_n \right) \theta, \quad J_n = 1_n 1_n^T$$

where $1_n^T = (1, \cdots, 1)$.

$$\begin{aligned} P(\theta_1, \cdots, \theta_n | y_1, \cdots, y_n) &\propto e^{-\frac{1}{2\sigma^2}(\theta^T\theta - 2\theta^T y)} e^{-\frac{1}{2\tau^2}\theta^T \left(I_n - \frac{1}{nJ_n}\right)\theta} \\ &\propto e^{-\frac{1}{2}\left(\theta^T D\theta - \frac{2}{\sigma^2}\theta^T y\right)} \\ &\propto e^{-\frac{1}{2}\left(\theta - \frac{1}{\sigma^2}D^{-1}y\right)^T D\left(\theta - \frac{1}{\sigma^2}D^{-1}y\right)} \end{aligned}$$

where $D = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right) I_n - \frac{1}{\tau^2}\left(\frac{1}{n} J_n\right)$.
Then $\theta | y \sim N(\frac{1}{\sigma^2} D^{-1} y, D^{-1})$.

Hierarchical Models VIII

Examples

1. **Hierarchical binomial model**
   - Example:
     $$y_j|\theta_j \sim Bin(n_j, \theta_j), \quad j = 1, \cdots, J$$
   - We could do inference on each θ_j separately. Problem: n_j may be small for some j. Not much information then about θ_j.
   - If you knew θ_j, would that give information about θ_i, $i \neq j$? If so, then inference about the parameters θ_j, $j = 1, \cdots, J$, may 'borrow strength' from each other.
   - Extreme case: assume $\theta_j = \theta$ for all j. Define $y = \sum_{j=1}^{J} y_j$ and $n = \sum_{j=1}^{J} n_j$. Straightforward to analyze θ with the usual Beta-Binomial approach.

Hierarchical Models IX

- Intermediate case: tie the θ's together by assuming a superpopulation/prior

$$\theta_j \sim^{iid} Beta(\alpha, \beta).$$

- Model summary

$$\begin{aligned} y_j|\theta_j &\sim Bin(n_j, \theta_j), \quad j = 1, \cdots, J \\ \theta_j &\sim Beta(\alpha, \beta) \\ \alpha &\sim Gamma(a_1, a_2) \\ \beta &\sim Gamma(b_1, b_2). \end{aligned}$$

- Sample from the joint posterior of $P(\theta, \alpha, \beta|y) = P(\theta|\alpha, \beta, y)P(\alpha, \beta|y)$ by sampling from:

Hierarchical Models X

- $\theta_j|\alpha, \beta, y$, $j = 1, \cdots, J$, which are independent *Beta* distributions.
- $P(\alpha, \beta|y)$ can be derived in closed form but cannot be sampled directly. Evaluate on grid and sample.

- Use WinBUGS.

2 **The one-way normal random effects model**

- Consider a model:

$$y_j|\theta_j \sim N(\theta_j, \sigma_j^2), \quad \sigma_j^2 \text{ known.}$$

- At on extreme: we may estimate each θ_j using the mean $\bar{y}_j$ of observations in the jth group.
- At the other extreme: we may assume $\theta_j = \theta$ for all j. Estimate θ with a pooling of group means $\bar{y}_j$.

Hierarchical Models XI

- Intermediate: a hierarchical model

$$\begin{aligned} & y_j|\theta_j \sim N(\theta_j, \sigma_j^2), \quad \sigma_j^2 \text{ known} \\ & \theta_j|\mu,\tau \sim N(\mu,\tau^2) \\ & P(\mu,\tau) = P(\mu|\tau)P(\tau) \propto P(\tau). \end{aligned}$$

- Here we do not assume equal group mean, yet the estimates of each θ_j borrow strength from each other.
- Use WinBUGS.

Empirical Bayes (EB) I

Instead of putting a uniform$(-\infty, \infty)$ prior for μ, estimate μ from the marginal distribution of y.

$$
\begin{aligned}
&y_1, \cdots, y_n \mid \theta_1, \cdots, \theta_n \sim^{ind} N(\theta_i, \sigma^2), \\
&\theta_1, \cdots, \theta_n \sim^{iid} N(\mu, \tau^2).
\end{aligned}
$$

Bayes estimator of θ is $\theta^{EB} = (1-B)y + B\mu 1_n$, $B = \frac{\sigma^2}{\sigma^2+\tau^2}$.
Marginally, $y_1, \cdots, y_n \sim^{iid} N(\mu, \tau^2 + \sigma^2)$. Estimate μ by $\bar{y}$.
Thus, EB estimator of θ is

$$
\begin{aligned}
&\hat{\theta}^{EB} = (1-B)y + B\bar{y}1_n = \hat{\theta}^{HB} = E(\theta|y) \\
&\theta|y \sim N((1-B)y + B\mu 1_n, \sigma^2(1-B)I_n).
\end{aligned}
$$

Empirical Bayes (EB) II

However, $V(\theta|y)$ is still estimated by $\sigma^2(1-B)I_n$. However, with the hierarchical Bayes procedure

$$V(\theta|y) = \sigma^2\left[(1-B)I_n + \frac{B}{n}J_n\right].$$

Thus,

$$var(\theta|y)^{(EB)} \leq var(\theta|y)^{(HB)}.$$

Empirical Bayes (EB) III

Consider again the HB model

$$
\begin{aligned}
& y|\theta, \mu \sim N(\theta, \sigma^2 I_n). \\
& \theta|\mu \sim N(\mu 1_n, \tau^2 I_n). \\
& \mu \sim \text{ uniform}(-\infty, \infty).
\end{aligned}
$$

$$
\begin{aligned}
V(\theta|y) &= E\left[V(\theta|y,\mu)|y\right] + V\left[E(\theta|y,\mu)|y\right] \\
&= E\left[\sigma^2(1-B)I_n|y\right] + V\left[(1-B)y + B\mu 1_n|y\right] \\
&= \sigma^2(1-B)I_n + B^2 V(\mu|y) J_n. \qquad (1)
\end{aligned}
$$

Empirical Bayes (EB) IV

Then

$$y_1, \cdots, y_n|\mu \sim N(\mu, \sigma^2 + \tau^2)$$
$$\mu \sim \text{ uniform}(-\infty, \infty).$$

The posterior distribution of μ given y is

$$\mu|y \sim N\left(\bar{y}, \frac{\sigma^2 + \tau^2}{n}\right) \equiv N\left(\bar{y}, \frac{\sigma^2}{nB}\right).$$

Since $B^2 V(\mu|y) J_n = B^2 \frac{\sigma^2}{nB} J_n = \frac{B\sigma^2}{n} J_n$, (1) is

$$V(\theta|y) = \sigma^2(1 - B) I_n + \frac{B\sigma^2}{n} J_n.$$

Empirical Bayes (EB) V

The EB procedure ignores the second term which is due to the fact that it fails to incorporate the uncertainty due to estimation of μ. The HB procedure rectifies that mistake by introducing a distribution (through flat) for μ.

Empirical Bayes (EB) VI

Empirical Bayes vs Hierarchical Bayes

- An EB scenario is one in which known relationships or structure of the parameter vector $\theta_1, \cdots, \theta_n$ allows one to estimate some features of the distribution $P(\theta|\lambda)$, the prior distribution. For example, one may believe that the prior is structurally known except for certain unknown parameter λ.
- An EB procedure is one which estimates λ from the marginal distribution of the observations and works with the estimated prior $P(\theta|\hat{\lambda})$.

Empirical Bayes (EB) VII

- Closely related to the EB procedure is the HB procedure which also recognize the uncertainty in the prior information. However, which the EB procedure estimates the prior parameters (hyperparameter) λ from the marginal distribution of the y_i's, the HB procedure assigns a prior (often diffuse) for λ.
- Typically the point estimates obtained under the two procedures are fairly close. However, an EB procedure typically underestimates the uncertainty as compared to the HB procedure by not taking into account the uncertainty due to estimation of the hyperparameters.