STAT 576 Bayesian Analysis

Lecture 5: Hierarchical Models

Chencheng Cai

Washington State University

Hierarchical Models

- Many statistical applications involve multiple parameters that are **related**.
- Example: (Multi-center study on the effectiveness of a drug)
  - Hospitals $j = 1, \ldots, J$.
  - Patients in hospital j has a probability of recovering of θ_j.
  - The $\theta_1, \ldots, \theta_J$ should be related.
- We use a prior distribution in which the θ_j's are viewed as a sample from the **population distribution**.
- If we observe y_{ij}, $i = 1, \ldots, n_j$ for hospital $j = 1, \ldots, J$.
- we can use the data y_{ij} to estimate aspects of the population distribution of θ.
- If furthermore, we approximation the population distribution by a parametric family, the corresponding parameters are called **hyperparameters**.

Example: Analysis using Historical Data

- Goal: estimate the probability of tumor in a population of female laboratory rats.
- Observation: $4/14$ rats show symptom of a tumor.
- Model: We assume the obervational model follows a binomial distribution:

$$y \sim \text{Binom}(14, \theta)$$

- with a conjugate prior for θ as $\text{Beta}(\alpha, \beta)$.
- The corresponding posterior is $\theta \mid y = 4 \sim \text{Beta}(\alpha + 4, \beta + 10)$.
- So far, the values for α and β are arbitrary.
- If we have a historical records of previous experiments, we can have better choices for α and β if we interpret the prior distribution as the population distribution.

Example: Analysis using Historical Data

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

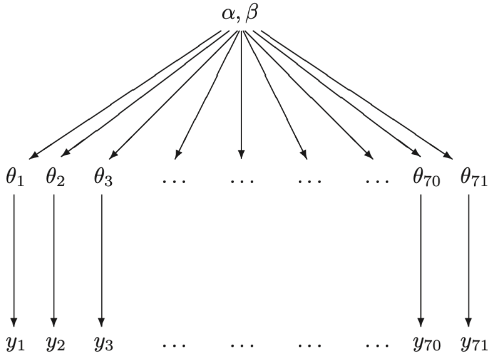
4/14

- For $j = 1, \ldots, 70$ experiments, we observed y_j out of n_j rats with the symptom.
- The estimated mean and standard deviation for y_j/n_j are 0.136 and 0.103.
- We may choose the hyperparameters (α, β) by (**Variance is overestimated!!**)

$$\frac{\alpha}{\alpha+\beta} = 0.135, \quad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.103^2.$$

Example: Analysis using Historical Data

- The solution is $\alpha = 1.4$, $\beta = 8.6$.
- The prior is Beta$(1.4, 8.6)$.
- The posterior is Beta$(5.4, 18.6)$.
- The posterior mean is 0.223 and the posterior s.d. is 0.083.



Note: the calculation demonstrated here is not a Bayesian calculation!

Example: Analysis using Historical Data

The demonstrated calculation is based on the assumption that the risk θ from the current observation is considered a **random** sample from a **common** distribution as previous 70 experiments.

Violations of this assumption:

- The risk θ changes over time. (More common for econometric and financial data).
- The historical data has other covariates.
    - Data collected in different hospitals/labs/centers.
    - Data collected for different sub-populations.

We may elaborate those factors into a more complicated model.

Example: Analysis using Historical Data

Question: Can we use the prior Beta$(1.4, 8.6)$ to do Bayesian inference for the 70 historical studies?

No!!

Remarks:

- The data is used twice: (1) estimating the prior (2) computing each's posterior.
- We ignored the uncertainty in estimating α and β. (In oppose to Bayesian inference, where the posterior measures uncertainty.)
- The prior distributions should be known **before** observing any data. Shall we really estimate them?

Example: Analysis using Historical Data

In addition:

- It definitely makes sense to estimate the population distribution from all the data than to estiamte them separately.
- The posterior for experiment j_1 and j_2 ($j_1 \neq j_2$) should be dependent because they are studying the same object.
- In order to retain the advantage of the hierarchical model and to get rid of the aforementioned trouble, we will build a full probability model for all parameters.
- The analysis using the data to estimate the prior parameters, which is sometimes called **empirical Bayes**, can be viewed as an approximation to the complete hierarchical Bayesian analysis.

Hierarchical Modeling

- Assume we have J experiments.
- For each experiment j, we have observation y_j, parameter θ_j and likelihood $p(y_j \mid \theta_j)$.
- If there is no additional information other than the observations y_j's, we assume the **exchangeability** of the parameters, that is

$$p(\theta_1, \ldots, \theta_J) \sim p(\theta_{\pi(1)}, \ldots \theta_{\pi(J)}),$$

  for any permutation $\pi : \{1, \ldots, J\} \to \{1, \ldots, J\}$.
- Furthermore, inspired by the De Finetti's Theorem, we can construct the prior on (θ_1, θ_J) in the following way:

$$\phi \sim p(\phi), \quad \theta_1, \ldots, \theta_J \mid \phi \sim p(\theta \mid \phi) \ i.i.d.$$

- Or in other words,

$$p(\theta_1, \ldots, \theta_J) = \int \left(\prod_{j=1}^{J} p(\theta_j \mid \phi) \right) p(\phi) d\mu(\phi)$$

Hierarchical Modeling

Back to the rat tumor example.

- If without any additional information on the historical data, we assume

$$\phi \sim p(\phi), \quad \theta_1, \ldots, \theta_{70} \mid \phi \sim p(\theta \mid \phi) \ i.i.d.$$

- If the experiments were conducted at 5 different centers, we assume (two-level hierarchical model)

$$\psi \sim p(\psi), \quad \phi_1, \ldots, \phi_5 \mid \psi \sim p(\phi \mid \psi) \ i.i.d., \quad \theta_{1j}, \ldots, \theta_{14j} \mid \phi_j \sim p(\theta \mid \phi_j) \ i.i.d.$$

- If each experiment j is equiped with covariate x_j, we assume

$$\phi \sim p(\phi), \quad \theta_1, \ldots, \theta_{70} \mid \phi, x_1, \ldots, x_{70} \sim \prod_{j=1}^{70} p(\theta_j \mid \phi, x_j)$$

Hierarchical Modeling

- Now the complete model is

$$
\begin{aligned}
&\phi \sim p(\phi) \\
&\theta_j \sim p(\theta \mid \phi) \ i.i.d. \text{ for } j = 1, \ldots, J \\
&y_j \sim p(y \mid \theta_j) \text{ independent for } j = 1, \ldots, J
\end{aligned}
$$

- The joint prior distribution:

$$p(\phi, \theta_1, \ldots, \theta_J) = p(\phi) \prod_{j=1}^{J} p(\theta_j \mid \phi)$$

- The observation model:

$$p(y_1, \ldots, y_J \mid \phi, \theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} p(y_j \mid \theta_j)$$

- The joint posterior distribution:

Hierarchical Modeling

- The distribution $p(\phi)$ is the "prior" distribution for the hyperparameter ϕ, which is called the **hyperprior** distribution.
- Due to the complexity in $\theta_1, \ldots, \theta_J$, it is often more convenient to look at the marginal posterior distribution for the hyperparameter.
- We often adopt a hybrid approach both analytically and numerically to conduct Bayesian inference.
- **Step 1 (analytic)**: get the marginal posterior distribution for ϕ.
- **Step 2 (numerical)**: draw samples of $(\phi, \theta_1, \ldots, \theta_J)$ from the joint posterior distribution.

Hierarchical Modeling

Step 1 Procedure:

1. Get the posterior in proportinal form:

$$p(\phi, \theta_1, \ldots, \theta_J \mid y_1, \ldots, y_J) \propto p(\phi) \prod_{j=1}^{J} p(y_i \mid \theta_j) p(\theta_j \mid \phi)$$

2. Determine the conditional posterior distribution of $(\theta_1, \ldots, \theta_J)$:

$$p(\theta_1, \ldots, \theta_J \mid \phi, y_1, \ldots, y_J) = A(\phi, y_1, \ldots, y_J) p(\phi) \prod_{j=1}^{J} p(y_i \mid \theta_j) p(\theta_j \mid \phi)$$

   for some normalizing coefficient A.

3. Determine the marginal posterior distribution of ϕ by

$$p(\phi \mid y_1, \ldots, y_n) = \frac{p(\phi, \theta_1, \ldots, \theta_J \mid y_1, \ldots, y_n)}{p(\theta_1, \ldots, \theta_J \mid \phi, y_1, \ldots, y_n)} \propto [A(\phi, y_1, \ldots, y_J)]^{-1}$$

Step 1 is analytical because $p(\theta_j \mid \phi)$ is chosen conjugate to $p(y_j \mid \theta_j)$.

Hierarchical Modeling

Step 2 Procedure:

1. Draw samples of ϕ from the marginal posterior distribution $p(\phi \mid y_1, \ldots, y_J)$.
2. Draw samples of $(\theta_1, \ldots, \theta_J)$ from the conditional distribution $p(\theta_1, \ldots, \theta_J \mid \phi, y_1, \ldots, y_J)$. This step can be done coordinate-wise because

$$p(\theta_1, \ldots, \theta_J \mid \phi, y_1, \ldots, y_J) \propto \prod_{j=1}^{J} p(y_i \mid \theta_j) p(\theta_j \mid \phi)$$

- With the samples from the joint posterior, we can estimate the posterior mean, median or other Bayesian estimators based on the empirical loss.
- To generate a prediction,
  - to predict a new observation for experiment j: draw new $\tilde{y}_j$ given a sample of θ_j.
  - to predict a new observation for a new experiment:
    (1) draw a new $\tilde{\theta}$ given a sample of ϕ
    (2) draw a new $\tilde{y}$ given $\tilde{\theta}$.

Example: Rat Tumor Risk

- Observation Model:

$$y_j \sim \text{Binom}(n_j, \theta_j), \text{ for } j = 1, \ldots, J.$$

- Joint prior distribution:

$$\alpha, \beta \sim p(\alpha, \beta), \quad \theta_j \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta) \text{ for } j = 1, \ldots, J.$$

that is,

$$p(\alpha, \beta, \theta_1, \ldots, \theta_J) = p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha - 1} (1 - \theta_j)^{\beta - 1}$$

- Joint posterior distribution:

$$p(\alpha, \beta, \theta_1, \ldots, \theta_J \mid y_1, \ldots, y_J) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n - y_j - 1}$$

Example: Rat Tumor Risk

- The conditional posterior of $\theta_1, \ldots, \theta_J$ is

$$p(\theta_1, \ldots, \theta_J \mid \alpha, \beta, y_1, \ldots, y_J) \propto \prod_{j=1}^{J} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n-y_j-1}$$

  This is the joint density of J independent $\text{Beta}(\alpha + y_j, \beta + n - y_j)$ distributions.
- The density is

$$p(\theta_1, \ldots, \theta_J \mid \alpha, \beta, y_1, \ldots, y_J) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y_j)\Gamma(\beta+n-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n-y_j-1}$$

- Then the marginal posterior distribution for α, β is

$$\begin{aligned}
p(\alpha, \beta \mid y_1, \ldots, y_J) &= \frac{p(\alpha, \beta, \theta_1, \ldots, \theta_J \mid y_1, \ldots, y_J)}{p(\theta_1, \ldots, \theta_J \mid \alpha, \beta, y_1, \ldots, y_J)} \\
&\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n-y_j)}{\Gamma(\alpha+\beta+n)}
\end{aligned}$$

Example: Rat Tumor Risk

- It is difficult to calculate the Fisher information matrix for α, β. Therefore, we choose the prior in an ad-hoc way:

$$p\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right) \propto 1$$

  That is,

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$$

- Or on the natural transformed scale:

$$p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}$$

- The marginal posterior:

$$p(\alpha, \beta \mid y_1, \ldots, y_J) \propto (\alpha+\beta)^{-5/2} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n-y_j)}{\Gamma(\alpha+\beta+n)}$$

- Remark: assigning a uniform prior on the natural transformed scale results in an improper posterior distribution

Normal with Exchangeable Parameters

- Suppose we have J independent experiments with the observations y_{ij} follows (with known σ^2)

$$y_{ij} \sim \mathcal{N}(\theta_j, \sigma^2) \text{ for } i = 1, \dots, n_j;\ j = 1, \dots, J.$$

- For each experiment, the sample mean is a sufficient statistics,

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

  with $\sigma_j^2 = \sigma^2 / n_j$.
- A conjugate prior is a normal distribution for θ's:

$$\theta_j \mid \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2)$$

- To a make it a full probability model, we need to assign hyperprior to (μ, τ). We assume

$$p(\mu, \tau) \propto p(\tau)$$

Normal with Exchangeable Parameters

- Let $y = \{y_{ij} : i = 1, \dots, n_j; j = 1, \dots, J\}$, $\theta = (\theta_1, \dots, \theta_J)$.
- Now the joint prior distribution is

$$p(\mu, \tau, \theta) \propto p(\tau)\tau^{-J} \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j - \mu)^2\right\}$$

- The observation model is

$$p(y \mid \mu, \tau, \theta) \propto \prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_j - \theta_j)^2\right\} \propto \exp\left\{-\sum_{j=1}^{J}\frac{1}{2\sigma_j^2}(\bar{y}_j - \theta_j)^2\right\}$$

- The joint posterior is

$$p(\mu, \tau, \theta \mid y) \propto p(\tau)\tau^{-J} \exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{(\mu - \bar{y}_j)^2}{\tau^2 + \sigma_j^2}\right\} \prod_{j=1}^{J} \exp\left\{-\frac{1}{2V_j}(\theta_j - \hat{\theta}_j)^2\right\}$$

$$\hat{\theta}_j = \frac{\frac{\mu}{\tau^2} + \frac{\bar{y}_j}{\sigma_j^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma_j^2}}, \quad V_j = \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma_j^2}}$$

Normal with Exchangeable Parameters

- The conditional posterior for θ is therefore

$$p(\theta \mid \mu, \tau, y) \propto \prod_{j=1}^{J} \exp\left\{-\frac{1}{2V_j}(\theta_j - \hat{\theta}_j)^2\right\}$$

  which is the density for J independent $\mathcal{N}(\hat{\theta}_j, V_j)$ variables.

- The density is

$$p(\theta \mid \mu, \tau, y) = \prod_{j=1}^{J} (2\pi V_j)^{-1/2} \exp\left\{-\frac{1}{2V_j}(\theta_j - \hat{\theta}_j)^2\right\}$$

- Hence, the marginal posterior for μ, τ is

$$p(\mu, \tau \mid y) = \frac{p(\mu, \tau, \theta \mid y)}{p(\theta \mid \mu, \tau, y)} \propto p(\tau)\left(\prod_{j=1}^{J}(\tau^2 + \sigma_j^2)\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{(\mu - \bar{y}_j)^2}{\tau^2 + \sigma_j^2}\right\}$$

Normal with Exchangeable Parameters

$$p(\mu, \tau \mid y) = \frac{p(\mu, \tau, \theta \mid y)}{p(\theta \mid \mu, \tau, y)} \propto p(\tau) \left(\prod_{j=1}^{J} (\tau^2 + \sigma_j^2) \right)^{-1/2} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{J} \frac{(\mu - \bar{y}_j)^2}{\tau^2 + \sigma_j^2} \right\}$$

- Now it is immediate:

$$\mu \mid \tau, y \sim \mathcal{N}(\hat{\mu}, V_\mu)$$

  with

$$V_\mu = \left(\sum_{j=1}^{J} \frac{1}{\tau^2 + \sigma_j^2} \right)^{-1} , \quad \hat{\mu} = V_\mu \sum_{j=1}^{J} \frac{\bar{y}_j}{\tau^2 + \sigma_j^2}$$

- Furthermore, the marginal posterior for τ is

$$p(\tau \mid y) \propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\tau^2 + \sigma_j^2)^{-1/2} \exp \left\{ -\frac{(\hat{\mu} - \bar{y}_j)^2}{2(\tau^2 + \sigma^2)} \right\}$$

- One can choose $p(\tau) \propto 1$. (Note that $p(\tau) \propto \tau^{-1}$ results in an improper posterior)