Bayesian hierarchical models

Hao Zhang
Xiaoshen Li
Yao Ji

Overview

- basic ideas about bayesian hierarchical models
- exchangeability
- bayesian analysis of conjugate hierarchical model
- bayesian analysis of normal hierarchical model

Basic ideas about bayesian hierarchical models

- Hyperparameter: parameter of the prior distribution
- Hyperprior: the prior distribution of hyperparameter

- Say a random variable Y follows a normal distribution with parameter θ as the mean and 1 as the variance, that is $Y \mid \theta \sim N(\theta, 1)$
- The parameter θ has a prior distribution given by a normal distribution with mean μ and variance 1, i.e. $\theta \mid \mu \sim N(\mu, 1)$. Till now,it's the bayesian model we
- are familiar with.
- Furthermore,if μ follows another distribution given, for example, by the standard normal distribution, $\mathrm{N}(0, 1)$.
- then the parameter μ is called the hyperparameter, while its distribution given by N(0,1) is an example of a hyperprior distribution.

- Bayesian analysis is about estimating parameters or something else.Significantly different from the traditional approach,by which we can get a point estimation of a parameter,bayesian model grants us a posterior density just by some simple assumptions.
- In bayesian hierarchical models,by assigning a hyperprior,we can finally get a joint posterior density,and a marginal posterior density for hyper parameter as well.

Example

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

Examples:To estimate the risk of tumor in a population of rats,scientists took an expriment like this.These data shows the outcome of historical control groups and the current group of rats.The numerator is the incidence of tumor in a group and the denominator is the amount of rats in a group.
Given a prior distribution and a likelihood,we can derive a posterior distribution given hyperparameter.Bayesian hierarchical model can be used to derive a hyperposterior once given a hyperprior.
First of all,we'd like to introduce a conventional way to estimate the hyperparameters—point estimation.

Point estimation for hyerparameter

Let vector (x1.....xn) denotes the incidence of tumor in each group

$$x_1, \ldots, x_n | \theta \overset{iid}{\sim} Bern(\theta)$$

$$\theta \sim Beta(\alpha, \beta)$$

from the previous lecture,we know the posterior of θ given hyperparameter follows a distribution like this:

$$Beta(\alpha + s, \beta + f)$$

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|\theta)p(\theta) \\ &= \theta^s(1-\theta)^f\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}. \end{aligned}$$

In our case,s is the number of tumor incidence in the current group,which equals 4,and f equals 10.Therefore the posterior for for θ follows Beta(α + 4, β + 10).

Approximate estimation for hyperparameter

- Using the frequency of tumor incidence in each historical control group,we can get a point estimation for α and β.
- By calculation,we get the mean value and standard deviation of $\frac{y_j}{n_j}$ is 0.136 and 0.103.
- The mean value and variance of a Beta(α, β) distribution is:

$$\mathrm{E}(\theta) = \frac{\alpha}{\alpha+\beta}$$
$$\mathrm{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Let the population mean value equals 0.136 and the variance equals 0.103,we can derive the point estimation for (α, β) is (1.4, 8.6).
- Therefore, θ 71 yields a Beta(5.4, 18.6) posterior distribution.

The problem of a point estimation

- These analyses require that the current tumor risk, $\theta 71$, and the 70 historical tumor risks, $\theta 1, \ldots, \theta 70$, be considered a random sample from a common distribution beta(α, β) which is an assumption that would be invalidated.Even if this assumption is validated,the frequency of each historical group can't be regarded as the observated value of θ.
- In one word, the point estimation for hyperparameter can only be treated as a rough approximation.
- However,bayesian hierarchical model offers us a solid estimation for hyperparameter.

Why we use exchangeability?

Generalizing from the example of the previous section, consider a set of experiments j=1,...,J,in which experiment j has data (vector) y_j and parameter (vector) θ_j,with likelihood

$p(y_j|\theta_j)$.Some of the parameters in different experiments may overlap.for example, each data vector y_j may be a sample of

observations from a normal distribution with mean u_j and common variance σ^2.in which case $\theta_j=(u,\sigma^2)$.In order to creat a joint probability model for all the parameters θ,we use the crucial idea of exchangeability.

Exchangeability

If no information—other than the data y—is available to distinguish any of the θ_j's from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution. This symmetry is of exchangeability.For example, rolling of a die.

Generally, the less we know about a problem, the more confidently we can make claims of exchangeability. the parameters $(\theta_1,.....,\theta_j)$ are exchangeable in their joint distribution if $p(\theta_1.....,\theta_j)$ is invariant to permutations of the indexes (1, . . . , J).For example, in the rat tumor problem, suppose we have no information to distinguish the 71 experiments, other than the sample sizes n_j , which presumably are not related to the values of θ_j ; we therefore use an exchangeable model for the θ_j 's.

The simplest form of an exchangeable distribution has each of the parameters θ_j as an independent sample from a prior (or population) distribution governed by some unknown parameter vector Φ. thus,

$$p(\theta \mid \phi) = \prod_{j=1}^{J} p(\theta_j \mid \phi)$$

In general, Φ is unknown, so our distribution for θ must average over our uncertainty in Φ:

$$p(\theta) = \int (\prod_{j=1}^{J} p(\theta_j \mid \phi)) p(\phi) d\phi$$

Example. Exchangeability and sampling

We have selected eight states out of the United States and recorded the divorce rate per 1000 population in each state in 1981. Call these $y_1,.....,y_8$. Since you have no information to distinguish any of the eight states from the others,you must model them exchangeably. You might use a beta distribution for the eight y_j 's, a logit normal, or some other prior distribution restricted to the range [0, 1].Unless you are familiar with divorce statistics in the United States, your distribution on $(y_1,.....,y_8)$ should be fairly vague.

1	2	3	4	5	6	7
5.8	6.6	7.8	5.6	7.0	7.1	5.4

y8, would probably be centered around 6.5 and have most of its mass between 5.0 and 8.0. Changing the indexing does not change the joint distribution.If we relabel the remaining value to be any other y_j the posterior estimate would be the same. y_j are exchangeable but they are not independent as we assume that the divorce rate in the eighth unobserved state is probably similar to the observed rates.

Arizon	Colorado	Idaho	Montana	Nevada	New Mexico	Utah	Wyoming

Now, before the seven data points were observed,the eight divorce rates should still be modeled exchangeably.it seems reasonable to assume that Utah, with its large Mormon population, has a much lower divorce rate, and Nevada, with its liberal divorce laws, has a much higher divorce rate, than the remaining six states.Now, even before seeing the seven observed values, you cannot assign an exchangeable prior distribution to the set of eight divorce rates, since you have information that distinguishes y8 from the other seven numbers, here suspecting it is larger than any of the others. Once $y_1, \ldots, y_7$ have been observed, a reasonable posterior distribution for y_8 plausibly should have most of its mass above the largest observed rate.The answer is Nevada's divorce rate in 1981.Incidentally, Nevada's divorce rate in 1981 was 13.9 per 1000 population.

Exchangeability when additional information is available on the units

Often observations are not fully exchangeable, but are partially or conditionally exchange-able:

- If observations can be grouped, we may make hierarchical model, where each group has its own submodel, but the group properties are unknown. If we assume that group properties are exchangeable, we can use a common prior distribution for the group properties.
- If y_i has additional information x_i so that y_i are not exchangeable but (y_i, x_i) still are exchangeable, then we can make a joint model for (y_i, x_i) or a conditional model for $y_i|x_i$.

EXCHANGEABILITY AND HIERARCHICAL MODELS

In general, the usual way to model exchangeability with covariates is through conditional independence:

$$p(\theta_1, \ldots\ldots, \theta_j \mid x_1, \ldots\ldots, x_j) = \int [\prod_{j=1}^{J} p(\theta_j \mid \phi, x_j)] p(\phi \mid x) d\phi$$

In this way, exchangeable models become almost universally applicable, because any information available to distinguish different units should be encoded in the x and y variables.

The full Bayesian treatment of the hierarchical model

Returning to the problem of inference, the key 'hierarchical' part of these models is that Φ is not known and thus has its own prior distribution, p(Φ). The appropriate Bayesian posterior distribution is of the vector (Φ, θ). The joint prior distribution is

$$p(\phi, \theta) = p(\phi)\, p(\theta \mid \phi)$$

And the joint posterior distribution is

$$p(\phi,\theta \mid y) \propto p(\phi,\theta)\, p(y \mid \phi,\theta)$$

$$= p(\phi,\theta)\, p(y \mid \theta)$$

with the latter simplification holding because the data distribution, p(y|Φ, θ), depends only on θ; the hyperparameters Φ affect y only through θ. Previously, we assumed Φ was known,which is unrealistic; now we include the uncertainty in Φ in the model.

EXCHANGEABILITY AND HIERARCHICAL MODELS

- In the rat tumor example, we have already noted that the sample sizes nj are the only available information to distinguish the different experiments. It does not seem likely that nj would be a useful variable for modeling tumor rates.

  So in the rat tumor example, yj were exchangeable as no additional knowledge was available on experimental conditions. If we knew that specific batches of experiments were made in different laboratories we could assume partial exchangeability and use two level hierarchical model to model variation within each laboratory and between laboratories.

Bayesian analysis of conjugate hierarchical model

- Application to the model for rat tumors:
- Model: $y_j \sim \text{Bin}(n_j, \theta_j)$ (i.i.d.)
  $\theta_j \sim \text{Beta}(\alpha, \beta)$

  The parameters θ_j are assumed to be independent samples from a single beta distribution.

  At last we can assign a noninformative hyperprior distribution $p(\alpha, \beta)$ to reflect our ignorance about the unknown hyperparameters.

- Joint posterior distribution for parameter and hyperparameter:

$$p(\theta,\alpha,\beta|y) \propto p(\alpha,\beta)p(\theta|\alpha,\beta)p(y|\theta,\alpha,\beta)$$
$$\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \quad (1)$$

- By integrating θ out of the joint posterior we can obtain a posterior distribution for hyperparameter(α, β):

$$p(\alpha,\beta|y) \propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}. \quad (2)$$

- Dividing equation (1) by equation (2) we can obtain the conditional posterior distribution for θ given (α, β):

$$p(\theta|\alpha,\beta,y) = \prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}.$$

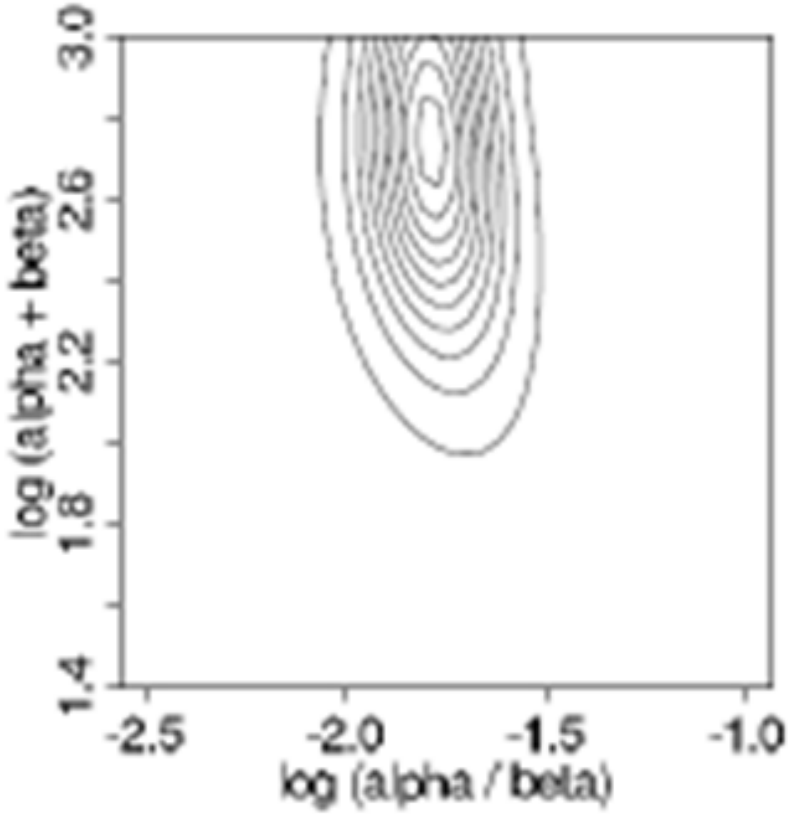
The estimation of hyperparameter

- Now hyperpassign a noninformative rior density:

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

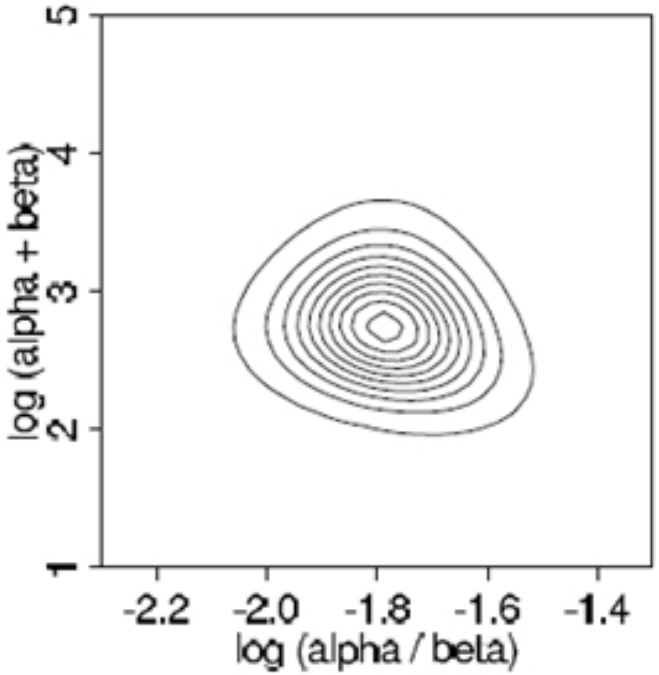
- on the natural transformed scale:

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



- First try at a contour plot of the marginal posterior density of $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right)$ for the rat tumor example. Contour lines are at 0.05, 0.15, . . . , 0.95 times the density at the mode.

- Computing the marginal posterior density of the hyperparameters. Now that we have established a full probability model for data and parameters, we compute the marginal posterior distribution of the hyperparameters. Figure shows a contour plot of the unnormalized marginal posterior density on a grid of values of $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right)$.To create the plot, we first compute the logarithm of the density function with prior density , multiplying by the Jacobian to obtain the density $p(log(\frac{\alpha}{\beta}), log(\alpha+\beta)|y)$. We set a grid in the range $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right) \in [-2.5,-1] \times [1.5,3]$, which is centered near our earlier point estimate $(-1.8,2.3)$ (that is, $(\alpha,\beta)= (1.4,8.6)$)



- Figure:Contour plot of the marginal posterior density of $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right)$for the rat tumor example. Contour lines are at 0.05, 0.15, . . . , 0.95 times the density at the mode.

- We recompute $p\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)|y\right)$, this time in the range $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right) \in [-2.3, -1.3] \times [1,5]$. The resulting grid, shown in Figure a, displays essentially all of the marginal posterior distribution. The graphs show that the marginal posterior distribution of the hyperparameters, under this transformation, is approximately symmetric about the mode, roughly (−1.75, 2.8). This corresponds to approximate values of $(\alpha, \beta) = (2.4, 14.0)$ which differs somewhat from the crude estimate obtained earlier.

- We can then compute posterior moments based on the grid of $\left(log(\frac{\alpha}{\beta}), log(\alpha+\beta)\right)$;for example,

$$\mathrm{E}(\alpha|y) \text{ is estimated by } \sum_{\log(\frac{\alpha}{\beta}),\log(\alpha+\beta)} \alpha \cdot p(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)|y).$$

- As a result ,E(α|y) = 2.4 and E(β|y) = 14.3.

Bayesian analysis of a exchangeable normal hierarchical model

- We assume (y1,.....,yj| θ 1,...., θ j)(population variance known) is independent,that is:

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

- Conditional prior distribution given hyperparameter follows a normal distribution with mean of μ and standard variance τ :

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2)$$

- We also assign a noninformative uniform hyperprior distribution to μ , given τ (we assume a uniform conditional density) :

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

The data structure

Consider J independent experiments, with experiment j estimating the parameter θ_j from n_j independent normally distributed data points, y_{ij}, each with known error variance σ^2; that is,

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, \sigma^2), \text{ for } i = 1, \ldots, n_j; \quad j = 1, \ldots, J. \tag{5.11}$$

Using standard notation from the analysis of variance, we label the sample mean of each group j as

$$\overline{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

with sampling variance

$$\sigma_j^2 = \sigma^2 / n_j.$$

- that's sample mean for group j given prior follows a normal distribution :

$$\overline{y}_{.j}|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2),$$

- Joint posterior distribution:

$$\begin{aligned} p(\theta, \mu, \tau|y) &\propto p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta) \\ &\propto p(\mu, \tau)\prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2)\prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\theta_j, \sigma_j^2) \end{aligned} \quad (1)$$

- By intergrating θj out of the joint posterior density,we can get the marginal posterior density of (μ, τ),which is p(μ, τ |y)(2).Then we divide joint posterior by p(μ, τ |y),we can obtain the marginal posterior distribution of θ conditional on (μ, τ), following a normal distribution:

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

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

- The mean value equation for parameter shows that the mean value is a precision-weighted average of the prior population mean and the sample mean of the jth group.

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

- The posterior variance equation shows that the variance is a combination of the prior precision and data precision.

- By integrating θ out of posterior,we get the marginal posterior of hyperparameter:

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\mu, \sigma_j^2 + \tau^2). \qquad (1)$$

- By integrating (μ , θ) out of the joint posterior,we can get the marginal posterior of τ ,which is p(τ |y).Then we divide marginal posterior of hyperparameter by it we can get the marginal posterior density of μ given τ which follows a normal distributi- - -

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

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} \overline{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}.$$

marginal posterior of τ

- Two ways to get marginal posterior of τ:
- Using the following equation (both of the numerator and denominator have been derived)

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\ &\propto \frac{p(\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\mu,\sigma_j^2+\tau^2)}{\mathrm{N}(\mu|\hat{\mu},V_\mu)}. \end{aligned}$$

- Alternatively we can integrate μ out of marginal posterior of hyperparameter density and we will get the marginal posterior distribution of τ:

$$\begin{aligned} p(\tau|y) &\propto \frac{p(\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\hat{\mu},\sigma_j^2+\tau^2)}{\mathrm{N}(\hat{\mu}|\hat{\mu},V_\mu)} \\ &\propto p(\tau)V_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\overline{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right), \end{aligned}$$