Hierarchical models I

May 8, 2023

Overview

1. Hierarchical models

2. Constructing a parameterized prior distribution

3. Exchangeability and hierarchical models

4. Bayesian analysis of conjugate hierarchical models

Hierarchical models

- Statistical problems might related to some structures. For example, the effectiveness of cardiac treatments for different patients in different hospital.
- Hierarchical models can have enough parameters to fit the data, while using a population distribution to structure some dependence into the parameters, thereby avoiding problems of overfitting.
- hierarchical thinking helps in understanding multiparameter problems and also plays an important role in developing computational strategies.

Constructing a parameterized prior distribution

Example: Estimating the risk of tumor in a group of rats

- Suppose the immediate aim is to estimate θ, the probability of tumor in a population of female laboratory rats of type 'F344' that receive a zero dose of the drug (a control group).
- The data show that 4 out of 14 rats developed a tumor.
- It is natural to assume a binomial model for the number of tumors, given θ. For convenience, we select a prior distribution for θ from the conjugate family, $\theta \sim \textit{Beta}(\alpha, \beta)$.

Analysis with a fixed prior distribution

Example: Estimating the risk of tumor in a group of rats

- From historical data, we can estimate the mean and standard deviation of θ.
- The parameters of the beta distribution α and β could be estimated through method of moments:

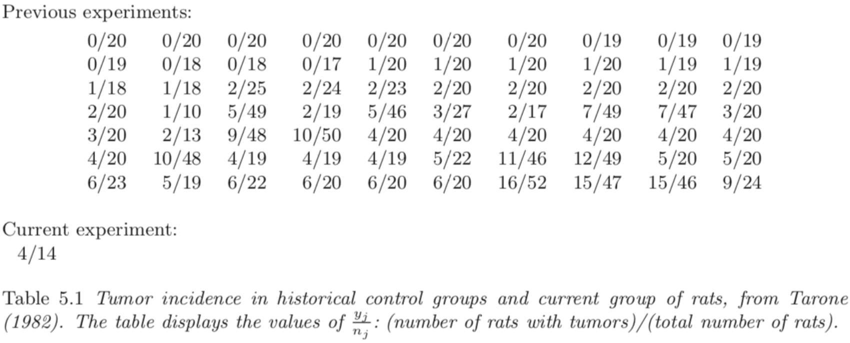
$$\alpha + \beta = \frac{\mathrm{E}(\theta)(1 - \mathrm{E}(\theta))}{\mathrm{var}(\theta)} - 1$$

$$\alpha = (\alpha + \beta)\mathrm{E}(\theta), \qquad \beta = (\alpha + \beta)(1 - \mathrm{E}(\theta)).$$

- Then, assuming a *Beta*(α, β) prior distribution for θ yields a *Beta*$(\alpha + 4, \beta + 10)$ posterior distribution for θ.

Approximate estimate of the population distribution using the historical data

The process is similar to meta-analysis as information combination.



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

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* $\frac{y_j}{n_j}$*: (number of rats with tumors)/(total number of rats).*

Approximate estimate of the population distribution using the historical data

- In the jth historical experiment, let the number of rats with tumor be y_j and the total number of rats be n_j.
- We model the y_j's as independent binomial data, given sample sizes n_j and probability θ_j.
- Assuming the beta prior distribution with parameters (α, β) is a good description of the population distribution of the θ_js in the historical experiments, we can display the hierarchical model schematically.

Approximate estimate of the population distribution using the historical data

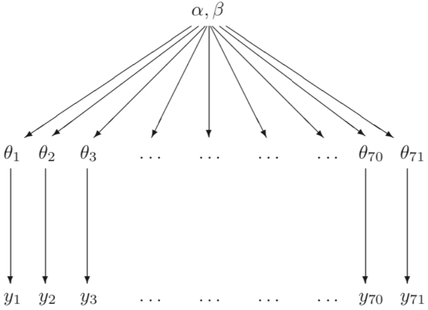

α, β
θ_1 θ_2 θ_3 ... ... θ_{70} θ_{71}
y_1 y_2 y_3 ... ... y_{70} y_{71}

Figure 5.1: *Structure of the hierarchical model for the rat tumor example.*

The observed sample mean and standard deviation of the 70 values $\frac{y_j}{n_j}$ are 0.136 and 0.103. Please calculate the estimation of α and β through method of moments.

Approximate estimate of the population distribution using the historical data

- The estimation of (α, β) is $(1.4, 8.6)$. (It is not a Bayesian estimation, we shall mention it later).
- Now please try to find the posterior distribution for θ_{71} and the mean and standard deviation. Comparing the posterior mean with frequentist mean.

Approximate estimate of the population distribution using the historical data

- The estimation would be

$$E(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$Var(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

- with updated $\alpha = 4 + 1.4$ and $\beta = 10 + 8.6$
- the frequentist estimation would be $\theta = 4/14 = 0.286$
- There might be potential difference between historical data and current experiment, such as different laboratory or time trend. The method is to systematically inflate the historical variance.
- For the beta model, inflating historical variance means decreasing $(\alpha + \beta)$ while holding $\frac{\alpha}{\beta}$ constant.

Exchangeability

- Consider a set of experiments $j = 1, \ldots, J$, 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.
- If no information other than the data y to distinguish any of the θ_js from any others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution.
- The symmetry is represented probabilistically by exchangeability.

Exchangeability

- The parameters $(\theta_1, \dots, \theta_J)$ are exchangeable in their joint distribution $p(\theta_1, \dots, \theta_J)$ in invariant to permutations of the indexes $(1, \dots, J)$
- The simplest form of an exchangeable distribution has each of the parameters θ_j as an independent sample from a prior distribution governed by some unknown vector ϕ

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

Exchangeability

- In general, ϕ is unknown, so our distribution for θ must average over our uncertainty in ϕ

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j|\phi) \right) p(\phi) d\phi,$$

- The mixture of independent identical distributions is usually all that we need to capture exchangeability in practice.
- de Finetti's theorem: in the limit as $J \to \infty$, any suitably well-behaved exchangeable distribution on $(\theta_1, \ldots, \theta_J)$ can be expressed as a mixture of independent and identical distributions.
- Parameters θ are drawn from a common 'superpopulation' that is determined by the unknown hyperparameters ϕ.

Exchangeability

- As a simple counterexample to the above mixture model, consider the probabilities of a given die landing on each of its six faces. The probabilities $\theta_1, \ldots, \theta_6$ are exchangeable, but the six parameters θ_j are constrained to sum to 1 and so cannot be modeled with a mixture of independent identical distributions.

Example: Exchangeability and sampling

- The authors have selected 8 states in the US and recorded the divorce rate per 1000 population in each state in 1981. $y_1, \ldots, y_8$
- The prior distribution you might use is the beta distribution for the y_js with restricted range $[0, 1]$
- Now 7 of the 8 states values are told as $5.8, 6.6, 7.8, 5.6, 7.0, 7.1, 5.4$
- Now guess the y_8. You may guess the value would probably be centered around 6.5 and have a range $[5, 8]$
- When we change the index, there will be no difference of the estimation. But y_js are not independent because we assume that the divorce rate in the states is probably similar.

Example: Exchangeability and sampling

- Now if we know the 8 states are Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming, but we cannot match them to the index. Will the estimation be changed?
- Probably yes if we know the culture background of the 8 states, especially, Utah is full of Mormon population, which may lead to a small divorce rate, while Nevada has liberal divorce law, which may lead to a large rate.
- since we have already know the previous 7 states' divorce rates are similar, it is reasonable to guess the last divorce rate might be way larger or smaller.

Example: Exchangeability and sampling

- Finally, if we know the last state is Nevada, the probability $p(y_8 > max(y_1, \ldots, y_7)|y_1, \ldots, y_7)$ should be large.
- In fact, the divorce rate of Nevada in 1981 was 13.9 per 1000 population.

Exchangeability

Often observations are not fully exchangeable, but are partially or conditionally exchangeable

- If observations can be grouped, we may make hierarchical model. 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 nor 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

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

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

  with $x = (x_1, \ldots, x_J)$.
- Any information available to distinguish different units should be encoded in the x and y variables.
- In the rat tumor example, we have already noted that the sample sizes n_j are the only available information to distinguish the different experiments.
- Plot $\frac{y_j}{n_j}$ and n_j to check the relationship.

The full Bayesian treatment of the hierarchical model

- The key hierarchical part of models is that ϕ is unknown and thus has its own prior distribution, $p(\phi)$.
- the joint prior distribution is

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

- the joint posterior distribution is

$$\begin{aligned} p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\phi, \theta) \\ &= p(\phi, \theta)p(y|\theta), \end{aligned}$$

- Simplification: $p(y|\theta, \phi) = p(y|\theta)$, which means the hyperparamter ϕ affect y only through θ.

Analytic derivation of conditional and marginal distributions

- We try to combine analytical and numerical methods to obtain simulations from the joint posterior distribution, $p(\theta, \phi|y)$, for the beta-binomial model for the rat-tumor example.
- We first perform the following three steps analytically.
  1. write the joint posterior density, $p(\theta, \phi|y)$, in unnormalized form as a product of the hyperprior distribution $p(\phi)$, the population distribution $p(\theta|\phi)$, and the likelihood $p(y|\theta)$
  2. Determine analytically the conditional posterior density of θ given the hyperparameters ϕ; for fixed observed y, this is a function of ϕ, $p(\theta|\phi, y)$
  3. Estimate ϕ using Bayesian paradigm; that is, obtain its marginal posterior distribution, $p(\phi|y)$

Analytic derivation of conditional and marginal distributions

- The third step can be performed by integrating the joint posterior distribution over θ:

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta.$$

- For many standard models, including the normal distribution, the marginal posterior distribution of ϕ can be computer algebraically using the conditional probability formula,

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

  This expression is useful because the numerator is the joint posterior distribution, and the denominator is the posterior distribution for θ if ϕ were known.

Drawing simulations form the posterior distribution

The following strategy is useful for simulating a draw from the joint posterior distribution, $p(\phi, \theta|y)$, for simple hierarchical models.

1. Draw the vector of hyperparameters, ϕ, from its marginal posterior distribution, $p(\phi|y)$.
2. Draw parameter vector θ from its conditional posterior distribution, $p(\theta|\phi, y)$,given the drawn value of ϕ.
3. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn θ.

Application to the model for rat tumors

- The data from experiments $j = 1, \ldots, J$, $J = 71$ are assumed to follow independent binomial distributions:

  $$y \sim Bin(n_j, \theta_j)$$

  with the number of rats, n_j, known.
- The parameters θ_j are assumed to be independent samples from a beta distribution:

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

  and the noninformative prior would be used.

Application to the model for rat tumors

Joint, conditional, and marginal posterior distributions.

- The joint posterior distribution of all parameters is

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

- given α and β, the components of θ have independent posterior densities that are of the form $\theta_j^A(1-\theta_j)^B$ that is beta densities. And the joint density is

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

Application to the model for rat tumors

- We can determine the marginal posterior distribution of (α, β) by substituting $p(\theta, \alpha, \beta|y)$ and $p(\theta|\alpha, \beta, y)$ (joint/conditional = marginal)

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

- The product in the equation cannot be simplified analytically but is easy to compute for any specific values of (α, β) using a standard routine to compute the gamma function.

Application to the model for rat tumors

- One reasonable choice of hyperprior density is uniform on $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$, which when multiplied by the appropriate Jacobian yields the following destities on the original scale,

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

  and on the natural transformed scale:

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

Computing the marginal posterior density of the hyperparameters

- We user the prior $p(\alpha, \beta)$ with the logarithm of the density function $p(\alpha, \beta|y)$, multiplying by the Jacobian to obtain the density $p(log(\frac{\alpha}{\beta}), log(\alpha + \beta)|y)$.
- We set a grid in the range $(log(\frac{\alpha}{\beta}), log(\alpha + \beta)) \in [-2.5, -1] \times [1.5, 3]$, which is centered near our earlier point estimate $(-1.8, 2.3)$ (that is $(\alpha, \beta) = (1.4, 8.6)$), yielding values of the unnormalized marginal posterior density.

Computing the marginal posterior density of the hyperparameters

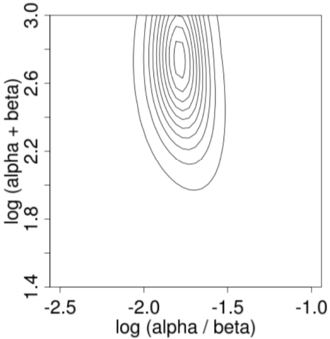


Figure 5.2 *First try at a contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode.*

Computing the marginal posterior density of the hyperparameters

- Because the important parts in previous graph lie outside, by rescaling the grid as $(log(\frac{\alpha}{\beta}), log(\alpha + \beta)) \in [-2.3, -1.3] \times [1, 5]$ the new graph could be drawn.
- Also the 1000 random draws from the numerically computed posterior distribution could be shown in a graph, with approximate values of $(\alpha, \beta) = (2.4, 14.0)$, which differs somewhat from the crude estimate obtained earlier.

Computing the marginal posterior density of the hyperparameters

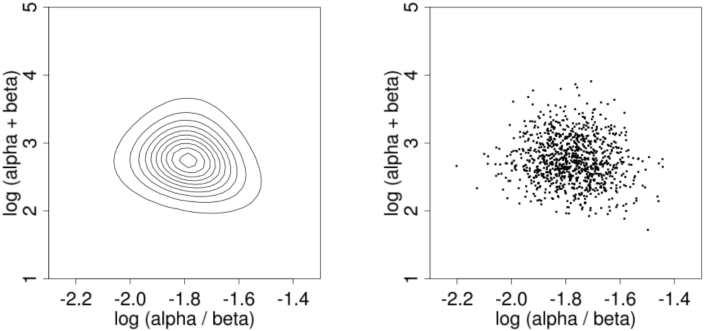


Figure 5.3 *(a) Contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode. (b) Scatterplot of 1000 draws* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *from the numerically computed marginal posterior density.*

Computing the marginal posterior density of the hyperparameters

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

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

- From appropriate grid, $E(\alpha|y) = 2.4$ and $E(\beta|y) = 14.3$

Sampling from the joint posterior distribution of parameters and hyperparameters

A more important consequence of averaging over the grid is to account for the posterior uncertainty in (α, β).

- We draw 1000 random samples from the joint posterior distribution of $(\alpha, \beta, \theta_1, \ldots, \theta_J)$, as follows
  1. Simulate 1000 draws of $(log(\frac{\alpha}{\beta}), log(\alpha + \beta))$ from the posterior distribution $p(log(\frac{\alpha}{\beta}), log(\alpha + \beta)|y)$ with discrete-grid sampling procedure.
  2. for each $j = 1, \ldots, J$ sample θ_j from its conditional posterior distribution, $\theta_j|\alpha, \beta, y \sim Beta(\alpha + y_j, \beta + n_j - y_j)$

Sampling from the joint posterior distribution of parameters and hyperparameters

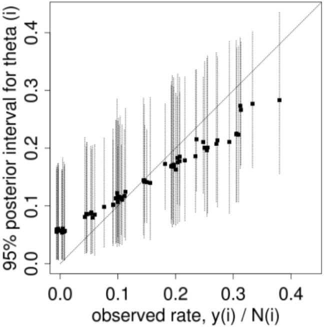


Figure 5.4 *Posterior medians and 95% intervals of rat tumor rates, θ_j (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45° line corresponds to the unpooled estimates, $\hat{\theta}_i = y_i/n_i$. The horizontal positions of the line have been jittered to reduce overlap.*