

# Bayes Homework3

Chen Xupeng

Textbook Chapter 5:

## Exchangeability:

### 5.5

Mixtures of independent distributions: suppose the distribution of  $\theta = (\theta_1, \dots, \theta_J)$  can be written as a mixture of independent and identically distributed components:

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

Prove that the covariances  $\text{cov}(\theta_i, \theta_j)$  are all nonnegative.

$$\begin{aligned} \text{cov}(\theta_i, \theta_j) &= E(\text{cov}(\theta_i, \theta_j | \phi)) + \text{cov}(E(\theta_i | \phi), E(\theta_j | \phi)) \\ &= 0 + \text{cov}(E(\theta_i | \phi), E(\theta_i | \phi)) \\ &= \text{var}(E(\theta_i | \phi)) \\ &\geq 0 \end{aligned}$$

### 5.4 (a)(b)

Exchangeable prior distributions: suppose it is known a priori that the  $2J$  parameters  $\theta_1, \dots, \theta_{2J}$  are clustered into two groups, with exactly half being drawn from a  $N(1, 1)$  distribution, and the other half being drawn from a  $N(-1, 1)$  distribution, but we have not observed which parameters come from which distribution.

(a) Are  $\theta_1, \dots, \theta_{2J}$  exchangeable under this prior distribution?

(b) Show that this distribution cannot be written as a mixture of independent and identically distributed components.

The density can be written as:

$$p(\theta_1, \dots, \theta_{2J}) = \binom{2J}{J}^{-1} \sum_p \left( \prod_{j=1}^J N(\theta_{p(j)} | 1, 1) \prod_{j=J+1}^{2J} N(\theta_{p(j)} | -1, 1) \right)$$

it is exchangeable

## Choosing Prior:

### 5.10 (a)(b)

10. Checking the integrability of the posterior distribution: consider the hierarchical normal model in Section 5.4.

(a) If the hyperprior distribution is  $p(\mu, \tau) \propto \tau^{-1}$  (that is,  $p(\mu, \log \tau) \propto 1$ ), show that the posterior density is improper.

(b) If the hyperprior distribution is  $p(\mu, \tau) \propto 1$ , show that the posterior density is proper if  $J > 2$ .

(c) How would you analyze SAT coaching data if  $J = 2$  (that is, data from only two schools)?

We have  $p(\mu|\tau, y)$  and  $p(\theta|\mu, \tau, y)$  have proper distributions, joint posterior density  $p(\theta, \mu, \tau|y)$  will also be proper if marginal posterior density  $p(\tau|y)$  is proper ( $\tau$  has a finite integral)

$$p(\tau|y) \propto \frac{p(\tau) \prod_{j=1}^J N(\bar{y}_j | \hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu} | \hat{\mu}, V_\mu)}$$

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

Everything multiplying  $p(\tau)$  in the above equation approaches a nonzero constant limit as  $\tau$  tends to zero. So the posterior density near  $\tau = 0$  is determined by the prior density. The function  $p(\tau) \propto 1/\tau$  is not integrable for a small interval including  $\tau = 0$  and so it leads to a nonintegrable posterior density.

If  $p(\tau) \propto 1$ , the posterior density is integrable near zero. For  $\tau \rightarrow \infty$ . The exponential term is less than or equal to 1. We can rewrite the remaining terms as  $\left(\sum_{j=1}^J \left[\prod_{k \neq j} (\sigma_k^2 + \tau^2)\right]\right)^{-1/2}$ . For  $\tau > 1$  making it bigger by dropping the denominator  $\sigma^2$  to yield  $(J\tau^{2(J-1)})^{-1/2}$ . An upper bound on  $p(\tau|y)$  for  $\tau$  large is  $p(\tau)J^{-1/2}/\tau^{J-1}$ . When  $p(\tau) \propto 1$  this upper bound is integrable if  $J > 2$ , and so  $p(\tau|y)$  is integrable if  $J > 2$

## Inference:

### 5.12

Conditional posterior means and variances: derive analytic expressions for  $E(\theta_j|\tau, y)$  and  $\text{var}(\theta_j|\tau, y)$  in the hierarchical normal model (and used in Figures 5.6 and 5.7). (Hint: use (2.7) and (2.8), averaging over  $\mu$ .)

$$E(\theta_j|\tau, y) = E[E(\theta_j|\mu, \tau, y) | \tau, y] = E\left[\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} | \tau, y\right] = \frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\hat{\mu}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$\text{var}(\theta_j|\tau, y) = E[\text{var}(\theta_j|\mu, \tau, y) | \tau, y] + \text{var}[E(\theta_j|\mu, \tau, y) | \tau, y] = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\right)^2 V_\mu$$

### 5.13

Hierarchical binomial model: Exercise 3.8 described a survey of bicycle traffic in Berkeley, California, with data displayed in Table 3.3. For this problem, restrict your attention to the first two rows of the table: residential streets labeled as ‘bike routes,’ which we will use to illustrate this computational exercise.

- a. Set up a model for the data in Table 3.3 so that, for  $j = 1, \dots, 10$ , the observed number of bicycles at location  $j$  is binomial with unknown probability  $\theta_j$  and sample size equal to the total number of vehicles (bicycles included) in that block. The parameter  $\theta_j$  can be interpreted as the underlying or 'true' proportion of traffic at location  $j$  that is bicycles. (See Exercise 3.8.) Assign a beta population distribution for the parameters  $\theta_j$  and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3. Write down the joint posterior distribution.

$$p(\theta, \alpha, \beta | y) \propto [p(y | \theta, \alpha, \beta)] [p(\theta | \alpha, \beta) p(\alpha, \beta)]$$

Let us consider  $y_1, y_2, \dots, y_{10}$  are the number of bicycles at ten different locations, and  $n_1, n_2, \dots, n_{10}$  are the total vehicles including bicycles at location 1 – 10, According to the statement,  $y_j$  can be considered to be generated from a binomial process with known  $\theta_j$

$$y_j | \theta_j \sim \text{Bin}(n_j, \theta_j), j = 1, 2, \dots, 10 \quad p(y | \theta, \alpha, \beta) = \prod_{j=1}^{10} \binom{n_j}{y_j} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j}$$

Similarly,  $\theta_j$  has Beta distribution with hyper-parameter  $\alpha$  and  $\beta$   $\theta_j \sim \text{Beta}(\alpha, \beta)$

$$p(\theta | \alpha, \beta) = \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1 - \theta_j)^{\beta-1}$$

One of the possible non-informative prior distribution for the hyper parameters is;

By putting things together,

$$\begin{aligned} p(\theta, \alpha, \beta | y) &\propto p(\alpha, \beta) \prod_{j=1}^{10} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j} \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1 - \theta_j)^{\beta-1} \\ p(\theta, \alpha, \beta | y) &\propto (\alpha + \beta)^{-5/2} \left( \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{10} \prod_{j=1}^{10} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j} \theta_j^{\alpha-1} (1 - \theta_j)^{\beta-1} \\ p(\theta, \alpha, \beta | y) &\propto (\alpha + \beta)^{-5/2} \left( \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{10} \prod_{j=1}^{10} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n_j - y_j - 1} \end{aligned}$$

- b. Compute the marginal posterior density of the hyperparameters and draw simulations from the joint posterior distribution of the parameters and hyperparameters, as in Section 5.3.

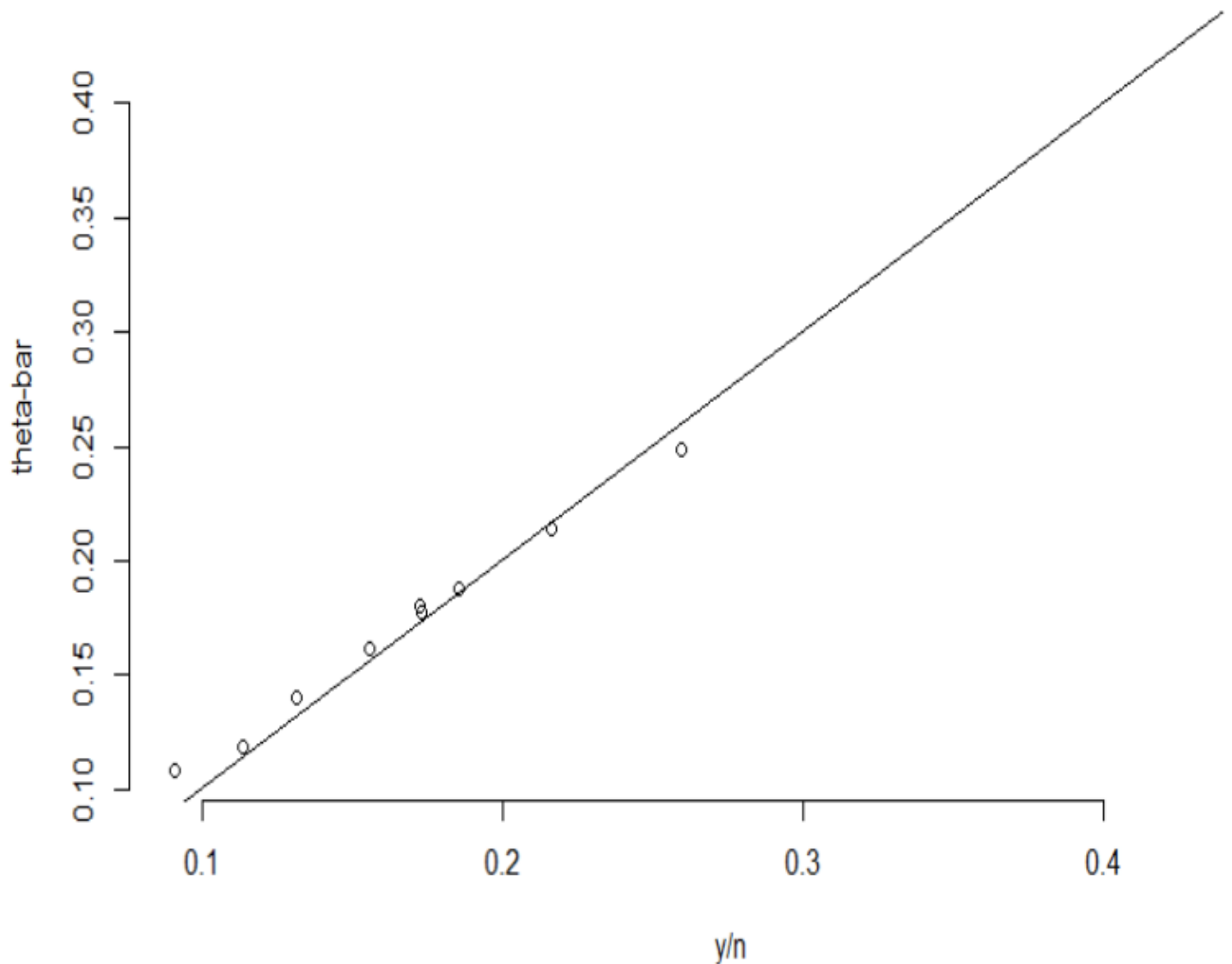
The marginal distribution of the hyperparameters can be computed through the following algebraic expression,

$$p(\alpha, \beta | y) = \frac{p(\theta, \alpha, \beta | y)}{p(\theta | \alpha, \beta, y)}$$

$$\begin{aligned} p(\theta, \alpha, \beta | y) &\propto (\alpha + \beta)^{-5/2} \left( \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{10} \prod_{j=1}^{10} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n_j - y_j - 1} \\ p(\theta | \alpha, \beta, y) &= \prod_{j=1}^{10} \frac{\Gamma(\alpha + \beta + n_j)}{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n_j - y_j - 1} \\ p(\alpha, \beta | y) &\propto (\alpha + \beta)^{-5/2} \left( \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{10} \prod_{j=1}^{10} \frac{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha + \beta + n_j)} \end{aligned}$$

For the given values of  $\alpha$  and  $\beta$ , we can simulate  $p(\theta_j | \alpha, \beta, y)$  because its normalized form is available.

- c. Compare the posterior distributions of the parameters  $\theta_j$  to the raw proportions, (number of bicycles / total number of vehicles) in location  $j$ . How do the inferences from the posterior distribution differ from the raw proportions?



image

d. Give a 95% posterior interval for the average underlying proportion of traffic that is bicycles.

quantile(p, c(0.025, 0.975)) = 0.1465 0.2938. 14.6% to 29.4% of the vehicles are bicycles.

e. A new city block is sampled at random and is a residential street with a bike route. In an hour of observation, 100 vehicles of all kinds go by. Give a 95% posterior interval for the number of those vehicles that are bicycles. Discuss how much you trust this interval in application.

Essentially, we are looking for posterior predictive distribution  $p(\tilde{y}|y)$  In order to get a sample from it, we need to generate new values of  $\tilde{\theta}$  from  $\alpha, \beta$  samples. Now using these values, we can sample from binomial distribution with  $n = 100$  i.e.  $\tilde{y}|\tilde{\theta} \sim \text{Bin}(100, \tilde{\theta})$

quantile(y.new, c(0.025, 0.975)) = 3 49. which means 3 to 49 bicycles will pass from the residential street with bike route (newly sampled) with the probability of 95%.

f. Was the beta distribution for the  $\theta_j$  's reasonable? The plot of  $\theta$  vs  $\frac{y}{n}$  fits well which indicates the values of  $\theta$  are reasonable.