Pei Lin Cui
Bayesian Final

Problem 1. Part 1)
Consider m observations $(y_1, n_1), \ldots, (y_m, n_m)$, where $y_i \sim Bin(n_i, \theta_i)$ are binomial variable. Assume that $\theta_i \sim w_1 Beta(\alpha_1, \beta_1) + w_2 Beta(\alpha_2, \beta_2)$ are mixture from two Beta distribution $(w_1 + w_2) = 1$.
Derive a Laplace approximation of the likelihood

Binomial distribution function: $\begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y}$

$$\log f(\theta) = \log \begin{pmatrix} n \\ y \end{pmatrix} + y \log \theta + (n-y)\log(1-\theta)$$

log-likelihood,

$$\propto \log \prod_{i=1}^{n} \theta_i^{y_i} + \log \prod_{i=1}^{n} (1-\theta_i)^{n_i - y_i}$$

$$\propto \sum y_i \log \theta_i + \sum (n_i - y_i) \log(1-\theta_i)$$

First derivative,

$$\frac{\partial}{\partial \theta_i} = \frac{y_i}{\theta_i} - \frac{n_i - y_i}{1-\theta_i} = 0,$$

$$\frac{y_i}{\theta_i} = \frac{n_i - y_i}{1-\theta_i}$$

$$y_i(1-\theta_i) = \theta_i(n_i - y_i)$$

$$y_i = \theta_i n_i$$

so, the mode is $\hat{\theta}_i = \dfrac{y_i}{n_i}$.

Second partial derivative, $\dfrac{\partial^2}{\partial \theta^2} = \left[-\dfrac{y}{\theta^2} - \dfrac{n-y}{(1-\theta)^2} \right]$.

To find the variance, we want to write Taylor expansion at mode,

$$\sigma^{-2} = -\frac{\partial^2}{\partial \theta^2} = \left[\frac{y}{\theta^2} + \frac{n-y}{(1-\theta)^2} \right] = \frac{y}{\left(\frac{y}{n}\right)^2} + \frac{n-y}{\left(\frac{n-y}{n}\right)^2} = \frac{n^2}{y} + (n-y) \bullet \frac{n^2}{(n-y)^2}$$

$$= \frac{n^2}{y} + \frac{n^2}{n-y} = \frac{n^2(n-y) + n^2 y}{y(n-y)} = \frac{n^3 - n^2 y + n^2 y}{y(n-y)} = \frac{n^3 - n^2 y + n^2 y}{y(n-y)}$$

$$= \frac{n^3}{y(n-y)}$$

so, $\sigma^2 = \frac{y(n-y)}{n^3}$

write a loglikelihood of normal with parameter theta, and theta^(-2)

The Laplace approximation is, $P(\theta) \approx N\left(\frac{y_i}{n_i}, \frac{y_i(n_i - y_i)}{n_i^3}\right)$

Each mixture component: $Beta(\alpha_1, \beta_1)$, $Beta(\alpha_2, \beta_2)$. I will use the general notation (α, β), results are the same in both cases.

$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\log P(\theta) \propto (\alpha - 1)\log\theta + (\beta - 1)\log(1-\theta)$

mode finding:

$\hat{\theta} = \arg\max[(\alpha-1)\log\theta + (\beta-1)\log(1-\theta)]$

$\frac{\partial L}{\partial \theta} = \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$

$\frac{\alpha-1}{\theta} = \frac{\beta-1}{1-\theta}$

$\hat{\theta} = \frac{\alpha-1}{\alpha+\beta-2}$, and $\hat{\theta}^{-2} = \frac{(\alpha_1+\beta_1-2)^3}{(\alpha-1)(\beta-1)}$.

These results were derived in the lecture, so, I am using them without giving my own derivation.

Laplace approximation for each mixture component is,

$w_1 Beta(\alpha_1, \beta_1) \approx w_1 N\left(\frac{\alpha_1 - 1}{\alpha_1+\beta_1-2}, \frac{(\alpha_1-1)(\beta_1-1)}{(\alpha_1+\beta_1-2)^3}\right)$ and

$w_2 Beta(\alpha_2, \beta_2) \approx w_2 N\left(\frac{\alpha_2 - 1}{\alpha_2+\beta_2-2}, \frac{(\alpha_2-1)(\beta_2-1)}{(\alpha_2+\beta_2-2)^3}\right)$.

Problem 1 Part 2)

Derive the empirical Bayes likelihood of the data by integrating out θ_i using the Laplace approximation, and leave the hyper-parameter (w_i, α_i, β_i) $(j = 1,2)$.

Our parameters-- of interest are $y_i, n_i, \alpha_1, \beta_1, \alpha_2, \beta_2$. So, we have

Using the Laplace approximation from part 1, we will use Gaussian instead of Beta.

$$P(y_i \mid n_i, \alpha_1, \alpha_2, \beta_1, \beta_2)$$

$$\propto \int_0^1 \theta_i^{y_i}(1-\theta)^{n_i-y_i}\left\{w_1\left[N\left(\frac{\alpha_1-1}{\alpha_1+\beta_1-2}, \frac{(\alpha_1-1)(\beta_1-1)}{(\alpha_1+\beta_1-2)^3}\right)\right] + w_2\left[N\left(\frac{\alpha_2-1}{\alpha_2+\beta_2-2}, \frac{(\alpha_2-1)(\beta_2-1)}{(\alpha_2+\beta_2-2)^3}\right)\right]\right\}d\theta_i$$

$$\propto \int_0^1 \theta_i^{y_i}(1-\theta)^{n_i-y_i} w_1\left[N\left(\frac{\alpha_1-1}{\alpha_1+\beta_1-2},\frac{(\alpha_1-1)(\beta_1-1)}{(\alpha_1+\beta_1-2)^3}\right)\right]d\theta_i$$
$$+\int_0^1 \theta_i^{y_i}(1-\theta)^{n_i-y_i} w_2\left[N\left(\frac{\alpha_2-1}{\alpha_2+\beta_2-2},\frac{(\alpha_2-1)(\beta_2-1)}{(\alpha_2+\beta_2-2)^3}\right)\right]d\theta_i$$

Let $\left(\mu_{w_1}=\frac{\alpha_1-1}{\alpha_1+\beta_1-2},\sigma_{w_1}^2=\frac{(\alpha_1-1)(\beta_1-1)}{(\alpha_1+\beta_1-2)^3}\right)$ denote the first mixture Gaussian.

Let $\left(\mu_{w_2}=\frac{\alpha_2-1}{\alpha_2+\beta_2-2},\sigma_{w_2}^2=\frac{(\alpha_2-1)(\beta_2-1)}{(\alpha_2+\beta_2-2)^3}\right)$ denote the second mixture Gaussian.

Many pages of work is skipped here, but, after the integration, we get

$$w_1\cdot p(y_i\mid n_i,\alpha_1,\beta_1)+w_2\cdot p(y_i\mid n_i,\alpha_2,\beta_2)$$
$$\propto w_1\exp\left(-\frac{\left(\left(\frac{y_i}{n_i}\right)-\left(\frac{\alpha_1-1}{\alpha_1+\beta_1-2}\right)\right)^2}{2\left(\frac{y_i(n_i-y_i)}{n_i^3}+\frac{(\alpha_1-1)(\beta_1-1)}{(\alpha_1+\beta_1-2)^3}\right)}\right)+w_2\exp\left(-\frac{\left(\left(\frac{y_i}{n_i}\right)-\left(\frac{\alpha_2-1}{\alpha_2+\beta_2-2}\right)\right)^2}{2\left(\frac{y_i(n_i-y_i)}{n_i^3}+\frac{(\alpha_2-1)(\beta_2-1)}{(\alpha_2+\beta_2-2)^3}\right)}\right)$$

The log likelihood is, $\prod_{i=1}^{m} w_1\cdot p(y_i\mid n_i,\alpha_1,\beta_1)+w_2\cdot p(y_i\mid n_i,\alpha_2,\beta_2)$.

Problem 1 Part 3)

Derive the EM algorithm to estimate the hyperparameters.
We are essentially doing EM for Gaussian mixtures.
EM algorithm:
Initialize the means, covariances and the weights, and evaluate the initial value of the log likelihood

1. E step: We want to evaluate responsibilities using the current parameter values
$$\gamma_{ik}=\frac{w_k N(\theta_i\mid\mu_k,\sigma_k)}{\sum_{j=1}^{2} w_j N(\theta_i\mid\mu_j,\sigma_j)}$$
2. M step: re-estimate the parameters using the current repsonsibilities
$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\theta_n$$
$$\Sigma_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}(\theta_n-\mu_k^{new})(\theta_n-\mu_k^{new})^T$$

$$w_k = \frac{N_k}{N}, \text{ where } N_k = \sum_{n=1}^{N} \gamma_{nk}.$$

3. Evaluate the log likelihood

$$\log p(\theta \mid \alpha, \beta, w, n) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} w_k N(\theta_k \mid \mu_k, \sigma_k) \right)$$

Problem 2 Part 1)

We are given data $(x, y) : i = 1, \ldots, n$ and the distribution $\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} u_i \\ v_i \end{pmatrix}, \Sigma \right)$, and

$v_i = a + bu_i$ with some parameter (a, b) and Σ is diagonal.

We can calculate the mean and variance of the likelihood by finding the conditional expectation for the mean and variance.

$$E[x_i] = E[E(x_i \mid \mu)] = E[\mu_i] = \mu$$

$$E[y_i] = E\left[E(y_i \mid v_i)\right] = E[a + b\mu_i] = a + b\mu$$

$$\text{var}(x_i) = \text{var}(E(x_i \mid u_i)) + E(\text{var}(x_i \mid u_i)) = \text{var}(u_i) + \sigma_{x_i}^2 = \tau^2 + \sigma_{x_i}^2$$

$$\text{var}(y_i) = \text{var}(E(y_i \mid u_i)) + E(\text{var}(x_i \mid u_i)) = \text{var}(a + bu_i) + E(\sigma_{y_i}^2) = b^2\tau^2 + \sigma_{y_i}^2$$

So, we have

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ a + b\mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma_{x_i}^2 & 0 \\ 0 & b^2\tau^2 + \sigma_{y_i}^2 \end{pmatrix} \right)$$

The likelihood model for bivariate normal

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left[\frac{(x_i - \mu)^2}{\sigma_{x_i}^2} \right] + \left[\frac{(y_i - (a + b\mu))^2}{\sigma_{y_i}^2} \right] \right)$$

Part 2)

The choices for non-informative prior is either a flat prior or Jeffery's prior. I will choose a flat prior, which is

$$P(a) \propto 1/a$$
$$P(b) \propto 1/b$$

Part 3) I would like to do a Gibbs sampling, but I was not able to write the full conditionals, thus, I did not do the simulation.

Problem 3 Part 1)

Describe your model, using logit link functions and flat prior, with intercept but without considering the interaction effects.

We have a flat prior on $\beta, \pi(\beta) \propto 1$

Posterior distribution

$$\ell(\beta \mid y, X) = \exp\left\{ \sum_{i=1}^{n} y_i X^{iT} \beta \right\} / \prod_{i=1}^{n} \left[1 + \exp\left(X^{iT} \beta \right) \right]$$

Metropolis-Hastings algorithm

In words, we use MLE $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ (corresponding to $\hat{\beta}$) as initial value for the proposal density, $\tilde{\beta} \sim N_k\left(\beta^{(t-1)}, \tau^2 \hat{\Sigma}\right)$

Step 1. Compute and set $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ (corresponding to $\hat{\beta}$) as initial value.

Step 2. Generate $\tilde{\beta} \sim N_k\left(\beta^{(t-1)}, \tau^2 \hat{\Sigma}\right)$.

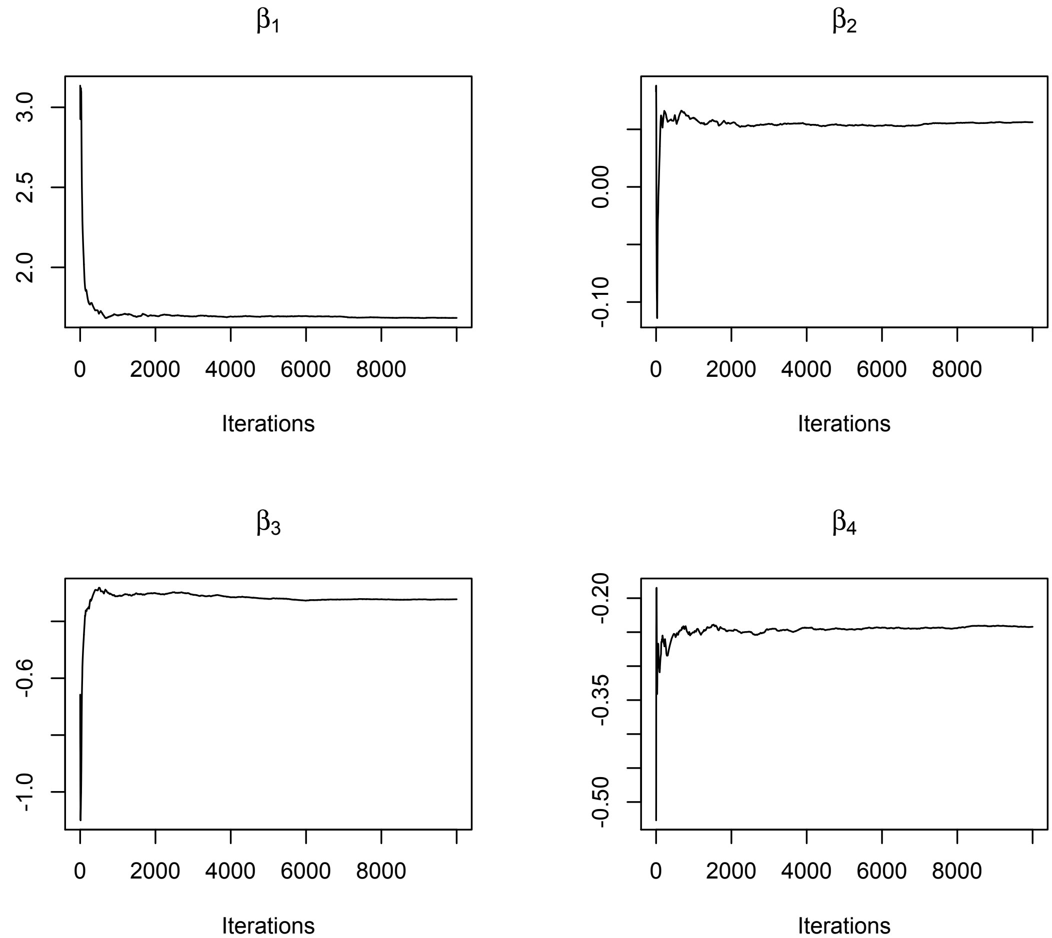
Step 3. Compute

$$\rho\left(\beta^{(t-1)}, \tilde{\beta}\right) = \min\left(1, \frac{\pi\left(\tilde{\beta} \mid y\right)}{\pi\left(\beta^{(t-1)} \mid y\right)} \right).$$

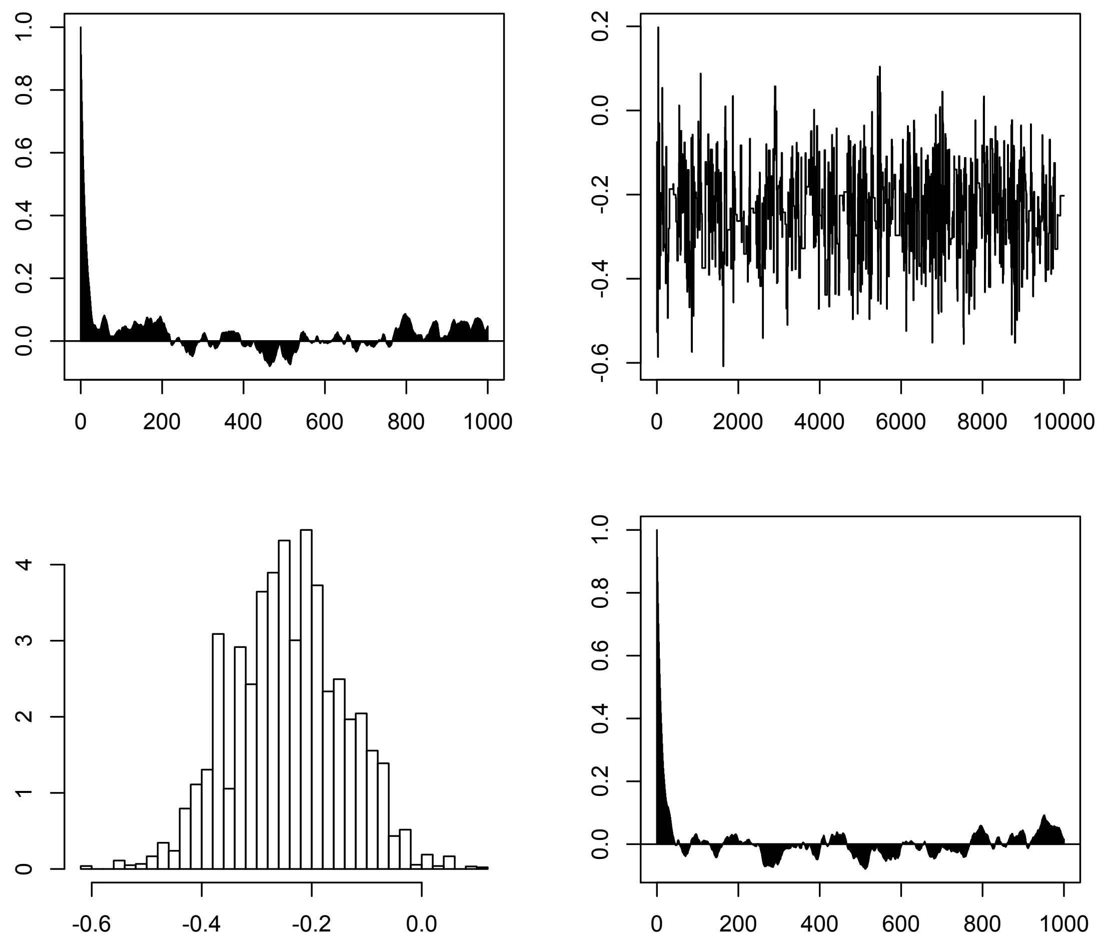
Step 4. Accept candidate point $\tilde{\beta}$ with probability $\rho\left(\beta^{(t-1)}, \tilde{\beta}\right)$, set $\beta^{(t)} = \tilde{\beta}$.

else, set $\beta^{(t)} = \beta^{(t-1)}$.

Part 3) Plotting



From the plot we can see that the burn-in period is about 1500 iteration, after that, the plots begin to converge for all β_i's.



I think this is about 100 iteration for the acf plot? Not sure.

Part 4) posterior mean and variance

The β_i's are the coefficients for the intercept term, Class, Sex, and Age, for $i = 0,1,2,3$, respectively. The posterior mean: $\beta_0 = 1.6823$, $\beta_1 = 0.0557$, $\beta_2 = -0.3235$, $\beta_3 = -0.2411$. The posterior variance: $\text{var}(\beta_0) = 0.0273$. $\text{var}(\beta_1) = 0.0023$, $\text{var}(\beta_2) = 0.0105$, $\text{var}(\beta_3) = 0.0106$.