AMS 207 HW Assignment 1

Mary Silva

April 14 2018

1. Beta Binomial

Obtain the mean and the variance for the beta-binomial distribution. Show that it tackles the overdispersion problem. Hint: use the formulas for conditional expectations and variances.

$$X|p \sim Binomial(n, \theta)$$

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

$$X \sim beta - binomial(n, \alpha, \beta)$$

$$E(X) = E_\theta(E_{X|\theta}(X|\theta)) = E_\theta(n\theta) = \frac{n\alpha}{\alpha + \beta}$$

$$\begin{aligned}
Var(X) &= E(Var(X|\theta)) + Var(E(X|\theta)) \\
&= E(n\theta(1-\theta)) + Var(n\theta) \\
&= n[E(\theta) - E(\theta^2)] + n^2 Var(\theta) \\
&= n[\frac{\alpha}{\alpha+\beta} - \frac{\alpha\beta + \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}] + n^2 \frac{\alpha\beta}{(\alpha+\beta^2)(\alpha+\beta+1)} \\
&= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}$$

To compare to binomial distribution, change the parameterization of beta distribution to $\alpha = \mu\tau, \beta = \tau(1-\mu)$. Then we have $E(\theta) = n\mu$ and $Var(\theta) = n\mu(1-\mu)\frac{\tau+n}{\tau+1}$.

Although the variance still depends on the expectation (there is a $n\mu$ term in the variance), there is a chance for the variance to be greater than the expectation depending on the scale parameter τ. The terms with μ in the beta-binomial distribution resembles the variance of binomial distribution since $0 < \mu < 1$ and can be considered as θ in the binomial model. Unlike the binomila, the variance of the beta-binomial can be controled by τ through $\frac{\tau+n}{\tau+1}$, thus solving the overdispersion problem.

2 Laplace Approximation

Obtain the Laplace approximation for the posterior expection of logit(mu) and log(tau) in the cancer mortality rate example (data available from the package LearnBayes).

Laplace approximation

The posterior expectation of $g(\theta)$ is calculated by taking the ratio of two integrals

$$E(g(\theta)|y) = \frac{\int_\Theta g(\theta)L(y|\theta)\pi(\theta)d\theta}{\int_\Theta L(y|\theta)\pi(\theta)d\theta}$$

where $L(y|\theta)$ is the likelihood, $\pi(\theta)$ is the prior and Θ is the support.

There are two ways to use Laplace approximation to evaluate the integral in the numerator depending on whether $g(\theta) > 0$. Here are the solutions to the two approaches will be presented and requirements associated.

When $g(\theta) > 0$

If $g(\theta) > 0$, we can write $h_1(\theta, x) = -\frac{1}{N}log(g(\theta)L(y|\theta)\pi(\theta))$ and use Taylor expansion on $h_1(\theta, x)$, N is the sample size. Then the integral becomes

$$\begin{aligned}
\int_\Theta g(\theta)L(y|\theta)\pi(\theta)d\theta &= \int_\Theta exp\{-Nh_1(\theta, y)\}d\theta \\
&= \int_\Theta exp\{-N(h_1(\hat{\theta_1}, y) + \frac{1}{2}(\theta - \hat{\theta_1})'H(\hat{\theta_1})(\theta - \hat{\theta_1}))\} \\
&\approx exp\{-N(log(h_1(\hat{\theta_1}, y)))\}(2\pi)^{d/2}|H_1^{-1}|^{\frac{1}{2}}N^{-d/2}
\end{aligned}$$

where $\hat{\theta_1} = argmax_\theta h_1(\theta, y)$ and $H_1 = D^2h_1(\hat{\theta_1}, y)$.

The second term in the expansion is a normal kernel and we get $(2\pi)^{d/2}|H^{-1}|^{\frac{1}{2}}N^{-d/2}$ as the normalizing constant when we integrate over the full support $(-\infty, \infty)$. This means transformation of variable is needed.

support of θ

When the support of θ is not $\mathcal{R}^d$, there should be an additional term in the form of normal CDF. Such terms are "cancelled" when we evaluate the expecation, since both of integrals are being evaluated in the same range.

We can evaluate the intergral in the denominator in the same fashion. Define $h_2(\theta, x) = -\frac{1}{N}log(L(y|\theta)\pi(\theta))$.

$$\begin{aligned}
\int_\Theta L(y|\theta)\pi(\theta)d\theta &= \int_\Theta exp\{-Nh_2(\theta, y)\}d\theta \\
&= \int_\Theta exp\{-N(h_2(\hat{\theta_2}, y) + \frac{1}{2}(\theta - \hat{\theta_2})'H(\hat{\theta_2})(\theta - \hat{\theta_2}))\} \\
&\approx exp\{-N(log(h_2(\hat{\theta_2}, y)))\}(2\pi)^{d/2}|H_2^{-1}|^{\frac{1}{2}}N^{-d/2}
\end{aligned}$$

where $\hat{\theta_2} = argmax_\theta h_2(\theta, y)$ and $H_2 = D^2h_2(\hat{\theta_2}, y)$.

$$E(g(\theta)) = (\frac{|H_1^{-1}|}{|H_2^{-1}|})^{\frac{1}{2}} \frac{exp\{-Nh_1\hat{\theta_1}, y\}}{exp\{-Nh_2(\hat{\theta_2}, y)\}}$$

When $g(\theta)$ can be smaller than 0

In this situation, we can no longer use the approach above since $log(g(\theta))$ can be undefined. Kass and Steffey [@kassApproximateBayesianInference1989] proposed a more general laplace formulation in which we do not fold $g(\theta)$ into $h(\theta, y)$. We have

$$\begin{aligned}
\int_{\Theta} g(\theta)L(y|\theta)\pi(\theta)d\theta &= \int_{\Theta} g(\theta)exp\{-Nh_2(\theta, y)\}d\theta \\
&\approx g(\hat{\theta_2})exp\{-N(log(h_2(\hat{\theta_2}, y)))\}(2\pi)^{d/2}|H_2^{-1}|^{\frac{1}{2}}N^{-d/2}
\end{aligned}$$

Where $\hat{\theta_2} = argmax_{\theta} h_2(\theta, y)$ and $H_2 = D^2 h_2(\hat{\theta_2}, y)$

Notice that when we evaluate the ratio of integral using this method, all the terms except for $g(\hat{\theta_2})$ will be canceled. Thus we have $E(g(\theta)) \approx g(\hat{\theta_2})$. Notice that the $\hat{\theta_2}$ is the poterior mode. Using this approach, we end up with the same result as using normal approximation of the posterior.

Cancer Motality

The posterior distribution in the cancer motality example is defined up to a constant by

$$f(\mu, \tau|y) \propto \prod_{j=1}^{n} \frac{Beta(\mu\tau + y_i, \tau(1-\mu) + n_j + y_j)}{B(\mu\tau, \tau(1-\mu))} \frac{1}{\mu(1-\mu)(1+\tau)^2}$$

Now change $t = logit(\mu) = log\frac{\mu}{1-\mu}, v = log(\tau)$.

$$\begin{aligned}
&log(f(y|v,t) * p(v,t)) = \\
&\sum_{j=1}^{n} log(Beta(\frac{e^{t+v}}{1+e^t} + y_j, e^v - \frac{e^{t+v}}{1+e^t} + n_j + y_j)) - log(Beta(\frac{e^{t+v}}{1+e^t}, e^v - \frac{e^{t+v}}{1+e^t})) \\
&- log(\frac{e^t}{(1+e^t)^2}) - 2log(1+e^v) + log(\frac{e^t}{(1+e^t)^2}) + log(e^v) \\
&= \sum_{j=1}^{n} log(Beta(\frac{e^{t+v}}{1+e^t} + y_j, e^v - \frac{e^{t+v}}{1+e^t} + n_j + y_j)) - log(Beta(\frac{e^{t+v}}{1+e^t}, e^v - \frac{e^{t+v}}{1+e^t})) \\
&+ log(\frac{e^v}{(1+e^v)^2})
\end{aligned}$$

Since we want to evaluate the posterior expectation of v, t, notice that $g(v) = v = log(\frac{\mu}{1-\mu})$ has support $(-\infty, \infty)$. Thus we have to use the second approach. Here we just need to find the posterior mode for v and t by maximizing the log of the function that is proportional to the posterior defined as above.

3 Failure time

The failure time of a pump follows a two-parameter exponential distribution, f(y|b,m) = 1/b exp(-(y-m)/b, when y>=m.

1. Obtain the likelihood for b and m based on an i.i.d. sample of size n

$$f(y_1 \cdots y_n|b, m) = (\frac{1}{b})^n exp\{-\sum_{i=1}^{n} \frac{y_i - m}{b}\} I_{(m,\infty)}(y_{(1)})$$

where $y_{(1)}$ is the minimal of $y_1, \cdots, y_n$.

2. Consider a suitable transformation that maps the parameters b and m to the plane

Notice that the support of posterior distribution of m is $(-\infty, y_{(1)})$. To map m to the entire real line, let $t = log(y_{(1)} - m), t \in R$. Let $v = log(b, v \in R$.

3.1 Multinomial-Direchlet

Given multinomial likelihood for $(y_1, \cdots, y_J)$ with $(\theta_1, \cdots, \theta_J)$ and direchlet prior, define $\alpha = \frac{\theta_1}{\theta_1+\theta_2}$

1. Write the marginal posterior distribuiton for α.

Let the prior for $(\theta_1, \cdots, \theta_J)$ be $\pi((\theta_1, \cdots, \theta_J) \sim Direchlet(a_1, \cdots, a_J)$, i.e. a priori equally likeliy. The posterior then for $(\theta_1, \cdots, \theta_J)$ is $Direchlet(a_1 + y_1, a_2 + y_2, \cdots, a_J + y_J)$. To find the posterior density for $\frac{\theta_1}{\theta_1+\theta_2}$, let's perform the following transformation, $u_1 = \frac{\theta_1}{\theta_1+\theta_2}, u_2 = \theta_1 + \theta_2, u_3 = \theta_3, \cdots, u_J = \theta_J$. The jacobian is $|J| = u_2$. Then we have

$$\begin{aligned} f(u_1, \cdots u_J | y_1, \cdots, y_J) &= \frac{\Gamma(\sum_{i=1}^J a_i + y_i)}{\prod_{i=1}^J \Gamma(a_i + y_i)} (u_1 u_2)^{a_1+y_1-1} (u_2 - u_1 u_2)^{a_2+y_2-1} u_3^{a_3+y_3-1} \quad \cdots (1 - u_2 - u_3 - \cdots - u_{J-1})^{a_J+} \\ &= \frac{\Gamma(\sum_{i=1}^J a_i + y_i)}{\prod_{i=1}^J \Gamma(a_i + y_i)} u_1^{a_1+y_1-1} (1-u_1)^{a_2+y_2-1} u_2^{a_1+y_1+a_2+y_2-1} u_3^{a_3+y_3} \quad \cdots (1 - u_2 - u_3 - \cdots - u_{J-1})^{a} \end{aligned}$$

Observe the equation above, we can see that u_1 and $u_i, i > 1$ are independent and the terms with $u_i, i > 1$ consist a kernel for $Direchlet(a_1 + y_1 + a_2 + y_2, a_3 + y_3, \cdots, a_J + y_J)$. Since we want the density of u_1, integrating over the rest of the terms yeilds the normalizing constant for this direchlet distribution.

$$\begin{aligned} f(u_1) &= \frac{\Gamma(\sum_{i=1}^J a_i + y_i)}{\prod_{i=1}^J \Gamma(a_i + y_i)} \frac{\Gamma(a_1 + y_1 + a_2 + y_2) \prod_{i=3}^J \Gamma(a_i + y_i)}{\Gamma(\sum_{i=1}^J a_i + y_i)} u_1^{a_1+y_1-1} (1-u_1)^{a_2+y_2-1} \\ &= \frac{\Gamma(a_1 + y_1 + a_2 + y_2)}{\Gamma(a_1 + y_1)\Gamma(a_2 + y_2)} u_1^{a_1+y_1-1} (1-u_1)^{a_2+y_2-1} \end{aligned}$$

Recogonize that $0 < u_1 < 1$ and this is the pdf of $beta(a_1 + y_1, a_2 + y_2)$. $\alpha = u_1 \sim beta(a_1 + y_1, a_2 + y_2)$

2. Show that this distribution is identical to the posterior distribution for α obtained by treating y1 as an observation from the binomial distribution with probability α and sample size $y_1 + y_2$, ignoring the data $y_3, ..., y_J$.

We find the prior for α by doing similar transformation to prior $\theta_1, \cdots, \theta_J \sim Direchlet(a_1, \cdots, a_n)$. The details are the same as in 1, except for we are using the prior instead of posterior distribuiton. Thus $\alpha \sim beta(a_1, a_2)$.

Then with a binomial likelhiood with α as the probability as sucssess, conjugacy tells us the posterior is $beta(a_1 + y_1, a_2 + y_2)$, which agrees with the solution in 1.

3.2 Election

Assume that $\theta_{bj}, \theta_{dj}, \theta_{nj}$ are the probability of voting for Bush, Dukakis and no opinion in the j^{th} survey. Then the posterior distribution of $\alpha_j = \frac{\theta_b j}{\theta_b j + \theta_d j} \sim beta(a_b + y_{bj}, a_d + y_{dj})$ according to our result from previous question, where a_b, a_d are from the prior $direchlet(a_b, a_j, a_n)$. To aquire a posterior inference on $\alpha_2 - \alpha_1$, let's get a sample of the posterior distribution of α_1 and α_2 respectively, and for each iteration i let $\delta^{(i)} = \alpha_2^{(i)} - \alpha_1^{(i)}$. This gives us a postetior sample of the difference. Assume non-informative prior, let

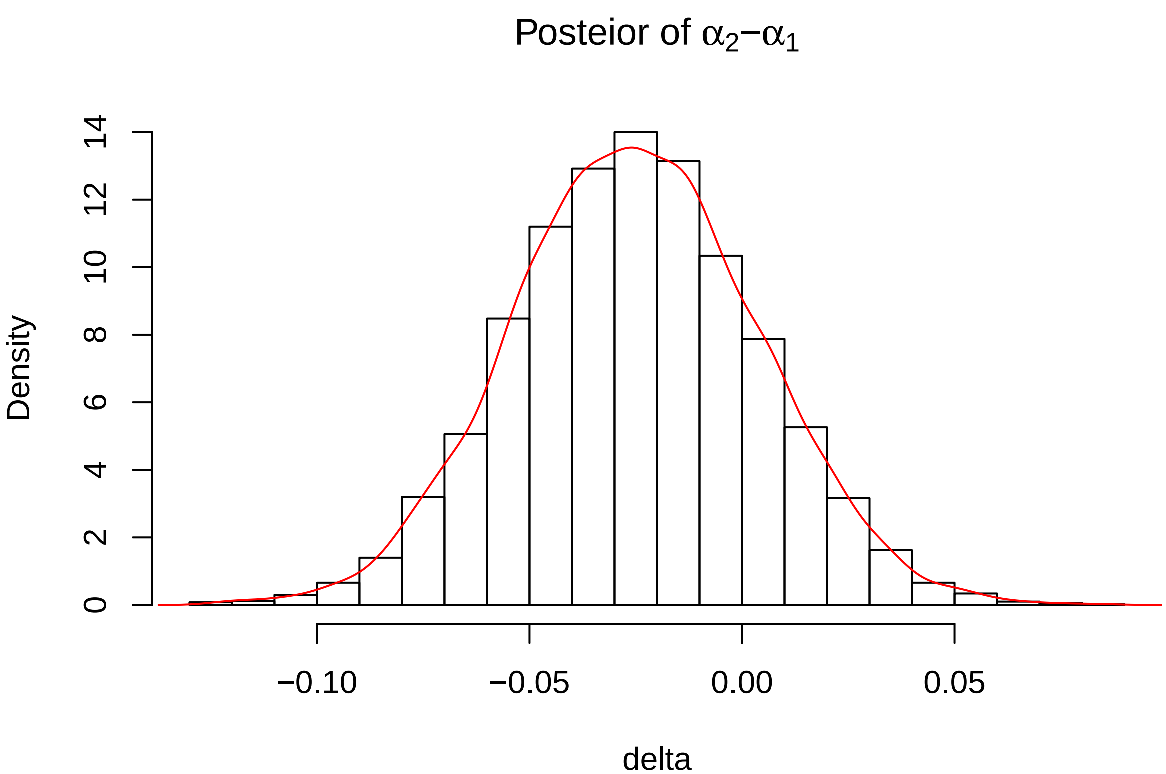
$a_b = a_d = a_n = 1$. Since the sample size is large, adding a prior sample size of one to each category doesn't seem unreasonable. We can easily sample the posteriors from a beta distribution in R.

To calculate the posterior shift towards bush, we want to essentially estimate $P(\alpha_2 - \alpha_1 > 0) = E(I_{[0,\infty]}(\alpha_2 - \alpha_1))$. Using WWLN, we have

$$E(I_{[0,\infty]}(\alpha_2 - \alpha_1)) \approx \frac{1}{N}\sum_{i=1}^{N} I_{[0,\infty]}(\alpha_2 - \alpha_1)$$

Let $N = 5000$, we find this posterior proability to be roughtly 0.1932.

```
alpha_1 = rbeta(5000, shape1 = 1+294, shape2 = 1+307)
alpha_2 = rbeta(5000, shape1 = 1+288, shape2 = 1+332)
delta = alpha_2 - alpha_1
hist(delta, freq=F, breaks=20, main=expression(paste("Posteior of ", alpha[2],"-", alpha[1])))
lines(density(delta), col="red")
```




```
sum(delta>0)/length(delta)
```

```
## [1] 0.191
```

3.5 Rounded Data

Assuming noninformative prior on μ, σ^2, the prior is $\pi(\mu, \sigma^2) \propto \sigma^2$.

1. Treating data as unrounded The joint posterior is given by

$$f(\mu, \sigma^2 | \bar{X}, S^2) \propto (\sigma^2)^{-\frac{n}{2}-1} exp\{-\frac{(n-1)S^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}\}$$

Here $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Integrating over σ^2 yields the marginal posterior for μ which is a t distribution with location $\bar{X}$ and scale $\frac{S^2}{n}$ with $n-1$ degrees of freedom. Thus the posterior expectation for μ is $\bar{X}$. The posterior variance is $\frac{(n-1)S^2}{(n-3)n}$

Intergrating over μ yields the marginal posterior for σ^2 which is a inverse gamma distribuition $IG(\frac{n-1}{2}, \frac{(n-1)S^2}{2})$. The posterior expectation for σ^2 is $\frac{(n-1)S^2}{n-3}$ and the posterior variance is undefined since $\frac{n-1}{2} - 2 = 0$

2. Treating data as rounded.

Treating the data as rounded, the real unrounded data y_i^* is in the range of $[y_i - 0.5, y_i + 0.5]$. So the joing posterior is given by

$$
\begin{aligned}
f(\mu, \sigma^2 | y_1, \cdots, y_n) &= \prod_{i=1}^{n} \int_{y_i-0.5}^{y_i+0.5} \phi(y|\mu, \sigma^2) dy \frac{1}{\sigma^2} \\
&= \prod_{i=1}^{n} (\Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})) \frac{1}{\sigma^2}
\end{aligned}
$$

3. How do the correct and incorrect posteriors differ? Compare mean, variance and contour plots.

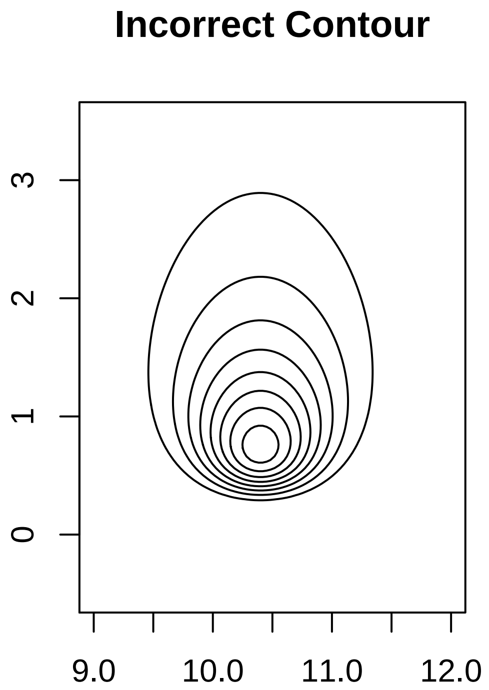
First method, grid approximation

```
dat = c(10, 10, 12, 11, 9)
x_bar = mean(dat)
grid_len = 300
nsample=1000
mu_grid = seq(5, 15, length.out = grid_len)
sigma_grid = seq(0.1, 10, length.out = grid_len)
joint_grid = expand.grid(mu_grid, sigma_grid)


normal_lp = function(mu, sigma_sqr, dat) {
    log_dens = (-1)*log(sigma_sqr) + sum(dnorm(dat, mu, sqrt(sigma_sqr),log = T))
    return(log_dens)
}
sum_stats = function(x){c(mean(x), sd(x), quantile(x, c(.025,.25,.5,.75,.975)))}

lp = apply(joint_grid, 1, function(x){exp(normal_lp(x[1], x[2], dat))})


lp_mat = matrix(lp, nrow = grid_len, ncol=grid_len)

incorrect_dens = lp_mat/(sum(lp_mat))
mu_density = apply(incorrect_dens, 1, sum)
sigma_density= apply(incorrect_dens, 2, sum)

mu_sample = sample(mu_grid, size=nsample, prob=mu_density, replace = T)
sigma_sample = sample(sigma_grid, size=nsample, prob=sigma_density, replace = T)

correct_log_post = function(mu, sigma, dat){
    llk = sum(sapply(dat, function(x){log(pnorm(x+0.5, mu, sigma) - pnorm(x-0.5, mu, sigma))}))
    log_post = llk - log(sigma^2)
    return(log_post)
}
```

```
clp =  apply(joint_grid, 1, function(x){exp(correct_log_post(x[1], sqrt(x[2]), dat))})

clp_mat = matrix(clp, nrow = grid_len, ncol=grid_len)


correct_dens = clp_mat/(sum(clp_mat))
mu_density = apply(correct_dens, 1, sum)
sigma_density= apply(correct_dens, 2, sum)

correct_mu_sample = sample(mu_grid, size=nsample, prob=mu_density, replace = T)
correct_sigma_sample = sample(sigma_grid, size=nsample, prob=sigma_density, replace = T)

output_table = rbind(sum_stats(mu_sample), sum_stats(correct_mu_sample), sum_stats(sigma_sample), sum_s

rownames(output_table) = c("Incorrect $\\mu$", "Correct $\\mu$", "Incorrect $\\sigma$", "Correct $\\sig

colnames(output_table) = c("Mean", "Sd", "2.5%", "25%", "50%", "75%", "97.5%")

knitr::kable(output_table)
```

	Mean	Sd	2.5%	25%	50%	75%	97.5%
Incorrect μ	10.382609	0.6217102	9.2140468	10.0167224	10.351171	10.719064	11.622910
Correct μ	10.390903	0.6445625	9.0468227	10.0418060	10.384615	10.752508	11.723244
Incorrect σ	2.084371	1.7158881	0.4642140	0.9277592	1.490636	2.483947	7.253495
Correct σ	1.912197	1.7099283	0.3979933	0.8284281	1.358194	2.285284	7.715385

```
par(mfrow=c(1,2))
contour(mu_grid, sigma_grid, lp_mat,drawlabels = F, xlim=c(9, 12), ylim=c(-0.5, 3.5), main="Incorrect Co
contour(mu_grid, sigma_grid, clp_mat, xlim=c(9, 12), ylim=c(-0.5, 3.5),drawlabels = F, main="Correct Con
```



4. Let $z = (z_1, \cdots, z_5)$ be the original unrounded measures corresponding to the five observations above. Draw simulation from the posterior distribution of z. Compute the posterior mean of $(z_1 - z_2)^2$.

Use inverse CDF sampling, $z|\mu, \sigma, y$ comes from a truncated normal with mean μ, variance σ^2 and truncated by $[y - 0.5, y + 0.5]$. The CDF of $Z|\mu, \sigma, y$ is given by

$$F_z(z_i) = \frac{1}{\Phi(y_i + 0.5|\mu, \sigma^2) - \Phi(y_i - 0.5|\mu, \sigma^2)}[\Phi(z|\mu, \sigma^2) - \Phi(y_i - 0.5|\mu, \sigma^2)]$$

Thus once sample a point between $\Phi(y_i + 0.5|\mu, \sigma^2$ and $\Phi(y_i - 0.5|\mu, \sigma^2)$, we can find z by taking the inverse.

```
lower_bound = dat-0.5
upper_bound = dat + 0.5

z = c()

for(i in 1:length(dat)){
    lower = pnorm(lower_bound[i], correct_mu_sample, sqrt(correct_sigma_sample))
    upper =  pnorm(upper_bound[i], correct_mu_sample, sqrt(correct_sigma_sample))
    this_z = qnorm (lower + runif(length(correct_mu_sample))*(upper-lower), correct_mu_sample, sqrt(cor
    z = cbind(z, this_z)
}

mean((z[, 1]-z[, 2])^2)
```

```
## [1] 0.1637309
```

The posterior mean of $(z_2 - z_1)^2$ is 0.16.

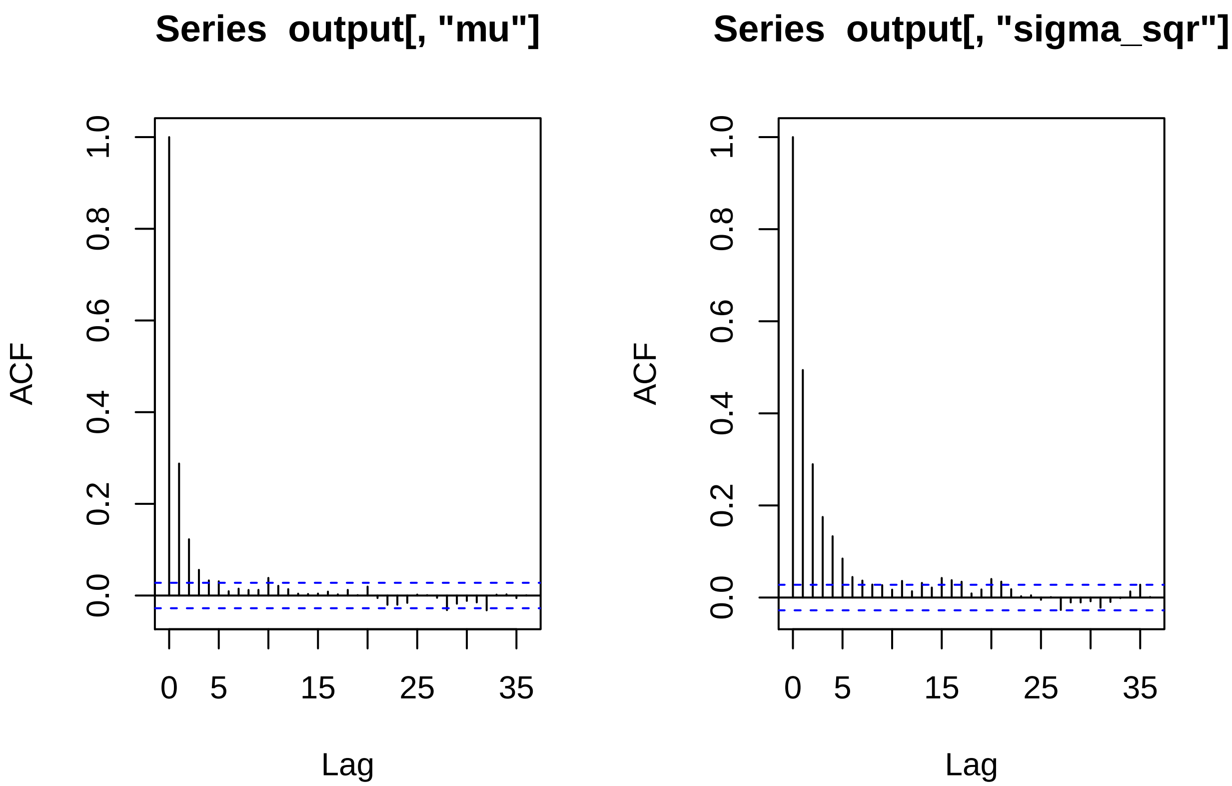
View z_i as the latent variable. Then the joint posterior of (z, μ, σ^2) becomes

$$f(z, \mu, \sigma^2|y) \propto \prod_{i=1}^{n} I_{[y_i - 0.5, y_i + 0.5]}(z_i)\phi(z_i|\mu, \sigma^2)\frac{1}{\sigma^2}$$

Thus alter the metroplis-hastings algorithm by adding a step proposing z_i from $unif(y_i - 0.5, y_i + 0.5)$ to get a sample of $z_1, \cdots, z_5$.

```
mh_to_sample_correct_mu_sigma_z = function(dat, num_iters=5000){
    output = c()


    n = length(dat)
    s_square = sd(dat)^2
    mu_mean = mean(dat)
    mu_var = 1

    sigma_mean = (n-1)*s_square/(n-3)
    sigma_var = 0.5

    uppder_bound = dat+0.5
    lower_bound = dat-0.5

    z_cur = runif(5, lower_bound, uppder_bound)
    mu_cur = 10
    sigma_cur = 2.6
    post_cur =  normal_lp(mu_cur, sigma_cur^2, z_cur)
```

```
    for (i in 1:num_iters){
        mu_pro = rnorm(1, mu_mean, mu_var)
        sigma_pro = sqrt(rnorm(1, sigma_mean, sigma_var))
        z_pro = runif(5, lower_bound, uppder_bound)
        post_pro = normal_lp(mu_pro, sigma_pro^2, z_pro)

        if(log(runif(1))<(post_pro-post_cur)){
            mu_cur = mu_pro
            sigma_cur = sigma_pro
            z_cur = z_pro
            post_cur = post_pro

        }

        output = rbind(output, c(mu_cur, sigma_cur^2, z_cur, post_cur))

    }

    colnames(output) = c("mu", "sigma_sqr", paste("z", seq(1,5)), "log_posterior")
    return(output)
}
output = mh_to_sample_correct_mu_sigma_z(dat)

par(mfrow=c(1, 2))
acf(output[, "mu"])
acf(output[, "sigma_sqr"])
```



```r
output_thin = output[seq(1000, 5000, by=5), ]
cmu_mean= mean(output_thin[, "mu"])
csig_mean = mean(output_thin[, "sigma_sqr"])

z_1 = output_thin[, "z 1"]
z_2 = output_thin[, "z 2"]

mean((z_2-z_1)^2)
```

```
## [1] 0.1589183
```

The posterior mean of $(z_2 - z_1)^2$ is approximately 0.17.

4.1 Cauchy

1. Find the posterior first and second derivative

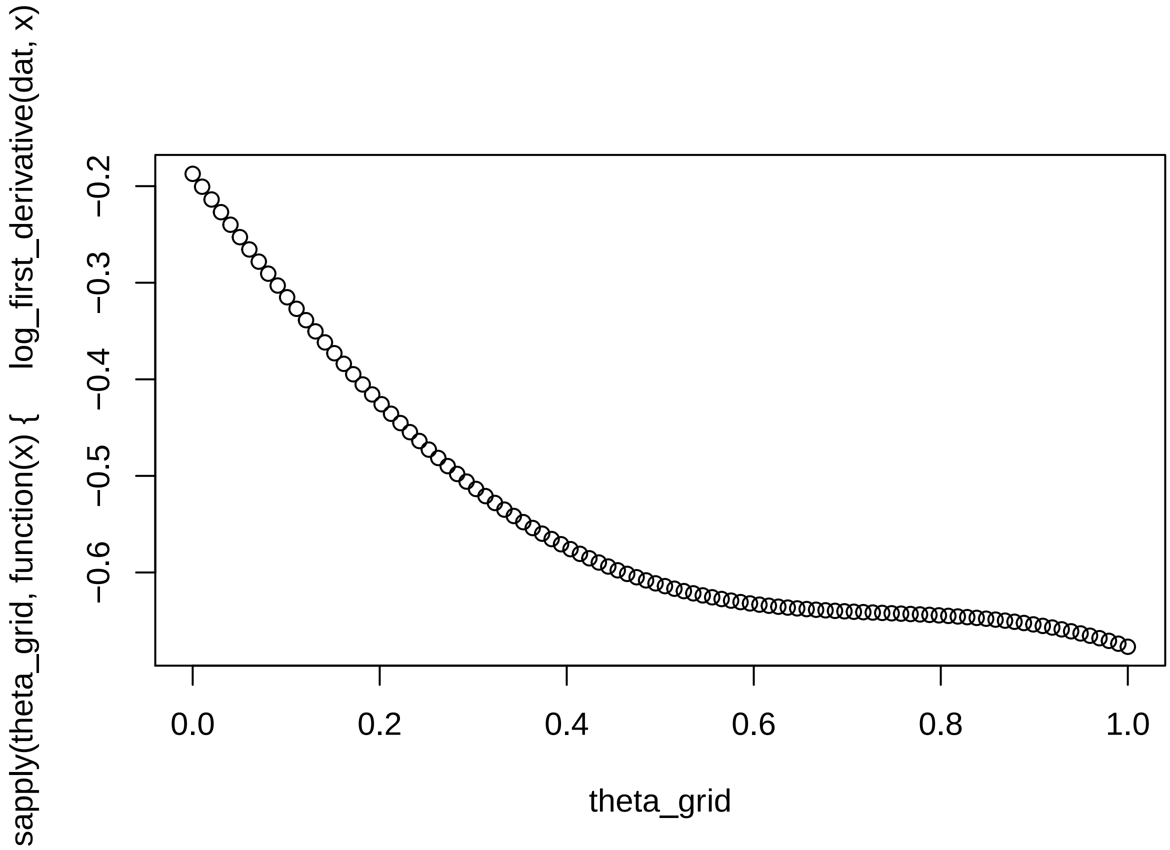
$$log(f(\theta|y)) = log(constant) - \sum_{i=1}^{n} log(1 + (y_i - \theta)^2)$$

$$\frac{d}{d\theta} log(f(\theta|y)) = \sum_{i=1}^{n} \frac{2(y_i - \theta)}{1 + (y_i - \theta)^2}$$

$$\frac{d^2}{d\theta^2} log(f(\theta|y)) = -\sum_{i=1}^{n} \frac{2 - 2(y_i - \theta)^2}{(1 + (y_i - \theta)^2)^2}$$

2. Find posterior mode

```r
dat = c(-2, -1, 0, 1.5, 2.5)

log_first_derivative = function(dat, theta){
    output = sum(sapply(dat, function(x){2*(x-theta)/(1+(x-theta)^2)}))
    return(output)
}

theta_grid = seq(0, 1, length.out =100)

plot(sapply(theta_grid, function(x){log_first_derivative(dat, x)})~theta_grid)
```



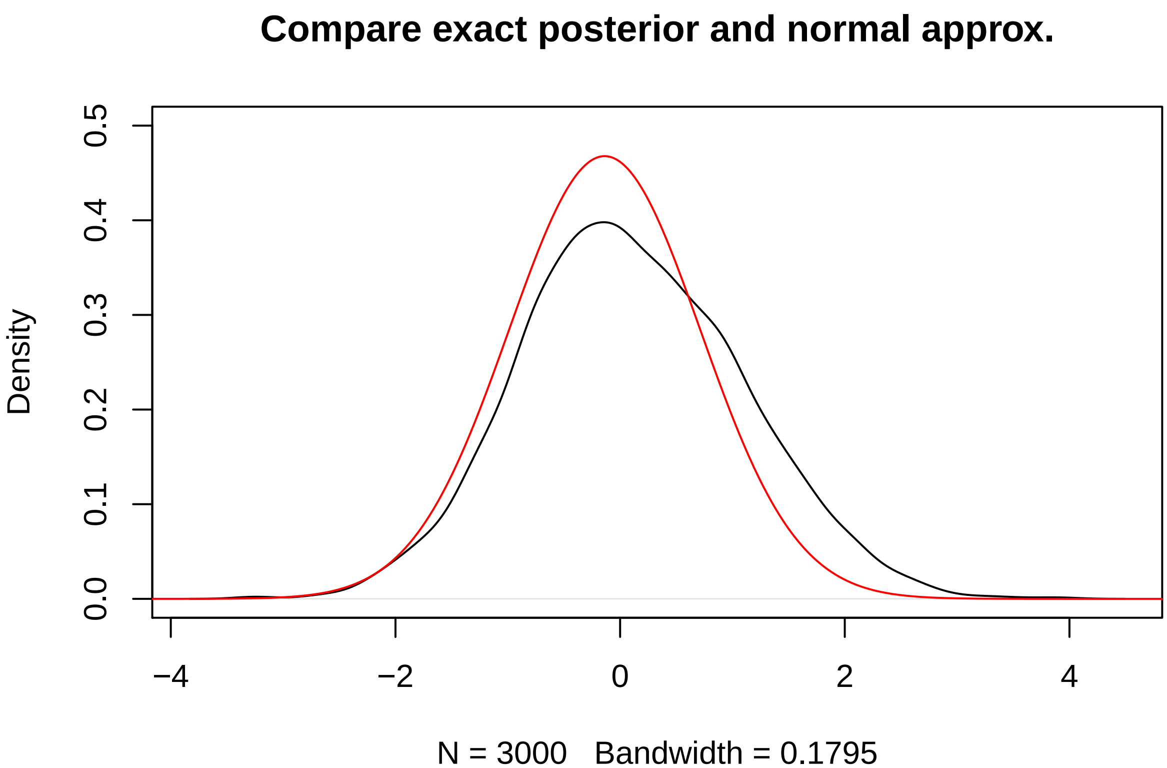

```
log_second_derivative = function(dat, theta){
    output = sum(sapply(dat, function(x){-(2-2*(x-theta)^2)/(1+(x-theta)^2)^2}))
    return(output)
}

get_theta_hat = function(dat, theta_init = 0.1, tol=1e-5){
    theta_cur = theta_init
    delta = 1

    while(delta > tol){
        theta_pro = theta_cur - log_first_derivative(dat, theta_cur)/log_second_derivative(dat, theta_cu
        delta = abs(theta_pro-theta_cur)
        theta_cur = theta_pro
    }

    return(theta_cur)
}

theta_hat = get_theta_hat(dat)
approx_sigma = -1/log_second_derivative(dat, theta_hat)
```

3. Use second derivative to do normal approximation.

$$\sigma^2 = -(\frac{d^2}{d\theta^2} log(f(\theta|y))|_{\theta=\hat{\theta}})^{-1} \approx 0.728$$

$$\theta|y \sim N(\hat{\theta}, \sigma^2)$$

```
expand_theta_grid = seq(-10, 10, length.out = 1000)
log_cauchy_posterior = function(dat, theta){
    output = - sum(sapply(dat, function(x){log(1 + (x-theta)^2)}))
    return(output)
}

unormalized_post = exp(sapply(expand_theta_grid, function(x){log_cauchy_posterior(dat, x)}))

normalized_post = unormalized_post/sum(unormalized_post)

plot(density(sample(expand_theta_grid,size=3000, prob=normalized_post, replace = T)), ylim=c(0, 0.5), ma
lines(dnorm(expand_theta_grid, theta_hat, sqrt(approx_sigma))~expand_theta_grid, col="red")
```



4.2 Normal approximation of logit

$$p(\alpha, \beta|y, n, x) \propto \prod_{i=1}^{K} \frac{exp\{\alpha + \beta x_i\}^{y_i}}{(1 + exp\{\alpha + \beta x_i\})^{n_i}}$$

$$log(p(\alpha, \beta|y, n, x)) = log(constant) + \sum_{i=1}^{K} (\alpha + \beta x_i)y_i - n_i log(1 + exp\{\alpha + \beta x_i\})$$

$$\frac{\partial}{\partial \alpha} log(p(\alpha, \beta|y, n, x)) = \sum_{i=1}^{K} y_i - n_i \frac{exp\{\alpha + \beta x_i\}}{1 + exp\{\alpha + \beta x_i\}}$$

$$\frac{\partial}{\partial\beta}log(p(\alpha,\beta|y,n,x)) = \sum_{i=1}^{K} x_i y_i - n_i \frac{x_i exp\{\alpha+\beta x_i\}}{1+exp\{\alpha+\beta x_i\}}$$

$$\frac{\partial^2}{\partial\alpha^2}log(p(\alpha,\beta|y,n,x)) = -\sum_{i=1}^{K} n_i \frac{exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2}$$

$$\frac{\partial^2}{\partial\beta^2}log(p(\alpha,\beta|y,n,x)) = -\sum_{i=1}^{K} n_i \frac{x_i^2 exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2}$$

$$\frac{\partial^2}{\partial\alpha\beta}log(p(\alpha,\beta|y,n,x)) = -\sum_{i=1}^{K} n_i \frac{x_i exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2}$$

So the information matrix is of the form

$$\begin{aligned} I(\alpha,\beta) &= \begin{bmatrix} -\sum_{i=1}^{K} n_i \frac{exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2} & -\sum_{i=1}^{K} n_i \frac{x_i exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2} \\ -\sum_{i=1}^{K} n_i \frac{x_i exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2} & -\sum_{i=1}^{K} n_i \frac{x_i^2 exp\{\alpha+\beta x_i\}}{(1+exp\{\alpha+\beta x_i\})^2} \end{bmatrix} \\ &= \begin{bmatrix} -a & -c \\ -c & -b \end{bmatrix} \end{aligned}$$

Evaluate $I(\alpha,\beta)$ at posterior mode $(\hat{\alpha},\hat{\beta})$ gives us the information matrix.

Invert $-I(\hat{\alpha},\hat{\beta})$ we have the normal approximation variance

$$\Sigma = [-I(\hat{\alpha},\hat{\beta})]^{-1} = \begin{bmatrix} \frac{b}{ab-c^2} & -\frac{c}{ab-c^2} \\ -\frac{c}{ab-c^2} & \frac{a}{ab-c^2} \end{bmatrix}$$

Reference