Rounded Data

Robert Espinoza

Abstract

In this case study we will be examining UV index data, which is rounded data to the nearest whole number. It is unreasonable to assume the rounded data is a random sample from a normal distribution, but it would be reasonable to assume the unobserved unrounded measurements are independently distributed from a normal distribution with mean μ and variance σ^2. We analyze the data in two different models, model (a) does not take into account the original $x_1, .., x_{10}$ unobserved unrounded data in the joint posterior distribution, and model (b) takes into account does take into account the $x_1, .., x_{10}$ in the joint posterior distribution. In both models we assume a non-informative prior $(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. The rounded data is as follows: 7, 6, 7, 5, 5, 3, 6, 5, 4, 3.

Posterior density, accounting for rounding

1. Model (a)

First we determine the probability of the individual observations given our parameters of interest (μ, σ^2).

$$Pr(y_i|\mu, \sigma^2) = \Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})$$

The joint posterior distribution for model (a) is as follows:

$$p(\mu, \sigma^2|\mathbf{y}) \propto p(\mu, \sigma^2)p(\mathbf{y}|\mu, \sigma^2)$$
$$= \frac{1}{\sigma^2} \prod_{i=1}^{10} [\Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})]$$

We are now interested in sampling from the joint posterior distribution, which will be executed by an grid approximation sampling algorithm. The sketch of the algorithm is as follows:

1. Generate a grid of points for μ and σ^2, and contour values.
2. Determine the target posterior distribution.
3. Generate a contour plot with the grid points and posterior distribution.
4. Normalize μ marginal density.
5. Sample μ indices from the μ grid of points with the normalized μ density.
6. Generate μ sample from sampled indices at the μ grid of points.
7. Generate σ^2 sample at the σ^2 grid of points with posterior density probabilities at the sampled μ indices.
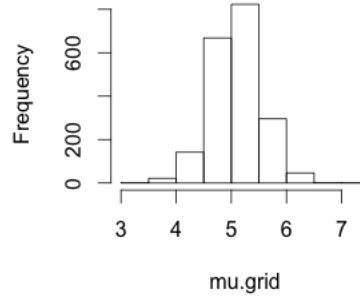
The above algorithm will provide us with an appropriate sample of (μ, σ^2)
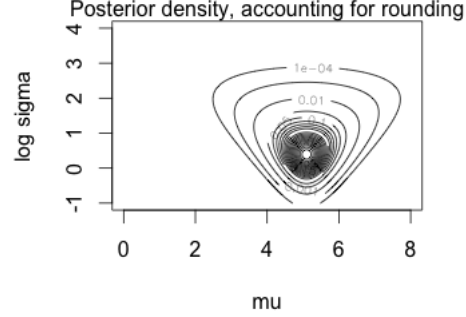
Histogram of mu.grid
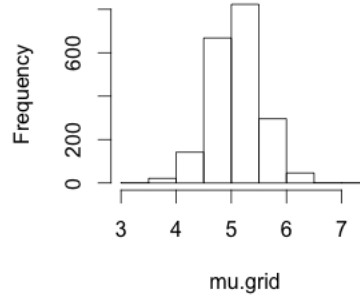
Histogram of mu.grid

2. Model (b)

The second model we will take into account is a posterior distribution with the unobserved unrounded measurements $x_1, .., x_{10}$ as random variables. Since the data involves x_i covariates in the posterior, we will require an inclusion distribution to derive the complete likelihood of the data. The posterior distribution is as follows:

$$p(\mu, \sigma^2, \mathbf{x}|\mathbf{y}, I) \propto p(\mu, \sigma^2)p(\mathbf{y}|\mu, \sigma^2)p(I|\mathbf{x}, \mathbf{y})p(\mathbf{x}|\mu, \sigma^2)$$

$$= \prod_{i=1}^{10}[\Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})]$$

$$\mathbb{1}_{y_i - 0.5 < x_i < y_i + 0.5}(\sigma^2)^{-6}$$

$$\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{10}(x_i - \mu)^2)$$

Histogram of mu.mcmc

In order to sample from this particular joint posterior distribution we will us a Markov Chain Monte Carlo sampling algorithm. We will utilize a Gibbs sampler with a random walk Metropolis Hastings algorithm to sample $(\mu, \sigma^2, x_1, .., x_{10})$. The following are the full conditional distributions that will be used in the Gibbs sampling algorithm:

$$p(x_i|y_i, \mu, \sigma^2, I) \propto \mathbb{1}_{y_i - 0.5 < x_i < y_i + 0.5}\exp(\frac{-1}{2\sigma^2}(x_i - \mu)^2)$$

Histogram of sig2.mcmc

$$p(\mu|\mathbf{y}, \mathbf{x}, \sigma^2) \propto \prod_{i=1}^{10}[\Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})]$$

$$\exp(\frac{-5}{\sigma^2}(\mu - x_i)^2)$$

3. Model Comparison

Below are summary statistics of model (a):

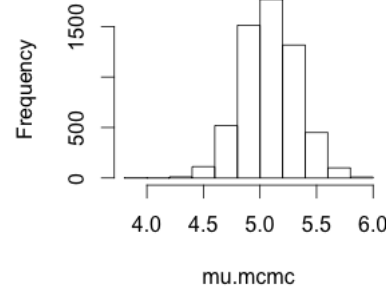| parameter	Mean	95% Posterior Interval
μ	5.09	[4.18, 5.98]
σ^2	1.95	[0.743, 4.65]

Below are the summary statistics of model (b):

$$p(\sigma^2|\mathbf{y}, \mathbf{x}, \mu) \propto \prod_{i=1}^{10}[\Phi(\frac{y_i + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_i - 0.5 - \mu}{\sigma})]$$

$$(\sigma^2)^{-6}\exp(\frac{-1}{2\sigma^2}\sum_{i=1}^{10}(x_i - \mu)^2)$$

Parameter	Mean	95% Posterior Interval
μ	5.08	[4.61, 5.57]
σ^2	1.04	[0.519, 2.04]
x_1	6.84	[6.51, 7.43]
x_2	5.92	[5.51, 6.45]
x_3	6.84	[6.51, 7.43]
x_4	5.01	[4.53, 5.47]
x_5	5.01	[4.52, 5.47]
x_6	3.16	[2.57, 3.49]
x_7	5.92	[5.52, 6.45]
x_8	5.01	[4.52, 5.47]
x_9	4.09	[3.55, 4.48]
x_{10}	3.15	[2.58, 3.48]

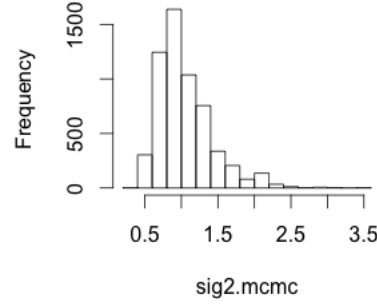The sketch of the Gibbs sampling algorithm is as follows:

1. set μ_0 and σ_0^2 starting values.
2. draw x_1 though x_10 using the full conditional of each of the x_i and use necessary current values to update, save draws as current values.
3. draw μ using the full conditional of μ which is approximated using a random walk Metropolis Hastings step using necessary current values to update, and save as current value.
4. draw a σ^2 using the full conditional of σ^2 which is approximated using a random walk Metropolis Hastings step using necessary current values to update, and save as current value.
5. burn-in and thin draws to eliminate correlation between draws.
6. repeat the above steps until necessary number of iterations has been completed.

The above algorithm will provide samples for $(x_1, ..., x_{10}, \mu, \sigma^2)$.

The confidence intervals of model (a) versus model (b) for μ and σ^2 are more tightly bounded for model2. This is possibly due to the fact that we introduced the inclusion parameter and unknown unrounded random variables into the joint posterior. The mean for σ^2 for model (b) was noticeably lower than model (a), which may be due to an error in my analysis,

or maybe be due to the fact that with the addition of the x_i random variables, it improved the estimate of the σ^2, which may indicate that the true σ^2 is near 1. With the tighter bounds around μ and σ^2, model (b) seems to be more robust, and should be the model to work with for future analysis. The DIC for model (a) is 41.35 and the DIC for model (b) i s40.75, therefore the DIC agrees with our assumption that model (b) is a better model.

REFERENCES

Gelman A., Robert C, Chopin N., Rousseau J.(2013), "Bayesian Data Analysis"