

# Stats4Astro 2017

## Model Building Lab

This lab will focus on writing MCMC samplers that account for selection effects. This problem is a simplified version of a cosmological techniques that uses Type Ia super novae (SNIa) as standard(izable) candles in the estimation of cosmological parameters that describe the expansion history of the universe. Again we use simulated data sets rather than real data to avoid technical difficulties that arise in fully accounting for the actual data generation mechanisms. Simulations and model fitting are based on the  $\Lambda$ -CDM model (using  $\Omega_m = 0.3$ ,  $\Omega_\kappa = 0$ ,  $H_0 = 67.3$  km/s/Mpc) which is provided in tabulated form.

file name	columns	description
<code>Lambda-CDM.txt</code>	$z$ = redshift, $\mu(z)$ = distance modulus	$\Lambda$ -CDM model as a function of $z$

Suppose the distribution of the absolute magnitude of SNIa follow a normal distribuiton,

$$M_i \sim \text{NORM}(\mu, \sigma^2),$$

where  $M_i$  represents absolute magnitude of SNIa  $i$ ,  $\mu$  is the mean absolute magnitude, and  $\sigma^2$  is the variance of the absolute magnitudes of SNIa. We are interested in estimating  $\mu$  and  $\sigma^2$  via a Bayesian analysis, using independent prior distributions on  $\mu$  and  $\sigma^2$ :  $\mu \sim \text{NORM}(-19.3, 20^2)$  and  $\sigma^2 \sim \beta^2/\chi_\nu^2$ , with  $\beta^2 = \nu = 0.02$ .

Unfortunately, we do not observe the absolute magnitudes, but rather observe the apparent magnitudes

$$m_i = \mu(z_i) + M_i,$$

where  $z_i$  is the observed redshift of SNIa. Values of  $\mu(z_i)$  are given in `Lambda-CDM.txt`. These values assume  $\Omega_m = 0.3$ ,  $\Omega_\kappa = 0$ , and  $H_0 = 67.3$  km/s/Mpc; we take these values as given throughout this exercise. Suppose, owing to instrumental constraints we only observe SNIa with  $m_i < 24$ .

For given true values of  $\mu$  and  $\sigma^2$  we can simulate a set of redshifts as well as absolute and apparent magnitudes for a hypothetical set of SNIa. Let  $\mu_{\text{true}}$  and  $\sigma_{\text{true}}^2$  represent these true values. We set  $\mu_{\text{true}} = -19.3$  throughout, but consider how the value of  $\sigma_{\text{true}}^2$  effect our results by varying its value.

Consider a set of  $n$  SNIa and let  $N$  be the subset of these with  $m_i < 24$ . Our observed dataset will be of size  $N$ . For given values of  $n$  and  $\sigma_{\text{true}}^2$  the R-code in Table 1 can be used to simulate a dataset. (This code assumes that the redshifts of SNIa are distributed with density  $\propto (1 + z_i)^2$ .) The output from the R-code is summarized in Table 2.

1. Simulate a data set with  $n = 200$  and  $\sigma_{\text{true}} = 3$ . Make a plot of your data with redshift on the horizontal axis and absolute magnitude on the vertical axis. Use color coding to indicate which SNIa are observed and plot a line to indicate the observation cut threshold (above which SNIa are not observed).

---

Table 1: R-code for simulating a dataset.

---

```
# Paramters
n          <- 200 # sample size before selection effects
var.true   <- 9   # intrinsic standard deviation of absolute magnitudes.

# sample the redshifts
z.pts      <- 1:100/100
z.prob     <- (1+z.pts)^2
z.sim      <- sample(z.pts, size=n, replace=TRUE, prob=z.prob)

# read in Lambda CDM model
LCDM       <- read.table("LambdaCDM.txt", header=TRUE)

# Simulate absolute magnituded
M.sim      <- rnorm(n, -19.3, sqrt(var.true))
# compute the apparent magnitudes, as absolute magnitude + mu (i.e., distance modulus)
m.sim      <- M.sim + LCDM[round(z.sim*100),2]

# Select if m.sim < 24
z.sel      <- z.sim[m.sim<24]
M.sel      <- M.sim[m.sim<24]
m.sel      <- m.sim[m.sim<24]

# observed sample size
N          <- length(m.sel)
```

---

2. Write a Gibbs sampler to sample from the joint posterior distribution,  $p(\mu, \sigma^2 \mid \mathbf{M.sel})$  ignoring selection effect. How does the posterior distribution compare with the true parameter values?
3. Suppose  $X$  follows an normal distribution with mean  $m$  and standard deviation  $s$ , but we only observe  $X$  if it is less than some threshold  $t$ . The observed variable follows a *truncated normal distribution*. Use the R-functions `dnorm` and `pnorm` to write a function that computes the density of this truncated normal distribution and the natural log of this density. An outline of the code:

```
dtnorm <- function(x, mean, sd, truncation, log=FALSE){

  value <- 0
  if (x <= truncation){
    value <- #compute the density here.
  }

  if(log == TRUE) value <- log(value)

  value
}
```

Table 2: Output from R-code for simulating a dataset.

variable name	description
<code>z.sim</code>	Simulated redshifts <i>before</i> selection effect
<code>M.sim</code>	Absolute magnitudes <i>before</i> selection effect
<code>m.sim</code>	Apparent magnitudes <i>before</i> selection effect
<code>z.sel</code>	Simulated redshifts <i>after</i> selection effect
<code>M.sel</code>	Absolute magnitudes <i>after</i> selection effect
<code>m.sel</code>	Apparent magnitudes <i>after</i> selection effect

In this code `dtnorm` stands for the density of a **t**runcated **n**ormal; `x` is the point at which the (log) density is evaluated, `mean` is the mean of the parent (untruncated) normal, `sd` is the standard deviation of the parent (untruncated) normal, `truncation` is the truncation point, and `log` should be set to `TRUE` to return the natural log of the density.

4. Write a Metropolis within Gibbs sampler to sample from the joint posterior distribution,  $p(\mu, \sigma^2 \mid \mathbf{M.sel})$  *accounting for selection effect*. How does the posterior distribution compare with the true parameter values? Compare these results with your answer to Question 2.
5. Suppose there were no selection effects and data were available for the full  $n = 200$  SNIa. Run your Gibbs sampler from Question 2 to sample the joint posterior distribution,  $p(\mu, \sigma^2 \mid \mathbf{M.sim})$  *ignoring selection effect*. (Note you are using the `M.sim` rather than `M.sel` in this analysis.) How does the posterior distribution compare with the true parameter values? Compare these results with your answer to Question 4.
6. You should have found the posterior variances of  $\mu$  and  $\sigma^2$  to be larger in Question 4 than in Question 5. Experimenting with the value of  $n$ , how large must  $n$  be for  $\text{Var}(\mu \mid \mathbf{M.sel})$  to be about the same size as  $\text{Var}(\mu \mid \mathbf{M.sim})$  computed with  $n = 200$  in Question 5? What is the corresponding value of  $N$ ?
7. In practice the apparent, not absolute magnitudes are observed. Write an MCMC sampler for  $p(\mu, \sigma^2 \mid \mathbf{m.sel})$ .
8. *Bonus:* In practice, the observed apparent magnitudes include observation errors. Suppose  $m_i^{\text{obs}} = m_i + e_i$ , where  $e_i$  are independent mean-zero Gaussian observation errors with known variance,  $\tau^2$ . Generalize the sampler you wrote for Question 7 to account for observation errors. Use a simulation study, varying the values of  $\tau^2$  and  $\sigma^2$  to explore how the observation errors effect the final error bars for  $\mu$ .