Math 459 Midterm 2
Qian Liu

1. We have introduced methods for exact and for approximate Bayesian inference. Briefly explain and comment on the relative advantages and disadvantages of the following approaches to Bayesian inference, making sure you clearly indicate whether each approach corresponds to exact or approximate inference.

(a) Analytic calculation of the posterior using a conjugate prior.

(b) The Bernstein-von Mises theorem.

(c) Markov Chain Monte Carlo.

Sol:

a) Explain and comment:
if the posterior distributions $p(\theta|x)$ are in the same family as the prior $p(\theta)$, then the prior is called a conjugate prior for $p(\theta|x)$.
Advantages:
It is easy to find the posterior when using conjugate priors because we know it must belong to the same family of distributions as the prior. Also be Interpretable as additional data.
Disadvantages:
Can be overly restrictive and whether a conjugate prior exists depends on the form of the likelihood function. Most cases do not have conjugate distributions.

b) **Explain and comment:**
In Bayesian inference, the Bernstein–von Mises theorem provides the basis for the important result that the posterior distribution for unknown quantities in any problem is effectively independent of the prior distribution (assuming it obeys Cromwell's rule) once the amount of information supplied by a sample of data is large enough.
Advantages:
It doesn't matter whether we have a good prior if we have a very large set of date. The posterior density can converge on the wrong result, but it should be noted that the posterior mode is consistent and will converge on the correct result when we have a very large data set.
Disadvantages:
Sometimes the choice of prior distribution is unimportant in practice, because it hardly influences the posterior distribution at all when there are moderate amounts of data. But if we do not try to find a better prior for the model, then bad things may happen if we were not given a reasonable good sample of data.

c) **Explain and comment:**
MCMC is a stochastic procedure that repeatedly generates random samples that characterize the distribution of parameters of interest. Markov chains are used to draw random samples from a distribution and Monte Carlo integration is used to generate summary estimates from those random samples. And the state of the chain after a number of steps is then used as a sample of the desired distribution.
Advantages:
MCMC offers an appealing approach to handling some difficult types of analyses. it is quite effective at handling complex models. minimal requirements on f .
convergence properties of Markov chains can be exploited to make things easier
Disadvantages:
Time consuming and some Markov chains fail to converge quickly.
there is no guarantee that the chain has converged after M draws, even for a very large M.

Math 459 Midterm 2
Qian Liu

2. Consider Bayesian inference for the parameters of a normal population with mean μ and variance σ^2. Suppose the joint prior $\pi(\mu, \sigma^2) = 1/\sigma^2$ is used. Show your work for each question.

(a) Is the prior proper?

(b) For a sample of size one, is the posterior proper?

(c) Is the posterior proper for a sample size $n > 1$?

Sol:

a) improper (uniform over some domain) prior, prior density may not integrate to unity over its domain.

b) estimating the mean of a normal distribution with known variance given a set of samples. Since we have

$$P(\mu, \sigma^2 |y) \propto P(y|\mu, \sigma^2)P(\mu, \sigma^2)$$

And $$P(\mu|y) = \int P(\mu, \sigma 2 |y)\, d\sigma^2$$

Consider the prior: $P(\mu, \sigma^2) \propto (\sigma^2)^{-1}$

When n= 1, $P(\mu, \sigma^2|y) \propto \sigma^{-3}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$

$$P(\mu|y) \propto \int \sigma^{-3}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) d\sigma^2$$

$P(\mu|y)$ here also belongs to exponential family, so the posterior is proper.

c) When n > 1, then this yields a posterior:

$$\begin{aligned} P(\mu, \sigma^2|y) &\propto \sigma^{-n-2}exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ &= \sigma^{-n-2}exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance of the y_i's.

$$P(\mu|y) = \int P(\mu, \sigma^2)d\sigma^2 \propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}$$

which is, by inspection, a student-t distribution $\theta \sim t_\nu(\psi, \rho^2)$ with $\nu = n-1$ degrees of freedom, mean $\psi = \bar{y}$ and scale $\rho^2 = \frac{s^2}{n}$. The pdf of the student-t distribution is given by

$$P(\mu|\psi, \nu, \rho) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\mu\pi\rho}}(1 + \frac{1}{\nu}(\frac{\mu-\psi}{\rho})^2)^{-(\nu+1)/2}$$

Under the sampling distribution $P(y|\mu, \sigma^2)$ the following relation holds when n is small.

$$\frac{\bar{y}-\mu}{s/\sqrt{n}}|\mu, \sigma^2 \sim t_{n-1}$$

so the posterior is proper.

Math 459 Midterm 2
Qian Liu

3. Explain in what sense Gibbs sampling is a special case of Metropolis-Hastings. Be specific. The best way to get full credit is to explain, in detail, how the algorithms are related.

Ans:

First I want to introduce the idea of Metropolis-Hastings and Gibbs sampling, then, based on those introduction, I will explain their relation.

1) Idea of Metropolis-Hastings :

the Metropolis-Hastings algorithm generates correlated variables from a Markov chain.

The algorithm is :

Given an initial value $x^{(t)}$ and conditional density $q(y|x^{(t)})$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Set

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, Y_t) \end{cases}$$

where

$$\rho(x, y) = \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

q is the proposal or candidate distribution And for every given q, we can construct a M-H kernel K such that f is its stationary distribution.

it is obvious that t if the proposed state is more likely than the old one, it is accepted with probability 1. If the proposed state is less likely than the current one, the probability of accepting depends on the likelihood ratio.

2) Idea of Gibbs sampling :

Suppose the parameter vector θ can be divided into d subvectors

- an iteration of the Gibbs sampler draws values of each subvector, conditional on the values of all the other subvectors, i.e. there are d steps in iteration t
- and P(θ)now are the joint posterior distribution

posterior distribution of each subvector, i.e.

$$p(\theta_j | \theta_{(-j)}^{t-1}, y)$$

with $\theta_{(-j)}^{t-1}$ all the components except for θ_j, at their current values:

$$\theta_{(-j)}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})^T$$

which are the iteration t values for those subvectors *already* updated and the iteration $(t-1)$ values for those subvectors not yet updated

3) their relation

Gibbs sampling is a special case of Metropolis-Hastings where the proposal q is based on the following two stage procedure.

- First, in an iteration of the Gibbs sampling procedure(there are d steps in one iter), single dimension i of θ is chosen randomly.
- Gibbs sampling performs a random walk where at each iteration the value along a randomly selected dimension is updated according to the conditional distribution

$p(\theta_j | \theta_{(-j)}^{t-1}, y)$ where $\theta_{(-j)}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1})^T$

the reason why the value is updated each iteration is that is

the acceptance rate

$$\frac{p(z')}{p(z^{(t)})}\frac{q(z^{(t)}|z')}{q(z'|z^{(t)})} = \frac{p(z'_i|z'_{-i})p(z'_{-i})}{p(z_i^{(t)}|z_{-i}^{(t)})p(z_{-i}^{(t)})}\frac{p(z_i^{(t)}|z'_{-i})}{p(z_i|z_{-i}^{(t)})} = \frac{p(z'_i|z'_{-i})p(z'_{-i})}{p(z_i^{(t)}|z'_{-i})p(z'_{-i})}\frac{p(z_i^{(t)}|z'_{-i})}{p(z_i|z'_{-i})} = 1$$

,which mean the acceptance rate is aways 1.

4. Perform Bayesian linear regression analysis for the dataset `stackloss` which is included in the base distribution of `R`. The response variable is `stack.loss`. Estimate a 95% HPD interval for each parameter. Clearly state your assumptions, and provide all `R` code. You can use whatever package you want, but you should briefly explain what assumptions (e.g. what type of priors and hyperparameters) are being used.

Ans:
R code:

```
source("http://bioconductor.org/biocLite.R")
biocLite(c("graph", "RBGL", "Rgraphviz"))
install.packages("gRain", dependencies=TRUE)
library(MCMCpack);
Nfit <- MCMCregress(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc. , data = stackloss,
          burnin = 1000, mcmc = 25000, thin = 25)

# summary for each parameter
> summary(Nfit)

Iterations = 1001:25976
Thinning interval = 25
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
  plus standard error of the mean:

            Mean      SD Naive SE Time-series SE
(Intercept) -40.5153 12.7952 0.404619      0.404619
Air.Flow      0.7131  0.1485 0.004697      0.004697
Water.Temp    1.3184  0.4048 0.012801      0.012801
Acid.Conc.   -0.1487  0.1681 0.005315      0.004979
sigma2       12.0119  4.7496 0.150194      0.150194

2. Quantiles for each variable:

            2.5%     25%     50%      75%    97.5%
(Intercept) -67.1977 -48.5110 -40.0959 -31.71670 -16.4090
Air.Flow     0.4226   0.6172   0.7076   0.81157   0.9961
Water.Temp   0.5438   1.0314   1.3260   1.57924   2.1097
```

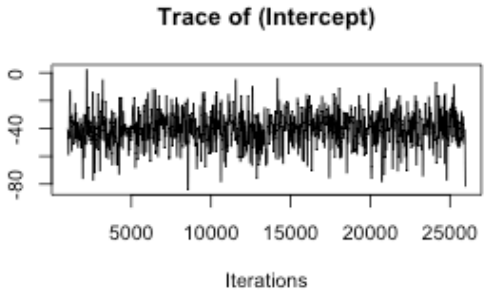
```
Acid.Conc.   -0.4757  -0.2557  -0.1570  -0.03933   0.1992
sigma2        5.9583   8.7963  11.0367  14.12567  22.8574
```

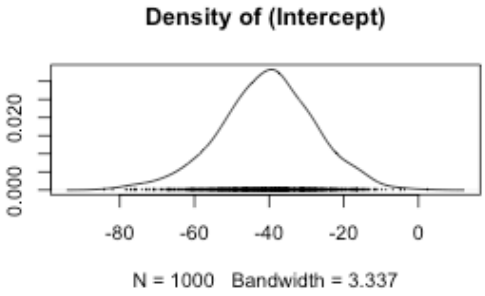
 # Estimate a 95% HPD interval for each parameter

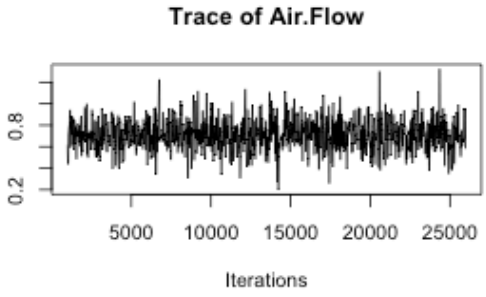
```
> intervals = HPDinterval(Nfit, prob = 0.95)
> intervals
                lower       upper
(Intercept) -67.1889713 -16.1749424
Air.Flow      0.4132705   0.9813900
Water.Temp    0.5891393   2.1316547
Acid.Conc.   -0.4880107   0.1816153
sigma2        5.4478441  21.2966820
attr(,"Probability")
[1] 0.95
```

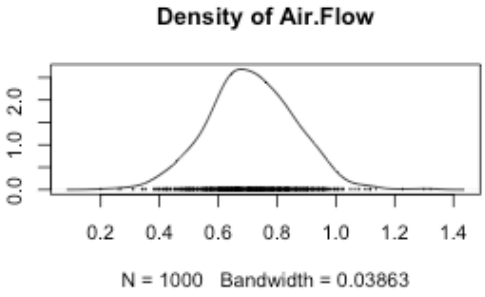
 # Plot each parameter

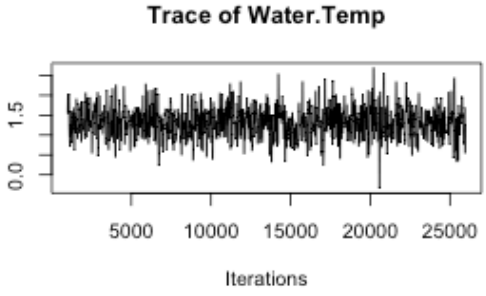
```
> plot(Nfit)
```

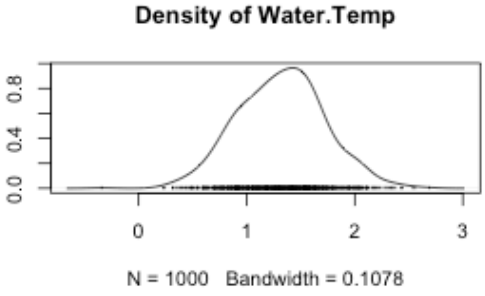




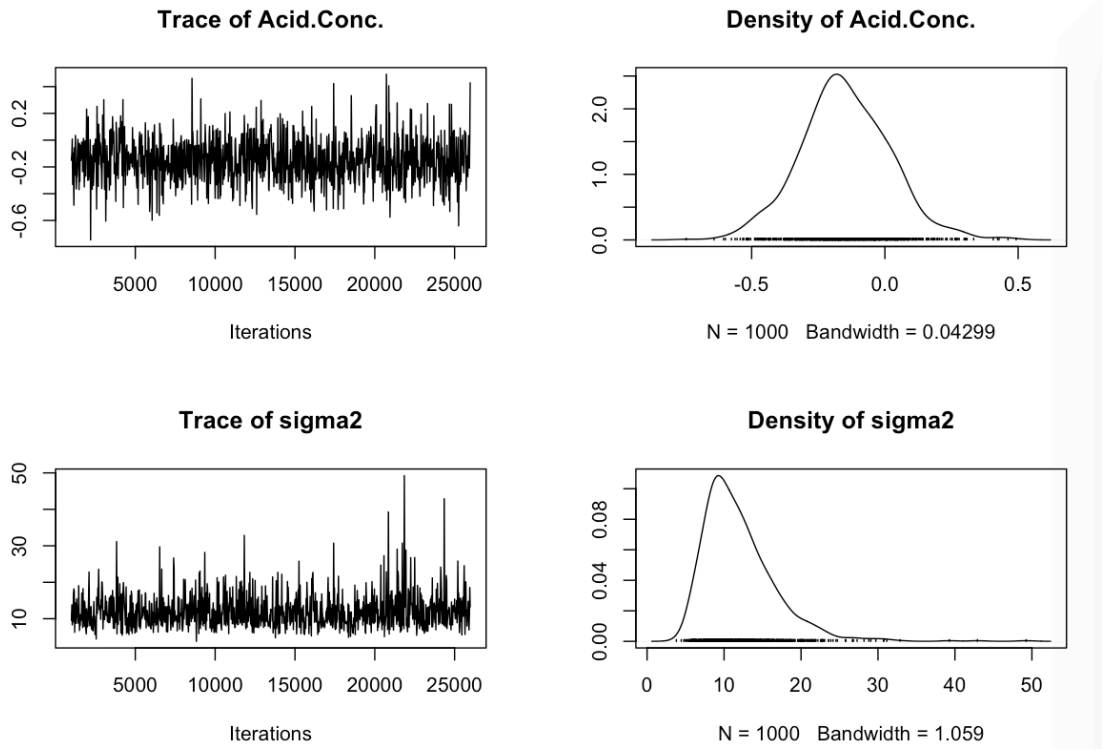








Math 459 Midterm 2
Qian Liu



Assumption:

As shown in class, If we know the full conditional distributions p(β|σ2, X, Y) and p(σ2|β, X, Y), we can sample from the joint posterior p(β, σ2|X, Y) using the Gibbs sampler. Simulates from posterior using Gibbs sampling.

- multivariate normal draw for β
- inverse gamma draw for conditional error variance $\sigma^{-2}|\beta$.

$$y_i = x_i'\beta + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\beta \sim \mathcal{N}(b_0, B_0^{-1}), \quad \sigma^{-2} \sim \text{Gamma}(\frac{c_0}{2}, \frac{d_0}{2})$$

and β, σ^{-2} assumed to be a priori independent.

and b0 prior mean for β; if scalar then all means the same.

default prior precision of β is B0 = 0; if scalar then it is value times identity matrix

c0/2: shape parameter for inverse gamma prior on σ^{-2}

d0/2: scale parameter for inverse gamma prior.