Gibbs sampling for Bayesian estimates of the normal mean and variance

Michael Rogers

June 6, 2013

Key Words: Bayesian estimation, Gibbs sampling, interval estimation, simulation

Abstract:

Estimation of the population mean and standard deviation of a normal distribution with unknown parameters can be achieved using the frequentist method, where only the sample data and knowledge of the expected distributions of the sample statistics are used; an uninformative Gibbs sampler, which incorporates an uninformative prior distribution into a simulation of two correlated Markov chains; and an informative Gibbs sampler, which uses an informative prior distribution in the correlated Markov chains simulations. The informative prior in the Gibbs sampler provides the most precise estimation of the unknown population mean and standard deviation. The uninformative Gibbs sampler yields only a slightly more precise parameter estimate over the frequentist estimate. The uninformative Gibbs sampler and frequentist methods produce parameter estimates that closely match each other though they use drastically different executions.

1-Introduction:

Gibbs sampling is a computational technique that uses the interrelated Markov chains of two unknown parameters to produce accurate estimates of both parameters. Consider the related parameters A and B. In Gibbs sampling, parameter A can be estimated by a Markov chain process that is dependent on the state of the parameter B, and likewise parameter B can be estimated by a Markov chain process dependent on the state of parameter A. A random seed is used in place of the parameter B in order to initiate the chain and produce the first estimate of parameter A. Next, the first estimate of parameter A is used to produce the next estimate for parameter B. By producing a large number of simulations of parameters A and B based on the current state of the other parameter, a steady state distribution is produced which can provide highly informative estimates of each parameter.

Prior information of the parameters can also be incorporated into a Gibbs sampler with the use of Bayes theorem to produce more precise distributions of the parameters using fewer simulations. In this paper, we will investigate the precision of the interval estimates provided by Gibbs sampling for the mean and standard deviation of a normal distribution when both the mean and standard deviation of the distribution are unknown. We will explore the effect of flat and informative prior distributions of the parameters on each interval estimate, and will compare the estimates with those achieved using frequentist methods.

2-Bayes Theorem for the mean and variance of a normal distribution

Bayes theorem incorporates prior knowledge of the distribution of a parameter with a set of sample data to produce more informative distributions and estimates of the parameter. Bayes theorem is given by the following equation:

Equation 2.1: $p(\theta|x) \propto p(\theta)\, p(x|\theta)$

Equation 2.1 indicates that the posterior distribution of parameter θ given the sample data *x* is proportional to the prior distribution of θ times the likelihood function of the sample data *x* based on the parameter θ.

In our estimations of the mean μ and variance σ^2 of an unknown normal distribution, we must first choose prior distributions for each parameter. The prior distribution for the mean μ should be in the form $p(\mu) \sim NORM(\mu_0,\sigma_0)$ where $\theta_0 = \sigma_0^2$; and the prior distribution for the variance should be in the form $p(\theta) \sim 1 / [GAMMA(\alpha_0, \kappa_0)]$ where $\theta = \sigma^2$. Next, we note that the sample data x are independently distributed as NORM(μ,σ) and are denoted as the following: $\mathbf{x} = (x_1, \ldots, x_n)$. This provides the likelihood functions for **x** in terms based on μ and θ:

$p(\mathbf{x}|\mu) \propto \exp[\ (-1/(2\ \sigma^2)\ \Sigma_{i=1}^{n}\ (x_i-\mu)^2\ ]$ and $p(\mathbf{x}|\theta) \propto \Pi_{i=1}^{n}\ [\theta^{-1/2}\ \exp(\ -x_i^2/(2\theta)\ )]$

Combining the likelihood functions with each parameter's prior distribution using Bayes theorem gives the following equations for the posterior distributions for μ and θ:

Equation 2.2: $\mu|\mathbf{x},\theta \sim NORM(\mu', sqrt(\theta'))$

$\theta' = 1/(n/\theta + 1/\theta_0)$ $\quad$ $\mu' = \theta'(n\bar{x}/\theta + \mu_0/\theta_0)$

Equation 2.3: $\theta|\mathbf{x},\mu \sim 1 / [GAMMA(\alpha',\kappa')]$

$\alpha' = \alpha^0 + n/2$ $\quad$ $\kappa' = \kappa_0 + [(n-1)s^2 + n(\bar{x}-\mu)^2]/2$

These posterior distributions can be incorporated into a Gibbs sampler so that alternating distributions of each parameter can be produced using the most recent distribution of the partnered parameter. In this way, the new distribution for μ will help create a new distribution for θ, and vice versa in a string of simulations.

3-Incorporating posterior distributions into the Gibbs sampler

A Gibbs sampler uses correlated Markov chains for two parameters, where the steady state distribution of the Markov chains reflects accurately the true distribution of the parameters. In the case of our normal mean and variance estimation the posterior distributions for μ and θ are used to model the Markov behavior where the new distribution of μ is only dependent on the current state of θ and the new distribution of θ is only dependent on the current state of μ. Our Gibbs sampler will set the prior distribution information, set the sample data from the n observations from the NORM(μ, σ) distribution, and then initialize an arbitrary estimate of θ. The initialized value for θ will be used in the posterior distribution equations to produce a new distribution for μ. A single random sample is taken from the new μ distribution and is stored, and then this μ distribution is used to produce a new distribution for θ using the posterior distribution equations. A single sample is then taken from the new θ distribution and stored. This process is repeated for each simulation where new distributions are found for μ and θ and single samples are taken from each and stored.

When simulating a large number of times from this Gibbs sampler, the first half of the observations are discarded. These observations represent the burn in period, where the initialized value of the sampler still affects the posterior distributions produced for the parameters and these Markov chains have not yet had enough time to reach their steady state distributions. After the burn in period has expired, sampled values from each parameter's posterior distributions are pooled into comprehensive distributions used to find point and interval estimates for each parameter. We will observe the effects of uninformative and informative prior distributions on the Gibbs sampler estimates.

4-Uninformative prior in Gibbs sampler

We will consider the following example to run our Gibbs sampler. The height of 41 men (n=41) is measured at the beginning and end of a day and the decrease in height (x_i) is recorded for each subject. We will assume that the distribution of height decreases follows the normal distribution with mean μ and variance $\sigma^2 = \theta$ both unknown. We will also suppose that the sample mean $\bar{x}$ of the 41 observations is 9.6 mm and the sample standard deviation s = 2.73 mm.

We first want to run our Gibbs sampler using an uninformative prior distribution. To do this, we let the prior distribution of the population mean of decreases in height μ follow the normal distribution NORM(μ_0,σ_0). We let μ_0 = 0mm and let σ_0 = 20mm, providing a very wide distribution for the mean relative to the sample data collected. We also let the prior distribution of the population variance $\sigma^2 = \theta$ be distributed as 1 / [GAMMA(α_0 ,κ_0)], where α_0 = 0.5 and κ_0 = 0.2. This prior distribution provides a 95% confidence interval for σ of 0.28 mm to 20.18 mm (sqrt(1/qgamma(c(.975,.025),.5,.2))). Again, this distribution provides a wide range of possible values relative to the sample data and is therefore uninformative about the correct value of the population variance.

We then incorporate our data and prior distributions into the Bayes posterior equations (Equations 2.2 and 2.3) using the Gibbs sampling method described in Section 3. The following code demonstrates the execution of our Gibbs sampler using 50,000 total simulations and where the first 25,000 simulations are discarded due to the sampler's burn-in period.

Program 4.1: Gibbs sampler for distribution with uninformative priors

```
> ## Gibbs sampling with uninformative priors
> set.seed(1237)                      # set randomizer seed
> m=50000                             # set num iterations
> MU=numeric(m);  THETA=numeric(m);   # sampled values
> THETA[1]=1;                         # inital value
> n=41; x.bar=9.6; x.var=2.73^2;      # data
> mu.0=0; th.0=400;                   # mu prors
> alp.0=1/2; kap.0=1/5;               # theta priors
>
> for (i in 2:m)
+ {
+   th.up=1/(n/THETA[i-1]+1/th.0)
+   mu.up=(n*x.bar/THETA[i-1]+mu.0/th.0)*th.up
+   MU[i]=rnorm(1,mu.up,sqrt(th.up))
+
+   alp.up=n/2+alp.0
+   kap.up=kap.0+((n-1)*x.var+n*(x.bar-MU[i])^2)/2
+   THETA[i]=1/rgamma(1,alp.up,kap.up)
```

```
+ }
>
> # Bayesian point and probability interval estimates
> aft.brn=(m/2+1):m                 # discard first half of simulations
> mean(MU[aft.brn])                 # point estimate mu
[1] 9.594313
> bi.MU=quantile(MU[aft.brn],c(.025,.975)); bi.MU
+                                   # 95% confidence interval mu
     2.5%     97.5%
 8.753027 10.452743
> mean(THETA[aft.brn])              # point estimate theta
[1] 7.646162
> bi.THETA=quantile(THETA[aft.brn],c(.025,.975)); bi.THETA
+                                   # 95% confidence interval theta
     2.5%     97.5%
 4.886708 11.810233
> SIGMA=sqrt(THETA)
> mean(SIGMA[aft.brn])              # point estimate of sigma
[1] 2.747485
> bi.SIGMA=sqrt(bi.THETA); bi.SIGMA # 95% confidence interval sigma
    2.5%    97.5%
2.210590 3.436602
```

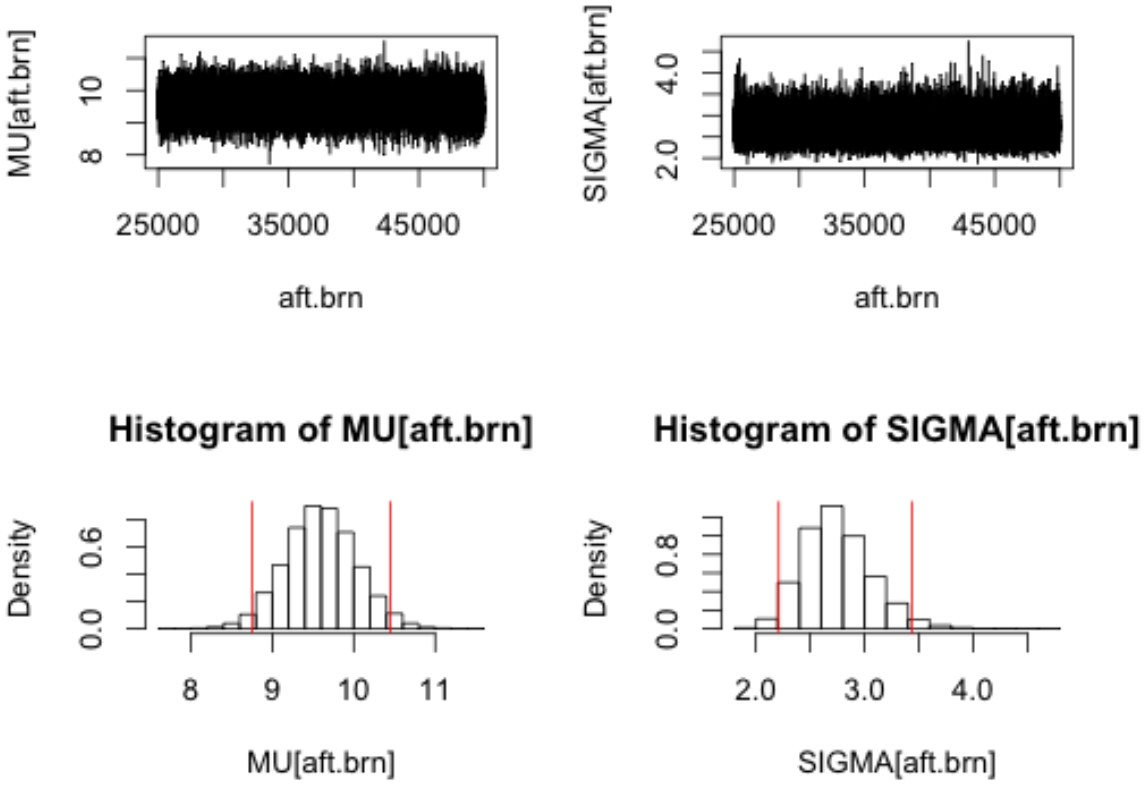
The Gibbs sampler with uninformative priors provides a point estimate for μ as 9.594 mm and a 95% confidence interval for μ as 8.753 mm to 10.453 mm; the Gibbs sampler gives the point estimate for σ as 2.747 mm and the 95% confidence interval for σ as 2.211 mm to 3.437 mm.

To ensure that the Gibbs sampler has indeed reached the steady state distribution of each Markov chain we observe the histories of the sequence of simulations of μ and σ as well as the histograms of the distributions for μ and σ for after the burn-in period. The histories for each parameter should show random simulations with constant variance throughout the sequence; no trending of the simulations should be detected through the progression of simulations. The histograms should reflect distributions closely following normal for μ and inverse gamma for σ. The following code produces these diagnostic graphs:

Program 4.2: Diagnostic graphs 1 for Gibbs sampler with uninformative priors

```
> par(mfrow=c(2,2))
>   plot(aft.brn,MU[aft.brn],type="l")
>   plot(aft.brn,SIGMA[aft.brn],type="l")
>   hist(MU[aft.brn],prob=T); abline(v=bi.MU, col="red")
>   hist(SIGMA[aft.brn],prob=T); abline(v=bi.SIGMA, col="red")
> par(mfrow=c(1,1))
```

Table 4.2: Diagnostic graphs 1 for Gibbs sampler with uninformative priors

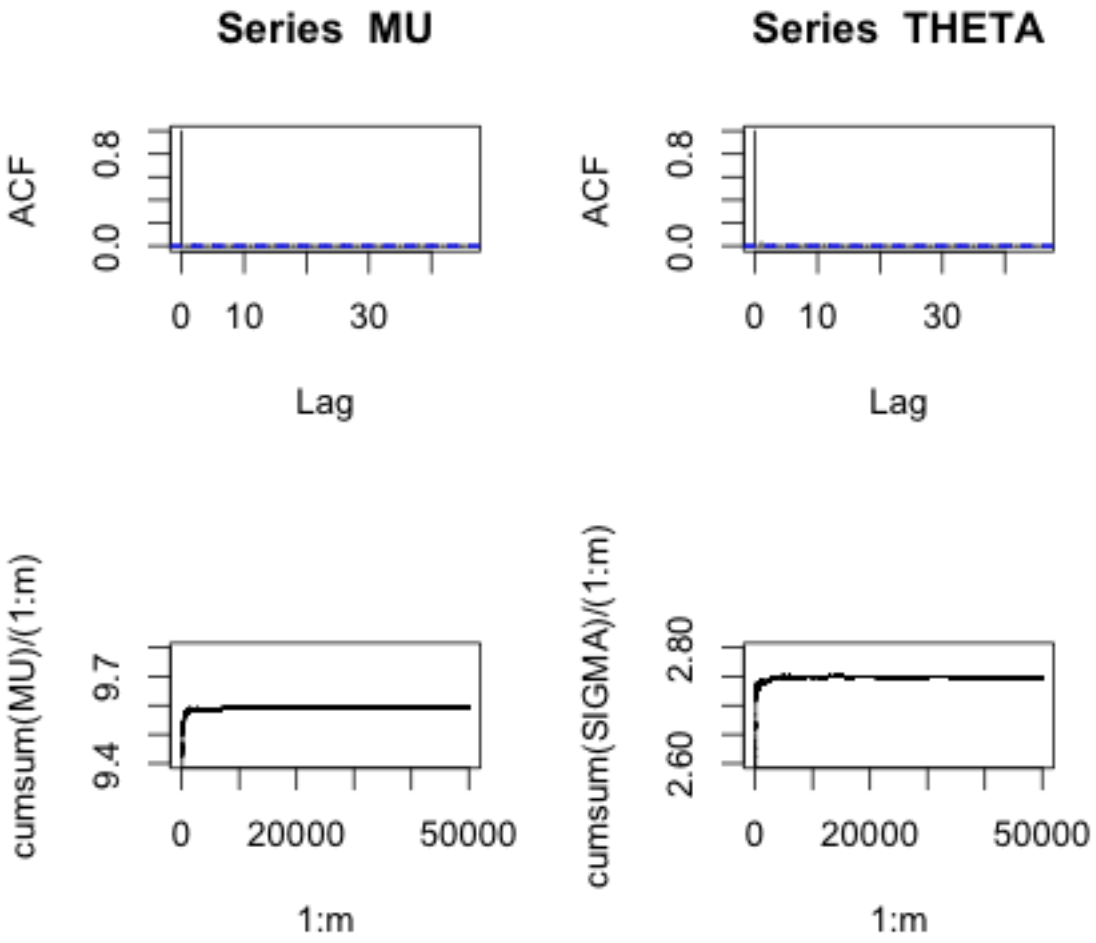


The top diagnostic charts do indeed show simulation histories with constant variance and without any trending over time, and the lower charts show show normal distribution for μ and an inverse gamma distribution for σ. We can also confirm the validity of the Gibbs sampler by observing the graphs of the auto-correlation function for each parameter's progression of simulations and of the cumulative averages of the sampled values.

Program 4.3: Diagnostic graphs 2 for Gibbs sampler with uninformative priors

```
> par(mfrow=c(2,2))
>   acf(MU); acf(THETA);
>   plot(1:m,cumsum(MU)/(1:m),type="l", ylim=c(9.4,9.8))
>   plot(1:m,cumsum(SIGMA)/(1:m),type="l",ylim=c(2.6,2.8))
> par(mfrow=c(1,1))
```

Table 4.3: Diagnostic graphs 2 for Gibbs sampler with uninformative priors



The auto correlation graphs on top indicate simulation quickly loses the influence of the initialized value and that progressive simulations are not correlated with samples before them. This behavior is expected for a Gibbs sampler that quickly brings each parameter to their steady state distribution. The cumulative average graphs below further corroborate this as the averages quickly settle to a single value.

5-Informative prior in Gibbs sampler

We next consider the Gibbs sampler from Section 4 using informative prior distributions. We again let the prior distribution of the population mean follow the normal distribution NORM(μ_0,σ_0) and the prior distribution of the population variance $\sigma 2 = \theta$ follow an inverse gamma distribution 1 / [GAMMA(α_0 ,κ_0)]. However, this time we let μ_0

= 10mm, σ_0 = 1mm, α_0 = 20 and κ_0 = 200. These prior distributions yield a point estimate for μ as 10mm and a 95% confidence interval for μ as 8.04mm to 11.96mm, and yield a point estimate for σ as 3.22mm and a 95% confidence interval for σ as 2.60mm to 4.05mm. Relative to the previous uninformative priors, these prior distributions provide much higher precision and are therefore much more informative. Furthermore, comparing the prior distributions to the sample data yields similar estimates of the values of the parameters, supporting the informative nature of these prior distributions. We rerun Program 4.1 using the updated program header below that includes the informative prior distributions for μ and θ:

Program 5.1: Gibbs sampler header for distribution with informative priors

```
> ## Gibbs sampling with informative priors
> set.seed(1237)                       # set randomizer seed
> m=50000                              # num iterations
> MU=numeric(m);  THETA=numeric(m);  # sampled values
> THETA[1]=1;                          # inital value
> n=41; x.bar=9.6; x.var=2.73^2;     # data
> mu.0=10; th.0=1;                     # mu priors
> alp.0=20; kap.0=200;                 # theta priors
>
> for (i in 2:m)
+ {
+   th.up=1/(n/THETA[i-1]+1/th.0)
+   mu.up=(n*x.bar/THETA[i-1]+mu.0/th.0)*th.up
+   MU[i]=rnorm(1,mu.up,sqrt(th.up))
+
+   alp.up=n/2+alp.0
+   kap.up=kap.0+((n-1)*x.var+n*(x.bar-MU[i])^2)/2
+   THETA[i]=1/rgamma(1,alp.up,kap.up)
+ }
>
> # Bayesian point and probability interval estimates
> aft.brn=(m/2+1):m                  # discard first half of simulations
> mean(MU[aft.brn])                  # point estimate mu
[1] 9.669655
> bi.MU=quantile(MU[aft.brn],c(.025,.975)); bi.MU
+                                    # 95% confidence interval mu
     2.5%     97.5%
 8.847174 10.507079
> mean(THETA[aft.brn])               # point estimate theta
[1] 8.928182
> bi.THETA=quantile(THETA[aft.brn],c(.025,.975)); bi.THETA
+                                    # 95% confidence interval theta
     2.5%     97.5%
 6.499256 12.163465
> SIGMA=sqrt(THETA)
> mean(SIGMA[aft.brn])                # point estimate of sigma
[1] 2.97847
> bi.SIGMA=sqrt(bi.THETA); bi.SIGMA  # 95% confidence interval sigma
    2.5%    97.5%
2.549364 3.487616
```

The Gibbs sampler with informative prior distributions provides a point estimate for μ as 9.670 mm and a 95% confidence interval for μ as 8.847 mm to 10.507 mm; the Gibbs sampler gives the point estimate for σ as 2.978 mm and the 95% interval for σ as 2.549 mm to 3.488 mm. The informative prior distribution does yield more precise estimates for μ and σ; however, the huge leap in additional information provided by these prior distributions over the uninformative priors does not produce significantly more precise estimates. The precision gains achieved by incorporating informative prior distributions appear tangible but modest.

In order to ensure that the Gibbs sampler with informative priors is also producing reliable estimates, we observe the diagnostic graphs. The history charts for μ and σ simulations again show sequences of random simulations with constant variance and lack of trending. The histograms indicate μ is indeed following a normal distribution and σ matches the inverse gamma distribution. In Table 5.3, the auto correlation function charts and cumulative average charts show how the succession of simulations quickly loses correlation from previous values and aligns itself to a set value in each parameter's steady state distribution. These graphs indicate that the informative prior distribution also provides good estimates for the parameters.

Table 5.2: Diagnostic graphs 1 for Gibbs sampler with informative priors

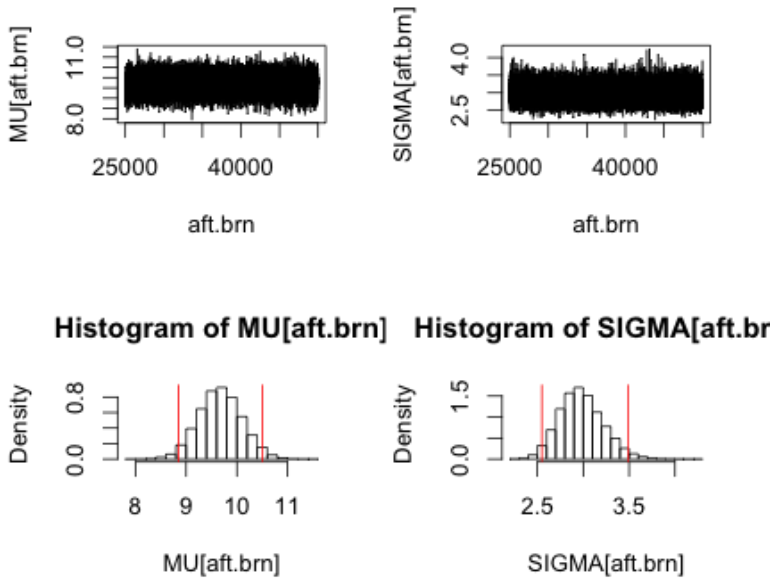
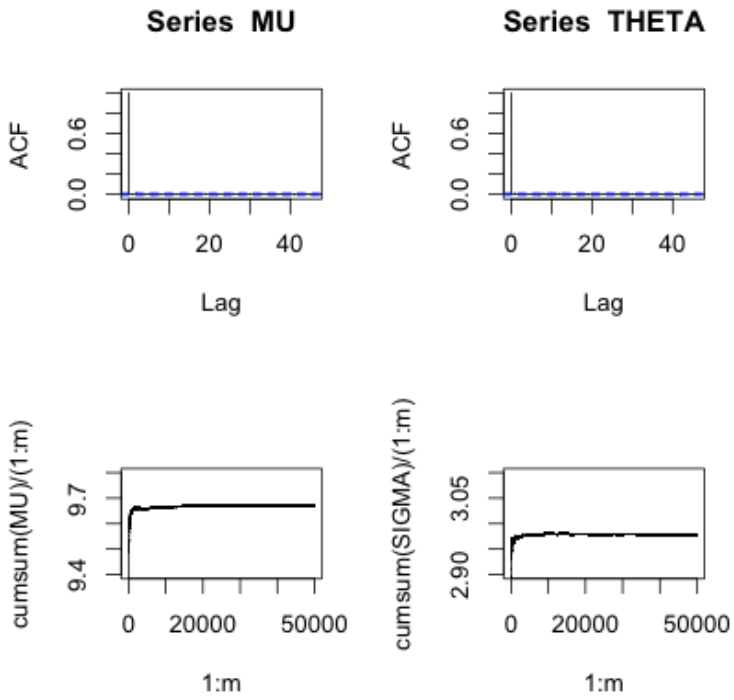


Table 5.3: Diagnostic graphs 2 for Gibbs sampler with informative priors



6-Frequentist method for parameter estimation

Gibbs sampling incorporates Bayes theorem, which allows prior knowledge of the distribution of unknown parameters to be taken into account to produce highly precise estimates of those parameters. Frequentist methods on the other hand rely solely on data to produce estimates of unknown parameters and neglect prior information of each parameter. In our sample case, we strive to produce estimates for the population mean μ and standard deviation σ of an unknown normal distribution when a sample of n=41 yields sample mean $\bar{x}$ of 9.6 mm and sample standard deviation s = 2.73 mm. Frequentist estimates for μ and σ rely only on the theoretical distribution of the sample mean and standard deviation for cases where each parameter is unknown:

Equation 6.1: $\left({\bar{x}-\mu}/{s\sqrt{n}}\right) \sim t(n-1)$

Equation 6.2: $\left({(n-1)s^2}/{\sigma^2}\right) \sim \chi^2(n-1)$

For frequentist methods we use the sample mean $\bar{x}$ as a point estimate for μ and the sample standard deviation s as the point estimate for σ. The distributions noted in

Equations 6.1 and 6.2 lead to the following (1-α)x100% confidence intervals for μ and σ respectively:

Equation 6.3: $\left[\bar{x} - t_{\frac{\alpha}{2}}(n-1)\left(\frac{s}{\sqrt{n}}\right), \bar{x} + t_{\frac{\alpha}{2}}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$

Equation 6.4: $\left[\sqrt{(n-1)s \Big/ \chi^2_{\frac{\alpha}{2}}(n-1)}, \sqrt{(n-1)s \Big/ \chi^2_{1-\frac{\alpha}{2}}(n-1)}\right]$

Incorporating the sample data with Equations 6.3 and 6.4 gives the frequentist point estimate of μ as 9.6mm, the 95% frequentist confidence interval for μ as 8.738mm to 10.462mm, and the frequentist point estimate of σ as 2.73mm, and the 95% frequentist confidence interval for σ as 2.241mm to 3.493mm. Tables 6.1 and 6.2 below compare the point and interval estimates obtained for μ and σ:

Table 6.1: Point and interval estimates for population mean

Method	**Point Estimate**	**95% Confidence Interval**	
Frequentist	9.6	8.738	10.462
Gibbs (uninformative prior)	9.594	8.753	10.453
Gibbs (informative prior)	9.67	8.847	10.507

Table 6.2: Point and interval estimates for population standard deviation

Method	**Point Estimate**	**95% Confidence Interval**	
Frequentist	2.73	2.241	3.493
Gibbs (uninformative prior)	2.747	2.211	3.437
Gibbs (informative prior)	2.978	2.549	3.488

These tables indicate that the Gibbs sampler with an informative prior distribution provides the most precise estimation of the population parameters. The Gibbs sampler with uninformative priors performed second best in estimate precision followed by the frequentist method. We should note however, that all three methods produced very similar estimates, with the uninformative prior Gibbs sampler closely matching the frequentist estimates. The gains in precision of the informative prior Gibbs sampler are

mild. The informative prior Gibbs sampler would likely achieve much higher precision over the frequentist and uninformative methods if the sample size were smaller. The smaller sample size would cause the frequentist's t and chi-squared distributions to expand and widen the parameter estimates. However, for this normal parameter estimation scenario each of the methods gives similar estimates with modest gains being achieved by adding more information to the Gibbs sampler.

Bibliography:

Hogg, Robert and Tanis, Elliot. *Probability and Statistical Inference, 8th Edition,* pgs. 280-308. Prentice Hall 2010.

Sincich, Levine, and Stephan. *Practical Statistics by Example, 2nd Edition,* pgs. 338-389. Prentice Hall, 2002.

Trumbo, Bruce and Suess, Eric. *Introduction to Probability Simulation and Gibbs Sampling with R*, pgs. 195-248. Springer 2010.

Trumbo, Bruce and Suess, Eric. "Gibbs Sampling", *Encyclopedia of methods.* 2013.