J. Appl. Environ. Biol. Sci., 6(2S)130-136, 2016

© 2016, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Markov Chain Monte Carlo Convergence Diagnostics For Gumbel Model

Nor Azrita Mohd Amin[1] And Mohd Bakri Adam[2]

[1]Institute of Engineering Mathematics, Universiti Malaysia Perlis
[2]Institute of Mathematical Research, Universiti Putra Malaysia

Received: January7, 2016
Accepted: March 2, 2016

ABSTRACT

Markov chain Monte Carlo (MCMC) has been widely used in Bayesian analysis for the analysis of complex statistical models. However, there are some isues on determining the convergence of this technique. It is difficult to determine the length of draws to make sure that the sample values converge to the stationary distribution and the number of n iterations should be discarded before the chain converge to the stationary distribution. Convergence diagnostics help to decide whether the chain converges during a particular sample run. Gelman and Rubin diagnostic is the most widely used method for convergence test. The MCMC technique, Metropolis-Hastings algorithm is used for posterior inferences of Gumbel distribution simulated data.

KEYWORDS: Markov chain Monte Carlo; convergence diagnostics; Metropolis-Hastings algorithm; Gumbel distribution.

1. INTRODUCTION

The explosion of interest in Bayesian methods over the last decades has been the result of the convergence of modern computing advances and the efficiency of Markov chain Monte Carlo (MCMC) algorithms for sampling from the posterior distribution (Carlin, 2011). The idea of MCMC is to construct a chain whose stationary distribution or the target distribution is the distribution from which you wish to sample and to run this chain for long enough that it has forgotten its starting state and is close to its stationary distribution. Research on the MCMC techniques are widely applied in various area but very little of practical use of computing the convergence. MCMC techniques does not give a clear indication whether the iterations have been converged. The fundamental theory of MCMC only guarantees that the distribution of the output will converge to the posterior distribution as the number of iterations increases to infinity. However, it is not guaranteed that the chain will converged after *N* iterations. Generally, MCMC samples is separated into two parts. The first part of the chains is a "burn in" period for which the samples are discarded and the rest of the iterations are considered to converge to the target distribution. The questions are, how long to discard the chains and how long the iterations are required to accurately estimate posterior quantities. Hence, convergence diagnostics assess the output from the MCMC sampling and analyze whether the draws have approximate the posterior distribution.

This works deal with the MCMC sampling for the inferences of Gumbel distribution (Coles, 2001) simulated data. Gumbel distribution is often used to model the distribution of the maximum or minimum level of a process which is practical and relevant for predicting the future extreme events such as flood, earthquake, air and water pollutions or other natural disaster. Gumbel distribution is a special case of the extreme value distribution in which the shape paremeter, ξ is equal to zero. There are two parameters in Gumbel distribution that are location, μ and scale, σ parameters. The distribution function for Gumbel distribution is given by Equation (1).

$$G(z)=\exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\};\quad -\infty<\mu<\infty,\sigma>0\,. \tag{1}$$

Inversion method is applied for simulating a series of data from Gumbel distribution. The MCMC techniques, Metropolis-Hastings (MH) method is developed to estimate the parameters. Some issues in implementing the algorithm are discussed and focus is on the convergence diagnostics based on Gelman and Rubin method.

* **Corresponding Author:** NOR AZRITA MOHD AMIN, Institute of Engineering Mathematics, Universiti Malaysia Perlis
norazrita@unimap.edu.my

2. **Markov chain Monte Carlo.** MCMC has been used to draw the random samples from complex, nonstandard distribution. If a process is producing points and the future value is independent from the past, then the sequence of values produced is called Markov chain. Monte Carlo in simple word refers to simulation technique of repeated sampling. Bayesian Markov chain has gained its popularity in statistical analysis for the inferences of posterior distributions as well as to predict the future probability. MCMC methods offer a great statistical tool and have been explored in diverse area. The execution of these approaches requires deep understanding and skills. Chib and Greenberg (1995) and Gamerman and Lopes (2006) provide a comprehensive preliminary details as well as an intensive development and applications of MCMC. Basically the idea of Bayesian MCMC arises when the target distribution, say $\pi(x)$ is complex such that it is difficult to sample from it directly. A series of samples generated from $\pi(x)$ will construct an aperiodic and irreducible Markov chain stationary to $\pi(x)$. The simulated values from the long enough chains can be treated as a dependent samples from the target distribution and use as a basis for summarizing $\pi(x)$ (Brooks, 1998).

Currently there are a wide variety of MCMC algorithms developed and practiced. But it is important to understand that each idea has its own distinct advantages and drawbacks between one another. MH method (Metropolis et al. ,1953, Hastings, 1970) and Gibbs sampling method (Geman, 1984) are very famous and most practical MCMC techniques. MH is the fundamental algorithm for many MCMC approaches while the Gibbs sampling is a good alternative if the full conditional distributions for each parameter are known. The Gibbs sampler is facing the difficulty to deal with the required conditional distributions. If the posterior doesn't look like any distribution or having no conjugacy, then MH is the most suitable method to use for generating random samples from a target distribution $\pi(x)$ for which direct sampling is cumbersome.

2.1 Metropolis-Hastings Algorithm. MH routine is capable to simulate a series from an arbitrary density as a basis for summarizing features of the equilibrium distribution which is a Bayesian posterior distribution for an unknown parameter θ (Smith, 1993). MH algorithm constructs a Markov chain by proposing a candidate value for θ from a proposal distribution, $q(.,\theta)$. The derivation of MH algorithm is discuss in Chib (1995) by exploiting the notion of reversibility. In general, the MH algorithm in algorithmic form can be summarize as follows.

1. Set the initial values for (θ_1^0, θ_2^0)
2. Given that the chain is currently at (θ_1^j, θ_2^j) :

Draw a candidate value $\theta^{can} \sim (\mu^j, \upsilon_\theta)$ for some suitably chosen variance υ_θ and take

$$\theta^{(j+1)} = \begin{matrix} \theta^{can} & \text{with probability } p \\ \theta^{(j)} & \text{with probability } 1 - p \end{matrix}$$

where p is the acceptance probability. This is implemented by drawing $u \sim Uniform(0,1)$ and taking $\theta^{(j+1)} = \theta^{(can)}$ if and only if $u < p$.

$$p = \frac{\pi(\theta^{can} \mid x)}{\pi(\theta \mid x)}$$

where $\pi(\theta \mid x)$ is the conditional posterior distribution for θ .

3. Iterate the updating procedure.

The variance of the candidate value, υ_θ is typically chosen by trial and error and aiming at an acceptance probability roughly around 20 % to 50%. There are some issues on simulation draws in Bayesian inference studies. First, the issue on how to decide whether the Markov chain has reached the stationary distribution. The draws in any MCMC method are regarded as a sample from the target density $\pi(x)$ only after the chains has passed the transient stage and the effect of the fixed starting value has become so small that it can be ignored. The second issue is to determine the number of iterations to keep after the Markov

chain has reached stationarity. Therefore, the questions arise on how large the initial sample should be discarded and how long the sampling should be run.

2. **MCMC Convergence Diagnostics Test**. Several convergence diagnostic tools are develop to validate a necessary but not sufficient condition for convergence. There are no conclusive tests that can tell exactly when the Markov chain has converged to its stationary distribution. Almost all of the applied work involving MCMC methods rely on applying the diagnostics tools to the output produced by the MCMC algorithms (Cowles, 1996). It is important to check the convergence of all the parameters before make any statisctial inferences. Certain parameters can appear to have a very good convergence behavior, but that could be misleading due to the slow convergence of other parameters. If some of the parameters have bad mixing, it is not possible to get accurate posterior inference for parameters that appear to have good mixing. More details discussions in Cowles and Carlin (1996) and Brooks and Roberts (1998). Several tests are available to determine if the chain appears to be converged.

2.1. **Visual Analysis via Trace plot and Density Plot**. The trace plot shows the iterations versus the draws of the parameters. The trace informs whether the chain has converged to the stationary distribution or not and gives an idea of the burn-in periods that should be discarded.

2.1. **Gelman and Rubin Diagnostic**. Gelman and Rubin diagnostics is introduce in Gelman and Rubin (1992) to compare several chains drawn from different starting points and checking that they are indistinguishable. This approach is based on the analysis of variance. Approximate convergence is diagnosed when the variance between the different chain, *B* is no larger than the variance within each individual chain, *W*. There are several steps required in Gelman and Rubin diagnostic. For each parameters, run $m \geq 2$ chains of length 2*n* from over dispersed starting values. Then discard the first *n* draws in each chain and calculate the within-chain variance and between-chain variance. Next step is to calculate the estimated variance $\hat{Var}(\theta)$ of the parameter as a weighted average of the within-chain and between-chain variance.

The convergence of the Markov chain can be monitored by estimating the factor by which the conservative estimate of the distribution of θ might be reduced. The ratio between the estimated upper and lower bounds for the standard deviation of θ, which is called the estimated potential scale reduction or shrink factor, $\hat{R}$. As the simulation converges, the shrink factor declines to 1, meaning that the parallel Markov chains are essentially overlapping. If the shrink factor is high, then one should proceed with further simulations.

Within chain variance, *W* is defined as in Equation (2)

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 , \qquad (2)$$

where

$$s_j^2 = \frac{1}{n-1}\sum \left(\theta_{ij} - \bar{\theta}_j\right)^2 .$$

s_j^2 is the variance of the j^{th} chain and *W* is the mean of the variance of each chain. For any finite *n*, the within chain variance *W* should underestimate the true variance of the stationary distribution since the chains have probably not reached all the points of the stationary distribution, and will have less variability. Then, between chain cariance, *B* is given as in Equation (3)

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_j - \bar{\bar{\theta}}\right)^2 , \qquad (3)$$

where

$$\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j .$$

The variance of the chain means is multiplied by n since each chain is based on n draws. Subsequently, the estimate variance of the stationary distribution is defined as a weighted average of W and B as in Equation (4),

$$\hat{Var}(\theta) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B \,. \qquad (4)$$

Because of overdispersion of the starting values, this overestimates the true variance, but is unbiased if the starting distribution equals the stationary distribution or the starting values were not overdispersed. Finally, the potential scale reduction factor is given by Equation (5),

$$\hat{R} = \sqrt{\frac{\hat{Var}(\theta)}{W}} \,. \qquad (5)$$

When $\hat{R}$ is high, perhaps greater than 1.1 or 1.2 then the chains should be run more longer to improve convergence to the stationary distribution.

3. RESULTS AND DISCUSSIONS

The simulation data are implement using Inversion method as developed in R programming in Appendix A. 1000 samples are generated from Gumbel distribution with parameters, $(\mu,\sigma) = (100,10)$. MH algorithm is developed to estimate the parameter of interest. The basic programming code of MH for Gumbel distribution is given in Appendix B. Figure 1 shows the trace plot of the MH draws for the inferences of Gumbel simulated data. It is clearly shows that at least the first 200 iterations have to be discard. The posterior mean and standard deviation for both parameters are $\mu = 99.06(0.31)$ and $\sigma = 9.583(0.22)$.

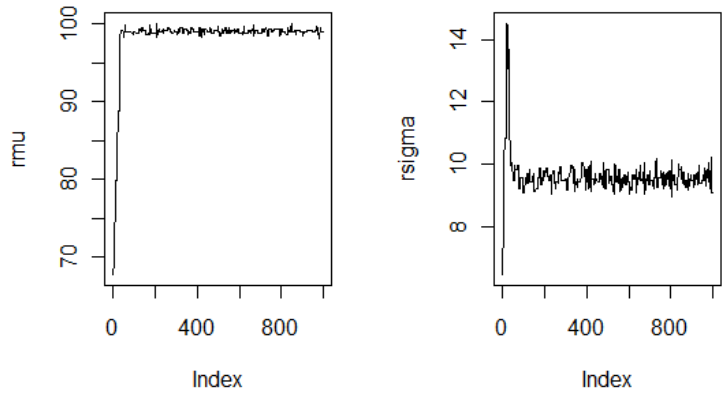

Figure 1. Trace plots for the location and scale parameters of Gumbel simulated data.

Then we proceed to the convergence diagnostics based on Gelman and Rubin test. Let assume the initial values for location parameter as Normal(mu0,1/kappa0) and for scale parameter as Gamma(alpha0,lambda0). Using the number of chains, m=5 with different initial values, let say set A as the following,

rwmetrop1<-function(data,mu0=70,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1),
rwmetrop2<-function(data,mu0=110,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1),
rwmetrop3<-function(data,mu0=100,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1),
rwmetrop4<-function(data,mu0=90,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1),
rwmetrop5<-function(data,mu0=120,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1).

Then, the associate potential scale reduction factors, $\hat{R}$ for each parameter are given as in Table 1. This result is based on 200 burn-in periods for 1000 and 3000 iterations. It is clear that the $\hat{R}$ value for location parameter is very far from 1.

Parameter	Point estimates	Upper confidence interval
1000 iterations; 200 burn-in periods		
μ	2.15	3.3
σ	1.17	1.4
3000 iterations; 200 burn-in periods		
μ	2.20	3.36
σ	1.13	1.31

Table 1: Potential scale reduction factors, $\hat{R}$ using the initial values, set A with 1000 and 3000 iterations.

Changing the initial value, set B, for each iteration and increase the number of burn-in periods and the number of iterations gives the following result as in Table 2.

```
rwmetrop1<-function(data,mu0=95,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1)
rwmetrop2<-function(data,mu0=97,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1)
rwmetrop3<-function(data,mu0=100,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1)
rwmetrop4<-function(data,mu0=103,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1)
rwmetrop5<-function(data,mu0=105,kappa0=0.25,alpha0=1,lambda0=1, vmu=4,vsigma=1)
```

Parameter	Point estimates	Upper confidence interval
10000 iterations; 1000 burn-in periods		
μ	1.07	1.19
σ	1.01	1.03
15000 iterations; 1500 burn-in periods		
μ	1.07	1.18
σ	1.01	1.02
20000 iterations; 5000 burn-in period		
μ	1.06	1.17
σ	1.01	1.02

Table 2: Potential scale reduction factors, $\hat{R}$ using the initial values, set B with 10000, 15000 and 20000 iterations.

There are only small differences results for different length of iterations, and all the $\hat{R}$ values can be acceptable. This diagnostics confirmed that the draws are converge to the stationary distribution at least after 10000 iterations and 1000 burn-in periods. However the factors are smaller for 20000 iterations and 5000 burn-in periods with the associates shrink factor plot as given in Figure 2. Finally, the better trace and density plots for both parameter are visualize in Figure 3.

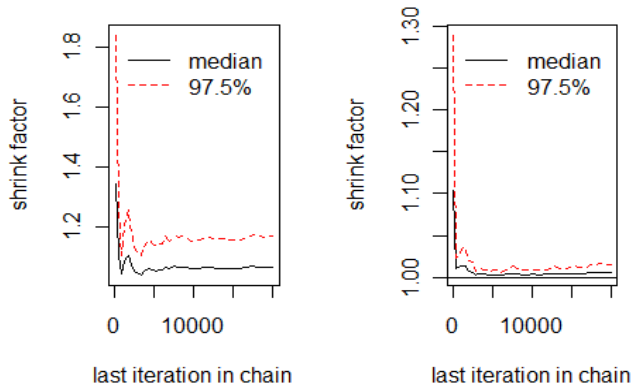

Figure 2. Shrink factor plot for 20000 iterations.

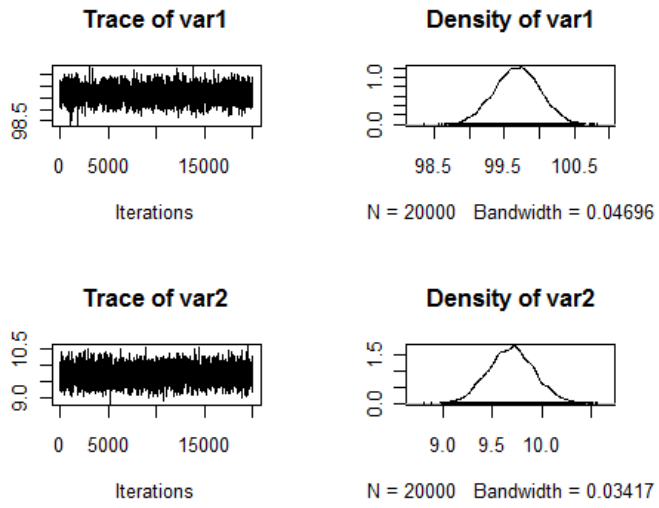

Figure3. Trace and density plot for location and scale parameters based on 20000 iterations.

3. **Conclusion**

Gelman-Rubin measures whether there is a significant difference between the variance within several chains and the variance between several chains by the potential scale reduction factors. It is obviously that this method needs at least two chains with different starting values. A factor, $\hat{R}$ equal to 1 means that between variance and within chain variance are equal while the larger values of $\hat{R}$ mean that there is still a notable difference between chains. The $\hat{R}$ value below 1.2 is consider acceptable. The shrink factor plots show the development of the scale-reduction over the iterations. The shrink factor plot in this study describe the stable reduction of the chains. In any real analysis, the bias that arises from the starting point of the chain (burn-in) should be discarded. For the inferences on the simulated Gumbel data, 20000 iterations and 500 burn-in periods give the best estimates while longer iterations need much computational time.

REFERENCES

[1] Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, *47*(1), 69-100.

[2] Brooks, S. P., & Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, *8*(4), 319-335.

[3] Carlin, B. P., & Louis, T. A. (2011). *Bayesian methods for data analysis*. CRC Press.

[4] Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, *49*(4), 327-335.

[5] Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). London: Springer.

[6] Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.

[7] Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

[8] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

[9] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.

[10] Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R news*, *6*(1), 7-11.

APPENDIX

Appendix A: Gumbel distribution simulated data

```
dsim<-function(n,inv.df){
inv.df<-function(x,mu=100,sigma=10)mu-sigma*log(-log(x))
u<-runif(n);y<-inv.df(u)
}
data<-dsim(100)
```

Appendix B: MH algorithm for Gumbel distribution

```
rwmetrop<-function(data,mu0=70,kappa0=0.25,alpha0=1,lambda0=1,nburn=0,ndraw=1000,vmu=4,vsigma=1
)
{
n<-length(data)
stdvmu<-sqrt(vmu);stdvsigma<-sqrt(vsigma)
mu<-rnorm(1,mu0,sqrt(1/kappa0)); sigma<-rgamma(1,alpha0,lambda0)

draws<-matrix(ncol=2,nrow=ndraw)
acceptmu<-0;acceptsigma<-0
it<- -nburn
while(it < ndraw){ it <- it+1;
⋮
if(it>0){;draws[it,1]<-mu;draws[it,2]<-sigma}
}
return(draws)
}
draw<-rwmetrop(data,nburn=0,ndraw=30000)
```