Estimating Variance Using MCMC When Sample Size Is One

a_statistician

March 31, 2017

Here is an exercise for a statistics class: Suppose that $Y_1, Y_2, ..., Y_n$ follow a normal distribution with unknown mean μ and unknown variance σ^2, where n is sample size. Please estimate the variance σ^2 when $n = 1$. I think most students will say that it is impossible to estimate the variance when sample size size is 1 and the mean μ is unknown.

This was correct before Markov Chain Monte Carlo (MCMC) algorithm was invented. Currently, it is easy to resolve the problem proposed in the exercise by using MCMC algorithm. Suppose $Y_1 = 4$. By running the following SAS program, you can get the estimate of the variance.

```
data mcmc;   y=4; run;
proc mcmc data=mcmc ntu=1000 nmc=20000 seed=246810;
   parms mu 0 sigma2 1;
   prior mu ~ normal(1, var=100000);
   prior sigma2 ~ uniform(0, 1000);
   model y ~ normal(mu,var=sigma2);
run;
```

Of course, other software, such as WinBUGS and R, with MCMC algorithm can also be used to get the answer. Here is the R code:

```
library(MCMCglmm)
mcmc=data.frame(y=4)
prior <- list(R = list(V=1, nu = 1), B=list(mu=0, V=1e+08))
model1 <- MCMCglmm(Y~1, random = NULL, data=mcmc, verbose=False,
             nitt=13000, burnin=3000, family = "gausian", prior=prior)
summary(model1)
```

For WinBUGS, specify the model by:

```
model{   y ~ dnorm(mu,invsigma)
mu ~dnorm(0,0.0000001)
inv ~ dnorm(0, 0.0001)
invsigma <- inv*inv}
```

and provide the data to MCMC by

```
list(x= 4)
```

Let's see how the MCMC algorithm get the estimate from the inestimable situation and evaluate its reliability of this estimate.

The MCMC algorithm was developed from Bayesian statistics which requires the prior distribution of the parameters. The Bayesian statistics combines the information from old prior distribution and the new data to get the posterior distribution of the parameters, and the estimate of the parameters is generated from the posterior distribution. For the MCMC algorithm, the prior distribution of all parameters is needed to start the algorithm. Some software of MCMC may not ask users to provide the prior, but the software itself provides the default prior.

In our example, the data do not provide any information to MCMC algorithm, but the prior distribution of σ^2 does. Therefore, the MCMC algorithm generates the posterior distribution according to the provided prior distribution without any modification because of the lack of the new information from the data. It means the estimate generated from MCMC algorithm is the same as that contained in the provided prior distribution.

You may think that we can provide the non-informative prior distribution, such that MCMC algorithm will not generate the absurd estimate. The fact is that real non-informative distribution does not exist, and non-informative prior distribution is a misnomer (I may discuss this issue later). In fact, we give the MCMC algorithm non-informative prior distribution that contains information.

So the validation of this kind of estimate totally depends on the reliability of prior distribution of the parameter. Generally, we do not have any information about parameters so the "non-informative prior distribution" is used. Under this situation the validation of this kind of estimate is null.

The purpose of this example is to give a warning for using MCMC algorithm or Bayesian estimate. Apart from checking the validation of the model and the properties of posterior distribution as described in many articles, it is also important to check if the data have the enough information to support the estimate of the parameters. The MCMC algorithm will give you the results given the syntax of language follows the specifications of software, regardless of the estimability of the parameters.

Two approaches can be used to check if the data support the estimate of parameters in the model. (1) Fit the model with traditional algorithm provided by the software. If traditional method fails and the MCMC algorithm produces the results, the estimate from MCMC may be problematic. (2) If the model is too complicated such that traditional methods do not work, try to specify the several prior distributions with big differences. If the results from MCMC algorithm basically do not depend on the prior distribution, the results may be reliable. If the estimate from MCMC algorithm heavily depends on the prior distribution, you cannot use this estimate. Obliviously, these approaches cannot detect all of the problems, but at least they can provide some help.

Previous Topics:

Interaction of Covariates in Statistics
Scope of the Appliaction for Survival Analysis
Log Transformation in Simple Linear Regression (3/3)
Log Transformation in Simple Linear Regression (2/3)
Log Transformation in Simple Linear Regression (1/3)