COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2021.1967985

A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario

Darshika Karunarasan, Roshini Sooriyarachchi, and Vimukthini Pinto

Department of Statistics, University of Colombo, Colombo, Sri Lanka

ABSTRACT

Multilevel modeling is a modern approach to deal with hierarchical or a nested data structure which can assess the variability between clusters. Bayesian Markov Chain Monte Carlo (MCMC) methods of estimations are advanced methods applicable for estimating multilevel models. However, these estimation methods are not as yet tested to identify its' performances as well as the properties associated with these estimation methods. This study targets to conduct a comparison of Bayesian MCMC methods which are developed for multilevel models where the response is normally distributed. The comparison is based upon extensive simulations and an application to a real-life dataset. The performance of Gibbs sampling (GS) and Metropolis Hastings (MH) methods are compared using a simulation study and additionally the factors which can affect the performance of both MCMC methods are identified. Practicality of these methods in real world scenario is confirmed through the application of MCMC method to a dataset. In the simulations though the Metropolis Hastings (MH) shows slightly better performance than Gibbs, there is no evidence to indicate that significant differences exist between these methods except for small samples where MH is superior. The results from the example are not as clear as from the simulations.

ARTICLE HISTORY
Received 22 May 2020
Accepted 7 August 2021

KEYWORDS
Estimation techniques; Goodness-of-fit; Markov Chain Monte Carlo; Multilevel modeling; Metropolis–Hastings; Gibbs Sampling

1. Introduction

1.1. Background of the study

1.1.1. Multilevel modeling

It is not always possible to obtain data where the observations are independent of each other. The reason for this dependency might be some inherent clusters within the data. Many types of data, including observational data collected in the human and biological sciences have a hierarchical, nested, or clustered structure. For example, animal and human studies of inheritance deal with a natural hierarchy where offspring are grouped within families, corporations are nested within nations, students are grouped within schools in education studies etc. There should be at least two levels of data within the multilevel structure and this study is based upon the two-level multilevel structure.

Even though there are several methods available to deal with variability between clusters of multilevel hierarchies such as multilevel modeling, replicated sampling techniques, sandwich estimation of standard errors and generalized estimating equations, multilevel modeling (MLM) is considered under this study. The category of multilevel models would be decided based on the

CONTACT M. R. Sooriyarachchi roshini@stat.cmb.ac.lk Department of Statistics, University of Colombo, Colombo 3, Sri Lanka.

type of data structure considered, distribution of the response variable and the variance structure (Rasbash, Steele, et al. 2017). Since a normally distributed response variable and a two-level multilevel structure are considered in this research, considering the variance structure, "two-level variance components model" is used for this study, which measures the proportion of total variability that is between clusters.

1.2. Bayesian Markov chain Monte Carlo (MCMC) methods

There are numerous strategies for estimating the parameters in multilevel models. Hox (2010) suggested several estimation methods such as maximum likelihood method, generalized least squares method, generalized estimating equations, Bayesian methods and Bootstrapping.

However, Bayesian Markov Chain Monte Carlo (MCMC) methods are selected for estimation of parameters in two-level variance components model considered in this study. MCMC methods are simulation-based procedures so that rather than simply producing point estimates the methods are executed for many iterations and at each iteration an estimate for each unknown parameter is produced (Browne, Charlton and Rasbash 2017). Two MCMC methods considered are namely;

1. Gibbs sampling
2. Metropolis-Hastings (MH) sampling

Recently, Pinto, I. V., and Sooriyarachchi, M. R. (2019) suggested the necessity of Bayesian MCMC methods in multilevel scenarios. It was the primary force which motivated the idea of considering comparision of MCMC methods in a multilevel scenario. As the preliminary initiative of this comparison in multilevel structure a two level model where the response is normally distributed is taken under consideration. According to, Browne, Charlton and Rasbash (2015), Gibbs sampling is available only for normal responses in MLwiN 2.19, the Statistical Package used for this research. Hence normal response multivariate modeling is considered in this study.

1.3. Objectives of the study

As selecting and tuning sampling is needed, the suitability of MCMC algorithms for a given problem remains challenging and a comprehensive comparison of different methods is so far not available. Thus the primary objective of this study is to compare the two main MCMC methods that are Gibbs and MH for a multilevel hierarchy, using simulation studies. Other than the main objectives, this study focuses on secondary objectives too. These secondary objectives are based on examining a real life dataset:

- To apply and verify the comparison of MCMC methods to a real life application.
- To describe about the posterior distribution of each MCMC methods to a real life data.
- To determine the effect of sample size on the two MCMC methods and compare the performance of small samples on each method.

2. Literature review

It is important to consider the hierarchical structure of individuals in groups when modeling such data because these data violate a crucial assumption of independence of observations that is widely used in statistical techniques. Goldstein (2011) improves the advantages of utilizing a statistical model which can demonstrate this clustered structure. Performing an analysis without

considering the existence of hierarchical structures would create technical problems, which may result in misleading conclusions. Rasbash, Steele, Browne and Goldstein (2009) proved these problems by carrying out analysis without considering the clustering for a pupil level analysis with no school terms, which caused standard errors of regression coefficients to be underestimated.

Jackson (1991), Guo and Zhao (2000), Rasbash, Steele, Browne and Goldstein (2017) and Snijders and Bosker (2012) are a few of the pioneer studies in multilevel data analysis in different fields of research. Moreover, Raudenbush and Bryk (2002) stated that, to manage the risky circumstances emerging from multilevel data, multilevel modeling has been developed to manage nested structures and Huang (2016) recommends multilevel modeling as a powerful and flexible approach to handle multilevel data. A clear overview of the logic and statistical theory behind multilevel models can be seen in Steenbergen and Jones (2002).

2.1. Methods of estimation

2.1.1. Description

Estimation of parameters in multilevel models is a fundamental step in statistical modeling thus this needs a more significant level of consideration. There are numerous estimation methods available for multilevel modeling such as Maximum Likelihood (ML), generalized least squares (GLS), Generalized estimating equations (GEE), Quasi Likelihood and Bayesian MCMC methods.

The usual method used to estimate both the regression coefficients and variance components is the ML method, because according to Hox (2010), it can yield the estimates of parameters which are both asymptotically efficient and consistent. However, ML estimation is not computationally feasible in multilevel models with discrete response variable. Hence, according to Goldstein (1999) quasi likelihood estimation methods are used in MLwiN in the case of multilevel data with discrete response variables.

GLS methods could be used for multilevel models for continuous responses, which is based on iterative processes recognized as iterative generalized least squares method (IGLS). However, Restricted Iterative Generalized Least Squares (RIGLS) method was established, in order to improve the results obtaining from IGLS. As stated by Rasbash et al. (2017), RIGLS is capable of producing more reliable estimates than IGLS when dealing with biased responses.

Another method which could be used to estimate the parameters in multilevel models with unknown correlation between outcomes is Generalized Estimating Equations (GEE) method. GEE was developed by Liang and Zeger (1986) for estimation purposes which are defined as an extension of Quasi-likelihood approach for repeated measures. Coelhoa, Infante, and Santos (2013) stated that, "For such cases (dependent data), the use of Generalized Estimating Equations (GEE) might be a valid alternative approach, as this modeling technique calculates a working correlation matrix that approximates the true correlation of the observations (Wang and Carey, 2003). A main advantage of GEE is that it comes up with a consistent estimate for the parameters even when the correlation matrix is not correctly specified.

2.1.2. Markov chain Monte Carlo (MCMC) methods

Bayesian MCMC methods are simulation based estimation procedures that will be compared for the multilevel scenario in this study. This section investigates the previous developments related to MCMC methods on which the foundation to this study is built.

Bayesian models for multilevel data using MCMC methods for Bayesian inference require numerical integration. These methods are designed specifically for hierarchical models although they can be adapted to fit other models (Browne, Charlton and Rasbash 2015). According to Goldstein, H. (2011), MCMC methods incorporate prior distribution assumptions and based

upon successively sampling from posterior distributions of the model parameters, yield a 'chain' which can then be used for constructing point and interval estimates of parameters. As stated by Geyer (1992), MCMC is a general method for the simulation of stochastic processes having probability densities known up to a constant of proportionality. Hence it may eventually have applications in every area of the statistic. The two most common procedures of MCMC methods in use are 'Gibbs sampling' and 'Metropolis-Hastings sampling'. Liu, Nordman, and Meeker (2014) mentioned about the introduction of both those methods in detail.

MLwiN (version 2.19) is used in this study, which is a specialized software and can be used for multilevel modeling. As mentioned in Rasbash et al. (2012), there are two families of simulation based estimation procedures available in MLwiN, which are MCMC sampling and bootstrapping. Bayesian MCMC methods such as Gibbs and MH are alternative to likelihood based estimation methods. Bayesian analyses depend on sampling based approximations to the distributions of interest via Markov chain Monte Carlo methods (Browne and Draper 2006). These MCMC methods can be implemented in MLwiN. The Bootstrap method could be used instead of MCMC methods for following reasons; explicitly for improving the accuracy of inferences about parameter values and correcting bias in the parameter estimates (Rasbash, Steele, Browne and Goldstein 2012). Goldstein (2011) reveals complete detail on bootstrap approach for multilevel generalized linear models.

2.2. Accuracy diagnostics

Any fitted model should be evaluated and confirmed on its performance before making inferences from the developed model. This is fundamentally to check if the predicted values from the fitted model provide values close to the observed data. This is alluded to as 'goodness of fit' (Hosmer et al. 2013).

The two main statistics which are usually used to check the GOF in one-level binary response models are deviance (likelihood ratio) and Pearson chi-square statistic. Hosmer and Lemeshow (1980) introduced a GOF test for single level logistic models that is most popular test used in these cases. However, Hosmer-Lemeshow test was modified by Lipsitz et al. (1996) where indicator variables were included to represent the deciles of risk. Lipstiz et al. (1996) introduced an extension for ordinal response models, while Fagerland et al. (2008) submitted a GOF test for multinomial response models. Perera et al. (2016) extended the GOF test for Multilevel Binary data based on the single level approaches mentioned above.

All the methods recommended in the above paragraph are not applicable to assess the GOF of normal response multilevel models with the application of MCMC methods. Therefore, there is a very limited amount of literature relevant to this field. According to Vivekananda Roy (2019), MCMC diagnostic tools are needed for deciding convergence of Markov chains from the stationarity. Also, although in general the longer the chain is run the better the Monte Carlo estimates it produces, in practice, it is desirable to use some stopping rules for prudent use of resources.

Monte Carlo Standard Error (MCSE) is an indicator of how much error has occurred in the estimates, usually regarding the expectation of posterior samples, from MCMC algorithms. MCSE is approximately a standard deviation throughout the expectation of the posterior samples, caused by uncertainty related with utilizing MCMC algorithms in general. Gelman et al. (2004) argued that MCSE is generally unimportant when the goal of inference is theta rather than expectation of posterior samples. The plot of estimated MCSE of the posterior estimate of the mean against the number of iterations could be attained from MCMC diagnostics in the MLwiN environment. According to Flegal, Haran, and Jones (2008), MCSE provides two desirable properties:

i. It gives useful information about the quality of the subsequent estimation and inference.

ii. it provides a theoretically justified, yet easily implemented, approach for determining appropriate stopping rules for their MCMC runs.

The effective sample size (ESS) provides estimates for the number of independent samples presented in the correlated MCMC chain. ESS is widely used in sample-based simulation methods for assessing the quality of a Monte Carlo approximation of a given distribution and of related integrals (Elvira, Martino and Christian 2018). According to Vehtari, Gelman, Simpson, Carpenter and Bürkner (2021), the ESS of a quantity of interest captures how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm. Clearly, the higher the ESS the better.

There are two contrasting accuracy diagnostics available in MLwiN. Raftery and Lewis (1992) propose a method for calculating an appropriate burn-in. They also discuss choosing a run-length so that the resulting probability estimate lies within a pre-specified interval which is called as "Raftery-Lewis diagnostic". According to Browne, Charlton, and Rasbash (2014), the statistic "Nhat" in Raftery-Lewis diagnostic is used to estimate the length of the Markov chain required to estimate a particular quantile to a given accuracy. In MLwiN the diagnostic is calculated for the two quantiles (the defaults are the 2.5% and 97.5% quantiles) that will form a central interval estimate. The Brooks-Draper diagnostic is a diagnostic based on the mean of the distribution. It is used to estimate the length of the Markov chain required to produce a mean estimate to k significant figures with a given accuracy Browne, Charlton, and Rasbash (2014).

In 1998, William Browne applied MCMC methods to multilevel models. The comparison of Bayesian MCMC method's performance is conducted through extensive simulations and an application to a real-life scenario in our study. However, Browne (1998) performed this task through three main steps which are; Fitting of multilevel models and investigation of maximum likelihood methods, deriving MCMC methods for those models and finally comparing the MCMC methods with the maximum likelihood method. While a two level multilevel model with normally distributed response is considered in this study, Browne (1998) considered N Level Gaussian models, Binary response multilevel logistic regression models and Gaussian models with complex variation at level 1. Moreover, Standard Errors of point estimates, 95% Confidence Intervals of Estimates and Accuracy Diagnostics (ESS and MCSE) are used to compare the performance of Gibbs and MH methods in this study. However, estimate bias and coverage properties were used by Brown (1998) to compare the MCMC methods with maximum likelihood methods.

2.3. Small samples

Hox, van de Schoot, and Matthijsse (2012), Kadane (2015), and McNeish and Stapleton (2016) have discussed the advantage of MCMC Bayesian methods over frequentist methods in the use of multilevel models for small samples. Therefore, it is of interest to validate their claim and examine which of Gibbs/MH, MCMC methods perform better for small samples.

3. Methodology

The simulation study and application to the real world dataset follow to analyze the performance as well as the applicability of the two main Bayesian MCMC methods; Gibbs sampling and MH method.

3.1 Simulation study

Performance of the MCMC methods such as Gibbs and MH are determined through the simulation study, where simulations have been developed by a varying number of clusters, observations

Table 1. Details for clusters.

Scenario	Cluster size (observations within a cluster)	Number of clusters
1	30	15
2	50	15
3	30	60
4	50	60

within each cluster, ICC values, burn-in length, MC length, prior distributions and acceptance rate. Two-level multilevel hierarchy is considered when generating the simulations for determining the performance of suggested MCMC methods. A dataset is generated under several specified conditions depending on properties considered in each scenario of the two-level hierarchy to guarantee that both MCMC methods are carried out on the same dataset. Macros in MLwiN version 2.19 are used to generate these datasets, where 1 is set as a seed value on each simulation scenario. The use of the seed value can confirm that running the same model with the same starting values and seed on a different machine will give the same answers (Browne, Charlton, and Rasbash 2015).

3.1.1 The fitted model in simulations

The two-level random intercept model is fitted with a single predictor (explanatory) variable in the simulations. As mentioned earlier, the normal response is considered in simulations, corresponding to the two-level multilevel model. Archer et al. (2007) mentioned that the explanatory variable can be generated from the Bernoulli distribution, normal distribution, and uniform distribution. But in this study, the single explanatory variable x_{ij} in the model is simulated from the normal distribution which was suggested by Perera et al. (2016).

The fitted model is

$$\boldsymbol{y_{ij}} = \boldsymbol{\beta_{0j}} + \boldsymbol{\beta_1}\ \boldsymbol{x_{ij}} + \boldsymbol{e_{ij}} \tag{1}$$

Where, $\boldsymbol{\beta_{0j}} = \boldsymbol{\beta_0} + \boldsymbol{u_j}$

$i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$ [k is the number of clusters]

The model stated above in (Archer et al. 2007) consists of both fixed and random components. β_1 is a fixed coefficient, β_0 is a fixed component and a random and u_j is a random component and are included into the intercept term β_{0j}, where the term u_j is exclusively used to indicate the random intercept. The random components, the error term in the Normal model $e_{ij} \sim N(0, \sigma_e^2)$ and the random component in the intercept $u_j \sim N(0, \sigma_u^2)$ follow the following distributions.

The model (Archer et al. 2007) is fitted for several combinations based on various conditions and then GOF and complexity of the models are evaluated under those conditions by utilizing the Deviance Information Criteria (DIC) values obtained from MLwiN.

3.1.2. Factors considered for simulations

The sample size of the model relies upon the number of clusters available in the model and the number of individuals in each designated cluster. Four scenarios for sample sizes are chosen as indicated by the rules determined by Maas and Hox (2005), Kreft and de Leeuw (1998) and Perera et al. (2016). Details of the selections of cluster sizes and the number of clusters for simulations are clearly provided in the following Table 1.

According to the values set out by Perera et al. (2016) for the standard deviation of level 2 residuals, three scenarios for the standard deviation are selected in this simulation study. Those different values of standard deviation of level 2 residuals allow to vary the intra cluster correlation (ICC) according to the equation, $ICC = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$, are shown in Table 2.

Table 2. Standard deviation of Level 2 residuals.

Scenario	Standard deviation (σ_u)	Variance (σ_u^2)
1	1.0	1.0
2	1.5	2.25
3	2.0	4.0

3.1.3. Incorporating factors related to MCMC algorithms

Five hundred (500) is the default value for burn-in length in MLwiN, Hence, for the comparative purpose, the values were selected for burn-in length such that those values are less than and greater than the default value of burn-in length. Likewise considering chain length, 5000 is the default value available for chain length in MLwiN, Hence, for the comparative purpose, the values were selected for chain length such that those values are less than and greater than the default value of chain length. These specifications of values result in 9 combinations of burn-in length and chain length to be used in the scenario of simulations. Above mentioned details of the selections are clearly summarized in the following Table 3.

According to Goldstein (2011), diffused priors are chosen for this study. Two distributions are selected for diffused priors based on the guidelines suggested by Browne (2015).

Considering acceptance rate, 50% is the default value available for acceptance rate in MLwiN, and Browne (2015) suggested that rates between 30% and 70% provide a useful compromise between a proposal variance that is too large and a variance that is too small through the literature. Thus, considering these aspects, three combinations of acceptance rate values are selected. Above mentioned details of the selections are clearly summarized in the following Table 4.

A dataset is simulated using macros in the MLwiN software for each of the scenarios which does not include the factors related to MCMC algorithms, hence all together 12 datasets are simulated. Then specific values for the MCMC factors listed under Table 5 are assigned when applying the Gibbs and MH methods separately. This process yields a total of 648 combinations in the simulation study. Thus, the combinations resulting from the choices made above are summarized in Table 5.

4. Results and conclusions

When applying the different specifications for simulations, the corresponding DIC values obtained from MLwiN output are noted for each combination to evaluate the GOF of the model. By looking at the DIC values of all 648 combinations, it has been found that the 322nd combination contains the minimum DIC value of 2705.82 for both Gibbs and MH methods. Thus, this combination is selected as best based on the suggestion of Francois and Laval (2018). Results obtained from MCMC methods in simulation and comparison of both MCMC methods will be carried out for this 322nd combination. Table 6 summarize the details of the selected best combination.

MCMC diagnostic outputs in the MLwiN environment are obtained for fixed and random effects under MH and Gibbs methods which can be used to compare the considered MCMC methods based on detailed diagnostic information about unknown parameters in the model.

Point and interval estimates of unknown parameters can be used to compare the estimation algorithms based on the standard error of the point estimates and length of the confidence interval. Table 7 summarizes the point estimates using the standard errors as well as the 95% interval estimates coming from Gibbs and MH fittings.

It could be seen that; Considering point estimates of fixed effects, the MH method yields the estimates of parameters with the smallest Standard errors compared to Gibbs, Considering point estimates of random effects, the MH method does better for estimation of between cluster variance (σ_u^2). But the standard error coming from MH looks slightly higher (0.003) than that from

Table 3. Details of burn-in length and chain length.

Scenario	Burn-in length	Chain length
1	300	3000
2	500	3000
3	700	3000
4	300	5000
5	500	5000
6	700	5000
7	300	7000
8	500	7000
9	700	7000

Table 4. Details of acceptance rates and priors.

Scenario	Acceptance rate	Priors
1	30%	Gamma
2	50%	Gamma
3	70%	Gamma
4	30%	Uniform
5	50%	Uniform
6	70%	Uniform

Table 5. Different combinations.

Factors considered	Selections	Combinations
Standard deviation of the random component (σ_u)	1, 1.5, 2	3
Number of clusters (Level 2)	15, 60	2
Cluster size (Level 1)	30, 50	2
Total combinations except factors related to MCMC algorithms	$3 \times 2 \times 2 = 12$	
Burn-in length	300, 500, 700	3
Chain length	3000, 5000, 7000	3
Priors	Gamma, Uniform	2
Acceptance rate	30%, 50%, 70%	3
Total combinations	$3 \times 2 \times 2 \times 3 \times 3 \times 2 \times 3 = 648$	

Table 6. Details of the selected best combination.

Factors considered	Values
Standard deviation of the random component (σ_u)	1.5
Number of clusters (Level 2)	15
Cluster size (Level 1)	30
Burn-in length	700
Chain length	3000
Priors	Gamma
Acceptance rate	70%

Table 7. Point estimates, standard errors and 95% confidence intervals of estimates.

	Gibbs		MH	
Parameters	Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
Fixed effects				
β_0	−0.161 (0.079)	−0.32,-0.053	−0.079 (0.037)	−0.122,0.005
β_1	0.444 (0.425)	−0.428,1.278	0.458 (0.419)	−0.324,1.281
Random effects				
σ_u^2	0.365 (0.175)	0.147,0.793	0.332 (0.168)	0.127,0.765
σ_e^2	78.754 (5.523)	69.15,91.10	79.193 (5.529)	69.76,91.20

Gibbs in the estimation of σ_e^2. The confidence intervals that are as narrow as possible can be concluded as better, thus MH method seems to be better than Gibbs as it yields the narrowest 95% interval estimates for both fixed and random effects in the two-level random intercept model.

Table 8. Details of the accuracy diagnostics.

	Gibbs		MH	
Parameters	ESS	MCSE	ESS	MCSE
Fixed effects				
β_0	4	0.038	5	0.020
β_1	2973	0.008	489	0.02
Random effects				
σ_u^2	2070	0.004	1178	0.004
σ_e^2	324	0.285	366	0.304

It is good to see that the best combination (322) is for the smallest sample. That is for the smallest number of clusters having the smallest cluster size. This indicates that the two MCMC methods perform better for small samples rather than large samples. Of the two methods the MH performs better than the Gibbs method with all parameters except σ_e^2 having smaller standard errors and narrower confidence intervals for the former method compared to the latter method.

Table 8 summarizes the accuracy diagnostics applicable for MCMC algorithms such as ESS and MCSE under both considered procedures, where Brooks-Draper diagnostic is not considered for the comparison as it was recognized that, this diagnostic isn't satisfied for all model parameters when having chain length equal to 3000. The higher value for ESS indicates the better fit as it estimates the number of independent samples presented in the correlated MCMC chain. MCSE can be achieved by dividing the standard deviation of the Markov chain values by square root of number of iterations (i.e., $\text{MCSE} = \text{SD}/\sqrt{\text{n}}$), so MCSE is an indicator of the accuracy of posterior mean estimate in Bayesian approaches.

It could be seen that in the estimation of β_0, there is no huge difference in ESS between both MCMC methods. So considering MCSE, MH does better than Gibbs as it provides the minimum MCSE in estimating the intercept parameter. In the estimation of β_1, Gibbs does better than MH based on the ESS and MCSE values. In the estimation of σ_u^2, MCSE values are equal for both MCMC methods. So considering ESS, Gibbs does better than MH as it is providing maximum ESS. In the estimation of σ_e^2, there is no huge difference in MCSE between both MCMC methods. So considering ESS, MH does better than Gibbs as it is providing maximum ESS in the estimation of σ_e^2.

The typical goal is to attain a large ESS. According to Drummond et al. (2006), it would be better if the value of ESS is near to an arbitrary cutoff of 200. This is achieved for all the parameters of this study except β_0. Since ESS denotes the number of independent samples presented in the correlated MCMC chain, there might be a chance to get a small number of independent samples occurring in the correlated MCMC chain.

MCMC diagnostics output for β_0 under the Gibbs and MH have also been considered for more clarification which are given in Figures 1 and 2. It is noticed that the generated values for β_0 seem somewhat auto correlated in the plot of parameter traces for β_0. This means each value of the Markov chain is highly correlated with the previous value. Moreover, the plot of ACF indicates that the generated chain consists of dependently distributed data. These patterns were not observed for the other parameters; hence, it might be the reason for getting small values of ESS for β_0.

4.1. Exploring the effect of the seed

In order to observe the subtle difference between MH and Gibbs the experiment was repeated over multiple seeds. Tables 7 and 8 are repeated over Tables 9 and 10 with different seeds.

Table 9 indicates that considering the standard error of the point estimates and the width of the interval estimates, overall conclusions are similar for different seeds used, that is MH shows a

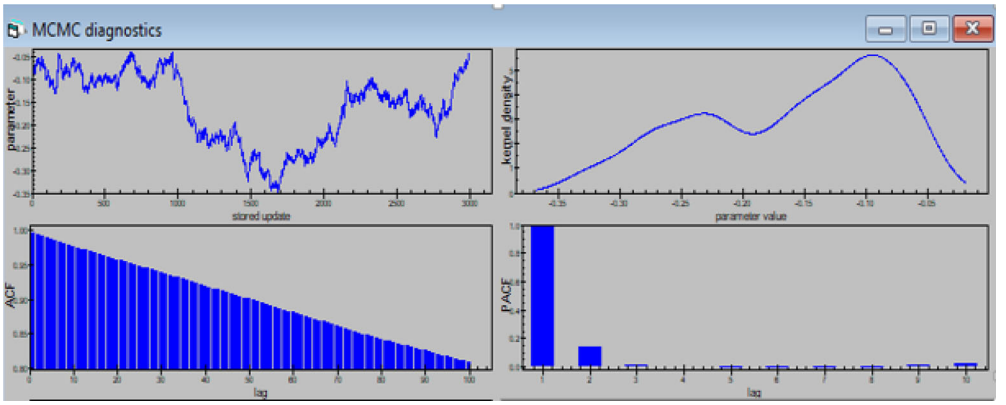

Figure 1. MCMC diagnostics for β_0 under the Gibbs Method

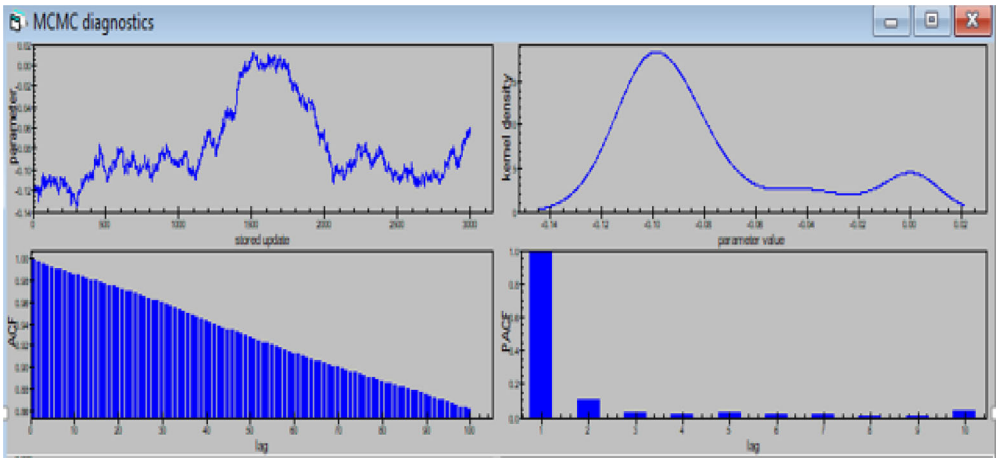

Figure 2. MCMC diagnostics for β_0 under the MH Method

slight superiority over Gibbs but nothing significant. However, from Table 10 considering accuracy diagnostics there are slight differences depending on the seeds.

4.2. Behavior of considered factors in simulations

There were several parameters such as standard deviation of random component, number of clusters, cluster size, burn-in length, chain length, priors and acceptance rate considered to designate the simulations in this study. Tables of standard errors (SE) of estimates against the specific combinations obtained from both Gibbs and MH algorithms are used to identify which factors are affecting the MCMC methods.

4.3. Practical application

Application of Gibbs and MH methods to a real-life scenario that consists of the multilevel hierarchy is important to verify the possibility of these algorithms in practical situations and also to compare the parameter estimates given by each method of estimation.

The dataset for this application was collected from the World Statistics Pocketbook published by the United Nations (UN) in 2005. Two hundred and nine countries which were available

Table 9. Summary of point and interval estimates under both MCMC methods.Where highlighted the narrowest interval estimate & smallest SE for each parameter

Seed = 10	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	−0.161 (0.079)	−0.32, −0.053	−0.079 (0.037)	−0.122,0.005
	β_1	0.444 (0.425)	−0.428,1.278	0.458 (0.419)	−0.324,1.281
	Random effects				
	σ_y^2	0.365 (0.175)	0.147,0.793	0.332 (0.168)	0.127,0.765
	σ_e^2	78.754 (5.523)	69.15,91.10	79.193 (5.529)	69.76,91.20
Seed = 100	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	0.297 (0.222)	−0.205,0.678	0.425 (0.094)	0.240,0.586
	β_1	0.216 (0.412)	−0.589,1.029	0.236 (0.422)	−0.674,0.996
	Random effects				
	σ_y^2	0.698 (0.323)	0.314,1.519	0.655 (0.279)	0.305,1.365
	σ_e^2	71.464 (4.693)	62.817,81.238	71.549 (4.856)	62.194,81.892
Seed = 1	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	0.228 (0.140)	−0.088,0.470	0.294 (0.076)	0.176,0.434
	β_1	−0.589 (0.293)	−1.171, −0.038	−0.601 (0.277)	−1.173,-0.048
	Random effects				
	σ_y^2	0.441 (0.194)	0.198,0.947	0.427 (0.180)	0.200,0.893
	σ_e^2	37.469 (2.584)	32.805,42.663	37.458 (2.657)	32.548,43.270
Seed = 500	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	0.094 (0.100)	−0.151,0.253	−0.018 (0.046)	−0.099,0.068
	β_1	0.100 (0.391)	−0.669,0.857	0.119 (0.366)	−0.586,0.850
	Random effects				
	σ_y^2	0.321 (0.151)	0.143,0.699	0.304 (0.146)	0.133,0.669
	σ_e^2	64.466 (4.363)	56.800,73.438	64.178 (4.340)	56.574,73.051
Seed = 10000	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	−0.139 (0.193)	−0.493,0.299	0.019 (0.096)	−0.132,0.203
	β_1	0.040 (0.578)	−1.090,1.179	0.046 (0.593)	−1.115,1.213

(*continued*)

	Random effects				
	σ_u^2	1.031 (0.456)	0.481,2.124	0.964 (0.421)	0.450,2.038
	σ_e^2	148.432 (9.696)	130.688,168.405	148.319 (9.637)	130.387,167.667
Seed = 2000	Parameters	Gibbs		MH	
		Point estimates with SE	95% interval estimates	Point estimates with SE	95% interval estimates
	Fixed effects				
	β_0	−0.154 (0.117)	−0.318,0.059	−0.105 (0.025)	−0.143,-0.059
	β_1	−0.981 (0.468)	−1.931, −0.091	−0.958(0.457)	−1.881, −0.072
	Random effects				
	σ_u^2	0.752 (0.325)	0.340,1.624	0.743 (0.318)	0.343,1.547
	σ_e^2	95.291 (6.611)	82.920,109.070	95.624 (6.368)	84.192,109.262

Table 10. Details of the accuracy diagnostics.Where mentioned the method which performed well for each parameter

Seed = 10	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_0	4	0.038	5	0.020	MH
	β_1	2973	0.008	489	0.02	Gibbs
	Random effects					
	σ_u^2	2070	0.004	1178	0.004	Gibbs
	σ_e^2	324	0.285	366	0.304	MH
	β_0	9	0.082	14	0.030	MH
Seed = 1	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_1	2780	0.008	301	0.023	Gibbs
	Random effects					
	σ_u^2	1696	0.006	2440	0.005	MH
	σ_e^2	453	0.220	525	0.223	MH
	β_0	9	0.069	4	0.036	MH
Seed = 1000	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_1	2533	0.005	537	0.013	Gibbs
	Random effects					
	σ_u^2	2196	0.004	1932	0.004	Gibbs
	σ_e^2	503	0.119	394	0.131	Gibbs
	β_0	15	0.034	9	0.016	MH
Seed = 1000	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_1	2532	0.008	616	0.015	Gibbs
	Random effects					
	σ_u^2	2177	0.003	1564	0.003	Gibbs
	σ_e^2	524	0.200	404	0.221	Gibbs
	β_0	7	0.072	4	0.034	MH
Seed = 50	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_1	2309	0.011	410	0.029	Gibbs
	Random effects					
	σ_u^2	2087	0.009	2498	0.008	MH
	σ_e^2	575	0.425	412	0.500	Gibbs
	β_0	3	0.109	3	0.013	MH
Seed = 2000	Parameters	**Gibbs**		**MH**		
		ESS	MCSE	ESS	MCSE	
	Fixed effects					
	β_1	2894	0.009	479	0.021	Gibbs
	Random effects					
	σ_u^2	2911	0.006	2482	0.006	Gibbs
	σ_e^2	429	0.320	416	0.331	Gibbs

under the UN as of 30th November 2004 are available in this dataset. The response variable in the dataset is the 'weighted average life expectancy at birth 2000-2005' which is a continuous variable. Weighted average life expectancy at birth (WLE) is a statistical measure of the average age at which an individual is expected to die. Countries available in the dataset are nested within several regions. All the explanatory variables except 'region' have been measured at the lowest level of the hierarchy. Hence, the dataset allows fitting the two-level multilevel model by considering the country at level-1 and the region at level-2.

The dataset consists of 209 UN countries with huge amount (i.e., 1424) of missing values. Initially the dataset consists of 50 explanatory variables. According to the earlier study which was carried out for this same dataset by Pathirathne and Sooriyarachchi (2019), 50 explanatory variables were reduced to 31 due to the removal of variables having a high proportion of missing values and several countries also have been omitted from the dataset due to those counties having

missing values for the important variables selected. Hence, the dataset resulted in only 115 counties and 31 explanatory variables for further analysis. A region that had only one observation and the country correspond to the irrelevant region was removed. In addition, the categories of regions related to 'Oceania' which had only a few observations were merged into one region to elude any opportunities of non-convergence with minor cluster sizes. Therefore, the dataset resulted with only 14 regions and 114 countries.

Regularization techniques such as the Lasso and Elastic Net (Bonaccorso 2017) were applied to the dataset to identify the important predictors by handling the multicollinearity, then the 18 variables selected from the Lasso method was considered for further analysis based on the smaller MSE calculated for the test set. Then the suitable two-level random intercept model was fitted and the backward selection was applied to find out the most suitable model which consists of significant variables only by using the Wald statistic at 5% level of significance. The model selection process was verified by checking the AIC values and the Test MSEs at each stage of backward elimination. Before applying the MCMC estimation methods to select the best model yielded from backward elimination, 54 combinations were designated by varying the burn-in length, chain length, acceptance rate and priors to find out the best combination of choices for the factors related to MCMC methods.

The results obtained from Gibbs and MH methods for the model fitted for best combination selected based on smaller DIC out of 54 combinations have been used to compare those methods. Summary statistics of the posterior distribution of estimates for each unknown parameter indicating that narrower confidence intervals arose for MH in estimating fixed effects, however, Gibbs has smaller SE of point estimates. Gibbs does better than MH in estimating the random effects, as the smaller SE of point estimates and the narrower confidence interval occurred for Gibbs. Considering the accuracy diagnostics results, Since ESS's were so much higher for Gibbs compared to MH, it could be decided that, Gibbs has the superiority over MH in estimating both random and fixed effects of the fitted model, which could be verified by smaller MCSE values for Gibbs. However, the picture is not much clearer for the practical applications as a simulation study, because it's only for one specific selected dataset.

5. Discussion

5.1. Important Conclusions

Theory, circumstantial situation and the logic behind the illustration of the Gibbs and MH algorithms are easily understood. The main idea of these Bayesian MCMC methods is to generate a chain of estimates from the joint posterior distribution of parameters in the model through an iterative process, which can be used to construct a useful statistical summary for a single parameter.

Simulation results indicate that only the parameters such as a number of clusters, priors and acceptance rates are affecting the performance of the MCMC methods.

- SE values of estimates showed a clear increment with the decrease in number of clusters for both MCMC methods
- 50% acceptance rate yields the better estimation through the MH algorithm.
- Overall Gamma priors give better estimation than Uniform priors for most of the model parameters.
- The smallest sample size (smaller number of clusters with the smaller sample size) yields the best results.

The entire simulation results indicate that there is no difference in the performance of both MCMC methods in estimating slope parameter and individual level variance (σ_e^2) but MH has the superiority over Gibbs in estimating the intercept parameter. Moreover, following conclusions on MCMC methods could be derived for the best combination;

- Generally, MH has the superiority over Gibbs in estimating the two-level random intercept model parameters based on the summary statistics of the posterior distributions.
- MH performs better than Gibbs in estimating the individual level variance (σ_e^2) and intercept parameter, but Gibbs does better than MH in estimating the between cluster variance (σ_u^2) and slope parameter of the two-level random intercept model based on accuracy diagnostic statistics.
- MH does better than Gibbs for small samples.

Overall, it could be concluded to go with MH generally to estimate the model parameters in the multilevel hierarchy, for small samples. However, for moderate to large samples there is no significant difference between MH and Gibbs.

There is no complexity in applying the MCMC methods to a real-life scenario, however, the performance of the MCMC methods can vary with the selected dataset.

5.2. Limitations

The study considers the multilevel model fitting using only two of the most common MCMC methods in use. However, there are several estimation methods such as ML, GLS, GEE, Quasi Likelihood and Bootstrapping is available for multilevel modeling. The study was established for two-level random intercept models because of the simplicity. Gibbs sampling is available for normal responses only in the software MLwiN (version 2.19), which was used in this study. There were a huge number of combinations in the simulation study. Thus, the tables with the whole combination's results were slightly unclear and the best combination was selected to compare the two MCMC methods. The regularization techniques haven't been applied in the presence of a multilevel model in the practical application of this study. The dataset used in the real-world application of this study consists of only a moderate numbers of observations.

5.3. Suggestions for further work

The following are a few suggestions for future researches associated with this field.

- The study could be improved with more than two levels in the hierarchy as well as the other responses other than normal distribution such as binomial, ordinal etc.
- More variations of cluster sizes, number of clusters, burn-in length, chain length and acceptance rate could be considered in the simulation study.

References

Archer, K., S. Lemeshow, and D. Hosmer. 2007. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis*.

Bonaccorso, G. 2017. *Machine learning algorithms*. Birmingham, UK: Packt Publishing Ltd.

Browne, W. 1998. *Applying MCMC methods to multi-level models*. University of Bath.

Browne, W., C. Charlton, and J. Rasbash. 2014. *MCMC estimation in MLwiN Version 2.31*. Centre for Multilevel Modelling, University of Bristol.

Browne, W., C. Charlton, and J. Rasbash. 2015. *MCMC estimation in MLwiN, Version 2.32*. Centre for Multilevel Modelling, University of Bristol.

Browne, W., C. Charlton, and J. Rasbash. 2017. *MCMC estimation in MLwiN, Version 3.01*. Centre for Multilevel Modelling, University of Bristol.

Browne, W., and Draper. D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 1 (3):473–514.

Coelhoa, R., P. Infante, and M. N. Santos. 2013. Application of Generalized Linear Models and Generalized Estimation Equations to model at-haulback mortality of blue sharks captured in a pelagic longline fishery in the Atlantic Ocean. *Fisheries Research* 145:66–75.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. *Relaxed phylogenetics and dating with confidence*. *PLoS Biology* 4 (5):e88. doi:10.1371/journal.pbio.0040088.

Elvira, V., L. Martino, and P. Christian. 2018. *Rethinking the Effective Sample Size, Computer Science, Mathematics*. arXiv:1809.04129. USA: Cornell University.

Fagerland, M. W., D. W. Hosmer, and A. M. Bofin. 2008. Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine* 27 (21):4238–53. doi:10.1002/sim.3202.

Flegal, J., M. Haran, and G. Jones. 2008. Markov chain Monte Carlo: Can We Trust the Third Significant Figure. *Statistical Science* 23 (2):250–60.

Francois, O., and G. Laval. 2018. *Deviance Information Criteria for Model Selection in Approximate Bayesian Computation*. University Joseph Fourier Grenoble, Centre National de la Research Scientific, TIMC-IMAG UMR 5525, 38042 Grenoble, France.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. *Bayesian Data Analysis*. *Texts in Statistical Science*, 2nd ed. London: Chapman and Hall.

Geyer, C. 1992. Practical Markov Chain Monte Carlo. *Statistical Science* 7 (4):473–83. doi:10.1214/ss/1177011137.

Goldstein, H. 1999. *Multilevel statistical models*. *Multilevel models project*. London: Institute of Education.

Goldstein, H. 2011. *Multilevel statistical models* (4th ed.). West Sussex, UK: John Wiley and Sons, Ltd

Guo, G., and H. Zhao. 2000. Multilevel modeling for binary data. *Annual Review of Sociology* 26 (1):441–62. doi:10.1146/annurev.soc.26.1.441.

Hosmer, D. W., and S. Lemesbow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics – Theory and Methods* 9 (10):1043–69.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied logistic regression* (3rd ed.). Chichester, UK: John Wiley and Sons, Inc.

Hox, J. J. 2010. *Multilevel analysis- techiniques and applications*. New York: Routledge.

Hox, J. J., R. van de Schoot, and S. Matthijsse. 2012. How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods* 6 (2):87–93. doi:10.18148/srm/2012.v6i2.5033.

Huang, F. 2016. Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education* 84 (1):175–96. doi:10.1080/00220973.2014.952397.

Jackson, J. E. 1991. Estimation of models with variable coefficients. *Political Analysis* 3:27–49. doi:10.1093/pan/3.1.27.

Kadane, J. B. 2015. Bayesian methods for prevention research. *Prevention Science: The Official Journal of the Society for Prevention Research* 16 (7):1017–25. Octdoi:10.1007/s11121-014-0531-x. PMID: 25468407

Kref, I., and J. de Leeuw. 1998. *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Liang, K., and S. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73 (1):13–22. doi:10.1093/biomet/73.1.13.

Lipsitz, S., G. Fitzmaurice, and G. Molenberghs. 1996. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45 (2):175–90.

Liu, J., D. J. Nordman, and W. Meeker. 2014. *The Number of MCMC draws needed to compute bayesian credible bounds*. Statistics, Iowa State University, Digital Repository, USA.

Maas, C., and J. Hox. 2005. *Sufficient sample sizes for multilevel modeling Methodology*. *Methodology* 1 (3):86–92. doi:10.1027/1614-2241.1.3.86.

McNeish, D. M., and L. M. Stapleton. 2016. The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review* 28 (2):295–314. doi:10.1007/s10648-014-9287-x.

Pathirathne, L., and M. R. Sooriyarachchi. 2019. Factors affecting life expectancy: A global perspective. *Journal of Environment Protection and Sustainable Development* 5 (1):4–21.

Perera, A., M. R. Sooriyarachchi, and S. Wickramasuriya. 2016. A goodness of fit test for the multilevel logistic model. *Communications in Statistics – Simulation and Computation* 45 (2):643–59. doi:10.1080/03610918.2013.868906.

Pinto, I. V., and M. R. Sooriyarachchi. 2019. Comparison of methods of estimation for use in goodness of fit tests for binary multilevel models. *International Science Index 148, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering* 13 (4):68– 73.

Raftery, A., and S. Lewis. 1992. How many iterations in the Gibbs sampler?. In *Bayesian Statistics 4*, eds. J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 763–73. doi:10.21236/ada640705

Rasbash, J., F. Steele, W. Browne, and H. Goldstein. 2017. *A user's guide to MLwiN, Version 3.00. Centre for multilevel modelling*. Bristol, UK: University of Bristol.

Rasbash, J., Steele, F. Browne, and W. Goldstein. 2009. *A User's Guide to MLwiN, Version 2.10. Centre for multilevel modelling*. Bristol, UK: University of Bristol.

Rasbash, J., Steele, F. Browne, and W. Goldstein. 2012. A User's Guide to MLwiN,Version 2.26. *Centre for Multilevel Modelling*. Bristol, UK: University of Bristol.

Raudenbush, S., and A. Bryk. 2002. *Hierarchical linear models: Applications and data analysis Methods (*2nd ed.*)*. London: Sage.

Roy, V. 2019. *Convergence diagnostics for Markov chain Monte Carlo*. Ames: Department of Statistics, Iowa state University Ames.

Snijders, T., and R. Bosker. 2012. *Multilevel Analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: SAGE Publications Ltd.

Steenbergen, M., and B. Jones. 2002. Modeling Multilevel Data Structures. *American Journal of Political Science* 46 (1):218. doi:10.2307/3088424.

Vehtari, A., Gelman, A. Simpson, B. Carpenter, and P.-C. Bürkner. 2021. Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC. *Bayesian Analysis* 16 (2):667–718. doi:10.1214/20-BA1221