Bayesian model averaging in longitudinal studies using Bayesian variable selection methods

YIMER, Belay Birlie; OTAVA, Martin; Degefa, Teshome; Yewhalaw, Delenasaw & SHKEDY, Ziv (2023) Bayesian model averaging in longitudinal studies using Bayesian variable selection methods. In: COMMUNICATIONS IN STATISTICS-SIMULATION AND COMPUTATION, 52(6), p. 2646-2665.

DOI: 10.1080/03610918.2021.1914088
Handle: http://hdl.handle.net/1942/34076

Bayesian Model Averaging in Longitudinal Studies using Bayesian Variable Selection Methods

Belay Birlie Yimer[ab], Martin Otava[c], Teshome Degefa[d], Delenasaw Yewhalaw[d], and Ziv Shkedy[b]

[a]Arthritis Research UK Centre for Epidemiology, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Manchester, UK M13 9PT; [b]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium[c]Statistics and Decision Sciences, Quantitative Sciences, Janssen Pharmaceutical companies of Johnson & Johnson, Czech Republic[d]School of Medical Laboratory Sciences, Faculty of Health Sciences, Jimma University, Jimma, Ethiopia

ARTICLE HISTORY
Compiled March 25, 2021

Abstract
Parameter estimation is often considered as a post selection problem, i.e. the parameters of interest are often estimated based on "the best" model. However, this approach does not take into account that "the best" model was selected from a set of possible models. Ignoring this uncertainty may lead to bias in estimation. In this paper, we present a Bayesian variable selection (BVS) approach for model averaging which would address the model uncertainty. Although averaging would be preferred approach, BVS can be used as well for model selection if the interest is to select one among the set of candidate models. The performance of Bayesian variable selection is compared with the information criterion based model averaging on real longitudinal data and through simulations study.

KEYWORDS
Clustering, Bayesian Modeling, Model selection, Multimodal Inference, Bayesian Variable Selection, Information criteria.

1. Introduction

In situations where the underlying goal of model selection is parameter estimation or prediction and/or no single model is overwhelmingly supported by the data, inferences and estimation can be made from several, even all, plausible models under consideration using model averaging techniques. Parameter estimates or predictions obtained by model averaging are advocated since they reduce model selection bias and account for model selection uncertainty (Draper 1995, Hoeting et al. 1999, Burnham and Anderson 2003, Johnson and Omland 2004, Claeskens and Hjort 2008, Lin et al. 2012).

The model averaging approach to incorporate model uncertainty into estimation and inferences relies on weighting model-specific parameter estimates with the posterior probability of the corresponding model. Various strategies have been suggested both under the frequentist (Burnham and Anderson 2003, Claeskens and Hjort 2008, Lin et al. 2012) as well as under the Bayesian paradigm (George and McCulloch 1993, Carlin and Chib 1995, Kuo and Mallick 1998). In the frequentist framework, the posterior model probabilities are estimated based on the Akaike information criterion (AIC, Akaike 1974) or the Bayesian information criterion (BIC, Schwarz 1978). Other information criterion, such as the Watanabe–Akaike information criterion (WBIC, Watanabe 2013), approximations to the marginal likelihood can also be used. In the Bayesian context, Bayesain variable selection (BVS) approaches that combine variable (model) selection with estimation of the unknown model parameters based on a spike and slab prior were proposed by George and McCulloch (1993), Carlin and Chib (1995), Kuo and Mallick (1998). Within the BVS, a variable indicator that allow for exclusion of possible predictors is introduced into the model. The resulting model averaged parameter estimates are a by-product of the variable selection procedure. In this paper we will investigate the Bayesian variable selection approach of Kuo and Mallick (1998)

Belay Birlie Yimer. Author. Email: belay birlie.yimer@manchester.ac.uk

extensively studied by Kasim et al. (2012) and Otava et al. (2014, 2017) in the context of order-restricted dose response modelling.

The current paper has two main objectives. The first is to illustrate the usage of Bayesian variable selection method for parameter estimation in longitudinal count data setting while taking in to account model uncertainty. The performance of the BVS method is compared with alternative methods based on information criteria in both real life case studies and simulation study. The second objective is to investigate the effect of prior specification for the model specific parameters on the performance of the BVS method.

For completeness, we will address the use of BVS as model selection tool. We present a simulation study in which the performance of the BVS model is compared with other model selection procedures based on the AIC, BIC and deviance information criterion (DIC, Spiegelhalter et al. 2002). All methods for model selection are applied for the case study as well.

This manuscript is organized as follows. The *Anopheles* mosquito count data set used to illustrate the methodology presented in this paper is described in Section 2. The methodological background for both the information criteria (IC) based methods and the BVS method is summarized in Section 3. The two approaches are applied to the *Anopheles* mosquito count data set in Section 4 and the results are evaluated. A large scale simulation study, conducted to compare the performance of the IC based methods with the BVS method is presented in Section 5. Finally, the findings are summarized and discussed in Section 6.

2. Case Study

Indoor resting *Anopheles* mosquitoes count data presented by Degefa et al. (2015) is used for illustration of the BVS method and for the comparison between the BVS and the information IC based model averaging approaches. The authors conducted a longitudinal entomological and parasitological study in Jimma town, Southwest Ethiopia to investigate the impact of resettlement on malaria incidence and transmission intensity. For a complete discussion about the study area and setting we refer to Degefa et al. (2015). Briefly, data were collected from two groups of villages where the first group (at risk) are villages from resettlement area that are recently inhabited which are believed to be prone to an increase in malaria transmission due to ecological transformation (i.e., suitable mosquito breeding sites created as a result of the resettlement) and the second group (control) are villages from the centre of the town which has been inhabited for long time. The data is displayed in Figure 1.

For the entomological study, adult *Anopheles* mosquitoes resting inside human habitations were collected monthly (June, 2013 - November, 2013) from 20 selected houses per village using pyrethrum spray catches (PSCs). The outcome of interest is the number of female *Anopheles* mosquitoes observed in the selected household at each observation time. The research question is whether or not the ecological transformation increases female *Anopheles* mosquito abundance. That is, do households from the resettled area have a higher mosquito count as compared to non-resettled area?

3. Methodology

Model-building for generalized linear models and their extensions involves choosing the independent variables, the link function, the variance function and the distribution of the response variable (McCulloch and Neuhaus, 2005). Each possible combination of independent variables (i.e., predictors), link function and variance function corresponds to a different model. In this paper, we consider the case where the model uncertainty is related to variable selection and therefore the models considered differ only in the form of the linear predictor. Accounting for the other sources of model uncertainty such as distributional assumption, the choice of the link function and variance function are beyond the scope of this paper.

The variable selection problem arises when there is an unknown subset of the proposed predictors for which the corresponding parameters equal to zero. Thus, this subset of predictors should not be included in the model. Within the frequentist modelling framework, many methods have been proposed for the selection of suitable predictors including those that sequentially delete or add predictors using the change in mean squared error (forward, backward, and stepwise procedures, Hocking 1976) and those based on information criterion such as BIC (Schwarz, 1978) and AIC (Akaike, 1974). Within the hierarchical Bayesian modelling framework, the deviance information criterion (DIC) was proposed by Spiegelhalter et al. (2002) for variable selection. For both approaches,

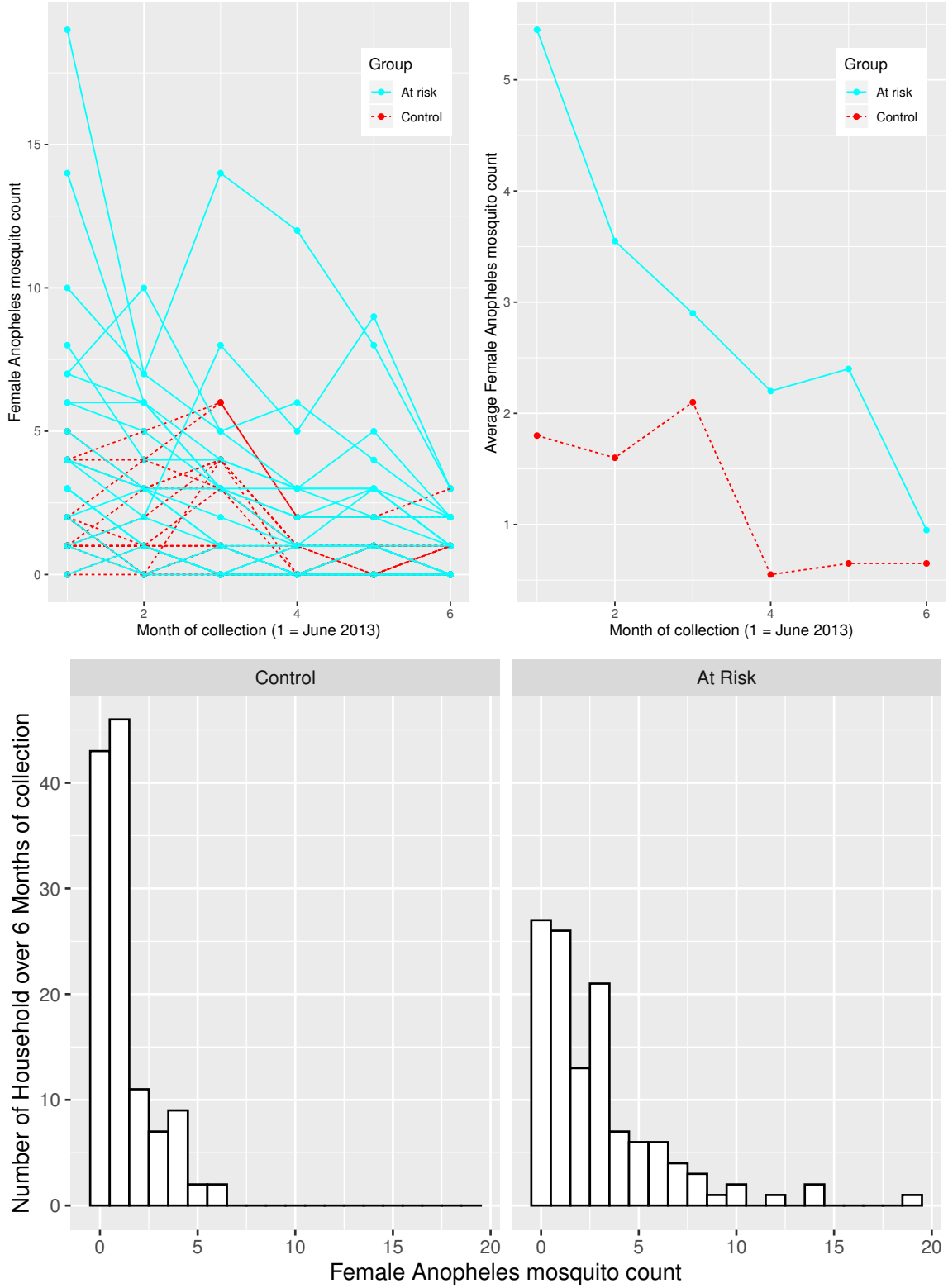

Figure 1. Female *Anopheles* mosquito count data. Household specific (top left), average profile (top right), histogram of total *Anopheles* mosquito count data (bottom) for at risk and control villages, Jimma town, South west Ethiopia (June to November 2013).

the estimation and inference of the unknown parameters of the final model is done after the variable selection and therefore the uncertainty due to the number of models fitted in the first stage is ignored. In this paper, we focus on a different approach, the Bayesian variable selection approach. Within this approach, the posterior probability that a single variable should be included in a model (the posterior inclusion probability) can be estimated and the estimation of individual contribution of the predictors can be done by averaging the parameter estimates obtained across several models. In Section 3.1, we briefly discuss the main concepts behind the BVS approach while in

Section 3.2 we formulate a BVS model for the *Anopheles* mosquitoes count data. Section 3.3 is devoted to model selection procedure based on information criteria.

3.1. Bayesian Variable Selection: An Introduction

3.1.1. Formulation for a Bayesian Variable Selection Model

The Bayesian variable selection (BVS) method allows us to formulate one model for a set of candidate models. Several Bayesian variable selection methods that combine variable selection and estimation through the use variable selection priors, i.e. spike and slab priors, have been developed in the last three decades (Miller, 2002). This approach was first proposed by Mitchell and Beauchamp (1988) for BVS with normal linear regression models where the spike and slab distribution for a parameter in the model is defined as a mixture of a point mass at 0 and a diffuse uniform distribution elsewhere. An alternative spike and slab procedure in which the subset selection is derived from a prior of a hierarchical normal mixture model were proposed by George and McCulloch (1993). Their methods, however, require to choose the tuning factors that specify the two variances in the normal mixture models in the first stage of the hierarchical prior. In this paper, we focus on the prior specification proposed by Kuo and Mallick (1998) which is based on a binary indicator variable embedded into the model's mean structure that determines whether the unknown parameter belongs to the slab or spike part of the prior.

Let Y_{ij} be a longitudinal outcome of interest for the i^{th} subject at time j. We further assume that $E(Y_{ij}) = \lambda_{ij}$, and we formulate a generalized linear mixed effects model (Molenberghs and Verbeke 2005) for Y_{ij} so that $g(\lambda_{ij}) = \eta_{ij} = \Sigma_{k=0}^{p}\beta_k X_{ik} + \Sigma_{q=0}^{Q} b_{iq} Z_{iq}$ where, g is a known link function, $\beta_0, \ldots, \beta p$ are the unknown fixed parameters and $b_{i0}, \ldots, b_{iQ}$ are subject specific random effects. Our aim is to select a subset of covariates from $X_1, \ldots, X_p$ that will be included in the model. Let γ_k be an indicator variable, $k = 1, \ldots, p$ such that $\gamma_k = 1$ if predictor X_k, is included in the model and $\gamma_k = 0$ otherwise and let $\theta_k = \gamma_k \beta_k$. Then, for a random intercept model (for brevity, we will simply write b_{i0} as b_i), the linear predictor is given by

$$\eta_{ij} = \beta_0 + \sum_{k=1}^{p} \gamma_k \beta_k X_{ik} + b_i = \beta_0 + \sum_{k=1}^{p} \theta_k X_{ik} + b_i, \tag{1}$$

where b_i is a normal distributed random intercept, $b_i \sim \mathrm{N}(0, \sigma_b^2)$, used to account for a possible correlation among the observations of the same subject (Molenberghs and Verbeke 2005). Note that we assumed all the competing models to include an intercept term β_0.

For a set of p covariates there is a set of 2^p candidates model $M_1, \ldots, M_{2^p}$ determined by the configuration of the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. The variable selection component within the model formulated in (1) entails estimating $\gamma_1, \ldots, \gamma_p$ while the parameter estimation component (accounting for model uncertainty) entails estimating $\theta_1, \ldots, \theta_p$. This can be done using MCMC simulation successively sampling from the joint posterior distribution of γ_k and θ_k.

Kuo and Mallick (1998) assumed that the indicators, γ_k, and the parameters under selection, β_k, are independent a priori, $P(\gamma_k, \beta_k) = P(\gamma_k)P(\beta_k)$. The usual approach is to assume $\beta_k, k = 1, \ldots, p$ are chosen independently, each with a normal prior $\mathrm{N}(\mu_{\beta_k}, \sigma^2_{\beta_k})$. The choice of μ_{β_k} and $\sigma^2_{\beta_k}$ reflects our prior belief about the mean and variance of β_k in the full model with all $\gamma_k = 1$. Kuo and Mallick (1998) treat $\sigma^2_{\beta_k}$ as fixed and given a moderately large constant value. Specifically, they suggest choosing a value in the range $[1/2, 4]$ for σ_{β_k}. For a cross-sectional linear regression case, Geweke (1996) assumed $\sigma^2_{\beta_k}$ as fixed and set its value to be equal to the changes of y_i divided by the changes of X_{ik} for each k. On the other hand, O'Hara and Sillanpää (2009) recommend to treat $\sigma^2_{\beta_k}$ as random rather than fixed and to place a hyperprior distribution, such as a

$$\sigma^{-2}_{\beta_k} \sim \Gamma(a, b) \quad \text{or} \quad \sigma_{\beta_k} \sim U(a, b).$$

O'Hara and Sillanpää (2009) advocate that hierarchical variance will pull the posteriors for the β_k's towards the right part of the parameter space, so that when $\gamma_k = 0, \beta_k$ will be sampled from close to the correct part of the parameter space, and will greatly improve mixing. The inclusion parameters γ_k are assumed to follow a Bernoulli distribution, that is

$$\gamma_k \sim B(\pi_k).$$

The probability π_k is called the inclusion probability and it reflects the preference for including the kth predictor in the model and it is assumed (Scott et al., 2010) that

$$\pi_k \sim U(0,1).$$

3.1.2. Posterior Model Probability

A MCMC simulation is used to estimate the model and, as mentioned above, variable (and model) selection is done based on the configuration of $\boldsymbol{\gamma}$. The posterior mean of γ_k, obtained through MCMC simulation, represents the posterior inclusion probability of β_k in the model (O'Hara and Sillanpää, 2009). Furthermore, the computation of the posterior model probability for each of the candidate models is achieved by defining an appropriate transformation function on the indicators γ_k that uniquely identify all the possible models (Ntzoufras, 2011; Kasim et al., 2012; Otava et al., 2014),

$$G = 1 + \sum_{k=1}^{p} \gamma_k 2^{(k-1)}. \tag{2}$$

Denoting the candidate models by $g_r, r = 0, ..., R$, $R = 2^{p-1}$, the posterior probability of $G = r+1$ defines uniquely the posterior probability of a specific model $g_r, P(g_r|Data)$, and it can be estimated by taking the proportion of times model g_r is selected over the total number of visited models during the MCMC simulation,

$$\hat{P}(g_r|\text{Data}) = \frac{1}{L}\sum_{\ell=1}^{L} I(g^\ell = g_r), \tag{3}$$

where g^ℓ is the model visited in iteration ℓ of the Markov chain, $I()$ is an indicator function which equal to 1 if $g^\ell = g_r$ and zero otherwise and L is the total number of MCMC iterations.

3.1.3. Model Averaged Estimates

One characteristic of the BVS approach is the ability to provide a model averaged parameter estimate as part of the MCMC simulation. The posterior mean of the parameter θ_k from the MCMC simulation is a model averaged estimate, weighted by the posterior probabilities of the models, i.e. $\overline{\theta}_k = 1/L \sum_{\ell=1}^{L} \hat{\theta}_k^\ell$, where in each iteration ℓ, one model g_r is considered and estimate $\hat{\theta}_k^\ell$ is obtained. In other words, each model contributes to the final estimate of the parameters to the extent determined by the posterior probability of that model being the true underlying model.

3.2. Model Formulation for Anopheles *Mosquitoes Count Data*

3.2.1. Hierarchical Bayesian Model

Let Y_{ij} be the mosquitoes count for the ith household at the jth time point, $i = 1, \ldots, 40$ and $j = 1, \ldots, 6$. We considered a hierarchical generalized linear mixed effects model. For the first stage of the model, we assume a Poisson likelihood for the count data, that is, $Y_{ij}|b_i \sim \text{Poisson}(\lambda_{ij})$ with linear predictor given by

$$\eta_{ij} = \log(\lambda_{ij}) = \beta_1 + \beta_2 I_i + (\beta_3 + \beta_4 I_i)t_{ij} + b_i, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ is the parameter vector of the "fixed effects", and the random effect b_i is a household specific parameter which captures a possible correlation among the observations from the same household over time, and I_i is an indicator variable which takes value 1 for a household from "at risk" (resettled area) village and 0 for a household from a control village. Hence, the linear predictor is given by

$$\eta_{ij} = \begin{cases} (\beta_1 + \beta_2) + (\beta_3 + \beta_4)t_{ij} + b_i & \text{if at risk,} \\ \beta_1 + \beta_3 t_{ij} + b_i & \text{if control.} \end{cases}$$

We denote the model formulated in (4) by g_3. The following priors and hyperpriors are assumed for the unknown parameters of the model:

$$\begin{aligned} \beta_\ell &\sim \mathrm{N}(0, \tau_{\beta_\ell}^{-1}), & \tau_{\beta_\ell} &\sim \Gamma(1,1),\ \ell = 1, ..., 4, \\ b_i &\sim \mathrm{N}(0, \tau_b^{-1}), & \tau_b &\sim \Gamma(10^{-3}, 10^{-3}). \end{aligned} \tag{5}$$

3.2.2. Bayesian Variable Selection Model

We considered four competing models for which the mean structures are presented in Table 1. Model g_3 is the full model discussed in the previous section. Model g_0 (the null model) assumes that both village types share the same mean structure (i.e., identical intercept and slope), model g_1 assumes identical slopes but different intercept and model g_2 assumes that both village types shares the same intercept but differ in their time trend.

Table 1. Configuration of $\boldsymbol{\gamma}$, G and the mean structure of all models.

Model	mean structure		$\boldsymbol{\gamma}$	G
g_0	$\eta_{ij} = \begin{cases} \beta_1 + \beta_3 t_{ij} & \text{if at risk,} \\ \beta_1 + \beta_3 t_{ij} & \text{if control.} \end{cases}$		(0,0)	1
g_1	$\eta_{ij} = \begin{cases} (\beta_1 + \beta_2) + \beta_3 t_{ij} & \text{if at risk,} \\ \beta_1 + \beta_3 t_{ij} & \text{if control.} \end{cases}$		(1,0)	2
g_3	$\eta_{ij} = \begin{cases} \beta_1 + (\beta_3 + \beta_4) t_{ij} & \text{if at risk,} \\ \beta_1 + \beta_3 t_{ij} & \text{if control.} \end{cases}$		(0,1)	3
g_3	$\eta_{ij} = \begin{cases} (\beta_1 + \beta_2) + (\beta_3 + \beta_4) t_{ij} & \text{if at risk,} \\ \beta_1 + \beta_3 t_{ij} & \text{if control.} \end{cases}$		(1,1)	4

Next, we formulate a BVS model that allows us to select between the competing models. The linear predictor for the model is given by

$$\eta_{ij} = \log(\lambda_{ij}) = \beta_1 + \gamma_1 \beta_2 I_i + (\beta_3 + \gamma_2 \beta_4 I_i) t_{ij} + b_i. \tag{6}$$

Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ be an indicator vector, for which the the configuration of $\boldsymbol{\gamma}$, defines uniquely one of the four possible model presented in Table 1,

$$\boldsymbol{\gamma} = \begin{cases} (0,0), & \text{neither } \beta_2 \text{ nor } \beta_4 \text{ are included in the model,} & g_0, \\ (1,0), & \text{only } \beta_2 \text{ is included in the model,} & g_1, \\ (0,1), & \text{only } \beta_4 \text{ is included in the model,} & g_2, \\ (1,1), & \text{both } \beta_2 \text{ and } \beta_4 \text{ are included in the model,} & g_3. \end{cases} \tag{7}$$

To complete the specification of the BVS model, we specify the same priors and hyperpriors defined in Equation (5) for $\beta_\ell, \ell = 1, ..., 4$, b_i and for the inclusion parameters we assume

$$\begin{aligned} \gamma_m &\sim Bernoulli(\pi_m), \\ \pi_m &\sim U(0,1),\ m = 1, 2. \end{aligned} \tag{8}$$

Alternatively, a $Beta(a, b)$ prior can be assumed for π_m, where the value of a and b are determined by the available prior information regarding the candidate models (see section 7 of the supplementary appendix).

For a given configuration of $\boldsymbol{\gamma}$, both the transformation function G given in (2) and the model posterior probability can be calculated. For example, the posterior probability of the null model (see Table 1) is given by

$$P(\boldsymbol{\gamma} = (0,0)|Data) = P(G = 1|Data) = P(g_0|Data).$$

The posterior probability of the other models can be calculated in the same way. Note that the inclusion probabilities of β_2 and β_4 are $P(\gamma_1 = 1)$ and $P(\gamma_2 = 1)$, respectively. The posterior means for the parameter vector $\boldsymbol{\beta}$ will be weighted by posterior probabilities and summed to obtain the model averaged estimate as described in Section 3.1.3.

3.3. Information Criteria Based Model Averaging

In the previous section, the posterior model probability was estimated using the distribution of the inclusion parameters $\boldsymbol{\gamma}$ over the MCMC simulation. Alternatively, one can fit all candidate models and calculate the model posterior probability using model averaging techniques. Burnham and Anderson (2003) addresses the problem of accounting model uncertainty through an information criteria (IC) approach. Let $g_0, \ldots, g_R$ be the collection of all candidate models and let IC_r denote an information criterion associated with the rth model. Our aim is to calculate the posterior probability for each candidate model. The marginal likelihood of rth model can be approximated by (Burnham and Anderson, 2003; Whitney and Ryan, 2009),

$$P(\text{Data}|g_r) = exp(-\frac{1}{2}\Delta IC_r), \tag{9}$$

where $\Delta IC_r = IC_r - IC_{min}$, with $IC_{min} = min_{r=0,\ldots,R} IC_r$. Assuming equal prior probabilities for all models, $P(g_r) = 1/R$, we get

$$P(g_r|\text{Data}) = \frac{exp(-\frac{1}{2}\Delta IC_r)}{\sum_{\ell=0}^{R} exp(-\frac{1}{2}\Delta IC_\ell)}. \tag{10}$$

Different information criteria can be used to calculate the posterior model probability in Equation (10). For example, Whitney and Ryan (2009) used the BIC while Lin et al. (2012) and Otava et al. (2017) used both AIC and BIC. Within the hierarchical Bayesian framework, the DIC (Spiegelhalter et al., 2002) can be used.

If model selection is of primary interest, one can select the model for which $P(g_r|Data)$ is maximal which is equivalent to selection of the model with the smallest IC. However, relaying on a single "best" model is often unsatisfactory because "the best" model is often highly variable. Alternatively, one can weight the parameter estimates obtained from each of candidate models by their corresponding posterior model probabilities, $P(g_r|Data)$. The model averaged estimate for a parameter of interest, say θ, can be obtained by (Burnham and Anderson 2003, Claeskens and Hjort 2008)

$$\hat{\bar{\theta}} = \sum_{r=0}^{R} P(g_r|Data)\hat{\theta}_r, \tag{11}$$

and the unconditional standard error given by

$$\hat{Var}(\hat{\bar{\theta}}) = \left[\sum_{r=0}^{R} P(g_r|Data)\sqrt{\hat{Var}(\hat{\theta}_r|g_r) + (\hat{\theta}_r - \hat{\bar{\theta}})^2}\right]^2, \tag{12}$$

where, $\hat{\bar{\theta}}$ denotes a model averaged estimate of θ and $\hat{\theta}_r$ is the estimate of θ obtained from model g_r. The $(1-\alpha)100\%$ unconditional confidence interval is then given by the endpoints $\hat{\bar{\theta}} \pm Z_{1-\alpha/2}\hat{se}(\hat{\bar{\theta}})$, where $\hat{se}(\hat{\bar{\theta}}) = \sqrt{\hat{Var}(\hat{\bar{\theta}})}$. This type of model averaging is applicable for prediction problems or in cases where a particular parameter (e.g., β_1 and β_3 in the hierarchical model (4) for the *Anopheles* mosquitoes count data) occurs in all the models in the set. For the parameters β_k associated with predictor variable X_k that appear only in some of the R possible models, the model averaged estimator, denoted by $\hat{\bar{\beta}}_k$, is given by (Burnham and Anderson, 2003)

$$\hat{\bar{\beta}}_k = \frac{\sum_{r=0}^{R} P(g_r|Data)I_k(g_r)\hat{\beta}_{k,r}}{\sum_{r=0}^{R} P(g_r|Data)I_k(g_r)}, \tag{13}$$

where,

$$I_k(g_r) = \begin{cases} 1, & \text{if predictor } X_k \text{ in the model } g_r, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\hat{\beta}_{k,r}$ is the estimate of β_k based on model g_r. Burnham and Anderson (2003) termed this estimate a "natural average" as it only averages β_k over models where the unknown β_k parameter appears. Note that, the equivalent for the BVS model of $I_k(g_r)$ is the inclusion parameters γ_k. However, in contrast with $I_k(g_r)$, the inclusion parameters in BVS are consider latent random variables which are estimated within the MCMC simulation and for which the posterior probability can be estimated according to (3). The main drawback of the approach discussed in the section is the necessity to fit all candidate models in order to compute the model posterior probabilities.

4. Application to *Anopheles* Mosquito Count Data

The four models discussed in Section 3.2.2 were fitted to the *Anopheles* mosquito count data and for each one of the models, the posterior probability were calculated using AICc (Sugiura, 1978), BIC, and DIC. The BVS model was fitted as well. For the Bayesian models, the `R` - package `runjags` (Denwood, 2016) and the `JAGS` software (Plummer et al., 2003) were used to fit the models. Three chains of 60,000 iterations, from which the first 30,000 were considered as the burn-in period and no thinning, were used to estimate the posterior model probabilities and posterior means for the unknown parameters. Model diagnostics, presented in Section 2 of the supplementary appendix of the paper, indicate convergence of the parameter of interest. To calculate the AICc and BIC, each of the competing mixed effect models were fitted using the `R` - package `lme4` (Bates et al., 2015).

4.1. Model Posterior Probabilities

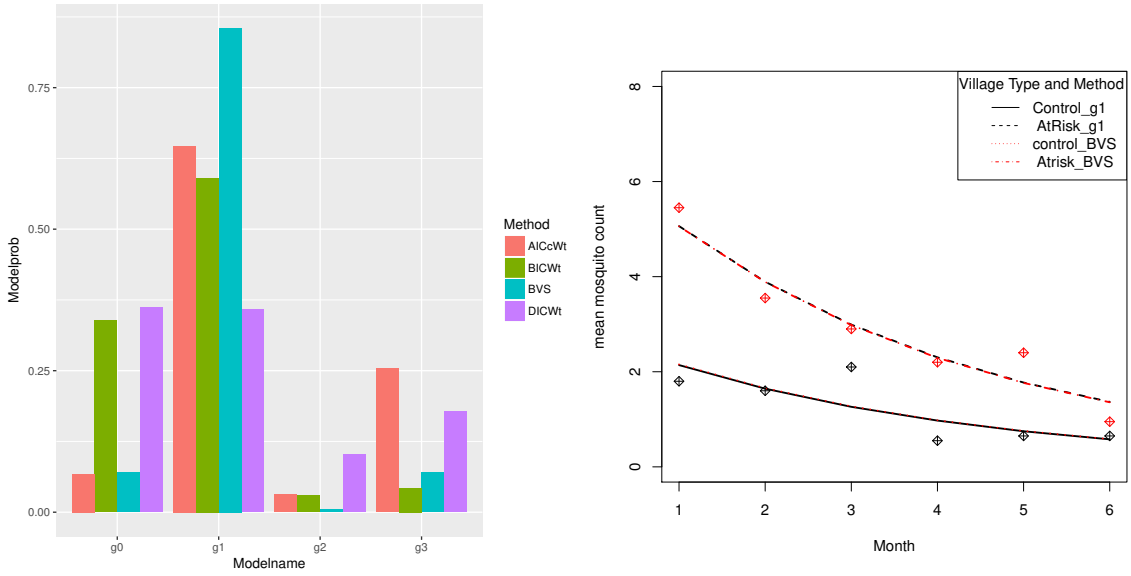

Figure 2. The *Anopheles* mosquito count data. Left panel: Posterior model probability for $g_0, g_1, g_2,$ and g_3 computed using AICc, BIC, DIC, and BVS method. Right panel: Sample means, posterior means of the BVS model, and model g_1 for the two groups.

The left panel of Figure 2 shows the posterior model probability of the set of candidate models computed using all approaches mentioned above. Clearly, model g_1 is selected as "the best" model for the data by the BVS, AICc and BIC approaches with $P_{BVS}(g_1|\text{Data}) = 0.82$, $P_{AICc}(g_1|\text{Data}) = 0.65$, and $P_{BIC}(g_1|\text{Data}) = 0.59$, respectively. The model posterior probabilities obtained using the DIC supports equally two models, model g_1 and model g_0, with $P_{DIC}(g_1|\text{Data}) = 0.37$ and $P_{DIC}(g_0|\text{Data}) = 0.37$, respectively. The right panel of Figure 2 shows the data, posterior means obtained for the BVS model and the posterior means for g_1. For the BVS model, at each MCMC

iteration one of the 4 candidate models are fitted to the data and the posterior mean is the average across all iterations. Due to the fact that the other competing models to g_1 have relatively small model posterior probabilities and the prediction is done conditional on the random effect, the effect of model averaging is not clearly visible in the figure.

4.2. Posterior Means for the Unknown Parameters

Table 2, presents the model averaged parameter estimates and their corresponding unconditional standard errors and confidence/credible intervals computed using the approaches described in Section 3. Parameter estimates, conditional standard errors and conditional credible interval obtained for model g_1 are presented for comparison as well. Figure 3 visualizes the result presented in Table 2. Depending on the parameter of interest, difference in the approaches on the point estimate, standard errors and confidence/credible intervals are observed. For example, the model averaged point estimates of the IC based approaches (except the AICc) for β_1 are about 17% larger than the estimate from BVS, the unconditional standard errors of the IC based approaches for β_1 are 12% - 37% larger than the BVS standard error, the model averaged point estimate of the IC based approaches for β_2 (the parameter under selection) are 10% - 19% larger than the estimate from BVS, and the unconditional confidence intervals of the IC based approaches for β_1 and β_4 are wider than credible intervals of BVS.

Table 2. Model averaged parameter estimates for the *Anopheles* mosquito count data for AICc, BIC, DIC, BVS, and model g_1 that was estimated within the hierarchical Bayesian modelling framework. Note that `lme4` doesn't provide standard errors of variance components.

	Model parameters																			
	β_1				β_2				β_3				β_4				σ_b			
Method	Mean	SD	LCL	UCL	Mean	SD	LCL	UCL	Mean	SD	LCL	UCL	Mean	SD	LCL	UCL	Mean	SD	LCL	UCL
AICcWt*	0.77	0.26	0.26	1.29	0.81	0.31	0.21	1.41	-0.26	0.04	-0.33	-0.19	-0.02	0.06	-0.14	0.10	0.85			
BICWt*	0.89	0.32	0.27	1.51	0.79	0.30	0.21	1.37	-0.27	0.03	-0.32	-0.21	0.00	0.07	-0.13	0.13	0.87			
DICWt	0.89	0.28	0.33	1.44	0.75	0.31	0.15	1.35	-0.26	0.04	-0.33	-0.19	0.00	0.06	-0.13	0.13	0.93	0.14	0.66	1.21
BVS	0.76	0.24	0.29	1.25	0.68	0.33	0.00	1.21	-0.26	0.03	-0.32	-0.20	0.00	0.02	-0.04	0.02	0.90	0.14	0.65	1.17
g_1	0.73	0.23	0.28	1.17	0.73	0.30	0.16	1.34	-0.26	0.03	-0.32	-0.21					0.90	0.13	0.65	1.16

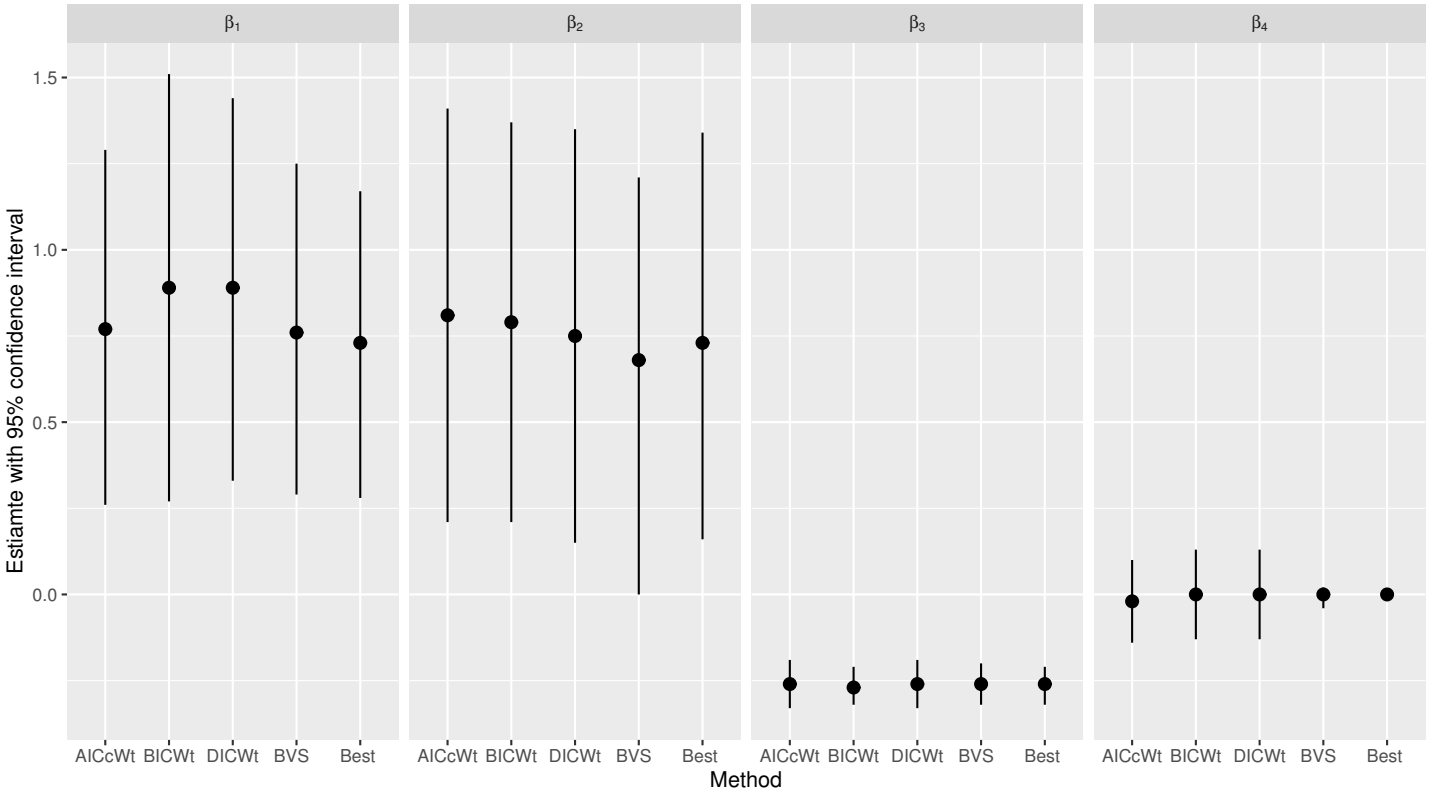

Figure 3. Model averaged parameter estimates for the unknown parameters with their 95% confidence/credible intervals. The parameter estimates with 95% credible intervals obtained from "the best" model (g_1) are presented as well.

In this Section, we assume a linear trend over time for the mean structure which leads to 4 possible models. A comparison of the BVS and IC based approaches using an unstructured mean instead is presented in Section 3 of the Supplementary appendix of this manuscript. Note that using an unstructured mean implies that 64 models should be fitted if the posterior model probabilities are calculated based on IC while only one model is fitted if the BVS model is used. Further, we illustrated the application of BVS approach for the random intercept and slope model and compare the results with the random intercept model reported in this section (see Section 4 of the supplementary material). Comparison of the two models indicates that, in general, the simpler model, the random intercept model should be preferred. Both models indicate that model g_1 has the best goodness of fit to the data.

5. Simulation Study

5.1. Simulation Setting

A simulation study was conducted in order to compare the performance of the BVS model with IC based methods using a longitudinal data with six time points in which 20 subjects were measured from two groups: half of the subjects were assumed to belong to each of the group. Data were generated according to the model

$$Y_{ij} \sim \text{Poission}(\lambda_{ij}),\ i = 1, \ldots, 20,\ j = 1, \ldots, 6.$$

Similar to the model specified in Equation (4), we assume a log-link function and the following linear predictor

$$log(\lambda_{ij}) = (\beta_1 + \beta_2 z_i) + (\beta_3 + \beta_4 z_i) t_{ij} + b_i.$$

Here, z_i is an indicator variable that take the value of 1 for the first group and zero for the second group. The random intercept is assumed to follow a normal distribution, $b_i \sim \text{N}(0, \sigma_b^2)$. Two values for the standard deviation were used, $\sigma_b = 0.1, 1$ corresponding to weak and strong intra-cluster correlation, respectively. Three set of true values, shown in Table 3, were used for the regression coefficients to generate the data. The resulting linear

predictor are shown in Figure S6 of the supplementary appendix for the manuscript. We vary the true value of the parameters under selection, β_2 and β_4, in order to study the behaviour of selection of 'weak (setting 1)', 'moderate (setting 2)' and 'strong (setting 3)' regressors relative to the noise. In total, 500 data sets were generated for

Table 3. True values for the regression parameters used for data generation.

	Model	β_1	β_2	β_3	β_4
Setting 1	g_0	2	0	-0.5	0
	g_1	2	0.2	-0.5	0
	g_2	2	0	-0.5	-0.75
	g_3	2	0.2	-0.5	-0.75
Setting 2	g_0	2	0	-0.5	0
	g_1	2	-2	-0.5	0
	g_2	2	0	-0.5	-0.5
	g_3	2	-2	-0.5	-0.5
Setting 3	g_0	2	0	-0.5	0
	g_1	2	2	-0.5	0
	g_2	2	0	-0.5	-1
	g_3	2	2	-0.5	-1

$3 \times 4 \times 2$ combinations of setting, model and σ_b, respectively.

For each simulated data sets, the posterior model probabilities, $\overline{\mathrm{P}}(g_r|\mathrm{Data})$, were computed according to the BVS, AICc, BIC and DIC methods. For the IC based methods, including the DIC approach, model averaged parameter estimates and confidence intervals were calculated according to Equation (13) and the formulation specified in Section 3.3, respectively. For the BVS model, the model average is the posterior mean obtained for the model (with the corresponding credible interval). The methods were evaluated based on three criteria: mean square error and achieved confidence interval coverage, and the correct identification of the true underlying model.

A second simulation study was conducted in order to explore the dependency of the posterior model probability on the specification of priors for the regression coefficients. Note that, in our BVS model formulation for *Anopheles* mosquito count data model (See Section 3.2.1) we assume the unknown parameters β_ℓ are drawn from $\mathrm{N}(0, \sigma^2_{\beta_\ell})$, where, $\sigma^2_{\beta_\ell}$ is a variance parameter to be estimated. Our aim is to investigate the impact of assuming that the variance, $\sigma^2_{\beta_\ell}$, is a fixed parameter compared with the setting in which a hyperprior is specified for $\sigma^{-2}_{\beta_\ell}$. For the first setting it was assumed that $\sigma^2_{\beta_\ell} = 0.1, 0.2, \ldots, 10$ while for the second setting it was assumed an inverse-gamma distribution for the variance parameter,

$$\sigma^{-2}_{\beta_\ell} \sim \Gamma(\alpha, \alpha),$$

with $\alpha = 0.001, 0.01, 0.1, \text{and}, 1$.

We extended the above simulation study for $n = 40$. The simulation setting and result are discussed in details in Section 8 of the supplementary appendix.

5.2. Simulation Results

5.2.1. Model Selection

We first evaluate the performance of the different approaches discussed in terms of their correct selection rate of the true data generating model. The results are presented in Figure 4 and Figure 5. Figure 4 shows the posterior model probability obtained for the true data generating model and Figure 5 shows the correct selection rate of the true model. Per simulation, the first is estimated by (3) for the BVS model and by (10) for the IC and averaged

across the 500 simulations and the second estimated by

$$\hat{\pi}(g_r) = \frac{1}{500}\sum_{s=1}^{500} I(g^s = g_r), \tag{14}$$

where g^s is the model with the highest posterior model probability ("best" model) for the s^{th} simulated data and $I()$ is an indicator function which equal to 1 if $g^s = g_r$ and zero otherwise (note that g_r is the true model used to generate the data).

Since the competing models $g_0 - g_2$ arise depending on whether β_2 and/or β_4 are excluded from the full model g_3, as expected, the correct selection of the true underlying model is highly dependent on the magnitude of these two parameters relative to noise. For strong effect size (i.e., setting 3 with $\beta_2 = 2$ and/or $\beta_4 = -1$) all approaches performs well, particularly in selecting the complex model (model g_3). As shown in Figure 4, for $\sigma_b = 0.1$, the mean posterior model probability for model g_3 is equal to one for all approaches (i.e., all approaches select model g_3 100% of the time as "the best" model) when the data is generated from model g_3. In setting 2, we used a strong intercept term ($\beta_2 = -2$) with a weak slope term ($\beta_4 = -0.5$). For this setting, the BIC method outperform the other methods when the data are generated using models g_0, g_1 and g_2 while failing to select the most complicated model. When the data are generated under g_3 both Bayesian methods outperform the AIC and BIC approaches (for both $\sigma_b = 0.1$ and 1). In setting 1, we used a weak intercept term ($\beta_2 = 0.2$) with a moderate slope term ($\beta_4 = -0.75$) and all approaches show poor performance in detecting models which include these parameters (model g_1 and g_3). Exact values for the results presented in Figure 4 and 5 are presented in Section 5 of the supplementary appendix for the manuscript.

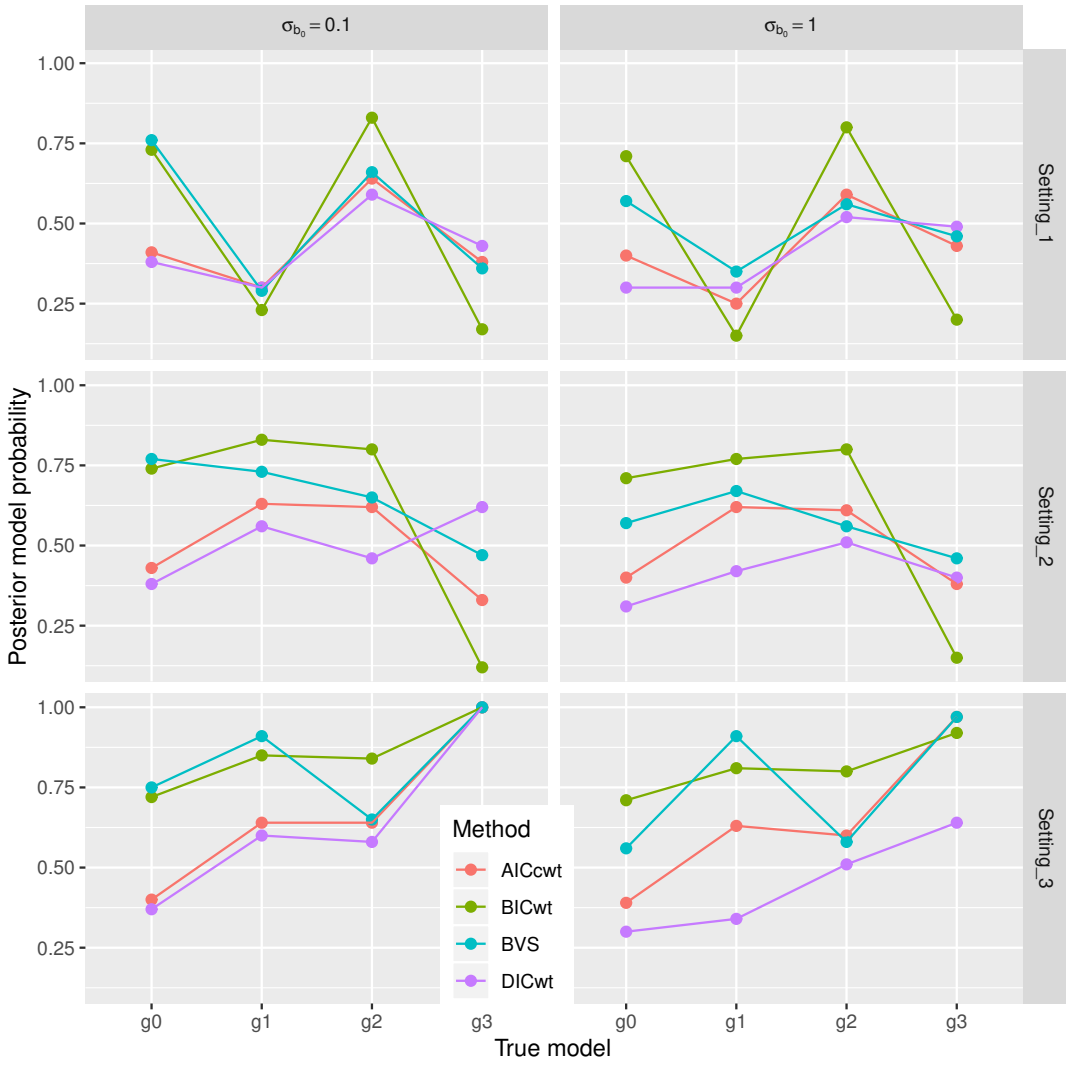

Figure 4. Mean posterior model probability for the true underlying model averaged over 500 simulations.

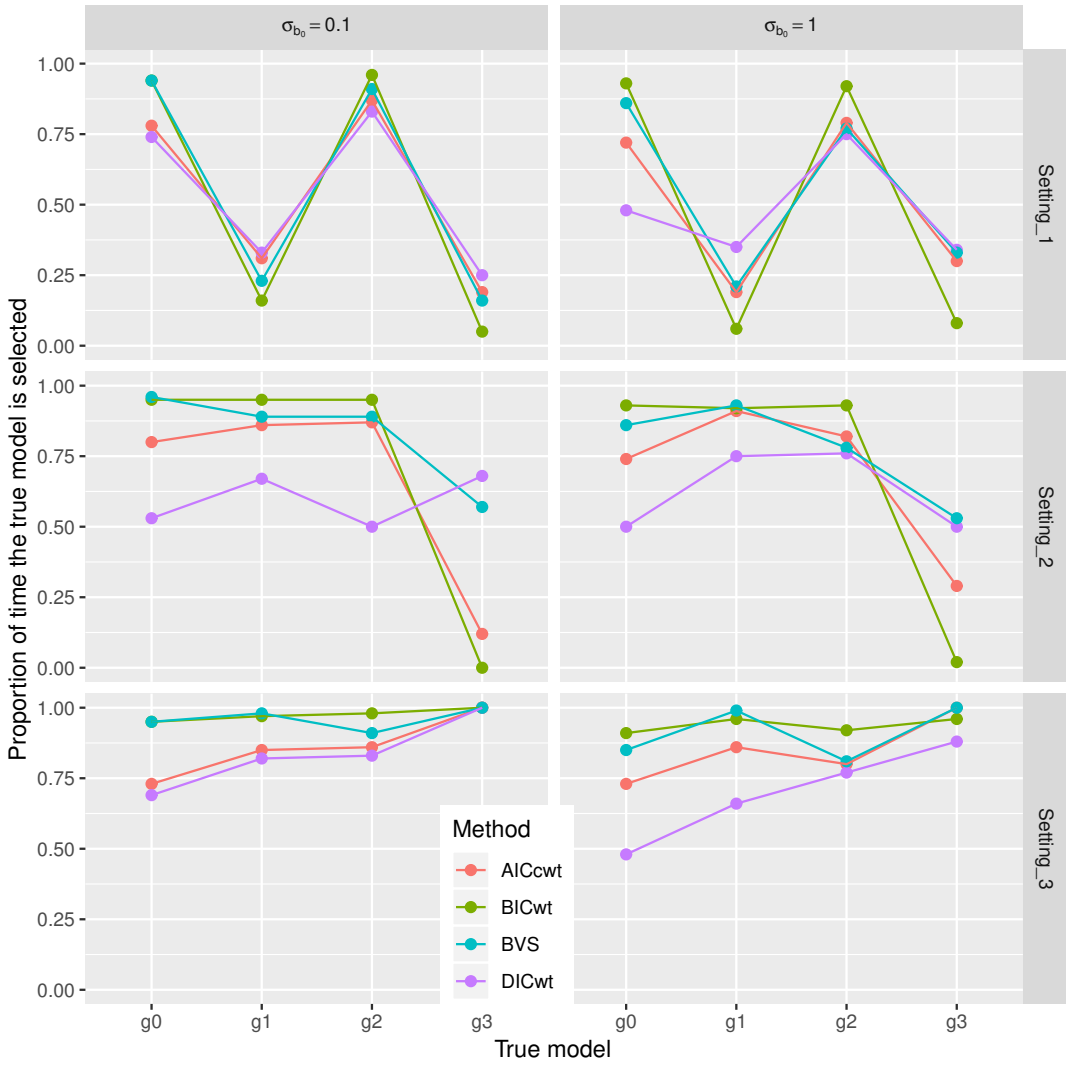

Figure 5. The proportion of times the true underlying model was selected as "best" model according to value of posterior probability over 500 simulations.

The results for the simulation setting with $n = 40$ indicate the same patterns. Further results are presented in Section 8 of the supplementary appendix. A simulation study for random intercept and slope model was conducted as well. The results of the simulation study lead, in general, for the same conclusions as reported in this section. An elaborate discussion about the simulation setting and result is presented in detail in Section 9 of the supplementary appendix.

5.2.2. Estimation - Model Averaging

Figure 8 - 6 shows the parameter estimates, standard error (SD), achieved confidence/credible interval coverage and associated mean squared error (MSE) of the parameters estimates for the simulation study using true values from setting 1 - 3. Overall, all approaches lead to a similar model averaged parameter estimates and comparable confidence/credible interval coverage. Compared to the IC based approaches, BVS leads to more precise (lower standard error and lower mean squared error) parameter estimates for all settings and data generating models. In particular, the standard error and mean squared error estimate for β_2 obtained from BVS is consistently lower than the estimates obtained from the IC based approaches. Exact values for the results presented in Figure 8 - 6 are presented in Section 5 of the supplementary appendix for the manuscript.

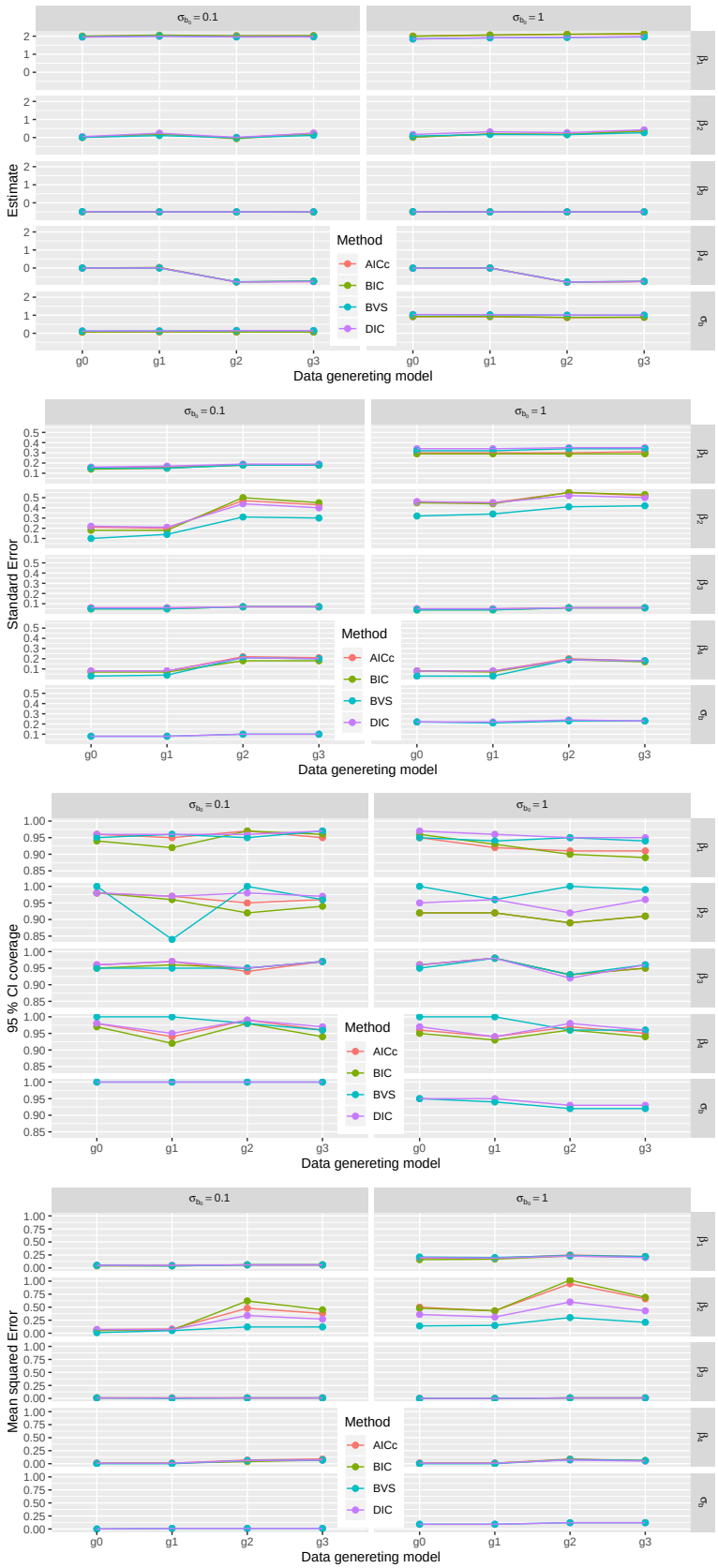

Figure 6. Results of simulation study for Setting 1 for 500 simulations. Top to bottom: model averaged parameter estimate, standard error of parameter estimates, 95% CI coverage, and mean squared error, respectively.

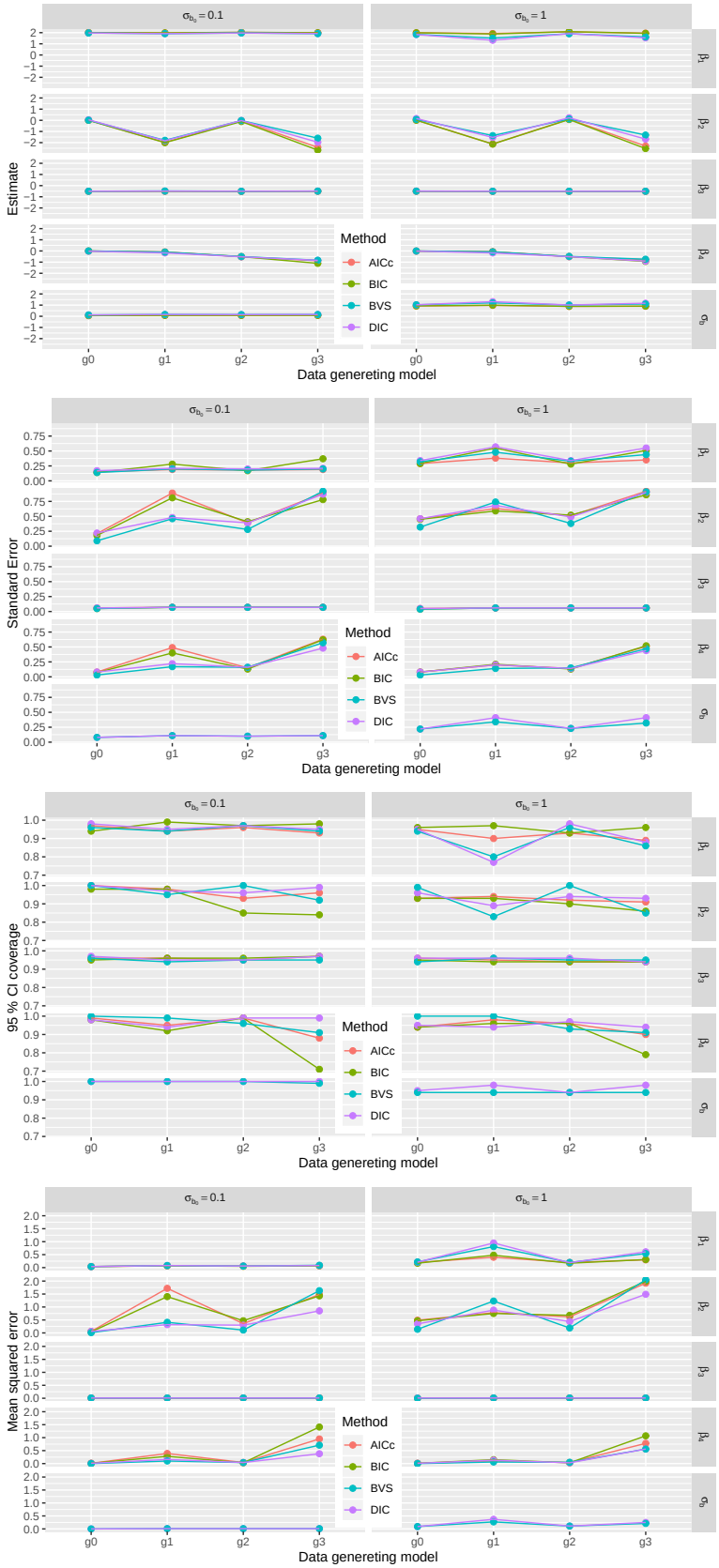

Figure 7. Results of simulation study for Setting 2 for 500 simulations. Top to bottom: model averaged parameter estimate, standard error of parameter estimates, 95% CI coverage, and mean squared error, respectively.

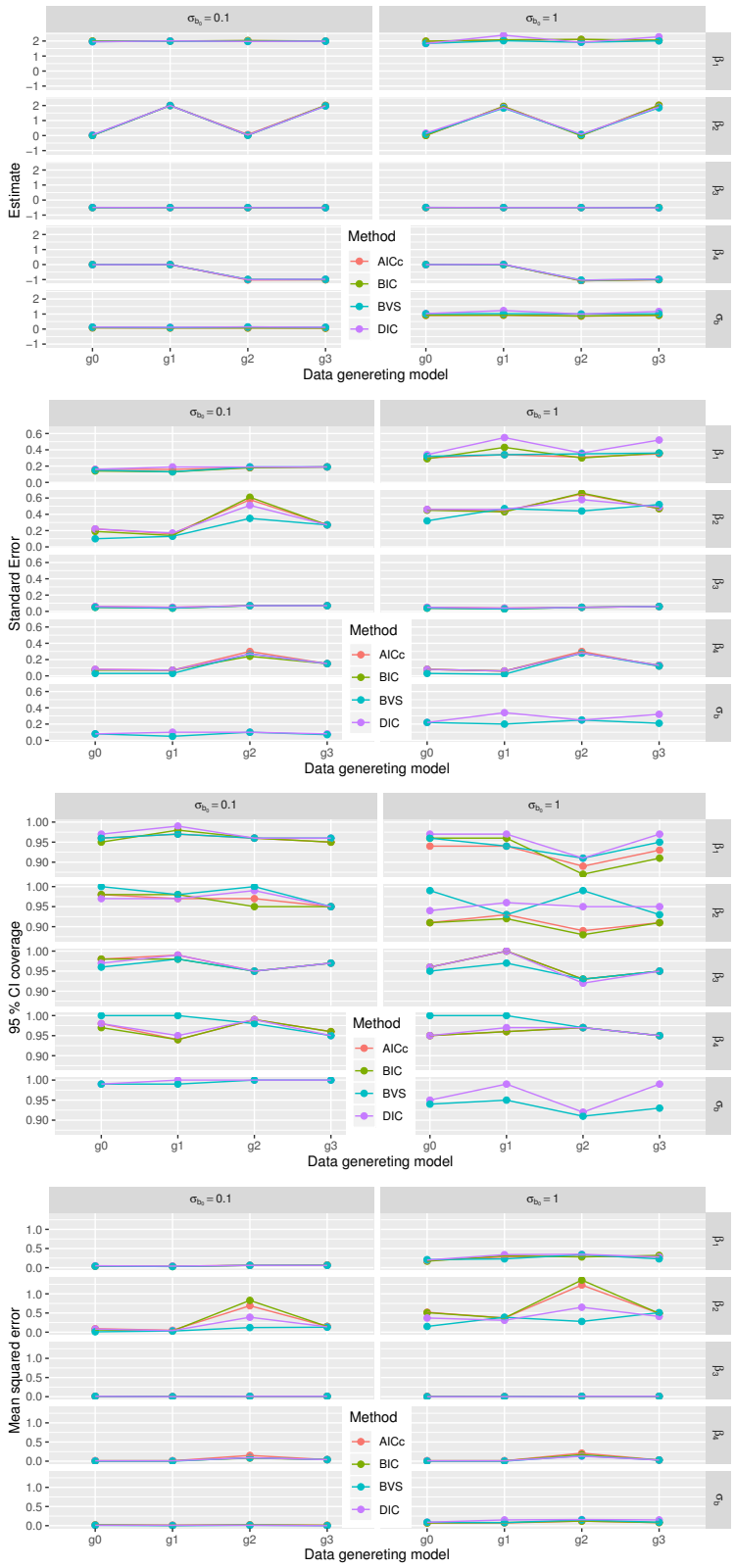

Figure 8. Results of simulation study for Setting 3 for 500 simulations. Top to bottom: model averaged parameter estimate, standard error of parameter estimates, 95% CI coverage, and mean squared error, respectively.

5.2.3. Prior Specification for the Regression Coefficients

In this section, we evaluate the dependence of the posterior model probability computed using BVS on the variance of the prior distribution of the unknown parameters. The results of the simulation study are shown in Figure 9. When we treat $\sigma^2_{\beta_\ell}$ as a fixed value the resulting posterior model probability from BVS is dependent on the constant value assumed for σ_{β_ℓ}. For model g_0 - g_2, assuming a higher value for σ_{β_ℓ} leads to a higher posterior

model probability for true data generating model and the result is comparable to the one where a hyper-prior is assumed for $\sigma^2_{\beta_\ell}$. Conversely, assuming large values for σ_{β_ℓ} leads to wrong selection of the true model when the true data generating model is model g_3. When we treat $\sigma^{-2}_{\beta_\ell}$ as a random effect coming from a $\Gamma(\alpha, \alpha)$ distribution, the resulting posterior model probability is robust against the choice of α value.

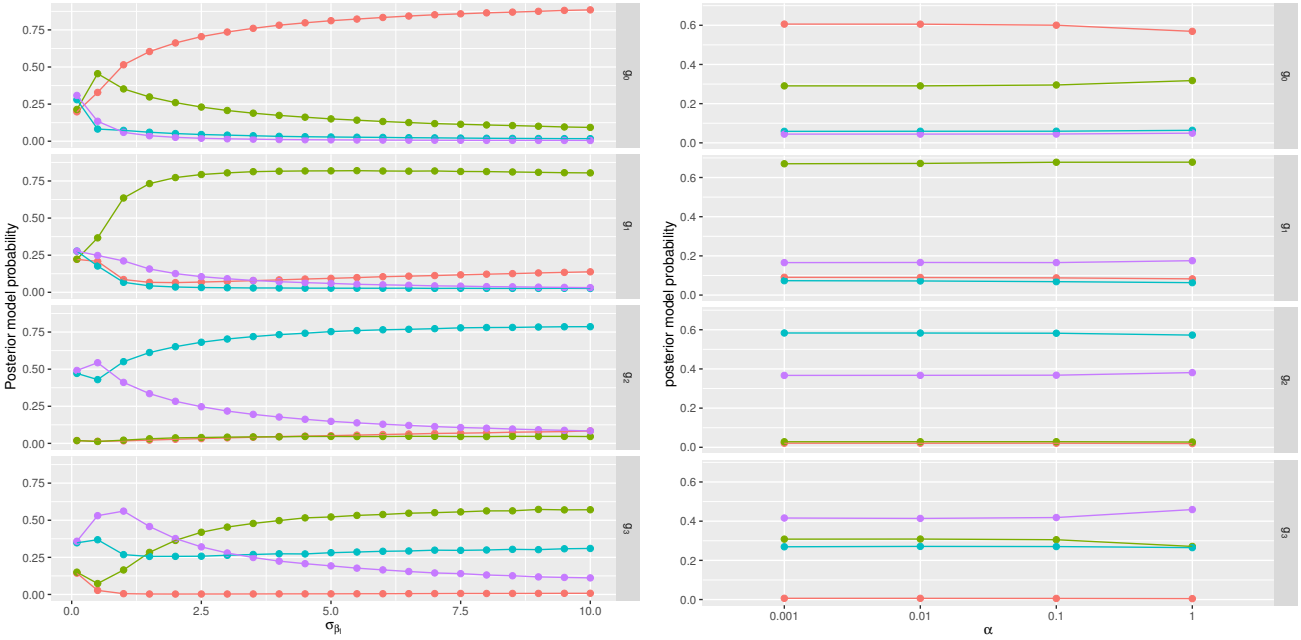

Figure 9. Dependency of posterior model probability on the prior specification for unknown parameters under selection by the true models. Left panels: $\beta_\ell \sim \mathrm{N}(0, \sigma^2_{\beta_\ell})$ and σ_{β_ℓ} is fixed as described in Section 5.1. Right panels: $\beta_\ell \sim \mathrm{N}(0, \sigma^2_{\beta_\ell})$ and $\sigma^{-2}_{\beta_\ell} \sim \Gamma(\alpha, \alpha)$]. Red line: g_0, green line: g_1, blue line: g_2, and pink line: g_3.

6. Discussion

The manuscript discusses the Bayesian variable selection method for the model averaging in context of longitudinal study. It demonstrates its performance using real world data and simulation study. A comparison with alternative methods based on information criteria (IC) using AICc, BIC, and DIC was conducted using both a real data set and simulation study. The performance of all approaches was comparable, nevertheless, BVS leads to more precise (lower standard error and lower mean squared error) parameter estimate for regression coefficients under all settings.

The advantage of BVS compared to the aforementioned IC based methods is its unified framework for model averaging and model selection. While posterior model probabilities or inclusion probabilities can be used as a model selection tool, the final parameter estimates returned from the MCMC simulation are weighted average of model-specific estimates according to the posterior model probabilities. Similarly, the model averaged estimates which takes model uncertainty into account can be obtained for IC methods by averaging model-specific estimates according to the appropriate model weights. However, computation become cumbersome and sometimes impossible as fitting each candidate models separately is required to compute the model averaged estimates, unconditional standard errors and unconditional confidence intervals.

In addition, the BVS approach has several advantages including (1) subjective prior can be applied if available, (2) additional functions of parameters can be easily computed from MCMC chains, and (3) model convergence is not an issue for complex models compared to the frequentist approach.

The magnitude of the true regression coefficient under selection was found to be an important factor for a correct model selection. All approaches perform well when the magnitude of the true regression coefficient under selection is large relative to noise. On the other hand, when the magnitude of the true regression coefficients under selection are small relative to noise, the AICc and BIC approaches failed to detect the true model when the true underlying model contained these parameters. Relatively, the BVS approach performs well in such situations. Regarding the model averaged parameter estimation, all approaches lead to similar result, except the difficulties

one faces in obtaining these results using the IC based approaches. Our simulation study also highlights the importance of assuming a prior distribution for the hierarchical variance parameter rather than treating it as a fixed value in BVS model formulation.

In this paper, we only considered the application of BVS on the mean structure, i.e., for the "fixed effect" components of the model. The BVS approach can be extended to selection of random effect components as well (see for example, Yang et al., 2020). Further, we assumed independence between the regression coefficients in our prior specification. The BVS model presented in this paper can be extended to situation where the predictors are correlated by considering priors that take in to account possible dependence among regression coefficients (Griffin et al., 2017).

Acknowledgement

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government.

Funding

Financial support from the Institutional University Cooperation of the Council of Flemish Universities (VLIR-IUC) is gratefully acknowledged.

Competing interest

The authors declare that they have no conflict of interest.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716–723.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48. doi:.

Burnham, K.P., Anderson, D.R., 2003. Model selection and multimodel inference: a practical information-theoretic approach. second ed., Springer-Verlag, New York.

Carlin, B.P., Chib, S., 1995. Bayesian model choice via markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Methodological) 57, 473–484.

Claeskens, G., Hjort, N.L., 2008. Model selection and model averaging. Technical Report. Cambridge University Press.

Degefa, T., Zeynudin, A., Godesso, A., Michael, Y.H., Eba, K., Zemene, E., Emana, D., Birlie, B., Tushune, K., Yewhalaw, D., 2015. Malaria incidence and assessment of entomological indices among resettled communities in ethiopia: a longitudinal study. Malaria journal 14, 24.

Denwood, M.J., 2016. `runjags`: An `R` package providing interface utilities, model templates, parallel computing methods and additional distributions for `MCMC` models in `JAGS`. Journal of Statistical Software 71, 1–25.

Draper, D., 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society: Series B (Methodological) 57, 45–70.

George, E.I., McCulloch, R.E., 1993. Variable selection via gibbs sampling. Journal of the American Statistical Association 88, 881–889.

Geweke, J., 1996. Variable selection and model comparison in regression. Bayesian statistics 5, 609–620.

Griffin, J., Brown, P., et al., 2017. Hierarchical shrinkage priors for regression models. Bayesian Analysis 12, 135–159.

Hocking, R.R., 1976. A biometrics invited paper. the analysis and selection of variables in linear regression. Biometrics 32, 1–49.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical science , 382–401.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends in ecology & evolution 19, 101–108.

Kasim, A., Shkedy, Z., Kato, B.S., 2012. Estimation and inference under simple order restrictions: hierarchical bayesian approach, in: Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., Bijnens, L. (Eds.), *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*. Springer, pp. 193–214.

Kuo, L., Mallick, B., 1998. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics,

Series B , 65–81.
Lin, D., Shkedy, Z., Aerts, M., 2012. Classification of monotone gene profiles using information theory selection methods, in: Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., Bijnens, L. (Eds.), *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*. Springer, pp. 151–163.
McCulloch, C.E., Neuhaus, J.M., 2005. Generalized linear mixed models. volume 4. Wiley Online Library.
Miller, A., 2002. Subset selection in regression. Chapman and Hall/CRC.
Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1032.
Molenberghs, G., Verbeke, G., 2005. Models for discrete longitudinal data. Springer Series in Statistics, Springer, New York.
Ntzoufras, I., 2011. Bayesian modeling using WinBUGS. volume 698. John Wiley & Sons.
O'Hara, R.B., Sillanpää, M.J., 2009. A review of bayesian variable selection methods: what, how and which. Bayesian analysis 4, 85–117.
Otava, M., Shkedy, Z., Hothorn, L.A., Talloen, W., Gerhard, D., Kasim, A., 2017. Identification of the minimum effective dose for normally distributed data using a bayesian variable selection approach. Journal of Biopharmaceutical Statistics , 1–16.
Otava, M., Shkedy, Z., Lin, D., Goehlmann, H.W., Bijnens, L., Talloen, W., Kasim, A., 2014. Dose–response modeling under simple order restrictions using bayesian variable selection methods. Statistics in Biopharmaceutical Research 6, 252–262.
Plummer, M., et al., 2003. JAGS: A program for analysis of bayesian graphical models using gibbs sampling, in: Proceedings of the 3rd international workshop on distributed statistical computing, Vienna. p. 125.
Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics 6, 461–464.
Scott, J.G., Berger, J.O., et al., 2010. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics 38, 2587–2619.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639.
Sugiura, N., 1978. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. Communications in Statistics-Theory and Methods 7, 13–26.
Watanabe, S., 2013. A widely applicable bayesian information criterion. Journal of Machine Learning Research 14, 867–897.
Whitney, M., Ryan, L., 2009. Quantifying dose-response uncertainty using bayesian model averaging. Uncertainty Modeling in Dose Response: Bench Testing Environmental Toxicity , 165–179.
Yang, M., Wang, M., Dong, G., 2020. Bayesian variable selection for mixed effects model with shrinkage prior. Computational Statistics 35, 227–243.