Hindawi
Journal of Mathematics
Volume 2024, Article ID 1459524, 22 pages
https://doi.org/10.1155/2024/1459524

Research Article

The Best Fit Bayesian Hierarchical Generalized Linear Model Selection Using Information Complexity Criteria in the MCMC Approach

Endris Assen Ebrahim,[1] **Mehmet Ali Cengiz**,[2] **and Erol Terzi**[2]

[1]*Debre Tabor University, Department of Statistics, College of Natural and Computational Science, South Gondar, Ethiopia*
[2]*Ondokuz Mayis University, Institute of Graduate Studies, Faculty of Science and Art, Department of Statistics, Samsun, Türkiye*

Correspondence should be addressed to Endris Assen Ebrahim; end384@gmail.com

Received 27 February 2023; Revised 9 January 2024; Accepted 18 January 2024; Published 1 February 2024

Academic Editor: Antonio Di Crescenzo

Both frequentist and Bayesian statistics schools have improved statistical tools and model choices for the collected data or measurements. Model selection approaches have advanced due to the difficulty of comparing complicated hierarchical models in which linear predictors vary by grouping variables, and the number of model parameters is not distinct. Many regression model selection criteria are considered, including the maximum likelihood (ML) point estimation of the parameter and the logarithm of the likelihood of the dataset. This paper demonstrates the information complexity (ICOMP), Bayesian deviance information, or the widely applicable information criterion (WAIC) of the BRMS to hierarchical linear models fitted with repeated measures with a simulation and two real data examples. The Fisher information matrix for the Bayesian hierarchical model considering fixed and random parameters under maximizing a posterior estimation is derived. Using Gibbs sampling and Hybrid Hamiltonian Monte Carlo approaches, six different models were fitted for three distinct application datasets. The best-fitted candidate models were identified under each application dataset with the two MCMC approaches. In this case, the Bayesian hierarchical (mixed effect) linear model with random intercepts and random slopes estimated using the Hamiltonian Monte Carlo method best fits the two application datasets. Information complexity (ICOMP) is a better indicator of the best-fitted models than DIC and WAIC. In addition, the information complexity criterion showed that hierarchical models with gradient-based Hamiltonian Monte Carlo estimation are the best fit and have supper convergence relative to the gradient-free Gibbs sampling methods.

1. Introduction

We think through the tricky of comparing complex hierarchical models in which linear predictors vary by grouping variables, and the number of model parameters is not noticeably distinct [1]. In natural structure, experimental data in cognitive science, education, public health, and social follow-up contain "clusters." These nested structures comprise experimental measurements that are much more correlated within the group than between them. Such clusters in clinical trials and experimental designs are subjects and experimental units (e.g., words, pictures, and measurements presented to the issues). These clusters arise because we have multiple (repeated) observations for each subject and item [2]. Incorporating this grouping structure in data analysis leads to the necessary use of a hierarchical model (also called a multilevel or mixed-effects model). This grouping structure and hierarchical modeling type are closely connected to the concept of exchangeability [3]. The exchangeability concept of hierarchical models is the Bayesian equivalent of the assumption "independent and identically distributed" one often encounters in classical statistics thinking.

The rapid advancement of complex statistical modeling and computing to fit real-world data structures need the best model selection criteria [4]. In statistical modeling, having attention to model convergence and interpretability, there is no specific reason to select a single best model according to some criterion [5]. Instead, it makes more sense to "deselect"

poor models which might have overfitting and underfitting, keeping a subset model for further inference. Often, this subset might consider a single model, but sometimes possibly not.

Model selection is a procedure for finding the best-fitted model from a subset of models. The predictor variables and covariates are related to the outcome/dependent variable [6]. Model selection approaches find the "best" trade-off between goodness-of-fit with data and model complexity [7, 8]. Based on the correct complexity interpretation of the models, these techniques can be categorized into different goals to realize high predictive density, low predictive error, high model probability, or minor looseness of information. These goals can be accomplished by ensuring unreliable or reliable model selection and integrating a piece of Bayesian prior information or not [9].

Hierarchical (i.e., mixed-effects) linear models include random intercepts, random slopes for all within-subjects, experimental units, and correlations between the random effects components [10]. The simulation results showed that complex models with more parameters, including Bayesian parameter priors, faced a more negligible convergence effect [11]. Although various model selection tools indicated the best-fitting model, proper model selection depends on the random effect structure of the Bayesian hierarchical model. That is why model selection should be built on goodness-of-fit and consider model complexity [12]. Fitted models with high complexity measures attempt to capture each deviation in the data points. Such models are supposed to have high variance structures, leading to overfitting the data [13].

Model selection, based merely on the fit to the trained dataset, leads to choosing an unnecessarily complex model that overfits the data and thus infers poorly. Model selection techniques must appropriately balance the consequence of overfitting [9, 11, 12, 14].

Many more information criteria are implemented in different modeling types and found in different software packages. These include Akaike Information Criteria (AIC), Adjusted Akaike Information Criteria (AICc), Schwartz Bayes Information Criteria (SBC, SBIC), and Bayesian Information Criteria (BIC) [15]. All these information criteria are based on the maximum likelihood (ML) point estimation, which can be expressed as the sum of the deviation and penalty terms [16]. These criteria select the "best-fitting" model with sufficient goodness of fit and few parameters, penalizing the over determined parameter for lack of appropriate measure [17]. The Bayesian model selection method used the posterior proposal to decide whether the fitted model maximizes the expected utility of the posterior distribution for the data and parameters [5, 18].

The model averaging approach used flat priors over the range of plausible values of the model parameters [19]. Another popular model selection method is the Bayes factor, which requires one unique priory nominated model by the decision of the statisticians or the researcher out of all available fitted models [20]. Nevertheless, the researcher's (or scientist's) decision might be wrong, making the Bayes factor irrelevant in limited situations. In addition, overlooking the information criteria computed based on the deviation terms in the maximum likelihood (ML) point estimation, the penalty terms rely on only the sample size and the number of parameters. None-point estimation of the full posterior estimation (the expected posterior estimator) considers the variance-covariance matrices [13, 17, 21].

In the Bayesian approach model compassion, the deviance information criterion (DIC) available in the MCMCglmm and the widely applicable information criterion (WAIC) of the BRMS in Stan are attempts to find the model with the best predictive models [22]. Like DIC, WAIC estimates adequate parameters to adjust for overfitting in the model. WAIC is a fully Bayesian pointwise version of the AIC, asymptotically equivalent to the Deviance Information Criterion (DIC) [20].

On the other hand, a novel model selection criteria known as information complexity (ICOMP) is calculated by the set of random vectors obtained from the information-based covariance complexity index for a general multivariate linear or nonlinear model estimated with the C-valued equation or by the structural complexity of an element. ICOMP represents and is a real-valued measure of complexity. ICOMP is the estimation of the covariance matrix of the parameter vectors in the model. In the general case of model selection criteria, the best model minimizes the criterion [23–25]. The main objective of this study is to evaluate the popular model selection criteria that consider the fixed parameterizations with ICOMP criteria that consider the covariance matrix of the parameter vectors for the different setups of hierarchical modeling with prior distributions. Therefore, this paper compares the fitted Bayesian hierarchical models using covariance complexity-based ICOMP and the number of estimated parameters based on DIC or WAIC under two popular MCMC approaches in three application datasets.

2. Methodology

Classical statistical linear models were often fitted by maximum likelihood estimation considering parameter point estimates. This paper demonstrates the theoretical and practical drive of the information complexity criterion for Bayesian hierarchical linear models, which fit by estimation methods of maximizing a posterior (MAP) distribution. The Bayesian hierarchical model considers a full posterior distribution supported by parameter prior information, and it has additional random terms besides the usual residual time in the standard linear model. Model complexity frequently refers to the number of features or variables incorporated in a specified predictive model in the machine learning concept. Based on the researcher's needs and the model structure, there are several freely available R-based Bayesian hierarchical model applications of bearing to the data in any field [26]. Here, we applied two freely available R packages for Bayesian hierarchical linear modeling: gradient-free Bayesian Monte Carlo and one gradient-based (Hamiltonian Markov Chains). The hybrid Hamiltonian in Stan is a pioneer and more effective Monte Carlo method than other Markov Chain Monte Carlo approaches [27]. Six different Bayesian hierarchical linear models, of which 3 used gradient-free Gibbs sampling approach and the other 3 used gradient-based Hamiltonian Monte Carlo estimation, were fitted for each of the three application datasets.

2.1. The General Form of the Fisher Information Matrix for a Linear Model. For a given vector of measurements y_{ij}, a general linear model can be written as follows:

$$Y = X\beta + e. \tag{1}$$

The likelihood $\mathscr{L}$ is a function of all model parameters $\theta = (\beta, \sigma_e^2)$ defined as $L(\theta \mid Y) = P(Y_{ij}, \theta)$. For a general linear regression model, the log-likelihood can be written as follows:

$$LL(\theta \mid Y) = \log P\left(Y_{ij}, \theta\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma_e^2\right) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{N}\left(y_{ij} - f\left(x_{ij}, \beta\right)\right)^2. \tag{2}$$

The partial derivative of the log-likelihood for the parameter θ is called the score. Under the general regularity condition, the expected value of the score is 0 [28]. Indeed, it is easy to show that

$$E\left(\frac{\partial}{\partial\theta}\log P\left(y_{ij}; \theta^*\right)\right) = 0, \tag{3}$$

where θ^* is the "true" unknown value of θ such that the observations y_{ij} were generated with model $P(.; \theta^*)$. The variance of the score is called the Fisher information matrix (FIM).

$$\mathrm{FIM}\left(\theta^*\right) = E\left[\frac{\partial}{\partial\theta}\log P\left(y_{ij}; \theta^*\right)\left(\frac{\partial}{\partial\theta}\log P\left(y_{ij}; \theta^*\right)\right)^t\right]. \tag{4}$$

Besides, when the log-likelihood LL is double differentiable concerning the parameter θ, the FIM is given by the following equation:

$$\begin{aligned}\mathrm{FIM}\left(\theta^*\right) &= -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log P\left(y_{ij}; \theta^*\right)\right)\\ &= -\sum_{i=1}^{n} E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log P\left(y_{ij}; \theta^*\right)\right).\end{aligned} \tag{5}$$

The variance of σ_e^2 is estimated by $I_y^{-1}(\widehat{\sigma}^2)$ as $I_y(\widehat{\sigma}^2) = (\partial^2/\partial(\sigma_e^2)^2)\,\mathscr{L}\mathscr{L}(\widehat{\beta}, \widehat{\sigma}_e^2) = -(n/2\widehat{\sigma}^4) + (n/2\widehat{\sigma}^6)\sum_{i=1}^{N}(Y_i - P(X_i - \widehat{\beta}))^2 = (n/2\widehat{\sigma}^4)$. And standard errors (SE) of $\widehat{\sigma}^2$ becomes $\widehat{\sigma}^2/\sqrt{n/2}$.

In this case, the estimated inverse Fisher information matrix (IFIM) (i.e., Cramér–Rao lower bound matrix) is as follows:

$$\widehat{\mathrm{Cov}}\left(\widehat{\beta}, \widehat{\sigma}_e^2\right) = F^{-1} = \begin{bmatrix}\widehat{\sigma}_e^2\left(X'X\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \dfrac{2\widehat{\sigma}_e^4}{n}\end{bmatrix}. \tag{6}$$

2.2. Bayesian Hierarchical Linear Modeling for Repeated Measures Data. Suppose the target variable Y_{ij} is the j^{th} repeated observation $(j = 1, 2, 3, \ldots, n_i)$ taken of the individual (subject) $i = 1, 2, 3, \ldots, m$ that is considered $N = \sum_{i=1}^{m} n_i$ as a total number of response measurements. Assuming the parameters (θ) of the model with the set of p explanatory variables $X_i = x_{ij}^{(1)}, \ldots, x_{ij}^{(p)}$ (fixed effects) and q random terms constant compliments of X_i as Z_i, the hierarchical linear model which considers a group (subject) i is an extension of equation (1) and can be written as follows:

$$Y_i = X_i\beta + Z_iU_i + \varepsilon_i, \tag{7}$$

where

$$\begin{aligned}Y_i &= \begin{bmatrix}y_{i1}\\ y_{i2}\\ \cdots\\ y_{in_i}\end{bmatrix}, X_i = \begin{bmatrix}1 & x_{i1}^{(1)} & \cdots & x_{i1}^{(p)}\\ 1 & x_{i2}^{(1)} & \cdots & x_{i2}^{(p)}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{in_i}^{(1)} & \cdots & x_{in_i}^{(p)}\end{bmatrix}, \beta = \begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{bmatrix}, U_i = \begin{bmatrix}U_{i0}\\ U_{i1}\\ \vdots\\ U_{iq}\end{bmatrix}, \varepsilon_i = \begin{bmatrix}e_{i1}\\ e_{i2}\\ \vdots\\ e_{in_i}\end{bmatrix},\\ Z_i &= \begin{bmatrix}x_{i1}^{(1)} & \cdots & x_{i1}^{(q)}\\ x_{i2}^{(1)} & \cdots & x_{i2}^{(q)}\\ \vdots & \ddots & \vdots\\ x_{in_i}^{(1)} & \cdots & x_{in_i}^{(q)}\end{bmatrix}.\end{aligned} \tag{8}$$

Since the random effects vary by group (g), $U_i \sim \text{Normal}(0, \sigma_g^2)$; $\text{Cov}(U_g, U_{g'}) = \sigma_{gg'}$ and $\varepsilon_i \sim \text{Normal}\ (0, \sigma_e^2)$; $\text{Cov}(\varepsilon_g, \varepsilon_{g'}) = \sigma_{ee'}$. Therefore, the variances and covariance among the random (group) effects are σ_g^2 and $\sigma_{gg'}$, respectively. The residual error variance and covariance between errors in the group are σ_e^2 and $\sigma_{ee'}$ [29–31]. In general form,

$$Y_{N\times 1} = \underbrace{X_{N\times p}\beta_{p\times 1}}_{\text{fixed effects}} + \underbrace{Z_{N\times mq}U_{mq\times \mathbf{1}}}_{\text{random effects}} + \underbrace{\varepsilon_{N\times 1}}_{\text{residuals (error term)}}, \quad (9)$$

where Y denotes the vector $(y_1', y_2', \ldots y_m')'$ of outcome variable, β denotes a vector of fixed effects parameters, U denotes a vector $(U_1', U_2', \ldots U_m')'$ of associated random effects (specific to each subject), X is a matrix of covariates (explanatory variables), Z denotes a block diagonal matrix of covariates for the random effects as a complement of X embraced of m blocks that each block has $n_i \times q$ dimension matrix, and ε denotes a column vector of residuals. We assumed that the random effects $U \sim N(0_q, \Omega = \sigma_g^2)$ and the residuals $\varepsilon \sim N(0_{n_i}, R = \sigma_e^2)$, where U and ε are independently distributed. Based on the unknown vector of φ_Ω and φ_R, the unknown random effects in Ω and R can be written as $\Sigma = (\varphi_\Omega, \varphi_R)$ [32].

As the population parameters of the hierarchical linear model, the model parameters are the vector of fixed effects β, the $q \times q$ variance-covariance matrix Ω for the random effects, and the variance σ_e^2 of the residual errors. Here, let $\theta = (\beta, \Omega, \sigma_e^2)$ be the set of $(p+1)$ fixed effect and $(mq+1)$ random effect model parameters and $y_i \sim N(X_i\beta, Z_i\Omega Z_i' + \sigma^2 I_{ni})$.

For a given set of parameter estimates $\widehat{\theta}$, the covariance matrix of Y can be computed as $\text{cov}(Y) = \widehat{V} = Z\widehat{\Omega}Z' + \widehat{\sigma}_e^2 I$ and the estimate of the covariance matrix is a positive definite matrix [33].

The Bayesian hierarchical linear model with fixed and random effect parameters can be noted as $\theta = (\beta, \sigma^2 = [\sigma_g^2, \sigma_e^2])$. This parameterization implies, for $Y \sim$ normal $(X\beta, V)$ in which $V = Z\Omega Z' + \widehat{\sigma}_e^2 I_n$, the marginal likelihood can be written as follows:

$$L(\beta, \sigma^2; Y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta). \quad (10)$$

The Best Linear Unbiased Estimator (BLUE) of fixed effects β and the Best Linear Unbiased Predictor (BLUP) of random effects can be expressed as follows:

$$\widehat{\beta} = \left(X'V^{-1}X\right)^{-1}X'V^{-1}Y, \qquad \widehat{u} = \Omega Z'V^{-1}(Y - X\widehat{\beta}). \quad (11)$$

2.3. Maximum a Posterior Estimator (MAP) of Model Parameters. The maximum likelihood (ML) estimators of θ maximize the log-likelihood function defined as follows:

$$\begin{aligned} \text{LL}(\theta \mid Y) &= \log(P(y, \theta)) \\ &= \sum_{i=1}^{N} \log(P(y_i, \theta)) \\ &= \sum_{i=1}^{N} \left\{ -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|Z_i\Omega Z_i' + \sigma^2 I_{ni}\right|\right) - \frac{1}{2}(y_i - X_i\beta)'\left(Z_i\Omega Z_i' + \sigma^2 I_{ni}\right)^{-1}(y_i - X_i\beta) \right\}. \end{aligned} \quad (12)$$

Due to the lack of a simple analytical solution to this optimization problem, various numerical methods such as the Newton–Raphson and the Expectation-Maximization (EM) algorithms can be used for maximizing $\text{LL}(\theta \mid Y)$. Among the likelihood methods, the unbiased estimates of variance and covariance parameters can be computed using the restricted maximum likelihood (REML) approach in contrast to the maximum likelihood (ML) method. The maximum likelihood estimates of the fixed effect parameters β_{px1} are given by the following equation:

$$\widehat{\beta}_{\text{MLE}} = \left(X'X\right)^{-1}X'y, \quad (13)$$

where (X'X) is nonsingular. Bayesian inference introduces prior distributions over all model parameters θ and is built entirely on posterior distributions of model parameters or conditional distributions $P(\theta_K \mid \text{data})$ for K^{th} parameter θ_K, given data, and other known quantities in the model.

In the Bayesian generalized hierarchical linear model, the joint full parameters, $\theta = (\beta, \Omega, \sigma_e^2)$, and posterior distribution $P((\theta \mid Y))$ can be articulated as follows:

$$P(\theta \mid Y) \propto P(\theta, Y) = \prod_{i=1}^{K} \int P\left(Y_{ij} \mid U_i, \beta\right) P\left(U_i \mid \Omega\right) P(\theta) dU_i, \quad (14)$$

where the regularizing constant is independent of the parameter components in θ. Estimation of the parameter θ can be derived from the full parameters posterior distribution, $P(\theta, Y)$, only through specified locational measures of the posterior, such as posterior mode, mean, or median. The prior for θ, $P(\theta)$, is a constant, then the posterior in equation (14) is

effectively proportional to the likelihood function in equation (10), and hence, the posterior mode is numerically identical to the maximum likelihood estimate. Thus, flat priors or noninformative priors for fixed effects, (β), and the random effect, (Ω), are usually preferred. In this case, the distribution of random effects assumed to be normal, specifically $U_1, U_2, \ldots, U_K$ are *i.i.d.* $q-$ dimensional $\mathrm{MVN}_q(\mathbf{0}, \Omega)$ [34].

It is computationally true that considering uniform distribution priors $p(\theta)$ for θ indicates that the posterior distribution in equation (14) is efficiently compared to the likelihood function. It provides identical results of the maximum likelihood estimate and the Bayesian maximum a posteriori (MAP) estimate of the posterior mode [35]. For this reason, flat priors or noninformative priors for fixed effect parameters β and random effect parameters Ω are typically chosen. Moreover, for the reason that the Bayesian hierarchical modeling dimensions complexity, the usual numerical approximation techniques cannot provide the maximization solution. Thus, a recent influential method for handling complex statistical integration is the Markov chain Monte Carlo (MCMC) algorithm. Then, with the noninformative prior specification, the joint posterior distribution can be written as follows:

$$P\left(\beta, \sigma_g^2, \frac{\sigma_e^2}{X}, Y\right) \propto p\left(X, \frac{Y}{\beta}, \sigma_g^2, \sigma_e^2\right) \times p\left(\frac{\sigma_g^2}{\sigma_e^2}\right) \times p(\beta) \times p\left(\sigma_g^2\right) \times p\left(\sigma_e^2\right). \tag{15}$$

The theoretical derivation of the Bayesian maximum a posteriori (MAP) estimator considers a Bayesian hierarchical linear model where all parameters are random at the individual level. Based on the marginal likelihood function of equation (10), one can derive separate scores for each component of θ as β and σ^2 (for $k = 1, 2, \ldots, K$) [35]. Therefore, the score for σ^2 from the gradient for the k^{th} pointwise entry of σ^2 is as follows:

$$\frac{\partial L\left(\beta, \sigma^2; Y\right)}{\partial \sigma_k^2} = -\frac{1}{2}\mathrm{tr}\left[V^{-1}\frac{\partial V}{\partial \sigma_k^2}\right] + \frac{1}{2}(Y - X\beta)^{'}V^{-1}\left(\frac{\partial V}{\partial \sigma_k^2}\right)V^{-1}(Y - X\beta). \tag{16}$$

To obtain the scores from every observed gradient i, we need to use the trace operator with a diag operator by component-wise multiplication known as Hadamard product. The score function $S_i(.)$ for each observation for the parameter vector θ is as follows:

$$S\left(\sigma_k^2; Y\right) = -\frac{1}{2}\mathrm{diag}\left[V^{-1}\frac{\partial V}{\partial \sigma_k^2}\right] + \left\{\frac{1}{2}(Y - X\beta)^{'}V^{-1}\left(\frac{\partial V}{\partial \sigma_k^2}\right)V^{-1}\right\}^{'} \odot (Y - X\beta). \tag{17}$$

Equation (17) gives $n \times 1$ score vector from the gradient of parameter σ_k^2 (a scalar).

Similarly, the score vector for the fixed effect parameter β can be obtained from the gradient.

$$\frac{\partial L\left(\beta, \sigma^2; Y\right)}{\partial \beta} = X^{'}V^{-1}(Y - X\beta). \tag{18}$$

Again replacing the matrix multiplication by the Hadamard product,

$$S(\beta; Y) = \left\{X^{'}V^{-1}\right\}^{'} \odot (Y - X\beta). \tag{19}$$

After the derivation of gradients for fixed and random effects, the complete set of scores can then be the combined matrix whose columns consist of the results of the separate score vectors from equations (17) and (19) [28].

2.4. Derivation of the Fisher Information Matrix for Hierarchical Linear Modeling. To derive the Fisher information matrix for the Bayesian hierarchical linear models (BHLM) in equation (7) or (9), we need to ponder the concept of Fisher information matrix for general regression and consider the design matrix Z can be partitioned into r submatrices. Suppose $q(i)$ denote the number of columns in Z_i, and then, $I_{q(i)}$ is an identity matrix with dimension $q(i) \times q(i)$. This implies $Z = [Z_1, Z_2, \ldots, Z_r]$ and $U^{'} = [U_1', U_2', U_3', \ldots, U_r']$ with $\mathrm{cov}(U_i) = \sigma_i^2 I_{q(i)}$ and $\mathrm{cov}(U_i, U_k) = 0$ for $i \neq k$. Therefore, the covariance matrix of the random effects U is a block diagonal matrix with blocks $\mathrm{cov}(U_i) = \sigma_i^2 I_{q(i)}$. As standard assumptions, $R = I_N$ and $R = Z_i Z_i'$. We can rewrite $\mathrm{cov}(Y) = V = \sum_{i=0}^{r} \sigma_i^2 Z_i Z_i'$, where $\sigma_0^2 = \sigma_e^2$. Thus, in equation (9), the covariance of the random effects U can be rewritten as follows:

$$\Omega = \sum_{i=1}^{r} \sigma_i^2 Z_i Z_i'. \tag{20}$$

Here, the unknown model parameters vector θ comprises the fixed effect parameters β and the random effect scalars $\sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$. Thus, as an alternative to estimating the matrix Ω, we need to estimate scalars $\sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$. Note that, besides the variance terms, the covariance terms, such as $\sigma_{01}, \sigma_{02}, \ldots . \sigma_{(r-1)r}$, are needed to be estimated.

The fixed portion of the model in equation (9) is analogous to the general linear model regression coefficients to be estimated by maximizing the log-likelihood. For the random part of the model (6), we assume that U had variance-covariance matrix Ω and that U is orthogonal to ε so that

$$Var\begin{pmatrix} U \\ \varepsilon \end{pmatrix} = \begin{bmatrix} \Omega & 0 \\ 0 & \sigma_e^2 R \end{bmatrix}. \tag{21}$$

Here, considering $V = \text{cov}\ (\theta)$ which encompasses the variance-covariances Ω_β, Ω_u, and Ω_e for fixed effect β, random effect U,, and the residual term ε, respectively, the full variance-covariance matrix for θ can be written as follows:

$$\Sigma = \begin{bmatrix} \Omega_\beta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Omega_u & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Omega_e \end{bmatrix}. \tag{22}$$

Suppose F denotes the Fisher information matrix for the hierarchical model in equation (9); without considering the constant 2π, the log-likelihood function for model (9) can be rewritten as follows:

$$\begin{aligned} \text{LL}(\theta \mid Y) &= \mathscr{L}\mathscr{L}\left(\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2 \mid Y\right) \\ &= -\frac{1}{2}\log|V| - \frac{(Y - X\beta)^{'V^{-1}(Y - X\beta)}}{2}. \end{aligned} \tag{23}$$

Now, our primary target is finding the matrix $F = -E_\theta \left\{\partial^2 \mathscr{L}\mathscr{L}(\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2 \mid Y)/\partial\theta\partial\theta^T\right\}$.

Thus, the Fisher information matrix for the Bayesian hierarchical linear model (BHLM) is a block diagonal matrix that combines the covariance terms of the fixed and random parts:

$$\begin{aligned} F &= -E_\theta\left\{\frac{\partial^2 \mathscr{L}\mathscr{L}\left(\beta, \sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2 \mid Y\right)}{\partial\theta\partial\theta^T}\right\} \\ &= \begin{bmatrix} X^{'}V^{-1}X & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}tr\left(V^{-1}Z_i Z_i^{'} V^{-1} Z_k Z_k^{'}\right) \end{bmatrix}. \end{aligned} \tag{24}$$

Then, from equations (23) and (24), we need to invert the information matrix F to get the full variance-covariance matrix concerning all parameters in the model [28].

2.5. Bayesian Deviance Information and Widely Applicable Information Criterion. Many researchers have scrutinized model selection from frequentist and Bayesian perspectives, and many tools for selecting the "best-fitted model" have been suggested [5]. The Bayesian deviance information criterion (DIC) developed by [1] is a Bayesian version of the Akaike Information Criterion that substitutes a maximized log-likelihood with the log-likelihood evaluated by the Bayes estimate. Nevertheless, it is not fully Bayesian logically because of its reducing behavior of the probability distribution down to point estimates. Bayesian deviance information criterion (DIC) can be computed as follows:

$$\begin{aligned} \text{DIC} &= -2\log p\left(y \mid \widehat{\theta}_{\text{MAP}}\right) + 2p_{\text{DIC}} \\ &= D_{\text{hat}} + 2p_{\text{DIC}} = D_{\text{bar}} + p_{\text{DIC}}, \end{aligned} \tag{25}$$

where $\widehat{\theta}_{\text{MAP}}$ is the maximizing a posterior (MAP) estimate that replaces the ML-point estimate in AIC. The new measure of the Bayesian predictive accuracy expressed as the expected log-point prediction intensity is the difference between the posterior mean of the deviance minus the deviance of the posterior means: $p_{\text{DIC}} = 2(\log p(y \mid \widehat{\theta}_{\text{MAP}}) - E_{\text{post}}(\log p(y \mid \theta)))$, and D_{bar} is the posterior mean of deviance and D_{hat} is a point estimate of the deviance obtained by substituting in the posterior means $\widehat{\theta}_{\text{MAP}}$, thus $D_{\text{hat}} = -2\log p(y \mid \widehat{\theta}_{\text{MAP}})$. This gives an adequate number of parameters, p_{DIC}, given by $p_{\text{DIC}} = D_{\text{bar}} - D_{\text{hat}}$.

Widely applicable information criterion (WAIC), also known as Watanabe–Akaike [36], could be achieved as an improvement over the deviance information criterion (DIC) for Bayesian models [25]. Widely applicable Bayesian information criterions penalty term is purely Bayesian and is computed pointwise as follows:

$$p_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}_{\text{post}}(\log p(y_i \mid \theta)). \tag{26}$$

Here, p_{WAIC}, the penalty term is the variance of the log-predictive-density terms aggregated over N data points. Thus, the Watanabe–widely applicable Bayesian information criterion can be calculated as follows:

$$\text{WAIC} = -2\log\left(p\left(Y \mid \theta_{\text{post}}\right)\right) + 2p_{\text{WAIC}}, \tag{27}$$

where θ_{post} is the joint posterior distribution of fitted model parameters and p_{WAIC} is an estimate of the number of effective parameters computed as in equation (26). In the general case, WAIC uses the logarithm of the pointwise predictive density, and its logarithm sums the overall observations. This estimation leads to the asymptotical equivalence of WAIC and DIC in Bayesian modeling. However, the penalization term for overfitted estimates the number of effective parameters [20].

2.6. The Information Complexity (ICOMP) Criterion for the Hierarchical Linear Model. Deviance can be defined as the likelihood difference between the fitted and perfect models. It is used to measure the deviance of the fitted binary logistic model for the saturated model for $P(Y = 1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$. It is statistically valid that the deviance is always larger than or equal to zero only if the fit is perfect [35]. The deviance of the worst (null) model, the one fitted without any predictor, to the ideal model can be written as $Y \mid (X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) \sim \text{Binary}(\beta_0)$.

As a background of model selection criteria, Akaike Information Criterion (AIC) [21, 37] is extensively used for different statistical models selection criteria and it is given by the following equation:

$$\text{AIC} = -2\text{LL}(\theta \mid Y) + 2k, \tag{28}$$

where $\text{LL}(\theta \mid Y)$ is the log-likelihood function of θ parameter in a probability distribution, and k is the number of parameters in a model M_θ. ICOMP (I, information-COMP, complexity) is a criterion developed by [38] for selecting

multivariate linear and nonlinear models, whereas the log-likelihood AIC is merely meant to strike a balance between the lack of fit and the penalty terms; however, ICOMP seeks to achieve this balance by taking into account a measure of complexity that assesses how the model's parameters interact with one another. As a result, it penalizes the covariance complexity of the model rather than simply punishing the number of parameters [39]. The information complexity (ICOMP) criterion is given by the following equation:

$$\text{ICOMP} = -2\text{LL}(\theta \mid Y) + 2C_{\max}(\Sigma). \tag{29}$$

The second part of equation (29) is called the measure of complexity of the model given in equation (9), which can be computed as follows:

$$C_{\max}(\Sigma) = \frac{d}{2}\log\left[\frac{\text{tr}(\Sigma)}{d}\right] - \frac{1}{2}\log|\Sigma|, \tag{30}$$

where $|\Sigma|$ represents the determinant of Σ and d is the dimension of Σ. As seen, $C_{\max}(\Sigma)$ includes the two most straightforward scales of multivariate scattering called determinant and trace in a single function [40]. Depending on the model structure, the ICOMP criterion can be computed by substituting the inverse Fisher information matrix (IFIM) F^{-1} to measure the covariance complexity. Information complexity (ICOMP) has various forms [41–43].

AIC, DIC, BIC, and ICOMP, among other model fit criteria, combine the goodness-of-fit term of the model, 2LL (Y), and the model's complexity, commonly referred to as the penalty term. All likelihood-based ML points and maximizing posterior estimation measures of model fit criteria also do this. However, many criteria used the number of parameters as a penalty for complexity; the penalty term of AIC is $2k$, two times the number of estimated parameters. In contrast, the penalty term of the ICOMP criterion is the measure of the covariance complexity for the fitted models [41]. Because the structured covariance matrix Σ of the estimated parameters $\widehat{\theta}$ is unknown in closed form and nonidentifiability, we then apply the estimated inverse Fisher information matrix to assess the complexity of the Bayesian hierarchical linear model [44]. The estimated inverse Fisher information matrix $\widehat{F}^{-1}$ can be computed with $\widehat{\theta}$ instead of θ in a matrix F^{-1}. Thus,

$$\text{ICOMP}_{(\text{IFIM})} = -2\text{LL}(\theta \mid Y) + 2C_{\max}\left(\widehat{F}^{-1}\right). \tag{31}$$

Combining equation (23) having constant 2π with equation (29), the inverse Fisher information matrix (IFIM)-based information complexity (ICOMP) [42] is computed as follows:

$$\begin{aligned}\text{ICOMP}_{(\text{IFIM})} &= -2\text{LL}(\theta \mid Y) + \frac{d}{2}\log\left[\frac{\text{tr}\left(\widehat{F}^{-1}\right)}{d}\right] - \frac{1}{2}\log\left|\widehat{F}^{-1}\right| \\ &= N \times \text{Log}(2\pi) + \log|\widehat{V}| \\ &\quad + (Y - X\widehat{\beta})'\,\widehat{V}^{-1}(Y - X\widehat{\beta}) \\ &\quad + \frac{d}{2}\log\left[\frac{\text{tr}\left(\widehat{F}^{-1}\right)}{d}\right] - \frac{1}{2}\log\left|\widehat{F}^{-1}\right|.\end{aligned} \tag{32}$$

For the ICOMP criterion implemented on the fitted models, the model with a minimum ICOMP measure value is known as the best model.

3. Application of Model Fitting to Categorical Outcome Repeated Measures Data

This article demonstrates and realizes the fitting of a categorical outcome Bayesian generalized hierarchical linear model for a trait with repeated observations using two different (MCMCglmm and BRMS) *R* packages in two real datasets and simulation data.

3.1. Bayesian Hierarchical Linear Model Specifications. Suppose the general Bayesian hierarchical model in equation (1) for the subject (group) i and measurement repeated at j. It is assumed that $Y_{ij} \sim N(\mu_{ij}, \sigma_y^2)\, u_i \sim N(\mu_g, \sigma_g^2)\, \beta_i \sim N(\mu_\beta, \sigma_\beta^2)$.

Every part of the hierarchical model can be seen in the graph as a node, as depicted in Figure 1. The solid arrows reflect stochastic (random) connections, such as those from σ_e^2 to Y_{ij}, whereas the dotted arrows indicate deterministic (fixed) relationships across the parameters, such as those from β_i to μ_{ij}.

3.1.1. Model 1: The Null or Unconditional Random Intercepts Model. In hierarchical modeling, the null model is the model with only grouping (clustering level) variables as a determinant of the intercept of the dependent variable [45]. This is an "unconditional" and random intercept model since it predicts the outcome variable's level one intercept without any predictor variable at any level. Furthermore, this model provides information on intraclass correlations, which helps to decide if hierarchical models are necessary for the data in the first place.

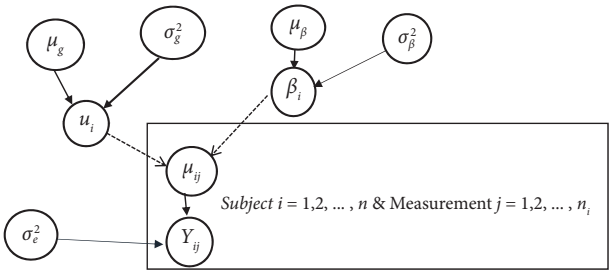


Figure 1: A full random intercept and slopes model (Bayesian framework).

$$\begin{aligned} Y_{ij} &= \text{Binomial}\left(n_i = 1, p_{ij}\right), \\ \text{Logit}\left(p_{ij}\right) &= \alpha_i + \alpha_j, \\ \alpha_i &\sim \text{Normal}\left(\alpha, \sigma_i\right), \\ \alpha_j &\sim \text{Normal}\left(0, \sigma_j\right), \\ \alpha &\sim \text{Normal}\left(0,\ 10\right), \\ \sigma_i &\sim \text{HalfCauchy}\left(1\right), \end{aligned} \tag{33}$$

where α is the overall intercept; α_i and α_j are the intercepts with σ_i and σ_j variance (standard deviation) components for each individual considering all measurements and for the repeated measures, respectively. In a Bayesian hierarchical model, the BRMS-Stan documentation package in *R* recommended using half-Cauchy as the prior, which is automatically constrained at zero, to lessen the likelihood of unreasonably large standard deviation (SD) values [45]. For the intercept, the default prior is a normal distribution.

3.1.2. Model 2: The Conditional Random Intercepts Model. A conditional random intercept model incorporates a cluster as a random effect, and only the intercept of the outcome variable is adjusted for the random effects. It is also conditional because predictor variables are added to the clustering variables [46]. A random intercept model is one in which intercepts are allowed to vary. As a result, the intercept that fluctuates across groups predicts the scores on the dependent variable for each unique observation [47].

However, the slopes in this model are assumed to be fixed (the same across different contexts). Given Y_{ij} = Binomial (n_i, p_{ij}). A random intercept logistic regression model is defined as follows:

$$\begin{aligned} \text{Logit}\left(p_{ij}\right) &= \alpha + \alpha_i + \alpha_j + \left(\beta_1 + \ldots + \beta_p\right)X_{ij}, \\ \alpha_i &\sim \text{Normal}\left(0, \sigma_i\right), \\ \alpha_j &\sim \text{Normal}\left(0, \sigma_j\right), \\ \alpha &\sim \text{Normal}\left(0,\ 10\right), \\ \left(\beta_1, \ldots \beta_p\right) &\sim \text{Normal}\left(0,\ 10\right), \\ \sigma_i &\sim \text{HalfCauchy}\left(0,\ 1\right), \\ \sigma_j &\sim \text{HalfCauchy}\left(0,\ 1\right), \end{aligned} \tag{34}$$

where α is the overall intercept; α_i and α_j are the intercepts with σ_i and σ_j variance (standard deviation) components for each individual considering all measurements and for the repeated measures at j, respectively. In a Bayesian hierarchical model, the BRMS-Stan documentation package in *R* recommended using half-Cauchy as the prior, which is automatically constrained at zero, to lessen the likelihood of unreasonably large standard deviation (SD) values at all levels [45]. The default prior is normal for the intercepts and slopes, with a mean of 0 and a standard deviation of 10 considered.

The mathematical notation of the variance-covariance matrix, Σ_f, considering the overall intercept α and slopes β, which are fixed effects, can be designed as follows:

$$\Sigma_f = \begin{bmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{bmatrix}, \tag{35}$$

where σ_α^2, σ_β^2, σ_β, and σ_α are the variance and standard deviation components of the intercept and coefficients and ρ is the relationship between intercepts and coefficients.

Similarly, the covariance matrix for the grouping (random) effects, Σ_r, can be expressed as follows:

$$\Sigma_r = \begin{bmatrix} \sigma_i^2 & \sigma_i^2\rho_{ij} \\ \sigma_i^2\rho_{ij} & \sigma_j^2 \end{bmatrix}, \tag{36}$$

where σ_i^2 and σ_j^2 are the variance components at the individual subject and repeated measurements, respectively; and ρ_{ij} is the relationship between j measurements within subject i. Hierarchical (multilevel) models adjust parameter estimates of the intercepts (mean) of one or more dependent variables at level 1 based on grouping variables defining the higher levels. And it also adjusts the slopes (β, coefficients) of one or more predictors (regressors) at any level, and ρ is the association between intercepts and coefficients at the subject level.

3.1.3. Model 3: The Conditional Random Coefficients Model. The random coefficients model also called the conditional random coefficient model is the model with all varying effects, adjusting the intercept (mean) of the outcome variable as well as the coefficients (slopes) of predictors for the random effects at any clustering (level) of the hierarchical data. The coefficient (slope) term in the random coefficients model should not obscure the fact that a random slope model estimates the intercept (mean) as well as the slopes (regression coefficients) at the appropriate hierarchy [48].

A random slope model is one in which the slopes are permitted to change, resulting in slopes that differ between groups. The most realistic model contains random intercepts and random slopes but can also be the most complex. Both intercepts and slopes can change among groups in this paradigm, implying distinct in different situations [49]. Thus, in Bayesian hierarchical modeling, the prior distribution parameters as hyperparameters and the distributions of hyperparameters as hyperprior distribution noticeably occurred.

It can be assumed that both the random intercepts and coefficient/slope model hyperparameters $\mu_g, \mu_\beta, \sigma_g, \sigma_\beta$, and ρ have uniform hyperprior distributions with assumptions suitable for parameters. A more common way to write the model by addressing the correlation between parameters is as follows: $Y_{ij} \sim f(u_i + X_{ij}\beta_i, \sigma_e^2)$, where the regression coefficient β_i vary by grouping variables.

$$\begin{aligned} Y_{ij} &= \text{Binomial}\left(n_i = 1, p_{ij}\right), \\ \text{Logit}\left(p_{ij}\right) &= \alpha + \alpha_i + \alpha_{ij} + \left(\beta_i\right)X_{ij}, \\ \alpha_i &\sim \text{Normal}\left(0, \sigma_i\right), \\ \alpha_j &\sim \text{Normal}\left(0, \sigma_j\right), \\ \alpha &\sim \text{Normal}\left(0,\ 10\right), \\ \left(\beta_1, \ldots \beta_p\right) &\sim \text{Normal}\left(0,\ 10\right), \\ \sigma_i &\sim \text{HalfCauchy}\left(0,\ 1\right), \\ \sigma_j &\sim \text{HalfCauchy}\left(0,\ 1\right), \end{aligned} \tag{37}$$

where α is the overall intercept and α_i and α_{ij} are the intercepts with σ_i and σ_j variance (standard deviation) components for each individual considering all measurements and for the repeated measures at j, respectively. Half-Cauchy distribution is used as the prior for standard deviation (SD) at the individual subject and measurement levels. The default prior is normal for the intercepts and slopes, with a mean of 0 and a standard deviation of 10 considered.

Considering the intercepts (means) α as μ for groups, the covariance matrix for the random effect, u and fixed effect, β of the model can be expressed as follows:

$$\begin{bmatrix} u \\ \beta \end{bmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_g \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_g^2 & \rho\sigma_g\sigma_\beta \\ \rho\sigma_g\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right), \tag{38}$$

where σ_g^2 and σ_β^2 are the variance components of the random and fixed effects, respectively, and ρ is the relationship between random and fixed effects.

3.2. Real Data Application One: Arterial Occlusive Disease Data. In this application, we consider a dataset related to surgical planning, specifically for arterial occlusive disease data. The two popular invasive methods, ultrasound imaging and reduced cuff pressure measures, classify each leg as either healthy (0) or diseased (1). The variables are indicated as health status: Y_{ij} which is the i^{th} patient's health status from the measurement on the j^{th} side of the leg. Ultrasound: X_{1ij} is the i^{th} patient's ultrasound measurement on the j^{th} side of the leg. RCP: X_{2ij} is the i^{th} patient's reduced cuff pressure (RCP) measurement on the j^{th} side of the leg. Here, individual patients $i = 1, 2, 3, \ldots, 16$ and leg side s, $j = 1, 2, 3, 4$, which indicates measurements at $j = 1$ for right leg upper-side, $j = 2$ for right leg lower-side, $j = 3$ for left leg upper-side, and $j = 4$ for left leg lower-side. In the body, the branches of the arteries that come out of the aorta and carry clean blood to the arms, legs, head, and organs, and the branches of the veins that connect to the main vein bring the dirty blood coming from them to the heart are called peripheral vessels [50].

Arterial occlusive diseases are occlusion or narrowing of the arteries in the legs (or rarely in the arms), usually caused by atherosclerosis resulting from reduced blood movement. Data were collected at Broadgreen Hospital in Liverpool, England, in 1988/89 [51]. Healthy peripheral arteries have a flatlining that prevents coagulation and promotes constant blood movement. Peripheral artery disease can affect all arteries but is most commonly seen in the legs. The data included 16 patients whose features were measured at 4 points on the lower and upper sides of their right and left legs. Of all patients' characteristics measured during data collection, we only considered the patient's health status, ultrasound measurements, and reduction in cuff pressure measurements, excluding variables measured at one point and missing values. The patient's health status was considered the outcome variable, while the ultrasound imaging score and cuff pressure measurement reduction were independent variables. The data structure for the variables in each measurement is shown in Table 1.

3.3. Real Data Application Two: The National Longitudinal Survey of Youth 1979. National Longitudinal Survey of Youth (1979–2012) is a longitudinal project that follows a sample of Americans on various life aspects collected from 1979 to 2012. The dataset has multiple characteristics, mainly socioeconomic status, employment, education, and marriage. For this modeling application, we considered 895 black Americans by year level, and the variables include poverty status, family size, and overall income extracted based on removing missing measurements on selected variables for two years. Bayesian hierarchical linear models were demonstrated, modeling the poverty status of young people as the response variable and family size and overall income varying over their identification (YID) as the predictor variables. The data can be accessed from https://dasil.sites.grinnell.edu/downloadable-data/.

3.4. Application Three: Simulation of Repeated COVID-19 PCR Test Results Scenario. This section considers the variation in three consecutive testing for people with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) PCR test results related to age and testing site (environment). The scenario can be easily simulated for a sample of 50 patients using the *R* package, as shown in the R-code of the data analysis.

4. Results and Discussion

4.1. Results of Fitted Model Estimates and Model Selection in Each Application Case. After fitting a set of models for each dataset, we need to know which model is more accurate and should be used to make inferences and draw conclusions. Choosing (using R^2, for example) the model with a better absolute fit to the real dataset can be a challenge, as this model does not necessarily perform well on new data.

Table 1: Arterial occlusive disease data structure and variables measurement.

i	Left leg						Right leg					
	Upper side			Lower side			Upper side			Lower side		
	X_{1i}	X_{2i}	Y_i	X_{1i}	X_{2i}	Y_i	X_{1i}	X_{2i}	Y_i	X_{1i}	X_{2i}	Y_i
1	170	0.25	1	45	0.66	0	170	1.0	1	45	0.81	1
2	−5	0.72	1	25	0.23	0	−5	0.78	1	25	0.37	1
3	30	0.29	0	−10	0.23	0	30	0.58	1	−10	0.12	0
4	160	0.47	0	100	0.37	0	160	0.91	1	100	1.0	1
5	50	0.63	1	70	0.61	1	50	0.69	0	70	1.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...

Instead, we may want to choose the model with the best predictive capabilities, that is, the model that performs best in predicting data that has not yet been observed. We call this ability the out-of-sample prediction performance of the model [3].

Model convergence was demonstrated in a complex parameter configuration using R-hat ($\widehat{R}$) statistics sometimes referred to as the potential scale reduction factor (PSRF) and effective sample sizes (ESS). The consistency of an ensemble of Markov chains was demonstrated by models with an effective sample size greater than 100 and an R-hat closest to 1.00 but not greater than 1.10 [52]. As model convergence diagnosis, we reported effective sample size cutoffs for hierarchical models as the bulk effective sample size (Bulk_ESS) and the tail effective sample size (Tail_ESS).

In the first case, based on the arterial occlusive disease real data application, among the fitted six models, results in Table 1 showed that the random coefficient model (M_{3H}) in the Hamiltonian Monte Carlo algorithm fits best relative to other models. Besides, the Bayesian information complexity criterion has smaller values than Bayesian deviance or widely applicable information criteria for each fitted model. Thus, the random coefficient model fitted by Hamiltonian Monte Carlo (HMC) algorithm has the least ICOMP value identified as the absolute best model or more accurate in applying dataset one.

In the second case, based on the National Longitudinal Survey of Youth 1979 real data application, among the fitted six models, results in Table 1 showed that the random intercept model (M_{2H}) in the Hamiltonian Monte Carlo algorithm fits best relative to other models. Besides, the Bayesian information complexity criterion has smaller values than Bayesian deviance or widely applicable information criteria for each model. Thus, the random intercept (fixed slopes) model fitted by Hamiltonian Monte Carlo (HMC) algorithm has the least ICOMP value, identified as the absolute best model or more accurate in the application two dataset. The results also showed that the varying effects of coefficients (fixed effects) do not provide additional information for the data.

In the third case, based on the simulation application dataset of the repeated COVID-19 PCR test results scenarios, among the fitted six models, the results in Table 2 showed that the random coefficient model (M_{2H}) in the Hamiltonian Monte Carlo algorithm fits best relative to other models. Thus, the random coefficient model fitted by Hamiltonian Monte Carlo (HMC) algorithm has the least ICOMP value identified as the absolute best model or more accurate in the application three simulation dataset. The three cases showed that Hamiltonian Monte Carlo (HMC) estimation is better than the Gibbs sampler for complex Bayesian hierarchical models. ICOMP criterion has smaller values and is a better model assessment tool than Bayesian deviance or widely applicable information criteria for each fitted model of two-level repeated measures data [53].

In Tables 3–5, the results of application datasets showed the estimates of the posterior mean (intercept), estimates of coefficients, and standard error as the standard deviation (SD) for each parameter. Model convergence was achieved well enough since Rhat = 1.00 in all cases, and both the bulk effective sample size (Bulk_ESS) and the tail effective sample size (Tail_ESS) for the 95% credential intervals were adequate [25]. Generally, each parameter is summarized by the posterior mean ("Estimate") and standard deviation of the population parameter ("Std. Err."). Moreover, a 95% credible interval can be used as lower and upper bounds based on posterior quintiles.

4.2. Assessment of Convergence and Conditional Effects in Selected Models. The information complexity (ICOMP) criterion assessed the best-fitted model based on the figures. The hierarchical model with gradient-based Hamiltonian Monte Carlo estimation provided the best fit and superconvergence models relative to the gradient-free Gibbs sampling approach. Although there is no specific best statistical package for a particular statistical model and dataset, complex hierarchical models can be estimated well using Bayesian Regression Models using Stan (BRMS) compared to MCMC generalized linear mixed/hierarchical models (MCMCglmm) in R software [27, 49, 52, 54].

Based on Application 1: Arterial occlusive diseases data, the model convergence diagnosis, paired plots, and marginal effects are visualized in Figures 2–4 for the models fitted by MCMCglmm and BRMS in Stan. Based on the convergence plots, the convergence assessment of the models selected above showed that the random coefficient model under the HMC approach best fitted and Bayesian hierarchical models converged well under the Hamiltonian Monte Carlo approach. The best-fitted model's marginal effects plot (Figure 2) showed that leg side and ultrasound imaging had a positive effect, while the reduced cuff pressure (RCP) measures had no effect in both the fixed and random parts. The probability of events (Figure 3) of a patient's illness positively affected the leg sides. There are

Table 2: Results of information criteria to evaluate the fitted Bayesian hierarchical models.

Application/data	MCMC approach	Fitted model type	ICOMP	DIC/WAIC
Application 1: arterial occlusive disease	Gibbs sampler	M_{1G} (null model)	140.79	1047.22
		M_{2G} (random intercept)	137.20	1210.81
		M_{3G} (random coefficient)	132.90	667.40
	Hamiltonian MC	M_{1H} (null model)	172.50	1254.68
		M_{2H} (random intercept)	83.65	172.89
		M_{3H} (random coefficient)	75.69*	125.96*
Application 2: National Longitudinal Survey of Youth 1979	Gibbs sampler	M_{1G} (null model)	1268.12	2423.22
		M_{2G} (random intercept)	1259.44	2389.12
		M_{3G} (random coefficient)	1218.83	2175.45
	Hamiltonian MC	M_{1H} (null model)	131.98	229.48
		M_{2H} (random intercept)	53.59*	117.64*
		M_{3H} (random coefficient)	62.50	132.74
Application 3: Simulation with repeated COVID-19 PCR test results	Gibbs sampler	M_{1G} (null model)	197.39	224.92
		M_{2G} (random intercept)	196.97	198.93
		M_{3G} (random coefficient)	160.12	667.23
	Hamiltonian MC	M_{1G} (null model)	198.91	1530.98
		M_{2G} (random intercept)	202.26	1354.48
		M_{3G} (random coefficient)	149.31*	180.97*

*The best-fitted model for each application dataset under each MCMC approach.

Table 3: Estimates of random coefficient models for application 1: arterial occlusive disease.

Covariates	Estimates (post. mean)	Std. Err.	95% cred. interval		Bulk_ESS	Tail_ESS
			Lower	Upper		
Population-level (location) effects: fixed effects						
Intercept	−3.33	0.86	−5.14	−1.74	3741	3257
Leg side	0.60	0.05	0.50	1.70	376	3468
Ultrasound	5.73	1.48	3.04	8.78	3760	3486
RCP	0.01	0.01	0.00	0.03	4062	3773
Group-level effects: random effects						
Level 2: patient ID	0.37	0.30	0.01	1.09	3847	3722
Level 1: leg side	0.04	0.01	0.02	0.06	6	2254
$\sigma_{\text{ultrason.ID}}$	0.73	0.65	0.02	2.30	2878	2841
$\sigma_{\text{RCP.ID}}$	0.01	0.01	0.00	0.03	3987	3377

Table 4: Estimates of random intercept models for application 2: NLS of Youth 1979.

Covariates	Estimates (post. mean)	Std. Err.	95% cred. interval		Bulk_ESS	Tail_ESS
			Lower	Upper		
Population-level (location) effects: fixed effects						
Intercept	0.10	6.70	−13.88	8.01	387	337
Year	5.73	1.48	3.04	8.78	3760	3486
Family size	0.01	0.01	0.00	0.03	4062	3773
Income	−0.01	0.01	−0.02	0.01	343	413
Group-level effects: random effects						
Level 2: youth ID	0.16	0.10	0.01	0.39	399	337
Level 1: years	1.86	3.20	0.08	10.14	358	314

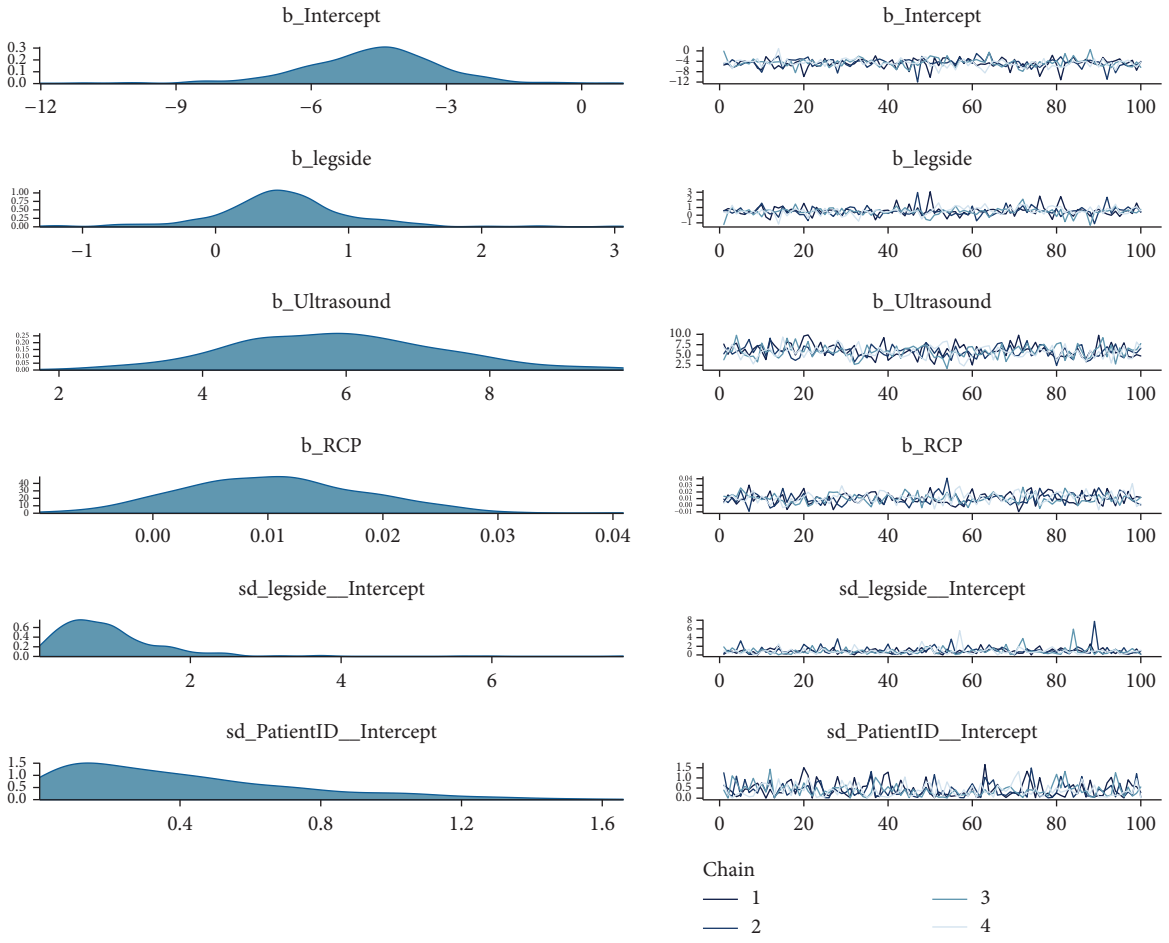
Table 5: Estimates of the random coefficient models for application 3: simulation data.

Covariates	Estimates (post. mean)	Std. err	95% cred. interval		Bulk_ESS	Tail_ESS
			Lower	Upper		
Population-level (location) effects: fixed effects						
Intercept	−3.95 354	55.59	−117.95	105.80	354	332
Time points	0.72	6.92	−12.75	14.37	467	443
Age	0.05	0.92	−1.81	1.90	357	399
Site	−0.76	6.94	−13.75	13.38	477	417
Level 2: ID: intercept	0.76	0.60	0.02	2.21	366	362
Level 1: time points	0.67	0.84	0.02	2.63	393	367
$\sigma_{\text{age.ID}}$	0.02	0.01	0.00	0.05	426	440
$\sigma_{\text{site.ID}}$	0.53	0.39	0.03	1.49	514	416

more disease occurrence variations between the legs within a patient than between patients. From Figures 3 and 4, ultrasound measurements showed a significant positive effect on patients' arterial occlusive disease status varying between four leg sides. At the same time, reduced cuff pressure (RCP) measures had no variation.

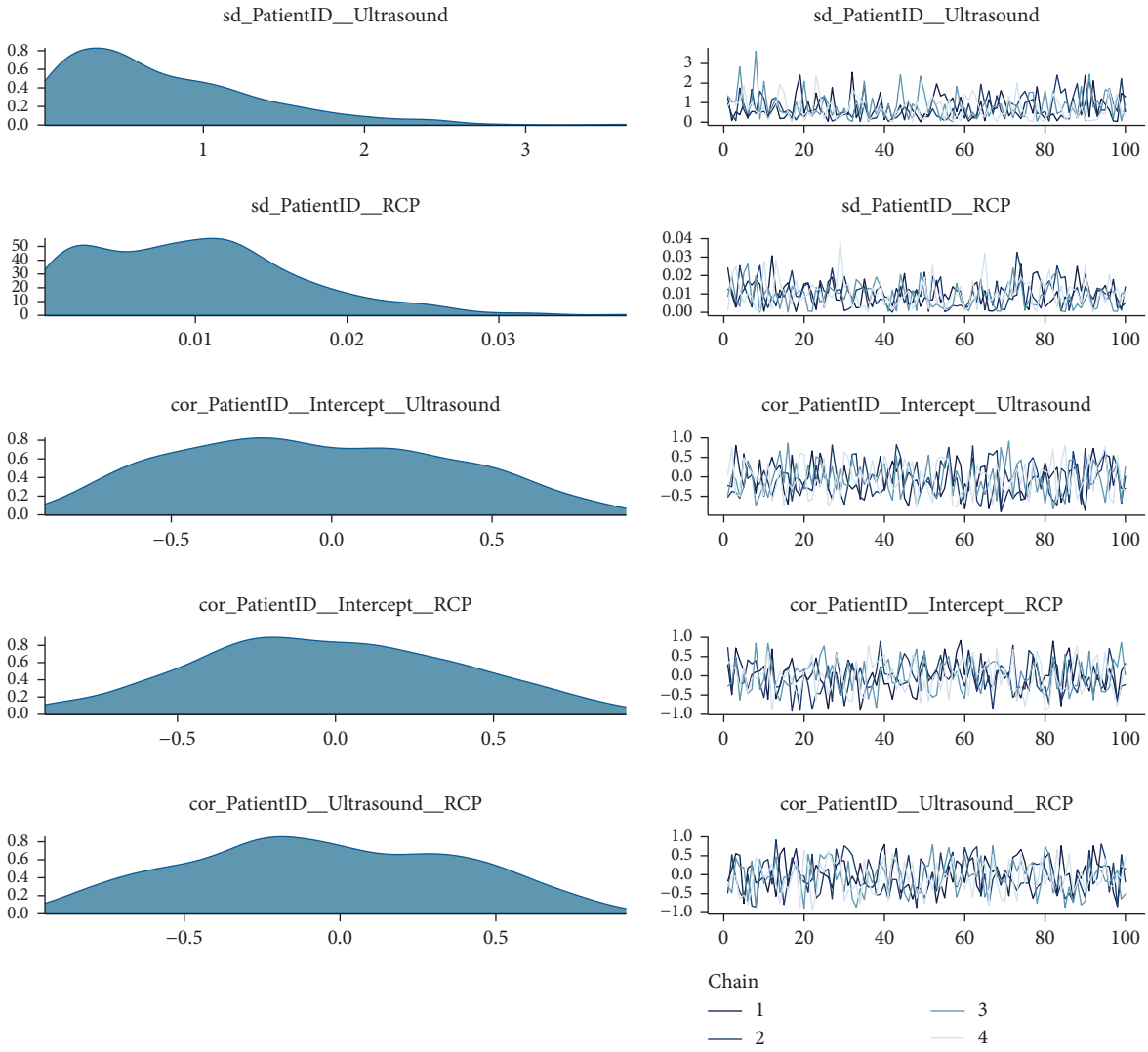
Application 2: National Longitudinal Survey of Youth 1979 data showed that the model convergence diagnosis paired plots and marginal effects are visualized in Figures 5 and 6 for the models fitted by MCMCglmm and BRMS in Stan. Based on convergence plots, the convergence assessment of the selected random intercept model under the HMC approach showed the best fit and Bayesian hierarchical models converged well under the Hamiltonian Monte Carlo approach. The marginal effects of year, family size, and income, as visualized in plot Figures 6 and 7 on the poverty status of youths in the best-fitted model, showed that year and family size had a positive effect. In contrast, income had almost no impact on poverty status.

Application 3: simulation data, the model convergence diagnosis paired plots, and marginal effects are visualized in Figures 8–10 for the models fitted by MCMCglmm and BRMS in Stan. Based on the convergence plots, the convergence assessment of the models selected above showed



(a)

Figure 2: Continued.



(b)

Figure 2: Trace and density plots of the best-fitted random coefficient model (application 1).

that the random coefficient model under the HMC approach was best fitted, and Bayesian hierarchical models converged well under the Hamiltonian Monte Carlo approach. The best-fitted model's marginal effects plot (Figure 8) showed that test time points and age had a positive effect, while the test site had (more negative) no impact in both the fixed and random parts. The probability of events (i.e., a positive SARS-CoV-2 PCR test) shown in Figures 3 and 10 of a patient's status in the three measurement time points increases with age but decreases with test site variation.

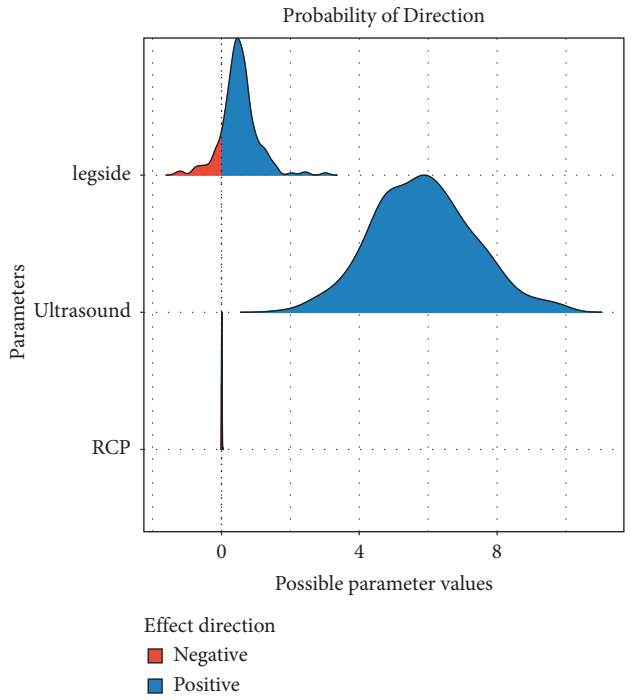


Figure 3: Marginal effect directions of the fixed effect parameters for the best-fitted model (application 1).

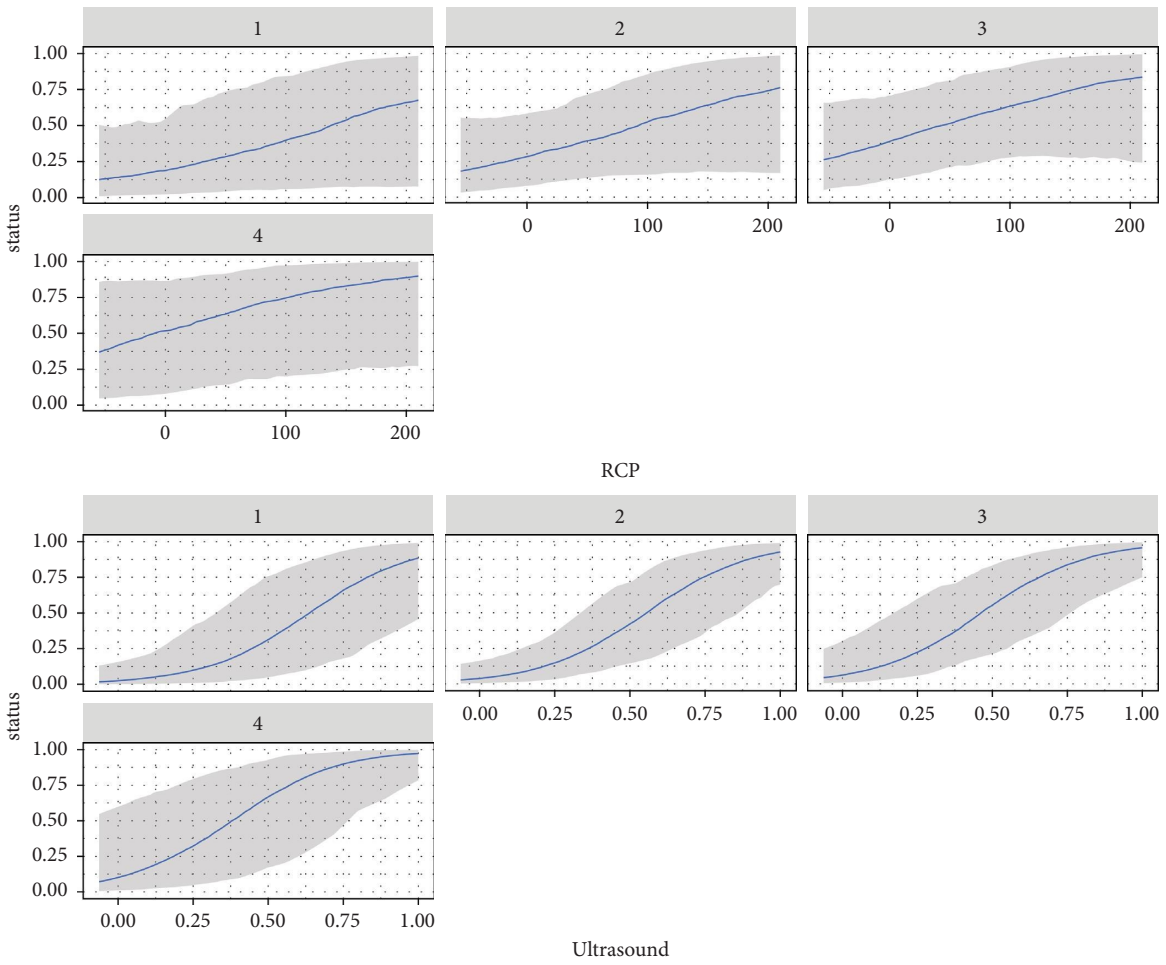


Figure 4: Marginal effects of RCP and ultrasound with measurements on the legs for the best-fitted model (application 1).

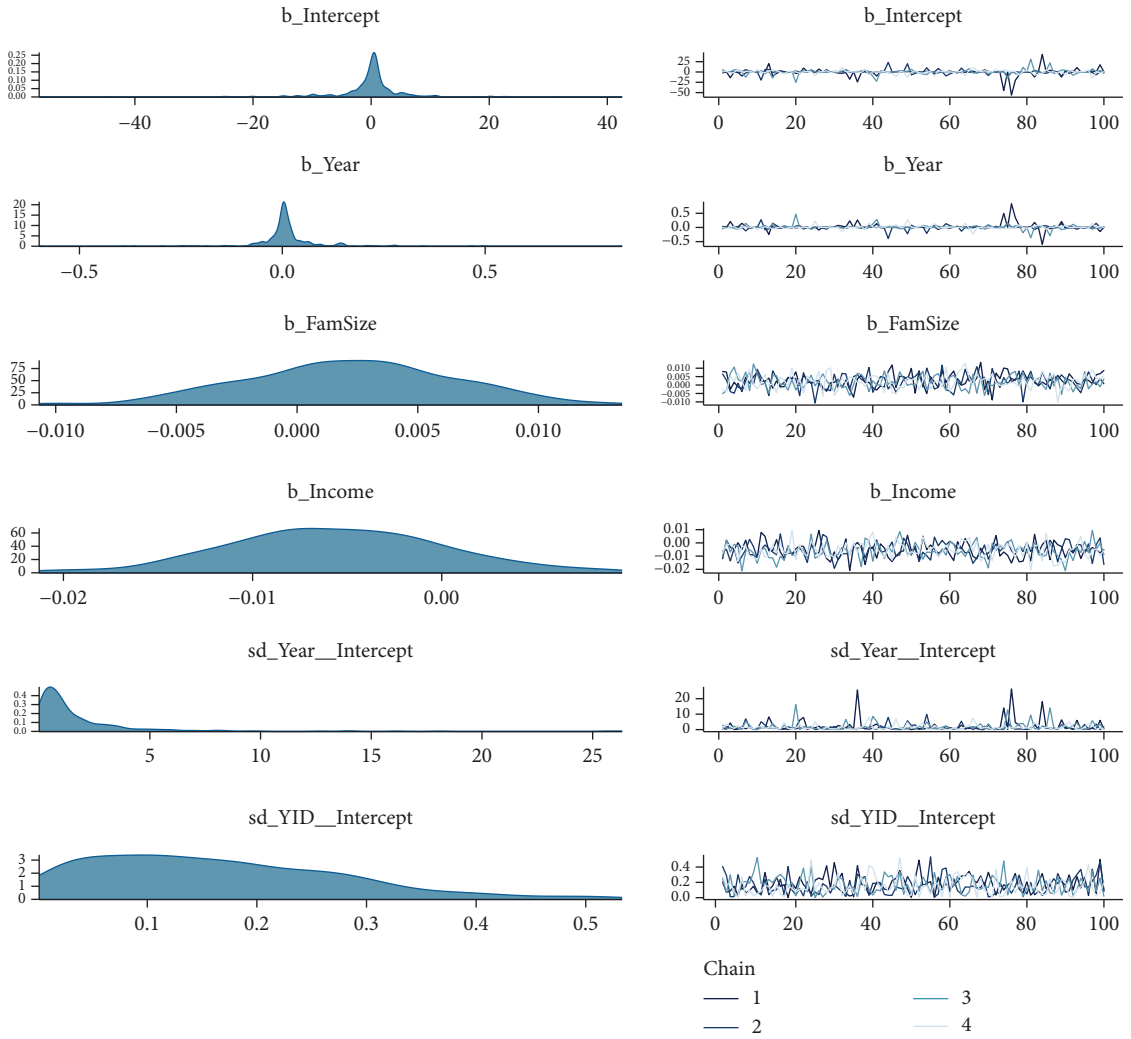


Figure 5: Trace and density plots of the best-fitted random intercept model (application 2).

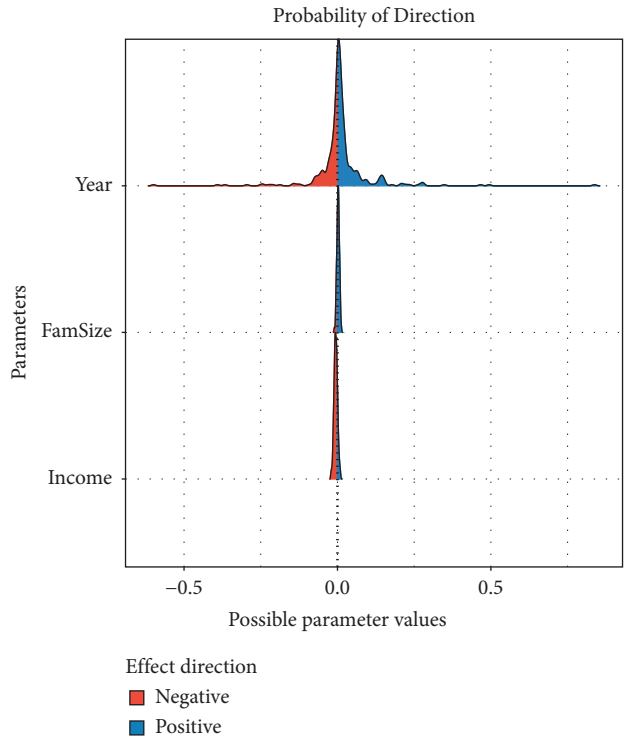


Figure 6: Marginal effect directions of the fixed effect parameters for the best-fitted model (application 2).

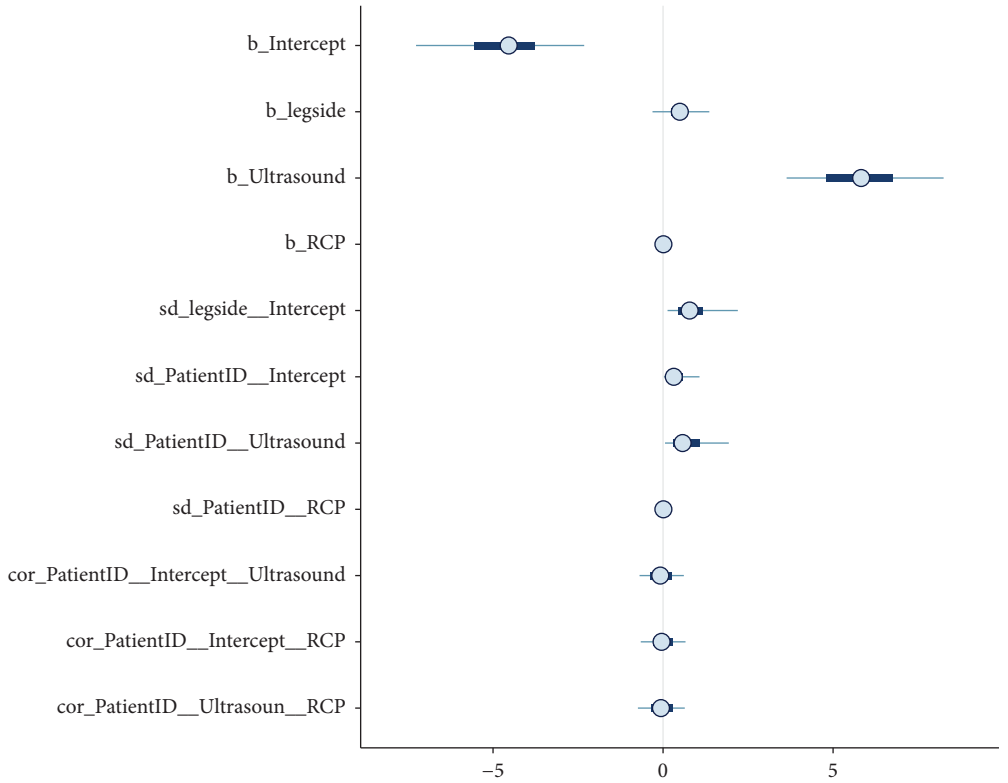
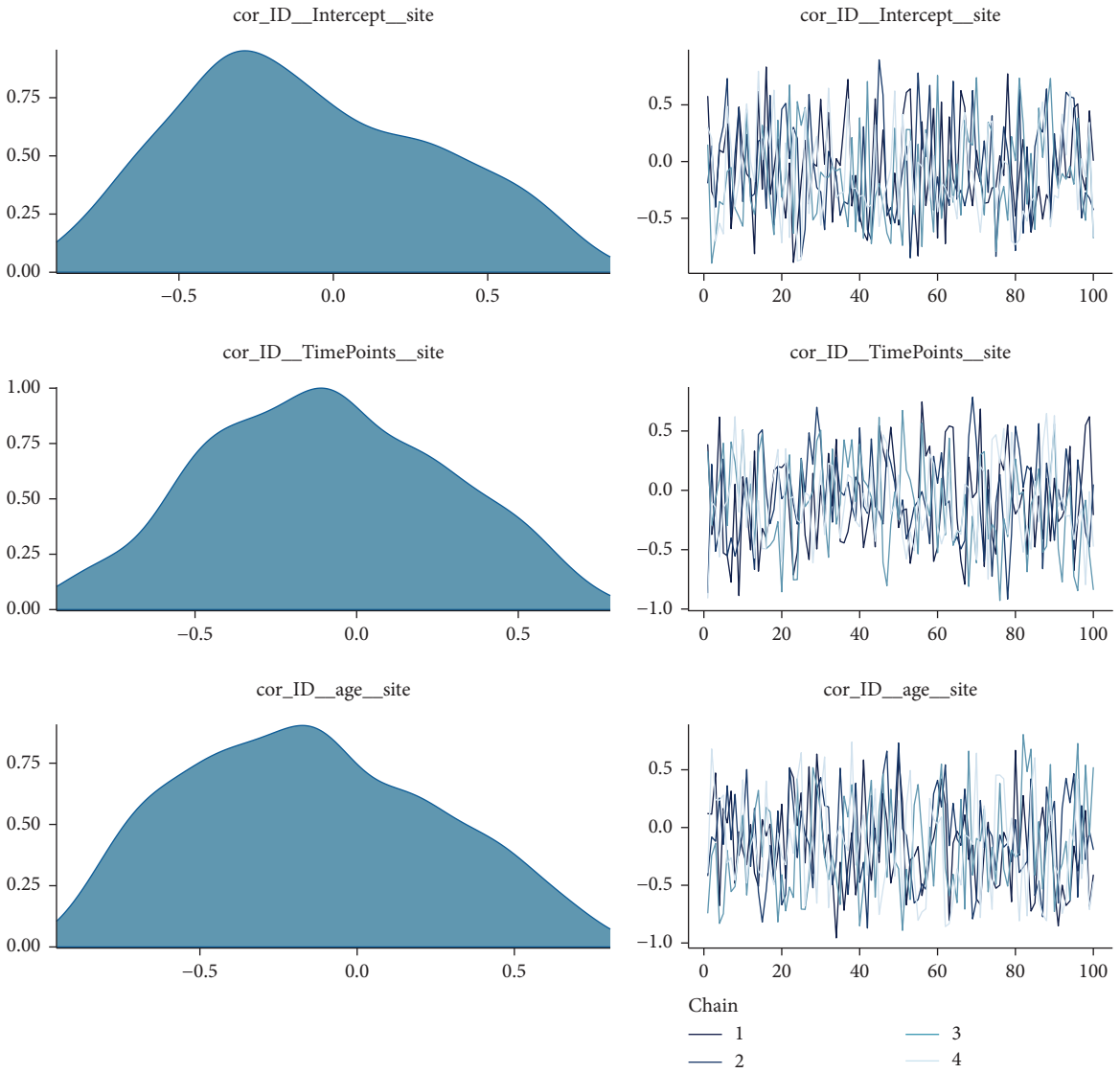
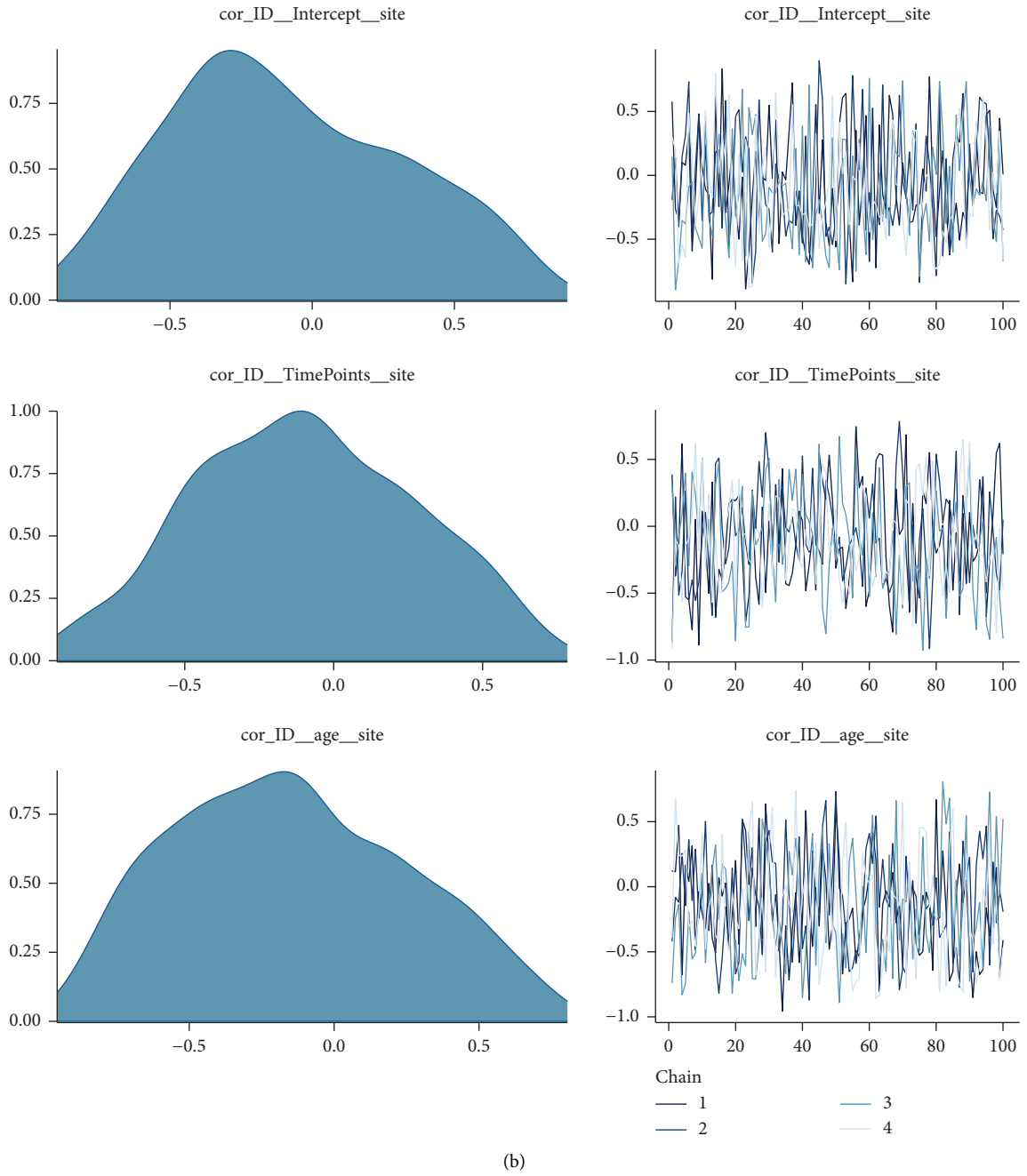


Figure 7: Marginal effects of fixed and random effects on all parameters for the best-fitted model (application 1).



(a)

Figure 8: Continued.



(b)

Figure 8: Trace and density plots of the best-fitted random coefficient model (application 3).

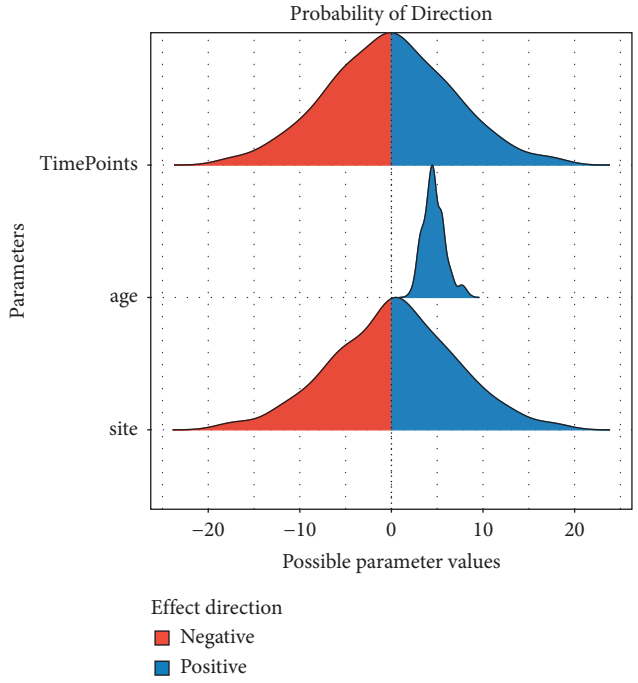


Figure 9: Marginal effect directions of the fixed effect parameters for the best-fitted model (application 3).

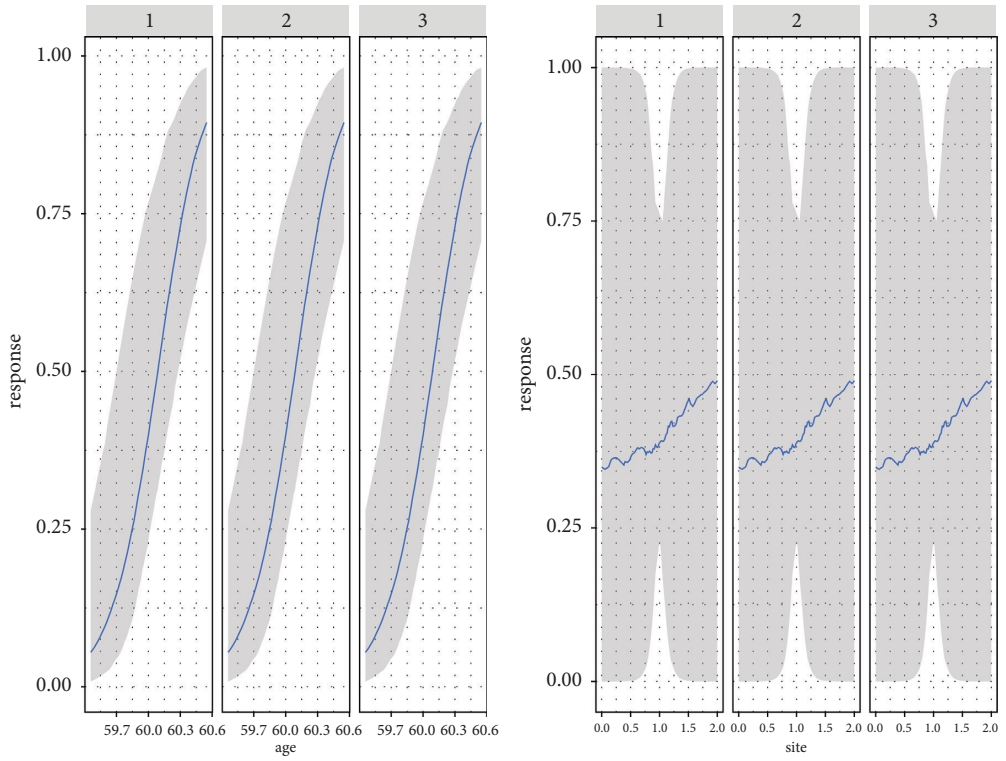


Figure 10: Marginal effects of age and site with measurement time points of the COVID-19 test for the best-fitted model (application 3).

5. Conclusion

Complex hierarchical models require a Bayesian computation (MCMC) that integrates the three components of the posterior: a prior, likelihood, and evidence (marginal likelihood). This study demonstrates two Bayesian computations (MCMC) methodologies to fit three distinct application datasets to three Bayesian hierarchical models: null, random intercept, and random coefficient models. All fitted models estimated by Gibbs sampler and Hamiltonian Monte Carlo (HMC)

approaches were compared to select the best-fitted model in each application case. In all cases, models fitted with the Hamiltonian Monte Carlo (HMC) method were the best-fitted models than models fitted with the Gibbs sampler. Hamiltonian Monte Carlo (HMC) in BRMS and Gibbs sampler in MCMCglmm R packages were used in this context.

That model convergence diagnosis demonstrated using effective sample size cutoffs for hierarchical models as the bulk effective sample size (Bulk_ESS) and the tail effective sample size (Tail_ESS) for the 95% credential intervals in each parameter estimation was adequate.

Moreover, model comparisons and sections were made using Bozdogan's information complexity measure, Bayesian deviance information: DIC (under MCMCglmm), and widely applicable information criterion (WAIC) of the BRMS in Stan [55]. Among the fitted candidate models, the information complexity (ICOMP) criterion showed the lowest measure values in all cases. A better-fit model has a smaller selection criterion measurement value. Hairy caterpillars were depicted in the model convergence assessment graphs, showing the model is well-fitted for the applied data. Performance assessment of ICOMP-type criteria and DIC or/and WAIC was validated using Bayesian hierarchical linear models. We are impressed that instead of DIC or/and WAIC, ICOMP deserves more exploration as a tool for model assessment and comparison with various thematic areas of repeat measures data of two hierarchical levels.

Data Availability

The labeled (Application 1 and Application 2) real datasets used to support the findings of this study and the R-code for the data analysis that had simulation scenarios (Application 3) and all fitted model applications are obtainable upon inquiry from the principal investigator.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Ebrahim EA engaged at every research stage: developing, writing the manuscript, managing data, and analyzing. EA Ebrahim wrote the initial format of the manuscript. EA Ebrahim and MA Cengiz performed material preparation, managing data, and analysis. MA Cengiz participated in editing and manuscript commenting. Erol Terzi participated in proofreading and revising the manuscript to improve the manuscript. The authors had substantial direct and intellectual involvement in the manuscript and approved it for publication.

References

[1] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 64, no. 4, pp. 583–639, 2002.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, CRC Press, Boca Raton, FL, USA, 3rd edition, 2013.

[3] R. McElreath, *Statistical Rethinking: A Bayesian Course With Examples In R And Stan*, CRC Press, Boca Raton, FL, USA, 2018.

[4] J. R. Busemeyer and Y.-M. Wang, "Model comparisons and model selections based on generalization criterion methodology," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 171–189, 2000.

[5] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 279–290, 2004.

[6] S. Müller, J. L. Scealy, and A. H. Welsh, "Model selection in linear mixed models," *Statistical Science*, vol. 28, no. 2, pp. 135–167, 2013.

[7] M. P. Wand, "Fisher information for generalised linear mixed models," *Journal of Multivariate Analysis*, vol. 98, no. 7, pp. 1412–1416, 2007.

[8] S. Watanabe, "Information criteria and cross validation for Bayesian inference in regular and singular cases," *Japanese Journal of Statistics and Data Science*, vol. 4, no. 1, pp. 1–19, 2021.

[9] M. Höge, T. Wöhling, and W. Nowak, "A primer for model selection: the decisive role of model complexity," *Water Resources Research*, vol. 54, no. 3, pp. 1688–1715, 2018.

[10] X. Liu, *Methods and Applications of Longitudinal Data Analysis, Uniformed Services University of the Health Sciences, Deployment Health Clinical Center, Defense Centers of Excellence, Walter Reed National Military Medical Center, Higher Education Press*, Elsevier Inc, Amsterdam, Netherlands, 2016.

[11] J. Park, R. Cardwell, and H. T. Yu, "Specifying the random effect structure in linear mixed effect models for analyzing psycholinguistic data," *Methodology*, vol. 16, no. 2, pp. 92–111, 2020.

[12] I. J. Myung, "The importance of complexity in model selection," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 190–204, 2000.

[13] E. Pamukçu and M. N. Çankaya, "Information complexity criterion for model selection in robust regression using A new robust penalty term," 2020, https://arxiv.org/abs/2012.02468.

[14] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[15] J. H. Ryoo, "Model selection with the linear mixed model for longitudinal data," *Multivariate Behavioral Research*, vol. 46, no. 4, pp. 598–624, 2011.

[16] A. E. McGlothlin and K. Viele, "Bayesian hierarchical models," *The Journal of the American Medical Association*, vol. 320, no. 22, pp. 2365-2366, 2018.

[17] L. Wasserman, "Bayesian model selection and model averaging," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, 2000.

[18] K. Vanbrabant, Y. Boddez, P. Verduyn, M. Mestdagh, D. Hermans, and F. Raes, "A new approach for modeling generalization gradients: a case for hierarchical models," *Frontiers in Psychology*, vol. 6, p. 652, 2015.

[19] F. M. Hollenbach and J. M. Montgomery, "Bayesian model selection, model comparison, and model averaging," in *The SAGE Handbook of Research Methods in Political Science and International Relations*, pp. 937–960, Sage Publications Ltd, Thousand Oaks, CA, USA, 2020.

[20] F. Korner-Nievergelt, T. Roth, S. von Felten, J. Guélat, B. Almasi, and P. Korner-Nievergelt, *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan*, Elsevier Inc, Amsterdam, Netherlands, 2015.

[21] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Selected Papers of Hirotugu Akaike*, Springer, pp. 199–213, Berlin, Germany, 1998.

[22] N. J. Evans, "Assessing the practical differences between model selection methods in inferences about choice response time tasks," *Psychonomic Bulletin and Review*, vol. 26, no. 4, pp. 1070–1098, 2019.

[23] J. Piironen and A. Vehtari, "Comparison of Bayesian predictive methods for model selection," *Statistics and Computing*, vol. 27, no. 3, pp. 711–735, 2017.

[24] A. Vehtari and J. Ojanen, "A survey of Bayesian predictive methods for model assessment, selection and comparison," *Statistics Surveys*, vol. 6, pp. 142–228, 2012.

[25] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[26] J. Brommer, B. Class, and G. Covarrubias-Pazaran, "Multivariate mixed models in ecology and evolutionary biology: inferences and implementation in R," 2019, https://ecoevorxiv.org/repository/view/4411/.

[27] J. K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, Elsevier Science, Amsterdam, Netherlands, 2nd edition, 2014.

[28] T. Wang and E. C. Merkle, "MerDeriv: derivative computations for linear mixed-effects models with application to robust standard errors," *Journal of Statistical Software*, vol. 87, 2018.

[29] S. N. Wood, "Linear mixed models," *Generalized additive models*, pp. 61–100, 2018.

[30] J. S. Hodges, "An opinionated survey of methods for mixed linear models," *Richly Parameterized Linear Model*, pp. 43–88, 2020.

[31] W. Zhang, C. Leng, and C. Y. Tang, "A joint modelling approach for longitudinal studies," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 77, no. 1, pp. 219–238, 2015.

[32] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, p. 963, 1982.

[33] Y. Lee and J. A. Nelder, "Hierarchical generalized linear models," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 4, pp. 619–656, 1996.

[34] P. XK. Song, "Mixed-effects models: bayesian inference," in *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, Berlin, Germany, 2007.

[35] W. W. Stroup, *Generalized Linear Mixed Models Modern Concepts, Methods and Applications*, CRC Press, Boca Raton, FL, USA, 2012.

[36] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, no. 116, pp. 3571–3594, 2010.

[37] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[38] H. Bozdogan, *Icomp: A New Model-Selection Criteria*, Elsevier Science Publishers, Amsterdam, Netherlands, 1988.

[39] H. Bozdogan, "Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions," *Journal of the Psychometric Society*, vol. 52, no. 3, pp. 345–370, 1987.

[40] H. Bozdogan, "Intelligent statistical data mining with information complexity and genetic algorithms," *Statistical Data Mining and Knowledge Discovery*, pp. 15–56, 2003.

[41] H. Bozdogan, "On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models," *Communications in Statistics-Theory and Methods*, vol. 19, no. 1, pp. 221–278, 1990.

[42] H. Bozdogan and D. M. A. Haughton, "Informational complexity criteria for regression models," *Computational Statistics and Data Analysis*, vol. 28, no. 1, pp. 51–76, 1998.

[43] H. Bozdogan, "A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation," *Istanbul University Faculty of Business Administration Journal*, vol. 39, no. 2, pp. 370–398, 2009.

[44] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 62–91, 2000.

[45] P. D. Congdon, *Book Bayesian Hierarchical Models: With Applications Using R*, CRC Press, Boca Raton, FL, USA, 2nd edition, 2020.

[46] S. M. O'Brien and D. B. Dunson, "Bayesian analyses of multivariate binary or categorical outcomes typically rely on probit or mixed-effects logistic regression models that do not have a marginal logistic structure for the individual outcomes," *Addition, Difficulties Arise when Simple N*, vol. 60, no. 3, 2004.

[47] J. Lin, M. F. Myers, L. M. Koehly, and C. S. Marcum, "A Bayesian hierarchical logistic regression model of multiple informant family health histories," *BMC Medical Research Methodology*, vol. 19, no. 1, 2019.

[48] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, Cambridge, UK, 2006.

[49] J. M. Hilbe, "Data analysis using regression and multilevel/hierarchical models," *Journal of Statistical Software*, vol. 30, pp. 530–548, 2009.

[50] J. P. Loenneke, C. A. Fahs, L. M. Rossow et al., "Effects of cuff width on arterial occlusion: implications for blood flow restricted exercise," *European Journal of Applied Physiology*, vol. 112, no. 8, pp. 2903–2912, 2012.

[51] D. F. Percy, "Blocked arteries and multivariate regression," *Biometrics*, vol. 48, no. 3, p. 683, 1992.

[52] A. Denton and S. Brownhill, "Doing bayesian data analysis: a tutorial with r and bugs," *Becoming a Brilliant Trainer*, Routledge, England, UK, pp. 114–128, 2019.

[53] M. Nishio and A. Arakawa, "Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values," *Genetics Selection Evolution*, vol. 51, no. 1, 2019.

[54] H. Xiong, S. Szedmak, and J. H. Piater, "Comparing binary Hamiltonian Monte Carlo and Gibbs sampling for training discrete MRFs with stochastic approximation," 2013, https://iis.uibk.ac.at/public/papers/Xiong-2014-AISTATS.pdf.

[55] B. B. Yimer and Z. Shkedy, "Bayesian inference for generalized linear mixed models: a comparison of different statistical software procedures," *RMS: Research in Mathematics and Statistics*, vol. 8, no. 1, Article ID 1896102, 2021.