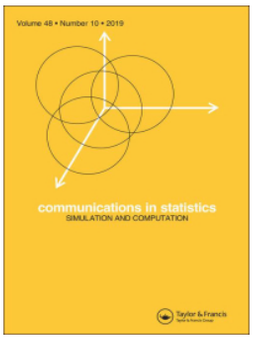

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors

Oludare Ariyo, Emmanuel Lesaffre, Geert Verbeke & Adrian Quintero

To cite this article: Oludare Ariyo, Emmanuel Lesaffre, Geert Verbeke & Adrian Quintero (2022) Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors, Communications in Statistics - Simulation and Computation, 51:4, 1591-1615, DOI: 10.1080/03610918.2019.1676439

To link to this article: https://doi.org/10.1080/03610918.2019.1676439

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2022, VOL. 51, NO. 4, 1591–1615
https://doi.org/10.1080/03610918.2019.1676439

Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors

Oludare Ariyo[a,b], Emmanuel Lesaffre[a], Geert Verbeke[a], and Adrian Quintero[a]

[a]Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium; [b]Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria

ABSTRACT

We explore the performance of three popular Bayesian model-selection criteria when vague priors are used for the covariance parameters of the random effects in a linear mixed-effects model (LMM) using an extensive simulation study. In a previous paper, we have shown that the conditional selection criteria perform worse than their marginal counterparts. It is known that for some "vague" priors, their impact on the estimated model parameters can be non-negligible, e.g., for the priors of the covariance matrix of the random effects in a longitudinal LMM. We evaluate here the impact of vague priors for the covariance matrix of the random effects on selecting the correct LMM using classical Bayesian selection criteria. We consider marginal and conditional criteria. For the random intercept case, we assign different vague priors to the variance parameters. With two or more random effects, we considered five different specifications of inverse-Wishart (IW) prior, five different separation priors and a joint prior. The results show again the better performance of the marginal over the conditional criteria and the superiority of joint and separation priors over IW in all settings. We also illustrate the performance of the selection criteria on a practical dataset.

ARTICLE HISTORY
Received 14 June 2019
Accepted 30 September 2019

KEYWORDS
Covariance matrices; Linear mixed-effects models; Model selection criteria; Vague priors

MATHEMATICAL SUBJECT CLASSIFICATION
62PXX

1. Introduction

The linear mixed-effects model (LMM) is a popular model to analyze longitudinal data with a Gaussian response, especially when the outcomes have been recorded at irregular time points. The model consists of fixed effects and random effects. The fixed effects represent the effect of covariates on the population average, while the random effects represent individual-specific deviations in profiles and account for the correlation among responses from the same individual. Selecting the appropriate LMM implies determining the appropriate fixed effects part and random effects part such that the model fits the current and future data well.

In a Bayesian framework, there is little agreement on the appropriate model selection criteria. Three criteria are currently popular in practice. The deviance information criterion (DIC) is by far the most popular criterion because it can be easily obtained with

CONTACT Oludare Ariyo oludaresamuel.ariyo@kuleuven.be Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KU Leuven, Kapucijnenvoer 35, block D, bus 7001 B-3000, Leuven, Belgium.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

the popular Bayesian software packages WinBUGS and OpenBUGS. The pseudo-Bayes factor (PSBF) and the widely applicable information criterion (WAIC) are increasingly in use but are not automatically obtained in the classical Bayesian packages, except for WAIC which is provided by Stan (Carpenter et al. 2017). These three model selection criteria may be computed on the hierarchical specification of the LMM, i.e., given the random effects. This then leads to the conditional version of the selection criteria. However, the marginal version of the LMM, i.e., the model averaged over the distribution of the random effects, can be analytically determined. Selection criteria based on this marginal likelihood are then referred to as marginal selection criteria. It is most popular, but also in general easier, to fit the hierarchical version of the LMM in Bayesian software thereby making use of the data augmentation algorithm. Indeed, the conditional version of DIC is provided by most Bayesian statistical packages, but also for the other selection criteria the conditional version is easy to compute from the generated Markov chain Monte Carlo (MCMC) samples. Consequently, marginal versions of the selection criteria are basically never reported. The conditional criteria have, however, been criticized in the literature (Chan and Grant 2016; Merkle, Furr, and Rabe-Hesketh 2018; Ariyo et al. 2019). Theoretical arguments and simulation results point out that model selection based on conditional criteria is inferior to model selection based on marginal criteria. This was for instance shown in Ariyo et al. (2019) for the LMM. Here, we examine the impact of vague priors on the model parameters on the performance of the model selection criteria. Given the inferior results of the conditional selection criteria, we are particularly interested to see whether the marginal selection criteria highly depend on the chosen vague priors for the model parameters. However, since we realize that the conditional selection criteria will remain popular despite the theoretical and empirical evidence, we also checked the impact of the vague priors on the conditional selection criteria.

While for the fixed effects most often normal priors with a large variance are chosen, there is no standard choice for the vague prior of the variance terms of the random effects in LMMs (Kass and Natarajan 2006). The impact of a vague prior on the posterior distributions can also be more pronounced when the dataset is small and/or the number of units contributing to the estimation of the between-unit variation is small (Lambert et al. 2005). In this situation, Lambert et al. (2005) argued that informative prior distributions are required.

When the LMM involves two or more random effects, a prior on their covariance matrix is required. The inverse-Wishart (IW) distribution is the natural choice for a covariance matrix due to its conditional conjugacy. However, problems have been reported with the use of the IW prior as it assumes the same amount of prior information for every variance parameter. More importantly, it assumes a prior relationship between the variances and correlations (Alvarez, Niemi, and Simpson 2014). These issues have a larger impact when the dimension of the covariance matrix increases. Several alternative priors for the covariance matrix have been suggested in the literature. Firstly, the IW prior has been given an hierarchical structure. Secondly, various priors have been suggested separating the priors on the variance and correlation parameters. Such separation priors has been shown in the literature to be more efficient than the classical IW prior (Huang and Wand 2013; Alvarez, Niemi, and Simpson 2014). Among

the merits of separation priors is their flexibility in incorporating informative prior information. However, things become somewhat more complicated with three or more random effects because certain restrictions must be imposed to ensure positive definiteness of the covariance matrix (Barnard, McCulloch, and Meng 2000; Huang and Wand 2013; Wei and Higgins 2013; Hurtado Rúa, Mazumdar, and Strawderman 2015). Other priors in which the variance terms of both the measurement errors and the variance-covariance of random effects are modeled jointly have been suggested. These priors have been shown to reduce bias and improve efficiency in the posterior inference (Demirhan and Kalaylioglu 2015; Kalaylioglu and Demirhan 2017). Just like for separation priors, certain restrictions are needed to ensure the positive-definite of the covariance matrix.

The aim of this study is to ascertain if the choice of the vague prior, especially on variance and covariance parameters, is important for model selection. More specifically, we wish to measure how much different vague priors impact the marginal selection criteria, but given their popularity we also checked this for the conditional criteria. The remainder of this article is organized as follows. In Sec. 2, we introduce the Bayesian linear mixed model for longitudinal data. We present the model selection criteria in Sec. 3. In Sec. 4, previous findings are discussed while in Sec. 5 we explore the vague prior for the covariance matrix of the random parameters. In Sec. 6, we assess the sensitivity of different vague covariance priors on random effects on the performance of the above-mentioned model selection criteria using a simulation study. An illustration on a practical dataset is shown in Sec. 7. Main conclusions and a discussion are given in Sec. 8.

2. The linear mixed-effects model (LMM)

Let $\mathbf{Y}_i = (y_{m_i i}, ..., y_{m_i, i})^T$ be an m_i-dimensional response vector of (longitudinal) measurements for the ith (independent) individual, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(m_i \times p)$ and $(m_i \times q)$-dimensional covariate matrices, respectively and $\boldsymbol{\beta}$ a p-dimensional vector of fixed effects. The classical LMM is then given as (Laird and Ware 1982)

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (i = 1, ..., n), \tag{1}$$

with the residual component vector $\boldsymbol{\epsilon}_i \sim N_{m_i}(0, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an $(m_i \times m_i)$ positive-definite covariance matrix with $\boldsymbol{\Sigma}_i = \sigma^2_\epsilon \mathbf{I}_{m_i}$ where $\mathbf{I}_{m_i}$ denotes the identity matrix of dimension m_i. The $q \times 1$ random effects vectors are also assumed normally distributed, i.e., $\mathbf{b}_i \sim N_q(0, \mathbf{D})$, where $\mathbf{D}$ is a $(q \times q)$ positive-definite covariance matrix. Model (1) is called the LMM because it combines the fixed-effects structure $\boldsymbol{\beta}$ with the subject-specific random effects $\mathbf{b}_1, ..., \mathbf{b}_n$. Inference may be focused on the regression coefficients $\boldsymbol{\beta}$, the unit-specific coefficients $\mathbf{b}_i$ or the variance components ($\boldsymbol{\Sigma_i} = \sigma^2 I_{m_i}$ and $\mathbf{D}$). Model (1) is the hierarchical version of the LMM, which provides the conditional LMM likelihood. The marginal version of the LMM is obtained as follows. Let $f(\mathbf{Y}_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ be the (Gaussian) density functions of $\mathbf{Y}_i$ and random effects respectively, then the marginal density function of $\mathbf{Y}_i$ is given by $f(\mathbf{Y}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{D}) = \int f(\mathbf{Y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \sigma^2) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$. It can easily be shown that, with Gaussian densities, the marginal version of (1), is a multivariate normal distribution given by

$$\mathbf{Y}_i \sim N_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i), \quad (i = 1, ..., n). \tag{2}$$

The Bayesian LMM is obtained when prior distributions are given for all model parameters. Hence, additional to model (1) or equivalently model (2) we specify priors for $\boldsymbol{\beta}, \mathbf{D}$ and σ^2. Classical choices for these priors are: $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \mathbf{B}_0), \mathbf{D} \sim \text{IW}(k, \mathbf{V})$ and $\sigma^{-2} \sim \text{Gamma}(\nu_0, \delta_0)$.

Typically, one needs MCMC methods to estimate the model parameters, such as Gibbs sampling or Metropolis–Hastings algorithm (Geman and Geman 1993; Lesaffre and Lawson 2012).

3. Bayesian model selection

Model selection is an important step in a statistical modeling exercise. In a frequentist context one distinguishes model selection for nested models versus model selection with non-nested models. In the first case formal tests, most often likelihood ratio tests, are used, while in the second case typically information criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are in use. In a Bayesian context, the same model selection criteria apply for nested and non-nested models.

The DIC is an adaptation of AIC to the Bayesian context. DIC is the most popular Bayesian model selection criterion, because it has been implemented in the popular software WinBUGS, and later also in other popular Bayesian software such as OpenBUGS. However, the literature has been critical about its theoretical foundations. This is sometimes reflected in practice when the associated degrees of freedom, p_{DIC}, is estimated negative thereby making the criterion useless (Spiegelhalter et al. 2014). As a result, there has been increasing interest to use other criteria, such as the PSBF and the WAIC. WAIC has been recently advocated as having a similar flavor as DIC but with better properties (Watanabe 2010; Millar 2018). There is, however, still no consensus about the best criteria for model selection in a Bayesian context.

For reasons of completeness, we will discuss the three most popular Bayesian model selection criteria in more detail.

3.1. The deviance information criterion

The DIC (Spiegelhalter et al. 2002) was developed for Bayesian model selection and is derived from AIC by replacing frequentist concepts by their Bayesian counterparts. As such, DIC expresses the predictive accuracy of the model in a Bayesian way. The frequentist mean is replaced by the posterior mean of the model parameter, i.e., $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y})$, and frequentist integration is replaced by Bayesian integration. DIC is then defined as

$$\text{DIC} = -2\log p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + 2p_{\text{DIC}}, \tag{3}$$

where p_{DIC} corresponds to the effective number of parameters, given by

$$p_{\text{DIC}} = -2\ E_{\boldsymbol{\theta}|\mathbf{y}}\big[\log p(\mathbf{y}|\boldsymbol{\theta})\big] + 2\log\big[p(\mathbf{y}|\bar{\boldsymbol{\theta}})\big],$$

which quantifies the number of parameters to be estimated after incorporating the prior information into the model. From (3) it is clear that low values of DIC indicate a better

fit of the model to the data. DIC is popular in practice because it is a by-product of the MCMC calculations and implemented in popular Bayesian software. Namely, with the deviance given by $\overline{D(\boldsymbol{\theta})} = -2\log p(\mathbf{y}|\boldsymbol{\theta})$, p_{DIC} and DIC can be approximated by making use of $\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^K$, which are the sampled values of $\boldsymbol{\theta}$ from a converged MCMC chain. We then have $p_{\text{DIC}} = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ and $\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_{\text{DIC}}$, where $\overline{D(\boldsymbol{\theta})} \approx \frac{1}{K}\sum_{k=1}^{K} D(\boldsymbol{\theta}^k)$ and $D(\bar{\boldsymbol{\theta}}) \approx D\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}^k\right)$. We note that there are different versions of DIC implemented in the popular Bayesian packages, where the difference is primarily due to a different definition of p_{DIC}.DIC (and p_{DIC}) have received considerable criticism in the statistical literature. First of all, DIC is not invariant to monotonic parameter transformations, i.e., DIC changes value when based on $\boldsymbol{\psi} = h(\boldsymbol{\theta})$ rather than on $\boldsymbol{\theta}$. Furthermore, it has been shown that the asymptotic properties upon which DIC is based, are not fulfilled in hierarchical models (Li, Zeng, and Yu 2012), see also Sec. 4. Note also that it is not clear how to compute DIC when there are missing responses. For this reason, different versions of DIC have been explored in Celeux et al. (2006).

3.2. The pseudo Bayes factor

A natural Bayesian selection mechanism is to choose the model with the largest posterior probability. Suppose that there are L models $M_1, ..., M_L$ to choose from, with prior probabilities $p(M_1), ..., p(M_L)$, respectively. The posterior probability of model M_ℓ is determined by computing the marginal likelihoods $p(\mathbf{y}|M_\ell) = \int p(\mathbf{y}|\boldsymbol{\theta}_\ell, M_\ell)\ p(\boldsymbol{\theta}_\ell|M_\ell)\ d\boldsymbol{\theta}_\ell$ $(\ell = 1, ..., L)$ and is given by

$$p(M_\ell|\mathbf{y}) = \frac{p(M_\ell)p(\mathbf{y}|M_\ell)}{\sum_k p(M_k)p(\mathbf{y}|M_k)} = \frac{p(M_\ell)BF_{1,2}[M_\ell : M_b]}{\sum_k p(M_k)BF_{1,2}[M_k : M_b]}, \tag{4}$$

where $BF_{1,2}[M_\ell : M_b]$ is the Bayes factor, which compares model M_ℓ to a reference model M_b and is given by

$$BF_{1,2}[M_\ell : M_b] = \frac{p(\mathbf{y}|M_\ell)}{p(\mathbf{y}|M_b)}.$$

The classical Bayes factor is difficult to use in practice because: (1) the marginal likelihood is not defined for improper priors, (2) priors must be well chosen otherwise the classical Lindley–Bartlett paradox (Bernardo 1980) comes into play, and (3) its computation can be very demanding, sometimes even worse than computing the posterior distribution (Lesaffre and Lawson 2012, p. 273). Several versions of the original Bayes factor have been suggested to make the computations feasible and practical. A popular version is the PSBF, where the numerator and denominator in (4) are replaced by the product of the marginal likelihoods over all subjects, whereby the marginal likelihood for the ith subject is evaluated in $\mathbf{y}_i$ and is based on the posterior of the model parameters obtained from all other subjects, i.e., from $\mathbf{y}_{(i)}$. This yields a ratio of two pseudo-likelihoods, each being the product of n conditional predictive ordinates (CPOs) (Gelfand and Dey 1994). The CPO for subject i under model M_ℓ is the probability of observing $\mathbf{y}_i$ given model M_ℓ fitted with all observations in the sample except for $\mathbf{y}_i$ i.e., $\text{CPO}_{i,\ell} = p(\mathbf{y}_i|\mathbf{y}_{(i)}, M_\ell)$. The CPO can be approximated making use of the converged

MCMC sample $\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^K$ as follows:

$$\text{CPO}_{i,\ell} \approx \left[\frac{1}{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}_\ell^k, \text{M}_\ell)}\right]^{-1}.$$

To compute the PSBF, the log-pseudo likelihood (LPML) for each model is computed by summing up $\text{CPO}_{i,\ell}$ across the n subjects, i.e., $\text{LPML}_\ell = \sum_{i=1}^{n} \log(\text{CPO}_{i,\ell})$. Then to compare two models M_1 and M_2, the pseudo-Bayes factor $\text{PSBF}_{1,2}$ favors model M_1 to model M_2 when $\text{PSBF}_{1,2} = \exp(\text{LPML}_2 - \text{LPML}_1) < 1$. Note, that we have adapted the original definition of PSBF in order that small values imply better models. In contrast to DIC, the PSBF is invariant to monotonic parameter transformations.

3.3. The widely applicable information criterion

The WAIC (Watanabe 2010) measures the predictive accuracy of the model based on the log-posterior predictive distribution $\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)$ of the parameter vector $\boldsymbol{\theta}$ for a future observation $\tilde{\mathbf{y}}_i$. The predictive accuracy for a future unknown $\tilde{\mathbf{y}}_i$ is expressed by the log-predictive distribution (elpd) as $\text{elpd}_i = E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)] = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i$, and f is the unknown distribution under the true model. The measure of predictive accuracy can also be described with a point estimate $\bar{\boldsymbol{\theta}}$, often taken equal to $E(\boldsymbol{\theta}|\mathbf{y})$, as the expected log predictive distribution given the point estimator $\text{elpd}_{\bar{\theta}} = E_f(\log p(\tilde{y}|\bar{\boldsymbol{\theta}}))$. The log pointwise predictive distribution (lppd) based on the observed data $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$, is calculated as follows $\text{lppd} = \log \prod_{i=1}^{n} p_{\boldsymbol{\theta}|\mathbf{y}}(\mathbf{y}_i) = \sum_{i=1}^{n} \log \int_{\boldsymbol{\theta}} p(\mathbf{y}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. In practice, lppd can be estimated with the converged MCMC sample $\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^K$ from the posterior distribution as $\widehat{\text{lppd}} = \sum_{i=1}^{n} \log\left[\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. The expected log pointwise predictive density elppd is estimated as the log pointwise predictive distribution lppd with a bias correction using the WAIC criterion $\widehat{\text{elppd}}_{\text{WAIC}} = \widehat{\text{lppd}} - p_{\text{WAIC}}$. The measure p_{WAIC} corresponds to the estimate of the effective number of parameters given by $p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left[\log(\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k)) - \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. WAIC can be alternatively expressed as $\widehat{\text{lppd}} = \sum_{i=1}^{n} \log\left[\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. $\text{WAIC} = -2\widehat{\text{lppd}} + 2p_{\text{WAIC}}$. As for PSBF, WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, with h a strictly monotone function. As for DIC, smaller values of WAIC indicate a better model.

4. Previous findings

In this article, we evaluate the dependence of vague priors on the performance of Bayesian selection criteria for the linear mixed model. As seen in, e.g., Quintero and Lesaffre (2018), the selection criteria can be based on the hierarchical or conditional version of the LMM given by model (1) or on the marginal version of the LMM based on model (2). In the first case, one speaks of a conditional selection criterion. We have for the above three popular criteria the conditional DIC (cDIC), the conditional PSBF (cPSBF) and the conditional WAIC (cWAIC). In the second case, we have the marginal

versions of the criteria denoted here as: mDIC, mPSBF, and mWAIC. It has been argued that the choice of likelihood (conditional or marginal) should be motivated by the aim of the study (Vaida and Blanchard 2005). For example, in a clinical trial that evaluates a new drug on patients enrolled within an hospital, cDIC may be used to conduct model selection if the interest lies on the efficacy of the new drug at the hospitals of the study. However, mDIC is the appropriate model selection criterion when one wishes to evaluate the efficacy of the new drug in all hospitals. In the statistical literature, there is evidence of the better performance of the marginal model selection criteria. In a slightly different setting, Chan and Grant (2016) observed in a simulation study that cDIC usually selects an overfitted model but that mDIC performs better. In general hierarchical models, Quintero and Lesaffre (2018) concluded via a simulation study that mDIC selects (much) more often the correct model than cDIC. They also provided R software to compute the marginal criteria via a dedicated sampling algorithm. The same result was obtained for cWAIC by Millar (2018) and therefore he recommended to use mWAIC. There is also evidence that the same is true for an item response model (Li et al. 2016; Merkle, Furr, and Rabe-Hesketh 2018; Millar 2018). Furthermore, Ariyo et al. (2019) compared the conditional and marginal versions of DIC, PSBF and WAIC for the Gaussian LMM, the skew-normal LMM (SNLMM) and the skew-*t* linear mixed model (STLMM) via an extensive simulation study. Both the balanced as well as the unbalanced case was studied for longitudinal data. Both the random intercept case as well as the 2- and 3-dimensional case for the random effects part were considered. The simulation results showed a strong advantage of the marginal criteria in selecting the true data-generating model. Since the marginal likelihood for the LMM, SNLMM and STLMM have a closed form it is relatively easy to compute these marginal criteria. In addition, the model selection performance of the conditional criteria decreases with increasing sample size (increasing number of random effects), while the performance of the marginal criteria improves with increase in sample size. Furthermore, to facilitate the computations of the marginal criteria in practice, R functions were developed, which can be downloaded from https://ibiostat.be/online-resources/bayesian. In the course of this study, it was observed that the choice of the prior may affect the ability of the criteria to select the appropriate data-generating model, especially for small sample sizes. The appropriateness of vague priors is most often checked by evaluating their effect on estimation, but here we check their impact on the performance of DIC, PSBF and WAIC in selecting the correct model. More specifically, we wish to check for the conditional and marginal criteria: (1) the impact of vague priors on selecting the best model and (2) whether there is a best vague prior in this context. In principle, we could have limited ourselves to the marginal version of the criteria since the conditional criteria showed repeatedly not to perform well. But, since many will continue to use the conditional criteria because of their practical advantage, the above aims are of practical interest.

5. Vague prior distributions for the LMM

An essential step in statistical modeling is the choice of the appropriate statistical model for the data at hand. This is not only an essential step, but also a notoriously complex part of statistical modeling involving statistical tools and substantive knowledge. In this

article, we look at model selection in a Bayesian context. That is, we assume that we have a rather limited number of models to choose from. In the model selection step it is customary to choose vague priors for the model parameters. In contrast, informative priors are typically chosen when an appropriate model is already available. For a LMM (vague) priors must be specified for the fixed effects and the variance components. For the fixed effects we have taken vague normal priors. Here, we focus on the vague priors for the covariance matrix of the random effects. We consider the univariate case of a random intercept and the multivariate case of several random effects.

5.1. Vague priors for the random intercept

Various vague priors have been suggested for the level-2 variance of the Gaussian hierarchical model. This model is a special case of the LMM with only a random intercept. In that case, (1) can be written as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta} + \mathbf{1}_{m_i} b_i, \sigma_\epsilon^2), \quad i = 1, ..., n, \tag{5}$$

with $\mathbf{1}_{m_i}$ is a $m_i \times 1$ vector of ones, and where the random intercept $b_i \sim N(0, \sigma_b^2)$. The improper prior $p(\sigma^2) \propto 1/\sigma^2$, suggested by Jeffreys for the simple case of $N(\mu, \sigma^2)$ yields an improper posterior for model (5) if applied to σ_b^2. This was recognized long time ago, see, e.g., Lesaffre and Lawson (2012). In the early days of the development and use of WinBUGS, this improper prior was replaced by $\sigma_b^2 \sim \text{IG}(0.001, 0.001)$, where $\text{IG}(\varepsilon, \varepsilon)$ refers to an inverse gamma distribution with two parameters equal to ε. Later on, it was realized that the posterior on σ_b^2 depends much on the choice of the value of ε. This was a trigger to suggest alternative vague but proper priors for σ_b^2. We note that, in contrast to above Jeffreys prior, the proper vague priors depend on the scale of the data. Hence, the vague prior distributions for σ_b listed below are not invariant to change of scale in the data. The following vague but proper priors for σ_b^2 have been considered in the literature:

1. $\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.001, 0.001)$. This was a popular prior distribution for variance terms used initially in the WinBUGS Examples I and II documents (Lunn et al. 2000);
2. $\log(\sigma_b^2) \sim \text{Uniform}(-10, 10)$. This prior distribution was suggested in the analysis of cluster randomized trials (Spiegelhalter 2001);
3. $\frac{1}{\sigma_b^2} \sim \text{Pareto}(1, 0.001)$. This prior was suggested in genetic epidemiology models (Burton et al. 1999; Scurrah, Palmer, and Burton 2000) and is equivalent to $\text{Uniform}(0, 1000)$ on the variance scale;
4. $\sigma_b \sim \text{Uniform}(0, 100)$. This prior was recommended by Spiegelhalter, Abrams, and Myles (2004);
5. $\sigma_b \sim \text{half}-t(0, 1, 1)$. Gelman (2006) suggested the use of half-t prior with $df = 1$ (half-Cauchy) on the standard deviation when the number of groups is small. Since a half-t prior appear to be completely harder to work with Huang and Wand (2013), we give the precision parameter a scaled gamma distribution which is equivalent to a half-Cauchy prior (with mean zero) on the standard deviation (Wand et al. 2011).

Note that the above priors are appropriate for the scale of the simulated data, but also for the scale of the data in the analysis of the chicken dataset in Sec. 7. Furthermore, the prior for the variance of the measurement error, σ_ϵ^2 is given an $\text{IG}(0.001, 0.001)$ prior, which is a classical choice.

5.2. Vague priors for the covariance matrix of the random effects

Specifying an appropriate prior for a covariance matrix has been the topic of intensive research in the last two decades. The mathematically convenient prior for a covariance matrix is given by the IW distribution. This prior is often used in Bayesian modeling for an unknown covariance matrix due to its conditional conjugacy and its implementation in most of the Bayesian statistical software, but there are practical problems with this prior. In next subsections we review the problems involved with the IW prior, then we discuss some generalizations of this prior to improve convergence properties and its ability to represent (absence of) prior knowledge in an appropriate manner. Note that the same inverse gamma prior for σ_ϵ^2 will be taken as in Sec. 5.1.

5.2.1. The IW prior and variations

The conditional conjugate prior for the covariance matrix $\boldsymbol{D}$ in the linear mixed model (1) is the IW distribution (Lesaffre and Lawson 2012; Schervish 2012)

$$\boldsymbol{D} \sim \text{IW}(k, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is a $q \times q$ positive semi-definite scale matrix and $k(\geq q)$ is the *df*. $\boldsymbol{V}$ is used to position the IW distribution in the parameter space, and k sets the certainty about the prior information in the scale matrix (Hurtado Rúa, Mazumdar, and Strawderman 2015). For instance, to obtain a minimally informative prior, $k \approx q$ appears appropriate (Gelman and Hill 2007; Gelman et al. 2014). When $k = q + 1$, the marginal distribution of the correlations is uniform, but their joint distribution is not (Tokuda et al. 2011). Further, the larger k, the more informative is the IW distribution (Gelman and Hill 2007; Gelman et al. 2014). In JAGS, the standard choice is to take small values for the diagonal elements of $\boldsymbol{V}$ with the degrees of freedom set equal to the dimension of the matrix. However, setting the diagonal elements to larger values also influences the position of the IW (Schnell et al. 2016). In other words, specifying an IW prior distribution requires balancing the size of $\boldsymbol{V}$ and the value of k, but it is not clear how to choose the diagonal elements in $\boldsymbol{V}$. In addition, varies studies have shown that the IW prior is problematic, namely: (1) there is over-dependence in the posterior distribution of the covariance matrix when data is sparse (i.e., small number of clusters), (Gelman 2006; Quintero and Lesaffre 2017); (2) the uncertainty for all variances is controlled by a single degree of freedom parameter (Gelman et al. 2004); (3) there is a priori dependence between the standard deviations and the correlation (Tokuda et al. 2011); and (4) the marginal distribution for the variances has low density in a region near zero (Gelman 2006). In addition, convergence may be difficult with the IW prior. This triggered Gelman et al. (2008) to suggest parameter expansion techniques, which primarily improve the convergence of the MCMC algorithm. Variations of the classical IW prior have been suggested to improve convergence of the MCMC computations, and to better

express (absence of) prior information. O'Malley and Zaslavsky (2008) suggested the scaled IW prior, which is based on the IW prior but with additional parameters to better specify the prior information on the variances. Another variation is suggested by Huang and Wand (2013), who suggested an hierarchical IW prior for $\boldsymbol{D}$:

$$\begin{aligned} &\boldsymbol{D} \mid d_1, ..., d_q \sim \text{IW}(v + q - 1, 2v\text{diag}(1/d_1, ..., 1/d_q)), \\ &d_k \sim \text{IG}(1/2, 1/A_k^2), k = 1, ..., q, \end{aligned} \tag{6}$$

where $\text{diag}(1/d_1, ..., 1/d_q)$ denotes a diagonal matrix with $1/d_1, ..., 1/d_q$ on the diagonal and v, $A_1, ..., A_q$ are positive scalars. The authors showed that (6) produces $\text{half}-\text{t}(v, A_k)$ distributions for each standard deviation of $\boldsymbol{D}$ and that it is a matrix generalization of the half-t prior of Gelman (Gelman 2006). Large values of A_k imply a weakly informative prior on standard deviations as in Gelman (2006). Huang and Wand (2013) also showed that the choice of $v = 2$ leads to marginal uniform distributions for correlation terms $\rho_{j,k}, j \neq k$. This prior will be evaluated in our simulated study and will be referred to as *HIW prior*, more specifically as $\text{HIW}(v, \boldsymbol{A})$, with $\boldsymbol{A} = \{A_1, ..., A_q\}$. The performance of both variations on the IW prior has been evaluated in a simulation study (Alvarez, Niemi, and Simpson 2014), who concluded that these priors show good performance and are definitely much better than the classical IW when the true variance is small relative to the prior mean, which holds for larger sample sizes.

5.2.2. Separation strategies for modeling covariance matrices

Another class of priors is based on the separation strategy, first suggested by Barnard, McCulloch, and Meng (2000). The idea is to decompose the variance covariance matrix $\boldsymbol{D}$ as $\boldsymbol{D} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^{\frac{1}{2}}$, where $\boldsymbol{S}^{\frac{1}{2}}$ is a diagonal matrix with standard deviations as elements and $\boldsymbol{R}$ is a $q \times q$ matrix of correlations. The next two vague priors are based on this separation technique. The two correlation priors will be combined with uniform priors on [0,100] for the elements of $\boldsymbol{S}$, i.e., the variances. In the first proposal, the correlation matrix $\boldsymbol{R}$ is factorized as $\boldsymbol{R} = \boldsymbol{L}^T\boldsymbol{L}$, where $\boldsymbol{L}$ is a $q \times q$ upper-triangular matrix. A prior is then placed on the $q(q+1)/2$ elements in $\boldsymbol{L}$, i.e., the Cholesky factors L_{ij} $(i = 1, ..., q, i \leq j)$. The following prior ensures unconstrained estimation of variance-covariance matrix and that the positive semi-definite condition is satisfied (Wei and Higgins 2013):

$$\begin{aligned} \boldsymbol{L}_{1j} &\sim U(-1, 1), \\ \boldsymbol{L}_{jj} &= \sqrt{1 - \sum_{i=1}^{j-1} L_{ij}^2}, \\ \boldsymbol{L}_{ij} &= U\left(-\sqrt{1 - \sum_{i=1}^{j-1} L_{kj}^2}, \sqrt{1 - \sum_{i=1}^{j-1} L_{ij}^2}\right), i < j \end{aligned} \tag{7}$$

for $j = 2, ..., q$, with $L_{11} = 1$ to ensure uniqueness. This prior will be referred to as the *Chol* prior.

Another approach is to use the spherical decomposition of the correlation matrix first suggested by Pinheiro and Bates (1996). In this approach, the Cholesky decomposition is parametrized by sine and cosine functions as follows. Setting $L_{11} = 1$ and let $k = 2, ..., q$, we have

$$
\begin{aligned}
L_{k1} &= \cos(\phi_{k2}),\\
L_{k2} &= \sin(\phi_{k2})\cos(\phi_{k3}),\\
&\vdots\\
L_{k,k-1} &= \sin(\phi_{k2})\sin(\phi_{k3})\ldots\cos(\phi_{kk}),\\
L_{k,k} &= \sin(\phi_{k2})\sin(\phi_{k3})\ldots\sin(\phi_{kk}).
\end{aligned}
$$

Uniform $(0,\pi)$ priors, with $\pi = 3.1415$, are given to the ϕ_{km} parameters to ensure the uniqueness of the spherical parametrization. Note that the (i, j)th element of $\boldsymbol{R}$ is the inner product $\boldsymbol{L}_i^T\boldsymbol{L}_j$ and $\boldsymbol{L}_k^T\boldsymbol{L}_k = 1$, where $\boldsymbol{L}_k$ is the kth column of $\boldsymbol{L}$. This prior is referred to as *spherical* prior.

Daniels and Kass (1999) proposed a separation prior that puts a distribution on the correlations so that they will end up shrinking toward 0. To this end, they proposed a normal distribution for the Fisher's z-transform on each of the $q(q-1)/2$ correlations ρ: $z(\rho) = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$. To guarantee a positive define matrix $\boldsymbol{D}$, the foregoing normal distributions on the z-transformed correlations needs to be truncated over the relevant values of the correlations (Daniels and Kass 1999). For a single ρ, assumed here and representing compound symmetry, a half-normal distribution for $z(\rho)$ is assumed. When the correlations are allowed to differ, the constraints to satisfy positive definiteness are more complicated. Further, the authors assigned a prior on the unknown variance σ_ρ^2 and flat priors on the diagonal elements of $\boldsymbol{D}$. Christiansen and Morris (1997) considered a hyper-prior on $\sigma_{\rho_b}^2$, with $\pi(\sigma_{\rho_b}^2) \propto (c + \sigma_{\rho_b}^2)^{-2}$ and c is a constant that represents a variance. For instance, c can be set to be $\frac{1}{n-3}$, the variance of the *Fisher-z* transformation Hurtado Rúa, Mazumdar, and Strawderman (2015). Similar to the approach of Hurtado Rúa, Mazumdar, and Strawderman (2015), we assigned IG(0.1,0.1) prior for variance parameters and a truncated normal distribution prior for $z(\rho)$. We refer this prior as the *Fisher-z* prior.

Barnard, McCulloch, and Meng (2000) proposed the separation strategy whereby the $q \times q$ correlation matrix $\boldsymbol{R}$ has a joint uniform distribution on $[-1, 1]^q$. However, the effective algorithm to draw $\boldsymbol{R}$ uniformly is computationally demanding for $q \geq 3$ due to the positive definite constraint. We used here the approach of Tokuda et al. (2011), which is based on the results shown by Joe (2006). He proved that a q-dimensional positive definite correlation matrix $\boldsymbol{R} = (\rho_{ij})_{i,j=1,\ldots,q}$ can be written in terms of the correlations $\rho_{i,i+1}$ and the partial correlations $\rho_{ij;i+1,\ldots,j-1}$ for $(j-1) \geq 2$. These parameters can take independently values in the $[-1, 1]$. Therefore, he concluded that one can generate a random positive definite correlation matrix by choosing independent distributions $\boldsymbol{F}_{ij}, 1 \leq j \leq q$ for these parameters (correlations and partial correlations). An appropriate choice for $\boldsymbol{F}_{ij}$ leads to a joint density for $\rho_{ij:1\leq i<j\leq q}$ that is proportional to $\det(\boldsymbol{R})^{\eta-1}$, where $\eta > 0$. When $\eta = 1$, Lewandowski, Kurowicka, and Joe (2009) proved that the marginal distribution of each correlation is a symmetric translated Beta(q/2,q/2)-distribution on the interval $[-1, 1]$. Consequently, the marginal distribution of each correlation becomes more concentrated around zero as q increases in order to satisfy the positive definite constraint. Note also that Joe proved that his algorithm is able to sample from a joint uniform distribution on $[-1, 1]^q$. Tokuda et al. (2011) visualized the implied distribution of $\boldsymbol{D}$ and they observed that for this prior the correlations are a priori independent of the standard deviations. There are several options for the prior

distribution on the diagonal elements of $\boldsymbol{D}$ (O'Malley and Zaslavsky 2008). Here, we assigned Gelman's folded half-t (Gelman et al. 2008) priors for elements of D. In our simulations we considered $q = 1, 2, 3$, then in each case we sampled (each) partial correlation from a translated Beta$(q/2, q/2)$ on $[-1, 1]$. We refer to this prior as the *partial* prior.

5.2.3. A joint prior for error variance and random effects variance-covariance matrix

Often, priors for error variance and variance-covariance matrix of the random effect are independently modeled. However, it has been shown by Demirhan and Kalaylioglu (2015) and by Kalaylioglu and Demirhan (2017) that a joint prior for these variance terms is more appropriate. Hence, we compare the performance of the conditional and marginal version of the criteria when variance terms are given a joint prior. Kalaylioglu and Demirhan (2017) utilized Cholesky decomposition to separate the random effects variance-covariance $\mathbf{D} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a $q \times q$ lower-triangular matrix. Further, the authors vectorized the diagonal and non-zero off-diagonal matrix $\mathbf{L}$ and the resulting column vectors are denoted by L_1 and L_2, respectively. Additionally, they considered a joint prior distribution for $(L_1^T, L_2^T, \sigma_\epsilon^2, \sigma_b^2)^T$ if the response variable is continuous and $(L_1^T, L_2^T, \sigma_b^2)^T$ for a dichotomous/polychotomous response. Furthermore, a multivariate distribution prior was assigned to the vector of log-transformed error variances, log-transformed L_1 and untransformed L_2. For theoretical details of this approach, the reader should consult Demirhan and Kalaylioglu (2015) and Kalaylioglu and Demirhan (2017). To ensure positive-definite of $\mathbf{D}$, priors on L_1 need to be positive while L_2 is left unconstrained. The authors' multivariate priors on $\mathbf{D}$ and σ_ϵ^2 are as follows:

$$(\log(L_1), L_2, \log(\sigma_\epsilon^2) \sim F(\delta, \nu, \lambda, \xi))^T,$$

and represent the generalized multivariate log gamma (G-MVLG) (Demirhan and Hamurkaroglu 2011). We refer to this prior as the G-MVLG prior.

6. Simulation study

We carried out simulation studies anchored on two longitudinal datasets. Two simulation studies were considered. In the first study, the guiding dataset is based on the well-known balanced dental growth study of Potthoff and Roy (1964). Measurements were taken on the jaw bi-annually from children between 8 and 14 years of age. The second study is based on the Jimma Infant Survival study, which was designed to evaluate the risk factors affecting infant survival in the Jimma town located in Ethiopia (Lesaffre, Asefa, and Verbeke 1999). This dataset is unbalanced due to missing responses, babies that dropped out of the study or died during the study.

In these simulation studies, we used two selection strategies based on (1) *minimum value* and (2) *absolute difference*. For the minimum value strategy, we selected the model having the lowest selection criterion. For the absolute difference strategy, the simplest model was selected when the absolute difference between these models is less than five. This has been suggested in the literature for AIC and BIC, but also for DIC (Lesaffre and Lawson 2012). We used the same threshold for WAIC and PSBF, however, our previous work (Ariyo et al. 2019) did not show justification for *absolute difference* outside DIC. Therefore, we report for WAIC and PSBF only the results using *minimum*

value. For both simulation substudies, convergence was evaluated using the Brooks–Gelman–Rubin (BGR) statistic Brooks and Gelman (1998); Gelman and Rubin (1992). All model parameters in the simulation study were estimated based on three chains of 15,000 iterations after discarding the first 7000 iterations as burn-in. The thinning factor was set at 10. When BGR was larger than 1.1, further sampling was performed until BGR $<$ 1.1. The JAGS code used in this study is provided in the Supporting Materials. Further details on the simulation settings are given below.

The aims of the simulation studies are ultimately to provide practical guidelines. More specifically, we are interested in:

- The impact of the particular choice of the vague prior on the conditional and marginal version of the selection criteria. This is the main aim of this article;
- Which of the criteria to choose in practice, taking also into account that DIC has some undesirable properties, such as non-invariance to parameter transformations and that sometimes $p_{\text{DIC}} < 0$ so that we cannot use DIC in that case;
- The difference in performance of the conditional and marginal version of the criteria. Previously, it has been shown that model selection should be done on the marginal criteria, but it is not immediately clear whether the priors affect the two versions of the criteria equally;
- If the conditional criteria are to be used, whether certain vague priors can still induce good performance of the conditional criteria;
- The impact of the sample size on the above conclusions.

6.1. The balanced case: the Potthoff and Roy dataset

In the Potthoff and Roy study, changes in pituitary-pterygomaxillary distances during growth of a child were examined at years 8, 10, 12, and 14 on 11 girls and 16 boys who underwent orthodontic treatment. The following LMM was fitted to the data as a function of age and sex (0 = girls, 1 = boys):

$$Y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_{ij} + b_{0i} + \epsilon_{ij}, \quad (i = 1, ..., 27; j = 1, ..., 4), \tag{8}$$

where Y_{ij} is the distance (mm) measure of the ith child at time j, b_{0i} is a random intercept with $b_{0i} \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The following restricted maximum likelihood estimates: $\widehat{\beta}_0 = 24.97, \widehat{\beta}_1 = 1.48, \widehat{\beta}_2 = -2.32, \widehat{\sigma}_b^2 = 2.05,$ and $\widehat{\sigma}_\epsilon^2 = 3.27$ were obtained and used as true parameters in the simulation study. We then considered two scenarios.

- **Scenario I**: We assumed that the random effects structure is known and considered models that differ in the fixed effects part. Besides the true data-generating model (8), we considered an overspecified model, which includes an interaction term age*sex and an underspecified model, which omits sex from the model.
- **Scenario II**: We assumed that the fixed structure is known and considered models that differ in the random effects. The overspecified model includes an additional random slope whereas the underspecified alternative ignores the random intercept in the data.

Table 1. Potthoff and Roy dataset (Scenario I): Sensitivity of the performance of the conditional and the marginal selection criteria to choose the correct LMM by varying the prior distribution on variance terms for different sample sizes.

		Sample sizes			
Prior	Criteria	5	10	25	100
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.0001, 0.0001)$	cDIC	26.8	27.2	64.0	79.4
	cPSBF	35.0	38.8	39.2	50.6
	cWAIC	21.4	24.8	62.6	72.4
	mDIC	55.2	60.8	77.4	83.2
	mPSBF	41.4	53.6	75.4	84.2
	mWAIC	46.8	59.4	75.2	83.2
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.1, 0.1)$	cDIC	55.4	56.2	64.6	79.8
	cPSBF	55.0	44.8	43.2	43.0
	cWAIC	53.8	55.6	56.8	65.4
	mDIC	56.0	63.6	78.0	83.2
	mPSBF	44.0	56.2	76.6	84.0
	mWAIC	46.8	61.2	75.6	83.2
$\log(\sigma_b^2) \sim \text{Uniform}(-10, 10)$	cDIC	25.0	27.4	67.2	70.0
	cPSBF	36.6	38.4	45.8	46.8
	cWAIC	33.2	38.4	62.0	68.8
	mDIC	58.8	61.6	78.0	82.8
	mPSBF	40.0	54.6	76.4	83.8
	mWAIC	48.8	57.6	75.8	82.6
$\log(\sigma_b^2) \sim \text{Uniform}(0.001, 100)$	cDIC	46.2	53.0	68.0	78.8
	cPSBF	43.8	45.0	46.8	48.0
	cWAIC	53.4	48.0	66.6	70.2
	mDIC	57.8	60.8	77.8	83.6
	mPSBF	40.0	58.0	75.6	84.6
	mWAIC	49.0	58.2	75.6	83.8
$\frac{1}{\sigma_b^2} \sim \text{Pareto}(1, 0.0001)$	cDIC	32.8	52.4	69.2	78.8
	cPSBF	33.4	44.4	43.6	42.4
	cWAIC	42.2	50.0	64.4	73.4
	mDIC	51.2	54.0	75.6	83.0
	mPSBF	40.8	51.2	76.0	83.0
	mWAIC	46.4	51.4	74.0	82.6
$\sigma_b \sim \text{Uniform}(0, 100)$	cDIC	31.6	46.6	69.4	79.4
	cPSBF	43.0	42.6	43.4	42.8
	cWAIC	41.4	45.6	61.6	71.8
	mDIC	55.0	56.4	77.6	83.2
	mPSBF	28.2	53.8	76.8	84.8
	mWAIC	33.8	55.6	75.8	83.8
$\sigma_b \sim \text{t}(0, 0.75, 1)$	cDIC	41.2	63.0	64.0	71.0
	cPSBF	31.8	43.7	52.0	64.8
	cWAIC	40.0	67.4	69.0	70.2
	mDIC	68.5	76.8	79.6	84.2
	mPSBF	61.4	74.8	75.2	80.6
	mWAIC	66.2	74.0	76.2	81.0

6.1.1. Data generation and prior specifications

Twenty simulation settings were considered for each of the four different sample sizes $n =$ (5, 10, 25, 100) and five signal-to-noise ratios ($\frac{1}{4}, \frac{1}{2}, 1$, 2, and 4 times the residual variance). Each time 500 datasets were generated from model (8). For each of the simulation settings, the regression coefficients were given a vague normal prior. Namely, $\beta_j \sim \text{N}(0, 10^6) \quad (j = 0, 1, 2)$. Further, seven prior distributions for variance terms were assigned. Each time, three models (correct, over, and under specified models) were fitted to evaluate the performance of both the marginal and conditional versions of the Bayesian model selection. We have taken the following vague priors for σ_b:

Table 2. Potthoff and Roy dataset (Scenario II): Sensitivity of prior distribution on variance terms on selecting the correct model (%) for different sample sizes and criteria DIC, PSBF, WAIC evaluated on conditional and marginal version of LMM.

		Sample sizes			
Prior	Criteria	5	10	25	100
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.0001, 0.0001)$	cDIC	27.8	42.6	50.8	44.2
	cPSBF	32.8	39.2	46.8	57.8
	cWAIC	22.6	37.6	46.6	43.0
	mDIC	41.0	54.0	82.4	82.6
	mPSBF	37.2	51.2	80.2	80.2
	mWAIC	42.6	57.4	76.8	81.2
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.1, 0.1)$	cDIC	54.6	50.6	52.2	49.2
	cPSBF	54.2	59.4	56.6	51.2
	cWAIC	46.4	40.0	46.0	28.4
	mDIC	52.0	72.8	81.4	84.6
	mPSBF	64.6	79.0	80.8	84.4
	mWAIC	59.8	76.6	77.8	82.6
$\log(\sigma_b^2) \sim \text{Uniform}(-10, 10)$	cDIC	53.2	55.2	46.6	41.0
	cPSBF	45.4	46.4	44.6	46.2
	cWAIC	48.8	45.0	51.4	48.8
	mDIC	40.8	66.6	82.4	82.8
	mPSBF	51.0	63.4	80.2	81.6
	mWAIC	52.6	68.0	79.0	80.6
$\log(\sigma_b^2) \sim \text{Uniform}(0.001, 100)$	cDIC	67.0	77.2	83.8	97.4
	cPSBF	66.8	80.6	77.0	94.4
	cWAIC	54.8	53.6	39.8	37.4
	mDIC	65.6	89.0	100.0	100.0
	mPSBF	75.0	92.6	100.0	100.0
	mWAIC	67.4	90.8	100.0	100.0
$\frac{1}{\sigma_b^2} \sim \text{Pareto}(1, 0.0001)$	cDIC	59.2	68.2	56.8	40.6
	cPSBF	61.4	66.4	52.6	47.0
	cWAIC	53.4	41.6	29.0	19.0
	mDIC	45.0	76.6	86.0	85.8
	mPSBF	40.0	87.4	86.0	85.2
	mWAIC	48.8	84.4	84.4	83.8
$\sigma_b \sim \text{Uniform}(0, 100)$	cDIC	61.4	62.2	53.8	42.2
	cPSBF	59.2	53.0	46.8	44.0
	cWAIC	51.4	42.0	36.6	35.0
	mDIC	48.8	76.4	84.8	85.6
	mPSBF	61.0	82.4	85.2	84.0
	mWAIC	50.4	80.4	81.6	83.4
$\sigma_b \sim \text{t}(0, 0.75, 1)$	cDIC	40.2	56.0	55.8	61.4
	cPSBF	41.2	43.7	47.4	50.0
	cWAIC	41.2	43.8	57.4	60.4
	mDIC	71.2	85.8	87.4	92.6
	mPSBF	70.6	84.4	86.4	86.9
	mWAIC	70.4	81.4	86.2	91.8

1. $\frac{1}{\sigma_b^2} \sim \text{Gamma}(a, a),\ (a = 0.001 \text{ and } a = 0.1)$;
2. $\log(\sigma_b^2) \sim \text{Uniform}(a, b),\ (a, b) = (-10, 10)$;
3. $\frac{1}{\sigma_b^2} \sim \text{Pareto}(a, b),\ (a, b) = (1, 0.001)$;
4. $\sigma_b \sim \text{Uniform}(a, b),\ (a, b) = (0, 100)$;
5. $\sigma_b \sim \text{half}-t(0, s, 1), s = (1, 0.75)$.

The motivation of the choices of a, b, and s is given in Sec. 5.1.

We focused on the independent prior distributions in which the variance terms of the random intercept and measurement errors are modeled independently to evaluate

the performance of both versions of the criteria. As these priors are commonly used in the literature. The impact of joint prior will be examined in the subsequent section.

6.1.2. Simulation results

Table 1 shows the percentage of correct selection for different sample sizes under Scenario I. The results show that the impact of the vague priors on the marginal criteria is minimal, but their impact on the conditional criteria is considerable. This conclusion holds irrespective of the sample size. However, the performance for the conditional criteria improves with increasing sample size. In addition, among the three conditional criteria, DIC is best for higher sample sizes (25 and 100), competing even with the marginal criteria for sample size 100. This in an inconsistent manner. For smaller sample sizes, the half-t prior performed best for the marginal version of all criteria. However, it is not clear which prior outperforms across the different settings and sample sizes.

In Table 2, the percentages of correct selection for different sample sizes under Scenario II are shown. We observed that a uniform prior for log(variance) performs well for both versions of DIC and PSBF, but the conditional WAIC performs poorly. The poor performance of the conditional WAIC is also seen with the other priors. Again, regardless of the scenario, the marginal criteria outperform the conditional criteria. Their performance increases with sample size while the conditional criteria often select over-specified models (not shown here). However, there is no clear winner among the marginal criteria in this scenario.

6.2. The unbalanced case: the Jimma Infant Growth study

The second dataset is obtained from an Ethiopian study designed to evaluate risk factors affecting infant survival.(Lesaffre, Asefa, and Verbeke 1999) The growth characteristics of the babies were examined approximately every 60 days, but there were occasional deviations from the planned visits. For the purpose of this analysis, we have taken weight as response with covariates age and sex (0 = girls, 1 = boys) of the child, and age of the mother at delivery (*agem*). The details of the original analysis can be found in Lesaffre, Asefa, and Verbeke (1999) and Lesaffre et al. (2000) where a sample of 495 children was selected to fit the model. This subset will also be the basis for this simulation study. As suggested, Lesaffre et al. (2000) the time variable age was transformed into $\text{newage}_{ij} = \sqrt{\text{age}_{ij}} - (\text{age}_{ij} + 1) - 0.02 \times \text{age}_{ij}$ in model to fit a LMM to the weight profiles. We select our model generating data to be

$$Y_{ij} = \beta_1 + \beta_2 \text{sex}_i + \beta_3 \text{newage}_{ij} + \beta_4 \text{agem}_i + b_{0i} + b_{1i} \times \text{newage}_{ij} + \epsilon_{ij}, \tag{9}$$

assuming $(b_{0i}, b_{1i}) \sim N_2(\boldsymbol{0}, \boldsymbol{D})$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The following parameter values were obtained by analysis Jimma data: $\widehat{\beta_1} = 2.8581, \widehat{\beta_2} = 0.1518, \widehat{\beta_3} = 0.8865, \widehat{\sigma_\epsilon} = 0.3465$, and $\mathbf{D} = \begin{pmatrix} 0.6813 & -0.0414 \\ -0.0414 & 0.0450 \end{pmatrix}$ where these parameters are used as population parameters for the simulated dataset. In total, 500 datasets were generated from model (9) with the covariate sex was generated from a Bernoulli distribution with probability of

Table 3. Jimma Infant Survival dataset (Scenario I ($q = 2$) and Scenario II ($q = 3$, 2, and 1): Performance of Bayesian model selection with six specifications of inverse-Wishart conjugate prior for over-specified, correct, and under-specified, respectively.

			Criteria					
Sample size	Scenario		cDIC	cPSBF	cWAIC	mDIC	mPSBF	mWAIC
10	I	$df = q, \mathbf{V} = 0.001$	49.0	47.8	41.8	59.0	59.0	56.8
		$df = q, \mathbf{V} = 1$	62.8	44.6	58.0	82.8	81.8	89.8
		$df = q + 1, \mathbf{V} = 0.001$	46.6	37.4	37.6	58.4	59.4	54.8
		$df = q + 1, \mathbf{V} = 1$	60.4	47.8	55.0	80.0	77.4	76.8
		$df = q + 2, \mathbf{V} = 0.001$	51.8	42.2	41.6	58.2	63.2	55.8
		$df = q + 2, \mathbf{V} = 1$	57.6	43.4	57.5	79.0	76.0	53.3
	II	$df = q, \mathbf{V} = 0.001$	61.0	52.6	63.2	63.2	69.4	68.2
		$df = q, \mathbf{V} = 1$	62.4	65.6	45.8	86.0	88.6	97.0
		$df = q + 1, \mathbf{V} = 0.001$	58.0	44.4	58.2	57.4	52.8	61.0
		$df = q + 1, \mathbf{V} = 1$	59.2	67.8	45.6	86.6	88.4	87.2
		$df = q + 2, \mathbf{V} = 0.001$	60.4	53.0	57.8	58.0	55.3	62.4
		$df = q + 2, \mathbf{V} = 1$	60.2	62.4	46.8	87.2	88.8	86.2
50	I	$df = q, \mathbf{V} = 0.001$	54.0	50.0	52.0	68.0	70.0	68.0
		$df = q, \mathbf{V} = 1$	68.8	72.4	67.8	90.0	90.0	90.0
		$df = q + 1, \mathbf{V} = 0.001$	64.4	72.4	67.8	76.4	77.6	76.0
		$df = q + 1, \mathbf{V} = 1$	69.4	70.6	68.4	90.0	90.0	90.0
		$df = q + 2, \mathbf{V} = 0.001$	77.0	71.4	77.8	79.0	86.6	79.0
		$df = q + 2, \mathbf{V} = 1$	68.8	73.0	77.8	90.0	90.0	90.0
	II	$df = q, \mathbf{V} = 0.001$	52.0	49.0	53.0	66.0	68.0	62.0
		$df = q, \mathbf{V} = 1$	53.6	47.8	56.0	80.4	76.4	79.6
		$df = q + 1, \mathbf{V} = 0.001$	60.0	44.2	55.4	66.4	71.0	65.4
		$df = q + 1, \mathbf{V} = 1$	64.2	42.4	57.4	82.0	75.8	79.0
		$df = q + 2, \mathbf{V} = 0.001$	57.8	49.0	78.2	65.2	69.6	64.6
		$df = q + 2, \mathbf{V} = 1$	62.0	40.8	51.6	79.6	74.6	75.8

success equal to 0.51, which is the proportion of boys in the dataset. The age of the mother was generated from a normal distribution $\text{agem}_i \sim N(24.49, 6.29)$ and we have taken $0, 60, 120, \ldots, 360$ days as the moments of measurements. The alternative models considered for each scenario are described below.

- **Scenario I**: We assumed that the random effects structure is known and considered the following models that differ in the fixed part parameters, namely
  - Model (9) and including an additional interaction (newage × sex) (overspecified),
  - Model (9) but ignoring the sex covariate (underspecified).
- **Scenario II**: We assumed that the covariates in the fixed part are known and considered the following models that differ in the random effects structure, i.e.,
  - Model (9) and including an additional random slope for newage^2 (overspecified),
  - Model (9) but ignoring the random slope for newage (underspecified).

6.2.1. Data generation and prior specifications

With the above specifications for generating data, we considered 24 simulation settings for five sample sizes. These settings correspond to twelve different prior choices for the covariance matrix for the two scenarios described above. For each of the settings and sample sizes, we generated 500 datasets from model (9). Model (9) is then fit using each of the following prior specifications:

Table 4. Jimma Infant Survival dataset: Sensitivity of the performance of the conditional and marginal selection criteria to choose the correct LMM by using separation priors, a joint prior with an IW prior for different sample sizes.

Scenario		I					II				
Prior		10	25	50	100	200	10	25	50	100	200
IW	cDIC	31	44	48	66	63	44	52	58	53	55
	cPSBF	36	40	46	68	69	42	39	37	42	46
	cWAIC	35	42	41	65	69	42	53	58	51	53
	mDIC	52	52	66	72	75	54	63	67	70	72
	mPSBF	52	50	67	72	75	54	54	68	73	73
	mWAIC	53	54	66	73	75	56	62	67	71	72
Chol	cDIC	60	61	61	56	52	70	72	79	78	78
	cPSBF	55	58	59	61	59	69	67	67	70	70
	cWAIC	59	61	60	52	50	66	68	72	76	78
	mDIC	89	89	88	92	96	100	100	100	100	100
	mPSBF	70	71	76	84	87	98	100	100	100	100
	mWAIC	79	79	80	83	87	98	100	100	100	100
HIW	cDIC	60	62	63	64	67	82	84	81	88	88
	cPSBF	60	60	66	66	67	72	76	71	78	77
	cWAIC	61	62	62	63	66	82	83	82	85	88
	mDIC	76	80	82	82	100	98	100	100	100	100
	mPSBF	74	81	85	84	100	96	100	100	100	100
	mWAIC	75	81	85	86	100	99	100	100	100	100
Partial	cDIC	66	61	64	62	61	54	72	62	68	66
	cPSBF	59	61	62	61	64	59	66	63	60	60
	cWAIC	64	64	64	66	67	51	69	63	67	69
	mDIC	76	84	82	86	84	73	84	86	87	80
	mPSBF	75	85	83	86	85	71	80	84	87	81
	mWAIC	83	83	79	86	85	69	80	85	86	80
Fisher-*z*	cDIC	74	75	80	79	82	71	69	71	80	87
	cPSBF	70	70	73	77	83	61	70	74	77	80
	cWAIC	67	70	74	76	78	67	66	68	77	83
	mDIC	74	79	83	100	100	70	82	98	100	100
	mPSBF	66	71	79	100	100	73	84	94	100	100
	mWAIC	75	74	81	100	100	67	80	96	100	100
G-MVLG	cDIC	62	69	64	66	67	71	86	79	89	89
	cPSBF	64	65	62	67	68	70	79	73	86	97
	cWAIC	61	68	63	68	67	69	85	69	86	87
	mDIC	69	73	80	88	100	100	100	100	100	100
	mPSBF	70	70	81	89	99	100	100	99	100	100
	mWAIC	71	72	83	89	100	99	100	100	100	100

- **Prior (1)**: Six specifications of the IW conditional conjugate prior described in Sec. 5.2.1 of the form $\text{IW}(df, \mathbf{V})$ using $df = q, q+1, q+2,$ and $\mathbf{V} = c\boldsymbol{I}_q$ for $c \in \{0.001, 1\}$, where $\boldsymbol{I}_q$ denotes $q \times q$ identity matrix. This is a relatively commonly used informative IW (Schnell et al. 2016).
- **Prior (2)**: Five separation strategies (Sec. 5.2.2) for covariance matrix **D**.
- **Prior (3)**: **(G-MVLG)** For joint variance prior $(\log(L_1), L_2, \log(\sigma_\epsilon^2)) \sim \text{GMVLG}(0.7, 1.42, \lambda, \xi)$ with $\lambda = (0.3, 0.3, 0.3, 0.4)^T$ and $\xi = (0.25, 0.35, 0.25, 0.1)^T$. This is a non-informative prior and the hyper-parameter values were selected to impose uncertainty on the variance parameters (Kalaylioglu and Demirhan 2017).

6.2.2. Simulation results

Table 3 shows the performance of different specifications of the IW prior to Jimma Infant Survival dataset. Regardless the scenario and sample size, the performance of the

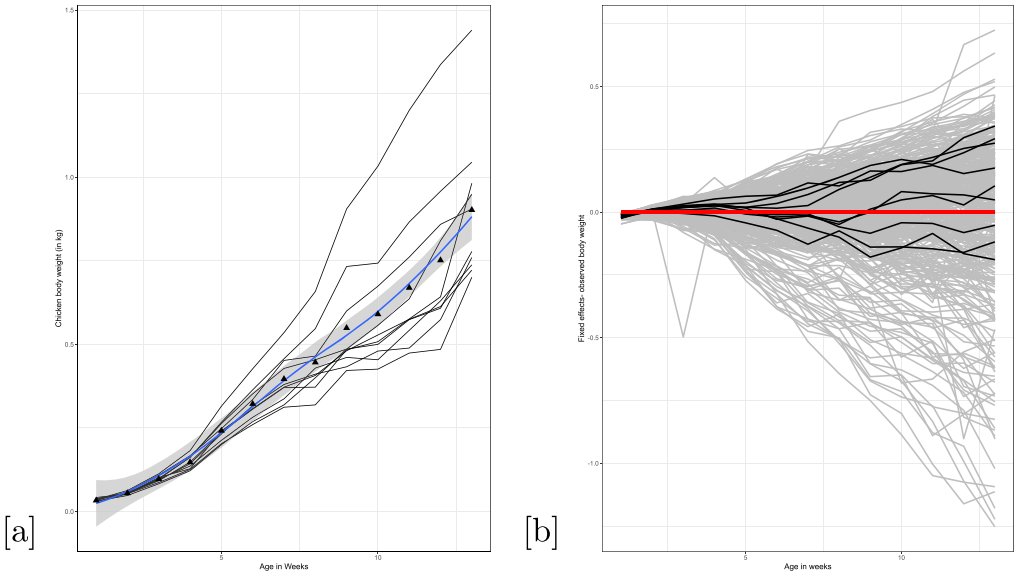

Figure 1. Nigerian indigenous chicken dataset: (a) individual and average profiles of 10 randomly selected chickens' body weight obtained by locally weighted regression using ggplot2 and (b) the deviation of the 10 randomly selected chickens' body weight from the mean structure.

conditional and marginal criteria varies with changing df and $\mathbf{V}$. We observed that both criteria perform better with a larger value of $\mathbf{c} = 1$ regardless of the value of df. The IW prior $df = q$ and $\mathbf{c} = 1$ performed relatively better in both scenarios.

The performance of the IW prior deteriorates with increasing dimension of the random effects covariance matrix, see also Ariyo et al. (2019). Additionally, to choose an appropriate scale matrix and degrees of freedom is not straightforward, since inconsistent performance is seen. Other disadvantages of the IW prior distribution have been discussed and alternatives proposed (Pourahmadi 1999; Barnard, McCulloch, and Meng 2000; Daniels and Pourahmadi 2002; Lu and Ades 2009; Wei and Higgins 2013; Schuurman, Grasman, and Hamaker 2016). Therefore, we considered the effect of some separation priors for the conditional and marginal versions of PSBF, DIC, and WAIC.

Table 4 shows the performance of both versions of selection criteria using a commonly used IW prior compared with the separation priors and a joint prior for different sample sizes. Since the Cholesky and spherical decomposition performed similarly, the results of the spherical decomposition are omitted here. For both scenarios, and especially for small sample sizes, the joint prior and separation priors outperformed the classical IW prior. In addition, for both versions of the criteria the impact of the sample size is less pronounced with a joint prior and the separation priors than for the IW prior. This result agrees with the conclusion in Alvarez, Niemi, and Simpson (2014), i.e., that the classical IW prior is less effective when compared with a separation prior.

Further, in Scenario I, the approach based on the *Fisher-z* transformation performed best for both the conditional and marginal versions of the criteria. For scenario II, there is no significant difference between the HIW prior based on the approach proposed by Huang and Wand (2013) and G-MVLG prior. For both scenarios, the G-MVLG prior and separation priors gave better performance for the conditional criteria when

Table 5. Nigeria indigenous chicken dataset: Sensitivity of the performance of the conditional and marginal selection criteria using separation priors and joint prior with an IW prior.

	Criteria	IW	HIW	Fisher-z	Chol	G-MVLG
Model 1	cDIC	−15,129.8	−14,791.2	−15,191.6	−15,189.5	−14,901.8
	cWAIC	−15,497.7	−15,334.3	−15,478.1	−15,470.4	−15,014.2
	clpml	−13,938.7	−12,995.1	−14,058.4	−14,073.7	−14,913.1
	mDIC	−13,648.5	−12,768.9	−13,695.0	−12,376.1	−13,012.0
	mWAIC	−13,639.5	−12,759.1	−13,685.7	−12,358.9	−13,085.8
	mlpml	−13,609.6	−12,727.2	−13,653.9	−12,333.5	−13,043.4
Model 2	cDIC	−16,726.7	−16,855.0	−16,755.0	−16,755.0	−16,045.7
	cWAIC	−17,110.1	−17,089.3	−17,049.3	−17,049.3	−16,042.3
	clpml	−15,401.3	−15,576.0	−15,516.0	−15,516.0	−16,519.1
	mDIC	−15,093.9	−15,131.8	−15,121.8	−15,121.8	−15,501.8
	mWAIC	−15,079.7	−15,117.1	−15,107.1	−15,107.1	−15,410.9
	mlpml	−15,036.7	−15,090.6	−15,060.6	−15,060.6	−15,462.2
Model 3	cDIC	−16,095.7	−19,095.7	−18,709.6	−18,702.6	−18,612.3
	cWAIC	−16,776.4	−19,776.4	−19,485.6	−19,482.6	−19,810.8
	clpml	−17,114.1	−17,914.1	−16,117.7	−16,111.7	−16,132.0
	mDIC	−16,509.1	−16,579.1	−15,365.3	−15,365.3	−15,369.0
	mWAIC	−16,505.7	−16,565.7	−15,351.6	−15,351.6	−15,350.9
	mlplm	−16,504.3	−16,524.3	−15,310.2	−15,310.2	−15,320.7
Model 4	cDIC	−16,186.8	−19,476.8	−18,796.3	−19,492.1	−20,047.4
	cWAIC	−16,851.8	−20,034.2	−19,533.5	−20,026.3	−20,126.8
	clplm	−16,777.5	−17,609.5	−16,193.5	−17,593.9	−17,784.0
	mDIC	−16,104.2	−16,788.4	−16,632.9	−16,878.2	−16,897.4
	mWAIC	−16,314.9	−16,770.5	−16,618.3	−16,663.6	−16,709.3
	mlpml	−16,466.7	−16,719.0	−16,572.4	−16,601.3	−16,700.4
Model 5	cDIC	−16,176.8	−16,676.8	−16,476.8	−16,476.8	−16,421.0
	cWAIC	−16,034.2	−17,434.2	−17,034.2	−17,034.2	−17,114.6
	clppd	−16,609.5	−17,609.5	−17,609.5	−17,609.5	−17,709.1
	mDIC	−16,108.4	−16,208.4	−16,498.4	−16,738.4	−16,715.0
	mWAIC	−16,400.5	−16,200.5	−16,470.5	−16,620.5	−16,631.3
	mlpml	−16,509.0	−16,309.0	−16,469.0	−16,710.0	−16,731.8

compared with IW prior. Additionally, the impact of sample sizes is less in both G-MVLG and separation prior compared with IW prior. While there is no best vague prior in both scenarios, we conclude that if the conditional version of the criteria is to be used, then the G-MVLG prior, hierarchical or separation priors are to be used. In fact, the use of IW prior to conditional criteria is strongly discouraged especially for smaller sample sizes. But, again the marginal version of the criteria outperformed the conditional criteria in all scenarios and sample sizes.

7. Analysis of the longitudinal evolution of Nigerian chickens

We analyzed of the Nigerian indigenous chicken (NIC) dataset and evaluated the sensitivity of separation priors and classical IW prior on the covariance matrix. These data concern the longitudinal evolution of body weight (BW) of chickens of different breeds raised in a university experimental farm. Four hundred and sixteen chickens were measured every week (age) from hatching up to twenty weeks to evaluate the growth of two progenies (breeds) of chicken. A first analysis can be found in Ariyo et al. (2019). We refer to Adeleke et al. (2011) for the rationale for the study and the experimental design. Figure 1a shows the evaluation of weight of the chicken and the average profile over time. The deviations between the observed chickens' body weight and the mean structure are presented in Figure 1b. It will be assumed that

$$Y_{ij} = \beta_0 + \beta_1 \text{breed}_i + \beta_2 \text{age}_{ij} + b_{0i} + b_{2i}\text{age}_{ij} + \epsilon_{ij}, \tag{10}$$

where Y_{ij} is the chicken body weight (kg); breed_i is the breed indicator (1 = pure breed, 2 = cross breed), age_{ij} represents the age (standardized). We limit the chicken's age to 13 weeks since a considerable amount of chicken died after this age.We fitted the following alternative models:

- Model 1: Linear model in fixed effects and linear in random effects
- Model 2: Quadratic model in fixed effects and linear in random effects
- Model 3: Linear model in fixed effects and quadratic in random effects
- Model 4: Quadratic in fixed effects and quadratic in random effects
- Model 5: Cubic in fixed effects and cubic in random effects.

The classical IW prior together with two separation (*Fisher-z* and *Chol*) priors and a Hierarchical prior (*HIW*) discussed in Sec. 5.2.2 were used for the covariance matrix.

Table 5 shows that there is some discrepancy in both the marginal and conditional criteria using different priors. The conditional DIC and WAIC select Model 2 using the IW prior. Contrary, the conditional PSBF as well as the marginal version of the criteria selection Model 3. This shows inconsistency in model selection among the conditional criteria when IW prior is used. However, both the models selected by these criteria seem to be incorrect as the average growth curve of the chicken seems quadratic and the individual growth curves differ from the average curve in a quadratic manner (see Figure 1). In contrast, all the separation priors as well as joint prior support Model 4 (i.e., the presence of quadratic terms in both fixed and random effects) which appears to be the appropriate model here based on Figure 1. This confirmed the results of the simulation that separation priors are more efficient than the IW prior.

8. Conclusion

We have performed simulation studies to determine if the choice of the vague prior for the variance or covariance matrix of the random effects in a longitudinal study is of great importance in model selection. In addition, we assessed whether different vague prior distributions have a different effect on the conditional and marginal version of DIC, PSBF, and WAIC. We made use of vague priors that were proposed in the literature. While the considered scenarios are still somewhat limited in scope, the performance of the criteria in our simulation study allows already for some clear conclusions.

The results can be broadly summarized as follows for the variance of the random intercept. The choice of the vague prior impacted both versions of the criteria but the impact is much less for the marginal version than for the conditional version of the criteria. In addition, the conditional criteria performed in an inconsistent manner often selecting over-specified models while the marginal version of the criteria showed much less dependence to the choice of parameter values of the prior and often selected the correct model. For longitudinal mixed models that involve two or more random effects, the joint prior, the hierarchical prior and the separation priors all outperformed the classical IW prior. These priors are also the choice when the conditional version of the

criteria is to be taken. We noted, to our surprise, that cWAIC was significantly poorer in some cases than the two other criteria.

Finally, we believe that a sensitivity analysis is necessary when using prior distributions that are intended to be vague for the level 2 variance parameters. This is especially important for small sample sizes. For models with more than one random effect, the joint prior, the hierarchical prior and separation priors are to be chosen for both the conditional and marginal versions of the criteria. For large sample sizes, the classical IW prior can still be used for model selection for computational convenience. Finally, the marginal version of the criteria outperformed the conditional version of the criteria, as was earlier recommended in the literature, see (Chan and Grant 2016; Li et al. 2016; Merkle, Furr, and Rabe-Hesketh 2018; Millar 2018; Quintero and Lesaffre 2018; Ariyo et al. 2019). We have added evidence to this recommendation in the context of longitudinal mixed models, which constitutes an important class of models in biomedical research.

Acknowledgments

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government Department EWI. The authors appreciate the Poultry Breeding Unit, Department of Animal Breeding and Genetics, Federal University of Agriculture, Abeokuta, Nigeria for the NIC dataset.

Disclosure statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The research of the first author was partially funded by Tertiary Education Trust Fund (TETFund)-AS&D grant of the Federal government of Nigeria through the Federal University of Agriculture, Abeokuta Nigeria.

ORCID

Oludare Ariyo http://orcid.org/0000-0003-3375-1831
Emmanuel Lesaffre http://orcid.org/0000-0001-7268-2221
Geert Verbeke http://orcid.org/0000-0001-8430-7576

References

Adeleke, M., S. Peters, M. Ozoje, C. Ikeobi, A. Bamgbose, and O. A. Adebambo. 2011. Genetic parameter estimates for body weight and linear body measurements in pure and crossbred progenies of Nigerian indigenous chickens. *Livestock Research for Rural Development* 23 (1): 1–7.

Alvarez, I., J. Niemi, and M. Simpson. 2014. Bayesian inference for a covariance matrix. arXiv Preprint arXiv:1408.4050.

Ariyo, O., A. Quintero, J. Muñoz, G. Verbeke, and E. Lesaffre. 2019. Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics* 1–24. doi:10.1080/02664763.2019.1657814.

Barnard, J., R. McCulloch, and X. Meng. 2000. Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage. *Statistica Sinica* 10:1281–312.

Bernardo, J. M. 1980. A Bayesian analysis of classical hypothesis testing. *Trabajos de Estadistica Y de Investigacion Operativa* 31 (1):605–47. doi:10.1007/BF02888370.

Brooks, S. P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7 (4):434–55. doi:10.2307/1390675.

Burton, P. R., K. J. Tiller, L. C. Gurrin, W. O. Cookson, A. W. Musk, and L. J. Palmer. 1999. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genetic Epidemiology* 17 (2):118–40. doi:10.1002/(SICI)1098-2272(1999)17:2<118::AID-GEPI3>3.0.CO;2-V.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (1):1–23. doi:10.18637/jss.v076.i01.

Celeux, G., F. Forbes, C. Robert, and D. Titterington. 2006. Deviance information criteria for missing data models. *Bayesian Analysis* 1 (4):651–73. doi:10.1214/06-BA122.

Chan, J., and A. Grant. 2016. On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics* 14 (4):772–802. doi:10.1093/jjfinec/nbw002.

Christiansen, C. L., and C. N. Morris. 1997. Hierarchical Poisson regression modeling. *Journal of the American Statistical Association* 92 (438):618–32. doi:10.1080/01621459.1997.10474013.

Daniels, M. J., and R. E. Kass. 1999. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* 94 (448):1254–63. doi:10.1080/01621459.1999.10473878.

Daniels, M. J., and M. Pourahmadi. 2002. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 89 (3):553–66. doi:10.1093/biomet/89.3.553.

Demirhan, H., and C. Hamurkaroglu. 2011. On a multivariate log-gamma distribution and the use of the distribution in the Bayesian analysis. *Journal of Statistical Planning and Inference* 141 (3):1141–52. doi:10.1016/j.jspi.2010.09.015.

Demirhan, H., and Z. Kalaylioglu. 2015. Joint prior distributions for variance parameters in Bayesian analysis of normal hierarchical models. *Journal of Multivariate Analysis* 135:163–74. doi:10.1016/j.jmva.2014.12.013.

Gelfand, A., and D. Dey. 1994. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)* 56 (3):501–14. doi:10.1111/j.2517-6161.1994.tb01996.x.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* 1 (3):515–34. doi:10.1214/06-BA117A.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7 (4):457–72. doi:10.1214/ss/1177011136.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2004. *Bayesian data analysis*. Boca Rotan, Florida: Chapman and Hall/CRC.

Gelman, A., D. Van Dyk, Z. Huang, and J. Boscardin. 2008. Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics* 17 (1):95–122. doi:10.1198/106186008X287337.

Geman, S., and D. Geman. 1993. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics* 20 (5–6):25–62. doi:10.1080/02664769300000058.

Huang, A., and M. P. Wand. 2013. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* 8 (2):439–52. doi:10.1214/13-BA815.

Hurtado Rúa, S. M., M. Mazumdar, and R. L. Strawderman. 2015. The choice of prior distribution for a covariance matrix in multivariate meta-analysis: A simulation study. *Statistics in Medicine* 34 (30):4083–104. doi:10.1002/sim.6631.

Joe, H. 2006. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis* 97 (10):2177–89. doi:10.1016/j.jmva.2005.05.010.

Kalaylioglu, Z., and H. Demirhan. 2017. A joint Bayesian approach for the analysis of response measured at a primary endpoint and longitudinal measurements. *Statistical Methods in Medical Research* 26 (6):2885–96. doi:10.1177/0962280215615003.

Kass, R. E., and R. Natarajan. 2006. A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Analysis* 1 (3):535–42. doi:10.1214/06-BA117B.

Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* 38 (4):963–74.

Lambert, P. C., A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 24 (15):2401–28. doi:10.1002/sim.2112.

Lesaffre, E., M. Asefa, and G. Verbeke. 1999. Assessing the goodness-of-fit of the Laird and Ware model an example: The Jimma Infant Survival Differential Longitudinal Study. *Statistics in Medicine* 18 (7):835–54. doi:10.1002/(SICI)1097-0258(19990415)18:7<835::AID-SIM75>3.0.CO;2-7.

Lesaffre, E., and A. Lawson. 2012. *Bayesian biostatistics (Statistics in practice)*. Chichester: Wiley.

Lesaffre, E., D. Todem, G. Verbeke, and M. Kenward. 2000. Flexible modelling of the covariance matrix in a linear random effects model. *Biometrical Journal* 42 (7):807–22. doi:10.1002/1521-4036(200011)42:7<807::AID-BIMJ807>3.0.CO;2-3.

Lewandowski, D., D. Kurowicka, and H. Joe. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100 (9):1989–2001. doi:10.1016/j.jmva.2009.04.008.

Li, L., S. Qiu, B. Zhang, and C. X. Feng. 2016. Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing* 26 (4):881–97. doi:10.1007/s11222-015-9577-2.

Li, Y., T. Zeng, and J. Yu. 2012. Robust deviance information criterion for latent variable models. *Research Collection School of Economics.* http://ink.library.smu.edu.sg/soe_research/1403 (accessed March 5, 2019).

Lu, G., and A. Ades. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 10 (4):792–805. doi:10.1093/biostatistics/kxp032.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10 (4):325–37. doi:10.1023/A:1008929526011.

Merkle, E., D. Furr, and S. Rabe-Hesketh. 2018. Bayesian model assessment: Use of conditional vs marginal likelihoods. *Psychometrika*. doi:10.1007/s11336-019-09679-0.

Millar, R. B. 2018. Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing* 28 (2):375–85. doi:10.1007/s11222-017-9736-8.

O'Malley, A. J., and A. M. Zaslavsky. 2008. Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association* 103 (484):1405–18. doi:10.1198/016214508000000724.

Pinheiro, J. C., and D. M. Bates. 1996. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* 6 (3):289–96. doi:10.1007/BF00140873.

Potthoff, R., and S. Roy. 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 5:313–26. doi:10.2307/2334137.

Pourahmadi, M. 1999. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* 83 (3):677–90. doi:10.1093/biomet/86.3.677.

Quintero, A., and E. Lesaffre. 2017. Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal* 59 (5):1047–66. doi:10.1002/bimj.201600193.

Quintero, A., and E. Lesaffre. 2018. Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine* 37 (16):2440–54. doi:10.1002/sim.7649.

Schervish, M. J. 2012. *Theory of statistics*. New York: Springer Science & Business Media.

Schnell, P. M., Q. Tang, W. W. Offen, and B. P. Carlin. 2016. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* 72 (4): 1026–36. doi:10.1111/biom.12522.

Schuurman, N. K., R. P. Grasman, and E. L. Hamaker. 2016. A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research* 51 (2–3):185–206. doi:10.1080/00273171.2015.1065398.

Scurrah, K. J., L. J. Palmer, and P. R. Burton. 2000. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genetic Epidemiology* 19 (2):127–48. doi:10.1002/1098-2272(200009)19:2<127::AID-GEPI2>3.0.CO;2-S.

Spiegelhalter, D. J. 2001. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine* 20 (3):435–52. doi:10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles. 2004. *Bayesian approaches to clinical trials and health-care evaluation*. West Sussex, UK: John Wiley & Sons.

Spiegelhalter, D., N. Best, N. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4):583–639. doi:10.1111/1467-9868.00353.

Spiegelhalter, D., N. Best, N. Carlin, and A. van der Linde. 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (3):485–93. doi:10.1111/rssb.12062.

Tokuda, T., B. Goodrich, I. Van Mechelen, A. Gelman, and F. Tuerlinckx. 2011. Visualizing distributions of covariance matrices. Tech. Rep., Columbia Univ., New York, USA, 18–21.

Vaida, F., and S. Blanchard. 2005. Conditional Akaike information for mixed-effects models. *Biometrika* 92 (2):351–70. doi:10.1093/biomet/92.2.351.

Wand, M. P., J. T. Ormerod, S. A. Padoan, and R. Frühwirth. 2011. Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* 6 (4):847–900. doi:10.1214/11-BA631.

Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11:3571–94.

Wei, J., and J. Higgins. 2013. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 32 (17):2911–34. doi:10.1002/sim.5745.