Bayesian model selection in linear mixed models for longitudinal data

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

KU Leuven, I-BioStat, U.Z. Sint-Rafal Kapucijnenvoer 35, blok D, bus 7001 B-3000 Leuven, Belgium
[b]Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria

ARTICLE HISTORY
Compiled February 7, 2023

ABSTRACT
Linear mixed models (LMMs) are popular to analyze repeated measurements with a Gaussian response. For longitudinal studies, the LMMs consist of a fixed part expressing the effect of covariates on the mean evolution in time and a random part expressing the variation of the individual curves around the mean curve. Selecting the appropriate fixed and random effect parts is an important modeling exercise. In a Bayesian framework, there is little agreement on the appropriate selection criteria. This paper compares the performance of the deviance information criterion (DIC), the pseudo-Bayes factor and the widely applicable information criterion (WAIC) in LMMs, with an extension to LMMs with skew-normal distributions. We focus on the comparison between the conditional criteria (given random effects) versus the marginal criteria (averaged over random effects). In spite of theoretical arguments, there is not much enthusiasm among applied statisticians to make use of the marginal criteria. We show in an extensive simulation study that the three marginal criteria are superior in choosing the appropriate longitudinal model. In addition, the marginal criteria selected most appropriate model for growth curves of Nigerian chicken. A self-written R function can be combined with standard Bayesian software packages to obtain the marginal selection criteria.

KEYWORDS
Deviance information criterion; Linear mixed models; Marginalized likelihood; Pseudo Bayes factor, Widely applicable information criterion.

1. Introduction

Longitudinal studies have become central in a great variety of research areas. The longitudinal study design is the only study design that allows to relate determinants measured at the start of the study to changes in the subjects' condition over time. Numerous books have recently appeared on longitudinal study designs, see e.g. (2; 12; 13; 21; 35). When the response is Gaussian, linear mixed-effects models (LMMs) are one of the most popular tools to analyze longitudinal data. Since its introduction by Laird and Ware (27), the LMM has been applied in a great variety of research areas and extended in many ways, e.g. to generalized linear mixed-effects models and non-linear mixed-effects models. Its popularity has much to do with its ability to describe both the impact of covariates on the mean longitudinal evolution as well as

Contact: Oludare Ariyo. Email: oludaresamuel.ariyo@kuleuven.be

how individual profiles differ over time from the mean curve. The impact on the mean longitudinal curve is evaluated by their regression coefficients, which are referred to as the fixed effects. The subject-specific profiles are expressed as latent variables, called random effects. In this way, the LMM fits subject-specific profiles and accounts for correlation among responses from the same subject. Another important feature is that the LMM allows for unbalanced data, i.e., when the number and timing of the observations per subject differ between subjects. The LMM parameters may be estimated using a frequentist approach. The properties of the estimated model parameters are then based on (restricted) maximum likelihood theory (54). Alternatively, one could use the Bayesian framework. In the Bayesian approach prior information on the model parameters is combined with information coming from the data. Using Bayes' theorem, an updated idea on the model parameters is obtained from the posterior distribution. The posterior distribution provides all information that is needed, and hence there is no need to refer to asymptotic normality properties for inference on the model parameters. This is especially useful in longitudinal studies with a small number of subjects and when the data are unbalanced (45). Since most posterior distributions are analytically intractable, they need to be determined in a numerical way. Most popular numerical techniques are based on sampling from the posterior distribution. The Markov chain Monte Carlo (MCMC) techniques provide an important class of such methods. In this paper we focus on fitting Bayesian LMMs to longitudinal data and compare the performance of different selection criteria. While in a Bayesian model all parameters are stochastic (and thus random), we will (as many others) still use the standard terminology of fixed and random effects.

A variety of LMMs can be fitted to the data at hand depending on several aspects such as: (i) the covariates that are considered in the fixed part of the model, (ii) the random effects structure to be included, e.g., random intercepts and/or random slopes, and (iii) possible transformations of the response. When considering several LMMs, it is important to select a parsimonious model that fits adequately the current and also future data. Unfortunately, there is little agreement on what criterion to choose for Bayesian model selection.

One of the first model selection criteria suggested in the literature is the Bayes factor (24), which is defined as the ratio of the marginal likelihood of two competing models. Although this criterion has a natural interpretation, its computation remains difficult in practice and the results can be sensitive to the choice of the prior distributions, presenting difficulties especially with improper priors. Gelfand and Dey (15) proposed the pseudo-Bayes factor (PSBF), which updates the (improper) prior to a proper posterior and calculates the Bayes factor using the generated posterior as prior. This alternative criterion, although relatively easy to compute, is not yet commonly used. The most popular Bayesian model selection criterion is the deviance information criterion (DIC) (48). The DIC is similar to the AIC often used in the frequentist framework, i.e., it represents a trade-off between model fit and model complexity. The aim of DIC is to estimate the predictive ability of the fitted model to future samples from the same population. More recently, the widely applicable information criterion (WAIC) was proposed (55) for model selection in the Bayesian framework. This criterion estimates the predictive accuracy of the model and includes a bias correction for using the data twice, i.e., to estimate the model and to evaluate model's accuracy. It has also been argued that WAIC is a more fully Bayesian approach (compared to DIC) and is suitable for singular models, such as LMMs for longitudinal data when the random effects are considered as parameters in the model (18).

Apart from the above three model selection criteria, a wide variety of (Bayesian) sta-

tistical approaches have been suggested to select the most appropriate LMM. While it is not the aim of this paper to give a comprehensive overview, the reader should be aware of the large number of alternative approaches proposed in the literature. For instance, a popular alternative approach is to use Bayesian variable selection techniques, often based on the SSVS approach of George & McCulloch (19). Examples of this approach can be found in Chen & Dunson (7),Cai & Dunson (5) and Gong et al (20).

Bayesian software for hierarchical models most often makes use of the data augmentation (DA) algorithm. For the LMM, this implies that the random effects are estimated jointly with the other parameters. Hereby, the DA algorithm avoids to take the integral over the distribution of the random effects, which is the classical approach in the frequentist framework. Thus, in the frequentist approach classically the marginal version of the LMM is fitted to the data, while in the Bayesian approach the hierarchical or conditional version of the LMM is usually fitted.

Whether the marginal or the conditional version of the LMM is fitted to the data, it has an impact on the performance of the model selection criteria even when the conditional and marginal LMM essentially lead to the same model. The model selection criteria applied to the hierarchical specification of the LMM is referred to as the conditional criterion. Hence, one has the conditional DIC (cDIC), and similarly the conditional PSBF (cPSBF) and the conditional WAIC (cWAIC). On the other hand when the model selection criterion is applied to the marginal specification of the LMM, one speaks of the marginal DIC (mDIC), marginal PSBF (mPSFB) and marginal WAIC (mWAIC). As will be shown in Section 5, these two versions of the model selection criteria are associated with different aims: cDIC (and similarly for cPSBF and cWAIC) considers the random effects as parameters of focus in the model whereas for mDIC (also mPSBF and mWAIC) the population of random effects represents the focus. In practice, this implies for mixed effects models that the conditional selection criteria evaluate the performance of the model when the population consists of all (future) measurements of the subjects included in the current study, while the marginal version of the criteria measures the performance of the model for all (future measurements of all) future subjects from the same population.

The problem is that in practice, model selection is most often based on cDIC (cPSBF, cWAIC) because of computational convenience. Indeed, cDIC can be immediately calculated using the conditional likelihood and it is automatically reported by WinBUGS (50) and other Bayesian software. However, most researchers are interested in knowing how well the model performs in the future. That is why one argues that conditional model selection criteria have the wrong focus, see e.g (52). Apart from not having the correct focus, model selection based on cDIC is questionable because the properties of DIC are based on the log-concavity of the likelihood, a condition that is violated in hierarchical models when the latent variables are considered as parameters in the model (33). The implication of using cDIC as model selection has been documented via simulations for financial volatility models (6). The authors concluded that in contrast to mDIC, cDIC tends to select overly complex models. For overdispersed count data, Millar (37) pointed out that the conditional-level DIC is an unreliable tool for model selection, while the same is true for the conditional WAIC (38). Merkle et al (36) advocated the use of marginal information criteria for item response models, and show that mWAIC corresponds to leave-one-cluster-out, whereas cWAIC corresponds to leave-one-unit-out.

While we focus in this paper on Bayesian model selection, we note that also in the frequentist paradigm the performance of the conditional versus marginal model selection

criteria has been compared extensively. A broad overview of a wide range of model selection criteria for the LMM is discussed in Mũller, Scealy and Welsh (39) for model selection in a frequentist content, including conditional and marginal information criteria. A short section in that paper is devoted to the Bayesian paradigm. Further, Fang (11) showed that the marginal AIC (mAIC) is asymptotically equivalent to the leave-one-cluster-out cross-validation while the conditional AIC (cAIC) is asymptotically equivalent to the leave-one-observation-out cross-validation. Srivastava & Kubokawa (51) derived three conditional AICs and showed theoretically and by simulations that their proposals outperform cAIC and mAIC of Vaida and Blanchard (52). Finally, Sefken et al (46) introduce the R-package cAIC4 for the calculation of the cAIC for LMMs estimated with lme4. To determine the marginal criteria extra computations are needed, which renders them less popular.
In practice, researchers are often not aware of the difference between the marginal and conditional version of the information criteria, therefore, rely on default software (36). That is why we have set up a simulation study that compares the performance of the two versions of the selection criteria for LMMs with longitudinal data. The first set of simulations makes use of the classical model LMM assumptions, i.e. when the random effects and measurement errors have a normal distribution. In the second set of simulations, we have simulated from LMMs with a skewed-normal and t-distribution for the random effects and measurement errors. Finally, we considered settings were we select both fixed and random effect jointly. All these sets of simulations clearly show the superiority of the marginal selection criteria. Moreover, in the analysis of a real data set, we again illustrate that the conditional criteria choose the least appropriate LMM. In order to promote the use of the marginal criteria for LMMs, we have written R software for the LMMs considered in our simulation study that can easily be combined with classical Bayesian software to compute the criteria mDIC, mPSBF and mWAIC for LMMs.

The rest of the article is organized as follows. In Section 2 we present the classical linear mixed model for longitudinal data. In Section 3 we treat the skew-normal LMM. The model selection criteria are introduced in Section 4 and the difference between conditional and marginalized versions is discussed in Section 5. In Section 6 we compare the criteria in an extensive simulation study, in order to give some practical recommendations. We also compared alternative versions of DIC and WAIC as suggested in the literature. In the same section we discuss the simulation results when the normality assumption in the LMM is relaxed. A comparison of the conditional and marginal criteria on a real data set is done in Section 7. We give concluding remarks in Section 8.

2. The linear mixed-effects model

The classical LMM (27) for longitudinal data can be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{Y}_i$ is an m_i-dimensional response vector of measurements for the i-th subject, $(i = 1, \ldots, n)$. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $m_i \times p$ and $m_i \times q$-dimensional covariate matrices, respectively, and $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects. The residual component vector $\boldsymbol{\epsilon}_i$ is distributed as $N_{m_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an $m_i \times m_i$ positive-definite covariance matrix. It is usually assumed that $\boldsymbol{\Sigma}_i = \sigma^2_\epsilon \mathbf{I}_{m_i}$, where $\mathbf{I}_{m_i}$ denotes the identity matrix

of dimension m_i.

The q-dimensional random-effects vectors $\mathbf{b}_i$ are assumed independent from the residuals and distributed as $N_q(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a $q \times q$ positive-definite covariance matrix. Model (1) is called a mixed-effects model because it combines the fixed-effects structure $\boldsymbol{\beta}$ with the subject-specific random effects $\mathbf{b}_1, \ldots, \mathbf{b}_n$. The LMM is advantageous because the data are not required to be balanced, and additionally, the within- and between-individual variations can be explicitly modeled through $\boldsymbol{\Sigma}_i$ and $\mathbf{D}$, respectively.

In the frequentist setting, the model parameters are estimated from the marginalized model for the response, after integrating out the random effects (54). The marginalized distribution has a closed form for model (1), namely

$$p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Sigma}_i) = N_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i). \tag{2}$$

In the Bayesian framework, inference is usually based on the hierarchical formulation of the model. In the first hierarchical stage, the response follows the conditional distribution $p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \mathbf{b}_i) = N_{m_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = N_{m_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i)$, whilst in the second stage, the subject-specific effects are specified with distribution $p(\mathbf{b}_i|\mathbf{D}) = N_q(\mathbf{0}, \mathbf{D})$.

3. The skew-normal linear mixed model

A $m-$dimensional random vector $\mathbf{Y}$ follows a m-variate skew-normal (SN) distribution with location vector $\boldsymbol{\mu}_0 \in \mathbb{R}^m$, $m \times m$ positive definite scale matrix $\mathbf{H}$ and $m \times q$ skewness matrix $\boldsymbol{\Delta}$, if its density function is given by

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\mu}_0, \mathbf{H}, \boldsymbol{\Delta}) =& 2^q \boldsymbol{\phi}_m\left(\mathbf{y}|\boldsymbol{\mu}_0, \mathbf{H} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right) \\ & \times \boldsymbol{\Phi}_q\left(\boldsymbol{\Delta}'\left(\mathbf{H} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right)^{-1}(\mathbf{y} - \boldsymbol{\mu}_0)|\mathbf{0}, \left(\mathbf{I}_q + \boldsymbol{\Delta}'\mathbf{H}^{-1}\boldsymbol{\Delta}\right)^{-1}\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\phi}_m$ and $\boldsymbol{\Phi}_q$ are the density function and the cumulative distribution functions of the m-dimensional and q-dimensional normal distribution, respectively. If we substitute $\boldsymbol{\Delta} = \mathbf{0}$, equation (3) reduces to the usual symmetric multivariate distribution $N_m(\boldsymbol{\mu}_0, \mathbf{H})$. Arellano et al. (3) denote $\mathbf{Y} \sim SN_{m,q}(\boldsymbol{\mu}, \mathbf{H}, \boldsymbol{\Delta})$ and $\mathbf{Y} \sim SN_m(\boldsymbol{\mu}, \mathbf{H}, \boldsymbol{\Delta})$ when $m = q$. Also, when $m = q$, $\boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_m)$ and $\mathbf{H}$ diagonal, equation (3) reduces to the multivariate skew-normal distribution, see e.g. (47). In practical settings, when the response and the covariate are highly skewed distributed, it might be more realistic to assume a multivariate SN for both random effects and measurement error (22).

The classical LMM (1) can be extended by assuming that

$$\mathbf{b}_i \sim SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\Delta}_b) \qquad \text{and} \qquad \boldsymbol{\epsilon}_i \sim SN_{m_i}(\mathbf{0}, \boldsymbol{\Psi}_i, \boldsymbol{\Delta}_{\epsilon_i}), \qquad i = 1, \ldots, n,$$

all independent. This results in the following skew-normal linear mixed model (SNLMM):

$$\begin{aligned} \mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}_i, \boldsymbol{\Delta}_{\epsilon_i} &\sim SN_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Delta}_{\epsilon_i}) \\ \mathbf{b}_i|\mathbf{D}, \boldsymbol{\Delta}_b &\sim SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\Delta}_b), \end{aligned}$$

where $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ is a dispersion matrix, usually associated with the between-units

variances, with $\boldsymbol{\alpha}$ unknown parameters in $\mathbf{D}$. In addition, $\boldsymbol{\Delta}_{\epsilon_i}$ and $\boldsymbol{\Delta}_b$ are diagonal matrices with unknown elements $\delta_{\epsilon_{i1}}, \ldots, \delta_{\epsilon_{im_i}}$ and $\delta_{b_1}, \ldots, \delta_{b_q}$, respectively. These components correspond to the skewness parameters. The marginal version of the SNLMM was shown by Arellano et al (4) to be equal to

$$f_{Y_i}(\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\vartheta}) = 2^{m_i+q}\phi_{n_i}(\mathbf{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Psi}_i)\boldsymbol{\Phi}_{m_{i+q}}(\boldsymbol{\mu}_{2i} - \boldsymbol{\Gamma}_i\boldsymbol{\mu}_{1i}|\mathbf{0}, \mathbf{R}_i + \boldsymbol{\Gamma}_i\boldsymbol{\Lambda}_i\boldsymbol{\Gamma}_i'),$$

where for $i = 1, \ldots, n$:

$$\boldsymbol{\Psi}_i = (\delta_\epsilon^2 + \sigma_\epsilon^2)\mathbf{I}_{m_i} + \mathbf{Z}_i(\boldsymbol{\Delta}_b^2 + \mathbf{G})\mathbf{Z}_i', \qquad \mu_{1i} = \frac{\boldsymbol{\Lambda}_i\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\delta_\epsilon^2 + \sigma_\epsilon^2},$$

$$\mu_{2i} = \left(\frac{\delta_\epsilon}{\sqrt{\sigma_\epsilon^2(\delta_\epsilon^2 + \sigma_\epsilon^2)}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right), \qquad \boldsymbol{\Gamma}_i = \begin{pmatrix} \frac{\delta_\epsilon}{\sqrt{\sigma_\epsilon^2(\delta_\epsilon^2+\sigma_\epsilon^2)}}\mathbf{Z}_i \\ -\boldsymbol{\Delta}_b(\Delta_b^2 + \mathbf{G})^{-1} \end{pmatrix},$$

$$\mathbf{R}_i = \begin{pmatrix} \mathbf{I}_{mi} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}_q + \boldsymbol{\Delta}_b\mathbf{G}^{-1}\boldsymbol{\Delta}_b)^{-1} \end{pmatrix}, \qquad \boldsymbol{\Lambda}_i = \left((\boldsymbol{\Delta}_b^2 + \mathbf{G})^{-1} + \frac{\mathbf{Z}_i'\mathbf{Z}_i}{\delta_\epsilon^2 + \sigma_\epsilon^2}\right).$$

Note that Arellano et al (4) also suggested a skew-t distribution whereby the basic Gaussian distribution is replaced by the t-distribution.

4. Bayesian criteria for model selection

Let $\boldsymbol{\theta}$ represent all model parameters of the LMM. For the marginal LMM, this includes the fixed effects and the parameters making up the covariance matrix of the random effects augmented with skewness parameters for the SNLMM. With the conditional LMM the random effects are part of $\boldsymbol{\theta}$. Further, we denote the collected (longitudinal) responses by $\mathbf{y}$ and the obtained covariate values by the matrix $\mathbf{X}$. The posterior distribution is $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X})p(\boldsymbol{\theta})/p(\mathbf{y} \mid \mathbf{X})$. Since the posterior distribution does not have a closed form for the LMM, it is approximated using MCMC methods. Namely, K (dependent) values $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ are sampled from the posterior distribution. The true posterior summary measures can then be approximated by their sampled versions.

When describing longitudinal data, a set of well-justified models can be established with different specifications for the fixed effects, random effects, covariance structure of the random effects and measurement error. Therefore, a model selection procedure is necessary to find an adequate model that explains current and future data. A variety of model selection procedures has been proposed in the Bayesian framework, but there is no consensus about the best criterion. Here we discuss the most popular criteria; they are also relatively easy to compute in practice.

4.1. The pseudo-Bayes factor

The Bayes factor (BF) could be viewed as the Bayesian equivalent of the likelihood ratio test. The Bayes factor can be used for testing the hypothesis that $\mathbf{y}$ is generated by model M_1 with parameters $\boldsymbol{\theta}_1$ versus the alternative model M_2 with parameters

$\boldsymbol{\theta}_2$. Hereby BF measures the change from prior to posterior odds in favor of the null model, namely

$$\mathrm{BF}_{1,2} = \frac{p(\mathrm{M}_1 \mid \mathbf{y})}{1 - p(\mathrm{M}_1 \mid \mathbf{y})} = \frac{p(\mathrm{M}_1 \mid \mathbf{y})}{p(\mathrm{M}_2 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid \mathrm{M}_1)\, p(\mathrm{M}_1)}{p(\mathbf{y} \mid \mathrm{M}_2)\, p(\mathrm{M}_2)},$$

where $p(\mathrm{M}_1)$ and $p(\mathrm{M}_2)$ are the prior model probabilities, commonly set as $p(\mathrm{M}_1) = p(\mathrm{M}_2) = 0.5$. In that case, the Bayes factor in favor of model M_1 is given by $\mathrm{BF}_{1,2} = p(\mathbf{y} \mid \mathrm{M}_1)/p(\mathbf{y} \mid \mathrm{M}_2)$ where $p(\mathbf{y} \mid \mathrm{M}_r) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_r, \mathrm{M}_r)\, p(\boldsymbol{\theta}_r \mid \mathrm{M}_r)\, d\boldsymbol{\theta}_r$ for $r = \{1, 2\}$. The use of the Bayes factor is, however, limited in practice since it has been shown to be quite sensitive to the choice of the prior distributions $p(\boldsymbol{\theta}_r \mid \mathrm{M}_r)$ and is not defined for improper priors, see e.g. (15).

Several alternatives for BF have been suggested to reduce the impact of $p(\boldsymbol{\theta}_r \mid \mathrm{M}_r)$. One proposal is PSBF, which is based on the partitions of the data set as follows. For the ith subject, one partitions the data set into a learning set $\mathbf{y}_L = \{\mathbf{y}_i : i \in L\}$ and a testing set $\mathbf{y}_T = \{\mathbf{y}_i : i \in T\}$ (14), whereby the testing and learning parts are defined respectively as $T = \{i\}$ and $L = \{1, ..., i-1, i+1, ..., n\}$. The pseudo-Bayes factor in favor of model M_1 with respect to model M_2 is then obtained as

$$\mathrm{PSBF}_{1,2} = \frac{\prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathrm{M}_1)}{\prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathrm{M}_2)},$$

where $\mathbf{y}_{(i)}$ is the total sample without $\mathbf{y}_i$. The component $p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathrm{M}_r)$ is the probability of observing $\mathbf{y}_i$ given the model M_r fitted with all observations in the sample except $\mathbf{y}_i$. Thus, the PSBF makes use of pseudo-marginal likelihoods in the numerator and denominator instead of the classical marginal likelihoods. The product terms are called conditional predictive ordinates (CPOs) (15). For the ith subject under model M_r, $\mathrm{CPO}_{r,i}$ is defined as $\mathrm{CPO}_{r,i} = p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathrm{M}_r)$. $\mathrm{CPO}_{r,i}$ is computed from the sampled values $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^K$ under model M_r as follows:

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

This statistic can be highly unstable for a very small value of the likelihood (44). To ensure stability, different approaches have been prescribed in the literature (9; 10; 15; 44). However, there is no perfect approach due to computational issues (25).
The log-pseudo marginal likelihood is then for each model equal to $\mathrm{LPML}_r = \sum_{i=1}^{n} \log(\mathrm{CPO}_{r,i})$. Therefore, the $\mathrm{PSBF}_{1,2}$ in favor of model M_1 respect to model M_2 can be computed as

$$\mathrm{PSBF}_{1,2} = \exp(\mathrm{LPML}_1 - \mathrm{LPML}_2).$$

4.2. The deviance information criterion

The DIC suggested by Spiegelhalter et al. (48) is based on the predictive accuracy of the estimated model defined as

$$\mathrm{DIC} = -2 \log p(\mathbf{y} | \bar{\boldsymbol{\theta}}) + 2 p_{DIC}, \tag{4}$$

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

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

which quantifies the number of parameters to be estimated after incorporating the prior information into the model. As seen above, the point estimator is the posterior mean of the parameters, but other estimates such as the median have also been suggested.

Defining the deviance as $D(\boldsymbol{\theta}) = -2\log\{p(\mathbf{y}|\boldsymbol{\theta})\} + 2\log\{f(\mathbf{y})\}$, the effective number of parameters can alternatively be written as $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance.

For practical purposes, we can ignore $f(\mathbf{y})$. The mean deviance $\overline{D(\boldsymbol{\theta})}$ can be approximated by $\frac{1}{K}\sum_{k=1}^{K} D(\boldsymbol{\theta}^k)$ and the plug-in deviance $D(\bar{\boldsymbol{\theta}})$ by $D(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}^k)$. This criterion is popular because it is easy to compute once we have an MCMC sample and can be directly obtained in several Bayesian packages such as WinBUGS. However, DIC has been criticized, see (49) for details. For instance, DIC is not invariant to non-linear transformations of $\boldsymbol{\theta}$ and negative values for p_{DIC} can occur in some cases.

4.3. The widely applicable information criterion

The widely applicable information criterion (WAIC) (55) is a fully Bayesian estimator that averages over the posterior distribution of $\boldsymbol{\theta}$ instead of conditioning on a point estimator $\hat{\boldsymbol{\theta}}(\mathbf{y})$ as done for DIC. For a future observation $\tilde{\mathbf{y}}_i$, this criterion 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 $\boldsymbol{\theta}$. Since $\tilde{\mathbf{y}}_i$ is unknown, predictive accuracy is defined by the expected log-predictive distribution (elpd) as

$$\text{elpd}_i = E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)] = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i,$$

where f is the unknown distribution under the true model. For each observation of a new data set, elpd is computed to establish the predictive accuracy of that data set. This is called the expected log-pointwise predictive density (elppd) defined as $\text{elppd} = \sum_{i=1}^{n} E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)]$.

Predictive accuracy can also be defined with a point estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$, often $\hat{\boldsymbol{\theta}}(\mathbf{y}) = E(\boldsymbol{\theta}|\mathbf{y})$, as the expected log predictive distribution given the point estimator $\text{elpd}_{\hat{\boldsymbol{\theta}}(\mathbf{y})} = E_f(\log p(\tilde{y}|\hat{\boldsymbol{\theta}}(\mathbf{y})) = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i$. The log pointwise predictive distribution (lppd) based on the observed data 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 using an MCMC sample from the posterior distribution as

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right].$$

With the WAIC criterion, the expected log pointwise predictive density elppd is estimated as the log pointwise predictive distribution lppd with a bias correction $\widehat{\text{elppd}}_{WAIC} = \widehat{\text{lppd}} - p_{WAIC}$. The measure p_{WAIC} corresponds to an estimate of the effective number of parameters given by

$$p_{WAIC} = 2 \sum_{i=1}^{n} \left[\log \left(\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right) - \frac{1}{K} \sum_{k=1}^{K} \log \ p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right].$$

Note that, WAIC can be alternatively expressed as

$$\text{WAIC} = -2\widehat{\text{lppd}} + 2p_{WAIC},$$

similar to DIC in (4).
One of the strengths of WAIC is its invariability to the scale of the model parameters, which implies that WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, with h a strictly monotone function.

5. Marginal and conditional criteria

In practice, the choice between conditional and marginal information criteria should be motivated by the aim of the study (52). Most often, this means that the marginal model selection criteria should be used since they estimate the predictiveness of the model when new clusters (in longitudinal studies, this implies new subjects) are involved, whereas the conditional criteria estimate the predictiveness of the model when new elements in the cluster (in longitudinal studies, new observations from the existing subjects) are involved. Nevertheless, when it comes to selecting the correct LMM it might still be that conditional criteria do a good job. In other words, it might be that the relative ordering of preference models is basically the same for both the conditional and marginal criteria. All of these comments apply to all three considered model selection criteria, but since cDIC is obtained automatically in most Bayesian software, it is the standard criterion in practice. Therefore, the literature shows some focus on DIC when examining the performance of conditional and marginal criteria. Despite the popularity of DIC, many have shown that the asymptotic justification of DIC (48) does not hold for hierarchical models, see e.g. Li et al (31).

6. Simulation studies

We have carried out three simulation studies. In the first two studies we based the simulated data on two classical data sets: the Potthoff and Roy data set (41) and the Jimma Infant Growth study (28). They were chosen because the first is representative for a balanced longitudinal study, while for the second study the time points are (somewhat) irregular and subjects drop out from the study. Using the fitted LMMs as population models, the performance of the conditional and marginal versions of DIC, PSBF and WAIC are contrasted using simulations. mDIC can be obtained from a WinBUGS run by working with the marginal model instead of the hierarchical model. To avoid specifying the marginal model in the estimation process, an R function was implemented, which computes the marginalized version of DIC, PSBF and WAIC for a

Gaussian, skew-normal and skew-t distribution of the random effects and measurement error. This R function takes the parameters sampled in the MCMC procedure from any Bayesian package and calculates the marginalized version using the closed form (2) and its extensions allowing for skew-normal and skew-t distributions. In addition, the conditional version of the three criteria is also computed by this function.

The main objective of the simulation study is to assess how well PSBF, DIC and WAIC select the correct model. According to the *minimum value* strategy, the model with the minimum value for the criterion is selected. Several simulation studies examining the performance of AIC and BIC, see e.g. (29), suggest to select the more complex model only if they differ in the criterion value with more than 5. This will be referred to as the *absolute difference* strategy. We will apply this strategy to all criteria. However, there is no evidence that this criterion is justified outside DIC.

6.1. The data sets and population models

In the dental study analyzed by Potthoff and Roy (41), the distance in (mm) from the pituitary to the pterygomaxillary fissure was measured at years 8, 10, 12, and 14 on 11 girls and 16 boys. We fitted the following linear mixed model as a function of *age* and *sex* (0= Female, 1=Male):

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_{ij} + b_{0i} + \epsilon_{ij}, \quad (i = 1, \ldots, 27; j = 1, \ldots, 4), \tag{5}$$

where y_{ij} is the distance (mm) measure of child i at time j and b_{0i} is a random intercept assumed to follow $b_{0i} \sim N(0, \sigma_b^2)$. Using the SAS procedure MIXED (34), we obtained the following maximum likelihood estimates: $\hat{\beta}_0 = 24.9688$, $\hat{\beta}_1 = 1.4831$, $\hat{\beta}_2 = -2.3210$, $\hat{\sigma}_b^2 = 2.0495$ and $\hat{\sigma}_\epsilon^2 = 3.2668$. These values were used as true parameters in this simulation study. The Jimma Infant Growth data set is based on the growth characteristics of about 8000 live births from South-West Ethiopia examined between September 1992 and September 1993. The growth characteristics height, weight and arm circumference of the babies were examined approximately every 60 days, but there were occasional deviations from the planned visits. Also, some children dropped out from the study for a variety of reasons such as relocation of their parents during the study or death of the child. This creates an unbalanced structure for the data. For the purpose of this simulation study, 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 (28; 30) where a sample of 495 children was selected to fit the model. This subset will also be the basis for this simulation study. The weight evolves in a non-linear way. To make use of an LMM, the time variable age was transformed into $\text{newage}_{ij} = \sqrt{\text{age}_{ij}} - (\text{age}_{ij} + 1) - 0.02 \times \text{age}_{ij}$ using fractional polynomials (30). Initially, our population model is based on the following random intercept and slope model:

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{newage}_{ij} + \beta_3 \text{agem}_i + b_{0i} + b_{1i} \times \text{newage}_{ij} + \epsilon_{ij}, \tag{6}$$

assuming $(b_{0i}, b_{1i})' \sim N(\mathbf{0}, \boldsymbol{D})$. Again, the estimates from this model (see Appendix) are used as the true values for the parameters in the simulation.

6.2. Simulation study 1

In the first simulation study, we consider the most popular setting of assuming normality for the random effects and measurement error. We believe that it is essential to show the performance of the selection criteria in this most popular setting. The performance of the model selection criteria may depend on whether the models differ in the fixed components or the random effects structure. Therefore, we examined the performance of the conditional and marginal criteria under two scenarios. For each of the two data sets we considered two scenarios. In *Scenario I* we assumed that the random effects structure is known but that the considered models differ from the true model in the fixed part. For *Scenario II* we assumed that the fixed part is known but the random effects part is unknown.

Regarding the prior distributions, we assigned independent vague normal priors, $N(0, 1000^2)$ for the regression coefficients and a vague inverse gamma prior for the residual variance, i.e. $\sigma^2 \sim IG(0.001, 0.001)$. The conditionally conjugate prior for the random-effects covariance matrix is the inverse Wishart distribution, but this choice has been shown to be problematic when the number of clusters (here subjects) is small (16; 42). Therefore, we have taken uniform priors $U(0, 100)$ for the standard deviation of the random effects, see (16). For the models with at least random intercept and slope, we assigned a uniform prior distribution $U(-0.5, 0.5)$ for all pairwise correlations between random effects to ensure positive definiteness of the covariance matrix $\mathbf{D}$ (40) following a proof in (8).

6.2.1. The balanced case: the Potthoff and Roy data set

As indicated above, we have considered two scenarios:

Scenario I: We assumed that the random effects structure is correct and considered models that differ in the fixed part. Besides the true data-generating model (5), we considered an overspecified model, which includes the interaction of age with sex and an underspecified model, which ignores the effect of sex. Hence, the alternative models are

- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + \beta_3 \text{age}_{ij} \times \text{sex}_i + b_{0i} + \epsilon_{ij}$ (overspecified),
- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + b_{0i} + \epsilon_{ij}$ (underspecified).

Scenario II: We assumed that the fixed structure is correct 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, more specifically

- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + b_{0i} + b_{1i} \times \text{age}_{ij} + \epsilon_{ij}$ (overspecified),
- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + \epsilon_{ij}$ (underspecified).

We simulated 500 data sets based on model (5). The covariate age was taken as in the original data set and sex was generated from a Bernoulli distribution with probability of success equal to 0.6, where 0.6 is the proportion of boys in the original data set. All the models in this simulation study were estimated based on three chains of $15,000$ iterations (discarding the first $5,000$ as a burn-in) and thinning equal to 10. Convergence of the MCMC samples was assessed with the Brooks-Gelman-Rubin (BGR) diagnostic. In cases where BGR was larger than 1.1, a new MCMC sample was selected with $10,000$ extra iterations until obtaining convergence.

In Table 1, we present for each criterion and for the two selection strategies, the percentage of times the correct, the overspecified or the underspecified model was chosen. The performance of the marginalized criteria is clearly better than the conditional counterparts in all cases. For instance, when using the *minimum value* selection rule, in most cases the percentage of correct selection for the marginalized version is almost twice that of the conditional counterpart. In addition, note that for the *absolute difference* rule in Scenario I, the percentage of correct model selections for the conditional version of DIC and of WAIC is basically zero. This strategy seems to work well also for PSBF and WAIC in Scenario II, but not in Scenario I. In Scenario II, the conditional versions of DIC, PSBF and WAIC favor overspecified models with additional random effects as also observed in (6) for financial volatility models.

[Table 1 appear here]

6.2.2. The unbalanced case: the Jimma Infant growth study

Again we considered two scenarios:

Scenario I: We assumed that the random effects structure is correct and considered the following models that differ in the fixed part parameters, namely

- Model (6) and including the interaction newage $\times$ sex (overspecified),
- Model (6) but ignoring the covariate sex (underspecified).

Scenario II: We assumed that the covariates in the fixed part are correct and considered the following models that differ in the random effects structure, i.e.

- Model (6) and including an additional random slope for newage^2 (overspecified),
- Model (6) but ignoring the random slope for newage (underspecified).

We generated 500 data sets from model (6). The covariate age was taken as in the original data set (i.e 8,10,12,14) and sex was generated from a Bernoulli distribution with probability of success equal to 0.6, where 0.6 is the proportion of boys in the original data set. 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. We created an unbalanced data set by allowing subjects to drop out randomly at days 240, 300 or 360.

As shown in Table 2, the marginalized criteria strongly outperform their conditional counterparts in both scenarios and selection strategies. We see again for Scenario II that all conditional criteria support the overspecified alternative with an additional random slope and that in this scenario the *absolute difference* strategy also works for PSBF and WAIC. With the *minimum value* rule, the probability of correctly selecting the data-generating model is about 1/3 with the conditional criteria. Hence, carrying out model selection based on the conditional criteria performs worse than selecting the models at random.

[Table 2 appear here]

6.3. Simulation study 2: additional simulations for the balanced case

We first evaluated the sensitivity of the results to some changes in the population model based on the Potthoff and Roy data. First, we varied the signal-to-noise ratio in model (5) by setting the value of σ^2_ϵ to be $\frac{1}{4}$, $\frac{1}{2}$, 1, 2 and 4 times of the estimated residual variance as specified in Section 6.1. Table 3 displays the results on model selection. Again, the marginal criteria outperform their conditional counterparts irrespective of the scenario and selection strategy. Note that the performance of mDIC decreases with increasing residual variance and using the *absolute difference* strategy.

[Table 3 appear here]

Second, we varied the number of subjects in the study as 25, 50, 75 and 100. As shown in Table 4, the marginal criteria perform best regardless of the sample size. Note also that the performance of the marginal criteria increases with increasing sample size in both scenarios and selection strategies, which is not the case for the conditional criteria. For instance, the percentage of correct model selection for cDIC decreases with sample size for Scenario II with both selection rules.

Our results are in line with the findings in (33), who pointed out asymptotic problems with cDIC. Our simulation study also indicates that cWAIC is not better in this sense.

[Table 4 appear here]

We additionally evaluated the model selection performance for alternative versions of DIC and WAIC. We denote as DIC_1 the criterion advocated in (48) where the complexity (p_{DIC1}) is defined in Section 4.2. The alternative version DIC_2 is the approximation to DIC_1 (17). The complexity penalty (p_{DIC2}) is a function of the variance of the deviance calculated as

$$p_{DIC2} = 2\text{var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log\{p(\mathbf{y}|\boldsymbol{\theta})\}). \tag{7}$$

Further, we modified DIC by letting the penalty term depend on the sample size. It has been suggested in (23) that the penalization should be defined based on the effective sample size n_e, which depends on the within-subjects error structure. In the context of the LMM, statistical software like SAS defines n_e as the total number of (independent) subjects, i.e. $n_e = n$. Otherwise, n_e is defined as the number of total data points, $n_e = n_T$. We defined the following DIC criteria as DIC_3 and DIC_4 with effective degrees of freedom defined as $p_{DIC3} = \log(n)\, p_{DIC1}$ and $p_{DIC4} = \log(n_T)\, p_{DIC1}$, respectively. These modifications are more a BIC-type as pointed out by a referee, however, we believe that it will be a useful exercise to evaluate their performance in this context.

The effective number of parameters of WAIC can be estimated in two ways (18); p_{WAIC1} as defined in Section 4.3 and the alternative version p_{WAIC2} given as the variance of the log posterior distribution as

$$p_{WAIC2} = \sum_{i=1}^{n} \text{var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}_i|\boldsymbol{\theta})).$$

We notice from Table 5 that Spiegelhalter's DIC (DIC_1) outperforms DIC_2 for the conditional versions. This may be expected since the alternative definition (7) is

explicitly based on approximate posterior normality, which is likely not satisfied in the hierarchical version of the model. The marginal versions of DIC_1 and DIC_2 perform similarly.

As expected, DIC_4 penalizes model complexity more heavily than DIC_3. Regardless of the selection strategy, we observed that by increasing the penalization, the percentage of correct model selection decreases under the marginal versions and increases under the conditional versions.

As for the different versions of WAIC, we observed that the percentage of correct selection for $WAIC_2$ is slightly higher in the conditional version whereas the performance of the marginal versions is similar irrespective of the scenario. *Absolute difference*, however, is not a good alternative to the conditional version of DIC and WAIC alternatives.

[Table 5 appear here]

6.4. Simulation study 3: extra simulation for possible extensions of LMM

6.4.1. Simulation study: jointly selection of both fixed and random effects

Depending on the data at hand, researchers are usually faced with the challenge of choosing the correct model. It is therefore important to select a parsimonious model that fits the data accurately. Since there is minimal agreement on which criteria to choose for Bayesian model selection, we evaluated the performance of the marginal and conditional criteria in choosing the correct model among other alternative models. Based on Potthoff & Roy data, we generated 500 data sets from Equation (5) and considered five possible alternative models for the data. We considered, namely, (i) different scale of the covariates (ii) distributional assumptions not satisfied for either or both random-effects and measurement error (iii) the nature of measurement error (heteroscedastic or heteroscedastic) (iv) wrong random effects structure. The following models were considered jointly with the model given by Equation (5).

- C1: The model generating data specified in Equation (5).
- C2: Equation (5) with age replaced by age^2 and including an additional random slope for age.
- C3: Equation (5) age replaced by age^2.
- C4: Equation (5) age replaced by log(age).
- C5: Equation (5) with the normality assumption for random effects replaced by the skew-normal assumption.
- C6: Equation (5) with the normality assumption for random effects replaced by the skew-normal assumption and heteroscedastic measurement error is assumed.

As seen in Table 6, the marginal criteria select the data-generating model (C1) in about 70% of the times contrary to the conditional criteria which select the true model in about 10% of the time. It is interesting to note that the conditional criteria select C5 (the model that assumes a skew-normal distribution for the random effects) in about 65% while the marginal criteria choose C5 in about 2%. The results show the superiority of the marginal criteria in selecting the true data-generating model.

[Table 6 appear here]

6.4.2. Simulation study: normality assumption for the random effects and measurement errors are relaxed

We also assessed the performance of the model selection criteria when the normality assumption for the random effects and measurement errors are relaxed. For this simulation study, we generated 500 data sets from the model

$$y_{ij} = \beta_0 + x_i\beta_1 + t_{ij}\beta_2 + b_{0i} + \epsilon_{ij}, i = 1, \ldots, n = 200, j = 1, \ldots, 6 \tag{8}$$

where $t_{ij} = j$, $\beta_1 = 2$, $\beta_2 = 1$ and $\epsilon_{ij} \sim SN_1(0, 0.5^2, 4)$.

First, we assumed that $\beta_0 + b_{0i} \sim N(4, 4)$, i.e, $\beta_0 = 4$ and $b_{0i} \sim N(0, 4)$. In addition, to show the advantages of the skew-normal distribution for the random effect it is penchant to accommodate skewness. Second, we have taken the previous one except now we generated the $\beta_0 + b_{0i}$ according to $Gamma(2, 1)$ distribution (as done also in (4) and (26)) with probability density $f(x) = x\exp(-x)$ yielding a highly skewed distribution. The subject-specific covariate x_i is binary with $x_i = 1$ if $i \leq n/2$ and is zero otherwise, while t_{ij} represents a covariate with values varying within individuals and the same for all individuals. For each of the 500 simulated data sets, model (8) was fit under alternative models as described in Section 6.2.1. We sampled 7000 iterations after discarding the initial 3000 iterations. The thinning factor was at 7 to avoid correlation problems in the generated chains

The following vague priors were assigned: $\beta \sim N(0, 10^2)$, $\sigma^2_\epsilon \sim IG(0.001, 0.001)$, $\sigma^2_b \sim IG(0.001, 0.001)$, $\delta_\epsilon \sim N(0, 10^2)\mathbb{I}\delta_\epsilon > 0$, $\delta_b \sim N(0, 10^2)\mathbb{I}\delta_b > 0$. The marginal distribution corresponding to Equation (8) is expressed in the closed form, as seen in Section 3. The simulation results shown in Table 7 confirm the results obtained above under the Gaussian distribution.

Finally, we repeated the above simulation when (i) both random effects and random error have a skew-normal distribution and when (ii) the random error follows a $t(3)$ distribution. The results (not shown) confirm the above simulation results.

7. Application

The Nigerian indigenous chicken (NIC) data set describes the longitudinal evolution of the body weight (BW) of chickens of different breeds raised in a university experimental farm. Four hundred and sixteen chickens were measured every week from hatching up to 20 weeks. The study aimed to evaluate the growth of different chicken breeds. Here we considered two classes of progenies. Two hundred and seventy chickens were produced from the same parent stock (pure breed), while 146 chickens have different parents (cross breed). The rational for the study and the experimental design can be found in (1). See Figure 1 for the evolution of weights of the chickens over time. Assuming a quadratic growth model with subject-specific random intercept and slopes, we fitted an LMM model to the weight at the jth measurement time of the ith chicken as

$$y_{ij} = \beta_0 + \beta_1 breed_i + \beta_2 age_{ij} + \beta_3 age^2_{ij} + b_{0i} + b_{1i} age_{ij} + b_{2i} age^2_{ij} + \epsilon_{ij}, \tag{9}$$

where y_{ij} is the chicken body weight (kg); breed_i is the breed indicator (1=pure breed, 2=cross breed), the age_{ij} represents the age (standardized). For the purpose of this study, we limited the chicken's age to 13 weeks since after that age a considerable

amount of chicken died. Thus, $\mathbf{x}_{ij} = (1, \text{breed}_i, \text{age}_{ij}, \text{age}_{ij}^2)'$, $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})'$ and $\mathbf{Z}_{ij} = (1, \text{age}_{ij}, \text{age}_{ij}^2)$, $i = 1, \ldots, 416$, $j = 1, \ldots, 13$.

We first used model (9) together with the classical Gaussian assumptions as model to fit the weights of the chickens over time, and we refer to this as Model 9(a). Based on the model fit, Figure 2 shows histograms and the corresponding Q-Q plots of the standardization posterior means of $\mathbf{b}_i$ and ϵ_{ij}, whereby the posterior means were divided by their corresponding posterior standard deviations. The plots show that there is apparently a non-normal pattern for subject-specific intercepts and slopes. Also, the residual plot suggests deviation from normality. We note that such plots may be difficult to interpret because the shrinkage effect depends on the number of measurements per subject, see e.g. (53). But here there were no missing responses up to week 13 and standardisation was applied. Nevertheless, these plots triggered us to consider three additional models with the same fixed effects structure but differing in the error and random effects distribution:

- **Model 9(b):** LMM with a univariate skew-normal distribution for measurement error and a trivariate Gaussian distribution for the random effects.
- **Model 9(c)**: LMM with model with a trivariate skew normal random effects with Gaussian measurement error.
- **Model 9(d):** LMM with a univariate skew-normal distribution for measurement error and a trivariate skew-normal distribution for the random effects.

The vague priors used are the same as those described in Section 6.4.1. We used 25,000 iterations after discarding the first 10,000 and thinning was set to 10. Convergence of the MCMC samples was assessed with the BGR criteria. Resulting parameter estimates are shown in Table 8.

It can be observed from Table 8 that the conditional criteria support Model 9(b), which seems to be an incorrect model based on Figure 2. In contrast, the marginal criteria favor Model 9(d), which appears to be also the most appropriate model here. We further evaluated the effect of the quadratic term in the fixed and random effects. The results (results not shown) of both versions of the criteria show that age^2 is more important in the random effects part than in the fixed part and there is an agreement between the conditional and the marginal criteria on this.

[Figure 1 appear here]
[Figure 2 appear here]

8. Discussion

We have compared three Bayesian selection criteria in the context of LMM for longitudinal data. In addition, we extended these settings to the skew-normal and t(3) distribution for random effects and measurement error. The simulation studies show that the marginal criteria outperform their conditional counterparts. Our results confirm the results of (6) for volatility models, (32; 36; 38) for item response models and (43) in hierarchical models.

It is important to remark that calculating the marginalized criteria does not represent an additional computational effort for LMM since the marginalized likelihood can be written in a closed form at least for a number of important distributions for the random effects and measurement errors. However, for generalized linear mixed models computing the marginalized likelihood is more involved and numerical integration

methods are needed (43). The performance of the conditional criteria will be examined in a subsequent paper.
We examined two selection rules: *minimum value* and *absolute difference* for all criteria. However, our results did not show justification for *absolute difference* outside DIC.
In our simulation study, the performance for the marginalized versions of DIC, WAIC and PSBF is similar. However, in contrast to DIC, WAIC and PSBF have the advantage of being non-invariant to non-linear transformations of the parameters in focus. For this reason, our advice is to base model selection on the marginal versions of WAIC or PSBF. Nevertheless, our R function computes both the marginal and conditional versions of all three selection criteria with no additional computational efforts. The function can be downloaded from https://ibiostat.be/online-resources/bayesian.

Another useful exercise is to evaluate the performance of the selection criteria when varying the vague prior for the covariance matrix of the random effects. This is under current examination.

Word counts

9798 words

Acknowledgement(s)

computational 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. We would like to thank the anonymous reviewers and associate editor whose suggestions lead to substantial improvement in the paper. The authors appreciate Dr Mathew Adeleke of the Discipline of Genetics, School of Life Sciences, University of KwaZulu-Natal South African 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 University of Agriculture, Abeokuta Nigeria.

Notes on contributor(s)

Nomenclature/Notation

Notes

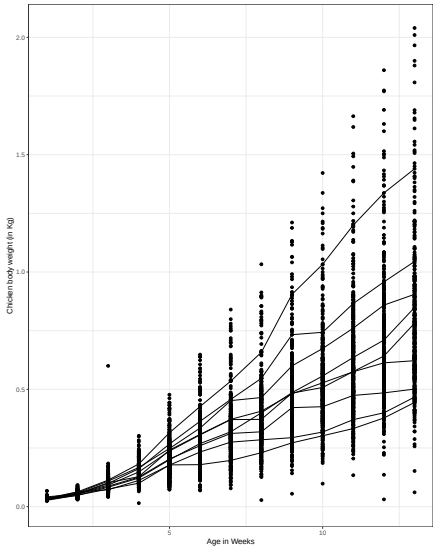

Figure 1. Nigerian indigenous chicken data set: Longitudinal profiles of body weight for 416 chickens highlighting 10 randomly chosen chickens

Table 1. Simulation study 1: Performance of the Bayesian model selection criteria for the Potthoff & Roy data set.

		Minimum value			Absolute difference		
Scenario	criteria	Over	Correct	Under	Over	Correct	Under
I	cDIC	18.6	67.6	13.8	2.4	1.0	96.6
	mDIC	16.8	76.4	6.8	1.4	55.2	43.4
	cPSBF	27.0	43.0	30.0	18.6	29.8	51.6
	mPSBF	17.6	75.2	7.2	2.8	65.2	32.0
	cWAIC	19.8	31.0	49.2	2.6	0.0	97.4
	mWAIC	18.8	75.0	6.2	1.4	58.4	40.2
II	cDIC	46.2	53.8	0.0	10.4	89.6	0.0
	mDIC	15.0	85.0	0.0	0.6	99.4	0.0
	cPSBF	52.4	47.6	0.0	32.0	68.0	0.0
	mPSBF	14.4	85.6	0.0	1.2	98.8	0.0
	cWAIC	63.2	36.8	0.0	16.0	84.0	0.0
	mWAIC	18.0	82.0	0.0	0.8	99.2	0.0

Table 2. Simulation study 1: Performance of the Bayesian model selection criteria for the Jimma Infant Growth data set.

Scenario		Minimum value			Absolute difference		
		Over	Correct	Under	Over	Correct	Under
I	cDIC	34.4	34.0	31.6	15.2	29.0	55.8
	mDIC	21.2	58.0	20.8	0.8	32.4	66.8
	cPSBF	33.0	32.8	34.2	47.0	31.8	21.2
	mPSBF	21.0	57.8	21.2	3.0	44.0	53.0
	cWAIC	36.2	31.2	32.6	14.4	26.4	59.2
	mWAIC	21.2	58.2	20.6	0.8	32.6	66.6
II	cDIC	63.2	36.8	0.0	43.2	56.8	0.0
	mDIC	26.4	73.6	0.0	0.2	99.8	0.0
	cPSBF	55.2	44.8	0.0	51.8	48.2	0.0
	mPSBF	28.0	72.0	0.0	2.8	97.2	0.0
	cWAIC	66.0	34.0	0.0	49.2	50.8	0.0
	mWAIC	27.4	72.6	0.0	0.2	99.8	0.0

Table 3. Simulation study 2: Percentage correct selection when changing the residual variance in the Potthoff & Roy data set.

Scenario		Minimum value					Absolute difference				
	Criteria	0.25	0.5	1	2	4	0.25	0.5	1	2	4
I	cDIC	64.6	70.2	77.0	77.8	79.2	0.6	1.2	3.2	10.8	24.6
	mDIC	81.6	83.0	83.0	82.8	82.0	93.0	92.8	92.6	88.6	78.6
	cPSBF	31.8	36.6	40.8	58.4	68.0	30.3	38.2	39.0	39.6	39.4
	mPSBF	91.2	94.0	83.2	90.8	87.8	95.4	97.8	93.0	97.8	94.0
	cWAIC	41.4	36.6	39.4	38.4	39.0	0.4	0.2	0.2	0.4	0.2
	mWAIC	81.2	81.6	82.4	82.0	81.6	92.2	93.0	92.8	89.0	79.0
II	cDIC	44.4	47.4	50.8	51.6	55.4	86.2	86.2	87.2	88.4	89.0
	mDIC	80.4	82.4	83.6	85.4	86.4	99.2	99.4	99.6	99.6	90.2
	cPSBF	60.4	58.4	44.8	62.2	73.4	52.0	55.8	65.8	67.5	69.6
	mPSBF	83.8	86.8	84.2	84.0	83.8	98.7	97.9	97.6	91.2	86.4
	cWAIC	34.4	32.8	34.2	36.6	36.2	81.4	81.8	83.4	81.0	82.6
	mWAIC	77.6	81.0	82.6	82.0	82.4	97.6	99.2	99.2	99.0	92.2

Table 4. Simulation study 2: Percentage correct selection when changing the sample size in the Potthoff & Roy data set.

Scenario		Minimum value				Absolute difference			
	Criteria	25	50	75	100	25	50	75	100
I	cDIC	67.6	77.0	79.0	80.6	1.0	3.2	7.2	19.0
	mDIC	76.4	83.0	84.2	82.8	52.2	92.6	98.4	99.0
	cPSBF	43.0	40.8	49.4	45.4	0.0	0.4	0.8	0.0
	mPSBF	75.2	83.0	84.4	83.0	83.1	93.0	93.2	96.1
	cWAIC	31.0	39.4	41.4	43.8	0.0	44.8	44.6	40.6
	mWAIC	75.0	82.4	83.8	82.0	56.2	92.8	98.8	98.8
II	cDIC	53.8	50.8	47.4	41.0	89.6	87.2	87.2	84.8
	mDIC	85.0	83.6	86.2	83.8	99.2	99.6	99.2	99.4
	cPSBF	47.6	44.8	47.8	53.0	65.2	65.8	66.2	65.8
	mPSBF	85.6	84.2	86.0	83.0	90.2	97.6	97.6	97.9
	cWAIC	36.8	34.2	34.2	31.8	83.8	83.4	83.2	82.6
	mWAIC	82.0	82.6	84.6	80.2	99.4	99.2	99.2	99.3

Table 5. Simulation study 2: Performance of alternative criteria for the Potthoff & Roy data set.

		Minimum value			Absolute difference		
Scenario	Criteria	Over	Correct	Under	Over	Correct	Under
I	$cDIC_1$	18.6	67.6	13.8	2.4	1.0	96.6
	$cDIC_2$	11.8	36.0	52.2	1.6	0.0	98.4
	$cDIC_3$	3.2	85.0	11.8	0.6	22.8	76.6
	$cDIC_4$	4.2	40.4	55.4	1.4	19.6	79.0
	$cWAIC_1$	19.8	31.0	49.2	2.6	0.0	97.4
	$cWAIC_2$	16.8	41.2	42.0	2.6	0.0	97.4
	$mDIC_1$	16.8	76.4	6.8	1.4	55.2	43.4
	$mDIC_2$	16.8	73.8	9.4	1.4	52.2	46.4
	$mDIC_3$	1.8	65.4	32.8	0.2	34.0	65.8
	$mDIC_4$	2.8	53.2	44.0	0.2	24.0	75.8
	$mWAIC_1$	18.8	75.0	6.2	1.4	58.4	40.2
	$mWAIC_2$	17.8	75.2	7.0	1.4	56.2	42.4
II	$cDIC_1$	46.2	53.8	0 .0	10.4	89.6	0.0
	$cDIC_2$	0.6	99.2	0.2	0 .0	99.4	0.6
	$cDIC_3$	0.0	47.8	52.2	0.0	36.8	63.2
	$cDIC_4$	0.0	0.8	99.2	0 .0	0.6	99.4
	$cWAIC_1$	63.2	36.8	0.0	16.0	84.0	0.0
	$cWAIC_2$	55.8	44.2	0.0	10.4	89.6	0.0
	$mDIC_1$	15.0	85.0	0.0	0.6	99.4	0.0
	$mDIC_2$	8.0	92.0	0.0	0.2	99.8	0.0
	$mDIC_3$	2.4	97.6	0.0	0.2	99.6	0.2
	$mDIC_4$	0.4	99.6	0.0	0.0	99.2	0.8
	$mWAIC_1$	18.0	82.0	0.0	0.8	99.2	0.0
	$mWAIC_2$	15.4	84.6	0.0	0.8	99.2	0.0

Table 6. Simulation study 3: Percentage of times the criteria selection select the required model described in Section 6.4.1 in the Potthoff &Roy data set.

	Model					
Criteria	C1	C2	C3	C4	C5	C6
cDIC	12.8	7.0	3.6	4.0	70.6	2.0
cWAIC	13.2	8.4	8.0	4.6	64.2	1.6
cPSBF	10.8	10.6	6.0	5.8	66.8	0.0
mDIC	76.2	18.4	1.2	2.8	1.4	0.0
mWAIC	67.4	20.4	2.2	3.0	4.2	2.8
mPSBF	74.8	8.6	11.4	3.4	1.8	0.0

Table 7. Simulation study 3: Performance of the Bayesian model selection criteria for Gamma(2,1) for random error and N(0,4) for random effect.

		Minimum Value			Absolute difference		
Scenario	Criteria	Over	Correct	Under	Over	Correct	Under
I	cDIC	29.6	43.2	27.2	39.8	60.2	0.0
	mDIC	13.0	60.8	26.2	22.4	77.6	0.0
	cPSBF	59.0	28.2	12.8	46.6	52.4	1.0
	mPSBF	11.0	67.4	21.6	44.2	55.8	0.0
	cWAIC	25.4	51.4	23.2	32.6	67.4	0.0
	mWAIC	11.0	62.4	26.6	20.2	79.8	0.0
II	cDIC	18.2	26.4	55.4	38.2	61.8	0.0
	mDIC	18.2	64.4	17.4	15.6	84.4	0.0
	cPSBF	19.2	56.4	37.2	47.2	51.4	1.4
	mPSBF	14.6	70.2	15.2	19.2	78.8	2.0
	cWAIC	15.6	20.4	64.0	32.2	67.8	0.0
	mWAIC	18.2	66.0	15.8	14.4	85.6	0.0

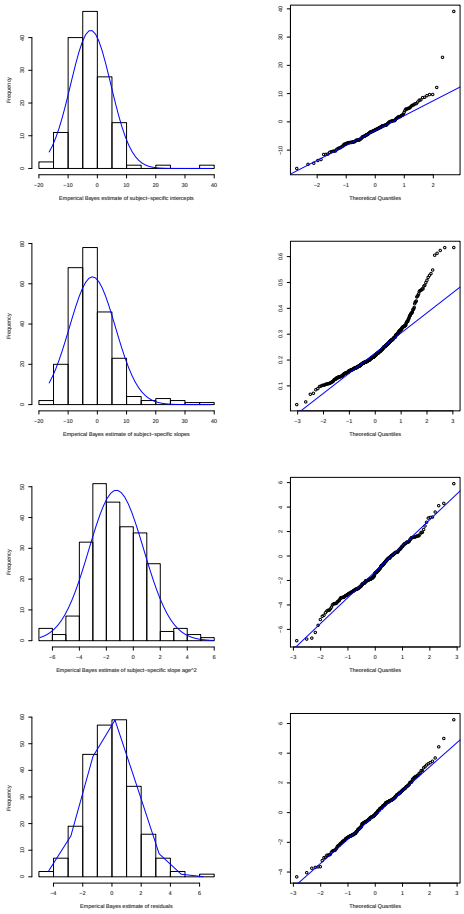

Figure 2. Nigerian indigenous chicken data set: Histogram and normal Q-Q plots for standardised posterior means of random effects based on Model 9(a): Subject-specific intercepts in the first row, subject-specific slope of age in the second row, subject-specific slope for the age^2 in the third row and residual in the fourth row.

Table 8. Nigeria indigenous chicken data set: Posterior mean (regression coefficients) & median (variance parts), 95% probability intervals and the conditional and marginal criteria under the four fitted models, see Section 7

	Model 9 a			**Model 9 b**			**Model 9 c**			**Model 9 d**		
	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%
β_0	0.335	0.321	0.349	0.369	0.284	0.848	0.359	0.353	0.374	0.315	0.299	0.329
β_1	-0.008	-0.014	-0.001	-0.009	-0.018	0.000	-0.028	-0.030	-0.021	-0.029	-0.035	-0.023
β_2	0.239	0.229	0.249	0.308	0.227	0.853	0.235	0.231	0.245	0.232	0.221	0.242
β_3	0.031	0.027	0.034	0.046	0.028	0.223	0.031	0.030	0.032	0.030	0.028	0.031
δ_{b1}	-	-	-	-	-	-	0.003	0.001	0.009	0.003	0.000	0.009
δ_{b2}	-	-	-	-	-	-	0.002	0.001	0.007	0.002	0.000	0.007
δ_{b3}	-	-	-	-	-	-	0.002	0.001	0.007	0.002	0.000	0.007
δ_ϵ	-	-	-	0.051	0.048	0.054				0.060	0.055	0.064
$d11$	0.013	0.011	0.015	0.013	0.011	0.319	0.015	0.014	0.017	0.014	0.012	0.016
$d12$	0.010	0.009	0.012	0.010	0.008	0.318	0.007	0.001	0.040	0.008	-0.012	0.031
$d13$	0.001	0.000	0.001	0.000	0.000	0.098	0.005	-0.002	0.023	0.004	-0.019	0.024
$d22$	0.010	0.008	0.011	0.010	0.008	0.383	0.008	0.003	0.122	0.009	0.001	0.085
$d23$	0.002	0.001	0.002	0.002	0.001	0.123	-0.003	-0.011	0.002	-0.003	-0.069	0.002
$d33$	0.001	0.001	0.001	0.001	0.001	0.039	0.008	0.004	0.081	0.006	0.001	0.068
σ_ϵ	0.001	0.001	0.001	0.000	0.000	0.001	0.002	0.002	0.002	0.001	0.001	0.001
cDIC		-19117.4			-19809.10			-19710.85			-18574.70	
cWAIC		-19782.2			-20361.42			-19117.48			-20128.30	
cplppd		-15242.3			-15945.86			-15414.23			-16113.33	
mDIC		-16821.6			-15673.10			-17269.46			-17362.04	
mWAIC		-16808.5			-15472.20			-17488.41			-17511.63	
mlppd		-16665.4			-16965.43			-16765.43			-17165.43	

References

[1] Adeleke, M., Peters, S., Ozoje, M., Ikeobi, C., Bamgbose, A., and Adebambo, O. A. 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 (2011).

[2] Anderson, S. J. Longitudinal Study Designs. *Handbook of Research Methods in Health Social Sciences* (2018), 1–20.

[3] Arellano-Valle, R., Bolfarine, H., and Lachos, V. Skew-normal linear mixed models. *Journal of Data Science 3*, 4 (2005), 415–438.

[4] Arellano-Valle, R., Bolfarine, H., and Lachos, V. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics 34*, 6 (2007), 663–682.

[5] Cai, B., and Dunson, D. B. Bayesian covariance selection in generalized linear mixed models. *Biometrics 62*, 2 (2006), 446–457.

[6] Chan, J., and Grant, A. On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics 14*, 4 (2016), 772–802.

[7] Chen, Z., and Dunson, D. B. Random effects selection in linear mixed models. *Biometrics 59*, 4 (2003), 762–769.

[8] Coakley, E. S., and Rokhlin, V. A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices. *Applied and Computational Harmonic Analysis 34*, 3 (2013), 379–414.

[9] Congdon, P. *Bayesian models for categorical data*. John Wiley & Sons, 2005.

[10] Dey, D. K., Chen, M.-H., and Chang, H. Bayesian approach for nonlinear random effects models. *Biometrics* (1997), 1239–1252.

[11] Fan, T.-H., Wang, Y.-F., and Zhang, Y.-C. Bayesian model selection in linear mixed effects models with autoregressive (p) errors using mixture priors. *Journal of Applied Statistics 41*, 8 (2014), 1814–1829.

[12] Funatogawa, I. *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Models*. Springer, 2017.

[13] Gayle, V., and Lambert, P. *What is Quantitative Longitudinal Data Analysis?* Bloomsbury Publishing, 2018.

[14] Geisser, S., and Eddy, W. F. A predictive approach to model selection. *Journal of the American Statistical Association 74*, 365 (1979), 153–160.

[15] Gelfand, A., and Dey, D. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society - Series B 56*, 3 (1994), 501–514.

[16] Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis 1*, 3 (2006), 515–534.

[17] Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian Data Analysis*. Chapman and Hall, 2004.

[18] Gelman, A., Hwang, J., and Vehtari, A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*, 6 (2014), 997–1016.

[19] George, E. I., and McCulloch, R. E. Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*, 423 (1993), 881–889.

[20] Gong, L., Flegal, J. M., Spindler, S. R., and Mote, P. L. Bayesian model selection on linear mixed-effects models for comparisons between multiple treatments and a control. *arXiv preprint arXiv:1509.07510* (2015).

[21] Hoffman, L. *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge, 2015.

[22] Huang, Y., and Dagne, G. Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses, and measurement errors in covariates. *Biometrics 68*, 3 (2012), 943–953.

[23] Jones, R. H. Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine 30*, 25 (2011), 3050–3056.

[24] Kass, R. E., and Raftery, A. E. Bayes factors. *Journal of the American Statistical*

Association 90, 430 (1995), 773–795.

[25] Lachos, V. H., Castro, L. M., and Dey, D. K. Bayesian inference in nonlinear mixed-effects models using normal independent distributions. *Computational Statistics & Data Analysis 64* (2013), 237–252.

[26] Lachos, V. H., Ghosh, P., and Arellano-Valle, R. B. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica* (2010), 303–322.

[27] Laird, N. M., and Ware, J. H. Random-effects models for longitudinal data. *Biometrics* (1982), 963–974.

[28] Lesaffre, E., Asefa, M., and Verbeke, G. 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 (1999), 835–854.

[29] Lesaffre, E., and Lawson, A. *Bayesian Biostatistics (Statistics in Practice)*. Wiley: Chichester, 2012.

[30] Lesaffre, E., Todem, D., and Verbeke, G. Flexible modelling of the covariance matrix in a linear random effects model. *Biometrical Journal 42*, 7 (2000), 807–822.

[31] Li, B., Bruyneel, L., and Lesaffre, E. A multivariate multilevel gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine 33* (2013), 1877–1899.

[32] Li, L., Qiu, S., Zhang, B., and Feng, C. X. Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing 26*, 4 (2016), 881–897.

[33] Li, Y., Zeng, T., and Yu, J. Robust deviance information criterion for latent variable models. *CAFE Research Paper No. 13.19 Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id* = 2316341(2013).

[34] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. *SAS for mixed models*. SAS institute, 2007.

[35] McArdle, J. J., and Nesselroade, J. R. *Longitudinal data analysis using structural equation models*. American Psychological Association, 2014.

[36] Merkle, E., Furr, D., and Rabe-Hesketh, S. Bayesian model assessment: Use of conditional vs marginal likelihoods. *arXiv preprint arXiv:1802.04452* (2018).

[37] Millar, R. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics 65*, 3 (2009), 962–969.

[38] Millar, R. B. Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing 28*, 2 (2018), 375–385.

[39] Müller, S., Scealy, J. L., Welsh, A. H., et al. Model selection in linear mixed models. *Statistical Science 28*, 2 (2013), 135–167.

[40] Plummer, M. Cannot invert matrix, November 2011. [Online; posted 11-November-2011].

[41] Potthoff, R., and Roy, S. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika 5* (1964), 313–326.

[42] Quintero, A., and Lesaffre, E. Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal 59*, 5 (2017), 1047–1066.

[43] Quintero, A., and Lesaffre, E. Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine 37*, 16 (2018), 2440–2454.

[44] Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics 8* (2007), 1–45.

[45] Raudenbush, S. W., and Bryk, A. S. *Hierarchical Linear Models: Applications and data analysis methods*, vol. 1. Sage, 2002.

[46] Säfken, B., Rügamer, D., Kneib, T., and Greven, S. Conditional model selection in mixed-effects models with cAIC4. *arXiv preprint arXiv:1803.05664* (2018).

[47] Sahu, S. K., Dey, D. K., and Branco, M. D. A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*

31, 2 (2003), 129–150.

[48] Spiegelhalter, D., Best, N., Carlin, N., and van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society - Series B 64*, 4 (2002), 583–639.

[49] Spiegelhalter, D., Best, N., Carlin, N., and van der Linde, A. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society - Series B 76*, 3 (2014), 485–493.

[50] Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. *WinBUGS User Manual*, 1.4 ed., 2003.

[51] Srivastava, M. S., and Kubokawa, T. Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis 101*, 9 (2010), 1970–1980.

[52] Vaida, F., and Blanchard, S. Conditional Akaike information for mixed-effects models. *Biometrika 92*, 2 (2005), 351–370.

[53] Verbeke, G., and Lesaffre, E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*, 433 (1996), 217–221.

[54] Verbeke, G., and Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, New York, 2000.

[55] Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11* (2010), 3571–3594.