Noname manuscript No.
(will be inserted by the editor)

Bayesian model selection for longitudinal count data

Oludare Ariyo · Emmanuel Lesaffre · Geert Verbeke · Adrian Quintero

Received: date / Accepted: date

Abstract We explore the performance of three popular model-selection criteria for generalised linear mixed-effects models (GLMMs) for longitudinal count data (LCD). We focus on evaluating the conditional criteria (given the random effects) versus the marginal criteria (averaging over the random effects) in selecting the appropriate data-generating model. We advocate the use of marginal criteria, since Bayesian statisticians often use the conditional criteria despite previous warnings. We discuss how to compute the marginal criteria for LCD by a replication method and importance sampling algorithm. Besides, we show via simulations to what extent we err when using the conditional criteria instead of the marginal criteria. To promote the usage of the marginal criteria, we developed an R function that computes the marginal criteria for longitudinal models based on samples from the posterior distribution. Finally, we illustrate the advantages of the marginal criteria on a well-known data set of patients who have epilepsy.

Keywords Replication sampling · Marginal likelihood · Bayesian model selection

Oludare Ariyo
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KU Leuven, Belgium
E-mail: ariyodare@gmail.com
Present address: Department of Statistics, Federal University of Agriculture, Abeokua, Nigeria

Emmanuel Lesaffre
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KU Leuven, Belgium

Geert Verbeke
Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KU Leuven, Belgium

Adrian Quintero
Icfes - Colombian Institute for Educational Evaluation, Statistics Department, Bogota, 111071, Colombia

1 Introduction

In a longitudinal study, subjects are monitored over time. Such a study type allows us to discover baseline or time-varying characteristics that have an impact on the outcome of interest. Generalised linear mixed models (GLMMs) are one of the most popular tools to analyze various types of outcomes repeatedly measured over time. The GLMM (McCullagh, 1989) is a generalisation of the linear mixed model including both fixed and random effects with a response having a distribution in the exponential family. In the frequentist approach, the model parameters are estimated by integrating out the random effects from the likelihood. Most often, this is done under the assumption of Gaussian random effects. The integral is then evaluated using non-adaptive or adaptive Gaussian quadrature methods. In contrast, in the Bayesian approach, the random effects are most often estimated together with the fixed effects. This implies that Bayesian computations are based on the conditional likelihood, which is the likelihood of the data given the random effects.

To find an appropriate GLMM for a (longitudinal) data set, one makes in the frequentist approach use of the likelihood ratio test for nested models or information criteria, such as AIC and BIC, for non-nested models. In the Bayesian approach, the same model selection criteria are used for both nested and non-nested models. One of such criteria to select between two models is the Bayes' factor (Kass and Raftery, 1995), defined as the ratio of the marginal likelihoods (marginalised over the prior of the model parameters) of the two competing models. While the Bayes' factor is an elegant Bayesian tool, there are serious issues with its computation in practice. Namely, it turns out that computing the Bayes' factor has proved to be at least as difficult as computing the posterior distribution, it cannot be computed with improper priors and is quite sensitive to the choice of the prior distribution. To overcome this problem, the pseudo-Bayes factor (PSBF) (Gelfand and Dey, 1994) has been suggested. To compute the PSBF one updates an (improper) prior to a proper posterior and calculates the Bayes' factor using the generated posterior as prior.

The most popular Bayesian model selection criterion is the Deviance Information criterion (DIC) (Spiegelhalter et al., 2002). The DIC aims to estimate the predictive ability of the fitted model to future samples from the same population, and like AIC and BIC, it represents a trade-off between the model fit and model complexity. However, the theoretical basis of DIC is not clear and several objections and alternatives have been formulated by the discussants of Spiegelhalter et al. (2002), see also Celeux et al. (2006) and Spiegelhalter et al. (2014). Recently, Watanabe's Widely Applicable Information Criterion (WAIC) (Watanabe, 2013) has been proposed. WAIC has been singled out as a worthy successor of DIC (Spiegelhalter et al., 2014). We consider PSBF, DIC and WAIC in this paper since there is little agreement in the statistical literature on the choice of these criteria for model selection.

Model selection criteria may be based on the conditional likelihood (given the random-effects) resulting in conditional criteria or on the marginal likelihood (integrating out the random effects) resulting in the marginal criteria.

The conditional criteria measure the predictiveness of the model for the subjects included in the current study, whereas the marginal criteria measure the predictiveness of the model for all subjects from the same population in a future study. Vaida and Blanchard (2005) pointed out that the choice of the criteria should be motivated by the research question. This implies that most often the marginal criteria should be used in practice. However, irrespective of that research question, the conditional criteria are most often used in practice because of convenience and their easy availability in most software. This usage has been questioned for LMMs (Ariyo et al., 2020, 2019) as well as for GLMMs (Millar, 2009; Christensen, 2017; Quintero and Lesaffre, 2018; Merkle et al., 2018). Ariyo et al. (2020) and Ariyo et al. (2019) explored the performance of the marginal model selection criteria for the LMM (based on the closed form of the marginal likelihood) and concluded their superior performance over the corresponding conditional criteria. An R program has been written that computes the marginal and conditional versions of PSBF, DIC and WAIC for any LMM based on MCMC output from a fitted model in our previous papers. However, for a GLMM there is no closed form for the likelihood. Hence, there is the need for an approach to compute the marginal model selection criteria for non-closed-form likelihoods such as for GLMMs.

Numerical methods have been developed that compute the marginal criteria for non-closed form likelihoods. For example, Chan and Grant (2014) proposed fast algorithms for computing the marginal DIC (mDIC) for a variety of high dimensional latent variable models and show that mDIC has much smaller numerical standard errors compared to the DIC based on the conditional likelihood (cDIC). Likewise, Chan and Grant (2016) proposed importance-sampling algorithms for computing mDIC under a variety of stochastic volatility models. In the INLA package, developed by Rue et al. (2009) for latent Gaussian models, the marginal posterior is computed by integrated nested Laplace approximations. Up to now, all these methods make use of a Gaussian assumption for the random effects. We will do that also in this paper; however, in principle, every method here can be easily adapted to a non-Gaussian distribution.

In this paper, we used a computational technique that computes the marginal criteria that involve specifying the marginalised likelihood components as an expectation of the conditional distribution. As such, the likelihood can be marginalised by generating replicate samples from the density of the random effects, which ought to be integrated out to estimate such expectation. This computational procedure can be carried out from the MCMC output of any Bayesian software; hence, it is widely applicable and easy to use. However, the major setback of this procedure is its computational complexity. When the number of observations and/or subjects increases, a large number of the replications may be required to obtain accurate results, hence the replication method can become impractical. As such, we give some recommendations for a trade-off between computational complexity and accuracy of the information criteria when usage of replication method is inevitable. To overcome the computational setback in the replication method, we have also used the importance sampling method and show via simulations and a real data set

of epilepsy patients that importance sampling is advantageous for computing marginal criteria in cases with a large number of observations as this method reduces the computation time by half. Nevertheless, the replication method is recommended for a smaller number of subjects.

Another contribution of this paper is to highlight the importance of the conditional/marginal criteria distinction for Bayesian model selection for generalised linear mixed-effects models (GLMMs) especially for overdispersed versions of such GLMMs and to recommend the need to utilise marginal criteria. Others have shown that the marginal criteria outperform the conditional counterparts for the classical (Gaussian) and less classical (non-Gaussian) LMM (Ariyo et al., 2020, 2019), in volatility models (Chan and Grant, 2016), for item response models (Merkle et al., 2018; Li et al., 2012) and for GLMMs in general (Quintero and Lesaffre, 2018; Millar, 2018). However, some settings have not been considered yet. Namely, we considered here settings that have been shown to affect the performance of the (frequentist) selection criteria (see for example Fitzmaurice, 1997; Howe et al., 2019; Chen et al., 2016; van Smeden et al., 2016, 2019) in Poisson mixed-effects models. Namely, we show the advantage of using the marginal criteria for longitudinal count data: (i) in the presence/absence of overdispersion, too many zeros, or both and (ii) when the number of repeated measurements is relatively small compared to the number of independent variables. Finally, to promote the usage of the marginal criteria, we developed an R function that computes the marginal criteria for a battery of longitudinal count models and this function is available in https://github.com/OludareAriyo/BayesselectGLMM.

The structure of the paper is as follows. Section 2 presents the general GLMM and introduces the Poisson mixed-effects model. In Section 3, we discuss the conditional and marginal selection criteria in generality. Section 4 presents and evaluates the sampling methods for the computation of marginal criteria of a GLMM. In Section 5, we discuss extensions of the Poisson mixed-effects model that deal with overdispersion in the repeated counts. In Section 6, our approach is illustrated on the well-known longitudinal epilepsy data set. Different simulation settings and scenarios are presented in Section 7. In the same section, we compare the performance of the conditional and marginal model selection criteria and evaluate the performance of the sampling techniques in computing the marginal criteria. The article concludes with a general discussion in Section 8.

2 Generalised linear models with cluster-specific effects

2.1 The generalised linear model

A random variable Y follows a distribution from the exponential family if its density is of the form

$$f(y) \equiv f(y \mid \lambda, \phi) = \exp\left\{\phi^{-1}[y\lambda - \zeta(\lambda)] + c(y, \phi)\right\}, \tag{1}$$

where λ and ϕ are termed "natural (canonical) parameter" and "dispersion parameter", respectively for unknown functions $\zeta(\cdot)$ and $c(\cdot,\cdot)$. As shown in Molenberghs and Verbeke (2005), the first two moments are functions of $\zeta(\cdot)$ as:

$$E(Y) = \mu = \zeta^{'}(\lambda), \tag{2}$$

and

$$Var(Y) = \sigma^2 = \phi\zeta^{''}(\lambda). \tag{3}$$

This implies that the mean and variance are related through $\sigma^2 = \phi\zeta''[\zeta^{'-1}(\mu)] = \phi\nu(\mu)$, with variance function $\nu(\cdot)$ describing the mean-variance relationship. Suppose that for the i^{th} subject $(i = 1, \ldots, n)$ a p-dimensional covariate vector $\mathbf{x}_i$ is available and that given $\mathbf{x}_i$, the response Y_i of that subject has the above exponential distribution with mean μ_i and that $\eta(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients. The first component of $\mathbf{x}_i$, x_{i1}, is usually equal to 1, representing the intercept. This defines a generalised linear model for the response, denoted as GLM. For some GLMs such as the binomial, Poisson and exponential distributions, the mean and variance parameters are forced to depend on a single parameter. However, in some applications, this assumption may be overly restrictive. A number of extensions to the Poisson model have been proposed by Hinde and Demétrio (1998); Breslow (1984); Lawless (1987) and Molenberghs and Verbeke (2005) that accommodate overdispersion, i.e. when the variance of the counts (much) exceeds their mean. Note that one way to deal with overdispersion is to allocate an overdispersion parameter $\phi \neq 1$ so that (3) produces $Var(Y) = \phi\nu(\mu)$. This leads to a quasi-likelihood approach, see Molenberghs et al. (2007). Here, we consider parametric generalisations of the Poisson model such as the Poisson mixed-model, Poisson-gamma mixed-model and their zero-inflated alternatives.

2.2 The generalised linear mixed model

The generalised linear mixed model extends the GLM by adding random effects, and thereby becomes a tool to analyse non-Gaussian repeated measurements, see e.g. Verbeke and Molenberghs (2000). Let y_{ij} be the j^{th} outcome $(j = 1, \ldots, m_i)$ measured on the i^{th} subject $(i = 1, \ldots, n)$, then a GLMM for y_{ij} is defined as a GLM conditional on random effects. More specifically, we assume that, in analogy with Section 2.1, conditional upon a $q-$dimensional random-effect $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, where $N_q(\cdot)$ is a q-variate variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)^{'}$, of dimension $q \times 1$, and $\mathbf{D}$ is a $q \times q$ positive-definite covariance matrix. The outcomes y_{ij} are distributed independently with densities of the form

$$f_i(y_{ij} \,|\, \mathbf{b}_i, \boldsymbol{\beta}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \zeta(\lambda_{ij})] + c(y_{ij}, \phi)\right\},$$

with

$$\eta[\zeta^{'}(\lambda_{ij})] = \eta(\mu_{ij}) = \eta[E(y_{ij} \,|\, \mathbf{b}_i, \boldsymbol{\lambda})] = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i,$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are $p-$dimensional and $q-$dimensional vectors of known covariates, respectively. As before, $\boldsymbol{\beta}$ is an unknown $p-$dimensional vector of fixed effects parameters, ϕ is a dispersion parameter and $\eta(\cdot)$ is a known link function. The distribution for the random effects $f(\mathbf{b}_i \,|\, D)$ is most often specified as $N_q(\mathbf{0}, \mathbf{D})$. Often, usually the Poisson distribution is taken as distribution for counts. With repeated measures, the Poisson mixed-effects model (PMM) in the context of a longitudinal study becomes

$$\begin{aligned} y_{ij}|\mathbf{b}_i &\sim Poi(\lambda_{ij} \,|\, \mathbf{b}_i),\ (i = 1, \ldots, n; j = 1, \ldots, m_i), \\ \log(\lambda_{ij}) &= \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i, \\ \mathbf{b}_i &\sim N_q(\mathbf{0}, \mathbf{D}). \end{aligned}$$

Here, we consider the GLMM for count data but also extensions in the longitudinal settings. These extensions will be discussed in Section 5.

3 Bayesian model selection criteria

The conditional version of these selection criteria is based on the conditional likelihood incorporating the random effects, i.e. they are based on the conditional likelihood $p(\mathbf{y} \,|\, \boldsymbol{\Theta}, \mathbf{b}) \equiv L(\boldsymbol{\Theta}, \mathbf{b} \,|\, \mathbf{y}) = \prod_i L(\boldsymbol{\Theta}, \mathbf{b}_i \,|\, \mathbf{y}_i)$, with $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ the total set of responses, $\boldsymbol{\Theta}$ the model parameters (fixed effects $\boldsymbol{\beta}$ and the variance parameters of the random effects, i.e. the elements of $\mathbf{D}$) and $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ the total set of random effects. In contrast, the marginal criteria are based on the marginal likelihood, which is simply the conditional likelihood integrated over the distribution of the random effects, i.e. the marginal likelihood is given by

$$L(\boldsymbol{\Theta} \,|\, \mathbf{y}) = \prod_{i=1}^{n} L(\boldsymbol{\Theta} \,|\, \mathbf{y}_i) = \prod_{i=1}^{n} \int L(\boldsymbol{\Theta}, \mathbf{b}_i \,|\, \mathbf{y}_i)\, p(\mathbf{b}_i \,|\, \boldsymbol{\Theta})\, d\mathbf{b}_i.$$

In many instances, integration over the distribution of the random effects requires numerical procedures, such as (non)-adaptive Gaussian quadrature methods.

Let $D(\boldsymbol{\psi})$ represent the deviance of the model evaluated in $\boldsymbol{\psi}$, i.e. $D(\boldsymbol{\psi}) = -2 \log p(\mathbf{y} \,|\, \boldsymbol{\psi}) + 2 \log f(\mathbf{y})$, where $f(\mathbf{y})$ represents the likelihood of a saturated model. The deviance information criterion is then defined as DIC $= D(\overline{\boldsymbol{\psi}}) + 2p_{DIC}$, where $D(\overline{\boldsymbol{\psi}})$ is the deviance (often) evaluated at the posterior mean. p_{DIC} is called the effective number of parameters of the model and is a contrast of the posterior mean of the deviance $\overline{D(\psi)}$ with the deviance at the posterior mean $D(\overline{\boldsymbol{\psi}})$ and is equal to $p_{DIC} = \overline{D(\boldsymbol{\psi})} - D(\overline{\boldsymbol{\psi}})$. Both DIC and p_{DIC} can be approximated from an MCMC run with a converged chain $\boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^K$. Namely, the deviance components are approximated as $\overline{D(\boldsymbol{\psi})} \approx \frac{1}{K} \sum_{k=1}^{K} D(\boldsymbol{\psi}^k)$ and $D(\overline{\boldsymbol{\psi}}) \approx D(\frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\psi}^k)$.

The conditional version of DIC (cDIC) is obtained by plugging the conditional deviance into the expression of DIC, and by taking the posterior mean of

$(\boldsymbol{\Theta}, \mathbf{b})$. The associated effective degrees of freedom is denoted then as p_{cDIC}. The marginal version of DIC (mDIC) is obtained by plugging in the marginal deviance into the expression of DIC together with the posterior mean of $\boldsymbol{\Theta}$ (which is the same as for the conditional likelihood). We denote the effective degrees of freedom now as p_{mDIC}. Note that the marginal deviance is the posterior mean of the log of the conditional likelihoods averaged over the distribution of the random effects, i.e.

$$\mathbf{E}_{\boldsymbol{\Theta}\,|\,\mathbf{y}}\left[-2\log p\left(\mathbf{y}_i\,|\,\boldsymbol{\Theta}\right)\right] = \sum_{i=1}^{n}\mathbf{E}_{\boldsymbol{\Theta}\,|\,\mathbf{y}}\left[-2\log \mathbf{E}_{\mathbf{b}_i\,|\,\boldsymbol{\Theta}}\big[p(\mathbf{y}_i\,|\,\boldsymbol{\Theta},\mathbf{b}_i)\big]\right].$$

Despite its popularity, DIC has suffered from some practical problems; see (Spiegelhalter et al., 2014) for more details. To accommodate DIC's setback, Watanabe (2010) has suggested the Widely Applicable Information Criterion abbreviated as WAIC. WAIC is an approximation to minus twice the expected log pointwise predictive density (elppd)

Hence, the elppd is given as

$$elppd = -2\sum_{i=1}^{n}\mathbf{E}_{\widetilde{\mathbf{y}}_i}\log\left[\mathbf{E}_{\boldsymbol{\Theta},\mathbf{b}\,|\,\mathbf{y}}p\left(\widetilde{\mathbf{y}}_i\,|\,\boldsymbol{\Theta}\right)\right].$$

Note that in practice we evaluate the criterion on the current data. An alternative would be to use a validation and test set, but that one would imply loosing an important part of the data to base the scientific conclusions on. When the responses y_{ij} are independent given the random effects (e.g. when there is no serial correlation), then the above expression can be written as:

$$elppd = -2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{E}_{\widetilde{y}_{ij}}\log\left[\mathbf{E}_{\boldsymbol{\Theta},\mathbf{b}\,|\,\mathbf{y}}p\left(\widetilde{y}_{ij}\,|\,\boldsymbol{\Theta},\mathbf{b}_i\right)\right]. \tag{4}$$

Based on a converged chain $\left\{\boldsymbol{\Theta}^1,\ldots,\boldsymbol{\Theta}^K,\mathbf{b}_1^1,\ldots,\mathbf{b}_1^K,\ldots,\mathbf{b}_n^1,\ldots,\mathbf{b}_n^K\right\}$ the conditional WAIC can be approximated as

$$\text{cWAIC} = -2\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right] + 2p_{\text{cWAIC}}, \tag{5}$$

with $p_{\text{cWAIC}} = 2\left(\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right] - \frac{1}{K}\sum_{k=1}^{K}\log\ p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right)$.
The WAIC of the marginal model, i.e. the marginal WAIC, is then approximated by

$$\text{mWAIC} = -2\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k)\right] + 2p_{\text{mWAIC}}, \tag{6}$$

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

Another criterion considered here is the pseudo-Bayes factor which is a version of the Bayes factor. Since the Bayes factor is based on the prior distribution of the model parameters, its computation becomes complicated with a vague prior parameter. Several "solutions" were proposed to solve this problem and some proposals imply applying the vague prior to a part of the data, and then use the resulting posterior as a prior for the calculation of the Bayes factor on the remaining data. The pseudo-Bayes factor deviates from this principle a bit by also involving cross-validation. Suppose we have two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with model parameters ψ_1 and ψ_2, respectively and data $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. The Bayes factor is based on the marginal likelihood, in the sense that the likelihood is marginalised over the model parameters' prior uncertainty. Namely, this marginal likelihood is given for model $\mathcal{M}$ and parameters ψ (leaving out the model subscript) by:

$$p(\mathbf{y} \,|\, \mathcal{M}) = \int \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \boldsymbol{\psi}, \mathcal{M})\, p(\boldsymbol{\psi})\, d\boldsymbol{\psi}. \tag{7}$$

However, (7) is not analytically available in general. Therefore, Geisser and Eddy (1979) suggested replacing (7) by the pseudo marginal likelihood (PML)

$$\widehat{p}(\mathbf{y} \,|\, \mathcal{M}) = \prod_{i=1}^{n} p(\mathbf{y}_i \,|\, \mathbf{y}_{-i}, \mathcal{M}), \tag{8}$$

where $p(\mathbf{y}_i \,|\, \mathbf{y}_{-i}, \mathcal{M})$ is called the ith conditional predictive ordinate (CPO_i) and is the predictive density calculated at the observed $\mathbf{y}_i$ given $\mathbf{y}_{-i}$, which is the set of all data except the ith observation. The pseudo-Bayes factor is then obtained by taking the ratio $\widehat{p}(\mathbf{y} \,|\, \mathcal{M}_1)/\widehat{p}(\mathbf{y} \,|\, \mathcal{M}_2)$ to evaluate the preference of model $\mathcal{M}_1$ over model $\mathcal{M}_2$. Low values of this ratio reflect preference of model $\mathcal{M}_2$ based on the current data. The conditional pseudo-Bayes factor (cPSBF) (given random effects) and the marginal pseudo-Bayes factor (mPSBF) (averaged over random effects) are based on (8) with corresponding conditional and the marginal likelihood plugged-in. In practice, one often evaluates the logarithm of expression (8), leading to the log pseudo marginal likelihood for model $\mathcal{M}_\ell$ equal to $\mathrm{LPML}_\ell = \sum_{i=1}^{n} \log(\mathrm{CPO}_{i,\ell})$ where

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

where $\boldsymbol{\Theta}_\ell^1, \ldots, \boldsymbol{\Theta}_\ell^k$ are the K draws from the posterior distribution of the model parameters for model $\mathcal{M}_\ell$.

We refer to Ariyo et al. (2020, 2019) for the detailed descriptions of the three considered model selection criteria.

4 Sampling methods for computing the marginal model selection criteria

The expression of the model selection criteria reveals that expected values over the distribution of the random effects need to be taken. In the simpler case of a linear mixed model, the computations are easy since the marginal LMM can be determined analytically, but this is not the case for a GLMM. For this reason, we explored the use of sampling methods to compute the model selection criteria for a GLMM. Here we combined the replication method, which is sampling from the prior of the random effects, with importance sampling to compute the marginal criteria. The former replaces the integral in $p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}) = \int p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \boldsymbol{\Theta})\, d\mathbf{b}_i = \mathbf{E}_{\mathbf{b}_i \,|\, \boldsymbol{\Theta}}[p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)]$ by sampling from the prior distribution of $\mathbf{b}_i$. We used vague priors throughout the document and assessed reliability of the MCMC samples based on the Monte Carlo errors. A large Monte Carlo error (larger than 5% of the posterior standard deviation for each parameter, see Koehler et al. (2009); Quintero and Lesaffre (2018) suggests the need of sampling for more iterations. With non-informative priors for the hyper-parameters of $\mathbf{b}_i$, the Monte Carlo errors reduced drastically (see also Merkle et al., 2018). This in turns ensures reliability of the information criteria. In this paper, we computed the marginal version of DIC, WAIC and PSBF based on these sampling techniques. An R function has been written and is available in https://github.com/OludareAriyo/BayesselectGLMM.

4.1 The replication method

The joint posterior $p(\boldsymbol{\Theta}, \mathbf{b} \,|\, \mathbf{y})$ can be approximated by making use of a MCMC sample $(\boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}^k)$, for $k = 1, \ldots, K$. $\boldsymbol{\Theta}^1, \ldots, \boldsymbol{\Theta}^K$ are the K draws from the posterior distribution $p(\boldsymbol{\Theta}, \mathbf{b}_{1:n} \,|\, \mathbf{y}_{1:n})$ as discussed in Section 3 (see supplementary documents for details). Since $p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}) = \int p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \boldsymbol{\Theta})\, d\mathbf{b}_i = \mathbf{E}_{\mathbf{b}_i \,|\, \boldsymbol{\Theta}}[p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)]$, the marginal criteria such as mDIC can be based on independent replicates $\widetilde{\mathbf{b}}_i^{\,k,l}$, $(l = 1, \ldots, L)$ from $p(\mathbf{b}_i \,|\, \boldsymbol{\Theta}^k)$ at each iteration k. To compute the plug-in deviance, we take replicates $\widetilde{\mathbf{b}}_i^m$ from $p(\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}})$ $(m = 1, \ldots, M)$ in order to approximate $\sum_{i=1}^n \log[p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}})] = \sum_{i=1}^n \log \mathbf{E}_{\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}}}[p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \mathbf{b}_i)]$. Thus, the components necessary to compute the marginal criterion mDIC are

$$\begin{aligned} \overline{D(\boldsymbol{\Theta})} &\approx -2 \sum_{i=1}^{n} \left(\frac{1}{K} \sum_{k=1}^{K} \log \left[\frac{1}{L} \sum_{l=1}^{L} p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}_i^{\,k,l}) \right] \right), \\ D(\overline{\boldsymbol{\Theta}}) &\approx -2 \sum_{i=1}^{n} \log \left[\frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \widetilde{\mathbf{b}}_i^m) \right]. \end{aligned} \tag{9}$$

The variability of (9) due to the replication method depends on several factors which include: (i) the number of observations in the sample, (ii) the variance of the latent variables induced by $p(\mathbf{b} \,|\, \boldsymbol{\Theta})$, and (iii) the posterior variance of the parameters. In Quintero and Lesaffre (2018), an expression of the variance

of mDIC is given. For a small value of Var(mDIC), the proposed estimator (9) provides a good approximation to mDIC. However, as pointed out in Quintero and Lesaffre (2018), this variance can be high for large sized clusters and when there are many clusters, which corresponds here to subjects with many repeated observations per subject or many subjects.

In the same spirit, the components necessary to compute marginal criterion mWAIC are

$$\begin{aligned}
\text{mlppd} &= \frac{1}{K}\sum_{i=1}^{n}\sum_{k=1}^{K}\log\left\{\frac{1}{L}\sum_{l=1}^{L}p\left(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\widetilde{\mathbf{b}}^{k,l}\right)\right\},\\
p_{\text{mWAIC}} &= 2\sum_{i=1}^{n}\left[\log\left\{\frac{1}{M}\sum_{m=1}^{M}p\left(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^m,\widetilde{\mathbf{b}}^m\right)\right\}-\frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{L}\sum_{l=1}^{L}\log p\left(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\widetilde{\mathbf{b}}^{k,l}\right)\right\}\right],
\end{aligned} \tag{10}$$

where mlppd is the log pointwise predictive density for the marginal model and p_{mWAIC} is the corresponding effective number of parameters to adjust for overfitting. The mPSBF consists of comparing the (marginal) log-pseudo likelihood (mLPML) for models M_1 and M_2, whereby mLPML is equal to $\sum_{i=1}^{n}\log(\text{mCPO}_{i,\ell})$ for model M_ℓ where

$$\text{mCPO}_{i,\ell}\approx\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\frac{1}{L}\sum_{l=1}^{L}p\left(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\widetilde{\mathbf{b}}^{k,l},\text{M}_\ell\right)}\right]^{-1}. \tag{11}$$

The replication method can be based on simple random sampling, but there is gain using instead importance sampling, see e.g. Tran et al. (2016) and Tokdar and Kass (2010) for an overview of the advantages of importance sampling over simple random sampling.

4.2 Importance sampling

Importance sampling consists of replacing an original integral over a distribution by an integral averaging of another easier-to-sample distribution, called the proposal density, and then replace the integral by sampling. Given that $p(\mathbf{y}_i\,|\,\overline{\boldsymbol{\Theta}})=\int g_i(\mathbf{b}_i)\,d\mathbf{b}_i$ with $g_i(\mathbf{b}_i)=p(\mathbf{y}_i\,|\,\overline{\boldsymbol{\Theta}},\mathbf{b}_i)\,p(\mathbf{b}_i\,|\,\overline{\boldsymbol{\Theta}})$ is replaced by $p(\mathbf{y}_i\,|\,\overline{\boldsymbol{\Theta}})=\int[g_i(\mathbf{b}_i)/q_i(\mathbf{b}_i)]\,q_i(\mathbf{b}_i)d\mathbf{b}_i$, with $q_i(\mathbf{b})$ an appropriate proposal density. Then $p(\mathbf{y}_i\,|\,\overline{\boldsymbol{\Theta}},\mathbf{b}_i)\,p(\mathbf{b}_i\,|\,\overline{\boldsymbol{\Theta}})$ is proportional to $p(\mathbf{b}_i\,|\,\mathbf{y}_i,\overline{\boldsymbol{\Theta}})$, the mean and the variance of $g_i(\mathbf{b}_i)$ can be estimated from an additional MCMC run fixing the parameters to $\boldsymbol{\Theta}=\overline{\boldsymbol{\Theta}}$.

We used this approach to evaluate $p(\mathbf{y}_i\,|\,\overline{\boldsymbol{\Theta}})$ for each observation unit to compute the plug-in deviance $D(\overline{\boldsymbol{\Theta}})$ and the mean deviance $\overline{D(\boldsymbol{\Theta})}$ as well as other components needed for the marginal criteria. As pointed out by Quintero and Lesaffre (2018), this posterior distribution is approximately normal for large sized observation units under regularity conditions, so it is adequate

to select a normal density for $q_i(\mathbf{b}_i)$ with the above mean and variance. This approach is based on independent draws from the proposal density $q_i(\mathbf{b}_i)$ which is easy to sample from. For small sized observation units, the function $g_i(\mathbf{b}_i)$ resembles the latent prior density, so it is appropriate to select $q_i(\mathbf{b}_i) = p(\mathbf{b}_i \mid \overline{\boldsymbol{\Theta}})$ (Quintero and Lesaffre, 2018).

Then, after sampling $\widetilde{\mathbf{b}}_i^m$ from $q_i(\mathbf{b})$ for $m = 1, \ldots, M$, different components for the marginal criteria are computed based on the plug-in deviance given by

$$\widehat{p}(\mathbf{y}_i \mid \overline{\boldsymbol{\Theta}}) = \frac{1}{M} \sum_{m=1}^{M} [p(\mathbf{y}_i \mid \overline{\boldsymbol{\Theta}}, \widetilde{\mathbf{b}}_i^m) p(\widetilde{\mathbf{b}}_i^m \mid \overline{\boldsymbol{\Theta}}) / q_i(\widetilde{\mathbf{b}}_i^m)],$$

and the mean deviance where $\boldsymbol{\Theta}^m$ is substituted with deviance for each iteration. Hence, the mean deviance is given by

$$\widehat{p}(\mathbf{y}_i \mid \boldsymbol{\Theta}^m) = \frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{L} \sum_{l=1}^{L} [p(\mathbf{y}_i \mid \boldsymbol{\Theta}^m, \widetilde{\mathbf{b}}_i^{m,l}) p(\widetilde{\mathbf{b}}_i^{m,l} \mid \boldsymbol{\Theta}^m) / q_i(\widetilde{\mathbf{b}}_i^{m,l})] \right].$$

Thus, to compute the marginal criteria components we use importance sampling based on MCMC for large-sized observation units, but for small-sized observation units, an independent sampling method can be used. This strategy for importance sampling simplifies and generalises the replication method in Chan and Grant (2016).

5 Extensions of the Poisson-mixed effects model

In this section, we illustrate the performance of the marginal and conditional model selection criteria on selecting the appropriate fixed effects. However, with count data there is always the possibility of overdispersion and occasionally of underdispersion. Overdispersion occurs when the data display more variability than is predicted by the assumed model. For counts, we usually start with a Poisson model that assumes that the mean and variance of the counts are equal. When the variance is larger (smaller) than the mean, we speak of overdispersion (underdispersion) compared to the Poisson model. Most often, counts encountered in medical data do not satisfy the Poisson assumption. However, ignoring over/underdispersion may influence the model estimates and therefore the (statistical) conclusions. Indeed, it is well-known that when overdispersion in the data is ignored, many of the regressors will indicate a wrong 'significance'. On the other hand, Fitzmaurice (1997) evaluated the performance of the classical frequentist model selection criteria AIC and BIC, but also of his proposed modified likelihood ratio statistics. The author observed that the considered selection criteria often prefer overdispersion models even when there is no overdispersion in the data set. Obviously, this can lead to a wrong interpretation of the model parameters. We refer to Lambert (1992) for more background on the issues of overdispersion and its impact on the conclusions of a statistical analysis. Therefore, we wished to evaluate the

performance of the above three model selection criteria when overdispersion is present or absent in the data set. We ignore here the case of underdispersion, since this occurs less frequently in practice, and thus focus on overdispersion. In order to detect such deviation from the Poisson model, we need to have statistical models for repeated count data that allow for overdispersion. Without clustering, some models have been suggested to model overdispersion. A popular choice is the negative binomial distribution, which arises as a continuous mixture of Poisson distributions with means that have a gamma distribution. If overdispersion is due to an excess of zeros, one could model the data with a zero-inflated Poisson distribution or a zero-inflated negative binomial distribution. Although, both are mixtures of the basic (Poisson/negative binomial) distribution with a degenerate distribution at zero. Also for longitudinal count data, models have been suggested that deal with overdispersion, see e.g. Booth et al. (2003); Aregay et al. (2013, 2015); Molenberghs et al. (2007). Here, we focus on the extensions suggested by Molenberghs et al. (2007, 2010), which we briefly describe in the sections below.

5.1 The Poisson-type models for count data with overdispersion

A natural extension of the random effects Poisson model is to make use of the generalisations suggested for a Poisson model (see Molenberghs et al. (2007)). That is, to allow for overdispersion by assuming a Poisson-gamma model or a zero-inflated Poisson/negative binomial model given the random effects. The first proposal was suggested in Molenberghs et al. (2007, 2010). More specifically, these authors suggest

$$\begin{aligned} y_{ij}|\,\mathbf{b}_i\,,\theta_{ij} &\sim Poi(\lambda_{ij}\,|\,\mathbf{b}_i),\ (i=1,\ldots,n; j=1,\ldots,m_i), \\ \lambda_{ij} &= \theta_{ij}\exp\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\,\mathbf{b}_i\right), \\ \mathbf{b}_i &\sim N_q(\mathbf{0},\mathbf{D}), \\ E(\boldsymbol{\theta}_i) &= E[(\theta_{i1},\ldots,\theta_{im_i})^T], \\ Var(\boldsymbol{\theta}_i) &= \Sigma_i, \end{aligned} \tag{12}$$

whereby θ_{ij} measures the overdispersion in the outcome for the i^{th} subject at the j^{th} occasion. When θ_{ij} has a Gamma(α_1,α_2) distribution, we call it a Poisson-gamma mixed effects model (PGMM). Alternatively, one could assume that the θ_{ij} has a lognormal distribution. In that case, it becomes the Poisson-lognormal mixed effects model (PLMM). Molenberghs et al. (2007) provide the expressions for the mean vector, the variance-covariance matrix and the joint marginal probability. Here we focus on the PGMM model.

5.2 Zero-inflated GLMM

Using the technique described in Sections 5.1, one could suggest a zero-inflated Poisson mixed effects model (ZIPMM) or a zero-inflated negative binomial

mixed-effects model (ZINBM). That is, given the random effects one could assume a zero-inflated Poisson/negative binomial. More specifically, the ZIPMM model for y_{ij} is given by:

$$\begin{aligned} &y_{ij}|\,\mathbf{b}_i \sim ZIP(p_{0,ij}, \lambda_{ij}\,|\,\mathbf{b}_i),\ (i=1,\ldots,n; j=1,\ldots,m_i) \\ &\text{with} \\ &y_{ij}|\,\mathbf{b}_i \sim \begin{cases} 0, & \text{with probability} \quad p_{0,ij} \\ Poi(\lambda_{ij}), & \text{with probability} \quad (1-p_{0,ij}), \end{cases} \end{aligned} \tag{13}$$

where $\lambda_{ij} = \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i)$. A ZIPMM will reflect the data accurately when overdispersion is caused by an excess of zeros (Adrion and Mansmann, 2012). The use of the ZIPMM is necessary when the nature of the source of zeros is not certain. However, overdispersion is attributed to factors beyond the inflation of zeros, a ZINBM is more appropriate (Yau et al., 2003). It is important to note that the rate of zero-inflation and the nature of the source of zeros may change over time, but such considerations will be ignored here.

5.3 Sampling methods for extended GLMM

With an extra random effect θ_{ij} in the model, the application of the desired sampling techniques remains the greatest priority. One approach is to apply the sampling techniques of Section 4 on both the θ_{ij}'s and $\mathbf{b}_i$'s jointly. Alternatively, one could first integrate θ_{ij}'s from the likelihood, and then apply the sampling techniques on $\mathbf{b}_i$. Given $\mathbf{b}_i$, the Poisson-gamma distribution averaged over θ_{ij} yields a conditional (on $\mathbf{b}_i$) negative binomial distribution. In Molenberghs et al. (2007) it is shown that

$$y_{ij} \mid \mathbf{b}_i \sim \mathrm{NB}(\alpha_1, \gamma_{ij}),$$

with $\gamma_{ij} = 1/(1+\lambda_{ij}\alpha_2)$, where α_1 and α_2 are the parameters of the gamma distribution and $\lambda_{ij} = \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i)$. Using this marginalised model (over θ_{ij}'s) allows us to compute the marginal model selection criteria with the above considered strategies since the only latent variables are the random effects.

However, in general, it is not possible to integrate out θ_{ij}'s from the likelihood analytically, as it is the case for the Poisson-gamma mixed effects model. Therefore, the θ_{ij}'s need to be integrated out simultaneously together with the random effects $\mathbf{b}_i$. This can be done by augmenting the set of latent variables which need to be integrated out ($\mathbf{b}_i$'s and θ_{ij}'s) and using the replication method or importance sampling as previously discussed in Section 4.

6 Application section

In this section, we evaluate the performance of the marginal criteria in longitudinal data with overdispersion. First, we illustrate how many replications

are needed to achieve convergence and reliability of the results. To achieve this, we performed a simple simulation study and give practical advice on the adequacy of the number of replications needed. Thereafter, we applied the methods in the analysis of the epilepsy data set.

6.1 Adequacy of the number of replications

From expressions (9), (10) and (11), it is clear that the number of subjects n in the data set impacts the variability of the estimators. A larger sample size leads to greater variability of $\overline{D(\boldsymbol{\Theta})}$, $D(\overline{\boldsymbol{\Theta}})$, lppd, p_{WAIC} and LPML since these estimators are the sums of the log-likelihoods pertaining to the observation units. In order to approximate well the true marginal model selection criteria, it is important to select L and M appropriately (not too small nor too high). For DIC, Quintero and Lesaffre (2018) suggested to take $L = 2M/\sqrt{K_{Eff}}$, where $K_{Eff} = K/(1+2\sum_{t=1}\rho_t)$ and $\rho_t = \text{Corr}(\sum_i \log \widehat{p}(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k), \sum_i \log \widehat{p}(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k + \mathbf{t}))$. Among others, these authors suggested to determine L and M such that the standard error for mDIC is smaller 0.5, such that the variability (measured by 95% CI) of mDIC can be expected to be smaller than 1. Here, we checked the adequacy of the choice of L and M for these selection criteria in a numerical exercise, see below. From this exercise, we tentatively conclude that when L and M are appropriate for DIC, they are likely to be appropriate for WAIC and LPML.

To illustrate the required number of replications, we performed a small simulation exercise. This is part of the simulation exercise described in more detail in Section 7.2. To this end, we have taken a Poisson mixed model. Namely, let y_{ij} be a count for the i^{th} subject ($i = 1, \ldots, n = 300$) at the j^{th} time point ($j = 1, \ldots, 5$) and b_{0i} the i^{th} random intercept with $b_{0i} \sim N(0, \sigma^2_{b_0})$. We allowed for time independent covariates for the i^{th} subject: age at baseline (age_i), baseline count (base_i), treatment (treat_i), interaction baseline count and treatment (basetreat_i) and the obvious time dependent covariate time (time_{ij}). That is, we assumed

$$y_{ij} \,|\, b_{0i} \sim \text{Poisson}(\lambda_{ij} \,|\, b_{0i}), \tag{14}$$

where $\lambda_{ij} = \beta_1 + \beta_2 trt_i + \beta_3 \log(base_i) + \beta_4 time_{ij} + \beta_5 \log(age_i) + \beta_6 trt_i \times \log(base_i) + b_{0i}$. The estimates of the model parameters are given in Section 7. It is suggested by Mason et al. (2012) to monitor the stability of the components of (9), (10) and (11) when increasing the number of replications. Figure 1 displays the marginal criteria components for the above Poisson model for increasing number of replications M. From this figure we can see that mDIC and mWAIC, and their components stabilise for $M = 8000$. Recall, that for mDIC we have also another basis to decide about its desired value, namely that the standard error of the estimated mDIC should be smaller than 0.5 (Quintero and Lesaffre, 2018). This is achieved for $M = 8000$ as then the standard error is 0.2. For mWAIC and mLMPL we judged the adequacy of M

purely graphically. Note that for mLMPL stability is already achieved with M around 7000.

Furthermore, we evaluated the dependence of the required M on the number of subjects and the number of observations/subject. For this, we have considered data sets with 10, 50 and 200 subjects combined with 4, 6 and 10 observations per subject. Each data set was generated according to the above Poisson mixed effects model. From basic principles, Quintero and Lesaffre (2018) concluded that the required M likely increases with increasing number of subjects and/or observations/subject. Note that, as before, $L = 2M/\sqrt{K_{Eff}}$. In Table 3, we show the model selection criteria when varying M from 5000 to 10 000. In contrast to the above conclusion, the suspected dependence does not show.

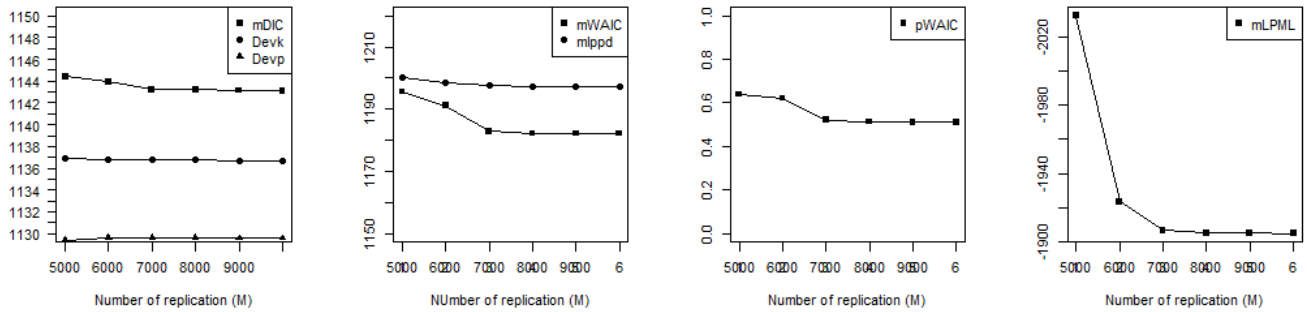

Fig. 1: Poisson model (14): Dependence of marginal model selection criteria on the number of replications M for: (a) mDIC, Devk=$\overline{D(\boldsymbol{\Theta})}$ and Devp=$D(\bar{\boldsymbol{\Theta}})$ as given in (9); (b) mWAIC and mlppd as given in (10): (c) p_{mWAIC} as given in (10): (d) mLMPL as given in (11)

6.2 Description of the epilepsy data set

We consider the analysis of data obtained from 89 epileptic patients that were randomised to a novel anti-epileptic drug (AED) in combination with one or two other AEDs (44 patients) or to placebo (45 patients) (Faught et al., 1996). A 12 weeks baseline period served as a stabilisation period. They were then measured on a weekly basis for 16 weeks, after which they were entered into a long-term open extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the last week. Booth et al. (2003) used this data set as an illustrating example when modelling longitudinal counts data with overdispersion. Others have also used this data set to illustrate their proposed statistical models, see e.g. Aregay et al. (2013); Rakhmawati et al. (2016); Faught et al. (1996); Molenberghs et al. (2007).
Figure 2 shows the individual curves and mean curves for both the treatment groups and this figure reveals substantial variability in counts between subjects. The graph also reveals the presence of extreme values. The presence of overdispersion in counts is seen in Table 1 where the sample mean, variance of the counts and the number of observations at each week for the treatment and placebo groups are shown. However, it is difficult to judge the presence of overdispersion towards the end of the study as there are fewer data. It is important to note here that overdispersion can be attributed to measured and unmeasured covariates. As such, overdispersion can also be checked by evaluating the Poisson residuals. However, in this paper, we checked overdispersion via model selection rather than through residuals.

Breslow and Clayton (1993) analysed the epilepsy data by considering the following covariates: logarithm of baseline seizure (base) count, treatment (trt), logarithm of age, visit, and the treatment by log(base) interaction. We fitted the model

$$\begin{aligned} y_{ij} \,|\, b_{0i} &\sim \text{Poisson}(\lambda_{ij} \,|\, b_{0i}), \\ \eta_{ij} = \log(\lambda_{ij}) &= \beta_1 + \beta_2 trt_i + \beta_3 \log(base_i) + \beta_4 visit_{ij} + \beta_5 \log(age_i) + \beta_6 trt_i \times \log(base_i) + b_{0i}, \end{aligned} \tag{15}$$

with $b_{0i} \sim N(0, \sigma^2_{b_0})$.

6.3 Analysis of the epilepsy data set

We fitted the Poisson mixed-effects model and extensions discussed in Section 5 to the data using *rjags*. An MCMC sample was drawn using Gibbs sampling in JAGS (Plummer, 2003). In this procedure, we sample from the conditional posterior for each parameter. For this case, JAGS used slice sampler for all parameters. Here, 60,000 iterations were sampled after discarding the initial 20,000 iterations as a burn-in. The thinning factor was set as 10. For all models considered, convergence was assessed using traceplots, the estimated potential

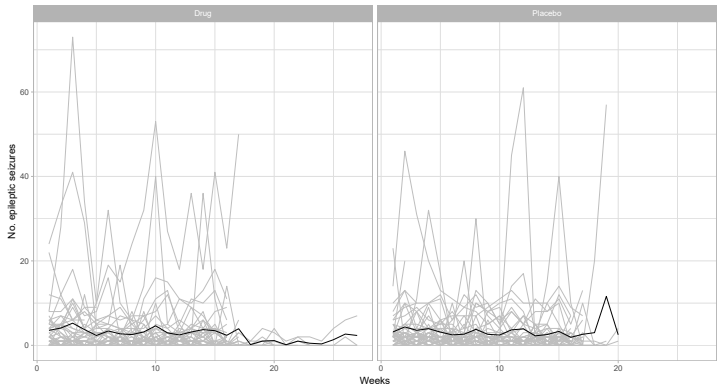

Fig. 2: Epilepsy data: Individual profiles for both treatment groups and mean curves (in bold).

scale reduction factor $\hat{R}$ and the Brooks, Gelman and Rubin's (BGR) diagnostic (Gelman and Rubin, 1992). Moreover, the estimated potential scale reduction factor $\hat{R}$ values for all the parameters were smaller than 1.1, which indicates convergence for all model parameters. The traceplots in Figure **S1** show a good mixing of the parameters. Additionally, the BGR plot (Figure **S2**) indicates convergence for all model parameters. Figure **S3** shows the running (ergodic) mean plots for β_1 and β_5, this indicates a stable behaviour for the parameters. The following vague priors were chosen, for the fixed effect parameters: $\beta_k \sim N(0, 100)\,(k = 0, \ldots, 6)$. For models with a random intercept and slope, we considered a separation prior for their covariance matrix i.e we have taken a half standard-Cauchy prior (Gelman, 2006) for the standard deviation of the random effects and Unif$(-1, 1)$ for correlation parameter. For the overdispersion parameter, we assumed $\theta_{ij} \sim \text{Gamma}(\alpha, 1/\alpha)\,(i = 1, \ldots, n; j = 1, \ldots, m_i)$ where $\alpha \sim \text{Unif}(0, 100)$. Finally, the zero-inflated probability $p_{0,ij} \sim \text{Beta}(0.5, 0.5)\,(i = 1, \ldots, n; j = 1, \ldots, m_i)$.
The time required to fit the models on a quad-core processor 3.0-GHz laptop was about 61 minutes. This shows that a quicker method is needed, we will explore this in future research. The results in Table 2 illustrates that all marginal criteria prefer the zero-inflated PGMM (ZIPGMM), which is in agreement with what was obtained by Warton (2005). For the conditional criteria the best two models (PGMM and ZIPGMM) are the same as for the marginal criteria, but they rank models PMM and ZIPMM in the opposite way compared to the marginal criteria. So, it seems that there is not much

difference between the solution offered by the conditional and the marginal criteria. The simulations in Section 7 check whether this is a general finding.

Table 1: Epilepsy data set: Sample mean (sample variance) and number of observations at selected time-points for each of the two treatment groups

Week	Mean (variance)	No of obs	Mean (variance)	No of obs
	Placebo		Treatment	
2	4.35(58.00)	45	4.09(45.24)	44
4	3.95(34.53)	42	3.72(46.24)	44
8	3.78(30.22)	41	2.55(17.43)	40
10	2.44(8.30)	41	4.63(109.37)	40
12	3.90(97.84)	40	2.95(27.49)	39
14	2.55(11.64)	40	3.71(43.31)	39
16	1.90(6.55)	40	2.39(22.63)	37
18	3.00(56.33)	7	0.18(0.16)	11
20	3.00(4.50)	2	1.13(2.41)	8
27	-	-	2.33(16.33)	3

Table 2: Epilepsy study: The value of both versions of the Bayesian selection criteria for each of the considered models for the epilepsy data sets: Poisson mixed effects model (PMM), Poisson-gamma mixed effects model (PGMM), Zero-inflated Poisson mixed-effects model (ZIPMM) and Zero-inflated -gamma mixed effects model (ZIPMM)

Criteria	PMM	PGMM	ZIPMM	ZIPGMM
cDIC	6045.68	4840.36	5331.78	4764.96
cWAIC	5966.37	4291.05	5215.23	4264.97
cLPML	-2132.48	-2607.61	-2145.52	-2983.19
mDIC	6203.09	6047.12	6383.50	6013.77
mWAIC	6213.70	6025.52	6472.96	6006.03
mLPML	-3016.79	-3049.28	-3085.34	-3087.93

7 Simulation study

Three simulation studies illustrate the conditional and marginal selection criteria' performance in identifying the true data-generating model. These simulations are based on equation (15) discussed above under varying conditions and settings. The simulation studies mimic some aspect of the data set described in Section 6.2 . Using the R procedure *glmer*, we obtained the maximum likelihood estimates : $\widehat{\beta}_1 = -3.96715$, $\widehat{\beta}_2 = -2.12053$, $\widehat{\beta}_3 = 0.94952$, $\widehat{\beta}_4 = -0.05872$, $\widehat{\beta}_5 = 0.89705$, $\widehat{\beta}_6 = 0.56223$, and $\widehat{\sigma}^2_{b_0} = 2.36045$. Unless specified, these values will be used as true parameters in the simulation studies described below. The same options in JAGS, but also the priors and so forth as described in Section 6.3 were used here. The simulation studies aim to confirm the superiority of the marginal criteria over the conditional criteria for

repeated count data as shown in the past for LMMs (Ariyo et al., 2020, 2019). Here, we also checked the performance of these criteria when overdispersion is of concern. More specifically, we were interested in exploring the performance of the conditional and marginal versions of the three model selection criteria:

- to select the correct data-generating model when: (i) the random effects structure is known and correctly specified, but the fixed effects part is unknown, (ii) the fixed effects structure is known and correctly specified, but the random effects structure is unknown;
- in the absence and presence of overdispersion;
- when the number of covariates is more than the number of subjects.

Also, we aimed to:

- evaluate the performance of the two sampling methods: replication method & importance sampling in calculating the marginal criteria;
- measure the impact of the number of subjects and number of observations per subject on the performance of the conditional and marginal criteria and the two sampling methods.

7.1 Simulation study 1

Here, we generated 300 data sets using the settings described above. We illustrate the performances of both versions of DIC, PSBF and WAIC by fitting three alternative models: (i) the true model ($\mathcal{M}_1$) give by equation (15), (ii) an under-specified model ($\mathcal{M}_0$) and (iii) over-specified model ($\mathcal{M}_2$). For the random effect scenario, $\mathcal{M}_0$ is given by (15) without $trt_i \times \log(base_i)$ interactions and $\mathcal{M}_2$ is given by (15) with additional covariate $trt_i \times \log(age_i)$. Likewise, for fixed effect scenario, $\mathcal{M}_0$ is given by (15) without the random intercept while $\mathcal{M}_2$ is given by (15) with the random intercept and slope. For these two scenarios, $\mathcal{M}_1$ is the true model.
Here, we illustrate the performance of the conditional and the marginal selection criteria in identifying the true model. Additionally, the effects of the number of subjects in the data were also evaluated in this simulation study under these two scenarios. Since the number of observations per subject may influence the performance of the replication method, we evaluated the performance of the Bayesian model selection criteria for a moderately large number of subjects ($n = 50$) and a varying number of observations per subject.
Table 4 presents the number of times the conditional and marginal criteria select the data-generating model for different number of observation per subject when $n = 50$. As seen in this table, the cluster size significantly influences the performance of both criteria regardless of the scenario. However, for the fixed effects scenario, the effect of the cluster on the performance of the marginal criteria is less pronounced. Additionally, the marginal criteria often select model $\mathcal{M}_1$ while the conditional criteria often select the wrong model regardless of the scenario used. Overall, the marginal criteria outperform the conditional criteria.

We also evaluated the performance of the Bayesian selection criteria under the scenarios discussed above with the assumption that the overdispersion in the data set is ignored. Here, we introduced an extra parameter θ_{ij} to simulate data with overdispersion. For the extra parameter, $\theta_{ij} \sim \text{Gamma}(\alpha, 1/\alpha)$ was assumed. High, moderate and low overdispersion level was induced by setting α to be $0.25, 1, 5$ respectively. We evaluated the model selection procedures when overdispersion is ignored in the data set. Here we used three wrong models: (ignoring overdispersion), models A, B and C which are the same with $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ described above. Where model B is the closest to the data-generating model without overdispersion, the number of times each model gave the smaller value for model selection criteria is presented in Table 5. As expected, both criteria performed poorly when overdispersion was ignored. As the overdispersion increases in the data set, the conditional criteria selected the model with extra fixed and random effects parameters while the marginal criteria selected the model without an extra parameter (model B), the closest model to the data-generating model that ignores overdispersion. These results are similar to the conclusion in Fitzmaurice (1997) that when overdispersion is ignored, model selection tends to select a model with too many parameters and can thus lead to over-interpretation of the parameters. In the Bayesian context, Millar (2009) advocated the use of the marginalized version of DIC and Bayes' factors as the use of the conditional DIC was misleading in the hierarchical modelling for overdispersed count data.

Table 3: Marginal selection criteria as a function of M for different numbers of subjects and observations/subject.

		Number of observations/subject								
		4			6			10		
# of subjects	M	mDIC	mWAIC	mLPML	mDIC	mWAIC	mLPML	mDIC	mWAIC	mLPML
	5000	114.562	116.762	-47.588	114.762	116.022	-47.359	114.902	116.212	-47.567
	6000	114.662	116.892	-47.365	114.342	116.002	-47.439	114.562	116.046	-47.234
10	7000	114.342	116.058	-47.288	113.392	115.122	-47.167	113.212	115.042	-47.162
	8000	113.462	115.478	-47.285	113.220	115.038	-47.169	113.108	115.002	-47.162
	9000	113.262	115.426	-47.285	113.210	115.026	-47.168	113.062	115.002	-47.072
	10000	113.062	115.426	-47.285	113.208	115.026	-47.167	113.062	115.002	-47.084
	5000	1146.404	1150.210	-537.070	1156.404	1158.210	-537.070	1146.404	1159.210	-547.072
	6000	1144.000	1150.110	-536.800	1154.100	1156.210	-536.800	1144.000	1157.210	-549.802
50	7000	1143.917	1149.210	-536.900	1153.017	1152.210	-526.723	1143.917	1152.210	-548.902
	8000	1143.111	1149.110	-536.900	1153.011	1152.315	-526.842	1143.111	1152.310	-540.102
	9000	1143.126	1149.120	-536.900	1153.026	1152.320	-526.812	1143.126	1152.320	-540.102
	10000	1143.117	1149.120	-536.900	1153.017	1152.320	-526.800	1143.117	1152.320	-540.202
	5000	1559.340	1660.010	-804.920	1564.012	1627.900	-804.200	1562.890	1616.794	-801.122
	6000	1560.210	1660.420	-804.890	1564.002	1625.010	-804.120	1562.320	1616.774	-801.102
200	7000	1562.120	1631.840	-804.100	1562.122	1621.840	-804.070	1562.120	1616.744	-801.072
	8000	1562.130	1621.880	-804.070	1560.592	1620.230	-804.070	1562.050	1616.634	-801.072
	9000	1562.130	1621.890	-804.070	1560.532	1620.210	-804.070	1562.043	1616.604	-801.072
	10000	1562.130	1621.900	-804.070	1560.532	1620.200	-804.070	1562.041	1616.604	-801.072

Table 4: Simulation 1: The number of times the selection criteria selects the data-generating model when varying the number of observations per subject for $n = 50$.

		Number of observations/subject								
		2			4			8		
Scenario	Criteria	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Fixed-Effects	cDIC	43	204	53	45	210	45	54	209	37
	cWAIC	41	203	56	78	168	54	81	170	49
	cLPML	50	202	48	57	204	39	39	217	44
	mDIC	49	248	3	42	253	5	39	252	9
	mWAIC	52	243	5	47	240	3	40	259	1
	mLPML	53	240	7	46	248	6	39	259	2
Random Effects										
	cDIC	41	52	207	38	47	215	27	49	224
	cWAIC	40	54	206	38	16	146	1	12	187
	cLPML	21	66	215	21	27	252	21	31	248
	mDIC	45	198	51	32	210	58	25	215	60
	mWAIC	46	196	58	34	208	53	26	217	57
	mLPML	45	198	57	37	211	52	36	214	50

Table 5: Simulation 1: The number of times three model specifications have the least value when overdispersion in the data set is ignored. For Low (L), Medium (M) and High (H) overdispersion.

		L			M			H		
Scenario	Criteria	A	B	C	A	B	C	A	B	C
Fixed-Effects	cDIC	33	82	185	31	72	197	29	70	205
	cWAIC	37	83	180	38	70	192	31	69	200
	cLPML	36	83	181	41	69	190	32	67	201
	mDIC	46	229	25	79	190	31	89	186	25
	mWAIC	49	221	30	73	199	28	87	180	33
	mLPML	40	230	30	77	192	31	84	183	33
Random Effects										
	cDIC	111	37	152	70	25	195	55	35	210
	cWAIC	109	39	152	67	37	196	67	35	198
	cLPML	114	46	140	58	42	200	67	36	197
	mDIC	54	167	79	91	141	68	114	133	53
	mWAIC	90	149	61	97	142	61	103	136	61
	mLPML	41	137	122	73	137	90	110	130	60

7.2 Simulation study 2

From each of the PMM, PGMM and ZIPMM described in Section 7.1 we generated 300 data sets. Data were simulated based on equation (15) together with parameter estimates and the extra parameters for PGMM are based on $y_{ij} \sim \text{Poi}(\lambda_{ij}\theta_{ij})$, with y_{ij} and λ_{ij} as defined above and $\theta_{ij} \sim \text{Gamma}(\alpha_j, 1/\alpha_j)$, where α_j takes the values 0.25 and $n = 50$. To ensure balanced data sets, we

simulated data sets with an equal number of observations per subject which is equal to 4. Using an appropriate number of replications as discussed in Section 6.1, we computed both the marginal and conditional criteria for all the three models via a self-written R code.

Different data settings were considered where data were generated from a particular model (true fit) and the model is evaluated with other alternative models. For instance, we generated data from the PMM (true fit) and considered PGMM, and ZIPMM as alternative models. Likewise, we generated data from the PGMM (true fit) and estimated with PMM and ZIPMM as alternative models and so on. The number of times when the selection criteria have a lower value for an alternative model against the true model was recorded and the percentage of misselection was calculated.
Figure 3 shows the histogram of the differences in selection criteria between the true model (PMM) and the alternative model fit (PGMM). The figure shows that the conditional criteria have a wider range of values as compared with the marginal criteria. Additionally, the marginal criteria have lower values for the PGMM model than the data-generating Poisson model in only about 2% of the times. This reflects the small penalty for the inclusion of an extra parameter in the PGMM model. Conversely, the fit of PGMM to PMM data has smaller values of the conditional criteria in 34.2%, 28.0% and 26.6% times respectively. That is, the conditional criteria select PGMM (model with extra parameter) which is the wrong model about 65.8%, 72.0% and 73.4 % respectively. This reflects the tendency of the conditional criteria to select a model with an extra parameter.

Similarly, Figure 4 shows that the conditional criteria prefer the ZIPMM model (model with extra parameter) often as against the Poisson data-generating model. In fact, the percentage of times the wrong model was selected increases from 34.2% to 52.0% for cDIC. Notwithstanding the narrow differences as shown in Figure 4. The marginal criteria, on the other hand, showed superior performance, preferring the ZIPMM model to the data-generating Poisson model less often.

When the PMM and ZIPMM models were fitted to the data generated from PGMM, the marginal criteria preferred more often the PGMM (the data-generating model) while the conditional criteria selected the PMM (see Figures 3 and 4). These results show the superior performance of marginal criteria in identifying the true data-generating model in count data sets. This is similar to the results earlier obtained for LMM (Ariyo et al., 2020, 2019; Ariyo and Adeleke, 2021), and for GLMM (Millar, 2009; Quintero and Lesaffre, 2018).

7.3 Simulation study 3

Here, we evaluated the performance of the two sampling techniques: replication and importance sampling. Following the simulation study described in Section 7.1, we generated 300 data sets from equation (15) under different number of subjects and observations/subject with $\alpha = 0.25$. The performance (in %) of

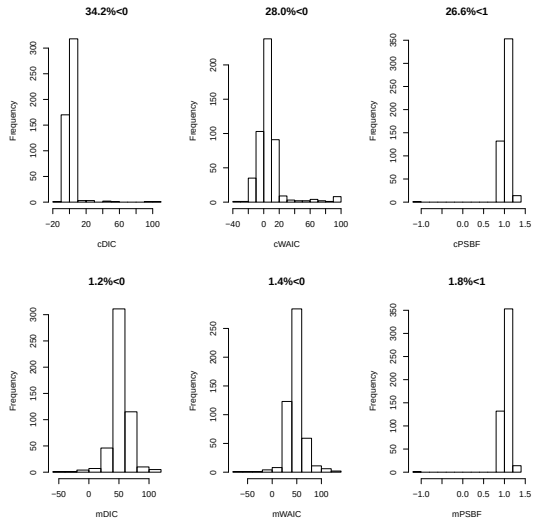

Fig. 3: Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the PGMM (fitted model) minus the selection criteria under the true model from 300 simulated PMM data sets. That is the percentages of times each criteria select wrong model.

the marginal criteria for both sampling methods in selecting the correct data-generating model is recorded in Table 6. It can been seem in the table that the performance of the importance sampling techniques fluctuate for a large subject (i.e $n \geq 200$) and a larger number of observations/subjects. This obviously shows that this sampling method needs improvement for $n \geq 200$. However, the advantage of importance sampling is shown when the number of subjects and/or observations/subject is large ($n < 200$) as the replication method becomes impracticable (i.e analysis taking too long time; i.e running several days) for a large number of subjects and/or observations/subject. This affirmed the results in Quintero and Lesaffre (2018).

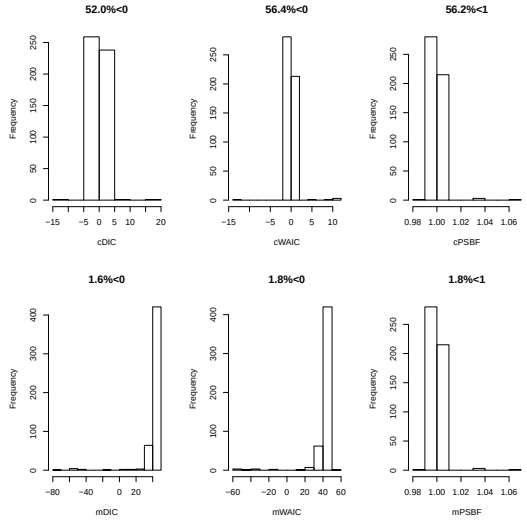

Fig. 4: Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the ZIPMM (fitted model) minus the selection criteria under the true model from 300 simulated PMM data sets.That is the percentages of times each criteria select wrong model.

Table 6: Simulation 3: The performance (in %) of the marginal criteria in selecting the data-generating model ($\mathcal{M}_1$) when varying the number of subjects and the number of observations/subject using two sampling methods: the replication method (Rep) and importance sampling (IS).

		5		10		30		60	
Criteria	# of subjects	Rep	IS	Rep	IS	Rep	IS	Rep	IS
mDIC	50	78.0	76.7	78.7	79.0	36.7	76.7	28.3	76.0
	100	80.7	82.0	80.0	80.0	31.7	78.7	24.7	74.7
	200	49.0	49.0	47.7	47.0	20.3	20.3	-	24.7
	500	14.3	45.0	1.0	45.3	-	45.7	-	46.3
mWAIC	50	77.7	77.7	78.0	80.7	35.3	75.3	26.7	71.0
	100	81.3	81.7	81.7	80.3	31.0	78.3	24.0	71.3
	200	50.7	50.3	46.3	48.3	19.7	20.7	-	29.7
	500	14.7	44.7	2.7	46.7	-	45.3	-	50.3
mLMPL	50	77.3	77.0	79.0	81.7	34.0	77.0	21.3	75.0
	100	80.0	81.7	82.3	79.0	36.3	78.7	20.7	70.7
	200	42.7	51.3	47.0	48.7	19.0	23.7	-	23.7
	500	15.0	45.7	2.3	46.3	-	45.0	-	42.3

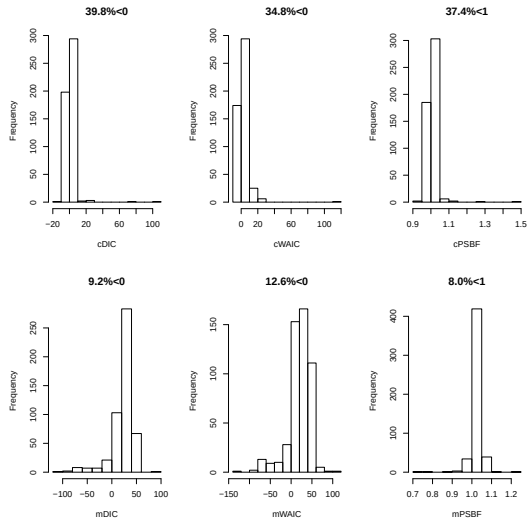

Fig. 5: Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the Poisson model minus the selection criteria under the true model from 300 simulated PGMM data sets.That is the percentages of times each criteria select wrong model.

8 Conclusion

In this paper, we aimed to promote the usage of the marginal criteria for model selection. As such, we showed the suboptimal performance of the conditional criteria when using Poisson mixed-effects models in case of overdispersion, too many zeros or both in longitudinal count data and when the number of repeated measurements is less than number of independent variables. It is shown here that these cases affect the performance of the information criteria. By evaluating the advantages/disadvantages of the two sampling techniques (i.e., replication and importance sampling), we showed that importance sampling is advantageous for calculating marginal criteria in cases with a large number of observations/subjects. An R function has been developed and is available that computes the marginal criteria for both the replication method and importance sampling.

The measures considered in this paper evaluated the selection of the best model among a number of proposed models. The best model is not necessarily a good model and needs to be examined for its goodness-of-fit to the data.Traditional steps in evaluating the goodness-of-fit of the model to the data is a residual analysis and the discovery of influential observations. But influential observations can affect the choice of the model when based on the considered model selection criteria. In De Oliveira et al. (2021) a new cross-validated Bayesian influence measure based on Bregman Divergence has been

suggested. The authors have applied their measure on a binary random-effects meta-analysis of 14 studies, aiming to find the influential study (result) based on the conditional version of the GLMM. This is the appropriate way to measure the impact of clusters on the estimated fixed effects. Likely, influential clusters on the conditional version of the GLMM will also have a high effect on the marginal version of the GLMM.

Acknowledgement(s)

The 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.

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 funded by Tertiary Education Trust Fund (TETFund)-AS&D grant of the Federal Government of Nigeria through Federal University of Agriculture, Abeokuta Nigeria.

References

Adrion, C. and Mansmann, U. (2012). Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: an example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology*, 12(1):137.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics & Data Analysis*, 57(1):233–245.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2015). Comparison of additive and multiplicative Bayesian models for longitudinal count data with overdispersion parameters: a simulation study. *Communications in Statistics-Simulation and Computation*, 44(2):454–473.

Ariyo, O., Lesaffre, E., Verbeke, G., and Quintero, A. (2019). Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors. *Communications in Statistics - Simulation and Computation*, 0(0):1–25.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., and Lesaffre, E. (2020). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, 47(5):890–913.

Ariyo, O. S. and Adeleke, M. A. (2021). Simultaneous Bayesian modelling of skew-normal longitudinal measurements with non-ignorable dropout. *Computational Statistics*, pages 1–23.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3(3):179–191.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):38–44.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673.

Chan, J. and Grant, A. (2014). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics and Data Analysis*, http://dx.doi.org/10.1016/j.csda.2014.07.018.

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

Chen, Q., Nian, H., Zhu, Y., Talbot, H. K., Griffin, M. R., and Harrell Jr, F. E. (2016). Too many covariates and too few cases?–a comparative study. *Statistics in Medicine*, 35(25):4546–4558.

Christensen, F. G. W. (2017). *New Approaches to Model Selection in Bayesian Mixed Modeling*. PhD thesis, UC Irvine.

De Oliveira, M. C., Castro, L. M., Dey, D. K., and Sinha, D. (2021). Bregman divergence to generalize Bayesian influence measures for data analysis.

Journal of Statistical Planning and Inference, 213:222–232.

Faught, E., Wilder, B., Ramsay, R., Reife, R., Kramer, L., Pledger, G., and Karim, R. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, 46(6):1684–1690.

Fitzmaurice, G. M. (1997). Model selection with overdispersed data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):81–91.

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

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

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

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Howe, E. J., Buckland, S. T., Després-Einspenner, M.-L., and Kühl, H. S. (2019). Model selection with overdispersed distance sampling data. *Methods in Ecology and Evolution*, 10(1):38–47.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

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

Mason, A., Richardson, S., and Best, N. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7(1):109–146.

McCullagh, P. (1989). *Generalized Linear Models*. Routledge.

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

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

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):375–385.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag; New York.

Molenberghs, G., Verbeke, G., and Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3):325–347.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

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

Rakhmawati, T. W., Molenberghs, G., Verbeke, G., and Faes, C. (2016). Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. *Biometrical Journal*, 58(6):1390–1408.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society - Series B*, 71:319–392.

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

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

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2016). Importance sampling squared for Bayesian inference in latent variable models. *arXiv preprint arXiv:1309.3339*.

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

van Smeden, M., de Groot, J. A., Moons, K. G., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1):163.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8):2455–2474. PMID: 29966490.

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

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The Official Journal of the International Environmetrics Society*, 16(3):275–289.

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

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897.

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(4):437–452.