Biometrics , 1–26 DOI: 10.1111/biom.13896
May 2023

Bayesian Model Selection for Generalized Linear Mixed Models

Shuangshuang Xu, Marco A. R. Ferreira*, Erica M. Porter, and Christopher T. Franck

Department of Statistics, Virginia Tech, Blacksburg, Virginia, 24060, U.S.A.

**email:* marf@vt.edu

Summary: We propose a Bayesian model selection approach for generalized linear mixed models (GLMMs). We consider covariance structures for the random effects that are widely used in areas such as longitudinal studies, genome-wide association studies, and spatial statistics. Since the random effects cannot be integrated out of GLMMs analytically, we approximate the integrated likelihood function using a pseudo likelihood approach. Our Bayesian approach assumes a flat prior for the fixed effects and includes both approximate reference prior and half-Cauchy prior choices for the variances of random effects. Since the flat prior on the fixed effects is improper, we develop a fractional Bayes factor approach to obtain posterior probabilities of the several competing models. Simulation studies with Poisson generalized linear mixed models with spatial random effects and overdispersion random effects show that our approach performs favorably when compared to widely used competing Bayesian methods including DIC and WAIC. We illustrate the usefulness and flexibility of our approach with three case studies including a Poisson longitudinal model, a Poisson spatial model, and a logistic mixed model. Our proposed approach is implemented in the R package GLMMselect (Xu et al., 2023) that is available on CRAN.

Key words: Approximate reference prior; Fractional Bayes factor; Generalized linear mixed model; Model selection; Pseudo likelihood method.

1. Introduction

Generalized linear mixed models (GLMMs) are widely used to model non-Gaussian data with dependent observations. This type of data is often found in many areas of application such as epidemiology (Meyer et al., 2017), meta-analysis of multiple clinical trials (Sauter and Held, 2015), survival analysis (Tawiah et al., 2020), and neuroimaging (Liu et al., 2016). Even though Bayesian estimation procedures for GLMMs are well established, there are just a handful of papers that address Bayesian model selection for GLMMs. Currently, most applied papers use the deviance information criterion (DIC) (Spiegelhalter et al., 2002) to perform Bayesian model selection for GLMMs (Nouvellet et al., 2021; Tredennick et al., 2021). Even though the DIC is widely applicable, we show in a simulation study that the DIC has some undesirable behaviors when applied to GLMMs. To provide more reliable results, here we develop a novel Bayesian model selection approach for simultanous selection of covariates and random effects for GLMMs.

Specifically, we focus on GLMMs where each random effect has a covariance matrix that is the product of an unknown variance component parameter and a known positive semi-definite symmetric matrix. The class of GLMMs we consider can be used for the analysis of spatial areal data (Clayton and Kaldor, 1987; Banerjee et al., 2014), genome-wide association studies (GWAS) (Williams et al., 2022), and longitudinal data (Breslow and Clayton, 1993; Xu et al., 2016). However, inference for GLMMs is difficult because the integrated likelihood function is not available in closed form. To deal with the issue of integration of random effects, we approximate the integrated likelihood function using a pseudo likelihood approach (Wolfinger and O'Connell, 1993) that leads to a Gaussian likelihood approximation. We then assign a flat prior for the vector of regression coefficients and an approximate reference prior (Ferreira et al., 2021) for the variance components of the GLMMs, which is inspired by the reference prior proposed by Keefe et al. (2019) for Gaussian data. In addition, we also consider a half-

Cauchy prior for the square root of variance components (Gelman, 2006; Polson and Scott, 2012). Because the prior on the vector of regression coefficients is improper, we develop a fractional Bayes factor (FBF) approach (O'Hagan, 1995). We note that Porter et al. (2023) have proposed FBF for Gaussian mixed models for the particular case of spatial areal data. In contrast, here we consider generalized linear mixed models. In addition, we consider not only spatial random effects but also many other types of random effects such as overdispersion random effects and longitudinal random effects. Because we use default priors combined with FBF, our proposed model selection approach is fully automatic, which obviates the need for subjective specification of hyperparameters and makes the method more accessible for practitioners. We call our two proposed model selection approaches the approximate reference method (ARM) and the half-Cauchy method (HCM).

To compare the performance of our methods ARM and HCM to the performance of the DIC, the Watanabe-Akaike information Criterion (WAIC) (Watanabe, 2010), and marginal likelihood computed by INLA under different parameter settings, we present a simulation study based on Poisson generalized linear mixed models with a spatial random effect and an overdispersion random effect. In this simulation study, we vary the sample size, coefficient of non-null covariates, level of spatial dependence, and overdispersion level. The simulation study shows that DIC and WAIC cannot reliably distinguish the random effect when there is another random effect. In contrast, our methods ARM and HCM perform well at detecting covariates and correct dependence structure. In particular, ARM and HCM always correctly detect the case of no random effects. Finally, while the performances of the DIC and WAIC do not improve much with large sample sizes, our proposed ARM and HCM have large improvement with increasing sample size. In addition, the simulation study shows that marginal likelihood computed by INLA has similar performance to our methods ARM and

HCM when selecting covariates. However, marginal likelihood computed by INLA does not perform well when selecting random effects.

Apart from the DIC, WAIC, and marginal likelihood, there are not many other Bayesian model selection approaches for GLMMs. One such approach proposed by Cai and Dunson (2006) for simultaneously selecting fixed and random effects in GLMMs assumes that the subject-specific random effects have a covariance matrix with all its elements being free parameters to be estimated. As a consequence, the method proposed by Cai and Dunson (2006) is only aplicable to problems with replications and can not be readily applied to problems where the vector of observations is a realization from a structured multivariate distribution such as GWAS data and spatial areal data. In contrast, because we assume that each random effect has a covariance matrix that is the product of an unknown variance component parameter with a known positive semi-definite covariance matrix, our methods ARM and HCM can be applied to longitudinal data, GWAS data, and spatial areal data.

The remainder of this paper is organized as follows. Section 2 describes the GLMMs that we consider. Section 3 outlines how the pseudo likelihood approach approximates GLMMs for non-Gaussian data by computing adjusted observations that are modeled using Gaussian LMMs. Section 4 introduces priors for model selection, the FBF approach for dealing with improper priors, and posterior computation. Section 5 presents the results of a simulation study. Section 6 illustrates our method with applications to two case studies. Section 7 concludes with a discussion and future directions.

The online supporting information contains details about the pseudo likelihood method (Web Appendix A), additional tables for the case studies (Web Appendix B), one additional case study (Web Appendix C), several additional simulation studies (Web Appendix D), and additional figures (Web Appendix E).

2. GLMMs

Consider a response vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ of n observations. Let $\boldsymbol{X}$ be an n by p matrix of explanatory variables and $\boldsymbol{\beta}$ be the corresponding p-dimensional vector of fixed effects. Let $\boldsymbol{Z}_j$ be an n by q_j design matrix and $\boldsymbol{\alpha}_j$ be the corresponding q_j-dimensional vector of random effects, $j = 1, \ldots, Q$. Let vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_{ij}$ be the ith rows of $\boldsymbol{X}$ and $\boldsymbol{Z}_j$, respectively. Conditional on linear predictors $\eta_1, \ldots, \eta_n$, the observations $y_1, \ldots, y_n$ are independent with probability density function belonging to the exponential family, that is $f(y_i|\eta_i) = \exp[y_i\eta_i - B_i(\eta_i) + C_i(y_i)]$, $i = 1, \ldots, n$, where the canonical parameter η_i is modeled as a linear function of fixed and random effects as $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \sum_j \boldsymbol{z}_{ij}^\top\boldsymbol{\alpha}_j$. Each observation y_i has mean $\mu_i = B_i'(\eta_i)$ and variance $v_i = B_i''(\eta_i)$. In addition, we assume that each vector of random effects $\boldsymbol{\alpha}_j$ has a multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\tau_j\boldsymbol{\Sigma}_j$, where the variance component parameter τ_j is unknown and $\boldsymbol{\Sigma}_j$ is a known symmetric positive semi-definite matrix. For example, if $\boldsymbol{\alpha}$ is a vector of overdispersion random effects then the corresponding matrix $\boldsymbol{\Sigma}$ is an identity matrix.

As another example, in the case of spatial areal data, we assume that $\boldsymbol{\alpha}$ is a vector of spatial random effects that follows a sum-zero constrained Gaussian Intrinsic Conditional Autoregressive Model (Keefe et al., 2018, 2019), that is,

$$\boldsymbol{\alpha}|\tau \;\sim\; N(\boldsymbol{0}, \tau\boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\Sigma}$ is a known positive semi-definite covariance matrix that depends on the neighborhood structure of the spatial subregions. Specifically, an adjacency matrix $\boldsymbol{W}$ is defined such that if subregion i and subregion j are adjacent, the entries in cells (i, j) and (j, i) are 1, otherwise 0. Let $\boldsymbol{D}_w$ be a diagonal matrix with each diagonal element equal to the summation of the corresponding row of $\boldsymbol{W}$. Then, the covariance matrix $\boldsymbol{\Sigma}$ is the Moore-Penrose inverse of $\boldsymbol{D}_w - \boldsymbol{W}$ (Keefe et al., 2018, 2019). We note that computations for this model may be performed using the precision matrix. In addition, we note that the knowledge about the covariance matrix $\boldsymbol{\Sigma}$ has allowed, for the case of Gaussian hierarchical models with ICAR random effects, the derivation of a reference prior for the parameters (Keefe et al., 2019), and formal Bayesian model selection (Porter et al., 2023).

3. Pseudo likelihood Function for GLMMs

A key step in Bayesian model selection is to integrate out random effects from the likelihood function. However, while for LMMs the random effects can be integrated out analytically, for GLMMs that is not possible. To overcome this difficulty, here we use a pseudo likelihood approach that approximates a GLMM for non-Gaussian data by computing adjusted observations that are modeled using an approximate Gaussian LMM.

Let $\boldsymbol{\alpha}$ represent all random effects and $\boldsymbol{\tau}$ represent all variance components. Then, the likelihood function with the relevant but intractable integral over random effects $\boldsymbol{\alpha}$ is

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\tau} | \boldsymbol{y}) &= \int p(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}|\boldsymbol{\tau}) \, d\boldsymbol{\alpha} \\
&= \int \prod_{i=1}^{N} \left[\exp\left\{ y_i \left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta} + \sum_j \boldsymbol{z}_{ij}^{\top}\boldsymbol{\alpha}_j \right) - B_i \left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta} + \sum_j \boldsymbol{z}_{ij}^{\top}\boldsymbol{\alpha}_j \right) + C_i(y_i) \right\} \right] \\
&\quad \prod_j \left[(2\pi)^{-\frac{q_j}{2}} |\tau_j \boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \exp\left\{ -\frac{\boldsymbol{\alpha}_j^{\top} \boldsymbol{\Sigma}_j^{-} \boldsymbol{\alpha}_j}{2\tau_j} \right\} \right] d\boldsymbol{\alpha}. \qquad (2)
\end{aligned}
$$

In Equation (2), the random effects $\boldsymbol{\alpha}$ cannot be integrated out analytically. Our method approximates the integral in Equation (2) with a Gaussian LMM via a pseudo likelihood approach. For a Gaussian LMM, the corresponding integral can be solved analytically, and then the likelihood function of parameters has an analytic expression.

The pseudo likelihood approach was first proposed by Wolfinger and O'Connell (1993). The pseudo likelihood approach is an iterative procedure that starts by writing the model as $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and $\boldsymbol{\epsilon}$ is a vector of errors with $cov(\boldsymbol{\epsilon}) = \boldsymbol{V} = diag(v_1, \ldots, v_n)$. Let $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{V}}$ be the current estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and $\boldsymbol{V}$. Here, $\widehat{\boldsymbol{\beta}}$ is initialized at the estimate from a GLM fit. Now, approximate μ_i with a first-order Taylor expansion around

$\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Rearrange all the terms in $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ such that the terms that depend on $\boldsymbol{y}$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\mu}}$ appear on the left side of the equation and the remaining terms appear on the right side of the equation. Multiply both sides by $\widehat{\boldsymbol{V}}^{-1}$. As a result, the left side of the equation will have $\boldsymbol{y}^{\star} = \widehat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \sum_j \boldsymbol{Z}_j \widehat{\boldsymbol{\alpha}}_j$. The vector $\boldsymbol{y}^{\star}$ is known as the vector of pseudo-observations or the vector of adjusted observations. Equating $\boldsymbol{y}^{\star}$ to the right side of the equation, we obtain the following model for the adjusted observations.

$$\begin{aligned} \boldsymbol{y}^{\star} &\approx \boldsymbol{X}\boldsymbol{\beta} + \sum_j \boldsymbol{Z}_j \boldsymbol{\alpha}_j + \widehat{\boldsymbol{V}}^{-1}\boldsymbol{\epsilon}, \\ \boldsymbol{\alpha}_j &\sim N(\boldsymbol{0}, \tau_j \boldsymbol{\Sigma}_j), \\ \boldsymbol{\epsilon} &\sim N(\boldsymbol{0}, \boldsymbol{V}). \end{aligned} \tag{3}$$

Thus, the pseudo likelihood approach assumes that $\boldsymbol{\epsilon}$ follows a Gaussian distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}$. Substituting $\boldsymbol{V}$ with $\widehat{\boldsymbol{V}}$ in Equation (3), $\boldsymbol{y}^{\star}$ can be approximately modeled with the LMM $\boldsymbol{y}^{\star} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sum_j \tau_j \boldsymbol{Z}_j \boldsymbol{\Sigma}_j \boldsymbol{Z}_j^{\top} + \widehat{\boldsymbol{V}}^{-1}\right)$. Therefore, we have the closed-form pseudo likelihood function

$$\begin{aligned} p(\boldsymbol{y}^{\star}|\boldsymbol{\beta}, \tau) &= (2\pi)^{-\frac{n}{2}} \left| \sum_j \tau_j \boldsymbol{Z}_j \boldsymbol{\Sigma}_j \boldsymbol{Z}_j^{\top} + \widehat{\boldsymbol{V}}^{-1} \right|^{-\frac{1}{2}} \\ &\exp\left\{ -\frac{1}{2}(\boldsymbol{y}^{\star} - \boldsymbol{X}\boldsymbol{\beta})^{\top} \left(\sum_j \tau_j \boldsymbol{Z}_j \boldsymbol{\Sigma}_j \boldsymbol{Z}_j^{\top} + \widehat{\boldsymbol{V}}^{-1} \right)^{-1} (\boldsymbol{y}^{\star} - \boldsymbol{X}\boldsymbol{\beta}) \right\}. \end{aligned} \tag{4}$$

Further details about the pseudo likelihood approach appear in Web Appendix A. To perform model selection, we first use the pseudo likelihood function in Equation (4) in an iterative manner to estimate the parameters and to obtain adjusted observations $\boldsymbol{y}^{\star}$. We then use these adjusted observations $\boldsymbol{y}^{\star}$ rather than the original observations $\boldsymbol{y}$ to perform model selection.

4. Model Selection

We perform model selection based on the pseudo likelihood function given in Equation (4). Similarly to Ten Eyck and Cavanaugh (2018), we use the same vector of adjusted observations to compare all candidate models' posterior probabilities. Specifically, we compute the vector of adjusted observations using the full model with all candidate regressors and all candidate random effects. In addition, consider the model space $\mathcal{M} = \{M_c,\ c = 1 \dots C\}$, with C possible models. We assume model M_c has K_c regressors, where $\boldsymbol{X}_c$ is the corresponding matrix of explanatory variables and $\boldsymbol{\beta}_c$ is the corresponding vector of coefficients. Further, model M_c has Q_c types of random effects. Let $\boldsymbol{\tau}_c = (\tau_{c,1}, \dots, \tau_{c,Q_c})$ be the vector of variance components of the Q_c types of random effects in the model M_c. The integrated likelihood based on the vector of adjusted observations $\boldsymbol{y}^\star$ is

$$p(\boldsymbol{y}^\star|M_c) \;=\; \int\int p(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c, \tag{5}$$

where $\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)$ is the prior distribution of $(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)$ conditional on model M_c. Let $\pi(M_c)$ be the prior probability of model M_c. Then, application of Bayes Theorem yields posterior model probabilities $P(M_c|\boldsymbol{y}^\star) \;=\; p(\boldsymbol{y}^\star|M_c)\pi(M_c)/\sum_{r=1}^{C} p(\boldsymbol{y}^\star|M_r)\pi(M_r) \;\propto\; p(\boldsymbol{y}^\star|M_c)\pi(M_c),$

In Section 4.1, we specify the priors for model parameters. In Section 4.2, we specify the priors on the model space. Section 4.3 discusses approximation of the integral in Equation (5). In Section 4.4, we propose an FBF approach (Porter et al., 2023) to perform model selection with improper priors.

4.1 *Priors for Model Parameters*

We consider the approximate reference prior proposed by Ferreira et al. (2021) in the context of LMMs for $\boldsymbol{\beta}$ and the reciprocal of τ, which is based on the reference prior proposed by Keefe et al. (2019). In what follows, we consider the implied reference prior for τ obtained by transformation of variables. For simple notation, let M without subscript represent a general model, $\boldsymbol{\beta}$ represent the corresponding vector of regressor coefficients, and τ represent the variance component. In the reference prior (Keefe et al., 2019), all the parameters are independent. The vector of regression coefficients $\boldsymbol{\beta}$ is assigned a uniform prior on $\mathcal{R}^p$. In addition, as τ goes to infinity the reference prior $\pi(\tau)$ is proportional to τ^{-2}. Further, as τ goes to 0 the reference prior is proportional to a constant. Based on the tail behavior of the reference prior for τ, Ferreira et al. (2021) proposed the approximate reference prior $\pi(\tau) \propto (1 + \frac{\tau}{a_\tau})^{-2}$, where a_τ is a hyperparameter. We set a_τ equal to 2. The choice of $a_\tau = 2$ is equivalent to the choice made by Ferreira et al. (2021) for Gaussian data. In addition, our simulation study shows that this choice also works well for GLMMs. Hence, for $\boldsymbol{\beta}$ we use the flat prior $\pi(\boldsymbol{\beta}|M) \propto 1$, and for τ we use the approximate reference prior

$$\pi_1(\tau|M) \ = \ \frac{1}{2(\tau/2+1)^2}, \quad \tau \geqslant 0. \tag{6}$$

This approximate reference prior is related to the half-Cauchy prior $\pi(\tau) \propto \frac{1}{\tau^2+1}$, which has the same tail behavior. Gelman (2006) proposed a half-Cauchy prior, however, for the standard deviation of random effects in a two-level Gaussian model. Assuming a half-Cauchy prior for the square root of the variance component parameter τ implies for τ the prior density $\pi_2(\tau) \propto \tau^{-\frac{1}{2}}(\tau+1)^{-1}$ (Polson and Scott, 2012). Thus, $\pi_2(\tau) = O(\tau^{-\frac{1}{2}})$ for $\tau \to 0$ and $\pi_2(\tau) = O(\tau^{-\frac{3}{2}})$ for $\tau \to \infty$. Hence, the half-Cauchy prior for $\sqrt{\tau}$ has more mass near zero and more mass for large values of τ than the approximate reference prior for τ given in Equation (6). Here, we consider two variants of our pseudo-likelihood-based method: ARM, which uses the approximate reference prior given in Equation (6); and HCM, which uses the half-Cauchy prior for $\sqrt{\tau}$. We compare our methods ARM and HCM to the DIC and WAIC in the simulation studies presented in Section 5.

4.2 *Priors on the Model Space*

Let K denote the number of candidate covariates and Q denote the number of candidate random effects types. For example, in an application where we may have spatial random effects and/or overdispersion random effects, $Q = 2$. In addition, let K_c denote the number of covariates in Model M_c. For fixed effects, we use priors from Scott and Berger (2010), which automatically correct for multiplicity. Specifically, the prior probability for model M_c with K_c covariates is $P(M_c \text{ with } K_c \text{ covariates}) = 1/\left[(K+1)\binom{K}{K_c}\right]$. With respect to random effects, there are 2^Q possibilities for inclusion and exclusion of random effects. Assuming that each random effect has 0.5 prior inclusion probability, the prior probability for Model M_c with Q_c types of random effects is $P(M_c \text{ with } Q_c \text{ types of random effects}) = 1/2^Q$. Because usually in practice the number of candidate random effects types Q is small, a discrete uniform prior for the inclusion of random effects is reasonable. Assuming *a priori* independence of inclusion of fixed effects and random effects, the prior probability for model M_c is $P(M_c) = 1/\left[2^Q(K+1)\binom{K}{K_c}\right]$.

4.3 *Integrated Likelihood Methods*

After the priors for parameters have been defined, the integrated likelihood given in Equation (5) based on the adjusted observations $\boldsymbol{y}^\star$ becomes

$$\begin{aligned} p(\boldsymbol{y}^\star|M_c) &= \int\int p(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\ &\propto \int\int \exp\left\{-\frac{1}{2}(\boldsymbol{y}^\star - \boldsymbol{X}_c\boldsymbol{\beta}_c)^\top \left(\sum_j^{Q_c}(\tau_{cj}\boldsymbol{Z}_{cj}\boldsymbol{\Sigma}_{cj}\boldsymbol{Z}_{cj}^\top) + \widehat{\boldsymbol{V}}^{-1}\right)^{-1}(\boldsymbol{y}^\star - \boldsymbol{X}_c\boldsymbol{\beta}_c)\right\} \\ &\quad \left|\sum_j^{Q_c}(\tau_{cj}\boldsymbol{Z}_{cj}\boldsymbol{\Sigma}_{cj}\boldsymbol{Z}_{cj}^\top) + \widehat{\boldsymbol{V}}^{-1}\right|^{-\frac{1}{2}} \pi(\boldsymbol{\tau}_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c. \end{aligned}$$

The vector of regression coefficients $\boldsymbol{\beta}_c$ can be integrated out analytically. After integrating out $\boldsymbol{\beta}_c$, we can write the integrated likelihood as

$$\begin{aligned} p(\boldsymbol{y}^\star|M_c) &= \int p(\boldsymbol{y}^\star, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\tau}_c \\ &\propto \int \exp\left[\frac{1}{2}\boldsymbol{y}^{\star\top}\left\{\boldsymbol{H}_c^{-1}\boldsymbol{X}_c(\boldsymbol{X}_c^\top\boldsymbol{H}_c^{-1}\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^\top\boldsymbol{H}_c^{-1} - \boldsymbol{H}_c^{-1}\right\}\boldsymbol{y}^\star\right] \\ &\quad \left|\boldsymbol{H}_c^{-1}(\boldsymbol{X}_c^\top\boldsymbol{H}_c^{-1}\boldsymbol{X}_c)^{-1}\right|^{\frac{1}{2}} \pi(\boldsymbol{\tau}_c)\, d\boldsymbol{\tau}_c, \end{aligned} \tag{7}$$

where $\boldsymbol{H}_c = \sum_j^{Q_c}(\tau_{cj}\boldsymbol{Z}_{cj}\boldsymbol{\Sigma}_{cj}\boldsymbol{Z}_{cj}^\top) + \widehat{\boldsymbol{V}}^{-1}$. Note that the vector of variance components $\boldsymbol{\tau}_c$ cannot be integrated out analytically. To compute the integral in Equation (7), we first perform a logarithm transformation on $\boldsymbol{\tau}_c$. Let $\boldsymbol{\delta}_c = \log(\boldsymbol{\tau}_c)$ be the vector obtained by applying the logarithm transformation to each element of $\boldsymbol{\tau}_c$. Then, we integrate out $\boldsymbol{\delta}_c$ using a Laplace approximation to obtain

$$\begin{aligned}\int p(\boldsymbol{y}^\star, \boldsymbol{\tau}_c | M_c)\, d\boldsymbol{\tau}_c &= \int p(\boldsymbol{y}^\star, \exp(\boldsymbol{\delta}_c)|M_c) \exp(\boldsymbol{\delta}_c)\, d\boldsymbol{\delta}_c \\ &\approx (2\pi)^{\frac{Q_c}{2}} \left| q''(\widehat{\boldsymbol{\delta}}_c) \right|^{-\frac{1}{2}} \exp\left\{ -q(\widehat{\boldsymbol{\delta}}_c) \right\}, \qquad (8)\end{aligned}$$

where $q(\boldsymbol{\delta}_c) = -\frac{1}{2}\boldsymbol{y}^{\star\top} \left[\boldsymbol{H}_c^{-1}\boldsymbol{X}_c(\boldsymbol{X}_c^\top \boldsymbol{H}_c^{-1}\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^\top \boldsymbol{H}_c^{-1} - \boldsymbol{H}_c^{-1} \right] \boldsymbol{y}^\star - \frac{1}{2}\log |\boldsymbol{H}_c^{-1}(\boldsymbol{X}_c^\top \boldsymbol{H}_c^{-1}\boldsymbol{X}_c)^{-1}|$ $-\log \pi(\exp(\boldsymbol{\delta}_c)) + \boldsymbol{\delta}_c$, $\widehat{\boldsymbol{\delta}}_c$ is the point that minimizes $q(\boldsymbol{\delta}_c)$, and $q''(\boldsymbol{\delta}_c)$ is the Hessian matrix.

4.4 *Fractional Bayes Factors*

In order to obtain the posterior model probabilities of interest, we use a fractional Bayes factor (FBF) approach. The FBF is a modification of the Bayes factor that allows for improper priors on parameters (O'Hagan, 1995).

To define the usual Bayes factor, let the baseline model M_l be the model with the largest integrated likelihood in the model space. Then, the Bayes factor BF_{cl} of model M_c versus the baseline model M_l is defined as the ratio of their integrated likelihoods $BF_{cl} = p(\boldsymbol{y}^\star|M_c)/p(\boldsymbol{y}^\star|M_l)$. Hence, we can compute the posterior probability of model M_c as proportional to its prior probability times its Bayes factor versus the baseline model, that is $P(M_c|\boldsymbol{y}^\star) \propto P(M_c)p(\boldsymbol{y}^\star|M_c)/p(\boldsymbol{y}^\star|M_l) \propto BF_{cl}P(M_c)$.

Note that the prior on the regression coefficients $\pi(\boldsymbol{\beta}_c|M_c) \propto d$ is improper, where d is an arbitrary constant. Thus, the Bayes factor computed with the integrated likelihood from Equations (7) and (8) is only defined up to an unspecified constant of proportionality and cannot be used to compare models directly.

To solve this problem, we use the fractional Bayes factor (FBF, O'Hagan (1995)) to approximate the Bayes factor. Porter et al. (2023) developed the fractional Bayes factor method for Gaussian hierarchical models with ICAR random effects. We use the FBF approach to train the improper prior so that we can compute a meaningful Bayes factor. By training the improper prior, we mean using Bayes Theorem to combine the improper prior with a fraction of the likelihood to obtain a proper distribution (O'Hagan, 1995; Porter et al., 2023). We can then use this latter distribution as a trained prior to compute a meaningful Bayes factor. Specifically, here we train the prior with a fraction b of the likelihood function. The trained prior density for model M_c is obtained by Bayes Theorem as $\pi^b(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c) = p^b(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)/\int p^b(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c$. The integrated likelihood is then computed as an integral of the product of the likelihood function raised to $1-b$ and the trained prior. Following O'Hagan (1995), the resulting integrated likelihood of model M_c, called the fractional integrated likelihood, is equal to

$$\begin{aligned} q_c(b, \boldsymbol{y}^\star) &= \int p^{1-b}(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi^b(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\ &= \int p^{1-b}(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\frac{p^b(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)}{\int p^b(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c \\ &= \frac{\int p(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}{\int p^b(\boldsymbol{y}^\star|\boldsymbol{\beta}_c, \boldsymbol{\tau}_c)\pi(\boldsymbol{\beta}_c, \boldsymbol{\tau}_c|M_c)\, d\boldsymbol{\beta}_c d\boldsymbol{\tau}_c}. \end{aligned} \tag{9}$$

The size of the training fraction b should be chosen carefully. When b is too small, the denominator in Equation (9) may diverge. If b is too large, a substantial part of the integrated likelihood is used to train the prior on the parameters, and then the remaining information in the integrated likelihood used to update the prior model probabilities will lead to less distinctive posterior model probabilities. Here, we consider a training fraction size equal to $b = m/n$, where m is the equivalent training size. To guide the choice of m in our considered GLMM context, we use the fact that for LMMs with the reference prior proposed by Keefe et al. (2019) the minimal value of m that guarantees propriety of the fractional integrated likelihood is $p+1$ (Porter et al., 2023). In particular, in all the GLMM applications we present in Section 6, the training fraction $b = (p+1)/n$ yields well-defined Bayes factors.

Then, the FBF of model M_c versus model M_l is defined as $BF_{cl}^b = \frac{q_c(b, \boldsymbol{y}^\star)}{q_l(b, \boldsymbol{y}^\star)}$. Next, we compute the posterior probability of model M_c as proportional to the FBF, BF_{cl}^b, times the prior probability of model M_c, that is $P^b(M_c|\boldsymbol{y}) = BF_{cl}^b \times P(M_c)/\left[\sum_{k=1}^C BF_{kl}^b \times P(M_k)\right]$.

5. Simulation Study

To investigate the performance of our proposed model selection methods ARM and HCM when compared to the widely used DIC, WAIC and marginal likelihood computed by INLA, we perform a simulation study for different combinations of parameter settings. Here we present results for Poisson GLMMs. In the Web Appendix C we present results for Bernoulli GLMMs. For each combination of parameter settings, we generate 100 datasets. We simulate samples on regular square grids and consider three sample sizes, $n = 100$, 400, and 900. Each sample may have spatial dependence based on a first-order neighborhood structure modeled with a vector of spatial random effects $\boldsymbol{\alpha}_1$ following the ICAR distribution given in Equation (1). For the variance component τ_1 of the spatial random effects, we consider values 0, 0.03, 0.05, 0.1, or 1, where $\tau_1 = 0$ implies no spatial dependence. We also consider the possibility of overdispersion random effect $\boldsymbol{\alpha}_2$ in the model. We set the variance component τ_2 of the overdispersion random effect to 0, 0.05, 0.1, 0.5, or 1, where $\tau_2 = 0$ implies no overdispersion. We consider 4 candidate covariates x_{1i}, x_{2i}, x_{3i} and x_{4i} sampled from a standard normal distribution. We assume that $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, 0, 0)^\top$, thus the last two covariates are not in the true model. Here β_0 is the intercept, with values equal to 1, 2, or 4. We let $\beta_1 = \beta_2$ with values 0, 0.1, 0.2, 0.3, 0.5, or 1. When β_1 and β_2 are both equal to 0, there is no covariate in the true model. Conditionally independent Poisson observations y_i are generated with the GLMM $y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$, $i = 1 \dots n$, with $\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \alpha_{1i} + \alpha_{2i}$, spatial random effects $\boldsymbol{\alpha}_1 \sim N(\mathbf{0}, \tau_1 \boldsymbol{\Sigma})$, and overdispersion random effects $\boldsymbol{\alpha}_2 \sim N(\mathbf{0}, \tau_2 \boldsymbol{I})$.

For each parameter setting, there are $C = 64$ candidate models in total. Specifically, there are 2^4 possible combinations of fixed effects. In addition, there are 4 possible combinations of random effects types, one with both spatial random effects and overdispersion random effects, one with only spatial random effects, one with only overdispersion random effects, and one without any random effects. We calculate posterior model probabilities for all 64 models, and we compute posterior inclusion probabilities for each of the 4 covariates, for the spatial random effect, and for the overdispersion random effect.

We compare our model selection methods ARM and HCM to the DIC, the WAIC and marginal likelihood computed by the R package INLA (Rue et al., 2009). For the ARM and HCM, we decide to include a component in the selected model if the posterior inclusion probability of that component is larger or equal to 0.5, that is, if that component is in the median probability model (Barbieri and Berger, 2004). For the criteria computed by INLA, we select the model with the lowest DIC and WAIC values, and the highest marginal likelihood, respectively. For the three criteria computed by INLA, we consider the INLA default prior specification as well as our proposed AR prior and HC prior.

[Figure 1 about here.]

Because currently the most widely used criteria for Bayesian selection of GLMMs are the DIC and WAIC computed with INLA default priors, here we compare these criteria with our ARM and HCM. We present a comparison of our methods ARM and HCM to DIC and WAIC computed using our AR and HC priors in Section D4 of Web Appendix D. The conclusions are similar to those for DIC and WAIC computed with default INLA priors presented here. Figure 1 presents the probability of each competing method selecting the correct covariate structure as a function of the value of their regression coefficients $\beta_1 = \beta_2$. Here, there are spatial random effects with $\tau_1 = 0.05$ and overdispersion random effects with $\tau_2 = 0.05$. Three sample sizes are considered: $n = 100, 400, 900$. Two values for the intercept are considered: $\beta_0 = 1$ and 4. Figure 1 shows that the ARM and HCM perform much better than the DIC and the WAIC computed with INLA's default priors . For example, in the most challenging case considered with $n = 100$ and $\beta_0 = 1$, the ARM and HCM have a higher probability than the DIC and WAIC of selecting the correct covariate structure when their regression coefficients β_1 and β_2 are zero. In addition, as the value of $\beta_1 = \beta_2$ increases, the probability of the ARM and HCM to correctly select the true non-null covariates x_1 and x_2 increases more quickly than that of the DIC and the WAIC. Finally, the probability of ARM and HCM to correctly select the two non-null regressors increases much closer to one than those of the DIC and WAIC as the sample size increases and as the intercept value increases. As the sample size increases, the probability of ARM and HCM detecting covariates with small coefficients increases substantially. For example, the left panels of Figure 1 show that when the intercept is equal to 1, the probabilities of our proposed methods choosing the correct covariates structure when the coefficient is equal to 0.1 are about 10%, 60%, and 90% for sample sizes 100, 400, and 900, respectively.

[Figure 2 about here.]

Figure 2 investigates the impact of different magnitudes of the variance components on the probability of selecting the correct covariate structure as a function of the value of the regression coefficients $\beta_1 = \beta_2$. Panels (a) and (b) of Figure 2 present settings with small ($\tau_1 = 0.01$ and $\tau_2 = 0$) and large ($\tau_1 = 1$ and $\tau_2 = 1$) variance components, respectively. In both panels, the sample size is $n = 400$ and the intercept is $\beta_0 = 1$. In the small variance components setting, ARM and HCM perform comparably to the DIC and WAIC for small values of $\beta_1 = \beta_2$, but our methods ARM and HCM greatly outperform the DIC and WAIC for moderate to large values of $\beta_1 = \beta_2$. Meanwhile, in the more challenging large variance components setting, when $\beta_1 = \beta_2 = 0$, our ARM and HCM correctly select the model with no regressor in the model for 100% of the simulated datasets samples. In contrast, when $\beta_1 = \beta_2 = 0$, the DIC and WAIC select the wrong covariate structure for 20% of the simulated datasets, respectively. Finally, as the magnitude of $\beta_1 = \beta_2$ increases, in comparison to the DIC and WAIC, ARM and HCM achieve much higher probabilities of selecting the correct model.

[Figure 3 about here.]

Figure 3 presents the probability of selecting correct spatial random effects structure as a function of the value of the variance component for the spatial random effects. Results are shown for sample sizes $n = 100$, 400 and 900, and variance of overdispersion random effects $\tau_2 = 0$ and 0.1. Figure 3 shows that the DIC and WAIC have low probability of selecting the model with no spatial random effects when the true model does not have spatial random effects; In addition, this performance does not improve much as the sample size increases from 400 to 900. In contrast, our methods ARM and HCM have large probabilities of selecting the correct spatial random effects structure when the true model does not have spatial random effects, and have large probabilities of selecting spatial random effects when the variance component for the spatial random effects is large. Finally, the performance of ARM and HCM at correctly detecting spatial dependence greatly improves as the sample size increases.

ARM, HCM, DIC and WAIC's performance when selecting overdispersion random effects is similar to selecting spatial random effects. Web Figure S1 in the supporting information presents the probability of selecting correct overdispersion structure as a function of the value of the variance for overdispersion. Web Figure S1 shows that the DIC and WAIC have low probability of selecting a model with no overdispersion random effects even when overdispersion is not present in the true model, and this undesirable performance does not improve much when the sample size increases. In contrast, our methods ARM and HCM have large probabilities of selecting correct overdispersion structure when overdispersion is not present in the true model, and have large probabilities of selecting overdispersion when the proportion of variance due to overdispersion is large. Finally, the performance of ARM and HCM at correctly detecting overdispersion greatly improves as the sample size increases.

Web Figures S12, S13, and S14 present a comparison of the performance of INLA marginal likelihood with our ARM and HCM methods. Web Figure S12 shows that INLA marginal likelihood with INLA's default priors is worse than our methods at selecting covariates when coefficients of covariates are small. INLA marginal likelihood with INLA's default prior or INLA marginal likelihood with our proposed priors are better than our methods ARM and HCM when the regression coefficient is large. For spatial random effects inclusion, Web Figure S13 shows that INLA marginal likelihood with any of the considered priors has difficulties to detect spatial random effects. For overdispersion random effects, Web Figure S14 shows that when there is no spatial random effects in the model, INLA marginal likelihood can correctly select overdispersion random effects. However, when there are spatial random effects in the model, marginal likelihood computed by INLA cannot correctly select overdispersion random effects. In summary, INLA marginal likelihood with our proposed priors works well for selection of regressors but does not work well for the selection of random effects. Meanwhile, our ARM and HCM methods work well in both cases.

6. Case Studies

6.1 *Longitudinal Epilepsy Seizure Data*

We analyze a dataset on epilepsy seizures previously analyzed by Thall and Vail (1990), Breslow and Clayton (1993), and others. The data were collected in four biweekly visits of 59 epileptics during a clinical trial to evaluate the effectiveness of a drug to control seizures (Leppik et al., 1987). The response variable is the number of seizures y_{ij} for patient i on visit j. The most general model we consider is $y_{ij}|\mu_{ij} \overset{ind}{\sim} \text{Poisson}(\mu_{ij})$, with $\log(\mu_{ij}) = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}+\alpha_{1i}+z_j\alpha_{2i}+\alpha_{3ij}$, $\boldsymbol{\alpha}_1 \sim N(\boldsymbol{0}, \tau_1 I_{59})$, $\boldsymbol{\alpha}_2 \sim N(\boldsymbol{0}, \tau_2 I_{59})$, and $\boldsymbol{\alpha}_3 \sim N(\boldsymbol{0}, \tau_3 I_{236})$, $i = 1 \ldots 59$ and $j = 1 \ldots 4$, where $\boldsymbol{x}_{ij}$ denotes a 6-dimensional vector with a one for intercept and 5 covariates. The 59 subjects were randomly assigned to a new drug or a placebo. The first covariate is the treatment indicator (Trt), where Trt=1 indicates that the patient received the treatment and Trt=0 indicates that the patient received the placebo. The second covariate is the baseline level of seizures (Base), equal to the logarithm of the average number of epileptic seizures per two weeks recorded in the 8-week period before treatment. The third covariate is the interaction term of Base and Trt. The fourth covariate is the logarithm of age (Age). And the fifth covariate is the visit number (Visit), with the 4 visits coded as -3, -1, 1 and 3. Breslow and Clayton (1993) mentioned that preliminary analysis indicated that the counts were substantially lower during the fourth visit. Thus, they also define a binary variable V4, such that V4=1 indicates the fourth visit and V4=0 indicates the other visits. In the model above, $\boldsymbol{\beta}$ is the vector of regression coefficients, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1\ 59})$ is the vector of patient-specific random effects, $\boldsymbol{\alpha}_2 = (\alpha_{21}, \ldots, \alpha_{2\ 59})$ is the vector of patient-specific random effects for the slope of the variable Visit with $\boldsymbol{z} = (-0.3,\ -0.1,\ 0.1,\ 0.3)$, and $\boldsymbol{\alpha}_3 = (\alpha_{311}, \ldots, \alpha_{3\ 59\ 1}, \alpha_{312}, \ldots, \alpha_{3\ 59\ 2}, \ldots, \alpha_{314}, \ldots, \alpha_{3\ 59\ 4})$ is the vector of overdispersion random effects.

The covariates Trt, Base, Age and Visit can be included in the model independently. However, the interaction term between Trt and Base is only included when both Trt and Base are in the model. Thus, there are 20 possible combinations of covariates. For the dependence structure, we follow the four cases considered by Breslow and Clayton (1993): no random effects in the model; only patient-specific random effects $\boldsymbol{\alpha}_1$; $\boldsymbol{\alpha}_1$ and overdispersion random effects $\boldsymbol{\alpha}_3$; $\boldsymbol{\alpha}_1$ and patient-specific random effects for the slope of the variable Visit $\boldsymbol{\alpha}_2$. Finally, we assume that the vectors of random effects $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are independent. Therefore, the model space has 80 models, composed by 20 combinations of covariates and 4 possible settings of random effects.

[Table 1 about here.]

Table 1 presents the posterior inclusion probabilities of the fixed and random effects. Both the ARM and the HCM indicate that the baseline level of seizures (Base) should be included in the model. However, the posterior inclusion probabilities do not provide support for any of the other covariates. Further, both ARM and HCM strongly indicate that $\boldsymbol{\alpha}_2$, the patient-specific random effects for the slope of the variable Visit should not be included in the model. Finally, both ARM and HCM strongly indicate the need to include the patient-specific random effect $\boldsymbol{\alpha}_1$ and overdispersion random effect $\boldsymbol{\alpha}_3$.

Web Table S1 in the supporting information presents a summary of the model selection results for the epilepsy data by comparing methods ARM, HCM, DIC and WAIC. A check mark appears next to the effects (rows) selected by each method (column). In addition, Web Table S1 presents the selection of fixed effects and variance components based on estimates and standard errors reported by Breslow and Clayton (1993) for two models fitted with PQL, which we denote by PQL1 and PQL2. Web Table S2 presents estimates and standard errors for the parameters based on the full model. Model PQL1 includes random effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ while Model PQL2 includes random effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$. Interestingly, while the original PQL method cannot choose between Model PQL1 or Model PQL2, our ARM and HCM clearly show that the data support exclusion of random effect $\boldsymbol{\alpha}_2$ and inclusion of random effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$. Further, the DIC and WAIC agree with the ARM and HCM and also choose random effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$. Furthermore, in terms of fixed effects the DIC and WAIC are the least parsimonious, choosing Base, Trt and Trt×Base, while PQL chooses Base and Trt. Finally, the ARM and HCM are the most parsimonious and choose only the Base covariate.

In addition to selecting more parsimonious models, our ARM and HCM provide more definitive support for the inclusion or exclusion of each effect in the form of Bayesian posterior probabilities. For example, the posterior inclusion probabilities of the patient-level random effects $\boldsymbol{\alpha}_1$, overdispersion random effects $\boldsymbol{\alpha}_3$, and the covariate Base are all equal to one. Further, there is a lot less support for the covariate V4, which has posterior inclusion probability of 0.12 by the ARM and 0.11 by the HCM. Furthermore, both ARM and HCM provide posterior inclusion probability equal to zero for the interaction between Trt and

Base. Finally, the simulation study presented in Section 5 shows that we can rely on the uncertainty quantification provided by the ARM and HCM.

6.2 *Spatial Lip Cancer Data*

In this section, we present an analysis of the Scottish lip cancer dataset previously analyzed by Clayton and Kaldor (1987), Breslow and Clayton (1993), Ferreira and De Oliveira (2007), among many others. This dataset provides the number of male lip cancer cases in the 56 counties of Scotland during the period 1975-1980, as well as the percentage of the work force employed in agriculture, fishing, or forestry (AFF) as a covariate. The most general model we consider is $y_i|\mu_i \overset{ind}{\sim} \text{Poisson}(\mu_i)$, $\log(\mu_i) = \log(n_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \alpha_{1i} + \alpha_{2i}$, $\boldsymbol{\alpha}_1 \sim N(\boldsymbol{0}, \tau_1 \boldsymbol{\Sigma})$, and $\boldsymbol{\alpha}_2 \sim N(\boldsymbol{0}, \tau_2 \boldsymbol{I}_{56})$, $i = 1 \ldots 56$, where n_i is the expected number of lip cancer cases in the i^{th} county, calculated based on the age distributions by counties. In this analysis, the n_i's are assumed to be known constants. In addition, the vector $\boldsymbol{x}_i$ is a two-dimensional vector with one as the first element and AFF for the i^{th} county as the second element. Further, $\boldsymbol{\alpha}_1$ is a vector of spatial random effects following a sum-zero constrained Gaussian Intrinsic Conditional Autoregressive model (Keefe et al., 2018) and modeled by Equation (1). Finally, $\boldsymbol{\alpha}_2$ is a vector of overdispersion random effects.

There are two possible combinations for the fixed effects: with or without the covariate AFF. For the random effects, we follow the options considered by Breslow and Clayton (1993). When $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are in the model at the same time, the PQL estimate of the overdispersion variance τ_2 is 0. Thus, we consider models with only three random effects combinations: spatial random effects $\boldsymbol{\alpha}_1$; overdispersion random effects $\boldsymbol{\alpha}_2$; and no random effects.

[Table 2 about here.]

Table 2 presents the posterior inclusion probabilities for the fixed and random effects. Both the ARM and HCM select the model with the covariate AFF and spatial random effect $\boldsymbol{\alpha}_1$.

Web Table S3 in the supporting information presents the DIC and WAIC for the 6 models considered. In contrast to the results of the ARM and HCM, DIC and WAIC select the model without the covariate AFF but with spatial random effect $\boldsymbol{\alpha}_1$. Web Table S4 in the supporting information summarizes model selection results for the ARM, HCM, DIC, WAIC, as well as the selection of model components based on PQL methods reported by Breslow and Clayton (1993) for two models: PQL1 includes $\boldsymbol{\alpha}_1$ and PQL2 includes $\boldsymbol{\alpha}_2$. Results from PQL for the AFF regressor agree with the results from the HCM and ARM. An advantage of the HCM and ARM over PQL is that they clearly indicate that the model should include a spatial random effect and not include overdispersion.

7. Discussion

We have proposed a novel Bayesian method for model selection for GLMMs. Our approach is based on a pseudo likelihood approximation of GLMMs by LMMs leading to a closed form solution for integrating out the random effects. We consider two priors for the model parameters. First, we use an approximate reference prior that is uniform for the fixed effects and has the tail behavior of the half-Cauchy prior for the variance parameters. Second, while keeping the improper flat prior for the fixed effects, we consider the half-Cauchy prior for the square root of the variance parameters (Gelman, 2006; Polson and Scott, 2012). Finally, to deal with the prior impropriety we have developed a fractional Bayes factor approach.

The simulation study has shown that our proposed methods ARM and HCM perform well for correctly selecting both covariates and dependence structure. ARM and HCM assign high posterior inclusion probability to covariates with large coefficients and also high posterior inclusion probability to random effects with large variance components. In particular, ARM and HCM are better than DIC and WAIC at correctly selecting covariates. In cases where random effects have large variances, the ability of DIC and WAIC to correctly select covariates is tremendously reduced. In contrast, ARM and HCM do not suffer as badly when selecting covariates in the presence of random effects with large variances. In addition, DIC and WAIC have high false positive rates and often select null fixed and random effects. In contrast, ARM and HCM correctly assign low posterior inclusion probability to null covariates and to null random effects. We also compared our methods with marginal likelihood computed by INLA. Our results show that when we use INLA with our priors instead of the default INLA priors, the marginal likelihood computed by INLA and the marginal likelihood computed by our pseudo likelihood approach work similarly for the selection of regression coefficients. However, the marginal likelihood computed by INLA does not work well for the selection of spatial random effects and overdispersion random effects. Therefore, it seems that our pseudo likelihood approximation works better than the INLA approximation to the marginal likelihood for the selection of random effects.

We illustrate the application of our proposed methods ARM and HCM with three case studies. In the first case study, we consider epilepsy seizures as a type of longitudinal count data. ARM and HCM are more parsimonious, selecting baseline covariate, patient-level random effects and overdispersion random effects. DIC and WAIC select two more covariates: treatment and interaction term between baseline and treatment. In the second case study, we study Scottish lip cancer data as a type of spatial count data. Our methods ARM and HCM select spatial dependence and covariate AFF, whereas DIC and WAIC select the model without covariate AFF but include spatial random effects. In the third case study, presented in Web Appendix C, we look at binary salamander mating data. For fixed effects, our methods ARM and HCM select WSF and WSF×WSM, whereas DIC and WAIC select all three covariates. For random effects, our two methods ARM and HCM have totally different results than DIC and WAIC: while DIC and WAIC select male random effect, our methods ARM and HCM select female random effect. Given the results from the simulation study, we recommend the models selected by ARM and HCM.

There are many potential avenues for future research. One possible future research topic is the use of Bayesian model averaging for computing credible intervals for regression coefficients. This can be facilitated by the fact that our methods provide posterior probabilities for different models. Another promising research direction is the use of nonlocal priors for the fixed effects. Finally, another possible research topic is model selection for GLMMs when the number of possible regressors is much larger than the sample size. We are currently working on the latter two research topics and will report the results in a future manuscript.

Acknowledgements

The work of Ferreira and Xu was supported in part by National Science Foundation awards DMS 1853549 and DMS 2054173. Computations for this manuscript have been performed on supercomputers of Advanced Research Computing at Virginia Tech. The authors would like to thank the associate editor and an anonymous referee for their insightful comments that helped substantially improve this article.

Data Availability Statement

The datasets analyzed in this paper are available in the R package mdhglm (Lee et al., 2018).

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall / CRC, Boca Raton, 2nd edition.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. The Annals of Statistics **32,** 870 – 897.

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

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

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics pages 671–681.

Ferreira, M. A. R. and De Oliveira, V. (2007). Bayesian reference analysis for Gaussian Markov random fields. Journal of Multivariate Analysis **98,** 789–812.

Ferreira, M. A. R., Porter, E. M., and Franck, C. T. (2021). Fast and scalable computations for Gaussian hierarchical models with intrinsic conditional autoregressive spatial random effects. Computational Statistics and Data Analysis **162,** 107264.

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

Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2018). On the formal specification of sum-zero constrained intrinsic conditional autoregressive models. Spatial Statistics **24,** 54–65.

Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2019). Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors. Bayesian Analysis **14,** 181 – 209.

Lee, Y., Molas, M., and Noh, M. (2018). The mdhglm package. https://CRAN.R-project.org/package=mdhglm.

Leppik, I., Dreifuss, F., Porter, R., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stackman, J., Graves, N., et al. (1987). A controlled study of progabide in partial seizures: methodology and results. Neurology **37,** 963–963.

Liu, Z., Berrocal, V. J., Bartsch, A. J., and Johnson, T. D. (2016). Pre-surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. Bayesian Analysis **11,** 599–625.

Meyer, S., Held, L., and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software **77,**.

Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. V., et al. (2021). Reduction in mobility and COVID-19 transmission. Nature Communications **12,** 1–9.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society: Series B (Methodological) **57,** 99–118.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. Bayesian Analysis **7,** 887–902.

Porter, E. M., Franck, C. T., and Ferreira, M. A. R. (2023). Objective Bayesian model selection for spatial hierarchical models with intrinsic conditional autoregressive priors. Bayesian Analysis to appear.

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 (Statistical Methodology) **71,** 319–392.

Sauter, R. and Held, L. (2015). Network meta-analysis with integrated nested Laplace approximations. Biometrical Journal **57,** 1038–1050.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics pages 2587–2619.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64,** 583–639.

Tawiah, R., McLachlan, G. J., and Ng, S. K. (2020). A bivariate joint frailty model with mixture framework for survival analysis of recurrent events with dependent censoring and cure fraction. Biometrics **76,** 753–766.

Ten Eyck, P. and Cavanaugh, J. E. (2018). An alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. Sankhya B **80,** 98–122.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. Biometrics pages 657–671.

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology **102,** e03336.

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.

Williams, J., Ferreira, M. A. R., and Ji, T. (2022). BICOSS: Bayesian iterative conditional stochastic search for GWAS. BMC Bioinformatics **23,** 475.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation **48,** 233–243.

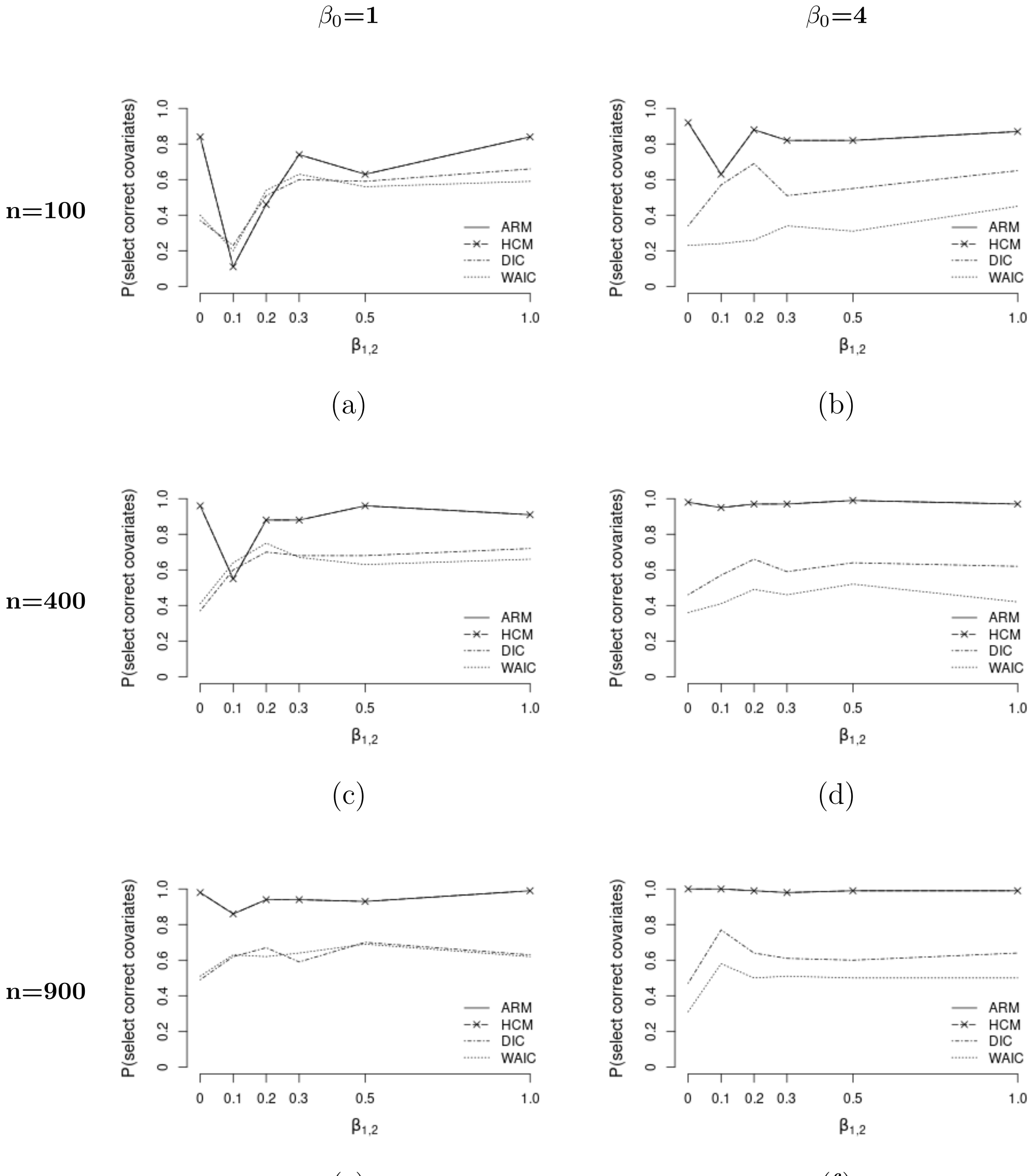
Xu, D., Chatterjee, A., and Daniels, M. (2016). A note on posterior predictive checks to assess model fit for incomplete data. Statistics in Medicine **35,** 5029–5039.

Xu, S., Ferreira, M. A. R., Porter, E. M., and Franck, C. T. (2023). The GLMMselect package. https://CRAN.R-project.org/package=GLMMselect. [accessed 20-April-2023].

Supporting information

Web Appendices, Tables and Figures referenced in Sections 1, 3, 5, and 6 are available with this paper at the Biometrics website on Wiley Online Library. The R package GLMMselect available at https://CRAN.R-project.org/package=GLMMselect implements our ARM and HCM methods. In addition, the source code of the R package GLMMselect and a vignette html file that analyzes the lip cancer dataset are available with this paper at the Biometrics website on Wiley Online Library.

Received September 2022.



(a)

(b)

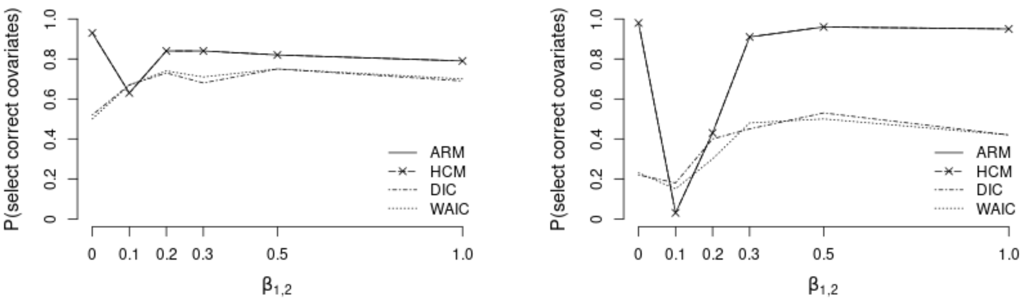
(c)

(d)

(e)

(f)

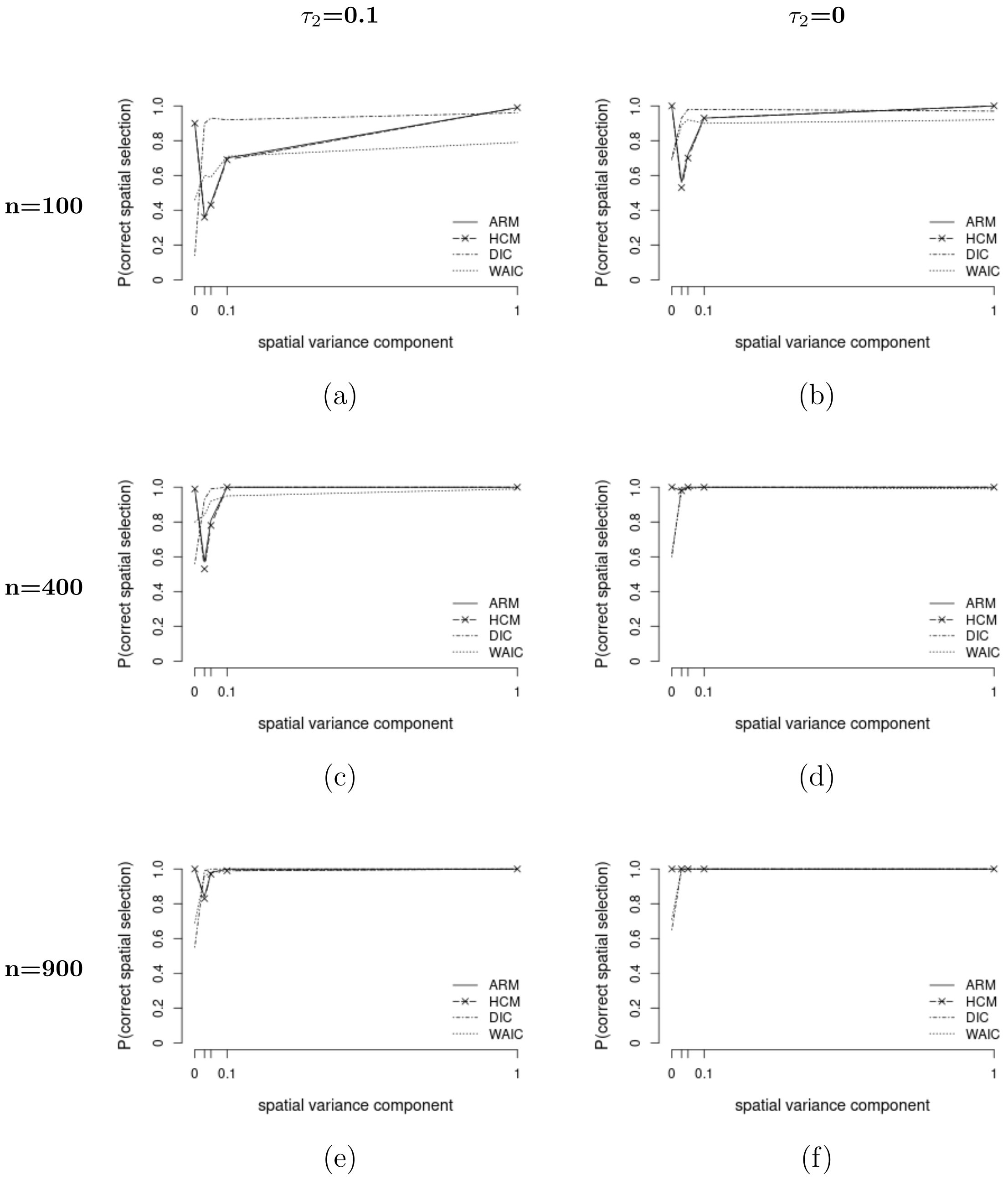
Figure 1. Probability of selecting the correct covariate structure as a function of the value of the regression coefficient, settings: $\tau_1 = 0.05$, $\tau_2 = 0.05$, n=100 (top row), n=400 (middle row), n=900 (bottom row), and $\beta_0 = 1$ (left column), $\beta_0 = 4$ (right column).



(a) $\tau_1 = 0.01$, $\tau_2 = 0$

(b)$\tau_1 = 1$, $\tau_2 = 1$

Figure 2. Probability of selecting the correct covariate structure as a function of the value of the regression coefficient, settings: (a) $\tau_1 = 0.01$ and $\tau_2 = 0$, and (b) $\tau_1 = 1$ and $\tau_2 = 1$, both with sample size $n = 400$ and intercept value $\beta_0 = 1$. (a) has weak dependence structure. (b) has strong dependence structure. Dependence structure can affect our method's performance for detecting covariates with small coefficients. However, the DIC and WAIC have difficulty detecting covariates even with large coefficients when spatial dependence is strong.



(a) (b) (c) (d) (e) (f)

Figure 3. Probability of selecting the correct spatial random effects structure as a function of the value of variance component for spatial random effects. Settings: $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, n=100 (top row), n=400 (middle row), n=900 (bottom row), and $\tau_2 = 0.1$ (left column), $\tau_2 = 0$ (right column). If the spatial variance proportion is zero then there is no vector of spatial random effects in the model, and the correct decision is to not select the vector of spatial random effects.

Table 1

Epilepsy data: posterior inclusion probabilities of fixed and random effects

	variable	ARM	HCM
fixed effect	Base	1	1
	Trt	0.14	0.04
	Trt × Base	0	0
	Age	0.03	0.01
	$\mathbf{V}4$	0.12	0.11
random effect	$\boldsymbol{\alpha}_1$	1	1
	$\boldsymbol{\alpha}_2$	0	0
	$\boldsymbol{\alpha}_3$	1	1

Table 2
Lip cancer data: posterior inclusion probabilities of fixed and random effects

	variable	ARM	HCM
fixed effect	AFF	0.93	0.92
random effect	$\boldsymbol{\alpha}_1$	1	1
	$\boldsymbol{\alpha}_2$	0	0