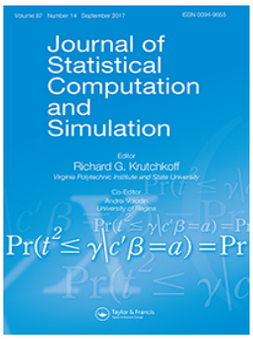
Fitting logistic multilevel models with crossed random effects via Bayesian Integrated Nested Laplace Approximations



Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Fitting logistic multilevel models with crossed random effects via Bayesian Integrated Nested Laplace Approximations: a simulation study

Leonardo Grilli & Francesco Innocenti

To cite this article: Leonardo Grilli & Francesco Innocenti (2017) Fitting logistic multilevel models with crossed random effects via Bayesian Integrated Nested Laplace Approximations: a simulation study, Journal of Statistical Computation and Simulation, 87:14, 2689-2707, DOI: 10.1080/00949655.2017.1341886

To link to this article: https://doi.org/10.1080/00949655.2017.1341886

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2017
VOL. 87, NO. 14, 2689–2707
https://doi.org/10.1080/00949655.2017.1341886

Taylor & Francis
Taylor & Francis Group

Fitting logistic multilevel models with crossed random effects via Bayesian Integrated Nested Laplace Approximations: a simulation study

Leonardo Grilli [a] and Francesco Innocenti [b]

[a]Dipartimento di Statistica, Informatica, Applicazioni 'G. Parenti', Università di Firenze, Firenze, Italy; [b]Department of Methodology and Statistics, Maastricht University, Maastricht, Netherlands

ABSTRACT

Fitting cross-classified multilevel models with binary response is challenging. In this setting a promising method is Bayesian inference through Integrated Nested Laplace Approximations (INLA), which performs well in several latent variable models. We devise a systematic simulation study to assess the performance of INLA with cross-classified binary data under different scenarios defined by the magnitude of the variances of the random effects, the number of observations, the number of clusters, and the degree of cross-classification. In the simulations INLA is systematically compared with the popular method of Maximum Likelihood via Laplace Approximation. By an application to the classical salamander mating data, we compare INLA with the best performing methods. Given the computational speed and the generally good performance, INLA turns out to be a valuable method for fitting logistic cross-classified models.

ARTICLE HISTORY
Received 20 July 2016
Accepted 9 June 2017

KEYWORDS
Binary response; crossed random effects; generalized linear mixed models; INLA; maximum Likelihood via Laplace approximation; non-hierarchical data; random effects; salamander mating data

AMS SUBJECT CLASSIFICATION
62F15; 62J12; 65C60

1. Introduction

Cross-classified data are non-hierarchical structures where lower level units belong to pairs or combinations of higher level units formed by crossing each other two or more higher level factors [1,2]. Examples include children cross-classified by primary and secondary schools [3] or by school and neighbourhood [4], and responses nested in the combination of test items and persons [5]. There are different degrees of cross-classification, that can be categorized essentially into two types [6]: (i) in a *complete cross-classification* the units in a cluster of one factor belong to all the clusters of the other crossed factor, and vice versa; (ii) in a *partial cross-classification* the units in a cluster of one factor belong to a subset of the clusters of the other crossed factor.

Linear cross-classified models have been widely studied in literature [6–8] and the related estimation issues have been satisfactorily addressed [4,9], as testified by the large number of published applications (see [10] for a detailed review).

On the other hand, fitting logistic cross-classified models is difficult for two reasons: (i) the distribution of the response conditional on the random effects is Bernoulli, thus the marginal likelihood is not in closed form; (ii) due to the cross-classification of the random effects, the variance-covariance matrix is not block-diagonal. In the simpler case of nested random effects (Generalized Linear Mixed Models), several methods are available to obtain Maximum Likelihood (ML) estimates, including linearization (MQL [11], PQL [12]) and numerical integration, such as Laplace Approximation (MLLA)

CONTACT Leonardo Grilli grilli@disia.unifi.it

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2017.1341886

[13] and Adaptive Gaussian Quadrature (AGQ) [14]. In general, ML methods tend to underestimate the variance components, especially in settings with a small number of clusters [15]. In the challenging case of crossed random effects, the above methods can still be used, even if better performances can be obtained by special algorithms based on data augmentation, such Monte Carlo Expectation Maximization (MCEM) [16] and Alternating Imputation Posterior (AIP) estimation [9,17].

Bayesian methods generally have a better performance in complex random effects models [18]. However, the standard Bayesian method, namely MCMC [19], has some practical limitations because of the computational burden and the difficulties in assessing convergence. A possible solution is represented by INLA, namely Integrated Nested Laplace Approximations [20]: indeed, INLA directly approximates the posterior distribution, thus avoiding complex simulation-based methods. INLA is promising because of the good performance observed in logistic models with nested random effects [21], where it is fast (nearly as frequentist quadrature methods) and accurate (slightly more than MCMC). The computational time ratio of MCMC over INLA depends on the sample size, for example Grilli et al. [21] obtained a ratio of about 10 in a data set with 10 clusters of size 50, and a ratio of about 50 in a data set with 100 clusters of size 50. Given these premises, it is worth to investigate the performance of INLA in logistic cross-classified multilevel models. In this setting, the only application we are aware of is reported in the Supplementary Material of Fong et al. [22], where INLA is used to fit model C of Karim and Zeger [19] on the classical salamander mating data. In that instance, INLA seems to underestimate the variance components, but a comprehensive evaluation of the method requires a systematic simulation study. We therefore devise a simulation study to evaluate INLA for a logistic model with two crossed random effects under several scenarios. The results are compared with those obtained with Maximum Likelihood via Laplace Approximation (MLLA), which is the default choice in many programs and it is similar to INLA in terms of computational time. We do not consider AGQ, which is generally superior to the Laplace Approximation, but it turns out to be infeasible in some scenarios. In the simulation study, we devote particular attention to situations with a small number of clusters, different degrees of cross-classification and low random effects variances. Furthermore, in order to compare INLA with a wide set of estimation methods (MCMC, MCEM, AIP), we apply it to the classical salamander mating data [17,19,23].

The rest of the paper is organized as follows. In Section 2 the INLA method is briefly introduced, whereas in Section 3 the simulation design is described. In Section 4 the findings of the simulation study are commented, comparing MLLA with INLA using two alternative prior distributions for the variance components. In Section 5 INLA is applied to the salamander mating data, allowing a comparison with several efficient algorithms. Section 6 offers some final remarks. The Supplementary Material collects further simulation results not reported in the paper.

2. The INLA method: a brief introduction

In this section we outline the INLA method; for a detailed illustration we recommend [20,24,25]. The INLA method is a deterministic approach to Bayesian inference in the wide framework of latent Gaussian models, which includes Generalized Linear Mixed Models [22]. Let $\boldsymbol{\theta}$ be a k-dimensional latent Gaussian random field, that is, a vector of k latent Gaussian variable, that is, the parameters of the model, then a Latent Gaussian Model can be constructed in three stages:

(1) Firstly, the observations $\boldsymbol{y}$ are assumed conditionally independent given $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_1$, a vector of hyperparameters.

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_1 \sim \pi(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_1).$$

(2) Secondly, the latent field $\boldsymbol{\theta}$ is assumed to be Normally distributed conditional to the hyperparameters $\boldsymbol{\gamma}_2$, with zero mean and precision matrix $\boldsymbol{Q}(\boldsymbol{\gamma_2})$

$$\boldsymbol{\theta} \mid \boldsymbol{\gamma}_2 \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\gamma_2})).$$

(3) Finally, a prior distribution for $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is specified

$$\boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}).$$

Therefore, assuming that the number of hyperparameters $\boldsymbol{\gamma}$ is small (say lower than 6 [20]), the targets of inference are:

$$\pi(\theta_i \mid \boldsymbol{y}) = \int \pi(\theta_i \mid \boldsymbol{\gamma}, \boldsymbol{y}) \pi(\boldsymbol{\gamma} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\gamma}, \tag{1}$$

that is, the marginal posterior distribution of parameter θ_i, with $i = 1, \ldots, k$, and

$$\pi(\gamma_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\gamma} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\gamma}_{-\boldsymbol{j}}, \tag{2}$$

the marginal posterior distribution of the hyperparameter γ_j, with $j = 1, \ldots, l$. Thus, the INLA algorithm is composed of three steps:

(1) Approximate the joint posterior distribution of the hyperparameters $\pi(\boldsymbol{\gamma} \mid \boldsymbol{y})$ with the following Laplace Approximation

$$\tilde{\pi}(\boldsymbol{\gamma} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{y})} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*(\boldsymbol{\gamma})}, \tag{3}$$

where $\tilde{\pi}_G(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$ is the Gaussian approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{y})$ derived by matching the mode and the curvature at the mode and $\boldsymbol{\theta}^*(\boldsymbol{\gamma})$ is the mean of $\tilde{\pi}_G(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{y})$. In order to facilitate the numerical integration in step 3, good evaluations points are selected exploring $\tilde{\pi}(\boldsymbol{\gamma} \mid \boldsymbol{y})$. Then, the marginal posterior distribution $\pi(\gamma_j|\boldsymbol{y})$ can be derived through numerical integration.

(2) Approximate $\pi(\theta_i \mid \boldsymbol{\gamma}, \boldsymbol{y})$ with one of the following alternative approaches:

(2.1) *Gaussian approximation*:

$$\tilde{\pi}_G(\theta_i \mid \boldsymbol{\gamma}, \boldsymbol{y}) = N(\theta_i; \mu_i(\boldsymbol{\gamma}), \sigma_i^2(\boldsymbol{\gamma})), \tag{4}$$

that can yield errors in location and/or lack of skewness.

(2.2) *Laplace approximation*:

$$\tilde{\pi}_{LA}(\theta_i \mid \boldsymbol{\gamma}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{y})}{\tilde{\pi}_{GG}(\boldsymbol{\theta_{-i}} \mid \theta_i, \boldsymbol{\gamma}, \boldsymbol{y})} \right|_{\boldsymbol{\theta_{-i}} = \boldsymbol{\theta^*_{-i}}(\theta_i, \boldsymbol{\gamma})}, \tag{5}$$

where $\tilde{\pi}_{GG}(\boldsymbol{\theta_{-i}}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y})$ is the Gaussian approximation to $\boldsymbol{\theta_{-i}}|\theta_i, \boldsymbol{\gamma}, \boldsymbol{y}$ centred at the modal configuration $\boldsymbol{\theta^*_{-i}}(\theta_i, \boldsymbol{\gamma})$. A drawback of this approach is that we must recompute $\tilde{\pi}_{GG}$ for each value of θ_i and $\boldsymbol{\gamma}$. Thus, k factorizations of the full precision matrix are needed.

(2.3) *Simplified Laplace approximation*: in simple terms, the Simplified Laplace Approximation $\tilde{\pi}_{SLA}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$ consists in doing a series expansion of the numerator and denominator of $\tilde{\pi}_{LA}(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$ up to the third order around $\theta_i = \mu_i(\boldsymbol{\gamma})$, this means to correct the Gaussian Approximation $\tilde{\pi}_G(\theta_i|\boldsymbol{\gamma}, \boldsymbol{y})$ for location and skewness. The benefit is purely computational.

(3) Combine 1 and 2 through numerical integration of $\tilde{\pi}(\boldsymbol{\gamma} \mid \boldsymbol{y}) \tilde{\pi}(\theta_i \mid \boldsymbol{\gamma}, \boldsymbol{y})$ to compute the marginal posterior distribution $\tilde{\pi}(\theta_i \mid \boldsymbol{y})$

$$\tilde{\pi}(\theta_i \mid \boldsymbol{y}) = \sum_s \tilde{\pi}(\theta_i \mid \boldsymbol{\gamma_s}, \boldsymbol{y}) \tilde{\pi}(\boldsymbol{\gamma} \mid \boldsymbol{y}) \Delta_s, \tag{6}$$

where the sum is over the values $\{\gamma_s\}$ which are found in step 1, with area weights Δ_s and $\tilde{\pi}(\theta_i \mid \boldsymbol{\gamma_s}, \boldsymbol{y})$ can be $\tilde{\pi}_{\mathrm{G}}$, $\tilde{\pi}_{\mathrm{LA}}$ or $\tilde{\pi}_{\mathrm{SLA}}$ depending on which approximation procedure we have chosen in step 2.

3. The simulation design

3.1. Model and sample structure

We consider a random intercept logistic model with two crossed random effects. Let $Y_{i(j_1j_2)}$ be a Bernoulli random variable for level 1 unit i (e.g. student) nested in two crossed classifications at level 2 (e.g. school and neighbourhood) with $j_1 = 1, \ldots, N_1$ and $j_2 = 1, \ldots, N_2$. Defining $\pi_{i(j_1j_2)} = P(Y_{i(j_1j_2)} = 1 | x_{1i(j_1j_2)}, x_{2i(j_1j_2)}, z_{j_1}, z_{j_2}, u_{j_1}, u_{j_2})$, the considered model is:

$$\text{logit}(\pi_{i(j_1j_2)}) = \alpha + \beta_1 x_{1i(j_1j_2)} + \beta_2 x_{2i(j_1j_2)} + \gamma_1 z_{1j_1} + \gamma_2 z_{2j_2} + u_{j_1} + u_{j_2} \tag{7}$$

$$u_{j_1} \sim N(0, \sigma^2_{u_{j_1}}) \quad u_{j_2} \sim N(0, \sigma^2_{u_{j_2}}),$$

where $x_{1i(j_1j_2)}$ is a continuous level 1 variable, $x_{2i(j_1j_2)}$ is a binary level 1 variable, z_{1j_1} and z_{2j_2} are binary level 2 variables (the former related to classification 1, the latter related to classification 2). The continuous covariate $x_{1i(j_1j_2)}$ is drawn from a standard Normal distribution, whereas the binary covariates are drawn from independent Bernoulli distributions with success probability equal to 0.5. In the Supplementary Material, we report simulation results for a success probability equal to 0.9, showing that highly skewed binary covariates tend to reduce the performance of the estimators, though the differences are often modest and the main patterns are preserved.

The true values of parameters in model (7) are set as follows: $\alpha = 0.1$ (so that the baseline individual has a probability of success slightly greater than 0.5), $\beta_1 = 0.1$, and $\beta_2 = \gamma_1 = \gamma_2 = 0.4$. Given that in a standard Normal distribution about 95% of the probability lies between -2 and $+2$, setting $\beta_1 = 0.1$ ensures that the continuous covariate has an effect comparable to that of the binary covariates. The values of the regression coefficients are constant across all configurations.

On the other hand, several values are considered for the variances of the random effects $\sigma^2_{u_{j_1}}$ and $\sigma^2_{u_{j_2}}$ since it is known that they strongly affect the performance of the estimation methods and the importance of the prior distribution [18,26]. Specifically, we consider four configurations yielded by setting the variances of the two random effects at either 0.01 (low impact of the random effects) or 0.25 (sizeable impact of the random effects). To see the impact of the random effects, note that a variance $\sigma^2_{u_{j.}} = 0.01$ corresponds to a standard deviation $\sigma_{u_{j.}} = 0.1$, thus under normality the random effect approximately has 95% probability of lying in the interval $[-0.2, 0.2]$, corresponding to the central interval $[0.45, 0.55]$ in terms of probability of the response variable. Similarly, $\sigma^2_{u_{j.}} = 0.25$ corresponds to $\sigma_{u_{j.}} = 0.50$ so that the central 95% interval of the probability is $[0.27, 0.73]$. In the Supplementary Material we also report simulations with both variances at 1.00, though we do not put much emphasis on the results since random effects of such size are rarely found in applications. Anyway, it is worth to note that inference about random effects with variance 1.00 is not problematic and, indeed, the performances of all the considered estimators are satisfactory.

In order to evaluate the influence of the degree of cross-classification on the performance of INLA and MLLA [6], we consider three scenarios ranging from complete cross-classification to an almost hierarchical structure:

- a complete cross-classified structure: we consider a square cross-classification matrix, namely the two classification factors have the same number of clusters $N_1 = N_2$. Since Bayesian and frequentist methods can differ substantially in scenarios with a small number of clusters, we focus our investigation on the case $N_1 = N_2 = 10$, considering four different values for the number of observations per cell n (1, 5, 10, 20). Moreover, in order to assess the asymptotic behaviour of the estimators, we consider four different values for $N_1 = N_2$ (10, 20, 50, 80), with $n = 10$ observations per cell in each scenario.
- two partial cross-classified structures: we generate those structures as follows [6]:

Table 1. Partial cross-classified structure with 10 feeders and 5 receivers (the symbol × denotes the presence of at least one observation).

		Receivers									
		1	2	3	4	5	6	7	8	9	10
	1		×	×		×		×			×
F	2		×	×				×		×	×
E	3	×	×	×	×				×		
E	4		×	×		×	×				×
D	5	×				×	×		×	×	
E	6		×	×	×	×				×	
R	7		×	×		×		×	×		
S	8		×		×	×	×	×			
	9	×		×	×		×				×
	10		×			×	×	×			×

(1) Generate a hierarchical three-level model: classification 1 is the third level with 10 clusters, classification 2 is the second level with 10 clusters within each third-level unit, and 100 observations are nested within each second-level unit.
(2) Randomly draw 10 second-level units (called *feeders*).
(3) Randomly draw k third-level units (called *receivers*), where k is set to either 2 or 5.
(4) For each feeder, randomly assign the observations to the receivers (50 observations per receiver in case of 2 receivers, and 20 observations per receiver in case of 5 receivers).

Table 1 represents a partial cross-classified structure with 10 feeders and 5 receivers. Note that the three structures outlined above have the same sample size (1000 observations), but they differ in the distribution of the empty cells (which is an indicator of the degree of cross-classification [6]) and in the number of observations per cell ($n = 10$ in the complete cross-classified structure, $n = 50$ in the structure with 2 receivers, and $n = 20$ in structure with 5 receivers).

3.2. Estimation methods and prior distributions

The simulation study is performed with the `R` software (version 3.2.2). We let our code available on the web. In particular, we exploit the following packages:

- the `inla` package (version 0.0-1440400394) for INLA – Bayesian inference through Integrated Nested Laplace Approximations [25,27];
- the `lme4` package (version 1.1-9) for MLLA – Maximum Likelihood via Laplace Approximation [28].

We rely on the default approximation method of the `inla` package, namely the Simplified Laplace Approximation. More complex methods, such as the Laplace Approximation and a Gaussian copula correction proposed by Ferkingstad and Rue [29], are available in the `inla` package, but simulations for the most challenging scenario show they are not worthwhile (see the Supplementary Material).

In Bayesian inference with INLA we specify a Normal prior distribution with zero mean and large variance for the regression coefficients (the default of the `inla` function). Since we focus on scenarios with small numbers of clusters and variances close to zero, the choice of the prior distribution for the variance components is crucial. As usual in Bayesian software, the `inla` function allows us to specify the prior distribution of the precision, instead of the variance. We avoid the default gamma prior Ga(1, 0.0005) because it has a poor performance in logistic models with nested random effects [21]. For the simulation study we choose two alternative priors for the precisions: Ga(0.001,0.001), namely the standard choice in the popular BUGS software, and Ga(0.5, 0.003737) specified according to the criterion proposed by Fong et al. [22], which consists in setting the parameters of the Gamma in

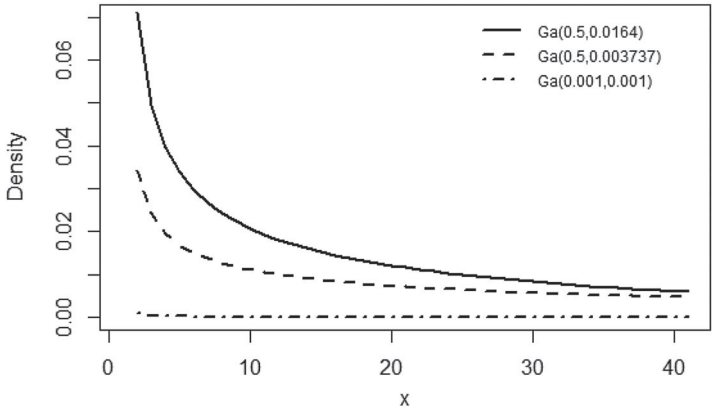


Figure 1. Probability densities of the considered prior distributions in [0, 40].

order to obtain a given marginal distribution for the random effects. In particular, Ga(0.5, 0.003737) yields a marginal Cauchy distribution having 95% probability of $e^{u_j} \in [\frac{1}{3}, 3]$, corresponding to a central interval for the probability of the response variable equal to [0.25, 0.75]. It is worth to note that the selected prior Ga(0.5, 0.003737) is different from the prior usually derived by applying the Fong et al. criterion (like in the simulations of Grilli et al. [21]), which is Ga(0.5, 0.0164). This prior amounts to random effects with a stronger impact, as it yields a marginal Cauchy distribution having 95% probability of $e^{u_j} \in [\frac{1}{10}, 10]$, corresponding to a central interval for the probability of the response variable equal to [0.09, 0.91]. In the simulations we tried both the priors derived by the Fong et al. criterion, but we retained only Ga(0.5, 0.003737) as it was outperforming the other one. The densities of the three mentioned priors are depicted in Figure 1, showing that Ga(0.5, 0.003737) is less informative than Ga(0.5, 0.0164), though more informative than Ga(0.001,0.001).

4. Simulation results

4.1. Measures of performance

The performances of INLA and MLLA are compared on the basis of the following measures of accuracy [6–8], where m is one of the scenarios defined in Section 3.1 and l is one of the L Monte Carlo replicates:

- relative bias for the estimates of the regression coefficients and variance components:

$$\mathrm{RB}(\hat{\theta}_m) = \frac{\hat{\theta}_m - \theta}{\theta},$$

  where $\hat{\theta}_m = \sum_{l=1}^{L} \hat{\theta}_{lm}/L$ is the Monte Carlo average of the estimates $\hat{\theta}_{lm}$ (point estimates for MLLA and posterior means for INLA) and θ is the population parameter;
- relative bias for the standard errors of the regression coefficients:

$$\mathrm{RB}(S_{\hat{\theta}_m}) = \frac{\overline{\mathrm{SE}}(\hat{\theta}_m) - \mathrm{SD}(\hat{\theta}_m)}{\mathrm{SD}(\hat{\theta}_m)},$$

  where $\overline{\mathrm{SE}}(\hat{\theta}_m) = \sum_{l=1}^{L} \mathrm{SE}(\hat{\theta}_{lm})/L$ is the Monte Carlo average standard error and $\mathrm{SD}(\hat{\theta}_m)$ is the Monte Carlo standard error, namely the standard deviation of the L estimates $\hat{\theta}_{lm}$.

The standard errors of the variance components are not considered because they are not provided by the `glmer` function in the `lme4` package (in general, it is not advisable to exploit the standard errors to make inference on the variance components).

4.2. Extreme estimates of the variance components

In order to give practical advice to applied researchers, it is worth to study when the two considered estimators yield extreme values for the estimates of the variance components. For MLLA it sometimes happens that $\hat{\sigma}^2_{u_{j.}} = 0$, namely the estimate is on the border of the parameter space. This problem does not occur with INLA since the priors push the estimates into the parameter space; however, INLA sometimes yields unrealistically large estimates. In the following, we label as aberrant the estimates larger than 2, namely $\hat{\sigma}^2_{u_{j.}} > 2$. Such threshold is necessarily subjective as it corresponds to the largest value that a researcher is willing to trust. For each scenario we report the percentages of null estimates of MLLA and aberrant estimates of INLA out of the 500 replicates.

In Bayesian inference with non-informative priors, the usual action in case of aberrant estimates is to change the priors, therefore we discard the replicates where at least one of the variance components is larger than 2 and compute the relative bias on the remaining replicates $L \leq 500$ (note that in most scenarios there are no aberrant estimates, thus $L = 500$). Discarding the replicates with aberrant estimates has a noticeable effect on the Monte Carlo relative bias of the variance components, whereas the effect on the regression coefficients is negligible.

4.3. Scenarios with few clusters

In our simulation study we devote particular attention to scenarios with few clusters because in these cases the estimation of the variances of the random effects is challenging and the influence of prior distributions is amplified, so that Bayesian and ML methods may yield considerably different results [18].

In Table 2 we compare the relative biases (net of aberrant estimates) for the regression coefficients yielded by INLA and MLLA in a complete 10×10 cross-classification matrix with varying number of observations per cell. Note that INLA and MLLA give similar results for the regression coefficients: both methods yield relative biases smaller than 10% even with $n = 5$ observations per cell and they decrease for larger cell sample sizes. However, the direction of the biases is hardly predictable. On the other hand, INLA and MLLA differ in the estimation of the standard errors of the regression coefficients: INLA yields more accurate standard errors for larger values of the variances of the random effects $\sigma^2_{j_1}$ and $\sigma^2_{j_2}$, whereas in this regard MLLA performs better when $\sigma^2_{j_1}$ and $\sigma^2_{j_2}$ are close to zero.

Figure 2 reports the relative biases (net of aberrant estimates) for random effects variances, highlighting the differences among MLLA and INLA with the two considered priors for the variance components. When both variance components are low (0.01), all the methods overestimate the population values, but the biases rapidly decline as the cell sample size n increases ($n = 10$ is enough for INLA, though not for MLLA as it underestimates the first variance). The two priors yield similar results. When both variance components are sizable (0.25), the three methods perform well even for small cell sample sizes. Note that, generally, MLLA underestimates the variance components, whereas INLA overestimates it, with the prior Ga(0.5, 0.003737) outperforming Ga(0.001, 0.001). The cases where the variance components have markedly different sizes (0.01 and 0.25) are troubling since the low variance component can be badly estimated even for $n = 10$ or $n = 20$. It is worth to note that these configurations have been considered as they are especially challenging, though they are unlikely in practice. In the configuration with variance components $\sigma^2_{j_1} = 0.01$ and $\sigma^2_{j_2} = 0.25$, INLA with the prior Ga(0.001, 0.001) shows an anomalous behaviour because the bias abruptly increases when moving from cell size $n = 10$ to $n = 20$. This pattern is analysed in detail in the Supplementary Material.

Table 2. Relative bias for regression coefficients (relative bias of standard errors in parenthesis).

n	INLA Ga(0.001, 0.001)	INLA Ga(0.5, 0.003737)	MLLA
		α	
1	−0.118 (0.014)	0.167 (−0.148)	−0.060 (−0.092)
5	0.078 (0.174)	0.380 (−0.020)	0.056 (−0.023)
10	−0.110 (0.102)	−0.030 (0.003)	−0.120 (−0.103)
20	−0.110 (0.002)	0.120 (−0.007)	−0.120 (−0.153)
		β_1	
1	0.364 (−0.141)	0.249 (−0.120)	0.270 (−0.078)
5	0.000 (0.000)	0.060 (−0.081)	−0.014 (0.016)
10	0.050 (−0.018)	0.060 (−0.047)	0.040 (−0.012)
20	0.030 (0.016)	−0.010 (−0.037)	0.020 (0.018)
		β_2	
1	0.113 (−0.081)	−0.057 (0.010)	0.035 (−0.011)
5	0.064 (0.015)	0.017 (−0.021)	0.050 (0.031)
10	0.025 (−0.006)	0.003 (0.065)	0.017 (0.008)
20	−0.003 (−0.038)	−0.003 (−0.024)	−0.005 (−0.036)
		γ_1	
1	0.220 (−0.003)	0.110 (−0.149)	0.092 (−0.081)
5	0.022 (0.120)	−0.105 (−0.044)	0.002 (−0.082)
10	0.008 (0.041)	0.080 (−0.009)	0.000 (−0.191)
20	−0.018 (0.069)	0.013 (0.002)	−0.020 (−0.111)
		γ_2	
1	0.230 (−0.062)	0.106 (−0.218)	0.142 (−0.148)
5	−0.053 (−0.132)	−0.030 (−0.089)	−0.064 (−0.158)
10	−0.008 (0.040)	0.065 (−0.025)	−0.013 (−0.193)
20	−0.005 (−0.013)	−0.018 (−0.015)	−0.008 (−0.022)

Notes: Logistic model of Equation (7) with $\sigma^2_{j_1} = \sigma^2_{j_2} = 0.25$. Complete cross-classification with $N_1 = N_2 = 10$ and varying number of observations per cell n.

Table 3. Percentage of extreme estimates out of the 500 replicates.

	INLA Ga(0.001, 0.001)		INLA Ga(0.5, 0.003737)		MLLA	
n	$\% \hat{\sigma}^2_{j_1} > 2$	$\% \hat{\sigma}^2_{j_2} > 2$	$\% \hat{\sigma}^2_{j_1} > 2$	$\% \hat{\sigma}^2_{j_2} > 2$	$\% \hat{\sigma}^2_{j_1} = 0$	$\% \hat{\sigma}^2_{j_2} = 0$
			$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.01$			
1	0.4	1.0	5.4	3.8	47.0	54.2
5	0.0	0.0	0.0	0.0	46.6	42.0
10	0.0	0.0	0.0	0.0	44.6	35.6
20	0.0	0.0	0.0	0.0	34.0	24.0
			$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.25$			
1	5.2	4.0	11.0	10.8	23.4	24.8
5	0.0	0.0	0.0	0.0	3.2	1.8
10	0.0	0.0	0.0	0.0	0.8	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0
			$\sigma^2_{j_1} = 0.01\, \sigma^2_{j_2} = 0.25$			
1	0.6	5.6	4.0	12.0	38.0	28.4
5	0.0	0.0	0.0	0.0	29.0	2.2
10	0.0	0.0	0.0	0.0	39.4	0.0
20	0.0	0.0	0.0	0.0	33.4	0.0
			$\sigma^2_{j_1} = 0.25\, \sigma^2_{j_2} = 0.01$			
1	6.0	0.6	12.6	4.4	26.6	38.0
5	0.0	0.0	0.0	0.0	5.2	22.8
10	0.2	0.0	0.0	0.0	1.0	19.6
20	0.2	0.0	0.0	0.0	0.0	20.0

Note: Complete cross-classification with $N_1 = N_2 = 10$ and varying number of observations per cell n.

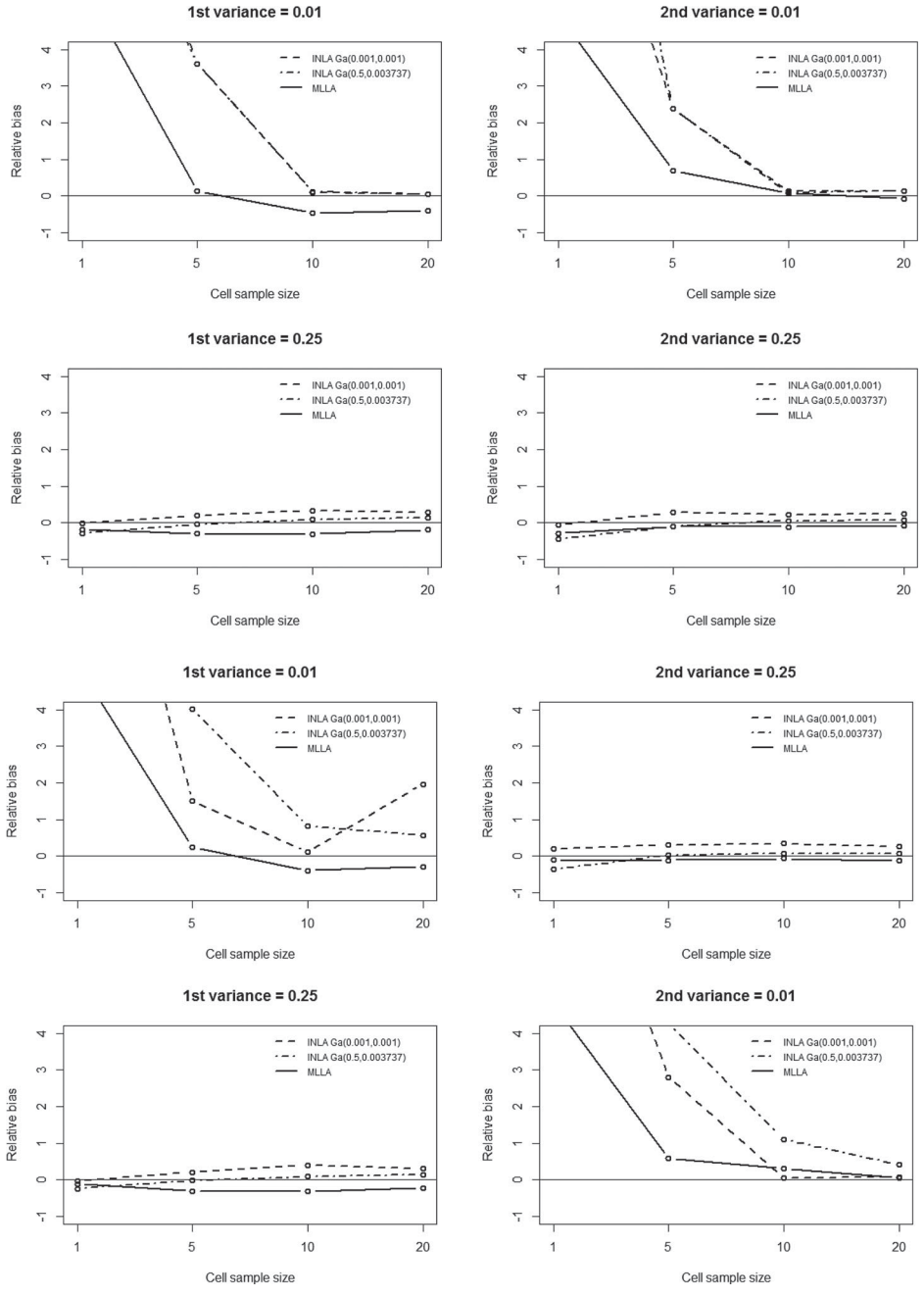


Figure 2. Relative bias for the variance components of the logistic model of Equation (7). Complete cross-classification with $N_1 = N_2 = 10$ and varying number of observations per cell n. Each pair of graphs corresponds to a combination of random effects variances $(\sigma^2_{j_1}, \sigma^2_{j_2})$: (0.01, 0.01), (0.25, 0.25), (0.01, 0.25), (0.25, 0.01).

As discussed at the end of Section 4.2, the considered estimation methods are prone to different kinds of extreme estimates for the variance components, namely MLLA may yield zero values, whereas INLA may yield very large values (here considered to be aberrant when larger than 2). Table 3 reports the percentage of extreme estimates out of 500 for each scenario. For MLLA the issue of zero estimates is severe (above 20%) for low variance components even in large sample sizes, whereas it is

Table 4. Relative bias for regression coefficients (relative bias of standard errors in parenthesis).

n	Receivers	INLA Ga(0.001, 0.001)	INLA Ga(0.5, 0.003737)	MLLA
			α	
50	2	0.230 (0.128)	0.020 (0.023)	0.220 (−0.051)
20	5	0.010 (0.028)	0.080 (0.244)	0.000 (−0.116)
10	10	−0.110 (0.102)	−0.030 (0.003)	−0.120 (−0.103)
			β_1	
50	2	−0.020 (0.049)	−0.030 (−0.037)	−0.020 (0.054)
20	5	0.010 (0.016)	0.070 (0.002)	0.010 (0.021)
10	10	0.050 (−0.018)	0.060 (0.065)	0.040 (−0.012)
			β_2	
50	2	−0.023 (−0.012)	0.005 (−0.037)	−0.028 (−0.006)
20	5	0.017 (−0.020)	−0.003 (0.002)	0.010 (−0.007)
10	10	0.025 (−0.006)	0.003 (0.065)	0.017 (0.008)
			γ_1	
50	2	0.017 (0.115)	0.032 (0.214)	0.008 (−0.078)
20	5	−0.020 (0.069)	−0.050 (0.164)	−0.028 (−0.108)
10	10	0.008 (0.041)	0.080 (−0.009)	0.000 (−0.191)
			γ_2	
50	2	0.000 (−0.056)	−0.005 (0.025)	−0.005 (−0.051)
20	5	0.025 (−0.011)	0.060 (−0.015)	0.020 (−0.006)
10	10	−0.008 (0.040)	0.065 (−0.025)	−0.013 (−0.193)

Notes: Logistic model of equation (7) with $\sigma^2_{j1} = \sigma^2_{j2} = 0.25$. structures with different degrees of cross-classification (10 feeders and varying number of receivers), and varying number of observations per cell n.

severe for variance components at 0.25 only in scenarios with cell size $n = 1$. The opposite issue for INLA, namely aberrant estimates, has noticeable percentages only for variance components at 0.25 and cell size $n = 1$. As expected, aberrant estimates are more likely with the more informative prior Ga(0.5, 0.003737).

4.4. Comparing scenarios with different degrees of cross-classification

It is well known that omitting a crossed factor in a linear model yields a bias on the variance of the remaining factor [7,30]. According to Luo and Kwok [6], the direction and magnitude of the bias are related to the degree of cross-classification. In the evaluation of INLA we do not consider the omission of a factor, anyway it is worth to check whether the degree of cross-classification plays a role in the performance of the estimators.

Following Luo and Kwok [6], we measure the degree of cross-classification by the number of receivers for a fixed number of feeders (see, e.g. Table 1). In particular, given 10 feeders, we consider three configurations, namely 10 receivers (complete cross-classification, i.e. without empty cells), 5 receivers (partial cross-classification with 50% empty cells), and 2 receivers (partial cross-classification with 80% empty cells, a situation close to a hierarchical structure). The three configurations have different numbers of observations per cell to ensure a total sample size of 1000. Table 4 reports the results obtained under different degrees of cross-classification for a scenario where the variances of the random effects are equal to 0.25, in order to compare these results to those of Table 2. The effect of the degree of cross-classification on the bias of the regression coefficients is conflicting, but remarkably modest. Therefore, in a situation with 1000 observations in a 10×10 matrix, the pattern of empty cells is practically uninfluential for the estimation of the regression coefficients: this is a noteworthy result since in many cross-classified data sets most of the cells are empty [8].

The effect of the degree of cross-classification on the estimation of the variances of the random effects is shown in Figure 3. The effect is more sizable than for regression coefficients, especially if the true variance is low (0.01): in such instances, the performance of INLA improves as the cross-classification matrix becomes closer to completeness (10 receivers), especially for the prior Ga(0.5,

Table 5. Percentage of extreme estimates out of the 500 replicates.

		INLA Ga(0.001, 0.001)		INLA Ga(0.5, 0.003737)		MLLA	
Receivers	n	$\%\,\hat{\sigma}^2_{j_1} > 2$	$\%\,\hat{\sigma}^2_{j_2} > 2$	$\%\,\hat{\sigma}^2_{j_1} > 2$	$\%\,\hat{\sigma}^2_{j_2} > 2$	$\%\,\hat{\sigma}^2_{j_1} = 0$	$\%\,\hat{\sigma}^2_{j_2} = 0$
				$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.01$			
2	50	0.0	0.0	0.0	0.0	36.6	35.0
5	20	0.0	0.0	0.0	0.0	39.4	30.6
10	10	0.0	0.0	0.0	0.0	44.6	35.6
				$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.25$			
2	50	0.2	0.0	0.0	0.0	1.4	0.8
5	20	0.0	0.0	0.0	0.0	0.8	0.6
10	10	0.0	0.0	0.0	0.0	0.8	0.0
				$\sigma^2_{j_1} = 0.01\ \sigma^2_{j_2} = 0.25$			
2	50	0.0	0.0	0.0	0.0	20.6	0.4
5	20	0.0	0.0	0.0	0.0	25.8	0.0
10	10	0.0	0.0	0.0	0.0	39.4	0.0
				$\sigma^2_{j_1} = 0.25\ \sigma^2_{j_2} = 0.01$			
2	50	0.0	0.0	0.0	0.0	0.8	22.2
5	20	0.0	0.0	0.0	0.0	0.8	17.0
10	10	0.2	0.0	0.0	0.0	1.0	19.6

Note: Structures with different degree of cross-classification (10 feeders and varying number of receivers), and varying number of observations per cell n.

0.003737). The effect on the performance of MLLA is more conflicting because in some scenarios increasing the number of receivers implies a larger bias (even for variance equal to 0.25), while in other scenarios the bias remains unchanged across the three degrees of cross-classification and in one case it slightly decreases (for the second variance in the scenario with $\sigma^2_{j_1} = 0.25$ and $\sigma^2_{j_2} = 0.01$). This contradictory behaviour can be explained by the high percentage of null estimates which undoubtedly affects the bias. It is worth to note that, for all the estimators, the degree of cross-classification plays a major role when the variances of the two factors are markedly different (0.01 and 0.25): in those cases, if the matrix is sparse (2 receivers) the estimation of the low variance is largely out of the target.

Table 5 reports the percentages of extreme estimates for the variance components, showing that the degree of cross-classification has a negligible role. Note that in the considered scenarios, all with a total sample size of 1000, INLA produces almost no aberrant estimates, whereas MLLA yields many zero estimates when the variance component is low (0.01).

4.5. Assessing the asymptotic behaviour

In order to evaluate the asymptotic behaviour of the considered estimation methods, we increase the number of clusters per classification in a setting with complete cross-classification and constant cell sample size $n = 10$. For the regression coefficients, Table 6 shows that INLA and MLLA have similar, satisfactory performances: the relative biases are smaller than 5% with 20 clusters per classification and the differences between the two methods decline sharply as the number of clusters increases.

Also for the variance components (Figure 4) the results of the three considered estimators become similar as the number of clusters increases, though the requirement for a satisfactory performance is higher (50 clusters per classification). As for the priors, INLA Ga(0.5, 0.003737) has the best performance among the three estimators when $N_1 = N_2 \geq 50$ regardless of the magnitude of the variance components. On the other hand, in small cross-classification matrices (10×10 or 20×20) the performance depends on the size of the variance components, with INLA Ga(0.001, 0.001) yielding the best performance when the variance is low (0.01). It is worth to note that INLA with the two considered priors always overestimates the variance components. On the other hand, MLLA underestimates a variance component with value 0.25, whereas the direction of the bias is unpredictable for a variance component with value 0.01.

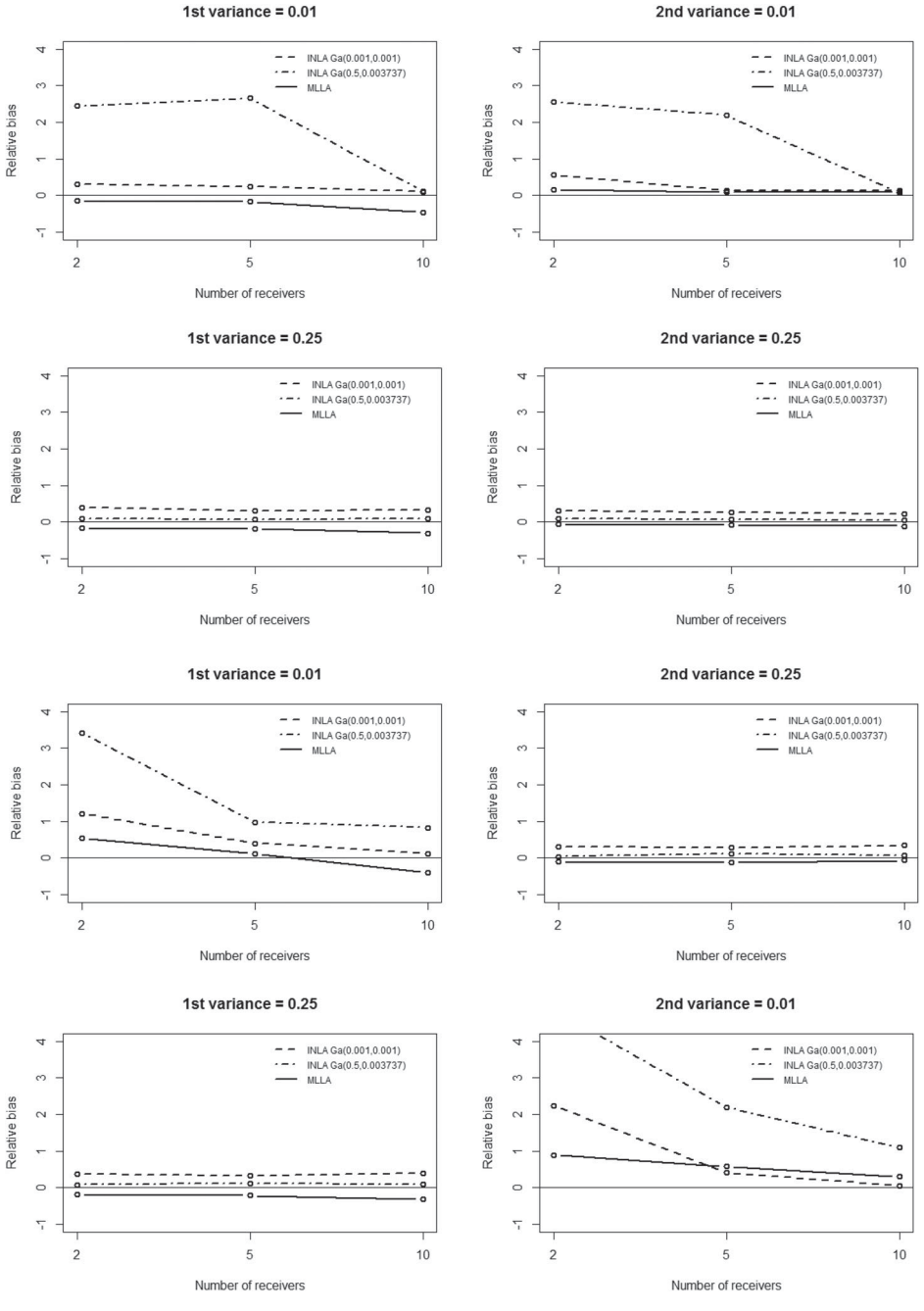


Figure 3. Relative bias for the variance components of the logistic model of Equation (7). Structures with different degree of cross-classification (10 feeders and varying number of receivers). The cell sample size n is set on the basis of the number of receivers to ensure a total sample size of 1000. Each pair of graphs corresponds to a combination of random effects variances $(\sigma^2_{J1}, \sigma^2_{J2})$: (0.01, 0.01), (0.25, 0.25), (0.01, 0.25), (0.25, 0.01).

The percentages of extreme estimates are reported in Table 7. INLA with the considered priors does not suffer from the issue of aberrant estimates even in the smallest design (10×10 with 10 observations per cell), while MLLA yields high percentages of zero estimates when estimating a low variance component in 10×10 and 20×20 designs.

Table 6. Relative bias for regression coefficients (relative bias of standard errors in parenthesis).

$N_1 = N_2$	INLA Ga(0.001, 0.001)	INLA Ga(0.5, 0.003737)	MLLA
		α	
10	−0.110 (0.102)	−0.030 (0.003)	−0.120 (−0.103)
20	0.080 (0.029)	−0.070 (−0.022)	0.080 (−0.046)
50	−0.020 (−0.013)	0.140 (−0.021)	−0.020 (−0.041)
80	−0.010 (−0.009)	−0.040 (−0.067)	−0.010 (−0.020)
		β_1	
10	0.050 (−0.018)	0.060 (−0.047)	0.040 (−0.012)
20	0.030 (−0.018)	0.020 (−0.009)	0.030 (−0.017)
50	−0.010 (−0.039)	0.010 (0.075)	−0.010 (−0.039)
80	0.000 (0.067)	0.010 (−0.005)	0.000 (0.067)
		β_2	
10	0.025 (−0.006)	0.003 (0.065)	0.017 (0.008)
20	−0.010 (0.088)	0.003 (−0.011)	−0.013 (0.090)
50	0.005 (−0.016)	−0.003 (−0.064)	0.005 (−0.016)
80	0.000 (0.002)	0.000 (0.011)	0.000 (0.002)
		γ_1	
10	0.008 (0.041)	0.080 (−0.009)	0.000 (−0.191)
20	−0.018 (0.023)	0.022 (−0.008)	−0.020 (−0.060)
50	0.020 (0.039)	−0.050 (0.014)	0.020 (0.012)
80	0.030 (−0.010)	0.022 (−0.008)	0.030 (−0.027)
		γ_2	
10	−0.008 (0.040)	0.065 (−0.025)	−0.013 (−0.193)
20	0.005 (−0.018)	−0.025 (−0.026)	0.003 (−0.018)
50	0.003 (−0.027)	0.000 (−0.045)	0.003 (0.002)
80	0.000 (−0.036)	0.003 (−0.010)	0.000 (−0.036)

Notes: Logistic model of equation (7) with $\sigma^2_{j_1} = \sigma^2_{j_2} = 0.25$. Complete cross-classification with varying number of clusters per classification $N_1 = N_2$ and $n = 10$ observations per cell.

Table 7. Percentage of extreme estimates out of the 500 replicates.

	INLA Ga(0.001, 0.001)		INLA Ga(0.5, 0.003737)		MLLA	
$N_1 = N_2$	$\% \hat{\sigma}^2_{j_1} > 2$	$\% \hat{\sigma}^2_{j_2} > 2$	$\% \hat{\sigma}^2_{j_1} > 2$	$\% \hat{\sigma}^2_{j_2} > 2$	$\% \hat{\sigma}^2_{j_1} = 0$	$\% \hat{\sigma}^2_{j_2} = 0$
			$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.01$			
10	0.0	0.0	0.0	0.0	44.6	35.6
20	0.0	0.0	0.0	0.0	19.8	12.4
50	0.0	0.0	0.0	0.0	0.0	0.0
80	0.0	0.0	0.0	0.0	0.0	0.0
			$\sigma^2_{j_1} = \sigma^2_{j_2} = 0.25$			
10	0.0	0.0	0.0	0.0	0.8	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0
50	0.0	0.0	0.0	0.0	0.0	0.0
80	0.0	0.0	0.0	0.0	0.0	0.0
			$\sigma^2_{j_1} = 0.01\, \sigma^2_{j_2} = 0.25$			
10	0.0	0.0	0.0	0.0	39.4	0.0
20	0.0	0.0	0.0	0.0	17.2	0.0
50	0.0	0.0	0.0	0.0	0.0	0.0
80	0.0	0.0	0.0	0.0	0.0	0.0
			$\sigma^2_{j_1} = 0.25\, \sigma^2_{j_2} = 0.01$			
10	0.2	0.0	0.0	0.0	1.0	19.6
20	0.0	0.0	0.0	0.0	0.0	12.0
50	0.0	0.0	0.0	0.0	0.0	0.0
80	0.0	0.0	0.0	0.0	0.0	0.0

Note: Complete cross-classification with varying number of clusters per classification $N_1 = N_2$ and $n = 10$ observations per cell.

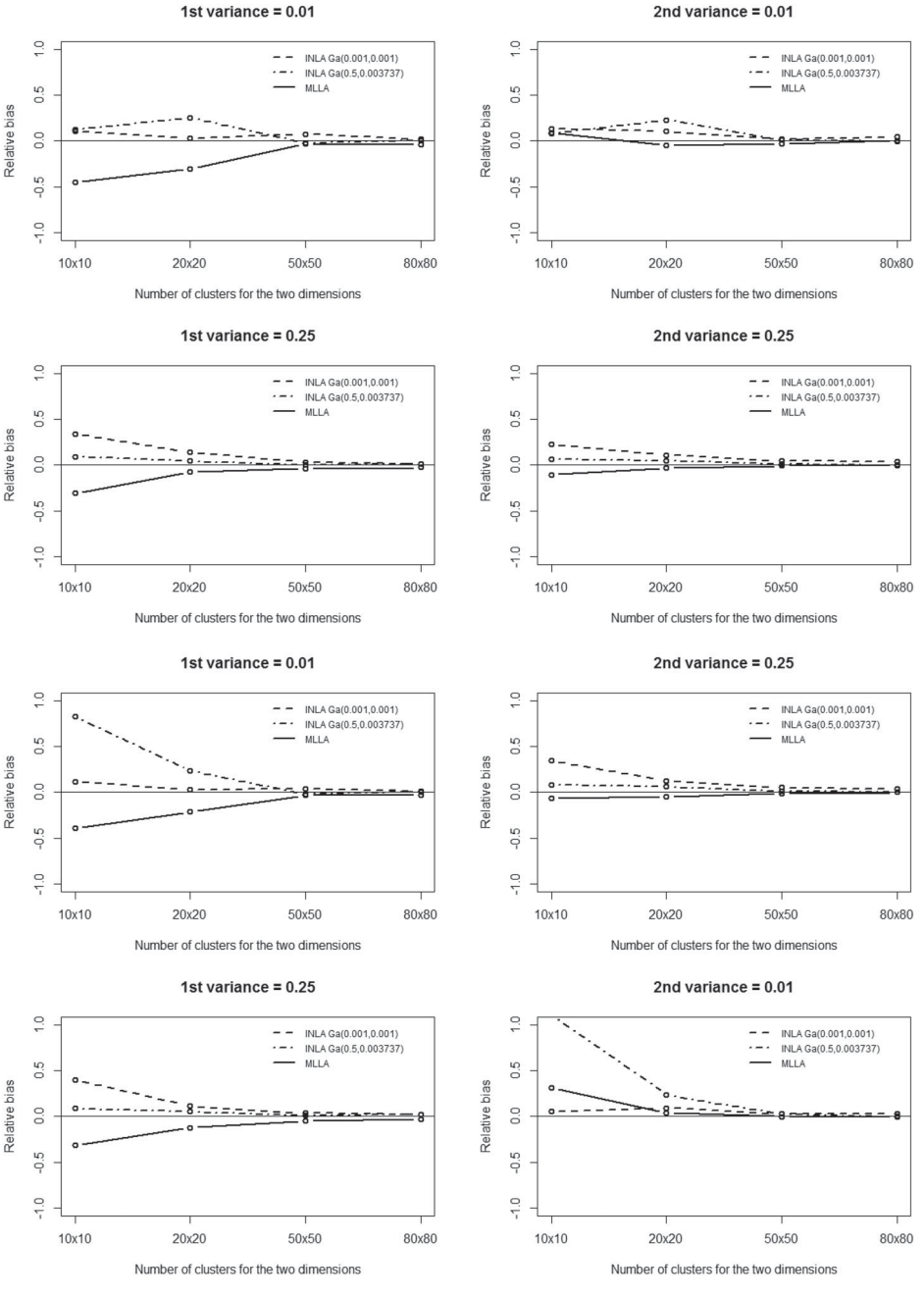


Figure 4. Relative bias for the variance components of the logistic model of Equation (7). Complete cross-classification with varying numbers of clusters per dimension $N_1 \times N_2$ and $n = 10$ observations per cell. Each pair of graphs corresponds to a combination of random effects variances $(\sigma^2_{j_1}, \sigma^2_{j_2})$: (0.01, 0.01), (0.25, 0.25), (0.01, 0.25), (0.25, 0.01).

5. The salamander mating data

In this Section we summarize the results obtained by applying INLA to the famous salamander mating data, which have become a standard test for estimation methods of cross-classified logistic models. The salamander mating data, presented for the first time by McCullagh and Nelder [23], were collected in 1986 by S. Arnold and P. Verell of the University of Chicago, Department of Ecology and

Table 8. The salamander mating data structure: columns represent factor 1 (male), rows represent factor 2 (female).

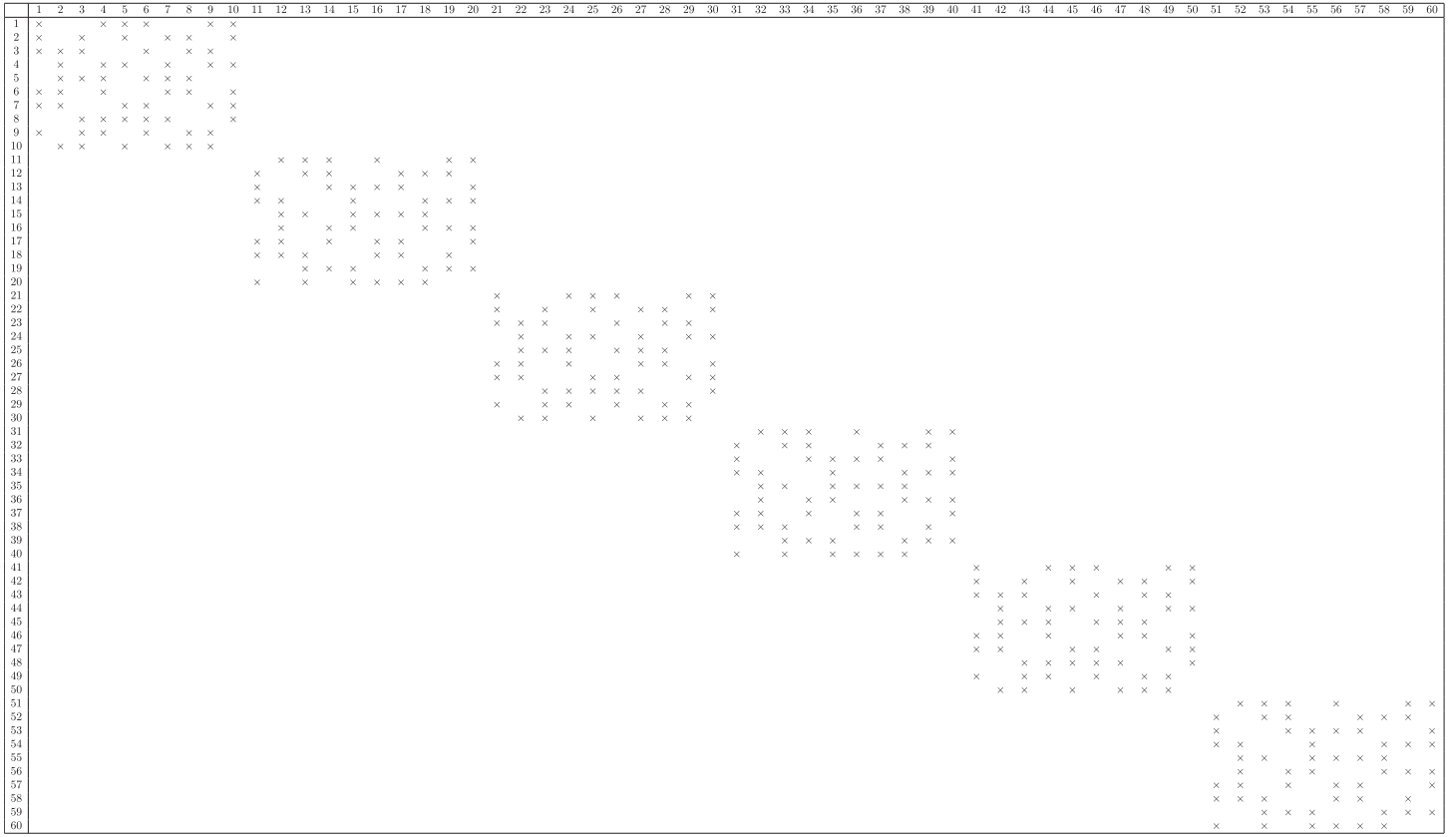


Table 9. Results for salamander data. standard errors in parenthesis.

	MCEM[a]	AIP with AGQ[a]	MLLA[a]	MCMC[b]	INLA Ga(0.001, 0.001)	INLA Ga(0.5, 0.003737)
β_0	1.02	1.02(0.41)	1.00(0.37)	1.03(0.43)	1.01(0.40)	0.98(0.38)
β_1	−0.69	−0.70(0.48)	−0.70(0.44)	−0.69(0.50)	−0.69(0.45)	−0.67(0.43)
β_2	−2.96	−2.96(0.58)	−2.91(0.50)	−3.01(0.60)	−2.94(0.55)	−2.84(0.54)
β_3	3.63	3.64(0.65)	3.59(0.54)	3.74(0.68)	3.61(0.60)	3.50(0.59)
σ_1	1.12	1.11	1.03	1.17	1.10	1.02
σ_2	1.18	1.17	1.08	1.22	1.17	1.10

[a] From [17, Tables 1 and 2.]
[b] From [19, Table 3] (uniform prior; posterior medians; SE is the range of the 90% CI divided by 3.3).

Evolution, through 3 experiments on 40 mountain dusky salamanders belonging to 2 different populations. The two populations, called Rough Butt and Whiteside from the names of the locations where they lived, were geographically isolated from one another, thus the aim of the three experiments was to investigate the extent to which Rough Butt and Whiteside would interbreed. In each experiment 40 salamanders were involved: they were divided in 2 groups each composed by 5 Rough Butt males, 5 Rough Butt females, 5 Whiteside males and 5 Whiteside females. Each salamander was paired with 6 partners, 3 belonging to the same population and 3 from the other, then across the 3 experiments 360 pairs were formed.

We consider model A of Karim and Zeger [19], which is a two-level random intercept logistic cross-classified model similar to the one defined by Equation (7). Specifically, the model for salamander mating has two covariates at level 2 and an interaction term:

$$\begin{aligned} \text{logit}(\pi_{i(j_1 j_2)}) &= \beta_0 + \beta_1 x_{1j_1} + \beta_2 x_{2j_2} + \beta_3 x_{1j_1} x_{2j_2} + u_{j_1} + u_{j_2} \\ u_{j_1} &\sim N(0, \sigma^2_{u_{j_1}}) \quad u_{j_2} \sim N(0, \sigma^2_{u_{j_2}}), \end{aligned} \tag{8}$$

where $\pi_{i(j_1 j_2)}$ is the probability of a successful mating between male j_1 (crossing factor 1) and female j_2 (crossing factor 2). The binary covariates x_{1j_1} and x_{2j_2} take the value 1 if the salamander is a Whiteside male or Whiteside female, respectively. Each factor of classification is composed by 60 clusters, within each cluster there are 6 level 1 units (i.e. male–female pairs), within each cell or pair of clusters belonging to the two factors there is a single level 1 unit. The data structure is sketched in Table 8, with 60 rows for the females, 60 columns for the males, and 6 blocks representing groups of salamanders across the three experiments. Therefore, the data are partially cross-classified with no replications within cells (i.e. the cell sample size is one) and 90% of empty cells. This structure is quite different from the structures considered in our simulation study, thus the findings of Section 4 do not necessarily carry over.

In order to specify the priors for the parameters in INLA, we choose the two distributions investigated in our simulation study, namely Ga(0.001, 0.001) and Ga(0.5, 0.003737), for the variance components and, following Karim and Zeger [19], a Normal distribution with zero mean and large variance for each regression coefficient.

Table 9 reports the point estimates of the parameters of model (8) for several estimation methods. We consider MLLA (as in our simulation study), MCEM (taken by Cho and Rabe-Hesketh [17] and Rabe-Hesketh and Skrondal [31] as the gold-standard), AIP with AGQ (the most accurate algorithm in [17]), as well as the Bayesian MCMC estimates of Karim and Zeger [19] obtained with uniform priors for both variance components (results with other priors are reported in Table 3 of Cho and Rabe-Hesketh [17]).

Table 9 shows that, taking MCEM as the benchmark, INLA with prior Ga(0.001, 0.001) has a good performance, similar to that of AIP with AGQ, and better than MLLA and MCMC with uniform prior. On the other hand, INLA with prior INLA Ga(0.5, 0.003737) has a less satisfactory performance, similarly to the results reported in the Supplementary Material of Fong et al. [22]. Overall, INLA

is a valuable method also in the peculiar framework of salamander data, confirming the encouraging findings of simulation studies (Section 4 for cross-classified models and [21] for nested random effects). In addition to accuracy, INLA is considerably faster than MCEM, AIP and MCMC, indeed computation with the salamander data requires only a few seconds.

6. Conclusions

We investigated the performance of INLA for fitting two-level random intercept logistic models with crossed random effects. The investigation exploited a detailed simulation study, entailing a comparison with MLLA, and an application to the classical salamander data, entailing a comparison with several competing methods (MCEM, AIP and MCMC).

In the simulation study we paid attention to scenarios with a small number of clusters, varying degrees of cross-classification, small magnitudes of random effects variances and different prior specifications. Both INLA and MLLA give quite accurate estimates of the fixed effects even in scenarios with a small number of clusters and it turns out that 5 units per cell are enough for a satisfactory performance. On the other hand, estimation of the variance components is challenging: in particular, when the true value is low (0.01) INLA has a severe upward bias if the cell sample size is less than 20. As discussed below, this behaviour is due to the role of the prior distribution. For sizable variance components (0.25) the performance of INLA is satisfactory, improving over MLLA when the prior Ga(0.5, 0.003737) is used. The degree of cross-classification has a little role on the performance of the estimators, even if in settings with small variances INLA yields better results when the data structure is closer to complete cross-classification.

The simulation study showed that INLA and MLLA sometimes fail in estimating the variance components, though in a different way. In fact, MLLA can yield zero estimates: when the variance component is close to zero (0.01) the issue of zero estimates is serious even with 20 observations per cell, whereas for sizable variance component (0.25) the issue of zero estimates becomes negligible with 10 observations per cell. On the other hand, INLA never yields zero estimates, though it occasionally provides aberrant estimates of the variance components in scenarios with one observation per cell. This problem, though relevant only in few scenarios, should be further investigated in order to prevent it.

For low to intermediate values of the variance components, namely $\sigma^2_{u_{j.}} = 0.01$ and $\sigma^2_{u_{j.}} = 0.25$, INLA tends to overestimate these parameters, even after eliminating aberrant estimates ($\hat{\sigma}^2_{u_{j.}} > 2$). This is a consequence of the Bayesian approach with a small sample and a non-informative prior on a parameter bounded to be positive when the true value is close to the bound. In this situation, very low estimates are prevented by the bound, while very large estimates occasionally appear since the flat prior, which is given a high weight in the posterior, does not smooth enough the contribution from the likelihood. For large values of the variance components the bound has little role, indeed for $\sigma^2_{u_{j.}} = 1$ considered in the Supplementary Material we found a negative bias, consistently with the results of Ferkingstad and Rue [29].

The application to the salamander data showed that INLA is competitive with respect to the most efficient algorithms for cross-classified random effects (MCEM, AIP with AGQ, MCMC) since it has similar accuracy, but lower computational time. In general, INLA has the advantage of directly approximating the posterior distribution, thus avoiding the subtle issue of assessing the convergence as in MCMC and AIP.

According to the findings of our simulation study, it is advisable to take care of the specification of the non-informative prior distribution for the variance components, especially when the true values are low and the sample size is small. In general, the prior distribution should be coherent with the plausible range of the size of the random effect. In this regard, the criterion of Fong et al. [22] is helpful as it allows to set the parameters of the prior in order to obtain a given marginal distribution for the random effect. It is worth to note that the difficulty of specifying a suitable prior distribution for the

variances of the random effects, which is common to all Bayesian methods, is alleviated with INLA since its computational speed enables a thorough sensitivity analysis [32,33]. For an overview of the issue of assigning priors in INLA we recommend Section 5 of Rue et al. [25], whereas Simpson et al. [34] provide a discussion of the methods to construct non-subjective priors in Bayesian hierarchical models and propose a widely applicable criterion in this framework.

In summary, INLA is an effective method for fitting logistic models with crossed random effects. It is preferable to MCMC in terms of speed and simplicity of implementation, and it can outperform MLLA depending on the chosen prior distribution for the variance components. In settings with limited sample sizes all methods have difficulties, which may hopefully be attenuated by improvements in the algorithms and in the specification of the prior distribution, see [29] and [35] for recent developments in INLA.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research has been supported by the grant Finite mixture and latent variable models for causal inference and analysis of socio-economic data (FIRB – Futuro in ricerca) funded by the Italian government[RBFR12SHVV].

ORCID

Leonardo Grilli http://orcid.org/0000-0002-3886-7705.
Francesco Innocenti http://orcid.org/0000-0001-6113-8992.

References

[1] Goldstein H. Multilevel cross-classified models. Sociol Methods Res. 1994;22:364–375.
[2] Leckie G. Cross-classified multilevel models – concepts. LEMMA VLEModule 12, 2013. p. 1–60 [cited 2015 Nov 4]. Available from: http://www.bristol.ac.uk/cmm/learning/course.html
[3] Browne WJ, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. Statist Model. 2001;1:103–124.
[4] Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. J Educ Behav Statist. 1994;19:337–350.
[5] Van den Noortgate W, de Boeck P, Meukders M. Cross-classification multilevel logistic models in psychometrics. J Educ Behav Statist. 2003;28:369–386.
[6] Luo W, Kwok O. The impacts of ignoring a crossed factor in analyzing cross-classified data. Multivariate Behav Res. 2009;44:182–212.
[7] Meyers JL, Beretvas SN. The impact of inappropriate modeling of cross-classified data structures. Multivariate Behav Res. 2006;41:473–497.
[8] Shi Y, Leite W, Algina J. The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. Br J Math Stat Psychol. 2010;63:1–15.
[9] Clayton DG, Rasbash J. Estimation in large crossed random effect models by data augmentation. J Roy Statist Soc Ser A. 1999;162:425–436.
[10] Fielding A, Goldstein H. Cross-classified and multiple membership structures in multilevel models: an introduction and review. Research Report RR791. London: Department for Education and Skills; 2006.
[11] Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika. 1991; 78:45–51.
[12] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Amer Statist Assoc. 1993;88:9–25.
[13] Raudenbush SW, Yang M, Yosef M. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. J Comput Graph Statist. 2000;9:141–157.
[14] Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. J Econom. 2005;128:301–323.
[15] Joe H. Accuracy of Laplace approximation for discrete response mixed models. Comput Statist Data Anal. 2008;52:5066–5074.
[16] McCulloch CE. Maximum likelihood variance components estimation for binary data. J Amer Statist Assoc. 1994;89:330–335.

[17] Cho SJ, Rabe-Hesketh S. Alternating imputation posterior estimation of models with crossed random effects. Comput Statist Data Anal. 2011;55:12–25.
[18] Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Anal. 2006;1:473–514.
[19] Karim MR, Zeger SL. Generalized linear models with random effects: salamander mating revisited. Biometrics. 1992;48:631–644.
[20] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J R Stat Soc Ser B. 2009;71:319–392.
[21] Grilli L, Metelli S, Rampichini C. Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. J Statist Comput Simul. 2015;85:2718–2726.
[22] Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010;11:397–412.
[23] McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman & Hall/CRC; 1989.
[24] Martins TG, Simpson D, Lindgren F, et al. Bayesian computing with INLA: new features. Comput Statist Data Anal. 2013;67:68–83.
[25] Rue H, Riebler A, Sørbye SH, et al. Bayesian computing with INLA: a review; 2016. arXiv:1604.00860 [stat.ME].
[26] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. 2006;1:515–534.
[27] Martino S, Rue H. Implementing approximate Bayesian inference using integrated nested Laplace approximation: a manual for the INLA program. Technical report. Trondheim: Norwegian University of Science and Technology; 2008.
[28] Bates DM. lme4: mixed-effects modeling with R. Springer; 2010.
[29] Ferkingstad E, Rue H. Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. Electron J Stat. 2015;9:2706–2731.
[30] Fielding A. Teaching groups as foci for evaluating performance in cost-effectiveness of GCE advanced level provision: some practical methodological innovations. Sch Effect Sch Improv. 2002;13:225–246.
[31] Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using stata. College Station (TX): Stata Press; 2012.
[32] Roos M, Held L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Anal. 2011;6:259–278.
[33] Roos M, Martins TG, Held L, et al. Sensitivity analysis for Bayesian hierarchical models; 2013. arXiv:1312.4797 [stat.ME].
[34] Simpson DP, Rue H, Martins TG, et al. Penalising model component complexity: a principled, practical approach to constructing priors. Trondheim (Norway): Norwegian University of Sciences and Technology; 2014. arxiv:1403.4630 (revised in 2015).
[35] Guo J, Riebler A, Rue H. Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors; 2015. arXiv:1512.06217 [stat.ME].