Stat Comput (2018) 28:1033–1051
https://doi.org/10.1007/s11222-017-9778-y

Markov chain Monte Carlo with the Integrated Nested Laplace Approximation

Virgilio Gómez-Rubio[1] · **Håvard Rue**[2]

Received: 26 January 2017 / Accepted: 22 September 2017 / Published online: 6 October 2017

Abstract The Integrated Nested Laplace Approximation (INLA) has established itself as a widely used method for approximate inference on Bayesian hierarchical models which can be represented as a latent Gaussian model (LGM). INLA is based on producing an accurate approximation to the posterior marginal distributions of the parameters in the model and some other quantities of interest by using repeated approximations to intermediate distributions and integrals that appear in the computation of the posterior marginals. INLA focuses on models whose latent effects are a Gaussian Markov random field. For this reason, we have explored alternative ways of expanding the number of possible models that can be fitted using the INLA methodology. In this paper, we present a novel approach that combines INLA and Markov chain Monte Carlo (MCMC). The aim is to consider a wider range of models that can be fitted with INLA only when some of the parameters of the model have been fixed. We show how new values of these parameters can be drawn from their posterior by using conditional models fitted with INLA and standard MCMC algorithms, such as Metropolis–Hastings. Hence, this will extend the use of INLA to fit models that can be expressed as a conditional LGM. Also, this new approach can be used to build simpler MCMC samplers for complex models as it allows sampling only on a limited number of parameters in the model. We will demonstrate how our approach can extend the class of models that could benefit from INLA, and how the **R-INLA** package will ease its implementation. We will go through simple examples of this new approach before we discuss more advanced applications with datasets taken from the relevant literature. In particular, INLA within MCMC will be used to fit models with Laplace priors in a Bayesian Lasso model, imputation of missing covariates in linear models, fitting spatial econometrics models with complex nonlinear terms in the linear predictor and classification of data with mixture models. Furthermore, in some of the examples we could exploit INLA within MCMC to make joint inference on an ensemble of model parameters.

Keywords Bayesian Lasso · INLA · MCMC · Missing values · Spatial models · Mixture models

1 Introduction

Bayesian inference for complex hierarchical models has almost entirely relied upon computational methods, such as Markov chain Monte Carlo (MCMC, Gilks et al. 1996). Rue et al. (2009) propose a new paradigm for Bayesian inference on hierarchical models that can be represented as latent Gaussian models (LGMs) that focuses on approximating marginal distributions for the parameters in the model. This new approach, the Integrated Nested Laplace Approximation (INLA, henceforth), uses several approximations to the conditional distributions that appear in the integrals needed to obtain the marginal distributions. See Sect. 2 for details.

INLA is implemented as an R package, called **R-INLA**. Model fitting usually takes a fraction of the time as compared to MCMC methods. **R-INLA** provides a simple interface to

✉ Virgilio Gómez-Rubio
Virgilio.Gomez@uclm.es

Håvard Rue
haavard.rue@kaust.edu.sa

[1] Department of Mathematics, School of Industrial Engineering, Universidad de Castilla-La Mancha, Avda España s/n, 02071 Albacete, Spain

[2] CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

implement models and it implements a number of likelihoods (including a few for survival analysis), many types of latent effects (such as random walks or spatial random effects) and a wide range of priors for the model parameters. Fitting models using INLA is restricted, in practice, to the classes of models implemented in the **R-INLA** package.

Despite its many features, INLA cannot easily tackle models with missing values in the covariates, as they are part of the latent effects. Similarly, INLA cannot handle mixture models (Marin et al. 2005) as they are often defined using a weighted combination of different distributions. In addition, INLA focuses on marginal inference of the model parameters, and it is not able to estimate the joint posterior distribution of an arbitrary ensemble of parameters and latent effects. In order to avoid some of the limitations of INLA, several authors have provided ways of fitting other models with INLA by fixing some of the parameters in the model so that conditional models are fitted with **R-INLA**. We have included a brief summary below.

Li et al. (2012) provide an early application of the idea of fitting conditional models on some of the model parameters with **R-INLA**. They developed this idea for a very specific example on a Poisson model with latent Gaussian spatiotemporal effects in which some of the model parameters are fixed at their maximum likelihood estimates, which are then plugged in to the overall model, thus ignoring the uncertainty about these parameters but greatly reducing the dimensionality of the model. However, they do not tackle the problem of fitting the complete model to make inference on all the parameters in the model.

Bivand et al. (2014, 2015) propose an approach to extend the type of models that can be fitted with **R-INLA** and apply their ideas to fit some spatial models. They note how some models can be fitted after conditioning on one or several parameters in the model. For each of these conditional models, **R-INLA** reports the marginal likelihood, which can be combined with a set of priors for the parameters to obtain their posterior distributions. For the remainder of the parameters, their posterior marginal distributions can be obtained by Bayesian model averaging (Hoeting et al. 1999) the family of models obtained with **R-INLA**.

Although Bivand et al. (2014, 2015) focus on some spatial models, their ideas can be applied in many other examples. They apply this to estimate the posterior marginal of the spatial autocorrelation parameter in some models. This parameter is known to be bounded, and computation of its marginal distribution is straightforward because the support of the distribution is a bounded interval.

For the case of unbounded parameters, the previous approach can be applied, but a preliminary search may be required to find the region of high probability of the posterior. For example, the (conditional) maximum log-likelihood plus the log-prior could be maximized to obtain the mode of the posterior marginal. This will mark the center of an interval giving the values of the parameter where the posterior marginal can be evaluated.

In this paper, we will propose a different approach based on Markov chain Monte Carlo techniques. Instead of trying to obtain the posterior marginal of the parameters we condition on, we show how to draw samples from their posterior distribution by combining MCMC techniques and conditional models fitted with **R-INLA**. This new INLA within MCMC algorithm provides several advantages, as described below, and will increase the number of models that can be fitted using INLA and its associated R package **R-INLA**. In particular, models that can be expressed as a conditional LGM could be fitted. The implementation of MCMC algorithms will also be simplified as only the important parameters will be sampled, while the remaining parameters are integrated out with INLA and **R-INLA**.

In the examples provided in Sect. 6, we discuss important applications. Firstly, we have considered an implementation of a Bayesian Lasso in which Laplace priors on the coefficients of the covariates are used. This example can be easily extended to other priors, such as objective, improper or multivariate priors. Next, a linear model with missing covariates is fitted in a way that imputation and model fitting are carried out at the same time. The third example considers a spatial econometrics model with complex nonlinear terms in the linear predictor. The last example focuses on classification of data using mixture models. All these examples have in common that the models involved can be expressed as a conditional LGM and they are susceptible to be fitted using INLA within MCMC.

Hubin and Storvik (2016a) have also effectively combined MCMC and INLA for efficient variable selection and model choice. Vanhatalo et al. (2013) have also successfully combined MCMC with the Laplace approximation to estimate the hyperparameters of a model when fitting Gaussian processes. In particular, they have resorted to MCMC when the space of the hyperparameters was too large for numerical integration (such as central composite design) to work well. Joensuu et al. (2014) have used this approach for the analysis of interval censored data, and Vehtari et al. (2016) give a summary of results when using MCMC and Laplace (and other methods) for leave-one-out cross-validation.

The paper is structured as follows. The Integrated Nested Laplace Approximation is described in Sect. 2. Markov chain Monte Carlo methods are summarized in Sect. 3. Our proposed combination of MCMC and INLA is detailed in Sect. 4. Some simple examples are developed in Sect. 5, and some real applications are provided in Sect. 6. Finally, a discussion and some final remarks are provided in Sect. 7.

2 Integrated Nested Laplace Approximation

We will now describe the types of models that we will be considering and how the INLA method works (for a recent review, see Rue et al. 2017). We will assume that our vector of n observed data $\mathbf{y} = (y_1, \ldots, y_n)$ is observations from a distribution in the exponential family, with y_i having a mean μ_i. We will also assume that a linear predictor on some covariates plus, possibly, other effects can be related to mean μ_i by using an appropriate link function. Note that this linear predictor η_i may be made of linear terms on some covariates plus other types of terms, such as nonlinear functions of the covariates, random effects or spatial random effects. All these terms will define some latent effects $\mathbf{x}$.

The conditional distribution of $\mathbf{y}$ given the linear predictors $\boldsymbol{\eta}$ will depend on a vector of hyperparameters $\boldsymbol{\theta}_1$. Because of the approximation that INLA will use, we will also assume that the vector of latent effects $\mathbf{x}$ will have a distribution that will depend on a vector of hyperparameters $\boldsymbol{\theta}_2$. Altogether, the ensemble of hyperparameters can be represented using a single vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

In addition, we will assume that observations are independent given the values of the latent effects $\mathbf{x}$ and the hyperparameters $\boldsymbol{\theta}$. That is, the likelihood of our model can be written down as

$$\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}). \tag{1}$$

Here, i is indexed over a set of indices $\mathcal{I} \subseteq \{1, \ldots, n\}$ that indicates observed responses. Hence, if the value of y_i is missing, then $i \notin \mathcal{I}$ (but the predictive distribution y_i could be computed once the model is fitted).

Under a Bayesian framework, the aim is to compute the posterior distribution of the model parameters and hyperparameters using Bayes' rule. This can be stated as

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta}). \tag{2}$$

Here, $\pi(\mathbf{x}, \boldsymbol{\theta})$ is the prior distribution of the latent effects and the vector of hyperparameters. As the latent effects $\mathbf{x}$ have a distribution that depends on $\boldsymbol{\theta}_2$, it is convenient to write this prior distribution as $\pi(\mathbf{x}, \boldsymbol{\theta}) = \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Altogether, the posterior distribution of the latent effects and hyperparameters can be expressed as

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}). \tag{3}$$

The joint posterior distribution, as presented on the left-hand side in Eq. (3), is seldom available in a closed form. For this reason, several estimation methods and approximations have been developed over the years.

Rue et al. (2009) have provided approximations based on the Laplace approximation to estimate the marginals of all latent effects and hyperparameters in the model. They develop this approximation for the family of latent Gaussian Markov random fields models. In this case, the vector of latent effects is a Gaussian Markov random field (GMRF). This GMRF will have zero mean (without loss of generality as any fixed mean can be introduced as an offset in the linear predictor) and precision matrix $\mathbf{Q}(\boldsymbol{\theta})$.

Assuming that the latent effects are a GMRF will let us develop Eq. (3) further. In particular, the posterior distribution of the latent effects $\mathbf{x}$ and the vector of hyperparameters $\boldsymbol{\theta}$ can be written as

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left\{ -\frac{1}{2}\mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i \in \mathcal{I}} \log\left(\pi(y_i|x_i, \boldsymbol{\theta})\right) \right\}. \tag{4}$$

With INLA, the aim is not the joint posterior distribution $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, but the marginal distributions of latent effects and hyperparameters. That is, $\pi(x_j|\mathbf{y})$ and $\pi(\theta_k|\mathbf{y})$, where indices j and k will take different ranges of values depending on the number of latent effects and hyperparameters.

Before computing these marginal distributions, INLA will obtain an approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$, $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. This approximation will later be used to compute an approximation to marginals $\pi(x_j|\mathbf{y})$. Given that the marginal can be written down as

$$\pi(x_j|\mathbf{y}) = \int \pi(x_j|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}, \tag{5}$$

the approximation is as follows:

$$\tilde{\pi}(x_j|\mathbf{y}) = \sum_g \tilde{\pi}(x_j|\boldsymbol{\theta}_g, \mathbf{y}) \times \tilde{\pi}(\boldsymbol{\theta}_g|\mathbf{y}) \times \Delta_g. \tag{6}$$

Here, $\tilde{\pi}(x_j|\boldsymbol{\theta}_g, \mathbf{y})$ is an approximation to $\pi(x_j|\boldsymbol{\theta}_g, \mathbf{y})$, which can be obtained using different methods (see, Rue et al. 2009, for details). In addition, $\boldsymbol{\theta}_g$ refers to an ensemble of hyperparameters that take values on a grid (for example), with weights Δ_g.

INLA is a general approximation that can be applied to a large number of models. An implementation for the R programming language is available in the **R-INLA** package at www.r-inla.org, which provides easy access to model fitting. This includes a simple interface to choose the likelihood, latent effects and priors. The implementation provided by **R-INLA** includes the computation of other quantities of interest. The marginal likelihood $\pi(\mathbf{y})$ is approximated, and it can be used for model choice. As described in Rue et al.

(2009), the approximation to the marginal likelihood provided by INLA is computed as

$$\tilde{\pi}(\mathbf{y}) = \int \left. \frac{\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y})}{\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}.$$

Here, $\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is a Gaussian approximation to $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and $\mathbf{x}^*(\boldsymbol{\theta})$ is the posterior mode of $\mathbf{x}$ for a given value of $\boldsymbol{\theta}$. This approximation is reliable when the posterior of $\boldsymbol{\theta}$ is unimodal, as is often the case for latent Gaussian models. Furthermore, Hubin and Storvik (2016b) demonstrate that this approximation is accurate for a wide range of models.

Other options for model choice and assessment include the Deviance Information Criterion (DIC, Spiegelhalter et al. 2002) and the Conditional Predictive Ordinate (CPO, Pettit 1990). Other features in the **R-INLA** package include the use of several likelihoods in the same model and the computation of the posterior marginal of a certain linear combination of the latent effects and others (see, Martins et al. 2013, for a summary of recent additions to the software).

3 Markov chain Monte Carlo

In the previous section, we have reviewed how INLA computes approximations of the marginal distributions of the latent effects and hyperparameters. Instead of focusing on an approximation to the marginals, MCMC methods could be used to obtain a sample from the joint posterior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. To simplify the notation, we will denote the vector of latent effects and hyperparameters by $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\theta})$. Hence, the aim now is to estimate $\pi(\boldsymbol{z}|\boldsymbol{y})$ or, if we are only interested in the posterior marginals, $\pi(z_i|\boldsymbol{y})$.

Several methods to estimate or approximate the posterior distribution have been developed over the years (Gilks et al. 1996). In the case of MCMC, the interest is in obtaining a Markov chain whose limiting distribution is $\pi(\boldsymbol{z}|\boldsymbol{y})$. We will not provide a summary of MCMC methods here, and the reader is referred to Gilks et al. (1996) for a detailed description.

The values generated using MCMC are (correlated) draws from $\pi(\boldsymbol{z}|\boldsymbol{y})$ and, hence, can be used to estimate quantities of interest. For example, if we are interested in marginal inference on z_i, the posterior mean can be estimated using the empirical mean of $\left\{z_i^{(j)}\right\}_{j=1}^{N}$. Similarly, estimates of the posterior expected value of any function on the parameters $f(\boldsymbol{z})$ can be obtained using that

$$E[f(\boldsymbol{z})|\boldsymbol{y}] \simeq \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{z}^{(j)}). \tag{7}$$

Multivariate inference is possible by using the multivariate nature of vector $\boldsymbol{z}^{(j)}$. For example, the posterior covariance between parameters z_k and z_l could be computed by considering samples $\left\{(z_k^{(j)}, z_l^{(j)})\right\}_{j=1}^{N}$.

3.1 The Metropolis–Hastings algorithm

This algorithm was first proposed by Metropolis et al. (1953) and Hastings (1970). The Markov chain is generated by proposing new moves according to a proposal distribution $q(\cdot|\cdot)$. A new point $\boldsymbol{z}^*$ is accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{z}^*|\boldsymbol{y})q(\boldsymbol{z}^{(j)}|\boldsymbol{z}^*)}{\pi(\boldsymbol{z}^{(j)}|\boldsymbol{y})q(\boldsymbol{z}^*|\boldsymbol{z}^{(j)})}\right\}. \tag{8}$$

If the proposed point is accepted, then $\boldsymbol{z}^{(j+1)}$ will become $\boldsymbol{z}^*$. Otherwise, $\boldsymbol{z}^{(j+1)}$ will be equal to $\boldsymbol{z}^{(j)}$. In the previous acceptance probability, the posterior probabilities of the current point and the proposed new point appear as $\pi(\boldsymbol{z}^{(j)}|\boldsymbol{y})$ and $\pi(\boldsymbol{z}^*|\boldsymbol{y})$, respectively. These two probabilities are unknown, in principle, but using Bayes' rule they can be rewritten as

$$\pi(\boldsymbol{z}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{z})\pi(\boldsymbol{z})}{\pi(\boldsymbol{y})}. \tag{9}$$

Hence, the acceptance probability α can be rewritten as

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{y}|\boldsymbol{z}^*)\pi(\boldsymbol{z}^*)q(\boldsymbol{z}^{(j)}|\boldsymbol{z}^*)}{\pi(\boldsymbol{y}|\boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(j)})}\right\}. \tag{10}$$

This is easier to compute as the acceptance probability depends on known quantities, such as the likelihood $\pi(\boldsymbol{y}|\boldsymbol{z})$, the prior on the parameters $\pi(\boldsymbol{z})$ and the proposal distribution. Note that the term $\pi(\boldsymbol{y})$ that appears in Eq. (9) is unknown, but that it cancels out as it appears both in the numerator and denominator.

In Eq. (10), we have described the move to sample from the joint ensemble of model parameters. However, this can be applied to individual parameters one at a time, so that acceptance probabilities will be

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{y}|z_i^*)\pi(z_i^*)q(z_i^{(j)}|z_i^*)}{\pi(\boldsymbol{y}|z_i^{(j)})\pi(z_i^{(j)})q(z_i^*|z_i^{(j)})}\right\}. \tag{11}$$

However, this expression is seldom used because of the difficulty in computing $\pi(\boldsymbol{y}|z_i)$.

4 INLA within MCMC

In this section, we will describe how INLA and MCMC can be combined to fit complex Bayesian hierarchical models. In

principle, we will assume that the model cannot be fitted with **R-INLA** unless some of the latent effects or hyperparameters in the model are fixed. This set of parameters is denoted by z_c so that the full ensemble of latent effects and hyperparameters is $z = (z_c, z_{-c})$. Here z_{-c} is used to denote all the parameters in z that are not in z_c. The posterior distribution of z can be split as

$$\pi(z|y) \propto \pi(y|z_{-c}, z_c)\pi(z_{-c}|z_c)\pi(z_c). \tag{12}$$

Note that integrating over z_{-c} conditional on z_c in the previous expression, we obtain

$$\pi(z_c|y) \propto \pi(y|z_c)\pi(z_c). \tag{13}$$

This means that conditional models (on z_c) can still be fitted with **R-INLA**, i.e., we can obtain marginals of the parameters in z_{-c} given z_c. The conditional posterior marginals for the k-th element in vector z_{-c} will be denoted by $\pi(z_{-c,k}|z_c, y)$. Also, the conditional marginal likelihood $\pi(y|z_c)$ can be easily computed with **R-INLA**.

4.1 Metropolis–Hastings with INLA

We will now discuss how to implement the Metropolis–Hastings algorithm to estimate the posterior marginal of z_c. Note that this is a multivariate distribution and that we will use block updating in the Metropolis–Hastings algorithm. This means that at each step a new value for the ensemble z_c is proposed and these values will be accepted or rejected altogether.

Say that we start from an initial point $z_c^{(0)}$; then, we can use the Metropolis–Hastings algorithm to obtain a sample from the posterior of z_c.

We will draw a new proposal value for z_c, z_c^*, using the proposal distribution $q(\cdot|\cdot)$. The acceptance probability, shown in Eq. (10), becomes now:

$$\alpha = \min\left\{1, \frac{\pi(y|z_c^*)\pi(z_c^*)q(z_c^{(j)}|z_c^*)}{\pi(y|z_c^{(j)})\pi(z_c^{(j)})q(z_c^*|z_c^{(j)})}\right\}. \tag{14}$$

Note that $\pi(y|z_c^{(j)})$ and $\pi(y|z_c^*)$ are the conditional marginal likelihoods on $z_c^{(j)}$ and z_c^*, respectively. All these quantities can be obtained by fitting a model with **R-INLA** with the values of z_c set to $z_c^{(j)}$ and z_c^*. Hence, at each step of the Metropolis–Hastings algorithm only a model conditional on the proposal needs to be fitted.

Furthermore, $\pi(z_c^{(j)})$ and $\pi(z_c^*)$ are the priors of z_c evaluated at $z_c^{(j)}$ and z_c^*, respectively, and they can be easily computed as the priors are known in the model. Values $q(z_c^{(j)}|z_c^*)$ and $q(z_c^*|z_c^{(j)})$ can also be computed as the proposal distribution is known. If the proposed point is accepted, then $z_c^{(j+1)} = z_c^*$, and $z_c^{(j+1)} = z_c^{(j)}$ otherwise. Hence, the Metropolis–Hastings algorithm can be implemented to obtain a sample from the (joint) posterior distribution of z_c. The marginal distributions of the elements of z_c can be easily obtained as well.

Regarding the marginals of $z_{-c,k}$, it is worth noting that at step j of the Metropolis–Hastings algorithm a conditional marginal distribution on $z_c^{(j)}$ (and the data y) is obtained: $\pi(z_{-c,k}|z_c^{(j)}, y)$. The posterior marginal can be approximated by integrating over z_c as follows:

$$\begin{aligned}\pi(z_{-c,k}|y) &= \int \pi(z_{-c,k}|z_c, y)\pi(z_c|y)\mathrm{d}z_c \\ &\simeq \frac{1}{N}\sum_{j=1}^{N}\pi(z_{-c,k}|z_c^{(j)}, y),\end{aligned} \tag{15}$$

where N is the number of samples of the posterior distribution of z_c. That is, the posterior marginal of $z_{-c,k}$ can be obtained by Bayesian model averaging (BMA, see Hoeting et al. 1999, for a summary) the conditional marginals obtained at each iteration of the Metropolis–Hastings algorithm.

Given an approximation to the posterior marginal of $z_{-c,k}$ computed using BMA, $\tilde{\pi}_{\mathrm{BMA}}(\cdot|y)$, point estimates and other quantities of interest can be estimated numerically. This is implemented in functions `inla.emarginal` (for the posterior expected value) and `inla.zmarginal` (for several posterior statistics) available in the **R-INLA** package. The numerical approximation is based on using Simpson's rule to approximate the different integrals that need to be computed. For example, the approximation to the posterior mean of $z_{-c,k}$ is

$$E[z_{-c,k}|y] \simeq \int z \cdot \tilde{\pi}_{\mathrm{BMA}}(z|y)\mathrm{d}z,$$

where the integral on the right-hand side is approximated using Simpson's rule.

4.2 Effect of approximating the marginal likelihood

So far, we have ignored the fact that the conditional marginal likelihood $\pi(y|z_c)$ used in the acceptance probability α is actually an approximation. In this section, we will discuss how this approximation will impact the validity of the inference.

The situations where a Metropolis–Hastings algorithm has inexact acceptance probabilities are often called pseudo-marginal MCMC algorithms (Beaumont 2003). These were first introduced in the context of statistical genetics where the likelihood in the acceptance probability is approximated using importance sampling. Andrieu and Roberts (2003) provided a more general justification of the pseudo-marginal

MCMC algorithm, whose properties are further studied in Sherlock et al. (2015) and Medina-Aguayo et al. (2016). These results show that if the (random) numerator and denominator of the acceptance probability are unbiased, then the Markov chain will still have as stationary distribution the posterior distribution of the model parameters.

In our case, the error in the acceptance rate is coming from a deterministic estimate of the conditional marginal likelihood; hence, the framework of pseudo-marginal MCMC does not apply. However, since it is deterministic, our MCMC chain will converge to *a* stationary distribution. This limiting distribution will be

$$\tilde{\pi}(z_c|\boldsymbol{y}) \propto \pi(z_c)\tilde{\pi}(\boldsymbol{y}|z_c), \tag{16}$$

where the '$\sim$' indicates an approximation. **R-INLA** returns an approximation to the conditional marginal likelihood term, which implies an approximation to $\pi(z_c|\boldsymbol{y})$. This raises the question as to how good this approximation performs. To evaluate this, we have to rely on asymptotic results, heuristics and numerical experience.

The conditional marginal likelihood estimate returned from **R-INLA** is based on numerical integration and uses a sequence of Laplace approximations (Rue et al. 2009, 2017). This estimate is more accurate than the classical estimate using one Laplace approximation. The Laplace approximation has, with classical assumptions, relative error $\mathcal{O}(n^{-1})$ (Tierney and Kadane 1986), where n is the number of replications in the observations. For our purpose, this error estimate is sufficient, as it demonstrates that

$$\frac{\tilde{\pi}(z_c|\boldsymbol{y})}{\pi(z_c|\boldsymbol{y})} \propto \frac{\tilde{\pi}(\boldsymbol{y}|z_c)}{\pi(\boldsymbol{y}|z_c)} = 1 + \mathcal{O}(n^{-1}) \tag{17}$$

for plausible values of z_c. However, as discussed by Rue et al. (2009, 2017), the classical assumptions are rarely met in practice due to 'random effects,' smoothing, etc. Precise error estimates under realistic assumptions are difficult to obtain; see Rue et al. (2017) for a more detailed discussion of this issue.

Hubin and Storvik (2016b) have studied empirically the properties and accuracy of the marginal likelihood estimate provided by INLA for a wide range of latent Gaussian models. They have compared the estimates with those obtained using MCMC, and in all their cases the approximates of the marginal likelihood provided by INLA were very accurate. For this reason, we believe that the approximate stationary distribution $\tilde{\pi}(z_c|\boldsymbol{y})$ should be close to the true one, without being able to quantify this error in more detail.

Although the error in Eq. (17) is pointwise, we do expect the error would be smooth in z_c. This is particularly important, as in most cases we are interested in the univariate marginals of $\tilde{\pi}(z_c|\boldsymbol{y})$. We expect that these marginals will typically have less error as the influence of the approximation error will be averaged out integrating out all the other components. A final renormalization would also remove any constant offset in the error.

Additionally, we will validate the approximation error in a simulation study in Sect. 5 where we fit various models using INLA, MCMC and INLA within MCMC and very similar posterior distributions are obtained. Furthermore, the real applications in Sect. 6 also support that the approximations to the marginal likelihood are accurate.

4.3 Some remarks

Common sense is still not out of fashion; hence, there is an implicit assumption that our INLA within MCMC approach should be only for models for which it is reasonable to use the INLA approach to do the inference for the conditional model. The procedure that we have just shown will allow INLA to be used together with the Metropolis–Hastings algorithm (and, possibly, other MCMC methods) to obtain the posterior distribution (and marginals) of z_c and the posterior marginals of the elements in z_{-c}. Hence, this will allow INLA to be used to fit models not implemented in the **R-INLA** package as well as providing other options for model fitting that we summarize here. Note also that this means that multivariate inference on the ensemble of parameters z_c will be possible as we will obtain samples from their joint posterior.

Furthermore, the Metropolis–Hastings algorithm will allow any choice of the priors on the set of parameters z_c. This is an advantage (as shown in the example in Sect. 6.1) of combining MCMC and INLA because priors that are not implemented in **R-INLA** can be used in the model. In particular, improper flat priors, multivariate priors and objective priors can be used.

The framework of conditional LGMs that we now can fit using our new approach is quite rich. It includes models with missing covariates that are imputed at each step of the Metropolis–Hastings algorithm (see example in Sect. 6.2), models with complex nonlinear effects in the linear predictor (see example in Sect. 6.3) or models that have a mixture of effects in the linear predictor (Bivand et al. 2015).

5 Simulation study

In this section, we develop simple examples to illustrate the method proposed in the previous sections, and we investigate how this new approach works in practice.

5.1 Bivariate linear regression

The first example is based on a linear regression with two covariates. Our aim is to use our proposed method to obtain

the posterior distribution of the coefficients of the two covariates and then compare the estimated marginals to the results obtained when the full model is fitted with MCMC and INLA.

The simulated dataset contains 100 observations of a response variable $\boldsymbol{y}$ and covariates $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The model used to generate the data is a typical linear regression, i.e.,

$$y_i = \alpha + \beta_1 u_{1i} + \beta_2 u_{2i} + \varepsilon_i;\ i = 1, \ldots, 100. \tag{18}$$

Here, ε_i is a Gaussian error term with zero mean and precision τ. The dataset has been simulated using $\alpha = 3$, $\beta_1 = 2$, $\beta_2 = -2$ and $\tau = 1$. Covariates u_{1i} and u_{2i} have also been simulated using a uniform distribution between 0 and 1 in both cases.

This model can be easily fitted using **R-INLA**, but we have chosen to condition on $\boldsymbol{\beta}$ to show how INLA within MCMC works. Given that we are using a Gaussian model, inference is exact in this case (up to integration error). For this reason, we can compare the marginal distributions of β_1 and β_2 provided by INLA and the ones obtained with our combined approach. Note that the Metropolis–Hastings algorithm will provide the joint posterior distribution of $\boldsymbol{\beta} = (\beta_1, \beta_2)$ that can be used to obtain the posterior marginals of β_1 and β_2. Furthermore, we can also compare the marginals of α and τ that will be estimated by averaging the different conditional marginals obtained in the Metropolis–Hastings steps.

In order to implement the Metropolis–Hastings algorithm to obtain a sample from $\pi(\boldsymbol{\beta}|\boldsymbol{y})$, we have chosen a starting point of $\boldsymbol{\beta}^{(0)} = (0, 0)$. The proposal distribution to obtain a candidate $\boldsymbol{\beta}^{(t+1)}$ at iteration t has been a bivariate Gaussian kernel centered at $\boldsymbol{\beta}^{(t)}$ with diagonal variance–covariance matrix with values $1/0.75^2$ in the diagonal as this provided a reasonable acceptance rate. The prior distribution of $\boldsymbol{\beta}$ has been the product of two Gaussian distributions with zero mean and precision 0.001 because these are the default priors for linear effects in **R-INLA**. Furthermore, α is assigned a Gaussian prior with zero mean and zero precision and τ a Gamma prior with parameters 1 and 5e−05 (the default priors in **R-INLA**). The prior on α used with **rjags** (Plummer 2016) has been a uniform between -1000 and 1000 to provide a very vague prior as **R-INLA** does.

Figure 1 summarizes the INLA within MCMC algorithm for this particular problem. Figure 2 shows a summary of the results. Given that both covariates are independent, their coefficients should show small correlation and this can clearly be seen in the plot of the joint posterior distribution of $\boldsymbol{\beta}$. Also, it can be seen how the marginals obtained with INLA within MCMC for β_1 and β_2 match those obtained with INLA and MCMC. In addition, we have included the estimates of the posterior marginals of the intercept α and the precision τ. When using INLA within MCMC, these are obtained by Bayesian model averaging over the fitted models at every step of the Metropolis–Hastings algorithm, while when computed with **R-INLA** these are obtained by using INLA alone. The three estimation methods provide very similar posterior distributions of the posterior marginals of the intercept and the precision, which again confirms the accuracy of INLA within MCMC.

5.2 Missing covariates

In the next example, we will discuss the case of missing covariates. In this example, we will consider a linear regression with covariate $\boldsymbol{u}_1$ only and we will assume that a number of values of the covariates are missing. The aim is to include the imputation of these variables into the model, so that the output is a marginal distribution of the missing values. We will not discuss here the different frameworks under which the values have gone missing, but this is something that should be taken into account in the model. In particular, we have removed the values of nine covariates, which is almost 10% of our data and summary plots can nicely be arranged in a three by three matrix of figures (as shown in Fig. 3). Hence, in this case the missingness mechanism is of the type missing completely at random (Little and Rubin 2002).

Now, we will treat the missing values as if they were parameters. We will use a block updating scheme as we can have a large number of missing covariates. The transition kernel will be a multivariate Gaussian with diagonal variance–covariance. The mean and variance for all values are the mean and variance of the observed covariates, respectively. The prior distribution is also a multivariate Gaussian, but now with zero mean and diagonal variance–covariance matrix with entries four times the variance of a uniform random variable in the unit interval (the one used to simulate the covariates). This is done so that the prior information is small compared to the information provided by the observed covariates.

Figure 3 shows the posterior marginals obtained from the samples. As it can be seen, most of them are centered at the actual values removed from the model. Note that this time the model with missing covariates cannot be fitted with **R-INLA** so that we can only compare the marginals to those obtained with MCMC. In all cases, the marginals obtained with INLA within MCMC and full MCMC are very similar.

5.3 Poisson regression

In this example, we consider a Poisson regression with two covariates:

$$y_i \sim Po(\mu_i);\ \log(\mu_i) = \alpha + \beta_1 u_{1i} + \beta_2 u_{2i};\quad i = 1, \ldots, 100. \tag{19}$$

INLA within MCMC algorithm

- Set $(\beta_1, \beta_2)^{(1)} = \boldsymbol{\beta}^{(1)} = (0, 0)$
- Fit model with INLA conditional on $\boldsymbol{\beta}^{(1)}$ to obtain:
  - $\tilde{\pi}(\mathbf{y}|\boldsymbol{\beta}^{(1)})$
  - $\tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^{(1)})$ and $\tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^{(1)})$
- For j from 2 to N, do:
  - Sample $\boldsymbol{\beta}^*$ from $q(\cdot|\boldsymbol{\beta}^{(j-1)})$
  - Fit model with INLA conditional on $\boldsymbol{\beta}^*$ to obtain:
    - $\tilde{\pi}(\mathbf{y}|\boldsymbol{\beta}^*)$
    - $\tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^*)$ and $\tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^*)$
  - Compute $\pi(\boldsymbol{\beta}^*)$, $q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(j)})$ and $q(\boldsymbol{\beta}^{(j)}|\boldsymbol{\beta}^*)$
  - Compute (approximate) log-acceptance probability:

$$\log(\tilde{\alpha}) = \log(\tilde{\pi}(\mathbf{y}|\boldsymbol{\beta}^*)) + \log(\pi(\boldsymbol{\beta}^*)) + \log(q(\boldsymbol{\beta}^{(j)}|\boldsymbol{\beta}^*)) - \\ - \log(\tilde{\pi}(\mathbf{y}|\boldsymbol{\beta}^{(j)})) - \log(\pi(\boldsymbol{\beta}^{(j)})) - \log(q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(j)}))$$

  - Sample u from uniform (0, 1)
  - If $\log(u) < \log(\tilde{\alpha})$, then
    - $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^*$
    - $\tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^{(j+1)}) = \tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^*)$
    - $\tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^{(j+1)}) = \tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^*)$

    else
    - $\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)}$
    - $\tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^{(j+1)}) = \tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^{(j)})$
    - $\tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^{(j+1)}) = \tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^{(j)})$
- Perform thinning to reduce autocorrelation: $\mathcal{N} \subseteq \{1, \ldots, N\}$
- Compute $\tilde{\pi}(\boldsymbol{\beta}|\mathbf{y})$ from $\{\boldsymbol{\beta}^{(i)}\}_{i \in \mathcal{N}}$ using *bivariate* kernel density estimation
- Compute $\tilde{\pi}(\beta_j|\mathbf{y})$ from $\{\beta_j^{(i)}\}_{i \in \mathcal{N}}$, $j = 1, 2$ using *univariate* kernel density estimation
- Compute $\tilde{\pi}(\alpha|\mathbf{y})$ as

$$\tilde{\pi}(\alpha|\mathbf{y}) = \sum_{i \in \mathcal{N}} \frac{1}{|\mathcal{N}|} \tilde{\pi}(\alpha|\mathbf{y}, \boldsymbol{\beta}^{(i)})$$

- Compute $\tilde{\pi}(\tau|\mathbf{y})$ as

$$\tilde{\pi}(\tau|\mathbf{y}) = \sum_{i \in \mathcal{N}} \frac{1}{|\mathcal{N}|} \tilde{\pi}(\tau|\mathbf{y}, \boldsymbol{\beta}^{(i)})$$

Notation

- $q(\cdot|\boldsymbol{\mu})$: bivariate Normal with mean $\boldsymbol{\mu}$ and precision matrix Σ:

$$\Sigma = \begin{bmatrix} 0.75^2 & 0 \\ 0 & 0.75^2 \end{bmatrix}.$$

- $q(\boldsymbol{\beta}|\boldsymbol{\mu})$: density of $q(\cdot|\boldsymbol{\mu})$ evaluated at $\boldsymbol{\beta}$.
- $\pi(\boldsymbol{\beta})$: prior density of the coefficients evaluated at $\boldsymbol{\beta}$.
- $\pi(\mathbf{y}|\boldsymbol{\beta})$: marginal likelihood of a model conditional on $\boldsymbol{\beta}$.
- $\pi(\cdot|\mathbf{y}, \boldsymbol{\beta})$: posterior marginal of a parameter conditional on $\boldsymbol{\beta}$.
- $|\mathcal{N}|$: cardinal of set $\mathcal{N}$.

Fig. 1 Algorithm of INLA within MCMC for the bivariate linear regression example

The values of the parameters used to simulate the dataset are $\alpha = 0.5$, $\beta_1 = 2$ and $\beta_2 = -2$. Covariates have been simulated as in the first example, using a uniform distribution between 0 and 1.

As in Sect. 5.1, our purpose is to estimate the joint posterior distribution of (β_1, β_2). The prior distributions on $\boldsymbol{\beta}$ and α used now are the same as in the first example in Sect. 5.1. Similarly, the posterior marginal of α is obtained by com-

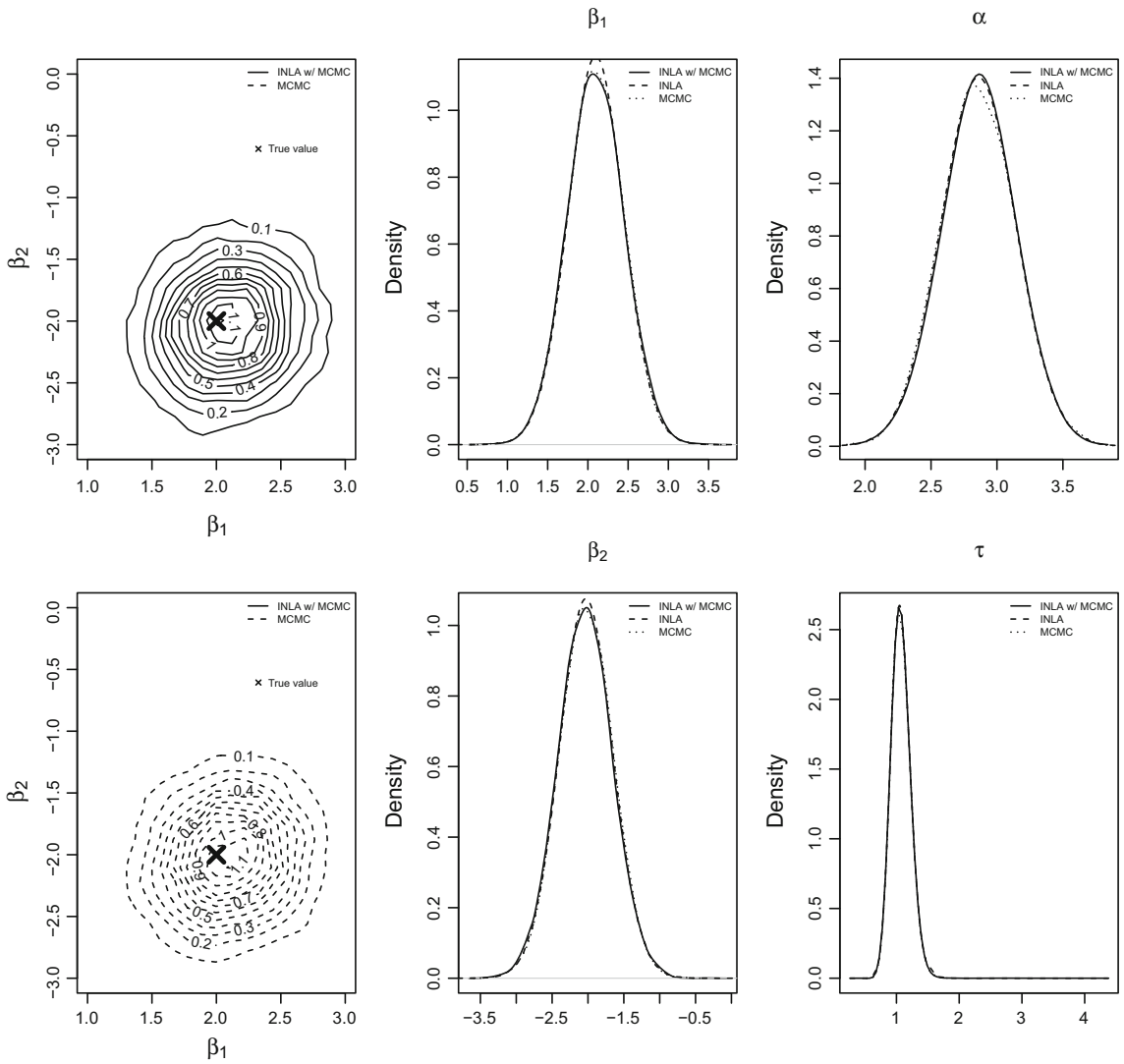

Fig. 2 Summary of results of model fitting combining INLA and MCMC in the bivariate case. Joint posterior distribution of (β_1, β_2) and posterior marginals of the model parameters

bining the different conditional marginals obtained at the different steps of the Metropolis–Hastings algorithm.

Figure 4 shows the estimates of the marginal distributions of the three parameters in the model, together with the joint posterior distribution of β_1 and β_2. In all cases, there is very good agreement between the estimates obtained with INLA, MCMC and INLA within MCMC of the marginals of the parameters in the model.

5.4 Computational gain

In terms of computational gain, the main advantage of INLA within MCMC is the ease to implement new and complex models to fit the data. This will be better illustrated in Sect. 6, where a few more examples on diverse topics have been included. In general, our approach allows us to focus on a reduced number of parameters because inference on the remainder of the parameters is already done by INLA and Bayesian model averaging.

In addition, effective sample size appears to be better with INLA within MCMC. We have compared the effective sample sizes obtained with INLA within MCMC and MCMC by computing the effective sample size for each variable given a fixed number of iterations. In order to make inference, the minimum effective sample size will give us a lower bound on the effective sample size of all the parameters involved.

The effective sample size has been computed using function `effectiveSize` in package **coda** (Plummer et al. 2006). Given an MCMC sample $\boldsymbol{x} = (x_1, \ldots, x_N)$ of length N, the effective sample size ESS is computed as

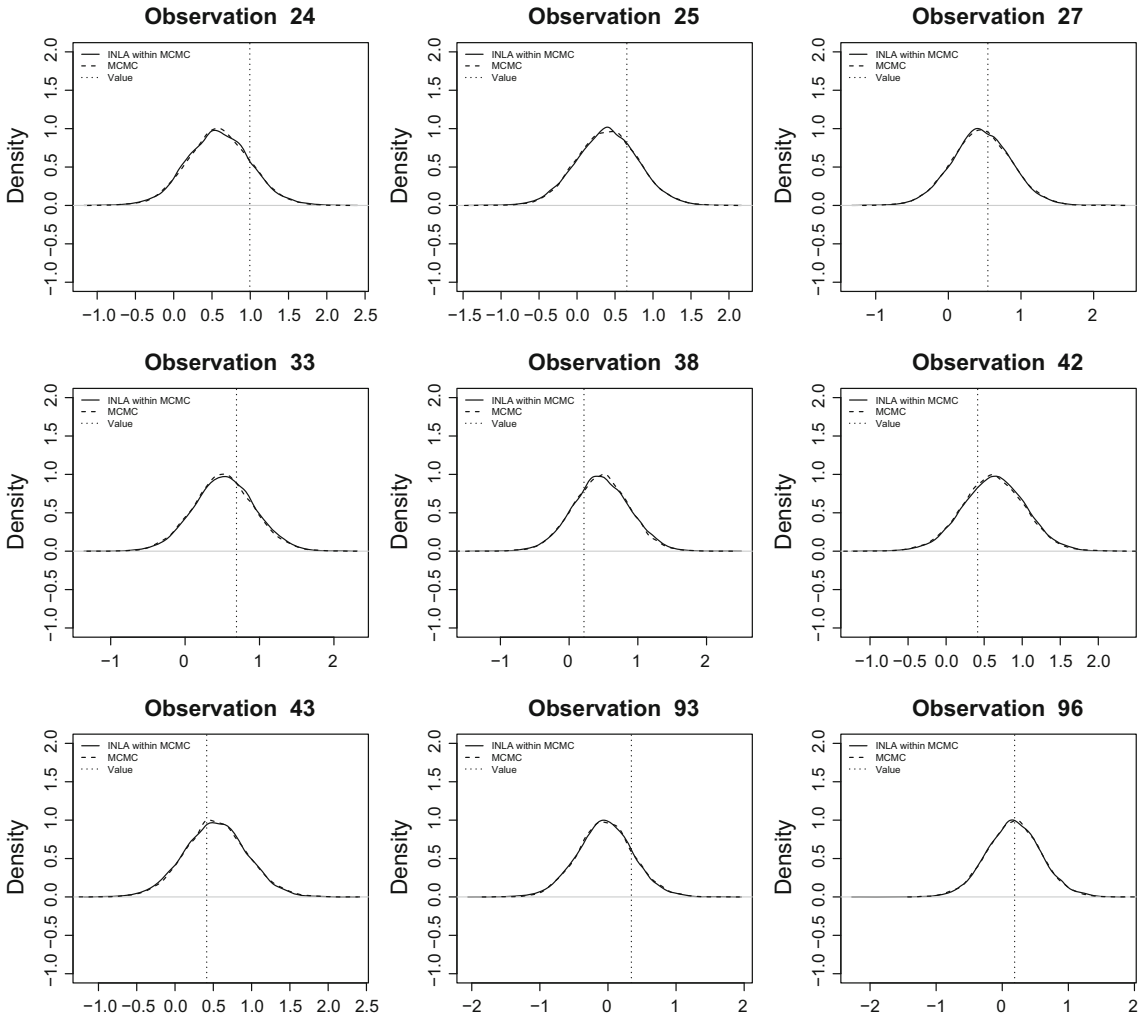

Fig. 3 Posterior marginals of the missing values in the covariates obtained by fitting a model with INLA within MCMC, and MCMC

$$\text{ESS} = \frac{S^2}{S_0^2/N},$$

where S^2 is the sample variance of $\boldsymbol{x}$ and S_0^2 is the estimated spectral density at frequency zero obtained by fitting an autoregressive model to $\boldsymbol{x}$ (computed using function `spectrum0.ar` in package **coda**). It is worth noting that S_0^2/N is an estimate of the variance of the sample mean of $\boldsymbol{x}$.

Figure 5 shows the minimum effective sample size for the examples on linear and Poisson regression. As it can be seen, INLA within MCMC provides higher effective sample sizes globally than MCMC for these particular examples. This means that our approach would require less iterations to achieve the same number of independent observations from the posterior.

However, we are not claiming that INLA within MCMC is uniformly better than MCMC. This gain in effective sample size can occur, for example, because of the block updating strategy that we use or the proposal distributions chosen for a particular problem. In this regard, it should be mentioned that **rjags** is essentially based on Gibbs sampling, so the two implementations compared are very different and difficult to compare directly.

Finally, in terms of actual computing time, it is difficult to make a fair comparison because of the differences in the actual implementations of the different approaches. MCMC with **rjags** is very fast in these examples. However, **R-INLA** is very fast to fit each conditional model, but there is a considerable overhead because of the temporary files that it creates each time a model is run. A tighter integration could be achieved by linking the part of the model that does MCMC to the C library GMRFlib, upon which the **R-INLA** package is built.

6 Applications

In this section, we will focus on some real life applications that provide a more realistic test of this methodology. In

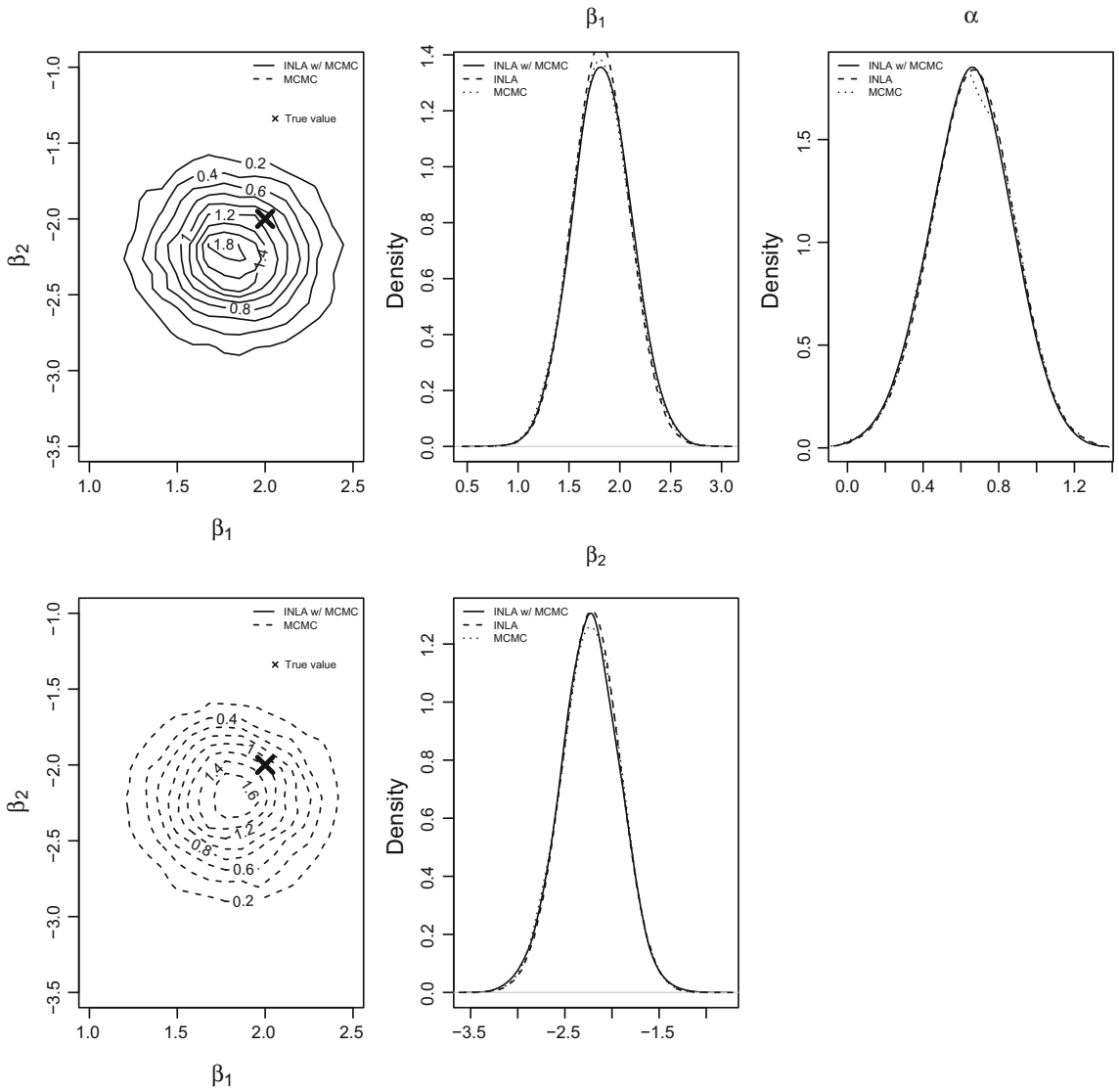

Fig. 4 Summary of results of model fitting combining INLA and MCMC for the Poisson regression example. Joint posterior distribution of (β_1, β_2) (left column) and posterior marginals of the model parameters

all the examples, we have run INLA within MCMC and MCMC for a total of 100,500 simulations and discarded the first 500. Then, we applied a thinning to keep one in ten iterations, to obtain a final chain of 10,000 samples. This includes samples from the missing observations and parameters of the fitted models. To fit the model using MCMC alone, we have used **rjags** with the same number of iterations and thinning. The implementation of INLA within MCMC is available as a new function `INLAMH()` that has been added to package **INLABMA** (Bivand et al. 2015). Bayesian model averaging will be done with the existing functions in the same package. Furthermore, the R code to reproduce the examples (and the simulation study) is freely available in a github repository (https://github.com/becarioprecario/INLAMCMC_examples). In order to test the code, users may want to reduce the number of iterations used in the examples so that the simulations finish in a shorter period of time.

6.1 Bayesian Lasso

The Lasso (Tibshirani 1996) is a popular regression and variable selection method. It has the nice property of providing coefficient estimates that are exactly zero, and hence, it performs model fitting and variable selection at the same time. For a linear model with a Gaussian likelihood, the Lasso is trying to estimate the regression coefficients by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

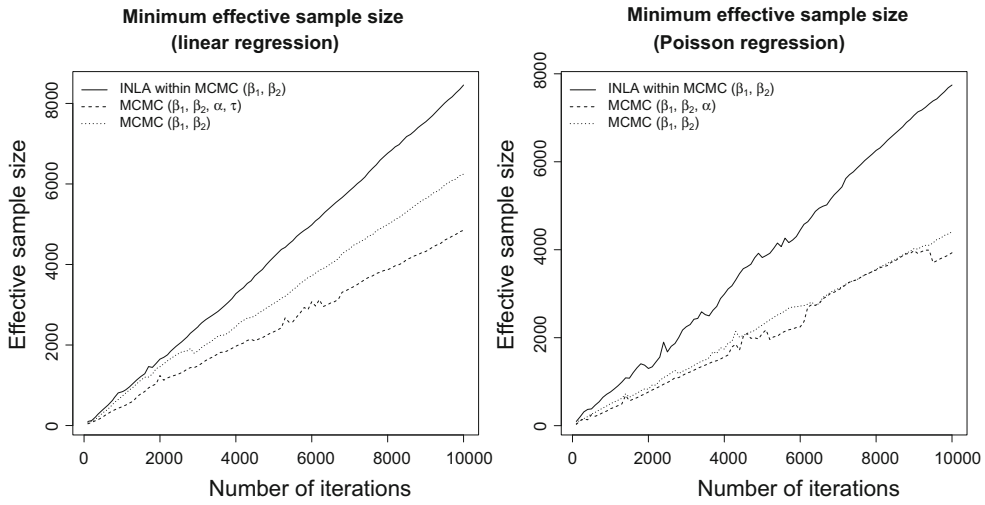

Fig. 5 Minimum effective sample size achieved with INLA within MCMC and MCMC

Here, y_i is the response variable and x_{ij} are associated covariates. n is the number of observations and p the number of covariates. Parameter λ is a nonnegative penalty term to control how the shrinkage of the coefficients is done. If $\lambda = 0$, then the fitted coefficients are those obtained by maximum likelihood, while higher values of λ will shrink the estimates toward zero.

The Lasso is closely related to Bayesian inference as it can be regarded as a standard regression model with Laplace priors on the variable coefficients. The Laplace distribution is defined as

$$f(\beta) = \frac{1}{2\sigma} \exp\left(-\frac{|\beta - \mu|}{\sigma}\right), \quad x \in \mathbb{R},$$

where μ and σ, a positive number, are parameters of location and scale, respectively. The Laplace prior distribution is not available for (parts of) the latent field in **R-INLA**. However, conditioning on the values of the $\boldsymbol{\beta}$-coefficients the model can be easily fitted with **R-INLA**.

We will apply the methodology described in this paper to implement the Bayesian Lasso by combining INLA and MCMC. We will be using the `Hitters` dataset described in James et al. (2013). This dataset records several statistics about players in the Major League Baseball, including salary in 1987, number of times at bat in 1986 and other variables. Our aim is to build a model to predict the player's salary in 1987 on some of the other variables recorded in 1986 (the previous season).

We will focus on a smaller model than the one described in James et al. (2013) and will consider predicting salary in 1987 on only five variables measured from the 1986 season: number of times at bat (AtBat), number of hits (Hits), the number of home runs (HmRun), number of runs (Runs) and the number of runs batted in (RBI).

For our implementation of the Bayesian Lasso, observations y_i will be assumed to have a Gaussian distribution with mean $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ and precision τ. We will be fitting models conditioning on the covariate coefficients $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)$. Also, we will assume that $\boldsymbol{\beta}$ and the error term precision τ are independent a priori, i.e., $\pi(\boldsymbol{\beta}, \tau) = \pi(\boldsymbol{\beta})\pi(\tau)$. This will provide a simpler way to compare our results with the Lasso, and it will also make computations a bit simpler. However, note that it is also possible to choose a prior so that $\pi(\boldsymbol{\beta}, \tau) = \pi(\boldsymbol{\beta}|\tau)\pi(\tau)$ (see, for example, Lykou and Ntzoufras 2011). The posterior distribution of these variables will be obtained using MCMC.

Regarding the prior on $\boldsymbol{\beta}$, we have assumed that the five coefficients $\beta_1, \ldots, \beta_5$ are independent a priori. Hence, the prior is the product of five Laplace distributions with $\mu = 0$ and $\sigma = 1/\lambda = 1/0.73$, because the Lasso provided an estimate of λ equal to 0.73. The proposal distribution for $\boldsymbol{\beta}$ is a multivariate Gaussian with zero mean and precision $4 \cdot \mathbf{X}^\intercal \mathbf{X}$, with $\mathbf{X}$ a matrix that has the covariates as columns. This proposal distribution resulted on a good acceptance rate. Finally, the prior on τ is the default in **R-INLA**, which is a Gamma distribution with parameters 1 and 5e−05.

The summary of the Lasso estimates is available in Table 1, and the posterior distributions of the coefficients are shown in Fig. 6. In all cases, there is agreement between the Lasso and Bayesian Lasso estimates. Also, the posterior distributions of the model coefficients are the same for MCMC and combining INLA with MCMC. For those coefficients with a zero estimate with the Lasso, the posterior distribution obtained with the Bayesian Lasso is centered at zero.

6.2 Imputation of Missing Covariates

van Buuren and Groothuis-Oudshoorn (2011) describe the R package **mice** that implements several multiple imputation

Table 1 Summary estimates of the Lasso and Bayesian Lasso (posterior mean and standard deviation, between parentheses)

Coefficient	Lasso	INLA w/MCMC	MCMC
AtBat	0.00	− 0.01 (0.08)	− 0.02 (0.08)
Hits	0.18	0.17 (0.11)	0.17 (0.12)
HmRun	0.00	0.03 (0.06)	0.02 (0.07)
Runs	0.00	0.07 (0.09)	0.07 (0.08)
RBI	0.23	0.20 (0.11)	0.22 (0.11)

methods. We will be using the `nhanes` dataset to illustrate how our approach can be used to provide imputation of missing covariates in a real dataset. This dataset contains data from Schafer (1997) on age, body mass index (`bmi`), hypertension status (`hyp`) and cholesterol level (`chl`). Age is divided into three groups: 20–39, 40–59 and 60+.

Our aim is to impute missing covariates in order to fit a model that explains the cholesterol level through age and body mass index. Although the values of age have been completely observed, there are missing values in body mass index and cholesterol level. INLA can handle missing values in the response (and will provide a predictive distribution of the missing response) but, as already stated, is not able to handle models with missing values in the covariates.

We will consider a very simple imputation mechanism by assigning a Gaussian prior to the missing values of body mass index. This Gaussian distribution is centered at the average of the observed values (26.56) and its variance is four times the variance of the observed values (71.07). With this, we expect to provide some guidance on how the imputed values should be but allowing for a wide range of variation. More complex imputation mechanisms could be considered (see, for example, Little and Rubin 2002). As in previous examples, we will fit the same model using MCMC in order to compare both results. The model that we will fit is:

$$\begin{aligned} \text{chl}_i &= \beta_0 + \beta_1 \text{bmi}_i + \beta_2 \text{age2}_i + \beta_3 \text{age3}_i + \varepsilon_i \\ \beta_0 &\propto 1 \\ \beta_k &\propto N(0, 0.001);\ k = 1, 2, 3 \\ \varepsilon_i &\sim N(0, \tau) \\ \tau &\sim Ga(1, 0.00005) \end{aligned} \quad . \tag{20}$$

Figure 7 shows the posterior marginal distributions of the imputed values of the body mass index. Both MCMC and our approach provide very similar point estimates. Table 2 summarizes the model parameters obtained both with MCMC and our approach, and Fig. 8 displays the posterior marginals of the model parameters obtained with our approach and MCMC. In all cases, the marginals agree, and the point estimates look very similar.

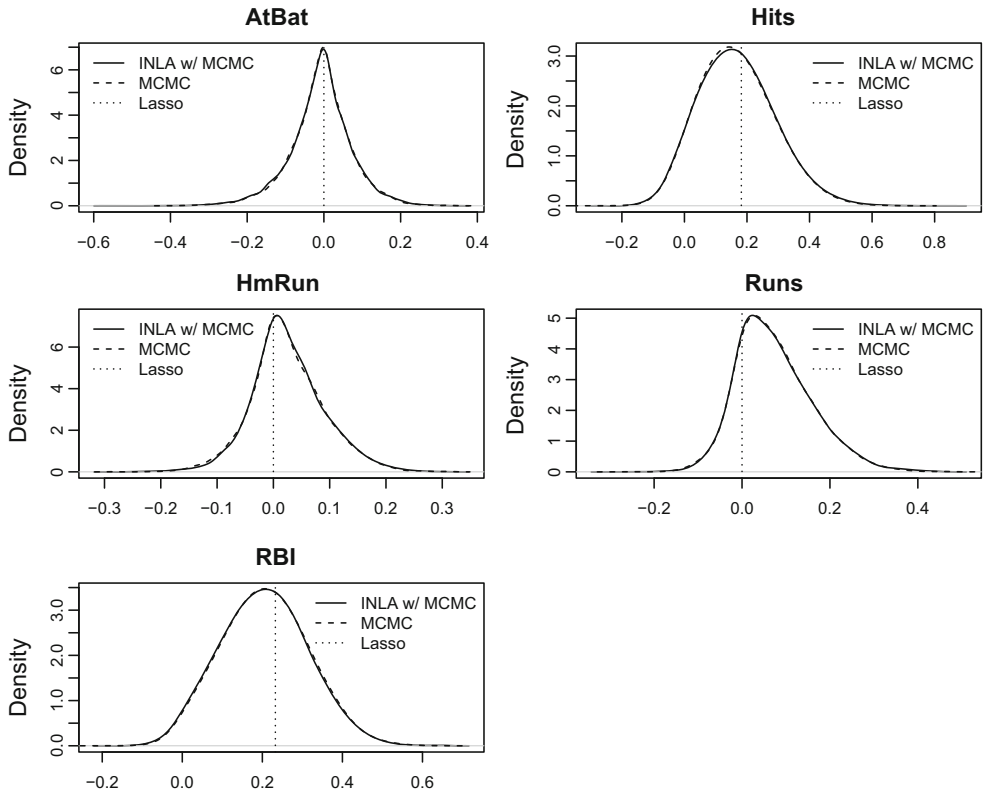

Fig. 6 Summary of results for the Lasso and Bayesian Lasso

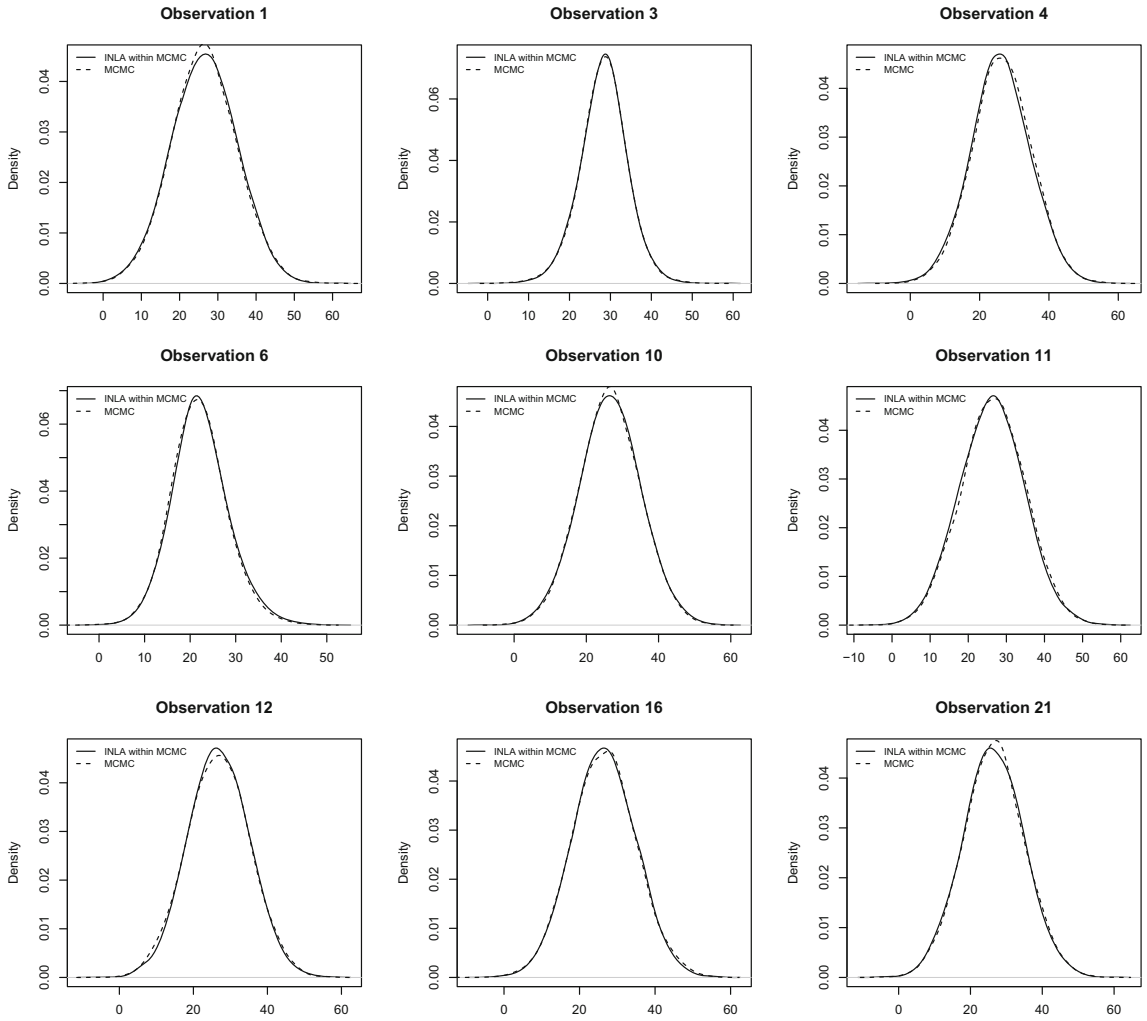

Fig. 7 Marginal distributions of the imputed values of body mass index

Table 2 Summary of model parameter posterior estimates: posterior mean and standard deviation (in parentheses), model with missing covariates

Parameter	MCMC	INLA w/MCMC
β_0	39.760 (61.463)	43.469 (62.603)
β_1	4.994 (2.167)	4.864 (2.206)
β_2	29.989 (17.542)	29.501 (17.871)
β_3	50.049 (23.277)	49.449 (23.207)
τ	0.001 (0.0005)	0.001 (0.0005)

6.3 Spatial econometrics models

Bivand et al. (2014) describe a novel approach to extend the classes of models that can be fitted with **R-INLA** to fit some spatial econometrics models. In particular, they fit several conditional models by fixing the values of some of the parameters in the model, and then, they combine these models using a Bayesian model averaging approach (Hoeting et al. 1999). Bivand et al. (2015) show a practical implementation with a spatial statistics model using R package **INLABMA**. Some of these models have already been included in **R-INLA** (Gómez-Rubio et al. 2017), but are still considered as experimental.

In this example, we will focus on one of the spatial econometrics models described in Bivand et al. (2014) to illustrate how our new approach to combine MCMC and **R-INLA** can be used to fit unimplemented models. In particular, we will consider the spatial lag model (LeSage and Pace 2009):

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u};\ \boldsymbol{u} \sim N(0, \frac{1}{\tau_u}\boldsymbol{I}).$$

Here, $\boldsymbol{y}$ is a vector of observations at n areas, $\boldsymbol{W}$ is an adjacency matrix, ρ a spatial autocorrelation parameter, $\boldsymbol{X}$ a $n \times p$ matrix of covariates with associated coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{u} = (u_1, \ldots, u_n)$ an error term. u_i, $i = 1, \ldots, n$, is Normally distributed with zero mean and precision τ_u. This model can be rewritten as follows:

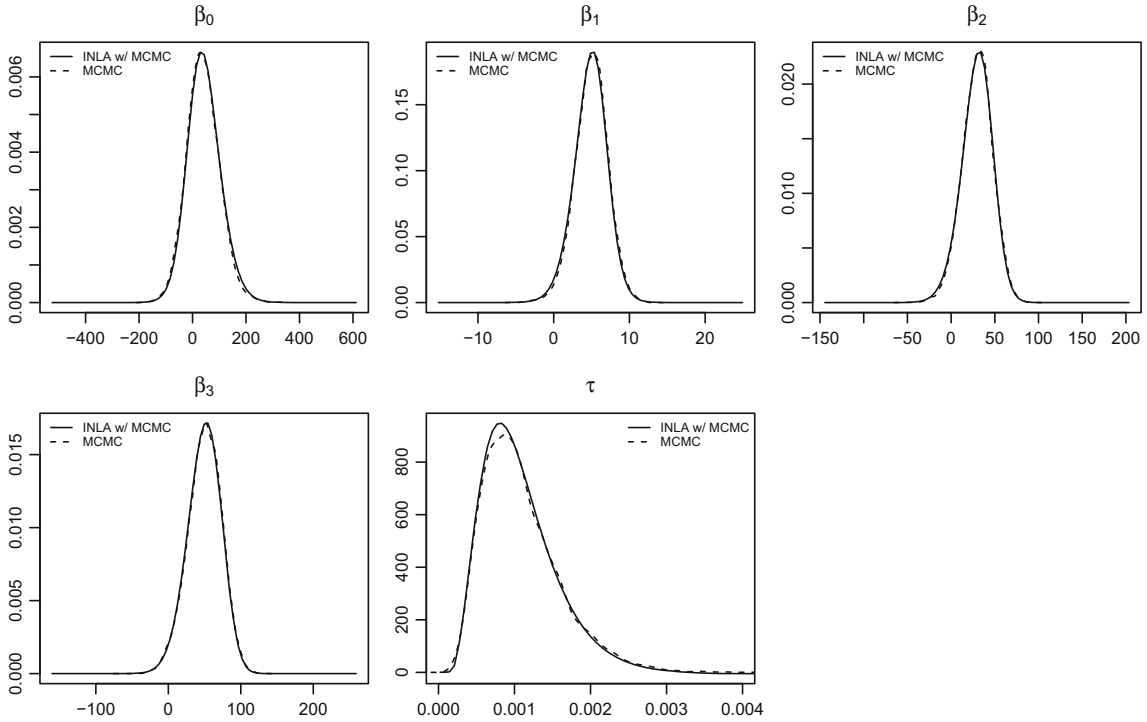

Fig. 8 Marginal distributions of the model parameters, model with missing values in the covariates

$$\boldsymbol{y} = (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon};$$
$$\boldsymbol{\varepsilon} \sim N\left(0, \frac{1}{\tau_u}\left[(\boldsymbol{I}_n - \rho \boldsymbol{W}')(\boldsymbol{I}_n - \rho \boldsymbol{W})\right]^{-1}\right).$$

This model is difficult to fit with any standard software for mixed-effects models because of parameter ρ. If the value of ρ is fixed, then it is straightforward to fit the model with **R-INLA** as it becomes a linear term on the covariates plus a random effects term with a known structure. Hence, by conditioning on the value of ρ we will be able to fit the model with **R-INLA**. In order to use our new approach, we will be drawing values of ρ using MCMC and conditioning on this parameter to fit the models with **R-INLA**.

Note that the adjacency matrix $\boldsymbol{W}$ is often taken to be row-standardized. This implies that ρ is constrained to the interval $(1/\lambda, 1)$, where λ is the minimum eigenvalue of $\boldsymbol{W}$ (see, Haining 2003, for details). This also means that ρ is not necessarily restricted to the interval $(-1, 1)$, as it might be expected.

We have fitted this model to the Columbus dataset available in **R** package **spdep**. This dataset contains information about 49 neighborhoods in Columbus (Ohio), and we have considered a model with crime rates as the response and household income and housing value as covariates. We have also fitted the spatial lag model using a maximum likelihood approach, the method proposed by Bivand et al. (2014) and MCMC using an implementation of the model for the Jags software included in package **SEMCMC**, which can be downloaded from Github.

Regarding prior distributions, ρ is assigned a uniform between -1.5 and 1, because in this case the inverse of the minimum eigenvalue of $\boldsymbol{W}$ is -1.5. Coefficients β_i, $i = 1, \ldots, p$, have been assigned Gaussian priors with zero mean and precision 0.001 (the default in **R-INLA**), and τ_u is assigned a Gamma distribution with parameters 1 and 0.00005 (the default for the precision of a 'generic0' latent class in **R-INLA**).

The results are shown in Table 3. All Bayesian approaches have very similar estimates, and these are also very similar to the maximum likelihood estimates.

6.4 Classification

In the previous examples, we have considered problems in which the number of latent parameters is small. In this new example, we will tackle the problem of classifying observations into a given number of groups. In particular, we will consider the eruption times of the Old Faithful geyser in Yellowstone National Park (Azzalini and Bowman 1990).

Waiting time since the previous eruption and eruption times is shown in Fig. 9, where a kernel density estimate of the eruption times has been displayed. It seems that there is a strong correlation between the time since the last eruption and eruption time, with longer waiting times leading

Table 3 Posterior means (and standard deviation) of the spatial lag model fitted to the Columbus data set using three different methods

Parameter	Max. Lik.	INLA w/ MCMC	MCMC	INLA+BMA
Intercept	61.05 (5.31)	60.62 (6.08)	58.53 (6.92)	60.81 (5.33)
$\beta_{\text{h. income}}$	− 1.00 (0.34)	− 0.97 (0.37)	− 0.91 (0.39)	− 0.98 (0.33)
$\beta_{\text{h. value}}$	− 0.31 (0.09)	− 0.31 (0.09)	− 0.30 (0.10)	− 0.31 (0.09)
ρ	0.52 (0.14)	0.55 (0.13)	0.55 (0.16)	0.54 (0.11)
τ_u	0.01 (–)	0.01 (0.002)	0.01 (0.002)	0.01 (0.00004)

INLA w/ MCMC refers to the results using the approach described in this paper, MCMC to the results from an implementation with Jags and INLA+BMA to the results using the method proposed by Bivand et al. (2014)

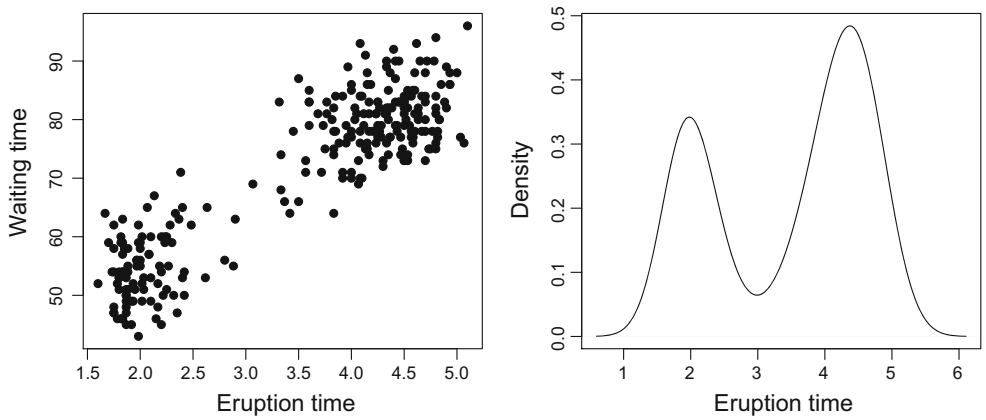

Fig. 9 Waiting times and eruption times of the Old Faithful geyser in Yellowstone National Park

to longer eruptions. Also, it seems that observations can be grouped into short and long eruptions.

We will label as 'group 1' the eruption times in the group with the lower mean, so that 'group 2' will be observations with longer eruption times. Furthermore, observations within each group will be assumed to follow a Gaussian distribution with mean μ_j and precision τ_j, with $j = 1, 2$. Classification will be done through a vector of latent index variables $\boldsymbol{z} = (z_1, \ldots, z_n)$. z_i indicates the group to which observation i belongs and the values it can take are either 1 or 2.

Hence, the aim is computing the posterior probabilities of $\boldsymbol{z}$ given the vector of eruption times $\mathbf{y}$, as well as the posterior distributions of the means and precisions of the Gaussian distributions that define the groups.

In general, this is a difficult problem (see, Marin et al. 2005, for a summary) where MCMC often struggles. A known phenomenon is that of label switching, which occurs when the observations are essentially assigned to the same groups, but the labels of these groups are swapped. This makes inference difficult because labels must be reassigned after the MCMC has been run, increasing computational time and postprocessing.

For this reason, we will use informative priors on μ_1 and μ_2 in order to avoid label switching. In particular, the prior on μ_1 will be a Gaussian distribution centered at 2 and the prior on μ_2 will also be Gaussian centered at 4.5. The precisions of both prior distributions will be 1. Although label switching may appear during burn-in, in this particular case it disappears once groups start to become identified. The priors on precisions τ_1 and τ_2 are the default in **R-INLA**, i.e., a Gamma with parameters 1 and 5e−05. Regarding index variables, they will have a prior distribution such as there is no preference a priori for any group, i.e., $\pi(z_i = 1) = \pi(z_i = 2) = 0.5,\ i = 1, \ldots, n$.

The proposal distribution will be defined such as the proposed values of the index variables depend on the proportion of observations allocated into each group and the estimates of the distributions that define the groups. This will follow the sampling distribution used by Gibbs sampling (Chib 1995). In addition, each z_i will be sampled separately, but the proposed ensemble value $\boldsymbol{z}^*$ will be accepted or rejected in a single movement. Hence, at iteration $k + 1$ a new value for z_i is proposed using the following probability distribution:

$$q(z_i^* | z_i^{(k)} = j) \propto \hat{w}_j^{(k)} N(y_i | \hat{\mu}_j^{(k)}, \hat{\tau}_j^{(k)}), \quad j = 1, 2,$$

where $\hat{w}_j^{(k)}$ is the proportion of observations in group j, $\hat{\mu}_j^{(k)}$ and $\hat{\tau}_j^{(k)}$ are the means of $\tilde{\pi}(\mu_j | \mathbf{y}, \boldsymbol{z}^{(k)})$ and $\tilde{\pi}(\tau_j | \mathbf{y}, \boldsymbol{z}^{(k)})$, respectively, at iteration k. That is, $\hat{\mu}_j^{(k)}$ and $\hat{\tau}_j^{(k)}$ are estimates of the parameters of the Gaussian distributions that define the observations in each group computed using the conditional marginals obtained at iteration k.

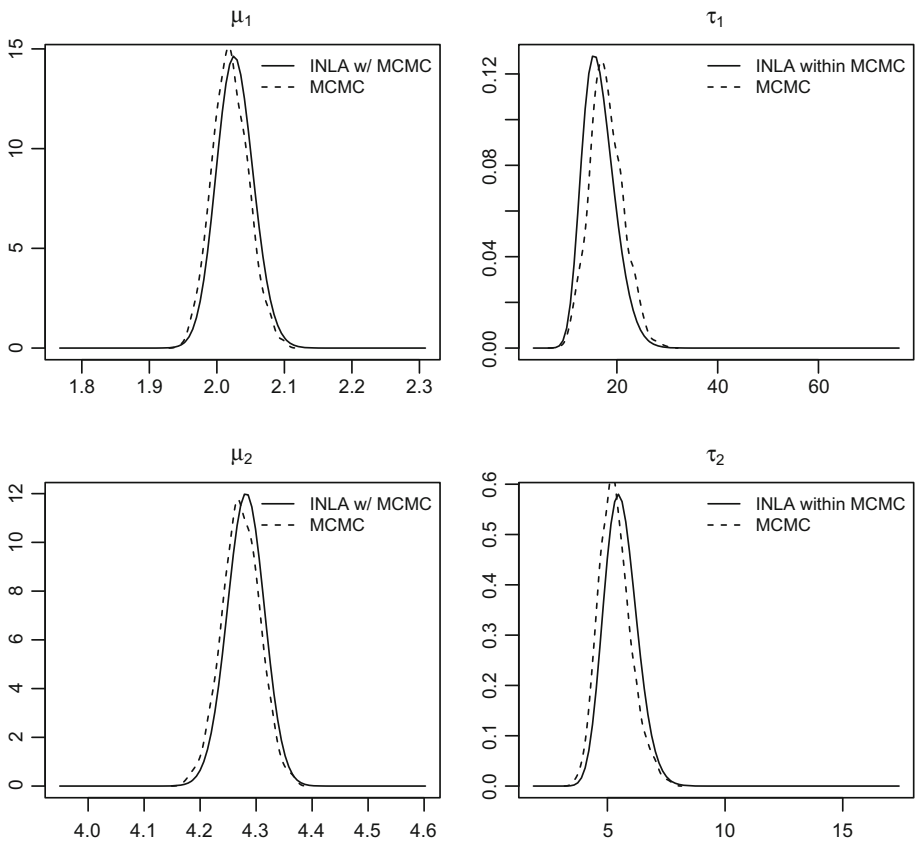

Fig. 10 Posterior marginals of the means and precisions of the Gaussian distributions that define the two groups

This model can be easily fitted using the approach that we have described before because, given z, the model is completely defined and it can be fitted with INLA. In particular, this would be a model with two likelihoods, one for each group, in which each group is defined by a different Gaussian distribution. Hence, each time a new proposal z^* is drawn, observations are reassigned to groups according to z^* and the model is refitted.

As a starting point, we have considered the observations in increasing order and we have assigned one third of the observations of group 1 and the others to group 2. When running the algorithms, we have used the same number of iterations as in the other examples, as described at the beginning of Sect. 6. In this case, the acceptance rate of INLA within MCMC has been 71.74%. Figure 10 shows the estimates of the posterior marginals obtained with INLA within MCMC and MCMC. Again, we find that there is a very good agreement between both approaches. However, we have also observed that the choice of the initial labeling is important to achieve a fast convergence.

7 Discussion

In this paper, we have developed a novel approach to extend the models that can be fitted with INLA. For this, the parameters are split into two sets and we have used INLA within the Metropolis–Hastings algorithm to sample only a small number of parameters to estimate their posterior distribution. For the remainder of parameters, the posterior marginals are estimated using Bayesian model averaging using the conditional posterior marginals obtained at with INLA the steps of the Metropolis–Hastings algorithm. The idea of dividing the parameter space of our model into two groups to estimate them using a combination of different methods has also been studied by other authors (for example, Vanhatalo et al. 2013). This is a convenient approach because of the ease to build and fit very complex models, and it is particularly important when specific approaches or software are good at a precise task.

We have shown four important applications of INLA within MCMC. In the first one, we have implemented a Bayesian Lasso using Laplace priors on the coefficients of the covariates. This example shows how other priors not available in **R-INLA** could be used on the latent effects and hyperparameters. This includes not only univariate priors, but also improper, objective and multivariate priors that are seldom available in **R-INLA**.

In our second example, we have tackled the problem of imputation of missing covariates in model fitting. Here, we have included a very simple imputation method for the missing values in the covariates, so that model fitting and

imputation were done at the same time. Compared to fitting this model with MCMC, we obtained very similar posterior estimates. In an ongoing work, we are exploring how this can be extended to larger problems and how different imputation models and missingness mechanisms can be properly addressed with INLA and MCMC.

In the third example, we have also shown how other models not included in the **R-INLA** software can be fitted with INLA and MCMC. In particular, we have fitted a spatial econometrics model by fitting conditional models on the spatial autocorrelation parameter. This method can be easily modified to suit any other models. In addition, Gibbs sampling could be used if the full conditionals are available for a subset of model parameters.

Finally, in the last example we have shown how INLA within MCMC can be used to fit mixture models with INLA. Although we have considered a mixture with two components, the methodology can be extended to fit mixtures with any number of components. However, fitting mixture models with our approach requires further investigation and we will focus on this particular topic in future research.

To sum up, INLA provides a simple way to reduce the dimension of the model so that estimation in the resulting low-dimensional parameter space can be tackled with a variety of other methods. In our opinion, this approach allows INLA to fit more complex models and perform multivariate inference on a small set of model paramaters, and it can also be combined with other MCMC algorithms to develop simple samplers to fit complex Bayesian hierarchical models. This method can work well when the conditional models are hard to explore with current approaches for which INLA provides a fast approximation, such as geostatistical models. Furthermore, INLA could be embedded into a Reversible Jump MCMC algorithm so that once the model dimension has been set, the resulting model is approximated with INLA. See, for example, Chen et al. (2000) for a comprehensive list of MCMC algorithms that could benefit from embedding INLA.

Acknowledgements Virgilio Gómez-Rubio has been supported by Grant PPIC-2014-001, funded by Consejería de Educación, Cultura y Deportes (JCCM) and FEDER, and Grant MTM2016-77501-P, funded by Ministerio de Economía y Competitividad. We would also like to thank Prof. Aki Vehtari for his comments on a preliminary version of this paper.

References

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach to efficient monte carlo computations. Genetics **37**(2), 697–725 (2003)

Azzalini, A., Bowman, A.W.: A look at some data on the Old Faithful geyser. Appl. Stat. **39**, 357–365 (1990)

Beaumont, M.A.: Estimation of population growth or decline in genetically monitored populations. Genetics **164**, 1139–1160 (2003)

Bivand, R.S., Gómez-Rubio, V., Rue, H.: Approximate Bayesian inference for spatial econometrics models. Spat. Stat. **9**, 146–165 (2014)

Bivand, R.S., Gómez-Rubio, V., Rue, H.: Spatial data analysis with **R-INLA** with some extensions. J. Stat. Softw. **63**(20), 1–31 (2015)

Chen, M.-H., Shao, Q.-M., Igrahim, J.G.: Monte Carlo Methods in Bayesian Computation. Springer, New York (2000)

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**(432), 1313–1321 (1995)

Gilks, W., Gilks, W., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. Chapman & Hall, Boca Raton (1996)

Gómez-Rubio, V., Bivand, R.S., Rue, H.: Estimating spatial econometrics models with integrated nested Laplace approximation (2017). arXiv preprint arXiv:1703.01273

Haining, R.: Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge (2003)

Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)

Hoeting, J., David Madigan, A.R., Volinsky, C.: Bayesian model averaging: a tutorial. Stat. Sci. **14**, 382–401 (1999)

Hubin, A., Storvik, G.: Efficient mode jumping MCMC for Bayesian variable selection in GLMM (2016a). arXiv preprint arXiv:1604.06398

Hubin, A., Storvik, G.: Estimating the marginal likelihood with integrated nested Laplace approximation (INLA) (2016b). arXiv preprint arXiv:1611.01450

James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. Springer, Berlin (2013)

Joensuu, H., Reichardt, P., Eriksson, M., Hall, K.S., Vehtari, A.: Gastrointestinal stromal tumor: a method for optimizing the timing of CT scans in the follow-up of cancer patients. Radiology **271**(1), 96–106 (2014). PMID: 24475826

LeSage, J., Pace, R.K.: Introduction to Spatial Econometrics. Chapman and Hall/CRC, Boca Raton (2009)

Li, Y., Brown, P., Rue, H., Al-Maini, M., Fortin, P.: Spatial modelling of Lupus incidence over 40 years with changes in census areas. J. R. Stat. Soc. Ser. C **61**, 99–115 (2012)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (2002)

Lykou, A., Ntzoufras, I.: WinBUGS: a tutorial. Wiley Interdiscipl. Rev. Comput. Stat. **3**, 385–396 (2011)

Marin, J.-M., Mengersen, K., Robert, C.P.: Bayesian modelling and inference on mixtures of distributions. In: Dey, D.K., Rao, C.R. (eds.) Handbook of Statistics, vol. 25. Elsevier, Amsterdam (2005)

Martins, T.G., Simpson, D., Lindgren, F., Rue, H.: Bayesian computing with INLA: new features. Comput. Stat. Data Anal. **67**, 68–83 (2013)

Medina-Aguayo, F.J., Lee, A., Roberts, G.O.: Stability of noisy Metropolis–Hastings. Stat. Comput. **26**, 1187–1211 (2016)

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machine. J. Chem. Phys. **21**, 1087–1091 (1953)

Pettit, L.I.: The conditional predictive ordinate for the normal distribution. J. R. Stat. Soc. Ser. B (Methodol.) **52**(1), 175–184 (1990)

Plummer, M.: **rjags**: Bayesian Graphical Models using MCMC. R package version 4-6 (2016)

Plummer, M., Best, N., Cowles, K., Vines, K.: CODA: Convergence diagnosis and output analysis for MCMC. R News **6**(1), 7–11 (2006)

Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). J. R. Stat. Soc. B **7**(2), 319–392 (2009)

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K.: Bayesian computing with INLA: a review. Annu. Rev. Stat. Appl. **4**, 395–421 (2017)

Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall, London (1997)

Sherlock, C., Thiery, A.H., Roberts, G.O., Rosenthal, J.S.: On the efficiency of pseudo-marginal random walk metropolis algorithms. Ann. Stat. **43**(1), 238–275 (2015)

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A.: Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. B **64**(4), 583–616 (2002)

Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**(1), 267–288 (1996)

Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**(393), 82–86 (1986)

van Buuren, S., Groothuis-Oudshoorn, K.: Mice: multivariate imputation by chained equations in R. J. Stat. Softw. **45**(1), 1–67 (2011)

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari, A.: GPstuff: Bayesian modeling with Gaussian processes. J. Mach. Learn. Res. **14**, 1175–1179 (2013)

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., Winther, O.: Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. J. Mach. Learn. Res. **17**(103), 1–38 (2016)