

Importance Sampling with the Integrated Nested Laplace Approximation

Item Type	Article
Authors	Berild, Martin Outzen; Martino, Sara; Gómez-Rubio, Virgilio; Rue, Haavard
Citation	Berild, M. O., Martino, S., Gómez-Rubio, V., & Rue, H. (2022). Importance Sampling with the Integrated Nested Laplace Approximation. Journal of Computational and Graphical Statistics, 1–25. https://doi.org/10.1080/10618600.2022.2067551
Eprint version	Publisher's Version/PDF
DOI	10.1080/10618600.2022.2067551
Publisher	Informa UK Limited
Journal	Journal of Computational and Graphical Statistics
Rights	Archived with thanks to Journal of Computational and Graphical Statistics under a Creative Commons license, details at: http://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2024-04-17 04:43:30
Item License	http://creativecommons.org/licenses/by-nc-nd/4.0/
Link to Item	http://hdl.handle.net/10754/668441

Importance Sampling with the Integrated Nested Laplace Approximation

Martin Outzen Berild^a, Sara Martino^a, Virgilio Gómez-Rubio^b , and Håvard Rue^c

^aDepartment of Mathematics, Norwegian University of Science and Technology, Trondheim, Norway; ^bDepartment of Mathematics, School of Industrial Engineering-Albacete, Universidad de Castilla-La Mancha, Albacete, Spain; ^cCEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ABSTRACT

The integrated nested Laplace approximation (INLA) is a deterministic approach to Bayesian inference on latent Gaussian models (LGMs) and focuses on fast and accurate approximation of posterior marginals for the parameters in the models. Recently, methods have been developed to extend this class of models to those that can be expressed as conditional LGMs by fixing some of the parameters in the models to descriptive values. These methods differ in the manner descriptive values are chosen. This article proposes to combine importance sampling with INLA (IS-INLA), and extends this approach with the more robust adaptive multiple importance sampling algorithm combined with INLA (AMIS-INLA). This article gives a comparison between these approaches and existing methods on a series of applications with simulated and observed datasets and evaluates their performance based on accuracy, efficiency, and robustness. The approaches are validated by exact posteriors in a simple bivariate linear model; then, they are applied to a Bayesian lasso model, a Poisson mixture, a zero-inflated Poisson model and a spatial autoregressive combined model. The applications show that the AMIS-INLA approach, in general, outperforms the other methods compared, but the IS-INLA algorithm could be considered for faster inference when good proposals are available. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received March 2021
Accepted April 2022

KEYWORDS

Bayesian lasso; Bayesian inference; INLA; Importance sampling; Mixture models; Spatial models

1. Introduction

The integrated nested Laplace approximation (INLA, Rue, Martino, and Chopin 2009) is a numerical method for approximated Bayesian inference on a well determined class of models named latent Gaussian models (LGMs). INLA focuses on providing approximate marginal posterior distributions for all parameters in the model. This is in contrast with the more traditional Markov Chain Monte Carlo (MCMC, Gilks et al. 1996) based inference that provides instead an estimate of the joint posterior distribution. INLA has become a widely used method because it is, in general, faster than MCMC while still providing accurate estimates. Moreover, INLA is implemented as an R package called R-INLA, that allows the user to do inference on complex hierarchical models often in a matter of seconds.

Implementing INLA from scratch may be a difficult task, therefore, fitting models with INLA is, in practice, restricted to the classes of models implemented in the R-INLA package. How to enlarge such selection has been the topic of many papers (see, e.g., Bivand, Gómez-Rubio, and Rue 2014, 2015; Gómez-Rubio, Bivand, and Rue 2020 and the references therein). One interesting approach is the one taken in Gómez-Rubio and Rue (2018) where they propose to combine INLA and MCMC methods. The basic idea is that certain models, named conditional LGMs, can be fitted with INLA, provided a (small) number of parameters are fixed to a given value. Gómez-Rubio and Rue (2018) propose to draw samples from the posterior distribution of the

conditioning parameters by combining MCMC techniques and conditional models fitted with R-INLA. This is made possible by the fact that INLA computes also the marginal likelihood of the conditional fitted model. The marginal likelihood is used in Gómez-Rubio and Rue (2018) to compute the acceptance probability in the Metropolis–Hastings (MH) algorithm, which is a popular MCMC method.

Combining INLA and MCMC allows to increase the number of models that can be fitted using R-INLA. The MCMC algorithm is simple to implement as only the conditioning parameters need to be sampled while the rest of the parameters are integrated out using INLA. The INLA-MCMC approach proposed by Gómez-Rubio and Rue (2018) relies on the MH algorithm and requires model fitting with R-INLA at every step. That may be slow in practice because the sequential nature of the MH algorithm makes parallelization difficult to implement. Gómez-Rubio and Palmí-Perales (2019) provide some insight on how to speed up the process of fitting conditional models with INLA, but it requires a good approximation to the posterior mode of the parameters of interest by relying, for example, on maximum likelihood estimates.

In this article we propose a new method for fitting conditional LGMs with INLA, similar in spirit to Gómez-Rubio and Rue (2018) but based on the importance sampling (IS) algorithm instead of on the MH one. The main advantage of the IS algorithm over MH is that it is easy to parallelize, thus, allowing for a great improvement in computational speed. The drawback

is that, lacking the adaptive nature of the MH algorithm, the performance of IS based inference relies on the choice of a good proposal distribution. This can be hard to determine in many practical cases. Therefore, we also propose an algorithm that is based on an adaptive multiple IS (AMIS, Corneut et al. 2012) that, for a slightly higher computing time than IS, has the advantage of requiring less human intervention.

By combining IS and AMIS with INLA it is possible to fit models with INLA that are highly parameterized or that escape the structure of a latent Gaussian Markov random field (GMRF) such as mixture models and models that include a hierarchical structure on the parameters of the likelihood or random effects. Furthermore, given that IS and AMIS are not sequential algorithms, it is possible to conduct model fitting in parallel in a short time.

The rest of the article is organized as follows. The class of models amenable to INLA is described in Section 2. A short description of how INLA works is also given in the same Section. Section 3 introduces IS while Section 4 shows how INLA and IS can be combined. In this section we also discuss numerical and graphical diagnostics to assess the accuracy of our algorithm. In Section 5 an adaptive version of the algorithm is presented while in Section 6 we show, in several examples, how our proposal works in practice. We end with a discussion in Section 7.

2. The Integrated Nested Laplace Approximation

Let our response $\mathbf{y} = (y_1, \dots, y_n)$ form a vector of observations from a distribution in the exponential family with mean μ_i . We assume that a linear predictor η_i can be related to μ_i using an appropriate link function:

$$\eta_i = g(\mu_i) = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \epsilon_i. \quad (1)$$

Here α is a common intercept, z_k indicate covariates with linear effect β_k to be estimated, ϵ_i is an independent noise term while $f^{(j)}(\cdot)$ indicates terms such as random effects, spatial effects, nonlinear effects of the covariates, etc., defined by some indices u_j . The terms $\mathbf{x} = (\eta, \alpha, \beta, f^{(1)}, f^{(2)}, \dots)$ define a latent field. The likelihood and the prior for \mathbf{x} will depend on some hyperparameters $\boldsymbol{\theta}$ and an appropriate prior $\pi(\boldsymbol{\theta})$ is assigned to these.

From Equation (1), it is clear that the observations are conditionally independent given the latent effect \mathbf{x} and the hyperparameters $\boldsymbol{\theta}$ so that the likelihood can be written as

$$\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}), \quad (2)$$

where i belongs to a set $\mathcal{I} = (1, \dots, n)$ that indicates observed responses.

In a Bayesian framework, the main interest lays in the posterior distribution:

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

This is usually not available in closed form, thus, several estimation methods and approximations have been developed. INLA, introduced by Rue, Martino, and Chopin (2009), is one of such

methods. INLA can be used for LGM provided the prior for the latent field \mathbf{x} is a Gaussian Markov random field (GMRF) model (Rue and Held 2005). We assume the latent GMRF to have 0 mean and precision (inverse of covariance) matrix $\mathbf{Q}(\boldsymbol{\theta})$. Equation (3) can then be rewritten as

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

INLA does not seek to approximate the joint posterior distribution $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, instead, it creates numerical approximations to the posterior marginals for the latent field $\pi(x_i|\mathbf{y})$ and the hyperparameters $\pi(\theta_j|\mathbf{y})$. To do this, the first step is to approximate $\pi(\boldsymbol{\theta}|\mathbf{y})$ by $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. Approximated marginal posteriors for the hyperparameters $\tilde{\pi}(\theta_j|\mathbf{y})$ can then be derived from $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ via numerical integration. Posterior marginals for the latent field $\pi(x_i|\mathbf{y})$ can be written as

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \quad (5)$$

and approximated as

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_g \tilde{\pi}(x_i|\theta_g, \mathbf{y})\tilde{\pi}(\theta_g|\mathbf{y})\Delta_g, \quad (6)$$

where θ_g are selected points and $\tilde{\pi}(x_i|\theta_g, \mathbf{y})$ is an approximation to $\pi(x_i|\theta_g, \mathbf{y})$, see Rue, Martino, and Chopin (2009) for details.

As a by-product of the main computations, INLA provides other quantities of interest. Of importance for this article is the marginal likelihood $\pi(\mathbf{y})$, which can be computed as

$$\tilde{\pi}(\mathbf{y}) = \int \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}_0(\boldsymbol{\theta})} d\boldsymbol{\theta}. \quad (7)$$

Here $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is a Gaussian approximation of $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ built by matching the mode and the curvature at the mode and $\mathbf{x}_0(\boldsymbol{\theta})$ is the posterior mode of $\mathbf{x}|\boldsymbol{\theta}$. Hubin and Storvik (2016) have investigated the performance of this approximation, finding it very accurate for a large class of models. Several authors (Bivand, Gómez-Rubio, and Rue 2014, 2015; Gómez-Rubio and Rue 2018; Gómez-Rubio and Palmí-Perales 2019; Gómez-Rubio, Bivand, and Rue 2020) have relied on the estimates of the marginal likelihood provided by INLA for model estimation and they have found them to be accurate enough in a number of scenarios.

3. Importance Sampling

Importance sampling (IS) is a popular Monte Carlo method where a mathematical expectation with respect to a target distribution is approximated by a weighted average of random draws from another distribution. IS relies on a simple probability result, which is stated next.

Let $\pi(x)$ be a probability density function for the random variable X defined on $\mathcal{D} \subseteq \mathbb{R}^d$, $d \geq 1$, and assume that we wish to compute μ_π defined as

$$\mu_\pi = \mathbb{E}_\pi[h(X)] = \int_{\mathcal{D}} h(x)\pi(x)dx, \quad (8)$$

where $h(\cdot)$ is some function of X . Then for any probability density $g(x)$ that satisfies $g(x) > 0$ whenever $h(x)\pi(x) > 0$, it holds that

$$\mu_\pi = \mathbb{E}_g[h(X)w(X)], \quad (9)$$

where the $w(x) = \frac{f(x)}{g(x)}$ and $\mathbb{E}_g[\cdot]$ indicates the expectation with respect to $g(x)$. Independent draws $\{x^{(j)}\}_{j=1}^N$ from $g(x)$ can then be used to approximate μ_π as

$$\hat{\mu}_{IS} = \frac{1}{N} \sum_{i=1}^N h(x_i)w(x_i). \quad (10)$$

In many cases $\pi(x)$ is only known up to a normalizing constant, in these cases $\hat{\mu}_{IS}$ is replaced by

$$\tilde{\mu}_{IS} = \sum_{i=1}^N h(x_i)\bar{w}(x_i) \quad (11)$$

where the so-called self-normalizing weights

$$\bar{w}(x_i) = \frac{w(x_i)}{\sum_{i=1}^N w(x_i)}, \quad (12)$$

can be computed as the normalizing constant cancels out. The estimator based on the self-normalizing weights is slightly biased but it tends to improve the variance of estimates (Robert and Casella 2004).

The performance of the IS estimator, both in its original and self-normalizing form, depends on the choice of the proposal distribution $g(\cdot)$, which should be as close as possible to $\pi(\cdot)$. In fact, an improper choice, for example, lighter tails in $g(\cdot)$, might lead to unbounded weights such that estimates only relies on few samples.

A common measure of the efficiency of the algorithm is the effective sample size (ESS). An estimate can be easily computed as

$$\widehat{ESS} = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2}. \quad (13)$$

This quantity is useful to assess the correlation of the simulated data and provides an overall estimate of the amount of data obtained with sampling. Effective sample size and estimation error are further discussed in Section 4.2.

4. Importance Sampling with INLA

In this Section we discuss how the class of models that INLA can fit can be extended by combining INLA and IS. Our approach follows the path presented in Gómez-Rubio and Rue (2018) with the key difference that we use IS instead of the MH algorithm.

Similar to Gómez-Rubio and Rue (2018) we collect all unknown parameters of the model in the vector $\mathbf{z} = (\mathbf{x}, \boldsymbol{\theta})$ which is split into two subsets $\mathbf{z} = (\mathbf{z}_{-c}, \mathbf{z}_c)$, where \mathbf{z}_{-c} indicates all parameters in \mathbf{z} that are not included in \mathbf{z}_c . The vectors \mathbf{z}_c and \mathbf{z}_{-c} are chosen such that the posterior distribution of \mathbf{z} can be written as

$$\pi(\mathbf{z}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{z}_{-c}, \mathbf{z}_c)\pi(\mathbf{z}_{-c}|\mathbf{z}_c)\pi(\mathbf{z}_c). \quad (14)$$

Furthermore, we assume that this model cannot be fitted with R-INLA unless the parameters in \mathbf{z}_c are fixed to some appropriate values, that is, we model \mathbf{z}_{-c} given \mathbf{z}_c . Conditional on \mathbf{z}_c , R-INLA can produce approximations to the conditional posterior marginals $\pi(z_{-c,k}|\mathbf{y}, \mathbf{z}_c)$, where k indicates the k th element of \mathbf{z}_{-c} , and to the conditional marginal likelihood $\pi(\mathbf{y}|\mathbf{z}_c)$, using Equations (7) and (6), respectively.

Unconditional posterior marginal for the elements of \mathbf{z}_{-c} could then be obtained integrating over \mathbf{z}_c as

$$\begin{aligned} \pi(z_{-c,k}|\mathbf{y}) &= \int \pi(z_{-c,k}, \mathbf{z}_c|\mathbf{y}) d\mathbf{z}_c \\ &= \int \pi(z_{-c,k}|\mathbf{y}, \mathbf{z}_c)\pi(\mathbf{z}_c|\mathbf{y}) d\mathbf{z}_c. \end{aligned} \quad (15)$$

Here, the conditional posterior marginals $\pi(z_{-c,k}|\mathbf{y}, \mathbf{z}_c)$ are approximated with R-INLA.

A naïve Monte Carlo estimate of the integral in Equation (15) is not a viable option; however, IS could be used to sample from a raw approximation $g(\mathbf{z}_c)$ of $\pi(\mathbf{z}_c|\mathbf{y})$, the posterior marginal in Equation (15) can be approximated as

$$\tilde{\pi}(z_{-c,k}|\mathbf{y}) \simeq \sum_{j=1}^n w_j \tilde{\pi}(z_{-c,k}|\mathbf{y}, \mathbf{z}_c^{(j)}), \quad (16)$$

where $\mathbf{z}_c^{(j)}$ are samples from a (multivariate) sampling distribution $g(\cdot)$, $\tilde{\pi}(z_{-c,k}|\mathbf{y}, \mathbf{z}_c^{(j)})$ are the approximated conditional posterior marginals obtained by INLA and w_j are the posterior weights defined as

$$w_j \propto \frac{\pi(\mathbf{z}_c^{(j)}|\mathbf{y})}{g(\mathbf{z}_c^{(j)})} \propto \frac{\pi(\mathbf{y}|\mathbf{z}_c^{(j)})\pi(\mathbf{z}_c^{(j)})}{g(\mathbf{z}_c^{(j)})}. \quad (17)$$

Note that we use the self-normalizing version of the IS algorithm as in Equation (17). When computing w_j we need the conditional marginal likelihood $\pi(\mathbf{y}|\mathbf{z}_c^{(j)})$ which, conveniently, is one of the outputs from R-INLA. See Section 4.2 for a discussion on this.

Finally, the joint posterior distribution of \mathbf{z}_c can be found with

$$\pi(\mathbf{z}_c|\mathbf{y}) = \sum_{j=1}^n w_j \delta(\mathbf{z}_c - \mathbf{z}_c^{(j)}), \quad (18)$$

where $\delta(\cdot)$ is the Dirac delta function. This has also been noted in Elvira, Martino, and Robert (2018). In a practical manner, as Equation (18) would require $n \rightarrow \infty$, the joint posterior distribution, $\pi(\mathbf{z}_c|\mathbf{y})$, is approximated using a weighted nonparametric kernel density estimation (Venables and Ripley 2002). A similar approach is used to find the posterior marginals $\pi(z_{c,k}|\mathbf{y})$ for the k th element of \mathbf{z}_c .

In practice, the choice of the \mathbf{z}_c set is problem dependent. One should aim at having a set that is as small as possible so that the heavy computational work is left to the efficient INLA algorithm. Alternatively, one could choose \mathbf{z}_c such that the complete model is split into a series of simpler, independent models each of which can be fit with INLA. In Section 6 we present examples from both cases.

4.1. Choice of the Sampling Distribution

The sampling distribution $g(\mathbf{z}_c)$ needs to be chosen with care in order to have a good performance of the IS algorithm. In principle, it should be as close as possible to $\pi(\mathbf{z}_c|\mathbf{y})$ but this may be difficult in practice.

We assume that \mathbf{z}_c is a vector of real valued parameters (transformations might be applied if necessary), therefore $g(\mathbf{z}_c)$ is a multivariate distribution. A reasonable proposal could be a multivariate Gaussian or Student's- t with v degrees of freedom. We indicate the location and scale parameters of both the Gaussian and Student's- t as $\boldsymbol{\lambda} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the Student's- t case, for $v > 2$ the covariance is defined as $\frac{v}{v-2} \boldsymbol{\Sigma}$. We want to choose $\boldsymbol{\lambda}$ such that the proposal is close to the target distribution. Moreover, for the Student's- t we want v to be low to guarantee heavy tails. We start therefore, from a preliminary proposal $g_0(\mathbf{z}_c)$, with parameters $\boldsymbol{\lambda}_0 = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Then, N_0 samples are generated from $g_0(\mathbf{z}_c)$ and used to build a rough approximation of the location and scale of the target as

$$\boldsymbol{\mu}_1 = \sum_{j=1}^{N_0} \bar{w}^{(j)} \mathbf{z}_c^{(j)} \quad (19)$$

$$\boldsymbol{\Sigma}_1 = \sum_{j=0}^{N_0} \bar{w}^{(j)} (\mathbf{z}_c^{(j)} - \boldsymbol{\mu}_1)(\mathbf{z}_c^{(j)} - \boldsymbol{\mu}_1)^\top, \quad (20)$$

where $\mathbf{z}_c^{(j)} \sim g_0(\mathbf{z}_c)$ and $\bar{w}^{(j)}$ is the normalized importance weight of the j th sample calculated with Equation (12).

The initial N_0 samples are then discarded and the new (improved) proposal distribution has parameters $\boldsymbol{\lambda}_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Other distributions than the Gaussian and the Student's- t could be used. For example, correction for skewness could be included in the previous approach or distributions with fatter tails could be employed.

For discrete parameters other distributions are required. For example, for binary random variables a Bernoulli distribution could be used. This will allow the probabilities to depend on, for example, a number of fixed effects, similarly as in a generalized linear model. The coefficients of these fixed effect will now take the role of the sampling distribution parameters to conduct IS or AMIS, so that they can be estimated and updated during the model fitting process if necessary. This approach can be extended to the case of discrete variables with more than two categories using a multinomial distribution, for example. Owen (2013) discusses importance sampling in detail for both continuous and discrete variables.

4.2. Estimation of the Error and Diagnostics

IS with INLA can be regarded as a particular type of IS in which INLA is used to integrate most of the latent effects and hyperparameter out, so that IS is applied to the low-dimensional parameter space of \mathbf{z}_c . As a result, IS weights are based on the conditional (on \mathbf{z}_c) marginal likelihood, which is estimated with INLA.

Similarly to what Gómez-Rubio and Rue (2018) point out in the case of INLA within MCMC, it may be difficult to provide an accurate estimate of the estimation error of IS with INLA. Instead, we will argue that the estimates of the

marginal likelihood provided by INLA are accurate, as several authors have pointed out in a wide range of applications. In particular, Hubin and Storvik (2016) conducted a thorough analysis and found the estimates to be very accurate. See, for example, Gómez-Rubio and Palmí-Perales (2019) and Gómez-Rubio, Bivand, and Rue (2020) and the references therein for other uses of the marginal likelihood estimated with INLA to fit different types of spatial models with success.

Hence, we may argue that the conditional marginal likelihoods are estimated with a tiny error, and that this leads to the error introduced when computing importance weights to be small as well. Furthermore, as weights are computed by averaging over a large number of values and then re-scaling, any error introduced is likely to fade out. This should make inference on \mathbf{z}_c accurate and reliable.

The error when estimating the posterior marginals of the elements in \mathbf{z}_{-c} is also difficult to estimate as this is obtained by using a convex combination of some posterior marginals obtained by conditioning on \mathbf{z}_c . Again, we do not expect the error to be large as the conditional marginals are usually estimated with a very small error by INLA, and the weights are likely to have a tiny error, as discussed above.

The first example in Section 6 has been specifically conducted to assess how accurate IS-INLA is when estimating the different posterior marginals of the parameters in the model. As it can be seen, the results provide compelling evidence as to the accuracy of the estimates for the posterior marginals of the elements of \mathbf{z}_c and \mathbf{z}_{-c} .

However, it is clear that the number of samples used in IS-INLA is crucial. For this reason, a number of numerical and graphical criteria should be used to assess that there is sufficient sample size to provide accurate estimates. Owen (2013) describes different ways to compute the effective sample size using the importance weights, as we have stated at the end of Section 3. Elvira, Martino, and Robert (2018) also discuss the estimation of an effective sample size for IS and make a number of important statements about how to compute this. First of all, the effective sample size must be computed separately for each function $h(x)$ involved in IS, that is, the sample size cannot only be computed based on the weights.

Most importantly, they state that the probability distribution $\pi(x)$ (i.e., the target distribution) is approximated by a random measure based on the sampled values of x and their associated weights. Hence, the discrepancy between the sampling distribution $g(x)$ and $|h(x)|\pi(x)$ is directly related to the quality of the IS estimators, with $|h(x)|$ the absolute value of $h(x)$. This discrepancy should then be assessed in some way as well. Note that this evaluations can be done for each element in \mathbf{z}_c separately.

Similarly, Owen (2013) discusses different IS diagnostics that can be used to assess that a sufficiently large sample size has been achieved and states that sample size estimation must include the $h(x)$ function. He proposes an effective sample size criterion dependent on $h(x)$ based on the following weights:

$$\tilde{w}_i(h) = \frac{|h(x_i)|\pi(x_i)/g(x_i)}{\sum_{i=1}^n |h(x_i)|\pi(x_i)/g(x_i)}.$$

The effective sample size, dependent on $h(x)$, is

$$n_e(h) = \frac{1}{\sum_{i=1}^n \tilde{w}_i(h)^2}.$$

This can be computed for each of the elements in \mathbf{z}_c so that a different per-variable effective sample size is obtained. In this particular case, $h(x)$ is taken as the identity function.

As stated above, Elvira, Martino, and Robert (2018) note that the IS sample and weights are implicitly used to estimate the joint posterior distribution of \mathbf{z}_c and their respective posterior marginals. The estimation of these posterior marginals can be regarded as the estimation of the quantiles of the posterior marginal distributions, which may be difficult. For this reason, we propose a graphical assessment based on a probability plot. This is produced by computing the empirical cumulative probability function for each element of \mathbf{z}_c and comparing it to its theoretical value, that is, the cumulative probability function of a discrete uniform distribution between 1 and n , with n the total number of samples. Departures from the identity line will indicate that the posterior marginals are not correctly estimated.

The empirical cumulative distribution for k th element in \mathbf{z}_c is obtained ordering in increasing order the simulated values, and their associated weights in the same order. Then the empirical cumulative distribution is simply the cumulative sum of the reordered weights. These values can be compared with the corresponding values of the theoretical cumulative distribution. For example, the cumulative sum of the reordered weights up to the l th value must be compared to value l/n .

5. Adaptive Multiple Importance Sampling with INLA

The nonadaptive nature of the IS algorithm makes the performance of IS based inference heavily dependent on a good choice of the sampling distribution. In Section 4.1 we suggest one preliminary sample step that could help locate the proposal close to the target distribution. In practice, such step might require several trial-and-error rounds before reaching a satisfactory proposal $g_1(\cdot)$. Moreover, the N_0 preliminary samples are discarded, which might require significant computational costs. It would be therefore, desirable to consider a more efficient design both more automatic and less wasteful of potentially valuable information.

To this end, we propose combining INLA with the adaptive multiple IS algorithm (AMIS) proposed in Corneut et al. (2012). This is one of several versions of adaptive IS algorithms proposed in the literature (see, e.g., Bugallo et al. 2017 and references therein) that has the advantage to employ a mixture of all past sampling distributions in the calculation of the importance weights such that samples can be kept after an adaptation. The proposal is updated several times in an automated way, in order to decrease the dissimilarity between target and proposal.

The algorithm starts with a proposal distribution $g_{\lambda_0}(\cdot)$ (here we will use Gaussian or Student's- t) with parameters $\lambda_0 = (\mu_0, \Sigma_0)$. At each iteration $t = 0, 1, \dots, T$, N_t samples are produced and a new, updated proposal $g_{\lambda_t}(\cdot)$ with parameters $\lambda_t = (\mu_t, \Sigma_t)$ is computed. The new parameters are computed similarly to what is done in Section 4.1 by matching the estimated moments of the target.

At each step, the proposal distribution $\psi_t(\mathbf{z}_c)$ is then a mixture:

$$\psi_t(\mathbf{z}_c) = \frac{\sum_{i=0}^t N_i g_{\lambda_i}(\cdot)}{\sum_{i=0}^t N_i} = \sum_{i=0}^t \rho_i g_{\lambda_i}(\cdot). \quad (21)$$

where $\rho_i = N_i / \sum_{i=0}^t N_i$ is the fraction of samples generated in iteration i . Let $\mathbf{z}_c^{(i,j)} \sim g_i(\cdot)$ be the j th sample generated in the i th iteration; then, the corresponding importance weight is

$$w^{(i,j)} \propto \frac{\tilde{\pi}(\mathbf{y}|\mathbf{z}_c^{(i,j)}) \pi(\mathbf{z}_c^{(i,j)})}{\psi_t(\mathbf{z}_c^{(i,j)})}, \quad (22)$$

where $\tilde{\pi}(\mathbf{y}|\mathbf{z}_c^{(i,j)})$ is the conditional marginal likelihood approximated with R-INLA and $\pi(\mathbf{z}_c^{(i,j)})$ the prior for \mathbf{z}_c evaluated at $\mathbf{z}_c^{(i,j)}$.

Note that the mixture changes after every adaptation and, thereby, the weighing must be updated for all prior samples before estimating new moments for the sampling distribution. To avoid unnecessary calculations a helper variable of the numerator in Equation (21) is used in the implementation. The full algorithm is shown in Algorithm 1.

6. Examples

In this section we present a series of examples to illustrate the methods proposed in the previous sections. The first two examples are taken from Gómez-Rubio and Rue (2018), while the third is described in Gómez-Rubio, Bivand, and Rue (2020). If not stated otherwise, the same strategy for running IS-INLA and AMIS-INLA will be used: they both start from the same preliminary proposal distribution, a Gaussian or Student's- t distribution with 3 degrees of freedom with location μ_0 and scale Σ_0 . IS-INLA uses then 800 samples to update the proposal and estimate the new parameters μ_1 and Σ_1 . The preliminary 800 samples are then discarded and 10,000 samples are generated from the new proposal distribution. AMIS-INLA generates a total of 10,000 samples by adapting the proposal distribution 27 times, to have a high number of adaptation steps. The initial number of samples is $N_0 = 250$. At each adaptation step N_t samples ($t = 1, \dots, T = 26$) are produced. N_t varies from 250 and 500 (with steps of size 10). No sample is discarded. For the MCMC-INLA algorithm, we collect 10,000 samples after convergence has been reached (500 burn-in iterations).

Effective sample size has been computed using the `effectiveSize()` function in the `coda` R-package (Plummer et al. 2006) for MCMC samples. For IS and AMIS samples, the effective sample size has been computed using expression $n_e(h)$ as detailed in Section 4.2. Note that effective sample size for MCMC is based on the actual samples and it will vary with the parameter for which it is computed while for IS and AMIS it is based on the weights so it will be the same for all parameters. When computing the running sample size for MCMC (i.e., the effective sample size after a given number of iterations) the value reported is the minimum among all effective samples size for all the parameters in the model.

For the examples in Sections 6.1–6.3, where we compare running times, a computer with a total of 28 CPUs with 3.2

Algorithm 1: A detailed description of the AMIS-INLA algorithm

- Initialize $N_0, N_1, \dots, N_T, g_{\lambda_0}(\cdot), \pi(z_c)$
- for** t from 0 to T **do**

 - for** j from 1 to N_t **do**

 - Generate sample $z_c^{(t,j)} \sim g_{\lambda_t}(\cdot)$
 - Fit INLA to the model conditional on $z_c = z_c^{(t,j)}$. This produces the quantities:

$\tilde{\pi}(y|z_c^{(t,j)})$ and $\tilde{\pi}(z_{-c,i}|y, z_c^{(t,j)})$, $\forall z_{-c,i} \in z_{-c}$

 - Compute:

$$\gamma^{(t,j)} = \sum_{l=0}^t N_l \cdot g_{\lambda_l}(z_c^{(t,j)}) \text{ and}$$

$$w^{(t,j)} = \frac{\tilde{\pi}(y|z_c^{(t,j)})\pi(z_c^{(t,j)})}{[\gamma^{(t,j)} / \sum_{l=0}^t N_l]}$$

if $t > 0$ **then**

- for** l from 0 to $t - 1$ **do**

 - for** j from 1 to N_l **do**

 - Update past importance weights:

$$\gamma^{(l,j)} \leftarrow \gamma^{(l,j)} + N_l g_{\lambda_l}(z_c^{(l,j)}) \text{ and}$$

$$w^{(l,j)} \leftarrow \frac{\tilde{\pi}(y|z_c^{(l,j)})\pi(z_c^{(l,j)})}{[\gamma^{(l,j)} / \sum_{k=0}^t N_k]}$$

- Calculate λ_{t+1} using the weighted set of samples:

$$(\{z_c^{(0,1)}, w^{(0,1)}\}, \dots, \{z_c^{(t,N_t)}, w^{(t,N_t)}\})$$

- Estimate $\pi(z_c|y)$ using kernel density estimation
- Estimate posterior marginals of z_{-c} :

$$\tilde{\pi}(z_{-c,i}|y) = \sum_{t=0}^T \sum_{j=1}^{N_t} w^{(t,j)} \tilde{\pi}(z_{-c,i}|y, z_c^{(t,j)}) \Bigg/ \sum_{t=0}^T \sum_{j=1}^{N_t} w^{(t,j)}$$

GHz clock speed, where a fixed number of 10 cores were used to prevent any major deviations in the computation speeds caused by the parallelization. The examples in Sections 6.4 and 6.5 have been run on a computer using 60 CPUs (out of 64) with 2.10 GHz clock speed. All our implementations and experiments are publicly available in the repository (<https://github.com/berild/inla-mc>). Additional examples are presented in the supplementary materials.

6.1. Bivariate Linear Model

In the first example, we repeat the simulated study in Gómez-Rubio and Rue (2018) and consider a simple linear model. 100 responses are simulated from

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \sim N(0, \tau), \quad \text{for } i = 1, \dots, 100.$$

Covariates x_1 and x_2 are simulated from a uniform distribution between 0 and 1 while the error terms ϵ_i are simulated from a

standard normal distribution (i.e., precision is $\tau = 1$). Moreover, we set $\beta_0 = \beta_1 = 1$, and $\beta_2 = -1$.

This model can be easily fitted using INLA, and since the likelihood is Gaussian, results are exact up to an integration error. Gómez-Rubio and Rue (2018) use this example to compare the MCMC-INLA approximations with the exact MCMC and INLA results and to show how MCMC-INLA gives also access to some joint posterior inference, for example the joint posterior of β_1 and β_2 that INLA cannot provide. We repeat this example to show that both IS-INLA and AMIS-INLA can reach the same results in just a fraction of the time used by MCMC-INLA.

For this model we have $\mathbf{z} = (\beta_0, \beta_1, \beta_2, \tau)$, and we set $\mathbf{z}_c = (\beta_1, \beta_2)$ and $\mathbf{z}_{-c} = (\beta_0, \tau)$. As in Gómez-Rubio and Rue (2018), the proposal in MCMC-INLA is a bivariate Gaussian with mean equal to the previous state $\boldsymbol{\beta}^{(j)}$ and variance of $0.75^2 \cdot \mathbf{I}$. We set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ as starting value. Both IS-INLA and AMIS-INLA use as first proposal distribution a bivariate Gaussian with mean $\boldsymbol{\mu}_0 = \mathbf{0}$ and covariance $\boldsymbol{\Sigma}_0 = 5 \cdot \mathbf{I}$. Figure 1 shows how the initial proposal distribution for β_1 changes after the preliminary step in IS-INLA and during the adaptation process in AMIS-INLA. In this case the preliminary step in IS-INLA seems to be sufficient to correctly locate the target. The adaptation process in AMIS-INLA could have been stopped earlier giving faster computing time.

Figure 2 shows the approximated posterior marginals of β_0 , β_1 , β_2 , and τ from the combined approaches, while Figure 3(a–c) show the estimated joint posterior for (β_1, β_2) . Posterior marginals from INLA alone and true values of the parameters are included for reference. All methods seem to be able to recover the parameters. MCMC-INLA seems to be the method most affected by Monte Carlo error, visible both in marginals and joint distributions.

Figure 3(d) shows the running ESS, as in Equation (13) for all combined approaches. Clearly, MCMC-INLA has achieved fewer effective samples in longer time. IS-INLA, which in this case is the most efficient method, achieved 49.2 effective samples per second, AMIS-INLA 19.5 effective samples per second, MCMC-INLA managed only 0.35 effective samples per second.

Finally, the different numerical and graphical diagnostics discussed in Section 4.2 have been computed to assess the quality of the estimates provided by IS with INLA. The sample sizes $n_e(h)$ are 9138 for IS-INLA and 9618 for AMIS-INLA. For MCMC, the minimum sample size achieved among all the parameters is 1121. Similarly, the probability plots (not shown) provide a curve that is very close to the identity line, which points to a very good estimate of the posterior marginal distributions.

6.2. Bayesian Lasso

The Lasso is a popular linear regression method that also provides variable selection (Tibshirani 1996). For a model with Gaussian likelihood, the Lasso tries to estimate the regression coefficients by minimizing

$$\sum_{i=1}^N \left(y_i - \alpha - \sum_{j=1}^{n_\beta} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{n_\beta} |\beta_j|, \quad (23)$$

where y_i is the response variable, and x_{ji} the associate covariates. N is the number observations and n_β the number of covariates.

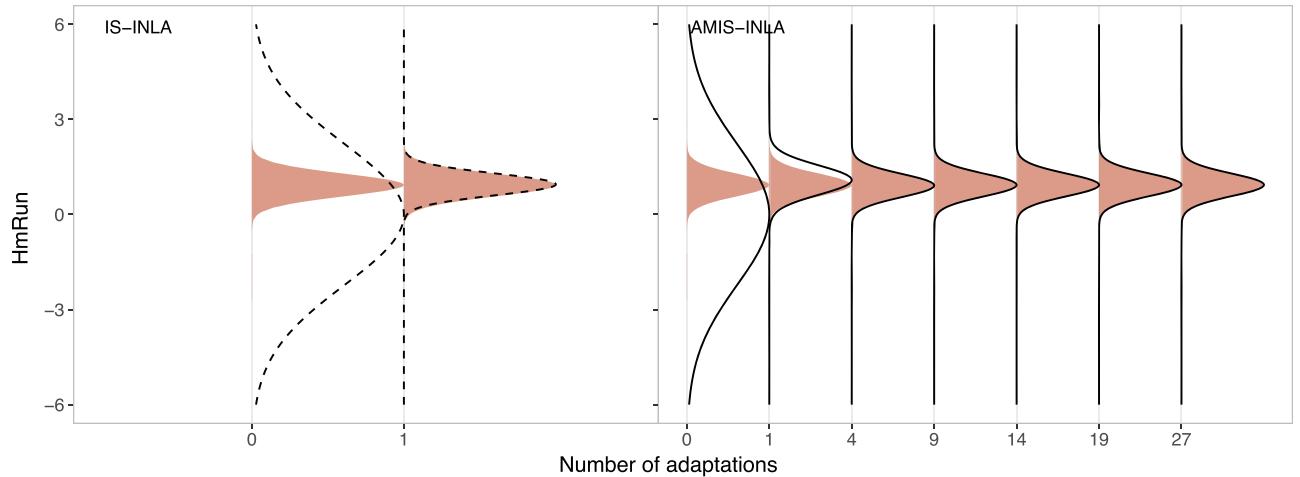


Figure 1. A visual representation of the initial search in IS-INLA (—, left) and the adaptation of proposal distribution in AMIS-INLA (—, right) for β_1 in the bivariate linear model. The x-axis is the number of adaptations of the proposal distribution. The lines (—, - -) are the proposal distributions and the filled area (■) denotes the target density.

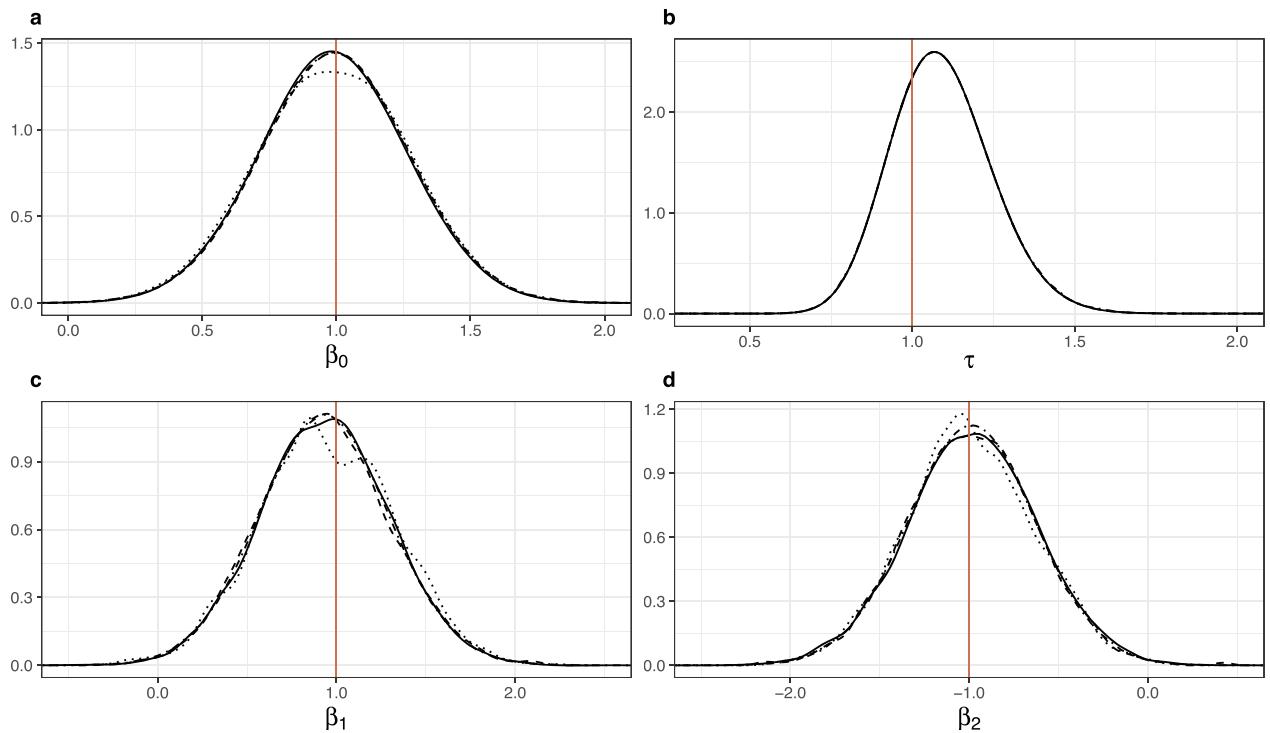


Figure 2. Posterior marginals of all parameters in the bivariate linear model approximated with AMIS-INLA (—), IS-INLA (- - -), MCMC-INLA (· · ·), and INLA (- · - · -). The line (|) is the value of the parameter chosen for the simulation of data.

The shrinkage of the coefficients is controlled by the regularization parameter $\lambda > 0$. Larger values of λ result in larger shrinkage that is, coefficients tend more toward zero. Using $\lambda = 0$ would yield the maximum likelihood estimates.

In a Bayesian setting, the Lasso can be regarded as a standard regression model with Laplace priors on the variable coefficients. The Laplace distribution is

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

where μ is a location parameter and $\sigma > 0$ a scale parameter corresponding to the inverse of the regularization parameter $\sigma = 1/\lambda$. The Laplace prior is not available for the latent field

in R-INLA, but the model is simple to fit if we condition on the values of the β coefficients.

We use the `Hitters` dataset (James et al. 2013), available in the `ISLR` R package (James et al. 2017), that contains several statistics about players in the Major League Baseball, including salary in 1987. Following Gómez-Rubio and Rue (2018), we want to predict the player's salary in 1987 based on $n_\beta = 5$ covariates, see Gómez-Rubio and Rue (2018) for details on the model and the choice of priors.

MCMC-INLA uses a multivariate Gaussian proposal distribution for $\beta^{(j)}$ with mean equal to the previous sample $\beta^{(j-1)}$ and precision $4 \cdot \mathbf{X}^\top \mathbf{X}$, as Gómez-Rubio and Rue (2018) reported good acceptance rates using this proposal. Here, \mathbf{X} is the model

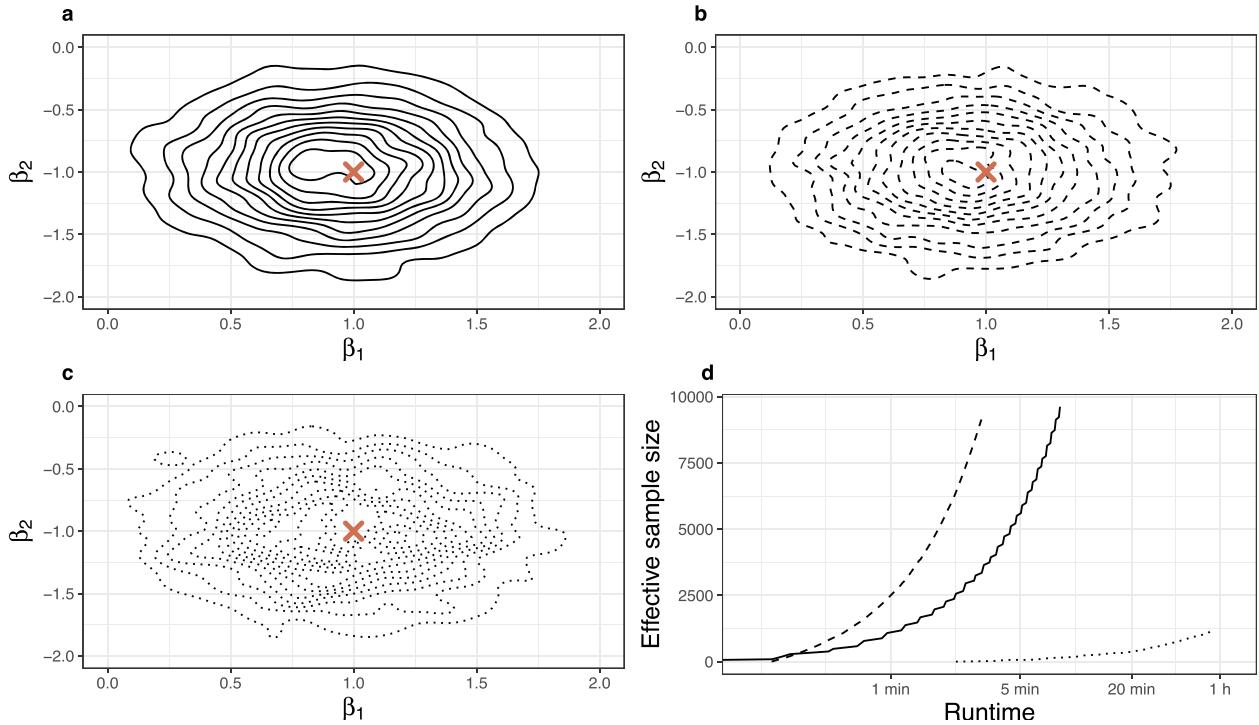


Figure 3. The joint posterior distribution of β in the bivariate linear model obtained using AMIS-INLA (—, a), IS-INLA (---, b), MCMC-INLA (· · · ·, c), and the running effective sample size (d) of the respective methods. The (X) denotes the values of β chosen for the simulation of data.

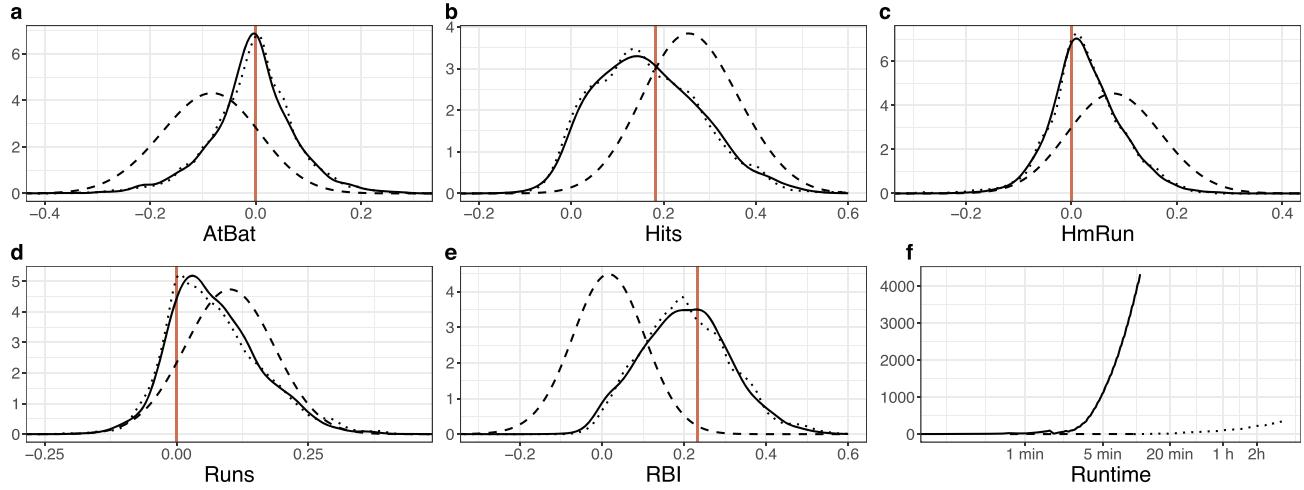


Figure 4. Approximate posterior marginals of the coefficients of the Bayesian Lasso model (a–e) fitted with AMIS-INLA (—), IS-INLA (---) and MCMC-INLA (· · · ·), and the Lasso estimates of the coefficients (|). The running effective sample sizes of the respective methods are shown in (f), where runtimes are presented in a logarithmic scale.

matrix with the individual observations as rows and the different covariates as columns. We set the initial state to $\beta^{(0)} = \mathbf{0}$. For the IS-INLA and AMIS-INLA methods, we use a multivariate Student's- t proposal with $\nu = 3$ and initial parameters $\mu_0 = \mathbf{0}$ and $\Sigma_0 = (\mathbf{X}^\top \mathbf{X})^{-1}$.

Figure 4(a)–(e) shows the estimated posterior marginals for the five coefficients. MCMC-INLA and AMIS-INLA provide similar estimates of the coefficients, with the posterior mode closely matching the Lasso regression estimates. On the contrary, IS-INLA does not provide accurate results. The problem here is that the preliminary 800 samples are not enough to correctly locate the proposal density. Figure 5 illustrates the problem occurring when the dimensionality of \mathbf{z}_c is high, as few

good samples are obtained in the preliminary steps the variance of the estimator for the mean and variance in Equation (20) is large and, thus, the estimated proposal distribution is poor. We could have used more samples in the preliminary step and make the IS-INLA work, but our point here is to show that AMIS-INLA requires less tuning in order to work well.

Figure 4(f) shows the running effective sample size. We get an effective sample size of 2784 for MCMC and one of 4321 for AMIS based on their 10,000 generated samples.

Effective sample sizes $n_e(h)$ for IS-INLA are about just 4, very low compared to the 4321 for AMIS-INLA. This points to the fact that AMIS-INLA provides more accurate estimates in this case. Note that in this way it is possible to assess the quality of

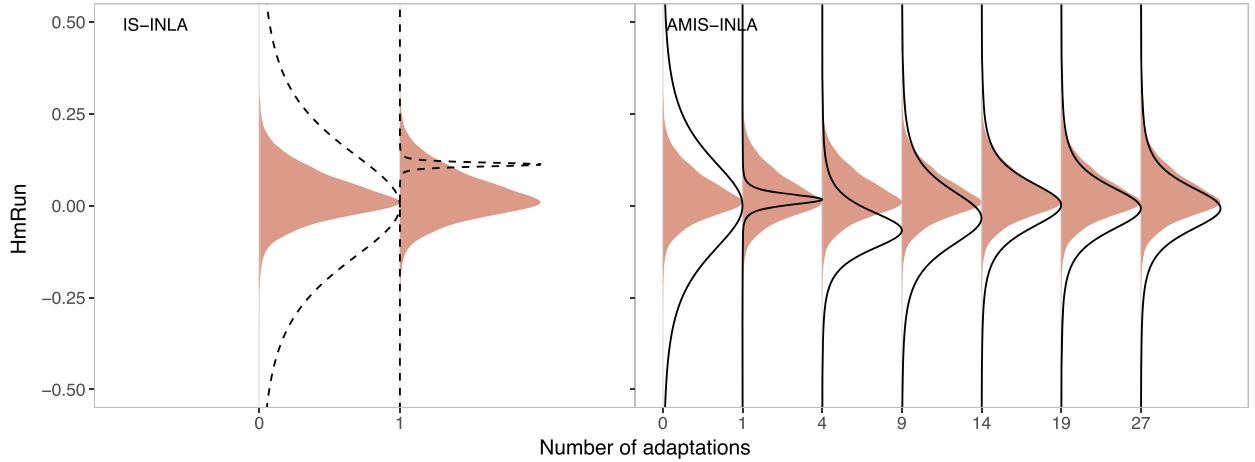


Figure 5. A visual representation of the initial search in IS-INLA (—) and adaptation of the proposal distribution in AMIS-INLA (—) for HmRun in the Bayesian lasso model. The x-axis is the number of adaptations of the proposal distribution and the fill (■) is the target density.

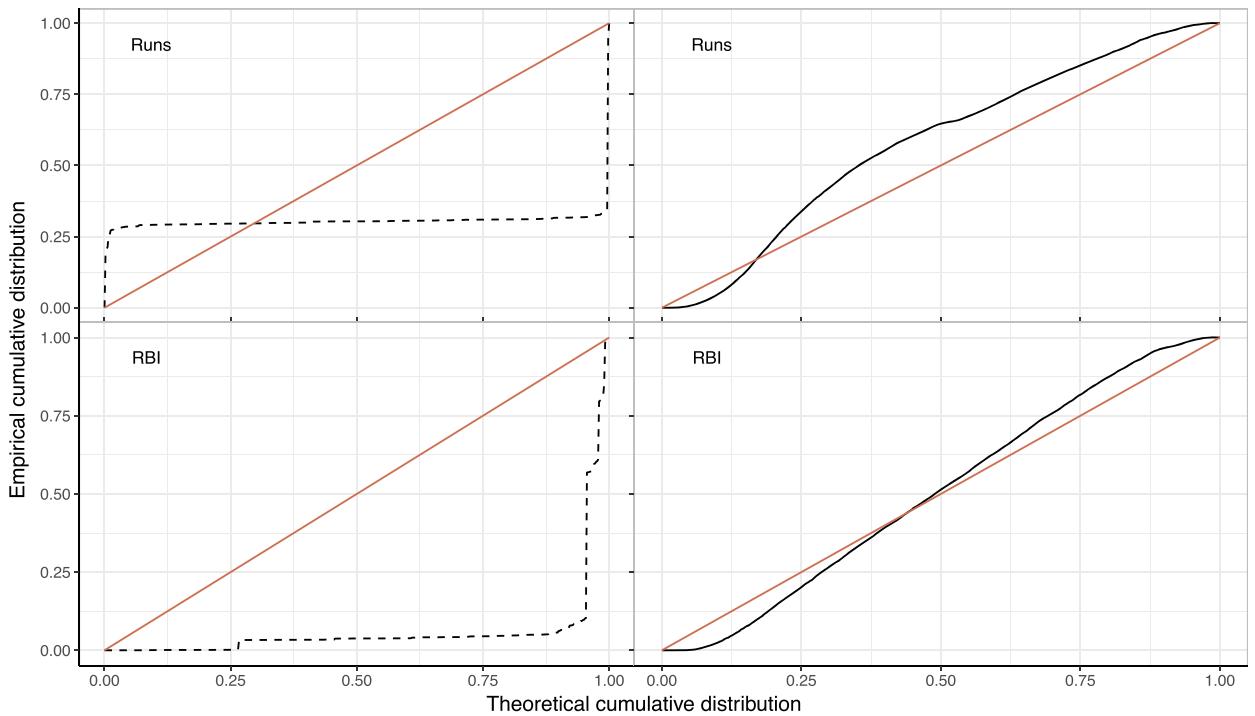


Figure 6. Probability plots for Runs and RBI parameters in the Bayesian lasso model obtained with IS-INLA (—) and AMIS-INLA (—). The comparison line (✓) denotes equivalent empirical and theoretical cumulative distributions.

the different IS estimates. Figure 6 shows the probability plots for β_4 and β_5 for IS-INLA and AMIS-INLA to assess the estimate of their posterior marginals from the weights and sample. This confirms that AMIS-INLA should be preferred in this case and illustrates the use of the IS diagnostics introduced in Section 4.2.

6.3. Spatial Autoregressive Combined Model

The next example is taken from Gómez-Rubio, Bivand, and Rue (2020) and deals with spatial econometric model (SEM; see, LeSage and Pace 2009 for a thorough account). In particular, we consider the spatial autoregressive combined (SAC) model

proposed by Manski (1993), where the response \mathbf{y} is modeled by an autoregressive term on the response:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\u}, \quad (24)$$

Here, the data is collected over n areas and \mathbf{X} are the covariates with effect $\boldsymbol{\beta}$, \mathbf{W} is the adjacency matrix of the n areas, ρ is a spatial autocorrelation parameter and $\mathbf{W} \mathbf{X}$ are the lagged covariates with effect $\boldsymbol{\gamma}$. Finally, $\boldsymbol{\u}$ is an error term modeled with a spatial autoregressive term on the error term as

$$\boldsymbol{\u} = \lambda \mathbf{W} \boldsymbol{\u} + \boldsymbol{\epsilon}_1, \quad (25)$$

where λ is another spatial autocorrelation parameter and $\boldsymbol{\epsilon}_1$ is Gaussian noise term with zero mean and precision $\tau \mathbf{I}$.

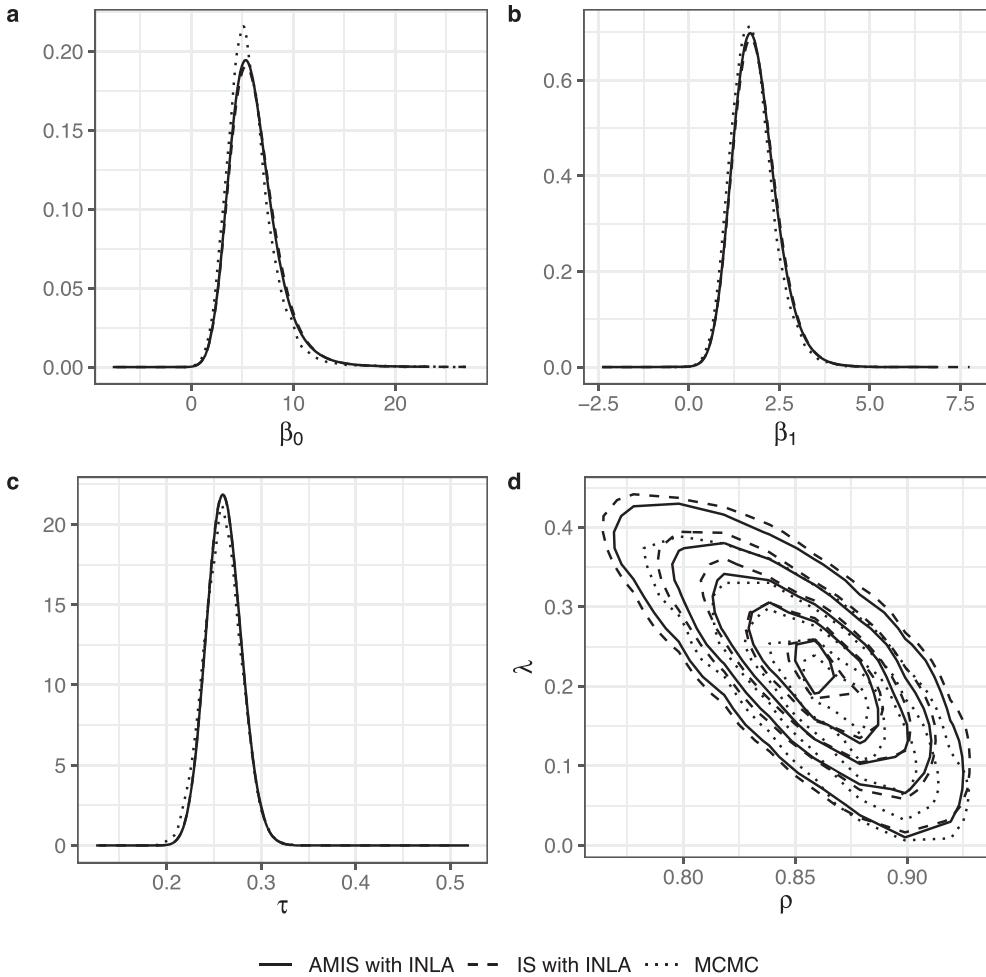


Figure 7. Posterior marginals of the intercept β_0 (a), coefficient of log GDP per capita β_1 (b), and the precision of the noise τ (c), and joint posterior distribution of the autoregressive terms ρ and λ (d) in the SAC model approximated with AMIS-INLA, IS-INLA, and MCMC.

The $n \times n$ adjacency matrix \mathbf{W} is constructed such that if the areas i and j are neighbors, the element (i, j) in \mathbf{W} is 1. The matrix is then standardized so that every row sums to one. This makes the spatial autocorrelation parameters ρ and λ bound to the interval $(1/\lambda_{\min}, 1)$, where λ_{\min} is the minimum eigenvalue of \mathbf{W} .

Equation (24) is then rewritten as

$$\mathbf{y} = (\mathbf{I} - \rho \mathbf{W})^{-1} (\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}) + \boldsymbol{\epsilon}_2, \quad (26)$$

with the revised error term:

$$\boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \tau(\mathbf{I} - \rho \mathbf{W}^\top)(\mathbf{I} - \lambda \mathbf{W}^\top)(\mathbf{I} - \lambda \mathbf{W})(\mathbf{I} - \rho \mathbf{W})). \quad (27)$$

Note that, because of the nonadditive term $(\mathbf{I} - \rho \mathbf{W})$ in Equation (26) and the complex structure in Equation (27), the model cannot be fit with INLA unless we condition on $\mathbf{z}_c = (\rho, \lambda)$.

As in Gómez-Rubio, Bivand, and Rue (2020), we consider the turnover dataset described in Ward (2008). This contains election turnovers in Italy from the 2001 elections, together with the GDP per capita (GDPCAP) from 1997 for $n = 477$ areas (*collegi* or single member districts). We apply the SAC model using turnover as response and log(GDPCAP) as covariate. See Gómez-Rubio, Bivand, and Rue (2020) for details about the data, the model and the choice of the priors.

Here we compare IS-INLA and AMIS-INLA with a standalone MCMC algorithm for SAC models available in the R package *spatialreg* (Bivand, Pebesma, and Gómez-Rubio 2013; Bivand and Piras 2015). After burn in, 10,000 MCMC samples are produced thinning a longer chain to reduce auto-correlation.

IS-INLA and AMIS-INLA algorithm use a bivariate Student's t proposal with initial parameters $\mu_0 = \mathbf{0}$ and $\Sigma_0 = 2 \cdot \mathbf{I}$. The posterior marginals of the intercept β_0 , the effect of log GDP per capita β_1 , the precision of the noise τ and the joint posterior of the spatial autoregressive terms in \mathbf{z}_c are shown in Figure 7. All estimates seem to agree very well.

The effective sample size obtained with the MCMC algorithm was 250 for the ρ parameter and 290 for the λ parameter in 99 sec a 2.52 effective sample size per second. Note, that the MCMC implementation in *spatialreg* is highly optimized for the SAC model. The IS-INLA method found 3222 effective samples in 62 min, 0.86 effective samples per second, and AMIS-INLA 4999 in 75 min, resulting in 1.1 effective samples per second.

MCMC-INLA has been omitted here because of its low performance. It managed to obtain just 64 effective samples in more than 8 hr.

6.4. Zero-Inflated Poisson

Excess of zeroes is a common problem when modeling count data. Zero-inflated distributions account for this excess of zeroes by including a mixture between a probability mass function with all its mass at zero (observed with probability p) and another distribution, usually a Poisson, from which a value is observed with probability $1 - p$. The probability p may depend on additional covariates, so that each observation has a different probability in the mixture. The R-INLA software includes different types of zero inflated distributions, including the zero-inflated Poisson (ZIP) distribution. However, while the mean of the Poisson distribution can depend on covariates or random effects, p is a common parameter, which means that all observations have the same probabilities of being a zero.

Here, we model the number of fishes y_i , $i = 1, \dots, 250$ caught by 250 groups of people that went to a park (data from <https://stats.idre.ucla.edu/r/dae/zip/>). The dataset includes, for each group: the number of children (child_i), the number of people (people_i), and whether or not the group brought a camper into the park (camper_i).

We assume the following model:

$$\begin{aligned} y_i | z_i, \mu &\sim \text{ZIP}(p_i, \mu_i), i = 1, \dots, 250 \\ \text{logit}(p_i) &= \gamma_0 + \gamma_1 \text{people}_i \\ \log(\mu_i) &= \beta_0 + \beta_1 \text{child}_i + \beta_2 \text{camper}_i \\ \gamma_0, \gamma_1 &\sim N(0, 0.001) \\ \beta_0, \beta_1, \beta_2 &\sim N(0, 0.001) \end{aligned}$$

Here, p_i represents the probability of observing $y_i = 0$. Otherwise, with probability $1 - p_i$, y_i comes from a Poisson with mean μ_i . Note that both p_i and μ_i are modeled in the appropriate scale by using suitable link functions. Vague priors are used for the intercepts and the coefficients of the covariates.

We take $\mathbf{z}_c = (\gamma_0, \gamma_1)$. Conditional on these two parameters, the model can be fitted in R-INLA using a different likelihood for each data point. In this way it is possible to modulate the p_i parameters according to the covariates and we use IS/AMIS-INLA to estimate the full model.

To sample γ_0 and γ_1 we use a Normal distribution centered at the maximum likelihood estimates, with standard deviation equal to the ML estimate multiplied by three to allow for ample variation in the samples. ML estimates have been obtained with the `zeroinfl()` function from the `pscl` package (Zeileis, Kleiber, and Jackman 2008). In particular the parameters (mean and standard deviation) of the sampling distributions for γ_0 and γ_1 have been $(1.30, 1.20)$ and $(-0.56, 0.48)$, respectively.

Table 1 shows the estimates of the model parameters using 10000 iterations of IS using the sampling distribution stated

Table 1. Summary of estimates for the zero-inflated model.

Parameter	ML		IS-INLA		AMIS-INLA		MCMC	
	Estimate	SE	Mean	SD	Mean	SD	Mean	SD
γ_1	1.297	0.374	1.323	0.408	1.330	0.386	1.354	0.390
γ_2	-0.564	0.163	-0.579	0.184	-0.585	0.170	-0.592	0.172
β_1	1.598	0.086	1.597	0.086	1.596	0.086	0.755	0.177
β_2	-1.043	0.100	-1.045	0.100	-1.048	0.100	-1.053	0.099
β_3	0.834	0.094	0.835	0.094	0.835	0.094	0.839	0.095

above (i.e., there is no initial adaptation), 10,000 iterations of AMIS (2000 initial simulations plus 4 adaptive steps with 2000 simulations each) and MCMC (using a burn-in of 10,000 iterations, plus 50,000 iterations of which only 1 in 10 has been kept). For completeness, we have also included the ML estimates (estimate and standard error). The results illustrate that IS and AMIS provide accurate estimates for all model parameters except β_1 when compared to the MCMC results. We can only argue that this difference may be due to the parameterization of the model in JAGS as the IS and AMIS results are also close to the ML estimates.

6.5. Poisson Mixture

Zucchini, MacDonald, and Langrock (2016) analyze the number of major earthquakes (magnitude equal to 7 or greater) per year in the period 1900–2006 using a Poisson mixture. The aim is to classify each year as “low rate” or “high rate.”

The number of major earthquakes y_i can be regarded as an observation from a mixture of two Poisson distributions with means μ_1 and μ_2 that represent low and high rate, respectively. Hence, it can be assumed that $\mu_1 < \mu_2$, and this will be encoded in the model by using appropriate priors (see below).

The model can be stated by using discrete indicator variables $z_i \in \{1, 2\}$ as follows:

$$\begin{aligned} y_i | z_i, \mu &\sim Po(\mu_{z_i}), i = 1, \dots, n \\ \Pr(z_i = j) &= 0.5, j = 1, 2 \\ \mu_1 &\sim N(10, 0.01) \\ \mu_2 &\sim N(30, 0.01) \end{aligned}$$

For given values of the indicator variables, the model simply consists of two separate models (one for each group) and can be easily fit with R-INLA using two likelihoods.

We set therefore, $\mathbf{z}_c = (z_1, \dots, z_n)$ and use IS to sample values of the indicator variables. The sampling distribution is:

$$g(z_i = j) \propto w_j Po(y_i | \mu'_j); j = 1, 2$$

with $\mu'_1 = 14.60$, $\mu'_2 = 26.18$, and $w_1 = w_2 = 0.5$. Note that the values of the means have been obtained with the k -means algorithm for two groups and that this probably provides a close approximation to the posterior distribution of the mixture. Also, note that the AMIS algorithm will update the parameters of the sampling distribution at each adaptive step.

IS has been run for 10,000 iterations (with no previous adaptation step) to achieve an effective sample size $n_e(h)$ of 313. Similarly, AMIS with INLA has been run for 10,000 iterations (using an initial 2000 iterations plus 4 adaptive steps with 2000 iterations each) to achieve an effective sample size $n_e(h)$ of 1077. MCMC estimates are based on 1000 samples obtained after 10,000 burn-in samples followed by another 10,000 samples (of which only 1 in 10 has been kept) obtaining 2517 effective samples for μ_1 and 2807 for μ_2 . The summary estimates of the model parameters can be seen in Table 2.

Table 2. Estimates of parameters of the mixture Poisson model fit to the earthquake data.

Parameter	IS-INLA		AMIS-INLA		MCMC	
	Mean	SD	Mean	SD	Mean	SD
μ_1	14.72	0.60	14.64	0.64	14.61	0.67
μ_2	24.99	0.86	24.84	0.91	24.82	0.91

7. Discussion

The integrated nested Laplace approximation is a suitable approach for approximate Bayesian inference for latent Gaussian models, as described in Rue, Martino, and Chopin (2009). Extending the use of INLA to other classes of models has been considered by several authors using INLA together with numerical integration or MCMC methods. Here, we have illustrated a novel approach to extend the models that INLA can fit by combining importance sampling and adaptive multiple importance sampling with INLA.

This new approach has a number of advantages over other similar approaches. First of all, importance sampling is a very simple algorithm that can also be easily parallelized, leading to a huge computational speed up. This means that, in practice, times for model fitting remain small. In the examples developed in this article we have illustrated how IS and AMIS with INLA are able to fit a wide range of models. Furthermore, the numerical experiments conducted show that the approximations of the posterior marginals obtained with IS and AMIS with INLA are also accurate and close to the actual posterior marginals.

This article also discusses numerical and graphical diagnostics to assess the accuracy of IS/AMIS when used in combination with INLA to fit models. We have observed that the different criteria usually agree, with small effective sample sizes associated to poor estimates of the posterior marginal distribution of some of the model parameters. Hence, these criteria can effectively be used to critically assess the quality of the estimates produced by IS/AMIS with INLA. In this sense, in the examples developed in the paper AMIS seemed to provide better estimates when used in combination with INLA for model fitting.

Supplementary Materials

INLA-IS-supplementary.pdf: Additional examples not included in this paper (Imputation of missing covariates and Bayesian quantile regression). **IS-INLA-code:** A folder containing the computer code used in this article. Within this folder you find the IS-INLA, AMIS-INLA, and INLA-MH algorithm within the *inlaMC* folder. The code for the respective examples are found in *toy*, *zip*, *sem*, *pqr*, *pois_mix*, *missing*, and *lasso*. All functions creating the figures and datasets that are not found in R-libraries are also available within. The figures used in the article are found in the *figures* folder and the results of the simulations are found in *sims*. Also included is a *readme.md*(*readme.html*) file explaining the code in more detail. This exact code is also publicly available in the GitHub repository <https://github.com/berild/inla-mc> which we have linked to in the article.

Funding

V. Gómez-Rubio has been supported by grant SBPLY/17/180501/000491, funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain)

and FEDER, and grants MTM2016-77501-P and PID2019-106341GB-I00, funded by Ministerio de Ciencia e Innovación (Spain).

ORCID

Virgilio Gómez-Rubio  <http://orcid.org/0000-0002-4791-3072>

References

- Bivand, R., and Piras, G. (2015), “Comparing Implementations of Estimation Methods for Spatial Econometrics,” *Journal of Statistical Software*, 63, 1–36. [10]
- Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2014), “Approximate Bayesian Inference for Spatial Econometrics Models,” *Spatial Statistics*, 9, 146–165. [1,2]
- Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2015), “Spatial Data Analysis with R-INLA with Some Extensions,” *Journal of Statistical Software*, 63, 1–31. [1,2]
- Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013), *Applied Spatial Data Analysis with R*, Volume 10 of Use R (2nd ed.), New York: Springer. [10]
- Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. (2017), “Adaptive Importance Sampling: The Past, the Present, and the Future,” *IEEE Signal Processing Magazine*, 34, 60–79. [5]
- Corneut, J.-M., Marin, J.-M., Mira, A., and Robert, C. P. (2012), “Adaptive Multiple Importance Sampling,” *Scandinavian Journal of Statistics*, 39, 798–812. [2,5]
- Elvira, V., Martino, L., and Robert, C. P. (2018), Rethinking the Effective Sample Size, arXiv:1809.04129 [stat.CO]. [3,4,5]
- Gilks, W., Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, Boca Raton, FL: Chapman & Hall. [1]
- Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2020), “Bayesian Model Averaging with the Integrated Nested Laplace Approximation,” *Econometrics*, 8, 1–15. [1,2,4,5,9,10]
- Gómez-Rubio, V., and Palmí-Perales, F. (2019), “Multivariate Posterior Inference for Spatial Models with the Integrated Nested Laplace Approximation,” *Journal of the Royal Statistical Society, Series C*, 68, 199–215. [1,2,4]
- Gómez-Rubio, V., and Rue, H. (2018), “Markov Chain Monte Carlo with the Integrated Nested Laplace Approximation,” *Statistics and Computing*, 28, 1033–1051. [1,2,3,4,5,6,7]
- Hubin, A., and Storvik, G. (2016), “Estimating the Marginal Likelihood with Integrated Nested Laplace Approximation (INLA),” arXiv:1611.01450 [stat.CO]. [2,4]
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: With Applications in R*, Volume 103 of Springer Texts in Statistics, New York: Springer. [7]
- James, G., Witten, D., Hastie, T., and Tibshirani, T.R. (2017), “ISLR: Data for an Introduction to Statistical Learning with Applications in R,” R package version 1.2. [7]
- LeSage, J., and Pace, R. K. (2009), *Introduction to Spatial Econometrics* (1st ed.), Boca Raton, FL: Chapman and Hall/CRC. [9]
- Manski, C. F. (1993), “Identification of Endogenous Social Effects: The Reflection Problem,” *The Review of Economic Studies*, 60, 531–542. [9]
- Owen, A. B. (2013), “Monte Carlo Theory, Methods and Examples,” available at <https://statweb.stanford.edu/~owen/mc/>. [4]
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006), “Coda: Convergence Diagnosis and Output Analysis for mcmc,” *R News*, 6, 7–11. [5]
- Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer Texts in Statistics (2nd ed.), New York: Springer. [3]
- Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields. Theory and Applications*, Boca Raton, FL: Chapman & Hall/CRC. [2]

- Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society, Series B*, 71, 319–392. [1,2,12]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [6]
- Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S* (4th ed.), New York: Springer. [3]
- Ward, M. D. (2008), *Spatial Regression Models*, Volume 07-155 of Quantitative Applications in the Social Sciences, Los Angeles, CA: Sage. [10]
- Zeileis, A., Kleiber, C., and Jackman, S. (2008), "Regression Models for Count Data in R," *Journal of Statistical Software*, 27, 1–25. [11]
- Zucchini, W., MacDonald, I., and Langrock, R. (2016), *Hidden Markov Models for Time Series: An Introduction Using R* (2nd ed.), Boca Raton, FL: CRC Press. [11]