Martin Outzen Berild

Integrated Nested Laplace Approximations within Monte Carlo Methods

Master's thesis in Applied Physics and Mathematics

Supervisor: Sara Martino

June 2020

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Martin Outzen Berild

Integrated Nested Laplace Approximations within Monte Carlo Methods

Master's thesis in Applied Physics and Mathematics
Supervisor: Sara Martino
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

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. In practice, applications of INLA are limited to the class of models implemented in the R package **R-INLA**. 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 thesis considers the three following *INLA within Monte Carlo methods*: Markov chain Monte Carlo (MCMC) with INLA, importance sampling (IS) with INLA, and a novel approach that combines INLA and the adaptive multiple importance sampling (AMIS) algorithm.

This thesis compares the INLA within Monte Carlo methods on a series of applications with simulated and observed datasets and evaluates their performance based on accuracy, efficiency, and robustness. The implementation of the methods are validated by exact posteriors in a simple bivariate linear model and tested on a spatial autoregressive combined model. Then, it presents a new approach to Bayesian quantile regression using AMIS with INLA, which is verified in a simulation study and applied to two observed datasets. Also, this thesis attempts to approximate the posteriors in a Gamma frailty model using AMIS with INLA.

The examples show that the AMIS with INLA approach, in general, outperformed the other methods on more complex models, but the IS with INLA algorithm could be considered for faster inference when good proposals are available. Also, the Bayesian quantile regression approach produced promising quantile curves in the simulation study, and the applications present a small portion of the large class of models that are facilitated through INLA for this type of quantile regression. In addition, the AMIS with INLA algorithm produced accurate posteriors in the Gamma frailty model with few clusters but attained a slight bias for a higher number of dimensions in the AMIS algorithm.

Sammendrag

Integrated Nested Laplace Approximations (INLA) er en deterministisk metode for å oppnå bayesiansk inferens på latente gaussiske modeller (LGMer) og fokuserer på raske og nøyaktige approksimasjoner av marginale posteriori-fordelinger for parametrene i en modell. I praksis er applikasjonene av INLA begrenset til de modellene som er implementert i R pakken **R-INLA**. I senere år har det blitt utviklet flere metoder for å utvide disse modellene til de som kan uttrykkes som betingede LGMer ved å fiksere noen av parameterne i modellen til beskrivende verdier. Metodene er forskjellige i hvordan de velger disse beskrivende verdiene. Denne oppgaven betrakter de tre følgende INLA med Monte Carlo-metodene: Markov chain Monte Carlo (MCMC) med INLA, importance sampling (IS) med INLA, og en ny metode som kombinerer INLA og adaptive multiple importance sampling (AMIS).

Denne masteroppgaven sammenligner INLA med Monte Carlo-metodene på flere applikasjoner med simulerte og observerte datasett, og vurderer deres ytelse basert på nøyaktighet, effektivitet og robusthet. Implementeringen av metodene er validert av eksakte posteriori-estimater i en enkel bivariat lineær modell og testet på en romlig autoregressiv kombinert modell. Deretter presenterer den en ny tilnærming til bayesiansk kvantileregresjon ved bruk av AMIS med INLA, som er evaluert i en simuleringsstudie og anvendt på to observerte datasett. Denne oppgaven prøver også å approksimere posteriori-verdier i en Gamma frailty modell ved å bruke AMIS med INLA.

Resultatene fra eksemplene indikerer at AMIS med INLA-metoden generelt gjorde det bedre enn de andre metodene på mer komplekse modeller, men IS med INLA-algoritmen kan vurderes for raskere inferens når det er lett å velge forslagsfordeling. Den bayesianske kvantileregresjonen produserte lovende kvantilekurver i simuleringsstudien, og applikasjonene på observerte datasett presenterer en liten del av alle modellene som er tilgjengelig gjennom INLA for denne typen for kvantileregresjon. I tillegg produserte AMIS med INLA-algoritmen nøyaktige posterioreri-resultater i Gamma frailty modellen med få grupper, men viste en viss grad av bias for et større antall dimensjoner i AMIS.

Preface

This Master's thesis concludes my five year Master of Science degree in *Applied Physics and Mathematics* with specialization in *Industrial mathematics* at the Norwegian University of Science and Technology (NTNU).

The work presented in my thesis is a continuation of my specialization project (Berild, 2020). It has been gratifying to work on a new method for Bayesian inference, and both exciting and challenging, having such a large playing field of models in the applications.

I would like to thank my supervisor, Associate Professor Sara Martino, for guiding me towards such an interesting topic and the numerous video calls during these special times. I would also like to thank my family for their continued support throughout the years, and my friends and fellow students for my incredible time at NTNU.

Martin Outzen Berild
Trondheim, June 2020

Contents

Chapter 1

Introduction

In the realm of Bayesian inference, there is no distinction between unknown quantities, and all are considered random variables. Prior knowledge about the phenomenon being modeled allows us to formulate prior distributions and likelihood functions relating the unknown quantities to observations, and inference is based on the posterior distribution obtained from Bayes' theorem. A large and frequently used model in Bayesian inference is hierarchical models.

The most common approach to inference on hierarchical models is Markov chain Monte Carlo (MCMC, Gilks et al., 1996). The technique constructs Markov chains with posteriors of interest as limiting distribution, and provides arbitrarily accurate results, depending on the number of posterior samples. However, widespread and revolutionary, MCMC has its drawbacks. The process requires a lot of CPU-time, and there is no great way to run computations in parallel. Also, the manual tweaking of model parameters and subsequent re-running of simulation to achieve convergence of the Markov chain, sum up to an underestimated time investment in view of the applied user.

An alternative well-known Monte Carlo technique is the class of importance sampling (IS, Robert et al., 2004) methods. The standard IS technique draws samples from a single proposal distribution and assigns them weights according to the dissimilarity between the target and the proposal distribution, and the performance of the IS methods highly depends on the choice of proposal distributions. In general, inference on hierarchical models using IS is relatively challenging because of the usually high dimensional target distributions. Several advanced IS methods have been proposed to produce more robust algorithms. One effective approach is to employ a population of proposal distributions, namely multiple importance sampling (MIS, Elvira et al., 2019), which avoids entrusting the performance of the algorithm to one single proposal distribution. Another effective method is to gradually increase the performance of the algorithm by sequentially adapting the proposal distribution to more accurately approximate the target distribution. This leads to the concept of adaptive IS (AIS; Bugallo et al., 2017) and, furthermore, employing the population of adapted proposal distribution outlines the promising strategy called adaptive multiple IS (AMIS, Corneut et al., 2012).

Rue et al. (2009) introduced a deterministic approach to approximate Bayesian inference for hierarchical models that can be represented as latent Gaussian models (LGMs). This new approach, called the integrated nested Laplace approximation (INLA), focuses on approximating the posterior marginals, and it is argued to outperform MCMC methods in both accuracy and speed (Rue et al., 2009). The INLA approach is implemented in the `R` package **R-INLA**, and is available at `http://www.r-inla.org` (Rue, 2020). In practice, fitting models with INLA is restricted to the class of models available in **R-INLA**, as the implementation of INLA is a demanding process. The framework allows for some user-defined models, but there are models and task that falls outside the scope of **R-INLA**. INLA is not able to provide joint inference on the unknown parameters, nor does it handle missing values in the covariates, and it can't have non-additive terms in the linear predictor.

Several methods have been proposed in the literature that fixes one or multiple unknown parameters in the model to representative values so that the conditional models can be fit with **R-INLA**. These methods differ in the way representative values are found. Li et al. (2012) fixed some of the parameters to their maximum likelihood estimates, thereby fitting models conditioned on the parameter estimates with **R-INLA**. However, this method ignores the uncertainty about the fixed parameters and does not produce inference about them.

Bivand et al. (2014), Bivand et al. (2015a) proposed a different method of constructing a grid on some of the parameters and fitting multiple models with **R-INLA** conditioned on the individual grid points. Inference about the parameters in the grid is achieved using the conditional marginal likelihoods approximated with INLA, and the prior distribution of the parameters. It is thereby constructing a weighted grid that can be normalized with numerical integration. The posterior marginals of the remaining parameters are obtained with Bayesian model averaging (BMA, Hoeting et al., 1999) using the conditional posterior marginals from the fitted models.

Recently, authors have proposed to generate these representative values using Monte Carlo techniques; thus, combining INLA and Monte Carlo methods, and we will refer to them in collection by the umbrella term *INLA within Monte Carlo methods*. Gómez-Rubio et al. (2018) proposed the use of Markov chain Monte Carlo techniques to generate samples from the posterior distribution of some of the parameters in the model, and apply INLA to fit the models conditioned on the generated samples. Similarly, Gómez-Rubio (2019) proposed the use of INLA within importance sampling, resulting in a more efficient sampling strategy than MCMC with INLA if provided with a suitable proposal distribution. However, finding a such distribution might be difficult in more complex models. For this reason, following the recent developments within importance sampling, we propose a novel approach based on the adaptive multiple importance sampling method.

The work presented in this thesis is a continuation of our specialization project (Berild, 2020). Therefore the theoretical development and implementation of the methods have carried over, and some code and parts of this thesis are similar.

1.1 Goals and structure

This thesis aims to give an introduction to the theory behind, and development of, the combined INLA and Monte Carlo approaches, with the core of the focus on the novel AMIS with INLA methodology. We aim to compare the INLA within Monte Carlo methods on efficiency, robustness, and accuracy, and to justify our implementations empirically by comparing the approximations to exact inference methods. Next, we aim to extend the set of models that can be fit through the **R-INLA** package, and introduce a novel application for Bayesian quantile regression. In addition, we present an application of the AMIS with INLA algorithm on a Gamma frailty model, testing the capabilities of the algorithm.
The thesis is divided into the following parts:

Chapter 2 establishes the overall theory behind all components in the combined approaches. It contains a simple introduction to Bayesian inference, and the class of models we consider in this thesis. The chapter also describes three simulation-based inference methods, MCMC, IS, and AMIS, and the approximate inference method INLA, with comments on their advantages and drawbacks.

Chapter 3 presents the INLA within Monte Carlo methods; MCMC with INLA (Gómez-Rubio et al., 2018), IS with INLA (Gómez-Rubio, 2019), and introduces the novel AMIS with INLA algorithm. Initially, we describe the approximation of conditional models in INLA, the combination of these approximations to obtain unconditional posteriors, and the general type of models compatible with the combined approaches. Then, we detail our implementation of the three methods and sketch them in pseudo-code.

Chapter 4 contains a collection of applications of the combined approaches. First, we consider a bivariate linear model, where we present the behavior of each algorithm and compare their approximations to exact posteriors. Next, we consider the spatial autoregressive combined model, comparing the results to the posteriors obtained with an MCMC algorithm. Then, we introduce a novel model-aware Bayesian quantile regression based on the INLA within Monte Carlo methods, focusing on the AMIS with INLA approach. Lastly, we attempt to approximate a Gamma frailty model for different number of clusters using AMIS with INLA.

Chapter 5 sum up the results presented in the thesis, discussing the methods advantages or drawbacks. In addition, it indicates some future applications and research.

1.2 Implementation

The INLA is only available through the R package **R-INLA**, so the conditional models are fit using its toolbox and, consequently, we use the programming language R in our implementations. The INLA within MCMC algorithm is available in the package **INLABMA**, but we have chosen to implement our function to have more control over input, output, and flow, which allows us to better compare the methods. For the AMIS with INLA and IS with INLA algorithms, there are no available tools, so their functions were implemented from scratch, and parallel computations are added wherever possible using the **parallel** package. In all simulations, a CPU with a total of 10 cores is used. All our implementations of the algorithms, models, and experiments are publicly available in the GitHub repository (`https://github.com/berild/master-thesis-code`).

Chapter 2

Bayesian Inference

In this chapter, the relevant inference methods and model architecture needed to develop the combined approaches in Chapter 3 are presented.

The Bayesian approach to inference assumes that the parameters of the model are random, and that predictive and parametric inference is achieved by updating our beliefs about the parameters in light of newly acquired information (Bernardo et al., 2000). Let us denote the parameter of interest x; for example, the effect of some covariate on response y. The core of Bayesian inference is to obtain the posterior distribution as a synthesis of our knowledge about the parameter of interest before observing the data, the prior distribution with density $\pi(x)$, and the likelihood function or conditional distribution of y given the effects $\pi(y \mid x)$, obtained from some model about the observed data. The relationship between these densities expressing the posterior density is called Bayes' rule:

$$\begin{aligned}\pi(x \mid y) =& \frac{\pi(x, y)}{\pi(y)} \\ =& \frac{\pi(y \mid x)\pi(x)}{\pi(y)} \\ =& \frac{\pi(y \mid x)\pi(x)}{\int \pi(y \mid x)\pi(x)\mathrm{d}x}.\end{aligned} \tag{2.1}$$

Here, $\pi(y)$ is the marginal likelihood which sometimes is referred to as the normalizing constant, and can be computed using the law of total probability, integrating $\pi(y|x)\pi(x)$ over all possible values of x. In general, the marginal likelihood can be hard to obtain; however, since it is not dependent on x, and is constant, many Bayesian inference methods employs Bayes' rule (2.1) in its unnormalized form:

$$\pi(x \mid y) \propto \pi(y \mid x)p(x).$$

2.1 Bayesian hierarchical models

The models we regard in this paper are of a hierarchical structure split into three stages. These models occur when the diversity of the prior information or the variability of the observations requires the introduction of several levels of prior distributions (Robert et al., 2004).

Let us consider n observations $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, where we define a likelihood model, conditioned on some latent variables $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta}_1$ as

$$\text{Stage 1: } \boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_1 \sim \pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_1).$$

This defines the first stage of our model. Furthermore, the set of latent variables given some hyperparameters $\boldsymbol{\theta}_2$ is distributed according to

$$\text{Stage 2: } \boldsymbol{x} \mid \boldsymbol{\theta}_2 \sim \pi(\boldsymbol{x} \mid \boldsymbol{\theta}_2),$$

and forms the seconds stage. To complete the hierarchical structure the third and last stage is to assign the hyperparameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ appropriate priors

$$\text{Stage 3: } \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

2.1.1 Additive Latent Gaussian models

A special class of Bayesian hierarchical models is the family of Latent Gaussian models (LGMs). These hold the necessary properties that are required by the approximate inference method detailed in Section 2.3, and are important for the development of the combined approaches presented in Chapter 3. In LGMs, the response y_i is assumed to belong to a distribution family (not necessarily the exponential family, Martins et al., 2013a), with a mean μ_i that is linked to a predictor through a link function, such that $g(\mu_i) = \eta_i$. In the predictor, the effects of the covariates are included in an additive manner, and this additive linear predictor is defined as

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f_k(u_{ki}) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

Here, α is the intercept, $\{\beta_j\}$ regulate the fixed effects of the covariates $\{\boldsymbol{z}_j\}$. Furthermore, the model components $\{f_k(\cdot)\}$ are unknown functions of the covariates $\{\boldsymbol{u}_k\}$, which map the kth covariate to the random effect or spatial effect on the response. $i = 1, \ldots, n$ represents the individual observations of the response and covariates, n_β is the total number of fixed effects, and n_f the total number of random effects and model components. Lastly, ϵ_i holds the unstructured terms. The components $f_k(\cdot)$ are used to model non-linear effects of the covariates, or spatial and temporal dependencies in the data.

We assume that the joint distribution of the unknown components in the linear predictor,

$$\boldsymbol{x} = (\alpha, \eta_1, \ldots, \eta_n, \beta_1, \ldots, \beta_{n_\beta}, f_1, \ldots, f_{n_f}),$$

is Gaussian conditioned on the hyperparameters $\boldsymbol{\theta}_2$. Note that the vector $\boldsymbol{x}$ corresponds to the second stage of a hierarchical model as defined in Section 2.1. Models where the latent field $\boldsymbol{x}$ is assigned a prior Gaussian distribution are called LGMs. The latent field in LGMs can therefore be expressed as

$$\begin{aligned} \boldsymbol{x}|\boldsymbol{\theta}_2 &\sim \mathcal{N}(\boldsymbol{x};\mathbf{0},\mathbf{Q}^{-1}(\boldsymbol{\theta}_2)) \\ &\propto |\mathbf{Q}(\boldsymbol{\theta}_2|^{1/2}\exp\{-\frac{1}{2}\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x}\}, \end{aligned} \tag{2.3}$$

where $\mathcal{N}(\cdot,\cdot)$ denotes a multivariate Gaussian distribution with zero mean and precision matrix (inverse of covariance matrix) $Q(\boldsymbol{\theta}_2)$.

We are interested in a particular type of LGMs, with conditional independence properties in the latent field, such that the precision matrix $Q(\boldsymbol{\theta}_2)$ attains a sparsity. A multivariate Gaussian field with a sparse precision matrix outline a Gaussian Markov random field (GMRF; see Rue et al., 2005). This model property is essential in the models applied to the approximate inference method detailed in Section 2.3, as it provides a substantial computational advantage (Rue et al., 2017).

Observations of the response are assumed conditionally independent given the latent field $\boldsymbol{x}$, and the vector of hyperparameters $\boldsymbol{\theta}_1$. Thereby, the first stage in the hierarchical structure of LGMs is the likelihood function defined as

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_1 \sim \prod_{i=1}^{n} \pi(y_i|\eta_i,\boldsymbol{\theta}_1), \tag{2.4}$$

where each observation y_i only depends on one element in the latent Gaussian field, the linear predictor η_i, and the hyperparameters $\boldsymbol{\theta}_1$.

To finalize the hierarchical structure of LGMs, appropriate priors are assigned to the hyperparameters of the model. With the stages set, the joint posterior distribution of all the unknown components $\boldsymbol{z} = (\boldsymbol{x},\boldsymbol{\theta})$ in the model is expressed as

$$\begin{aligned} \pi(\boldsymbol{z}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i\in\mathcal{I}}\pi(y_i|x_i,\boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\mathbf{Q}(\boldsymbol{\theta})|^{1/2}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\mathbf{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n}\ln\pi(y_i \mid \eta_i,\boldsymbol{\theta})\right\}, \end{aligned} \tag{2.5}$$

The posterior $\pi(\boldsymbol{z} \mid \boldsymbol{y})$ is most often very high dimensional, such that analytical results are not achievable. Several methods have been developed to perform inference on (2.5) and, in the following sections, we will review some of them. These include two sampling based methods (Markov chain Monte Carlo and importance sampling), and one method for approximate inference (integrated nested Laplace approximations). Each of these methods will be describe with reference to the posterior distribution $\pi(\boldsymbol{z} \mid \boldsymbol{y})$ in (2.5); even though one should be aware that both sampling based methods are general algorithms that also can be applied outside the scope of LGMs. The aim is for the reader to have an overview of such methods, as it is necessary to understand why and how, in Chapter 3, we propose to merge some of them.

2.2 Monte Carlo Methods

The classical Monte Carlo method is based on generating independent realizations of $\boldsymbol{z}$ from its probability distribution $\pi(\cdot)$. In general, this is achieved by simulating a number of independent samples for the distribution of interest and, then, using these samples as the basis for our inference on $\pi(\boldsymbol{z} \mid \boldsymbol{y})$, which is in this setting referred to as the target distribution.

Monte Carlo methods can be used to achieve different tasks, for example: approximating the target distribution

$$\boldsymbol{z} \mid \boldsymbol{y} \sim \pi(\boldsymbol{z} \mid \boldsymbol{y}), \tag{2.6}$$

estimating some quantity of interest

$$\mathbb{E}_{\pi}[(f(\boldsymbol{z})] = \int f(\boldsymbol{z})\pi(\boldsymbol{z} \mid \boldsymbol{y})\mathrm{d}\boldsymbol{z}, \tag{2.7}$$

where $f(\cdot)$ is any integratable function with respect to the target. Or optimize, i.e. obtaining posterior modes

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \ \pi(\boldsymbol{z} \mid \boldsymbol{y}). \tag{2.8}$$

However, if the distribution of interest is unknown or a non-standard distribution, realizations are unobtainable with the aforementioned approach. In this case, we must employ more sophisticated but related methodologies. We will collect these in the umbrella term of their antecedent, Monte Carlo methods. More specifically, we will describe three Monte Carlo methods; Markov Chain Monte Carlo (MCMC), Importance Sampling (IS), and Adaptive Multiple Importance sampling (AMIS), as they are relevant for the development of the combined approaches detailed in Chapter 3.

2.2.1 Markov Chain Monte Carlo

The most common methods for inference on Bayesian hierarchical models are Markov chain Monte Carlo (MCMC). It comprises many different algorithms that are all variations of the general framework proposed by Metropolis et al. (1953), and generalized to its current form by Hastings (1970), namely the *Metropolis-Hastings* algorithm by the surname of its authors. We will only describe MCMC from a general point of view, covering only the Metropolis-Hastings algorithm, and the reader is referred to Robert et al. (2004) for an extensive introduction to the theory of MCMC.

In essence, an MCMC algorithm produces an ergodic Markov chain (Robert et al., 2004, chap. 6.6) with the target distribution as limiting distribution. The Markov chain is a stochastic system with states governed by transition probabilities $p(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^{(j-1)} \dots)$, and the order of the chain depict the number of previous states, $\boldsymbol{z}^{(j-1)}, \ldots, \boldsymbol{z}^{(1)}$, the current state, $\boldsymbol{z}^{(j)}$, is dependent on. The target distribution is in this setting unknown, and using (2.1), it is generally known up to the normalizing constant such that the likelihood is known, and we can assert some prior on the parameters.

To construct the Markov chain, a simpler distribution $q(\boldsymbol{z}^{(j+1)} \mid \boldsymbol{z}^{(j)})$ is employed, which is simpler in that it is explicitly available; i.e., we can sample from it and obtain probabilities by evaluation. Henceforth, this simpler distribution will be referred to as the proposal distribution. New candidate states, $\boldsymbol{z}^*$, are drawn from the proposal distribution conditioned on the current state, $\boldsymbol{z}^{(j)}$. The candidate state is then accepted or rejected according to the acceptance probability $\alpha(\boldsymbol{z}^*, \boldsymbol{z}^{(j)})$. This probability is derived using the assumed properties of the Markov chain and the detailed balanced condition $\pi(\boldsymbol{z}^*|\boldsymbol{y})p(\boldsymbol{z}^{(j)}|\boldsymbol{z}^*) = \pi(\boldsymbol{z}^{(j)}\boldsymbol{y})p(\boldsymbol{z}^*|\boldsymbol{z}^{(j)})$, where the transition probability is given by $p(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^*) = q(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^*)\alpha(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^*)$; which can be interpreted as: the probability of going from state $\boldsymbol{z}^*$ to state $\boldsymbol{z}^{(j)}$ and oppositely going from state $\boldsymbol{z}^{(j)}$ to $\boldsymbol{z}^*$ is equivalent. Thereby, the acceptance probability can be formulated as

$$\alpha(\boldsymbol{z}^* \mid \boldsymbol{z}^{(j)}) = \min\left\{1, \frac{\pi(\boldsymbol{z}^* \mid \boldsymbol{y})q(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^*)}{\pi(\boldsymbol{z}^{(j)} \mid \boldsymbol{y})q(\boldsymbol{z}^* \mid \boldsymbol{z}^{(j)})}\right\}. \tag{2.9}$$

If the candidate is accepted, then the candidate state is set as the new state, $\boldsymbol{z}^{(j+1)} = \boldsymbol{z}^*$, and oppositely if the candidate is rejected the current state is set as the new state, $\boldsymbol{z}^{(j+1)} = \boldsymbol{z}^{(j)}$.

The posterior distribution $\pi(\boldsymbol{z} \mid \boldsymbol{y})$ in (2.5) can be expressed using Bayes' rule (2.1), where the unknown normalizing constant $\pi(\boldsymbol{y})$ conveniently cancels out as it occurs in both the numerator and denominator. The resulting acceptance rate and the general representation of the acceptance rate in the Metropolis-Hastings algorithm is given by

$$\alpha(\boldsymbol{z}^* \mid \boldsymbol{z}^{(j)}) = \min\left\{1, \frac{\pi(\boldsymbol{y} \mid \boldsymbol{z}^*)\pi(\boldsymbol{z}^*)q(\boldsymbol{z}^{(j)} \mid \boldsymbol{z}^*)}{\pi(\boldsymbol{y} \mid \boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)})q(\boldsymbol{z}^* \mid \boldsymbol{z}^{(j)})}\right\}, \tag{2.10}$$

Here, $\pi(\boldsymbol{z})$ is the prior of $\boldsymbol{z}$ and $\pi(\boldsymbol{y} \mid \boldsymbol{z})$ the known likelihood.

In MCMC, starting at an initial state $\boldsymbol{z}^{(0)}$, the development of the Markov chain is a sequential process of accepting/rejecting candidate states until the convergence of the chain is promised under mild conditions. Once the stationary state has been reached, one can consider the MCMC samples as correlated samples from the target distribution. The initial part of this chain, before reaching this stationary state, is indicated as burn-in and removed prior to inference. The issue of convergence is an essential topic when constructing the MCMC algorithm for a particular problem, and the investigation of this property is paramount for the effectiveness of the algorithm. Usually, a diagnostic of convergence can be determined by looking at the trace of state values, and if an equilibrium of these values is reached. Another important measure to determine the quality of the Markov chain is to inspect the acceptance rate. Too low acceptance rate is an indication that the Markov chain could get stuck in some local maxima, and convergence is slow; oppositely, too high acceptance rate might indicate that the sampler moves very slowly and, therefore, takes a long time to explore the parameter space fully. An approach in improving convergence is to adjust the proposal distribution $q(\cdot)$; for example, by altering its variance and rerun the simulation. Less variance allows for smaller jumps in the parameters space and, oppositely, a larger variance allows for bigger jumps.

Let us now assume that the Markov chain has been constructed according to the Metropolis-Hastings algorithm, and that the chain has converged to our target distribution. Then, estimating quantities of interest, (2.7), can be obtained empirically as

$$\mathbb{E}_{\pi}[f(\boldsymbol{z})\,|\,\boldsymbol{y}] = \int_{\mathcal{Z}} f(\boldsymbol{z})\pi(\boldsymbol{z}\,|\,\boldsymbol{y})\mathrm{d}\boldsymbol{z} \approx \frac{1}{M}\sum_{j=1}^{M} f(\boldsymbol{z}^{(j)}), \tag{2.11}$$

where $\mathcal{Z}$ is the state space of $\boldsymbol{z}$. With the introduction of (2.11), another useful diagnostic is the *effective sample size* of the Markov chain. As the samples produced by the MCMC method typically will be autocorrelated, the variance of the estimator in (2.11) is increased. Thereby, given the dependent states of the Markov chain, the effective sample size is the number of independent states with the same estimator variance as produced by the autocorrelated states. The effective sample size generated by a MCMC simulation is defined as

$$\widehat{\mathrm{ESS}} = \frac{N}{1 + 2\sum_{t=1}^{\infty}\rho_t},$$

where ρ_t is the autocorrelation function at lag t. In a practical setting, the upper bound of the sum is a finite number $t = T$, where the autocorrelation is close to zero.

To achieve numerical results of tasks (2.6) and (2.8), the samples are generally placed into bins according to their sample values, where each bin is weighted according to the number of samples within. Thereby, the kernel of the target distribution can easily be approximated with these points, yielding a solution to (2.6). The optimization of the target distribution is simply solved by picking the bin containing the highest number of samples. It is important to note that these are approximations and will, in a practical manner, carry some numerical errors. However, if the number of samples grows to infinity, and the bin-size tends towards zero, the approximations intuitively becomes exact.

2.2.2 Importance Sampling

Importance sampling (IS) may be considered a precursor to MCMC methods, and was first introduced by Kahn (1950) to estimate the probability of nuclear particles penetrating shields. The method is based on the identity

$$\int f(\boldsymbol{z})\pi(\boldsymbol{z}\,|\,\boldsymbol{y})\mathrm{d}\boldsymbol{z} = \int \frac{f(\boldsymbol{z})\pi(\boldsymbol{z}\,|\,\boldsymbol{y})}{q(\boldsymbol{z})}q(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \tag{2.12}$$

and is commonly used to compute (2.7) in situations where the domain $f(\boldsymbol{z})$ lies in the area of low probability of the target distribution $\pi(\boldsymbol{z} \mid \boldsymbol{y})$. In this setting, the classical Monte Carlo approach, which generates samples from $\pi(\boldsymbol{z} \mid \boldsymbol{y})$, would obtain poor approximations of (2.7) because most of its samples would be in a region where $f(\boldsymbol{z}) = 0$. It is apparent that, similar to MCMC, a simpler distribution (or proposal distribution) $q(\boldsymbol{z})$ must be employed to generate samples that eclipse the important region, the region where $f(\boldsymbol{z})\pi(\boldsymbol{z} \mid \boldsymbol{y}) \neq 0$. Then, by taking advantage of the identity in (2.12), the estimate of (2.7) is adjusted to account for the use of this proposal distribution. The IS method is commonly used in high energy physics, rare event simulation, and rendering in computer graphics. Moreover, it can also substitute the accept-reject design in MCMC, and be used for Bayesian inference.

In Bayesian inference, the interest lies in approximating the target distribution or a particular moment about it, such that the important region must overweigh the sample space of the target distribution instead of the domain of $f(\boldsymbol{z})$. The question then arises; why invoke this proposal distribution to generate samples as the classical Monte Carlo method could be used on this problem? However, similar to the setting in MCMC, our target distribution is unknown, such that samples are unobtainable from the target distribution itself.

Consider the M generated samples $\{\boldsymbol{z}\}_{j=1}^{M}$ from the proposal distribution; an unbiased and consistent estimator (Bugallo et al., 2017) of the expected value of a function $f(\boldsymbol{z})$ with respect to the target distribution $\pi(\boldsymbol{z} \mid \boldsymbol{y})$, can be expressed as

$$\begin{aligned} \mathbb{E}_{\pi}[f(\boldsymbol{z})] &= \int_{\mathcal{Z}} f(\boldsymbol{z}) \frac{\pi(\boldsymbol{z} \mid \boldsymbol{y})}{q(\boldsymbol{z})} q(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\ &\simeq \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)} \mid \boldsymbol{y})}{q(\boldsymbol{z}^{(j)})}. \end{aligned} \tag{2.13}$$

Here, $q(\boldsymbol{z})$ is a multivariate proposal distribution, where $q(\boldsymbol{z}) > 0$ whenever $f(\boldsymbol{z})\pi(\boldsymbol{z} \mid \boldsymbol{y}) \neq 0$, such that the tail of the proposal is heavier than the target. The choice of proposal is important, as the variance of the estimator in (2.13) directly depends on the dissimilarity between the shape of the proposal and the target distribution (Robert et al., 2004). This dissimilarity, representing the significance of one sample in approximating the target, is generally referred to as the importance weight.

$$\omega^{(j)} = \frac{\pi(\boldsymbol{z}^{(j)} \mid \boldsymbol{y})}{q(\boldsymbol{z}^{(j)})}. \tag{2.14}$$

To compute the importance weight in (2.14) the normalizing constant of $\pi(\boldsymbol{z}|\boldsymbol{y})$ needs to be obtainable. This is not the case in many practical situations and, in general, is not the case for LGMs. An alternative is then to employ the *self-normalized* importance weights:

$$
\begin{aligned}
\bar{\omega}^{(j)} &= \frac{\pi(\boldsymbol{y} \mid \boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)})}{q(\boldsymbol{z}^{(j)})} \Bigg/ \sum_{j=1}^{M} \frac{\pi(\boldsymbol{y} \mid \boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)})}{\pi(q(\boldsymbol{z}^{(j)})} \\
&= \omega^{(j)} \Bigg/ \sum_{j=1}^{M} \omega^{(j)} ,
\end{aligned}
\tag{2.15}
$$

where the unknown normalizing constant conveniently cancels out. The resulting estimator for the quantity of interest is

$$
\tilde{\mathbb{E}}_{\pi}[f(\boldsymbol{z})] = \sum_{j=1}^{M} \bar{\omega}^{(j)} f(\boldsymbol{z}^{(j)}). \tag{2.16}
$$

It can be shown that (2.16) is biased for finite M but consistent (Geweke, 1989). Considering the importance weight $\bar{\omega}^{(j)}$ represents the target distribution evaluated at $\boldsymbol{z}^{(j)}$, i.e. $\pi(\boldsymbol{z}^{(j)}; \boldsymbol{z} \mid \boldsymbol{y}) \simeq \bar{\omega}^{(j)}$, and assuming that the number of samples $M \to \infty$, the approximation of the target distribution, and the solution of (2.6), can be found with the expression

$$
\tilde{\pi}(\boldsymbol{z} \mid \boldsymbol{y}) = \sum_{j=1}^{M} \bar{\omega}^{(j)} \delta(\boldsymbol{z} - \boldsymbol{z}^{(j)}),
$$

where $\delta(\cdot)$ is the Dirac delta function. In a practical manner, where $M \to \infty$ is infeasible, the target distribution can approximated with non-parametric kernel density estimation (see Silverman, 1986), using the self-normalized weights and their corresponding samples. Furthermore, the mode can be found with the argument of the maximum value of the now approximated kernel density. The mode can also be found by viewing the corresponding sample value of the maximum weight, but the accuracy of this method is highly dependent on the number of samples within the probability mass of the target distribution.

In extreme settings, some weights might be significantly larger than others, achieving a limited number relevant samples; similarly, all weights might be zero, ruling all samples insignificant; in other settings, the conclusion about the quality of the samples might be difficult to draw. In this latter case, a common diagnostic is the *effective sample size*:

$$
\begin{aligned}
\widehat{\text{ESS}} &= \frac{1}{\sum_{j=1}^{M} \bar{\omega}^{(j)2}} \\
&= \frac{\left(\sum_{j=1}^{M} \omega^{(j)}\right)^2}{\sum_{j=1}^{M} \omega^{(j)2}},
\end{aligned}
\tag{2.17}
$$

which is the number of independent samples generated from the target distribution required to obtain the same estimator variance of (2.7) as the self-normalized estimator (2.16) using the M generated samples from the proposal distribution. The theoretical development of (2.17) will not be detailed here, and the reader is referred to Martino et al. (2017) for a thorough account.

Compared to MCMC, IS has the advantage to be easily parallelized since the samples are drawn independently from each other. On the other hand, the performance of the basic IS algorithm highly depends on the choice of proposal distribution, which remains constant during the whole simulation.

2.2.3 Adaptive Multiple Importance Sampling

In the IS method, the validity of the method is promised under mild conditions; however, the variance of the estimator (2.16) is dependent on the dissimilarity between the shape of the target distribution and the proposal (Elvira et al., 2019; Robert et al., 2004). Following the development of more robust IS schemes, we will, in this section, present the *adaptive multiple* importance sampling methodology (AMIS) proposed by Corneut et al. (2012). The AMIS method combines two modern concepts in IS; *multiple* importance sampling (MIS), employing a mixture of distributions as proposal distribution (Owen et al., 2000); *adaptive* importance sampling (AIS), adapting the proposal distribution to better approximate the target (Cappé et al., 2004). The AMIS merge these concepts by constructing a mixture distribution through the adaptation of the proposal. We will first describe the MIS method and the estimators associated with it. Then, we sequentially develop the AMIS algorithm and, lastly, outline its convergence properties.

The main idea behind the MIS is to use a series of T proposal densities $\{q_t(\cdot)\}_{t=1}^{T}$ combined in a mixture as

$$\psi(\cdot) = \sum_{t=1}^{T} \rho^{(t)} q_t(\cdot), \tag{2.18}$$

where $\rho^{(t)}$ are the mixture weights, such that $\sum_{t=1}^{T} \rho^{(t)} = 1$ to ensure that $\psi(\cdot)$ is probability density. The result is that the performance of MIS depends on a series of proposals, instead of entrusting the approximation of the target distribution to the dissimilarity with one single proposal distribution (IS). Assume that N_t samples $\{\boldsymbol{z}^{(t,j)}\}_{j=1}^{N_t}$ are generated from the corresponding proposal distribution $q_t(\cdot)$, such that the total number of samples is $\sum_{t=1}^{T} N_t = N$. The mixture weight $\rho^{(t)}$ is determined by the fraction of samples drawn from the tth proposal distribution, i.e. $\rho^{(t)} = N_t/N$, and the resulting mixture distribution can be expressed as

$$\psi(\cdot) = \frac{1}{N} \sum_{t=1}^{T} N_t q_t(\cdot). \tag{2.19}$$

Similar to the importance weights calculated in (2.14), the importance weights in the MIS scheme is calculated by

$$\omega^{(t,j)} = \frac{\pi(\boldsymbol{z}^{(t,j)} \mid \boldsymbol{y})}{\psi(\boldsymbol{z}^{(t,j)})}, \tag{2.20}$$

where the single proposal distribution in the denominator is replaced with the mixture of many proposal distributions. The estimator for a quantity of interest in MIS is computed with (2.13), where the single proposal distribution is replaced with the mixture of proposal distribution:

$$\begin{aligned} \widehat{\mathbb{E}}_\pi[f(\boldsymbol{z})\mid \boldsymbol{y}] &= \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{N_t}\frac{f(\boldsymbol{z}^{(t,j)})\pi(\boldsymbol{z}^{(t,j)}\mid \boldsymbol{y})}{\psi(\boldsymbol{z}^{(t,j)})} \\ &= \frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{N_t}\omega^{(t,j)}f(\boldsymbol{z}^{(t,j)}), \end{aligned} \tag{2.21}$$

which is an unbiased and consistent estimator (Elvira et al., 2019). When the normalizing constant of the target distribution $\pi(\boldsymbol{z}\mid\boldsymbol{y})$ is unknown, we can rely on the self-normalizing importance weights:

$$\begin{aligned} \bar{\omega}^{(k,l)} &= \frac{\pi(\boldsymbol{y}\mid \boldsymbol{z}^{(k,l)})\pi(\boldsymbol{z}^{(k,l)})}{\psi(\boldsymbol{z}^{(k,l)})}\Bigg/\sum_{t=1}^{T}\sum_{j=1}^{N_t}\frac{\pi(\boldsymbol{y}\mid \boldsymbol{z}^{(t,j)})\pi(\boldsymbol{z}^{(t,j)})}{\psi(\boldsymbol{z}^{(t,j)})} \\ &= \omega^{(k,l)}\Bigg/\sum_{t=1}^{T}\sum_{j=1}^{N_t}\omega^{(t,j)}\ . \end{aligned} \tag{2.22}$$

Here, $k \in (1, T)$ denotes the proposal distribution the sample $\boldsymbol{z}^{(k,l)}$ is drawn from, and $l \in (1, N_k)$ the lth sample from the kth proposal distribution. The self-normalized estimator of a quantity of interest, using weights calculated by (2.22), is expressed in (2.16), which is biased for finite $\sum_{t=1}^{T} N_t$ but consistent as shown in Elvira et al. (2019).

In both IS and MIS, the proposal distribution(s) are static throughout the whole sampling process; thereby, bad proposals will lead to a low-quality inference. A strategy to improve the algorithm would be to let the algorithm "learn" a better proposal during the sampling process. This is exactly the idea behind the AIS methods, where the proposal is adapted sequentially, gradually increasing the accuracy in approximating the target distribution. Consider the proposal distribution belonging to a parametric family of distributions $\{q(\cdot;\boldsymbol{\phi})\mid\boldsymbol{\phi}\in\boldsymbol{\Phi}\}$, where $\boldsymbol{\Phi}$ is the parameter space. The initial proposal distribution is assigned the parameters $\boldsymbol{\phi}_1\in\boldsymbol{\Phi}$. From this proposal, N_1 samples are generated and weighted according to (2.15). The proposal distribution is then adapted by updating the parameters $\boldsymbol{\phi}_1$ following some criterion. This updating procedure is repeated T times, obtaining the sequence of parameters $\boldsymbol{\phi}_1 \to \boldsymbol{\phi}_2 \to \cdots \to \boldsymbol{\phi}_T$ for the proposal distribution, where the use of the last parameters $\boldsymbol{\phi}_T$, should best explain the probability mass of the target distribution.

A common approach in adapting the proposals is the *moment matching criterion* (Corneut et al., 2012), using (2.16) to estimate first and second moments (mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$) of the target distribution, and assign them to the parameters of the new proposal distribution $q(\cdot;\boldsymbol{\phi}_2) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. An alternative criterion, is the minimization of the *Kullback-Leibler* divergence between the proposal and target distribution (Cappé et al., 2008; Corneut et al., 2012).

The AMIS combines both the adaptive and multiple proposals ideas, and in the following, we will describe the sequential development of the mixture in the AMIS methodology. Similar to the AIS method, the algorithm starts with a single proposal distribution $q_1(\cdot; \boldsymbol{\phi}_1)$, where N_1 samples are drawn and weighted according to (2.15). Then, $\boldsymbol{\phi}_1$ is updated according to one of the aforementioned criterion's, and N_2 new samples are generated from the new proposal distribution $q_2(\cdot, \boldsymbol{\phi}_2)$. Here, the standard AIS and AMIS method diverge. Borrowing strength from MIS weighting scheme, the AMIS method weights the new samples according to (2.22), where the mixture distribution in (2.19) comprise of the previous proposal $q_1(\cdot)$ and the new $q_2(\cdot)$. Furthermore, to gain this strength in all samples, the past N_1 samples are re-weighted using (2.22) and this mixture.

In general, the AMIS method follows this sequential process for a predetermined number of epochs T; where a epoch refers to one cycle of generating samples, expanding the mixture with the new proposal distribution, calculating and updating weights, and adapting the proposal. Thereby, similar to MIS, the mixture in the AMIS weighting scheme will ultimately consists of the T proposal distributions, where posterior estimates is calculated with (2.16). The number of epochs T and the generated samples in each epoch, $N_1, N_2, \ldots, N_T$, is referred to as our sampling strategy, and the values are related to the dimensions d of the target distribution $\pi(\boldsymbol{z} \mid \boldsymbol{y})$. In Corneut et al. (2012), it is recommended that $N_t >= 25$ when d is small (around $d = 2$), and $N_t >= 500$ when d is large (> 20). Corneut et al. (2012) also note that an increasing sample size after each adaptation, i.e. $N_1 < N_2 < \cdots < N_T$, is favorable because this increases the importance of later proposals in the mixture (2.19), and more samples are drawn from essentially better proposal distributions.

The performance of the AMIS algorithm is still highly dependent on the choice of initial proposal distribution $q_1(\cdot)$, as the algorithm only sees the sample space of the proposal distribution. Therefore, if the probability mass of the target distribution is outside the sample space of the proposal, the method would require many adaptations and long computing times to move the proposal distributions accordingly. The convergence and unbiasedness of (2.21) as established by Owen et al. (2000) and Corneut et al. (2012), implies that the convergence of the AMIS algorithm is promised. However, with the introduction of adaptive proposal distributions, the importance weights of new samples are dependent on prior samples, impeding the unbiasedness property. Also, even the convergence is challenged as it would require the compactness restriction on the sample space (Corneut et al., 2012). One could show unbiasedness of (2.21) in the AMIS setting, by letting $N_1, \ldots, N_{T-1}$ and T be finite when $N_T \to \infty$, making (2.21) only dependent on the last proposal distribution $q_t(\cdot)$; thus, removing the dependency in the samples and bias in the weights (Corneut et al., 2012). However, this is infeasible in practical situations, and not a recommended application of AMIS according to Corneut et al. (2012).

Similar to IS, the quality of the samples generated by the AMIS algorithm can be evaluated by the estimated effective sample size, which is calculated by (2.17) using the self-normalized mixture weights from (2.22). The estimate refers to the number of independent samples drawn from the target distribution required to obtain the same estimator variance of (2.7) as the (2.16) using all AMIS samples.

In the implementation of the algorithm, the updating of past samples is done in an inexpensive way to avoid multiple evaluations of the prior and previous proposal distribution for one sample; thereby, only the new proposal distribution is evaluated when updating past weights. The updating scheme, introducing the $\delta^{(t,j)}$ parameter, and the individual steps of the generic AMIS algorithm as proposed by Corneut et al. (2012) is presented in Algorithm 1.

Algorithm 1: Generic AMIS as proposed by Corneut et al. (2012)

```
- Initialize $\boldsymbol{N}_t = (N_1, \dots, N_T)$, $q_1(\cdot; \boldsymbol{\phi}_1)$
for $j$ from 1 to $N_1$ do
    - Generate sample $\boldsymbol{z}^{(1,j)} \sim q_1(\cdot; \boldsymbol{\phi}_1)$
    - Compute:
        $\delta^{(1,j)} = N_1 q_1(\boldsymbol{z}^{(1,j)}; \boldsymbol{\phi}_1)$ and $\omega^{(1,j)} = \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(1,j)})\pi(\boldsymbol{z}^{(1,j)})}{q_1(\boldsymbol{z}^{(1,j)}; \boldsymbol{\phi}_1)}$
- Calculate $\boldsymbol{\phi}_2$ using the weighted set of samples:
        $(\{\boldsymbol{z}^{(1,1)}, \omega^{(1,1)}\}, \dots, \{\boldsymbol{z}^{(1,N_1)}, \omega^{(1,N_1)}\})$

for $t$ from 2 to $T$ do
    for $j$ from 1 to $N_t$ do
        - Generate sample $\boldsymbol{z}^{(t,j)} \sim q_t(\cdot; \boldsymbol{\phi}_t)$
        - Compute:
            $\delta^{(t,j)} = \sum_{l=1}^{t} N_l q_t(\boldsymbol{z}^{(t,j)}; \boldsymbol{\phi}_t)$ and $\omega^{(t,j)} = \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(t,j)})\pi(\boldsymbol{z}^{(t,j)})}{\left[\delta^{(t,j)} / \sum_{l=1}^{t} N_l\right]}$
    for $l$ from 1 to $t-1$ do
        for $j$ from 1 to $N_l$ do
            - Update past importance weights:
                $\delta^{(l,j)} \leftarrow \delta^{(l,j)} + N_l q_t(\boldsymbol{z}^{(l,j)}; \boldsymbol{\phi}_t)$ and $\omega^{(l,j)} \leftarrow \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}^{(l,j)})\pi(\boldsymbol{z}^{(l,j)})}{\left[\delta^{(l,j)} / \sum_{k=1}^{t} N_k\right]}$
    - Calculate $\boldsymbol{\phi}_{t+1}$ using the weighted set of samples:
            $(\{\boldsymbol{z}_c^{(1,1)}, \omega^{(1,1)}\}, \dots, \{\boldsymbol{z}_c^{(t,N_t)}, \omega^{(t,N_t)}\})$
```

2.3 Integrated Nested Laplace Approximations

An alternative approach to inference is the *integrated nested Laplace approximation* (INLA) proposed by Rue et al. (2009). INLA differs from MCMC in many ways. First of all, its use is limited to the LGM class described in Section 2.1.1. INLA is a deterministic algorithm that relies upon Laplace approximations to compute integrals of specific densities that are coupled together to obtain approximations to the posterior marginals of $\pi(\boldsymbol{z} \mid \boldsymbol{y})$. For LGMs, when compared to MCMC, INLA is much faster and reliable, not having to deal with convergence issues (Rue et al., 2009). INLA deals in different ways with the elements of $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{\theta}$, so we will in this section keep them separate.

For clarity, we sum up the critical assumptions about the LGMs required by INLA (Rue et al., 2017):

1. Each observation y_i only depends on one components of the latent field $\boldsymbol{x}$, the linear predictor η_i, resulting in the likelihood (2.4).
2. The dimensions of the hyperparameters $\boldsymbol{\theta}$ is small (2-5, not >20)
3. The latent field $\boldsymbol{x} \mid \boldsymbol{\theta}$ is Gaussian, and can be high dimensional but is required to be a Gaussian Markov random field, such that the precision matrix $Q(\boldsymbol{\theta})$ is sparse.
4. The linear predictor is in the form (2.2), i.e. additive with the effects of covariates.

In this section, we will outline the basic ideas behind the INLA, and for a thorough introduction see Martino et al. (2019).

2.3.1 Laplace Approximations

A classic approach to approximations of posterior moments and marginals is the Laplace method (Tierney et al., 1986). Consider a probability density function $\pi(x)$ of the random variable $X \in \mathcal{X} \subseteq \mathbb{R}$, and suppose that we are interested in the integral

$$\int_{\mathcal{X}} \pi(x)\mathrm{d}x = \int_{\mathcal{X}} \exp(ng(x))\mathrm{d}x, \tag{2.23}$$

where n is a samples size or a parameter allowing $n \to \infty$. To find a numerical approximation of (2.23), the second order Taylor series expansions of $g(x)$ about a point $x = x_0$ is computed as

$$g(x) \simeq g(x_0) + (x - x_0)g'(x_0) + \frac{(x - x_0)^2}{2} g''(x_0) + R(x), \tag{2.24}$$

where the remainder is $R(x) = \mathcal{O}\left((x - x_0)^3\right)$. Choosing x_0 to be the global maximum of $g(x)$, which is a stationary point if it is not a endpoint of $\mathcal{X}$, such that $g'(x_0) = 0$ removing the linear term in (2.24). By substituting the Taylor expansion (2.24) with $g(x)$ in (2.23), we have the approximation

$$\int_{\mathcal{X}} \pi(x)\mathrm{d}x \simeq \exp(ng(x_0)) \int_{\mathcal{X}} \exp\left(\frac{n(x - x_0)^2}{2} g''(x_0)\right)\mathrm{d}x. \tag{2.25}$$

Note that (2.25) is only valid in the neighborhood of x_0. We observe that the integrand in (2.25) is the core of the Gaussian probability density function, denoted $\phi(\cdot)$, with mean x_0 and variance $\sigma_0^2 = -(ng''(x_0))^{-1}$. Then, by taking the integral over the interval $[\alpha, \beta] \subseteq \mathcal{X}$, the approximation in (2.25) can be expressed as

$$\begin{aligned}\int_\alpha^\beta \pi(x)\mathrm{d}x &\simeq \pi(x_0)\sqrt{2\pi\sigma_0^2}\int_\alpha^\beta \phi(x; x_0, \sigma_0^2)\mathrm{d}x \\ &= \pi(x_0)\sqrt{2\pi\sigma_0^2}\left(\Phi(\beta; x_0, \sigma_0^2) - \Phi(\alpha; x_0, \sigma_0^2)\right),\end{aligned} \tag{2.26}$$

where $\Phi(\cdot)$ denotes the Gaussian cumulative density function.

2.3.2 Approximate inference with INLA

The INLA approach does not attempt to estimate the joint posterior distribution in (2.5), but rather the posterior marginals of the components in the latent field and hyperparameters, expressed as

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta} \tag{2.27}$$

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}_{-j}. \tag{2.28}$$

The main idea in the INLA methodology is to build an approximation for $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$, and solve the integrals in (2.27) and (2.28) numerically by

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_j \tilde{\pi}(x_i \mid \theta_j, \boldsymbol{y})\tilde{\pi}(\theta_j \mid \boldsymbol{y})\boldsymbol{\Delta}_j \tag{2.29}$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \sum_k \tilde{\pi}(\theta_j, \boldsymbol{\theta}_{-j}^{(k)} \mid \boldsymbol{y})\boldsymbol{\Delta}_k, \tag{2.30}$$

where $\boldsymbol{\Delta}_j$ and $\boldsymbol{\Delta}_k$ are appropriate weights. We now need to obtain approximations to $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$, and find representative values of $\boldsymbol{\theta}$ to solve (2.29) and (2.30).

First, we consider the approximation of the joint posterior of the hyperparameters as

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \\ &\simeq \left.\frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}_0(\boldsymbol{\theta})} := \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}),\end{aligned} \tag{2.31}$$

where we substitute the denominator $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$, which is hard to compute explicitly, with its Gaussian approximation $\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$. This approximation is built by matching the mode and the curvature at the mode as

$$\begin{aligned}\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\mathbf{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i\in\mathcal{I}} \ln \pi(y_i \mid x_i, \boldsymbol{\theta})\right) \\ &\simeq |\mathbf{P}(\boldsymbol{\theta})|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0(\boldsymbol{\theta}))^T\mathbf{P}(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{x}_0(\boldsymbol{\theta}))\right).\end{aligned} \tag{2.32}$$

Here, $\mathbf{P}(\boldsymbol{\theta}) = \mathbf{Q}(\boldsymbol{\theta})+\text{diag}(\mathbf{C}(\boldsymbol{\theta}))$ is the precision and $\boldsymbol{x}_0(\boldsymbol{\theta})$ is the mode of $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ for a given $\boldsymbol{\theta}$. The matrix $\mathbf{C}(\boldsymbol{\theta})$ is the negative second derivative of the log-likelihood evaluated at the mode, which is the inverse of σ_0^2 from (2.26) in a multivariate setting. The Gaussian approximation is accurate on $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ since it is a-priori distributed as a GMRF, and $\boldsymbol{y}$ only shifts the mean, reduces the variance and presents some skewness.

The main use of $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$ is to estimate the posterior marginals $\pi(\theta_j \mid \boldsymbol{y})$ and $\pi(x_i|\boldsymbol{y})$ by integrating out the uncertainty about $\boldsymbol{\theta}$ according to (2.29) and (2.30). This is achieved by using the approximation $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$ to locate the area of high density and, thereby, choose some representative points in the space of $\boldsymbol{\theta}$. The posterior mode of $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ is obtained using a quasi-Newton method to maximize $\ln \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$ with respect to $\boldsymbol{\theta}$. Furthermore, the negative Hessian matrix $\mathbf{H} > 0$ is computed using finite differences at the mode $\boldsymbol{\theta}^*$. Next, a reparametrization of $\boldsymbol{\theta}$ is performed to correct for scale and rotation, and simplify the numerical integration:

$$\boldsymbol{\theta} = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z},$$

where the inverse Hessian is decomposed using the eigenvalue decomposition $\mathbf{H}^{-1} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$.

Two methods can be used to locate the representative points: the first one, constructs a grid of step size h around the mode, where points, $\boldsymbol{z}^*$, are kept only if

$$\left|\ln \tilde{\pi}(\boldsymbol{\theta}(\mathbf{0}) \mid \boldsymbol{y}) - \ln \tilde{\pi}(\boldsymbol{\theta}(\boldsymbol{z}^*) \mid \boldsymbol{y})\right| < \delta,$$

and δ is a given threshold. The second alternative, is to create a central composite design (CCD, see George E. P. Box, 1987) around $\boldsymbol{\theta}(\mathbf{0})$. This method strategically chooses relevant points, given the mode $\boldsymbol{\theta}^*$ and the Hessian $\mathbf{H}$, to perform a second-order approximation to a response variable. The CCD method is generally used when the dimensions of the hyperparameters is high, because it utilizes much less points than the grid exploration but still manages to capture the variability of the hyperparameters (Rue et al., 2009, Section 6.5). Following this grid exploration the posterior marginals $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$ is found using an interpolation algorithm on the weighted points $\{\boldsymbol{\theta}(\boldsymbol{z}_j), \pi(\boldsymbol{\theta}(\boldsymbol{z}_j) \mid \boldsymbol{y})\}$ (Martins et al., 2013b).

To obtain an approximation to the posterior marginals of the latent components, the hyperparameters need to be integrated out according to (2.29) using the weighted set of $\boldsymbol{\theta}$. Here, $\tilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ is an approximation to $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$. Rue et al. (2009) propose three different methods for this approximation. The first easy possibility is to use the marginal of the Gaussian approximation $\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ described in (2.32). Thereby, the marginals can be computed, where the Cholesky decomposition is used for the precision matrix (Rue et al., 2007). Despite being computationally very fast, it generally isn't very precise. The second option is to split the latent components, $\boldsymbol{x} = (x_i, \boldsymbol{x}_{-i})$, and use Bayes' rule as

$$\begin{aligned} \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\ &\approx \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\tilde{\pi}(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i} = \boldsymbol{x}_{-i,0}(x_i, \boldsymbol{\theta})} := \tilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}), \end{aligned} \tag{2.33}$$

where $\pi(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})$ is approximated using the Laplace method describe in Section 2.3.1. This approach is computationally very expensive since $\pi(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})$ needs to be re-estimated for each value $\boldsymbol{\theta}$ and x_i, but the approximations are typically very accurate by the denominator being fairly Gaussian (Rue et al., 2009). The third and last method is termed *simplified Laplace approximation* and it relies on a Taylor expansion around the mode of the Laplace method. This adds a linear and cubic term to the Gaussian approximation as

$$\ln \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx -\frac{1}{2} x_i^2 + b_i(\boldsymbol{\theta}) x_i + \frac{1}{6} c_i(\boldsymbol{\theta}) x_i^3. \tag{2.34}$$

Moreover, skew-Gaussian distribution is assigned to (2.34), such that (2.29) is approximated with a mixture of skew-Gaussian distributions, where linear term provides correction to the mean and the cubic term provide corrections to the skewness. This method increases the computational speed, but with a slight loss in accuracy from the second method in (2.33). All three approaches are described in Section 3.2 in Rue et al. (2009), and in Section 3.2 in Rue et al. (2017).

As a byproduct of the previous computations, the INLA methodology allows for approximations of other quantities useful for example in model comparison. The marginal likelihood is one of such quantities and can be derived as

$$\tilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x} = \boldsymbol{x}^*(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}, \tag{2.35}$$

where $\boldsymbol{x}_0(\boldsymbol{\theta})$ is the mode of $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ for a given $\boldsymbol{\theta}$, and $\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation described in (2.32). The approximation in (2.35) is derived from (2.30) as the normalizing constant of $\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})$. This method of approximating the marginal likelihood using INLA has proven to be quite accurate when compare with the computational speed of other approaches (see Hubin et al., 2016), and will be very useful in the development of the algorithms in Chapter 3.

As previously mentioned, INLA provides fast and accurate inference on posterior marginals for the additive LGMs described in Section 2.1.1. The efficiency of the INLA procedures relies on a careful implementation of the different algorithms within. A such implementation is available in the R package **R-INLA**, which allows us to fit complex models in a matter of seconds.

Implementing INLA from scratch is a daunting task so, in practice, the applications of INLA are limited to the (large) class of models implemented in **R-INLA**. Although **R-INLA** offer the possibility for some user-defined models shown in Gómez-Rubio (2020), there are models that do not fit the scope of **R-INLA**.

In the next chapter, we will present an approach that allows us to extend the class of models that can benefit from the fast inference of INLA by coupling INLA with some of the Monte Carlo algorithms presented earlier in this chapter.

Chapter 3

INLA within Monte Carlo Methods

As described in Section 2.3, INLA obtains posterior inference on the LGMs detailed in Section 2.1.1, and is restricted to the class of models implemented in **R-INLA**. However, many models are excluded from this list, and it is difficult to add new models to the framework in R. Multiple methods to extend the set of models that can be fit with INLA through the **R-INLA** package have been proposed. They are similar in that they fix some of the unknown parameters to suitable values, and fit the models conditioned on these parameters with **R-INLA**. The parameters are conveniently chosen, such that the conditional models are LGMs, and we will refer to these models as *conditional* LGMs. That is, we assume that while the model $\boldsymbol{z} \mid \boldsymbol{y}$ cannot be approximated with **R-INLA**, the conditional model $\boldsymbol{z}_{-c} \mid \boldsymbol{y}, \boldsymbol{z}_c$ can. Here, we indicate $\boldsymbol{z} = (\boldsymbol{z}_{-c}, \boldsymbol{z}_c)$ as the vector of all unknown parameters of the model $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{z}_c$ is the subset of parameters we condition on, and $\boldsymbol{z}_{-c}$ is its complement.

Assume that we have a way to sample a series of values $\boldsymbol{z}_c^{(j)}$ for $j = 1, \ldots, N$, and that the model $\boldsymbol{z}_{-c} \mid \boldsymbol{y}, \boldsymbol{z_c} = \boldsymbol{z}_c^{(j)}$ can be fitted with INLA. We can then obtain approximate conditional posterior marginal $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c = \boldsymbol{z}_c^{(j)})$ for all $z_{-c,i} \in \boldsymbol{z}_{-c}$. In addition, we can recover the approximate conditional marginal likelihood $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c = \boldsymbol{z}_c^{(j)})$. Assuming that multiple values of $\boldsymbol{z}_c$ are chosen, and that the conditional models are fit with INLA obtaining the conditional posterior marginals and conditional posterior likelihoods, such that an approximation of the posterior distribution $\pi(\boldsymbol{z}_c \mid \boldsymbol{y})$ is found; then, the posterior marginals of the elements of $\boldsymbol{z}_{-c}$ can be approximated using (2.7) as

$$\begin{aligned}\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}) &= \mathbb{E}_{\pi}[\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c)] \\ &= \int \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c)\pi(\boldsymbol{z}_c \mid \boldsymbol{y})\mathrm{d}\boldsymbol{z}_c.\end{aligned} \tag{3.1}$$

This approximation is generally referred to as *Bayesian model averaging* (BMA; see Hoeting et al., 1999). Posterior quantities of interest, e.g. posterior moments, about the estimated posterior marginals $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y})$ in (3.1) is estimated with

$$\mathbb{E}[f(z_{-c,i}) \mid \boldsymbol{y}] \simeq \int f(z)\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y})\mathrm{d}z, \tag{3.2}$$

where $f(\cdot)$ is any integrated function on the domain of the approximated posterior marginal. In practice, the integral is solved by numerical integration; for example, using Simpson's rule. The approximation of (3.2) is also available in the **R-INLA** package, where the function `inla.emarginal` can be used for a self-defined function $f(\cdot)$, and `inla.zmarginal` for posterior statistics.

In this chapter, we will describe three algorithms that combine Monte Carlo techniques and INLA; MCMC with INLA proposed by Gómez-Rubio et al. (2018), IS with INLA introduced by Gómez-Rubio (2019), and a new methodology combining AMIS with INLA. In common, these methods employ a Monte Carlo method to generate samples to obtain the posterior distribution of $\boldsymbol{z}_c$. Moreover, they use INLA to approximate conditional posterior marginals of $\boldsymbol{z}_{-c}$, which are combined with (3.1) using the result of the Monte Carlo simulations. These combined approaches are entirely made possible by the approximation of the conditional marginal likelihood (2.35) provided by INLA, $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c)$, which actuates the accept-reject design in MCMC and weighing scheme in IS.

The situations described below is an overview of models that can be represented as conditional LGMs. These models can be applied to INLA within Monte Carlo methods and are either covered by other authors, or in our experiments. The INLA methodology requires all elements in (2.2) to be additive and Gaussian, such that the linear predictor η also is Gaussian. This is, for example, not the case in a Bayesian lasso model, where the linear coefficients $\boldsymbol{\beta}$ in (2.2) are assigned Laplace priors. The Bayesian lasso, therefore, does not fulfill the requirements in **R-INLA**. Gómez-Rubio et al. (2018, Section 6.1) show how to perform Bayesian lasso combining INLA and MCMC. Another situation where INLA cannot be directly applied is when there exists non-additive terms in the predictor, e.g the autocorrelation parameters in spatial lag models (Gómez-Rubio et al., 2018, Section 6.3), or spatial autoregressive combined models (Gómez-Rubio et al., 2019 or Section 4.2). Moreover, INLA does not handle models where one or more parameters $\boldsymbol{\theta}_1$ in (2.4) are dependent on some of the covariates in the model. These situations might occur in time-series data, where the variance of the likelihood changes over time; or in measurements, where the variance of the measurement increase with distance from the location being measured. Another case is missing values in the covariates. **R-INLA** provides a predictive distribution of missing values in the response but does not handle missing values in its covariates. However, if these covariates are fixed to some imputed values, the model can be approximated (Gómez-Rubio et al., 2018, Section 6.2). Lastly, INLA approximates posterior marginals of the parameters in $\boldsymbol{z}$, and if one is interested in the joint posterior distribution of a subset of these $\boldsymbol{z}_c$; then, joint inference can be provided by another inference method, and the remaining conditional LGM is approximated with INLA (Gómez-Rubio et al., 2019, Section 4.2).

3.1 MCMC with INLA

To extend the number of models that can be fit with INLA, Gómez-Rubio et al. (2018) proposed the combination of MCMC algorithm with the INLA. More specifically, they use the Metropolis-Hastings algorithm detailed in Section 2.2.1. We will adopt the hierarchical model architecture described in Section 2.1 and the ensemble notation, $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\theta}) = (\boldsymbol{z}_c, \boldsymbol{z}_{-c})$. Furthermore, we assume that the conditional latent field $\boldsymbol{z}_{-c} \mid \boldsymbol{z}_c$ is Gaussian distributed for $\boldsymbol{z}_c$ fixed to some value $\boldsymbol{z}_c^{(j)}$. The Metropolis-Hastings algorithm then tries to construct a Markov chain of the parameters $\boldsymbol{z}_c$, with their joint posterior distribution, $\pi(\boldsymbol{z}_c \mid \boldsymbol{y})$, as limiting distribution. The trick is now to fix $\boldsymbol{z}_c$ to the generated states of the Markov chain, and employ the INLA to fit the conditional LGMs.

The MCMC algorithm start from a chosen initial state, $\boldsymbol{z}_c^{(0)}$; then, the conditional model given this initial state is fit with INLA, obtaining the initial posterior marginals $\tilde{\pi}(\boldsymbol{z}_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(0)}))$ and the conditional marginal likelihood $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(0)})$. Consider the proposal distribution $q(\boldsymbol{z}_c^* \mid \boldsymbol{z}_c^{(j)})$, used to draw candidate states $\boldsymbol{z}_c^*$ dependent on the previous state $\boldsymbol{z}_c^{(j)}$. The candidates states are then accepted according to a acceptance probability, which found by rewriting (2.10) as

$$\alpha = \min\left\{1, \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^*)\pi(\boldsymbol{z}_c^*)q(\boldsymbol{z}_c^{(j)} \mid \boldsymbol{z}_c^*)}{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j)})\pi(\boldsymbol{z}_c^{(j)})q(\boldsymbol{z}_c^* \mid \boldsymbol{z}_c^{(j)})}\right\}. \tag{3.3}$$

Note that we use a block update, which means that all the parameters in $\boldsymbol{z}_c$ is either rejected or accepted altogether (see Gilks et al., 1996, Chapter 1.4.1). In (3.3), $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j)})$ and $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^*)$ denotes the conditional marginal likelihoods of the current state, $\boldsymbol{z}^{(j)}$, and the candidate state, $\boldsymbol{z}^*$, obtained by the approximation in (2.35). $\pi(\boldsymbol{z}_c)$ is the known prior distribution, which is set before starting the simulation and hold our prior knowledge about the parameters $\boldsymbol{z}_c$.

As described in Section 2.2.1, if the candidate is accepted, then $\boldsymbol{z}_c^{(j+1)} = \boldsymbol{z}_c^*$, and if the candidate is rejected, then $\boldsymbol{z}_c^{(j+1)} = \boldsymbol{z}_c^*$. Similarly, for the conditional posterior marginals and the conditional marginal likelihoods, if the candidate is accepted, then $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j+1)}) = \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^*)$ and $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j+1)}) = \tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^*)$, or if rejected, then $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j+1)}) = \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j)})$ and $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j+1)}) = \tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j)})$. This accept/reject design continues until convergence is reached and a predetermined number of samples from the target distribution are generated. After the simulation, the diagnostics described in Section 2.2.1 are considered, and conclusions are drawn about the results.

Assume now that we have N samples $\boldsymbol{z}_c^{(j)}$ for $j = 1, \ldots, N$ of the target distribution, and that for each sample we have the conditional posterior marginal $\{\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j)})\}_{j=1}^N$ for all $z_{-c,i} \in \boldsymbol{z}_{-c}$. The posterior distribution and posterior quantities of $\boldsymbol{z}_c$ is obtained using the standard MCMC estimates described in Section 2.2.1. Moreover, the posterior marginal of $z_{-c,i}$ is found by BMA, estimating

(3.1) with (2.11) as

$$\begin{aligned}\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y}) &= \int \tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c)\pi(\boldsymbol{z}_c\,|\,\boldsymbol{y})\mathrm{d}\boldsymbol{z}_c \\ &\simeq \frac{1}{M}\sum_{j=1}^{M}\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c^{(j)}).\end{aligned} \tag{3.4}$$

Posterior quantities of interest about the posterior marginals approximated with (3.4) can then found with numerical integration in (3.2). The steps of the INLA within Metropolis-Hastings algorithm are summarized in Algorithm 2.

The MCMC with INLA methodology is computationally very slow, as it is a sequential algorithm, wherein each iteration an approximation of the model is made with INLA. Although the INLA alone is quite fast, only a few seconds to a minute, running the approximations in sequence sums up to an underestimated time investment. We will address the MCMC with INLA algorithm as a "proof of concept" for the other INLA within Monte Carlo methods.

3.2 IS with INLA

Gómez-Rubio (2019) proposed the use of a different Monte Carlo method combined with INLA, namely importance sampling (IS; Section 2.2.2). The IS algorithm benefits from its samples being independently drawn from a proposal distribution, allowing INLA to fit the conditional LGMs in parallel. We adopt the model notation from Section 3.1, where $\boldsymbol{z} = (\boldsymbol{x},\boldsymbol{\theta}) = (\boldsymbol{z}_c,\boldsymbol{z}_{-c})$, and we assume that the conditional latent field $\boldsymbol{z}_{-c}\,|\,\boldsymbol{z}_c$ is Gaussian if $\boldsymbol{z}_c$ is fixed to some value $\boldsymbol{z}_c^{(j)}$. The IS algorithm is in this setting responsible for generating these values of $\boldsymbol{z}_c$, and INLA will fit the conditional LGMs on these samples; ultimately, obtaining all posteriors of interest.

Say that the we have the proposal distribution $q(\boldsymbol{z}_c)$ that eclipse the region of all values $\boldsymbol{z}_c$, where the posterior distribution $\pi(\boldsymbol{z}_c\,|\,\boldsymbol{y}) \neq 0$. A sample of $\boldsymbol{z}_c^{(j)}$ is then generated from this proposal distribution and fixed $\boldsymbol{z}_c = \boldsymbol{z}_c^{(j)}$. Then, the conditional LGM on $\boldsymbol{z}_c(j)$ is fit with INLA to obtain the conditional posterior marginals, $\tilde{\pi}(\boldsymbol{z}_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c = \boldsymbol{z}_c^{(j)})$ for all $z_{-c,i} \in \boldsymbol{z}_{-c}$, and the conditional marginal likelihood $\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c = \boldsymbol{z}_c^{(j)})$.

Assuming that the normalizing constant in this model is unknown, the importance weights of the jth sample is calculated by

$$\omega^{(j)} \propto \frac{\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(j)})\pi(\boldsymbol{z}_c^{(j)})}{q(\boldsymbol{z}_c^{(j)})}, \tag{3.5}$$

where the $\boldsymbol{z}$ in (2.14) is replaced with $\boldsymbol{z}_c$. In (3.5), $\pi(\boldsymbol{z}_c^{(j)})$ is the evaluation of our chosen prior distribution for the parameters in $\boldsymbol{z}_c$, $\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(j)})$ the conditional

Algorithm 2: Metropolis-Hastings algorithm with INLA (Gómez-Rubio et al., 2018)

```
- Set $\boldsymbol{z}_c = \boldsymbol{z}_c^{(0)}$
- Fit INLA to model model conditioned on $\boldsymbol{z}_c^{(0)}$:
      $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(0)}) \quad \text{and} \quad \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(0)}),\ \forall z_{-c,i} \in \boldsymbol{z}_{-c}$
for $j$ from $2$ to $N-1$ do
    - Generate proposal $\boldsymbol{z}_c^* \sim q(\cdot \mid \boldsymbol{z}_c^{(j)})$
    - Fit INLA to model model conditioned on $\boldsymbol{z}_c^*$:
          $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^*) \quad \text{and} \quad \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^*),\ \forall z_{-c,i} \in \boldsymbol{z}_{-c}$
    - Compute acceptance probability (usually log scale):
          $\alpha = \min\left\{1, \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^*)\pi(\boldsymbol{z}_c^*)q(\boldsymbol{z}_c^{(j)} \mid \boldsymbol{z}_c^*)}{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(j)})\pi(\boldsymbol{z}_c^{(j)})q(\boldsymbol{z}_c^* \mid \boldsymbol{z}_c^{(j)})}\right\}$
    - Sample $u \sim \mathcal{U}[0,1]$
    if $u < \alpha$ then
        - $\boldsymbol{z}_c^{(j+1)} \leftarrow \boldsymbol{z}_c^*$
        - $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j+1)}) \leftarrow \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^*),\ \forall z_{-c,i} \in \boldsymbol{z}_{-c}$
    else
        - $\boldsymbol{z}_c^{(j+1)} \leftarrow \boldsymbol{z}_c^{(j)}$
        - $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j+1)}) \leftarrow \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j)}),\ \forall z_{-c,i} \in \boldsymbol{z}_{-c}$
- Estimate $\tilde{\pi}(\boldsymbol{z}_c \mid \boldsymbol{y})$ from $\{\boldsymbol{z}_c^{(j)}\}_{j=1}^N$ using kernel density estimation
- Compute posterior marginals using BMA:
      $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}) = \frac{1}{N}\sum_{j=1}^{N} \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j)}),\ \forall z_{-c,i} \in \boldsymbol{z}_{-c}$
```

marginal likelihood approximated with INLA, and $q(\boldsymbol{z}_c^{(j)})$ the evaluation of the proposal distribution.

Suppose that the simulation of the IS with INLA algorithm has obtained the weighted set of N samples $\{\boldsymbol{z}_c^{(j)}, \omega^{(j)}\}_{j=1}^N$, and the corresponding N conditional posterior marginals approximated with INLA for all $z_{-c,i} \in \boldsymbol{z}_{-c}$. Then, the importance weights calculated with (3.5) is normalized according to (2.15), and the self-normalized estimator from (2.16) is given as

$$\begin{aligned} \mathbb{E}_\pi[f(\boldsymbol{z}_c)] &= \int f(\boldsymbol{z}_c)\pi(\boldsymbol{z}_c \mid \boldsymbol{y})\mathrm{d}\boldsymbol{z}_c \\ &\simeq \sum_{j=1}^{N} f(\boldsymbol{z}_c^{(j)}) \cdot \bar{\omega}^{(j)} = \widehat{\mathbb{E}}_\pi[f(\boldsymbol{z}_c)], \end{aligned} \tag{3.6}$$

where $\bar{\omega}^{(j)}$ is the self-normalized version of the importance weights in (3.5). Note that to compute the normalization term $\sum \omega^{(j)}$ we need to have access to all simulated elements.

An approximation of the posterior distribution $\pi(\boldsymbol{z}_c \mid \boldsymbol{y})$ is found using nonparametric kernel density estimation as described in Section 2.2.2, and the posterior mode is found with the resulting kernel. Other posterior quantities, for example, mean, variance, and correlation is estimated with (3.6). Similar to MCMC, the posterior marginals of $z_{-c,i}$ is approximated using BMA, which in IS with INLA is achieved by solving (3.1) with (3.6) as

$$\begin{aligned} \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}) &= \int \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c)\pi(\boldsymbol{z}_c \mid \boldsymbol{y})\mathrm{d}\boldsymbol{z}_c \\ &\simeq \sum_{j=1}^{N} \bar{\omega}^{(j)} \cdot \tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(j)}), \ \ \forall \ z_{-c,i} \in \boldsymbol{z}_{-c}. \end{aligned} \tag{3.7}$$

Posterior quantities about these marginals are estimated using the numerical integration method in (3.2).

If it is easy to choose a good enough proposal for $\pi(\boldsymbol{z}_c \mid \boldsymbol{y})$; then, the standard IS with INLA algorithm described above can be used. However, in some applications, the dimensions of $\boldsymbol{z}_c$ is high, and it is not easy to choose an appropriate proposal distribution. To account for this, we have, in our implementation of IS with INLA, decided to add an initial search for the probability mass of the target distribution. This search is not mentioned in Gómez-Rubio (2019), but we have observed in our experiments that it generally improves the effectiveness of the algorithm. The search is carried out similarly to the adaptation in the AMIS algorithm, and we assume that the proposal distribution belongs to a parametric family of distributions $\{q(\boldsymbol{z}_c; \boldsymbol{\phi}) \mid \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$, where $\boldsymbol{\Phi}$ is the parametric space. Then, consider the initial parameters $\boldsymbol{\phi}_0$ describing the initial proposal distribution, whereby N_0 samples are generated and weighted according to the IS with INLA scheme. Using the now obtained weighted set of samples, moments can be approximated with (3.6), and the parameters $\boldsymbol{\phi}_0$ are updated with these directly or some transformation of them; for example, location, shape, and scale. Following the search mentioned above, the initial N_0 samples are thrown away, and the IS with INLA algorithm employs the updated proposal distribution $q(\boldsymbol{z}_c; \hat{\boldsymbol{\phi}})$ to simulate the posteriors of the model. The individual steps of our implementation of IS with INLA are summarized in Algorithm 3.

We want to mention that this initial search does not guarantee any improvements of the proposal distribution. In some situations, it might even prove detrimental to the posterior estimates. Its accuracy also highly relies on the choice of N_0 and the vagueness of the initial proposal distribution. Even though increasing N_0 might improve the search, throwing away more samples is a waste.

Algorithm 3: IS with INLA (Gómez-Rubio, 2019), with search

```
- Initialize $N_0$, $q_0(\cdot;\boldsymbol{\phi}_0)$, $N$
for $j$ from 1 to $N_0$ do
    - Generate sample $\boldsymbol{z}_c^{(0,j)} \sim q_0(\cdot;\boldsymbol{\phi}_0)$
    - Fit INLA to model conditioned on $\boldsymbol{z}_c^{(0,j)}$:
        $\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(0,j)})$ and $\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c^{(0,j)}),\ \forall\ z_{-c,i}\in\boldsymbol{z}_{-c}$
    - Compute
        $\omega^{(0,j)} = \dfrac{\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(0,j)})\pi(\boldsymbol{z}_c^{(0,j)})}{q_0(\boldsymbol{z}_c^{(0,j)};\boldsymbol{\phi}_0)}$
- Compute parameter estimates $\hat{\boldsymbol{\phi}}$ for the new proposal $q(\cdot;\hat{\boldsymbol{\phi}})$ from the
  weighted set of samples :
    $(\{\boldsymbol{z}_c^{(0,1)},\omega^{(0,1)}\},\ldots,\{\boldsymbol{z}_c^{(0,N_0)},\omega^{(0,N_0)}\})$
for $j$ from 1 to $N$ do
    - Generate sample $\boldsymbol{z}_c^{(j)} \sim q(\cdot;\hat{\boldsymbol{\phi}})$
    - Fit INLA to model conditioned on $\boldsymbol{z}_c^{(j)}$:
        $\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(j)})$ and $\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c^{(j)}),\ \forall\ z_{-c,i}\in\boldsymbol{z}_{-c}$
    - Compute
        $\omega^{(j)} = \dfrac{\tilde{\pi}(\boldsymbol{y}\,|\,\boldsymbol{z}_c^{(j)})\pi(\boldsymbol{z}_c^{(j)})}{q(\boldsymbol{z}_c^{(j)};\hat{\boldsymbol{\phi}})}$
- Estimate $\tilde{\pi}(\boldsymbol{z}_c\,|\,\boldsymbol{y})$ from $\{\boldsymbol{z}_c^{(j)}\}_{j=1}^{N}$
- Compute posterior marginals using BMA:
    $\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y}) = \sum_{j=1}^{N}\omega^{(j)}\tilde{\pi}(z_{-c,i}\,|\,\boldsymbol{y},\boldsymbol{z}_c^{(j)}) \Big/ \sum_{j=1}^{N}\omega^{(j)}$
```

3.3 AMIS with INLA

Following the development of more robust IS schemes, we now propose the *adaptive multiple* importance sampling (AMIS) with INLA algorithm to extend the models that can be fit with INLA through the **R-INLA** package. The AMIS algorithm is presented in Section 2.2.3, and is similar to the initial search for the probability mass of the target distribution in our implementation of the IS with INLA method, where AMIS performs many of these adaptations without discarding the generated samples. There is a significant computational cost of fitting a model with INLA, and to throw away potentially valuable information is not favorable, particularly in very complex models. The samples are kept by constructing a mixture of all adapted proposal distribution, taking advantage of the strengths the multiple IS weighing scheme in (2.20) carries.

We maintain the ensemble notation of the unknown parameters in the hierarchical model described in Section 2.1, $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{\theta}) = (\boldsymbol{z}_c, \boldsymbol{z}_{-c})$, and with $\boldsymbol{z}_c$ fixed to some value $\boldsymbol{z}_c^{(j)}$ assume that the latent field is Gaussian. We will use AMIS to generate a series of these values for $\boldsymbol{z}_c$, and use INLA to fit the conditional models. Consider the initial parameters $\boldsymbol{\phi}_1$ of the parametric proposal distribution $q_1(\boldsymbol{z}_c; \boldsymbol{\phi}_1)$, whereby N_1 samples of $\boldsymbol{z}_c$ are generated. Conditioned on these samples the model is fit with INLA obtaining N_1 conditional posterior marginals, $\tilde{\pi}(z_{-c,1} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(1,j)})$, and conditional marginal likelihoods $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(1,j)})$ for $j = 1, \dots, N_1$. Furthermore, the samples are weighted according to

$$\omega^{(t,j)} \propto \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(t,j)})\pi(\boldsymbol{z}_c^{(t,j)})}{\psi_t(\boldsymbol{z}_c^{(t,j)})}, \tag{3.8}$$

where in this particular case, $t = 1$. In (3.8), $\pi(\boldsymbol{z}_c)$ is the chosen prior distribution of $\boldsymbol{z}_c$, and $\psi_t(\boldsymbol{z}_c)$ is the mixture distribution (2.19) at step t. For $t = 1$, the mixture consists of only the initial proposal distribution $q_1(\cdot; \boldsymbol{\phi}_1)$. Note that we assume that the normalizing constant $\pi(\boldsymbol{y})$ is unknown, such that during the simulation the weights are set equal to (3.8); then, after the simulation or when a estimate is required, the weights are normalized similar to (2.22).

To adapt the proposal distribution, we will use the *moment matching criterion* between the proposal and the joint posterior distribution of $\boldsymbol{z}_c$. In the AMIS with INLA algorithm, the posterior expected value of any function on $\boldsymbol{z}_c$ is estimated as

$$\begin{aligned}\mathbb{E}_\pi[f(\boldsymbol{z}_c)] &= \int f(\boldsymbol{z}_c)\pi(\boldsymbol{z}_c \mid \boldsymbol{y})\mathrm{d}\boldsymbol{z}_c \\ &\simeq \sum_{l=1}^{t}\sum_{j=1}^{N_l} f(\boldsymbol{z}_c^{(l,j)}) \cdot \bar{\omega}^{(l,j)} = \widehat{\mathbb{E}}_\pi[f(\boldsymbol{z}_c)].\end{aligned} \tag{3.9}$$

Here, $\bar{\omega}^{(l,j)}$ denotes the normalized weights of the jth sample drawn from the lth proposal distribution. For example, a approximation of the mean $\bar{\boldsymbol{z}_c}$ is obtained with $f(\boldsymbol{z}_c) = \boldsymbol{z}_c$ in (3.9), and the variance is estimated with $f(\boldsymbol{z}_c) = (\boldsymbol{z}_c - \bar{\boldsymbol{z}_c})^2$.

Conclusively, the moments are matched by applying these estimated moments to the moments of the proposal distribution and, in effect, the parameter $\boldsymbol{\phi}_1$ is updated and the proposal distribution is adapted to $q_2(\boldsymbol{z}_c; \boldsymbol{\phi}_2)$

Ensuing this initial adaptation, the new proposal distribution is added to the mixture distribution $\psi_2(\cdot)$ according to (2.19), and the process continues by drawing N_2 new samples and weighing them with (3.8) for $t = 2$. In addition, the importance weights of all past samples are update with this new mixture in (3.8). The AMIS with INLA algorithm comprise of T such adaptations or epochs, and in an epoch arbitrary epoch $t \in (1, T)$, N_t samples are generate and weighed according to (3.8); then, an estimate of (3.9) is calculated to adapt the proposal; lastly, the proposal distribution is added to the mixture (2.19), and the weights of all accumulated samples are update with this new mixture. At epoch T, the algorithm has generated $N = \sum_{t=1}^{T} N_t$ weighted samples $\{\boldsymbol{z}_c^{(t,j)}\}_{t=1,j=1}^{T,N_t}$, and the corresponding conditional posterior marginal $\{\tilde{\pi}(z_{-c,i} \,|\, \boldsymbol{y}, \boldsymbol{z}_c^{(t,j)})\}_{t=1,j=1}^{T,N_t}$.

The posterior distribution of $\boldsymbol{z}_c$ is approximated with non-parametric kernel density estimation (Silverman, 1986), and the mode is obtained by the maximum value of this kernel. Using BMA, the posterior marginals $\pi(z_{-c,i} \,|\, \boldsymbol{y})$ are obtained. They are attained by estimating (3.1) with (3.9):

$$\begin{aligned}\tilde{\pi}(z_{-c,i} \,|\, \boldsymbol{y}) &= \int \tilde{\pi}(z_{-c,i} \,|\, \boldsymbol{y}, \boldsymbol{z}_c)\pi(\boldsymbol{z}_c \,|\, \boldsymbol{y})\mathrm{d}\boldsymbol{z}_c \\ &\simeq \sum_{t=1}^{T}\sum_{j=1}^{N} \bar{\omega}^{(t,j)} \cdot \tilde{\pi}(z_{-c,i} \,|\, \boldsymbol{y}, \boldsymbol{z}_c^{(t,j)}), \;\; \forall \; z_{-c,i} \in \boldsymbol{z}_{-c}.\end{aligned} \tag{3.10}$$

If posterior estimates of the expected value of some function $f(\cdot)$ about these posterior marginals is of interest, they can obtained with the method in (3.2).

Note that it is also possible to determine if the proposal distribution needs to continue adapting by calculating the effective sample size per sample, i.e., dividing (2.17) by the number of generated samples. Thereby, if the estimate is higher than a set threshold, the adaptation stops, and the last proposal distribution generates samples until some predetermined number of effective samples is obtained. However, we have decided not to add this criterion to our implementation.

Similar to IS, the samples in the AMIS algorithm are drawn independently of each other within an epoch. This allows the approximation of the conditional models with INLA to be computed in parallel between the adaptations, resulting in a significant increase in computation speeds. The individual steps in the AMIS with INLA algorithm are shown in Algorithm 4.

Algorithm 4: AMIS with INLA

```
- Initialize $\boldsymbol{N}_t = (N_1, \dots, N_T)$, $q_1(\cdot; \boldsymbol{\phi}_1)$
for $j$ from $1$ to $N_1$ do
    - Generate sample $\boldsymbol{z}_c^{(1,j)} \sim q_1(\cdot; \boldsymbol{\phi}_1)$
    - Fit INLA to model conditioned on $\boldsymbol{z}_c^{(1,j)}$:
          $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(1,j)})$ and $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(1,j)}), \ \forall \ z_{-c,i} \in \boldsymbol{z}_{-c}$
    - Compute:
          $\delta^{(1,j)} = N_1 q_1(\boldsymbol{z}_c^{(1,j)}; \boldsymbol{\phi}_1)$ and $\omega^{(1,j)} = \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(1,j)}) \pi(\boldsymbol{z}_c^{(1,j)})}{q_1(\boldsymbol{z}_c^{(1,j)}; \boldsymbol{\phi}_1)}$
- Compute parameter estimates $\boldsymbol{\phi}_2$ of the weighted set of samples:
      $(\{\boldsymbol{z}_c^{(1,1)}, \omega^{(1,1)}\}, \dots, \{\boldsymbol{z}_c^{(1,N_1)}, \omega^{(1,N_1)}\})$

for $t$ from $2$ to $T$ do
    for $j$ from $1$ to $N_t$ do
        - Generate sample $\boldsymbol{z}_c^{(t,j)} \sim q_t(\cdot; \boldsymbol{\phi}_t)$
        - Fit INLA to model conditioned on $\boldsymbol{z}_c^{(t,j)}$:
              $\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(t,j)})$ and $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(t,j)}), \ \forall \ z_{-c,i} \in \boldsymbol{z}_{-c}$
        - Compute:
              $\delta^{(t,j)} = \sum_{l=1}^{t} N_l q_t(\boldsymbol{z}_c^{(t,j)}; \boldsymbol{\phi}_t)$ and $\omega^{(t,j)} = \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(t,j)}) \pi(\boldsymbol{z}_c^{(t,j)})}{\left[\delta^{(t,j)} / \sum_{l=1}^{t} N_l\right]}$
    for $l$ from $1$ to $t-1$ do
        for $j$ from $1$ to $N_l$ do
            - Update past importance weights:
                  $\delta^{(l,j)} \leftarrow \delta^{(l,j)} + N_l q_t(\boldsymbol{z}_c^{(l,j)}; \boldsymbol{\phi}_t)$ and $\omega^{(l,j)} \leftarrow \frac{\tilde{\pi}(\boldsymbol{y} \mid \boldsymbol{z}_c^{(l,j)}) \pi(\boldsymbol{z}_c^{(l,j)})}{\left[\delta^{(l,j)} / \sum_{k=1}^{t} N_k\right]}$
    - Compute parameter estimates $\boldsymbol{\phi}_t$ of the weighted set of samples:
          $(\{\boldsymbol{z}_c^{(1,1)}, \omega^{(1,1)}\}, \dots, \{\boldsymbol{z}_c^{(t,N_t)}, \omega^{(t,N_t)}\})$
- Estimate $\tilde{\pi}(\boldsymbol{z}_c \mid \boldsymbol{y})$
- Compute posterior marginals using BMA:
      $\tilde{\pi}(z_{-c,i} \mid \boldsymbol{y}) = \sum_{t=1}^{T} \sum_{j=1}^{N_t} \omega^{(t,j)} \tilde{\pi}(\boldsymbol{z}_{-c,i} \mid \boldsymbol{y}, \boldsymbol{z}_c^{(t,j)}) \Big/ \sum_{t=1}^{T} \sum_{j=1}^{N_t} \omega^{(t,j)}$
```

Chapter 4

Examples

This chapter will present some application of the INLA within Monte Carlo methods and compare each algorithm on efficiency, accuracy, and robustness. To determine the effectiveness of a method, we will use the effective sample size per second. Note that this estimate may not be the best method for comparing the efficiency of the methods (Elvira et al., 2018); however, as the methods highly rely on Monte Carlo integration (2.7), and since the effective samples size convey the variance of this integration, it will give some indication of efficiency. The accuracy of the methods is deduced by either comparing posterior densities or posterior statistics to the truth, exact methods, or more established methods (e.g. MCMC). The robustness is evaluated by the degree of manual tweaking and re-simulations needed in the methods to reach convergence and obtain good approximations.

Section 4.1 will illustrate the performance and behavior of the combined INLA and Monte Carlo methods on a simple bivariate linear model. The posteriors are compared to the INLA, which is exact for this model up to an integration error. In the second example Section 4.2, we apply the algorithms on a spatial autoregressive combined model, presented in Gómez-Rubio et al. (2019). They use a grid exploration with INLA to obtain posteriors, and our goal is to investigate the performance of the INLA within Monte Carlo methods on the same problem. Next, we present a novel approach to Bayesian quantile regression using AMIS with INLA, and validate it in a simulation study before applying it to existing datasets. In the last example, we attempt an approximation of a Gamma frailty model using AMIS with INLA, and analyze it in a simulation study.

4.1 Bivariate linear model

In the first example, we consider a bivariate linear model on a simulated dataset. A dataset of 100 observations is obtained by drawing samples of the covariates x_{1i} and x_{2i} from a uniform distribution between zero and one. Furthermore, we apply a Gaussian noise term ϵ_i with mean zero and precision τ, such that the response is calculated by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \text{ for } i = 1, \dots, 100.$$

We have chosen the parameters $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -1$ and $\tau = 1$. In all methods, we have assigned the linear effects $\boldsymbol{\beta} = (\beta_1, \beta_2)$ the default prior in **R-INLA**; a product of two Gaussian distributions with zero mean and precision 0.001. Moreover, we have applied the default priors in **R-INLA** to the intercept β_0 and precision τ; a Gaussian distribution with zero mean and zero precision and Gamma distribution with parameters 1 and $5e-5$, respectively.

This model can be easily fitted using INLA alone, and since the likelihood is Gaussian, the INLA approximations are exact up to an integration error. However, the INLA will only obtain the posterior marginals so, for example, joint inference on β_1 and β_2 is not possible. Therefore, the scope of this example is double; on one side, we can compare the results of the combined algorithms with the exact INLA results; on the other side, we can show how the combined strategy also allows for joint inference. For this example the vector of unknown parameters is $\boldsymbol{z} = (\beta_0, \beta_1, \beta_2, \tau)$, and we set $\boldsymbol{z}_c = (\beta_1, \beta_2)$ and $\boldsymbol{z}_{-c} = (\beta_0, \tau)$.

In the INLA within Metropolis-Hastings algorithm, we have chosen a bivariate Gaussian proposal distribution with mean equal to the previous state $\boldsymbol{\beta}^{(j)}$ and variance of $0.75^2 \cdot \mathbf{I}$, and we set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ as starting value. For the proposal distribution in the AMIS with INLA and IS with INLA methods, we have chosen a bivariate Gaussian with initial parameters $\boldsymbol{\phi}_1 = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_1 = \mathbf{0}$ is the initial mean and $\boldsymbol{\Sigma}_1 = 5 \cdot \mathbf{I}$ is the initial variance-covariance matrix. These proposal distributions are a little vague but will be illustrative for the adaptation in the algorithms.

In the MCMC with INLA algorithm, 10500 samples are drawn, where a burn-in of 500 is used. Figure 4.1 shows the trace of samples values for β_1 and β_2. The MCMC appears to converge and mix well. The IS with INLA algorithm generates $N_0 = 800$ samples and adapts the proposal distribution throwing away these samples. Then, it generates $N = 10000$ samples. For the AMIS with INLA algorithm, a total of 10000 samples are drawn by adapting the proposal distribution $T = 27$ times. The initial distribution draws $N_1 = 250$ samples, and the remaining distributions generates $N_t = (250, 260, \dots, 490, 500)$ samples, where $t = 2, \dots, T$. Figure 4.2 shows the adaptation of the proposal distribution in AMIS with INLA and IS with INLA. The methods have no issues finding the probability mass of the posterior distribution in just one adaptation, and in this example, it would be better to stop the adaptation in AMIS with INLA at $t = 2$. This could have been done using the effective sample size per sample stopping criterion described in Section 3.3.

In Figure 4.3, the approximated posterior marginals of β_0, β_1, β_2, and τ from the combined approaches are presented. The posterior marginal from INLA alone and the true values of the parameters are also included for reference. Here, all posterior marginals seem to follow the guidelines set by INLA, where the MCMC with INLA provides some small inaccuracies around the posterior mode of β_1 and

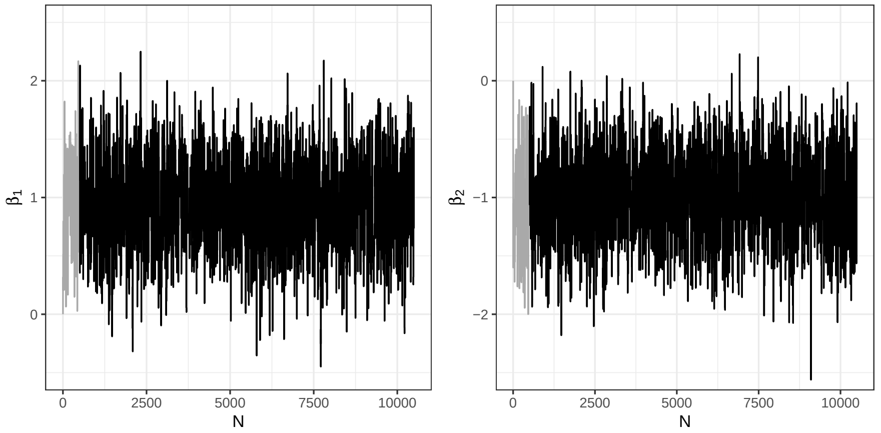

Figure 4.1: Trace plot of the Markov chain of the $\boldsymbol{\beta}$s in the bivariate linear model constructed by the MCMC with INLA method with the burn-in highlighted (grey).

β_2. Similar inaccuracies are present in Figure 4.4, where the joint posterior distribution of $\boldsymbol{\beta}$ is shown.

In addition, Figure 4.4 presents the running effective samples size obtained with the combined approaches, and the MCMC with INLA method has achieved much fewer effective samples than the other methods in a much longer time. The IS with INLA method achieved an effective sample size of 9137 in 3 minutes and 5 seconds, 49.2 effective samples per second; AMIS with INLA obtained 9618 effective samples in 8 minutes and 12 seconds, 19.5 effective samples per second; Lastly, the MCMC with INLA found a Markov chain of 1120 effective samples in 53 minutes, 0.35 effective samples per second.

4.2 Spatial autoregressive combined model

In this example, we will apply the IS with INLA and AMIS with INLA algorithms on a specific kind of spatial econometric model (SEM; see LeSage et al. (2009) for a thorough account). These models comprise of one or many spatial autoregressive terms that control the spatial dependencies and interactions in the data. We will consider the spatial autoregressive combined (SAC) model proposed by Manski (1993), where the response $\boldsymbol{y}$ is modelled by a autoregressive term ρ on the response:

$$\boldsymbol{y} = \rho \mathbf{W}\boldsymbol{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{u}. \tag{4.1}$$

Here, the data is collected over n areas and $\mathbf{X}$ are the covariates of effect $\boldsymbol{\beta}$, $\mathbf{W}$ is the adjacency matrix of the n areas, and $\mathbf{WX}$ are the lagged covariates of effect $\boldsymbol{\gamma}$. The adjacency matrix $\mathbf{W}$ of size $n \times n$ is constructed such that if the area i and area j are neighbors, the element (i, j) in $\mathbf{W}$ will be 1. Subsequently, the matrix is row-standardized such that every row sum to one, which in turn makes the spatial autocorrelation parameters bound to the interval $(1/\lambda_{\min}, 1)$, where $\lambda_{\min}$ is the

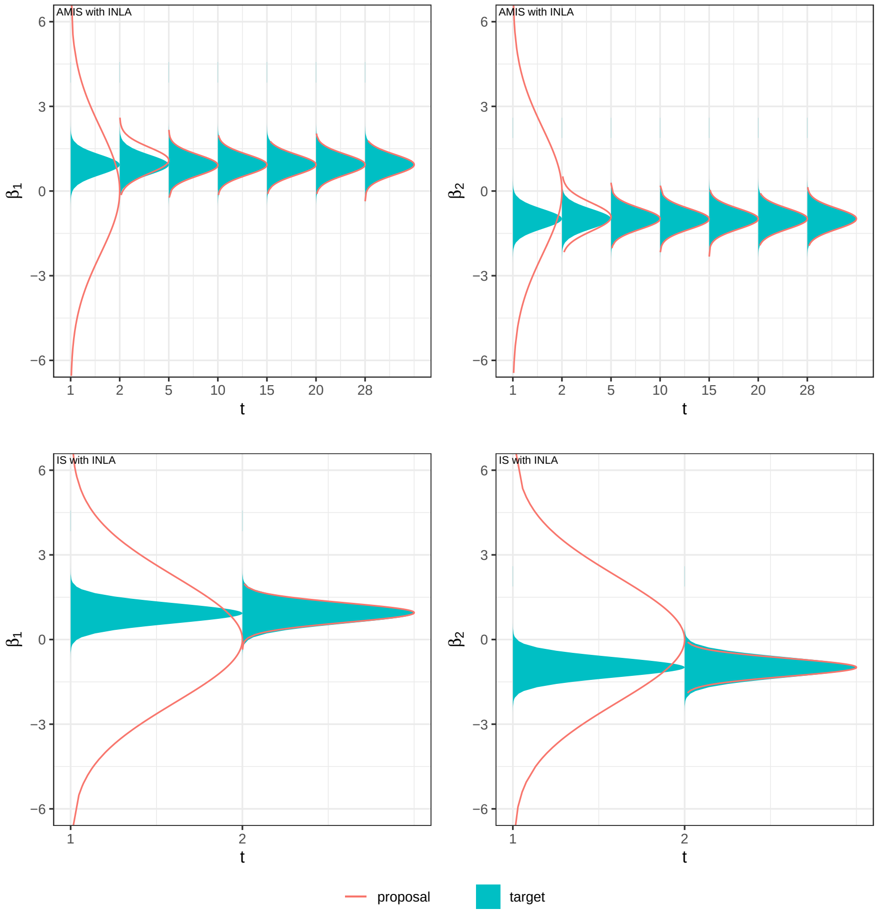

Figure 4.2: A visual representation of the adaptation of proposal distribution in AMIS with INLA (top), and the initial search in IS with INLA (bottom) for the $\boldsymbol{\beta}$s in the bivariate linear model. The x-axis is the number of adaptations T of the proposal distribution.

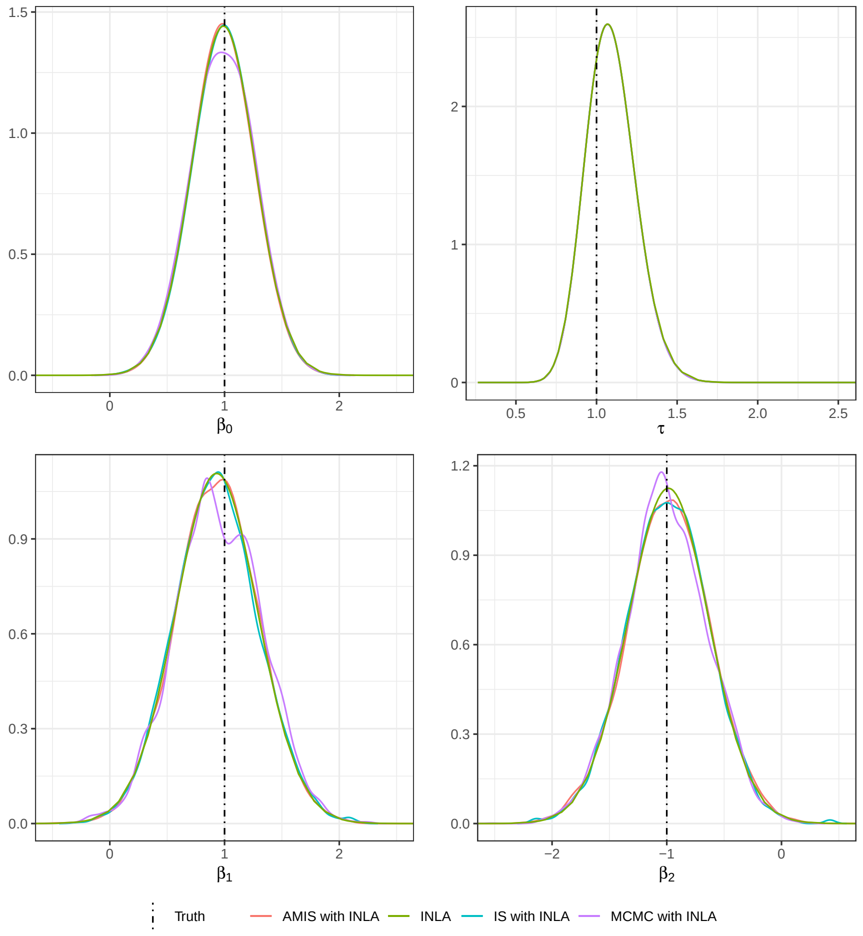

Figure 4.3: Posterior marginals of all parameters in the bivariate linear model approximated with AMIS with INLA, INLA, IS with INLA, and MCMC with INLA. The line (-·-·) are the values of the parameters chosen for the simulation of data.

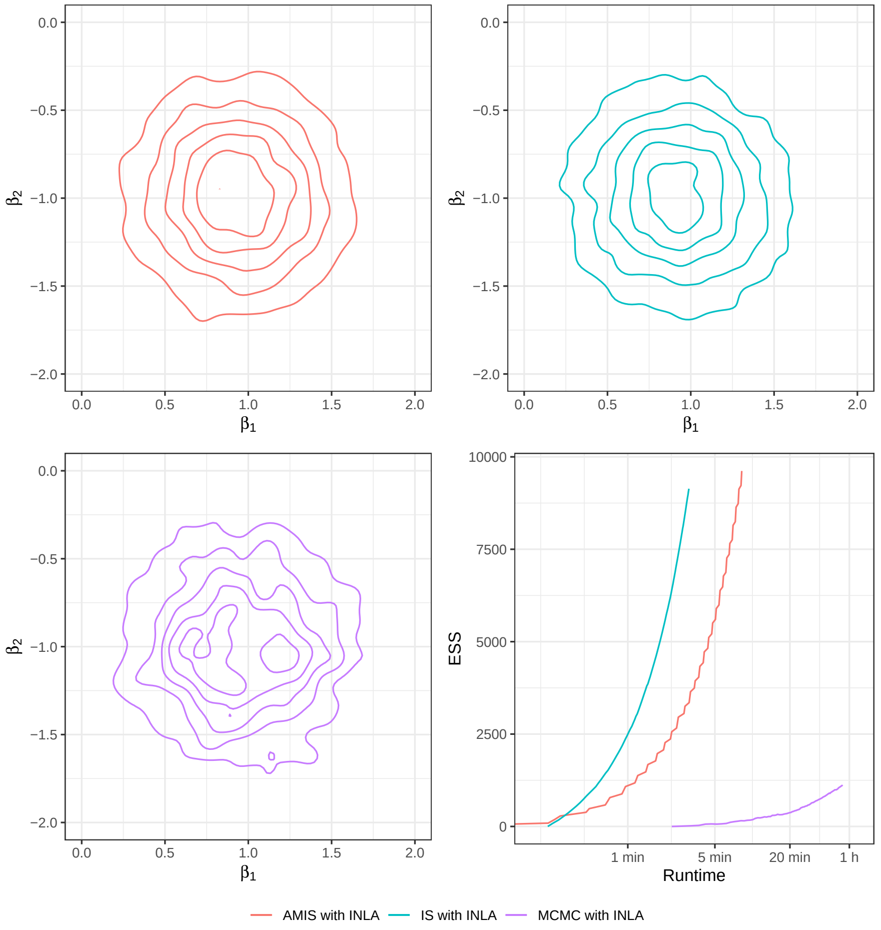

Figure 4.4: The joint posterior distribution of $\boldsymbol{\beta}$ in the bivariate linear model, and the running effective sample size (bottom right) obtained using AMIS with INLA, IS with INLA, and MCMC with INLA.

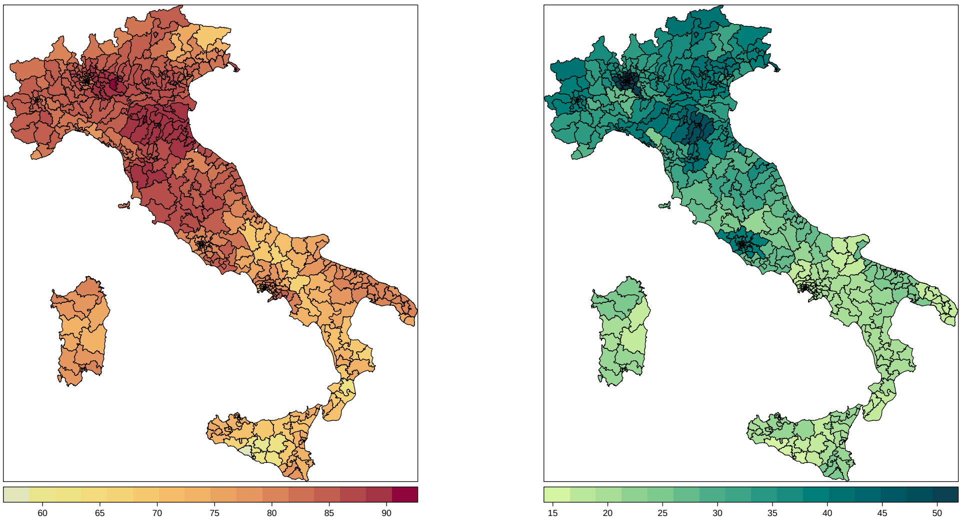

Figure 4.5: Values of the election turnover in 2001 (left) and GDP per capita in 1997 (right) from single-member districts in Italy.

minimum eigenvalue of $\mathbf{W}$. In this model, $\boldsymbol{u}$ is an error term that is modelled with a spatial autoregressive term λ on the error term as

$$\boldsymbol{u} = \lambda \mathbf{W}\boldsymbol{u} + \boldsymbol{\epsilon}_1, \tag{4.2}$$

where $\boldsymbol{\epsilon}_1$ is Gaussian noise term with zero mean and precision τ.

The response in (4.1) is rewritten as

$$\boldsymbol{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}) + \boldsymbol{\epsilon}_2, \tag{4.3}$$

with the revised error term:

$$\boldsymbol{\epsilon}_2 \sim \mathcal{N}\left(\mathbf{0}, \tau(\mathbf{I} - \rho\mathbf{W}^T)(\mathbf{I} - \lambda\mathbf{W}^T)(\mathbf{I} - \lambda\mathbf{W})(\mathbf{I} - \rho\mathbf{W})\right). \tag{4.4}$$

Observe that we have the non-additive term $(\mathbf{I} - \rho\mathbf{W})$ in (4.3), and that the error term in (4.4) have a very complex structure. Hence, the model cannot be fit with INLA directly unless we condition on the spatial autoregressive terms ρ and λ and, thus, $\boldsymbol{z}_c = (\rho, \lambda)$.

In this example, we consider the turnover dataset described in Michael D. Ward (2008), which is obtainable at `http://ksgleditsch.com/srm_book.html`. The dataset contains election turnovers in Italy from 2001, and the GDP per capita (GDPCAP) from 1997 for $n = 477$ areas. These areas are the single-member districts in Italy. The spatial distribution of the variables, turnover and GDP per capita, are shown in Figure 4.5, and we want to model the turnover using the SAC model with ln(GDPCAP) as covariate and a intercept, i.e. the effect β_1 for ln(GDPCAP) and β_0 for the intercept.

This example is similar to the one presented in Gómez-Rubio et al. (2019), where they construct a grid on ρ and λ, and use INLA to obtain weighted grid points. The conditional models from INLA are then combined with Bayesian model averaging. Their aim is to estimate the impacts of covariates on the neighboring responses; however, we will, in our example, focus on posterior statistics. We have used the same adjacency of areas as Gómez-Rubio et al. (2019), where areas with centroids closer than 50 km are neighbors. The exact dataset we used, with this adjacency matrix included, is found in the code repository of Gómez-Rubio et al. (2019) (`https://github.com/becarioprecario/SAC_INLABMA`).

We know apriori that the autoregressive terms are in the range $(1/\lambda_{\min}, 1)$, and that the minimum eigenvalue of **W** is $\lambda_{\min} = -0.82$. Thus, the auto regressive terms in $\boldsymbol{z}_c$ are assigned a uniform prior distribution between -1 and 1. Furthermore, the parameters β_0 and β_1 are given a Gaussian prior with mean zero and variance 1000, and the precision τ of $\boldsymbol{\epsilon}_1$ is assigned a Gamma prior with shape 0.01 and rate (inverse-scale) 0.01. These are the same priors used by Gómez-Rubio et al. (2019).

In the AMIS with INLA and IS with INLA algorithm, we have used the bivariate Student's t proposal distribution with three degrees of freedom for $\boldsymbol{z}_c$, where initial parameters for the proposal distribution in both methods are a zero mean and variance $2 \cdot \mathbf{I}$. We apply an initial search of $N_1 = 800$ samples for the probability mass of the posterior distribution of $\boldsymbol{z}_c$ in the IS with INLA algorithm; then, $N_2 = 10000$ is drawn from the new proposal. In the AMIS with INLA algorithm, we initially generate $N_1 = 250$ samples. Next, the proposal distribution is adapted $T = 27$ times, and each new proposal distribution draws $N_t = (250, 260, \ldots, 490, 500)$ samples. This results in a total of $N = 10000$ samples and conditional model approximations with INLA.

To compare our results, we have estimated the model with a standalone MCMC algorithm for SAC models available in the `R` package **spatialreg** in the function `spBreg_sac` (Bivand et al., 2013; Bivand et al., 2015b), which was also used for comparison in Gómez-Rubio et al. (2019). In the MCMC simulation, 110000 samples where drawn with a burn in of 10000, and every 10th samples are kept to reduce auto-correlation. Resulting in Markov chain of 10000 samples obtained with the MCMC algorithm.

The posterior marginals of the intercept β_0, the effect of log GDP per capita β_1, and the precision of the noise τ are presented in Figure 4.6. The INLA with Monte Carlo methods approximated close to similar posterior marginals as the MCMC method. Figure 4.6 also show the approximated joint posterior distribution of the spatial autoregressive terms in $\boldsymbol{z}_c$. Again, the distribution is quite similar between the methods. Overall, the MCMC algorithm estimates a slightly smaller variance than the combined approaches, and there are also some noticeable differences in the posterior mean. The effective sample size obtained with the MCMC algorithm was 295 for the ρ parameter. The IS with INLA method found 3222 effective samples in 62 minutes, 0.86 effective samples per second, and AMIS with INLA 4999 in 75 minutes a resulting in 1.1 effective samples per second.

We have omitted the results of the MCMC with INLA algorithm from tables and figures because of its low performance compared to the others. It obtained an effective sample size of 64 in 8 hours, such that even with the unfeasible number of 10000 effective samples, it would sill have lower effective samples size per second than the AMIS with INLA and IS with INLA algorithms. The posterior distributions obtained with MCMC with INLA is compared to those from AMIS with INLA in Figure A.1.

Parameter	MCMC	IS with INLA	AMIS with INLA
β_0	5.76(2.34)	6.17(2.46)	6.11(2.42)
β_1	1.75(0.59)	1.84(0.61)	1.83(0.61)
ρ	0.86(0.04)	0.84(0.07)	0.84(0.07)
λ	0.21(0.11)	0.25(0.13)	0.24(0.13)
τ	0.26(0.02)	0.26(0.02)	0.26(0.02)

Table 4.1: Posterior mean and standard deviation (in parenthesis) estimated by the MCMC, IS with INLA, and AMIS with INLA algorithm on the SAC model.

4.3 Model-aware Bayesian quantile regression

Quantile regression is used to understand the relationship between the quantiles of the response and the covariates and was introduced by Koenker et al. (1978). The frequentists' approach to quantile regression is well developed and relying on minimizing a loss function. In the Bayesian framework, a common approach to quantile regression is to employ a likelihood function based on the asymmetric Laplace distribution (Yu et al., 2001). These approaches are generally non-parametric (no unknown parameters in the model) or semi-parametric, and have a likelihood but no model for the data. Padellini et al. (2019) proposed a fully parametric approach by modifying the link function in **R-INLA** to link the linear predictor to the quantiles of the response. Their approach is model-based and assumes Poisson distributions for the data, which conveniently has only one likelihood parameter because the mean is equivalent to the variance and the model can, therefore, be approximate with **R-INLA**.

Our approach is similar to Noufaily et al. (2013) and Padellini et al. (2019) in that we assume a likelihood model that describes the data generation process and is not a mere working likelihood as the asymmetric Laplace. Moreover, we follow Noufaily et al. (2013), which show that by modeling all likelihood parameters as a function of covariates, interesting shapes are found in the quantile curves. **R-INLA** does not allow the user to link more than one likelihood parameter to the covariates, and for this reason, we employ our combined methods.

Consider the Gaussian likelihood $y \mid x \sim \mathcal{N}(\mu, \sigma^2)$. The p quantile y_p is given by $y_p = \sigma y_p^* + \mu$, where y_p^* is the p quantile of a standard Gaussian distribution. It is clear that having a fixed scale parameter σ, as is the case in INLA, results in

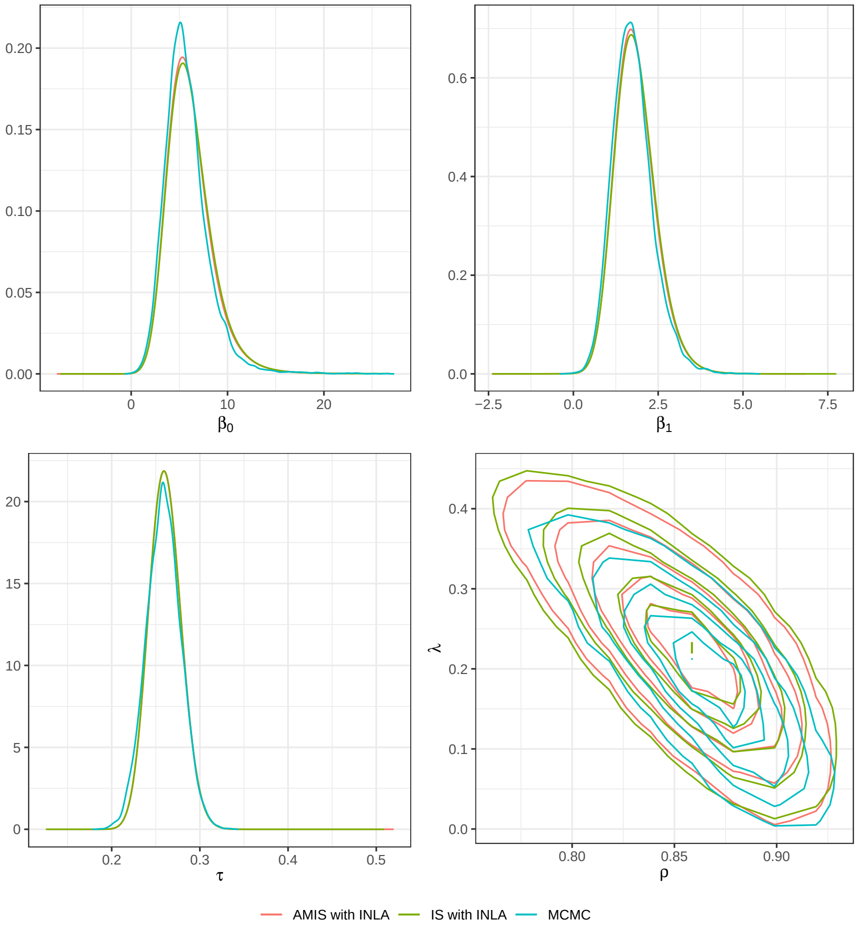

Figure 4.6: Joint posterior distribution of the autoregressive terms ρ and λ (bottom right) and the posterior marginals of the intercept β_0, log GDP per capita β_1, and the precision of the noise τ in the SAC model approximated with AMIS with INLA, IS with INLA, and MCMC.

parallel quantile curves, which are not so interesting. We instead use the model

$$\begin{aligned}\mu(x) &= a + bx\\ \log(\sigma) &= c + dx,\end{aligned}$$

such that both μ and σ are allowed to vary with the covariates, and we have the likelihood

$$y = a + b \cdot x + e^{c} e^{d \cdot x} \cdot \epsilon, \tag{4.5}$$

where ϵ is a realization of the standard normal distribution. Thus, the model is non-additive and cannot be approximated with INLA alone unless we fix c and d to some value.

In the following, we will only consider Gaussian or Gamma likelihoods but many other distributions for the response can be used. For the Gaussian case, similar to (4.5), the likelihood is modeled by the mean $\mu(x)$ and precision $\tau(x)$ as

$$\begin{aligned}y_i &\sim \mathcal{N}\left(\mu(x_i),\ \frac{1}{\tau(x_i)}\right)\\ &= \mathcal{N}\left(a + b \cdot x_i,\ \frac{1}{\exp(c + d \cdot x_i)}\right).\end{aligned} \tag{4.6}$$

In the case of a Gamma likelihood, the mean is modeled by $\mu(x) = \exp(a + b \cdot x)$, and instead of the precision we use the shape $k(x) = \exp(c + d \cdot x)$. The resulting Gamma likelihood is

$$y_i \sim \text{Gamma}\left(k(x_i), \frac{\mu(x_i)}{k(x_i)}\right), \tag{4.7}$$

where $\mu(x)/k(x)$ is the scale, and the precision can be found with $k(x)/\mu(x)^2$. Assuming that a model follows the Gaussian likelihood in (4.6), the associated quantile function can be calculated by

$$y_p(x) = a + b \cdot x + \frac{1}{\sqrt{\exp(c + d \cdot x)}} Q_{\mathcal{N}}(p), \tag{4.8}$$

where p denotes the quantile and $Q_{\mathcal{N}}(p)$ the quantile function of the standard normal distribution. Lastly, the quantile function of the Gamma likelihood in (4.7) is formulated as

$$y_p(x) = \frac{\mu(x)}{k(x)} \cdot Q_{\text{Gamma}}(p\ ; 1, k(x)), \tag{4.9}$$

where $Q_{\text{Gamma}}(p; 1, k(x))$ is the Gamma quantile function with scale one and shape $k(x)$.

As shown in (4.5), the change in the precision as an effect of the covariates is not handled by **R-INLA** alone; however, if conditioned on the values of precision the model can be fit with **R-INLA**. Thus, we propose a new approach to Bayesian quantile regression using the AMIS with INLA algorithm. The other INLA within Monte Carlo methods are also used, but most of their results are presented in the Appendix A. This is due to the fact that IS with INLA algorithm struggles to obtain good approximations when employed with the same proposal distribution as AMIS with INLA, and requires more investigation of the data. The same could be said for the MCMC with INLA, which generally has low acceptance rates and requires many re-runs and manual tweaking. In addition, we choose not to clutter the figures with quantile curves from many methods. We collect the parameters in $\boldsymbol{z} = (a, b, c, d)$ and follow the notation in Section 3.3, where inference about the elements in $\boldsymbol{z}_{-c} = (a, b)$ are obtained with INLA by conditioning on the generated samples of $\boldsymbol{z}_c = (c, d,)$ obtained with the AMIS algorithm. We assume that the parameters $\boldsymbol{z}_{-c}$ are Gaussian given $\boldsymbol{z}_c$, such that inference with INLA is feasible.

Note that this approach is fully parametric for the linear predictors used in (4.6) and (4.7). However as described in Section 2.1.1, INLA can model smooth effects $f(\cdot)$ of the covariates in the linear predictor (2.2), which is a non-parametric estimation (Gómez-Rubio, 2020, Chapter 9). Thus, our quantile regression approach is in some models semi-parametric and, in Section 4.3.2, we demonstrate the application of such a smooth effect. An important consequence of our approach to quantile regression is that the quantile curves cannot cross, which is a major issue in many quantile regression methods and a problem covered in many studies (Rodrigues et al., 2017). If quantile curves were to cross, it would suggest a negative probability, which is highly unreasonable.

In Section 4.3.1, we will conduct a simulation study on two Gaussian and Gamma models to validate our quantile regression approach. Next, the method is applied to a LIDAR dataset, which is modeled with a smooth effect on the covariate, and we assume a Gaussian model. Lastly, we test the approach on the Immunoglobulin G dataset used in Noufaily et al. (2013), and assume a Gamma likelihood. If not otherwise specified, we have in all these examples used a vague prior for the parameters in $\boldsymbol{z}_c$, a Gaussian distribution with mean zero and precision 0.01. Furthermore, b is assigned a Gaussian prior with mean zero and precision 0.001, and a is assigned a Gaussian prior with zero mean and zero precision (the default priors in **R-INLA**).

In the AMIS with INLA algorithm, we have employed a bivariate Student's t-distribution with three degrees of freedom as proposal distribution, where the initial parameters $\boldsymbol{\phi}_1$ are set to mean zero and variance-covariance matrix $10 \cdot \mathbf{I}_2$. Similar to the simulation strategy used in Section 4.1, a total of 10000 by the AMIS with INLA algorithm. Initially $N_1 = 250$ are generated; then, the proposal distribution is adapted $T = 27$ times, where in each adaptation $N_t = (250, 260, \ldots, 500)$ samples are drawn with the new proposal for $t = 2, \ldots, 27$.

4.3.1 Simulation study

In the first examples, we will perform Bayesian quantile regression using the AMIS with INLA algorithm on four simulated datasets of $n = 500$ observations. The covariate x is generated from a uniform distribution between zero and one, and the parameters in the models are set according to Table 4.2. Then, the response is simulated using the associated likelihood, (4.6) or (4.7), which is also highlighted for the respective models in Table 4.2.

Model	a	b	c	d
M1 Gaussian	1	-0.1	-1.5	-2
M2 Gaussian	-2	-5	-2	3.5
M1 Gamma	3	-1	-1	6
M2 Gamma	-3	2	4	-1

Table 4.2: Coefficients of the simulated data for quantile regression.

The sampling strategy and proposal distribution of the AMIS with INLA algorithm, and the priors of the Bayesian parametric quantile regression model is detailed in the introduction of Section 4.3. The posterior statistics and effective sample sizes of all the simulations are presented in Table 4.3. There are some deviations from the parameters set during simulation, and some vary as much as ±0.4 from its actual value. However, since only $n = 500$ observations are included in the dataset, the deviations between the true values of the parameters and the estimated values can be caused by the randomness in the dataset.

Model	a	b	c	d	$\widehat{\text{ESS}}$
M1 Gaussian	1.28(0.23)	0.14(0.53)	-1.49(0.12)	-2.02(0.22)	7879
M2 Gaussian	-1.64(0.14)	-5.38(0.18)	-1.99(0.12)	3.48(0.22)	7961
M1 Gamma	3.03(0.05)	-1.02(0.06)	-0.87(0.11)	5.73(0.20)	7653
M2 Gamma	-3.02(0.01)	2.07(0.03)	3.94(0.12)	-0.88(0.22)	7549

Table 4.3: Estimated posterior statistics of the coefficients in the quantile regression models with simulated data. The standard deviations are presented in the parenthesis, and the last column are the estimated effective sample sizes in the models.

The approximated quantile curves of the four models found using (4.8) or (4.9), and the estimated parameters obtained with AMIS with INLA is presented in Figure 4.7. Here, the quantile curves seem to explain the relationship between the response and the covariates in the data well. Again, there are some slight deviations between the true quantile curves and the estimated ones, but they are quite similar in shape.

4.3.2 Ratio of received light in LIDAR measurements

We will now consider an existing dataset of light detection and ranging (LIDAR) measurements found in Sigrist et al. (1994), where they use the LIDAR data to monitor pollutants. It contains $n = 221$ observations of two variables; the first is the log of the ratio of light received by two lasers, and the second is the distance the light has traveled before it is reflected back to its source. The data are plotted in

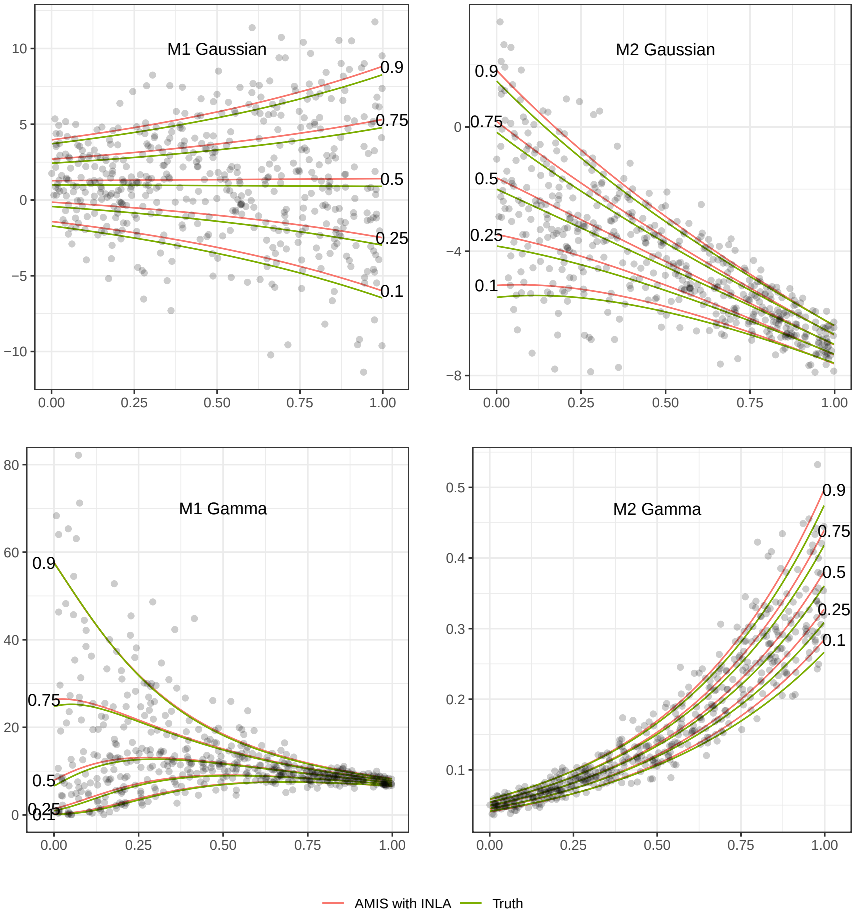

Figure 4.7: Quantile curves of the simulated data sets detailed in Table 4.2 approximated using the AMIS with INLA approach to Bayesian parametric quantile regression and the true quantile curves.

Figure 4.8, and it is apparent that both the mean and dispersion of the observation depends on the observed value of the covariate. In our model, we will use the log ratio of received light as the response y (logratio), and model it by a smooth effect of the distance the light has traveled before it is reflected x (range):

$$\eta_i = a + f(x_i) \tag{4.10}$$

Smooth effects are simple to use in **R-INLA**, and we choose the second order random walk model as prior distribution for the smooth effects $f(\cdot)$ as

$$\pi(\boldsymbol{f} \mid \theta) \propto \theta^{(n-2)/2} \exp\left(-\frac{\theta}{2}\sum_{n=1}^{n-2}(f(x_i) - 2f(x_{i+1}) + f(x_{i+2}))^2\right).$$

Here, $\boldsymbol{f} \mid \theta$ is Gaussian distributed with mean zero and precision given by $\mathbf{Q}(\theta)$ (see Martino et al., 2019, Section 3). The model is completed by assigning a log Gamma prior for $\log(\theta)$ with parameters 1 and $5e-5$.

A Gaussian likelihood is assumed for the model, and its precision is modelled by the parameters c and d as described in (4.6), and the mean $\mu(x)$ is given by (4.10). Like in the other models, we will generate samples of $\boldsymbol{z}_c = (c, d)$ using AMIS, and conditional on these to obtain the conditional marginal likelihood with INLA. After the simulation, an estimate of the posterior mean of $\boldsymbol{z}_c$ is found, and **R-INLA** is used to predict the mean $\mu(x)$. Then, the quantile curves are calculated using (4.8).

Before running the AMIS with INLA algorithm, we have scaled the covariate x by its maximum value to have a higher chance of convergence with the proposal distribution described in the introduction of this section. After the simulation, the range covariate and the parameters associated with it are re-scaled, such that there is no clear effect of the scaling other than the convergence. In addition, we have used the same priors for a, c, d as described in Section 4.3. In this example, we have also applied the MCMC with INLA and IS with INLA algorithms. Sufficient acceptance rates were achieved with the MCMC with INLA algorithm using a bivariate Gaussian proposal distribution with the variance-covariance matrix $2 \cdot \mathbf{I}$. For the IS with INLA algorithm, acceptable results were only obtained when the bivariate Student's t proposal distribution was moved to fit the posterior distribution better, and we specified a mean $\boldsymbol{\mu} = (10, -10)$ and variance-covariance matrix $3 \cdot \mathbf{I}$.

Table 4.4 shows the posterior statistics of the parameters in the random walk model for the LIDAR data. They are approximated by fitting $N = 10000$ conditional models with INLA on the samples drawn with the Monte Carlo methods. Here, we observe very similar posterior estimates between the methods. However, the effective sample size, also presented in Table 4.4, is very small for the IS and MCMC with INLA methods. This is after several reruns and changing of initial parameters. In addition, the posterior marginals shown in Figure A.2 are also pretty poor for the IS with INLA and MCMC with INLA methods. The resulting quantile curves obtained with the AMIS with INLA algorithm are shown in Figure 4.8. We have also included the joint posterior distributions of $\boldsymbol{z}_c$ in Figure A.3, where we observe a significant negative correlation between the parameters.

Method	a	c	d	$\widehat{ESS}$
AMIS w/ INLA	-0.291(0.006)	13.679(0.598)	-0.014(0.001)	6340
IS w/ INLA	-0.291(0.005)	13.656(0.589)	-0.014(0.001)	1770
MCMC w/ INLA	-0.291(0.005)	13.644(0.622)	-0.014(0.001)	144

Table 4.4: Posterior means and standard deviations (in parenthesis) estimated with all INLA within Monte Carlo methods for the coefficients in the random walk model on the LIDAR data.

4.3.3 Serum immunoglobulin G concentrations in children

The dataset used in this example is collected from the research by Isaacs et al. (1983), where parametric quantile regression models are fit to the data. Moreover, it is also used in a Bayesian semi-parametric method in Jara et al. (2011), and in the parametric quantile regression method using the generalized Gamma likelihood by Noufaily et al. (2013). The dataset contains measurements of the serum immunoglobulin G concentrations (IgG; in grams per liter) from 298 children, where their age was also recorded.

Consider the model with response y of immunoglobulin G concentrations and covariate x holding the children's age. The aim is to perform Bayesian quantile regression with the INLA within AMIS algorithm to explain the effect of the age on the quantiles of immunoglobulin levels. Furthermore, we assume that the response is Gamma distributed according to (4.7) and, thereby, the quantile curves can be obtained with the quantile function in (4.9). We also assume that the parameters $\boldsymbol{z}_c = (c, d)$ and $\boldsymbol{z}_{-c} = (a, b)$ have the same prior distributions as described earlier, and we use the same Student's t proposal distribution of three degrees of freedom for the AMIS with INLA algorithm.

We have also attempted to approximate the model using the IS with INLA, and MCMC with INLA approaches. In the former, we choose a bivariate Gaussian proposal distribution with mean $\boldsymbol{\mu} = (2, 0)$ and variance $\boldsymbol{\sigma} = (1, 0.5)$, which is very close to the posterior distribution of $\boldsymbol{z}_c$. The MCMC with INLA algorithm is assigned a bivariate Gaussian with variance $\sigma = (0.1, 0.15)$.

The posterior statistics estimated with the INLA within Monte Carlo methods on the Immunoglobulin G dataset are presented in Table 4.5, and we observe that these estimates are almost identical. However, looking at the effective sample size, the MCMC with INLA algorithm seems to have highly correlated samples. The approximated posterior marginals presented in Figure A.4 is also pretty poor for the MCMC with INLA algorithm. Using the coefficients found by the AMIS with INLA method in Table 4.5, and the quantile function in (4.9), we obtain

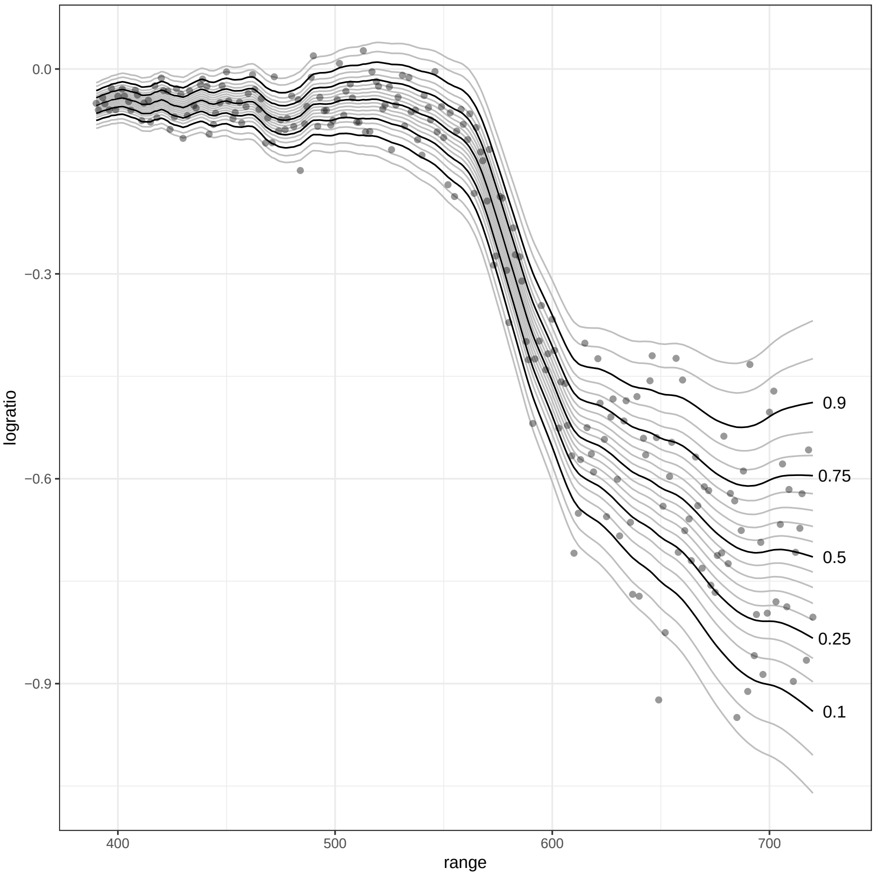

Figure 4.8: Estimated quantile curves of the second order random walk model on the LIDAR dataset obtained using the AMIS with INLA algorithm. The light grey lines are quantile curves in the range $p \in (0.025, 0.975)$.

Method	a	b	c	d	$\widehat{ESS}$
AMIS w/ INLA	1.28(0.05)	0.13(0.01)	1.68(0.15)	0.08(0.04)	6260
IS w/ INLA	1.28(0.05)	0.13(0.01)	1.69(0.15)	0.08(0.04)	8344
MCMC w/ INLA	1.28(0.05)	0.13(0.01)	1.69(0.15)	0.08(0.05)	212

Table 4.5: Posterior statistics estimated with all INLA within Monte Carlo methods for the coefficients in the Gamma model for immunoglobulin G concentrations.

the quantile curves shown in Figure 4.9. We observe that these quantile curves are quite similar to the results presented in Noufaily et al. (2013) and Jara et al. (2011) but diverge for small values of x, where their estimates attain a substantial drop in quantile values. This is most likely caused by the added bias from our assumption that the response follows a Gamma likelihood.

4.4 Gamma Frailty Model

In this example, we will consider a parametric proportional hazard model with Gamma frailty terms (Duchateau et al., 2008) on simulated datasets. The frailty term is used to describe the effect of unobserved covariates in the survival model; for example, it could explain the resilience a specific family has on a disease, i.e. there is dependency within groups (see Gómez-Rubio, 2020).

Consider the data $\boldsymbol{y}$ with only right censored data at time $\boldsymbol{t}$, and the fixed effects β_1 of some covariates $\boldsymbol{x}$. We will model the survival data with a proportional hazard as

$$\begin{aligned} h_{ij}(t) &= \alpha t^{\alpha-1} \exp\left(\beta_0 + x_{ij}^T\beta + w_i\right) \\ &= \alpha t^{\alpha-1} u_i \exp\left(\beta_0 + x_{ij}^T\beta\right), \end{aligned} \tag{4.11}$$

where h_{ij} is the conditional hazard function of subject $j = 1,\ldots,N_i$ and cluster $i = 1,\ldots,M$, and t is its censoring time. The random effect of cluster i, w_i, is the logarithm of u_i, which are the frailty parameters. The linear predictor can in this model be formulate as $\eta_{ij} = \beta_0 + x_{ij}^T\beta_1 + w_i$. In addition, we consider a Gamma frailty u_i with prior distribution

$$\pi(u_i \mid \gamma) = \text{Gamma}(\gamma, \gamma). \tag{4.12}$$

Here, γ is the shape and rate of the Gamma distribution, and is a unknown parameter in the model.

Frailty models can be fit with INLA alone if one assumes that w_i is Gaussian distribution, such that the frailty u_i is log-normal. However, with the Gamma frailty, the linear predictor is not Gaussian unless we fix the frailty terms to some representative values. Therefore, we will employ the AMIS with INLA method to provide posterior inference about the parameters of the model. Samples of $\boldsymbol{z}_c = (\boldsymbol{u}, \gamma)$ are

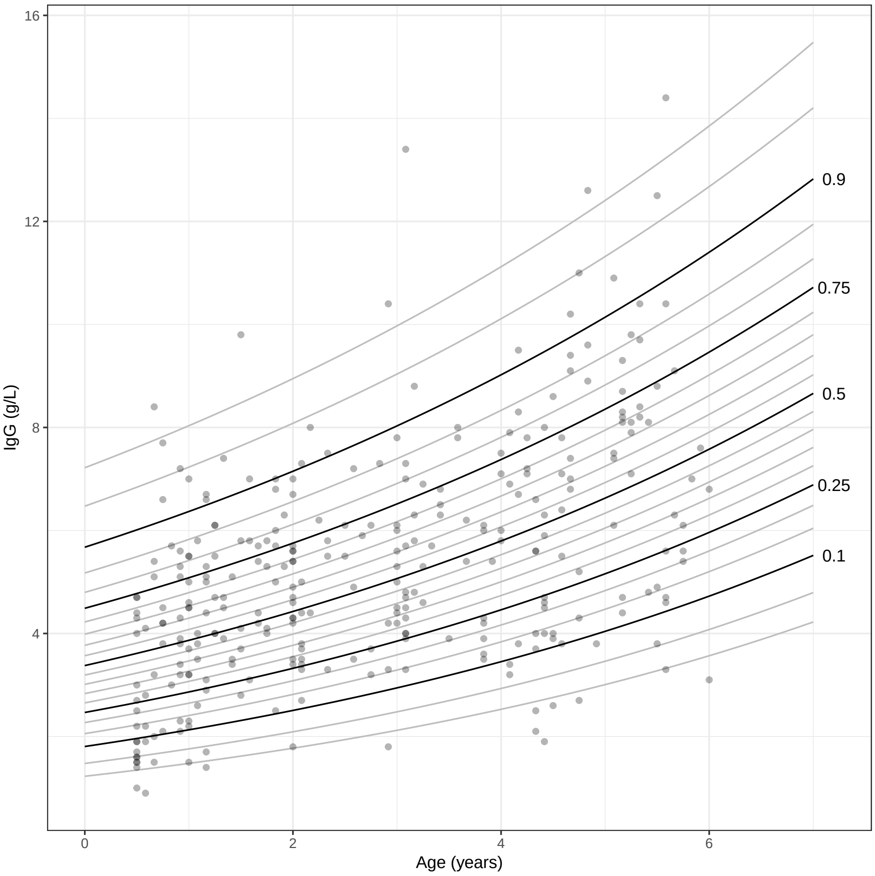

Figure 4.9: Approximated quantile curves of the immunoglobulin G concentration in children modeled by their age. The points are the observations in the dataset. The light grey lines are quantile curves in the range $p \in (0.025, 0.975)$.

generated using AMIS, and the conditional model $\boldsymbol{z}_{-c} | \boldsymbol{z}_{\boldsymbol{c}}$ is fit with **R-INLA**, where $\boldsymbol{z}_{-c} = (\alpha, \beta_0, \beta_1)$. We assume a Gamma prior for γ with parameters 1 and 0.01. The other parameters are assigned their default priors in **R-INLA**; $\log(\alpha)$ a log-Gamma with parameters 25 and 25, β_1 a Gaussian with zero mean and precision 0.001, and β_0 a Gaussian with zero mean and precision.

For the proposal distribution of $\boldsymbol{z}_c$, multivariate log-normal distribution is employed. Note that the dimension of $\boldsymbol{z}_c$ in some model might be very high (100-200+) because of the number of clusters/groups in u_i; therefore, the AMIS algorithm might struggle to find the probability mass of the joint posterior distribution of $\boldsymbol{z}_c$. To account for this, we set the initial parameters of the proposal distribution to the posterior estimates of the frailty from a INLA approximation using its default log-normal prior, which has proven to be quite similar to the Gamma frailty estimates. To clarify, the AMIS weights are in this example calculated as

$$\omega^{(k,t)} = \frac{\pi(\boldsymbol{y} \mid \boldsymbol{u}^{(k,t)}, \gamma^{(k,t)})\pi(\boldsymbol{u}^{(k,t)} \mid \gamma^{(k,t)})\pi(\gamma^{(k,t)})}{\psi_t(\boldsymbol{u}^{(k,t)}, \gamma^{(k,t)})},$$

where k and t denotes the kth sample from the tth epoch, and $\psi_t(\cdot)$ the mixture of all adapted proposal distributions at epoch t. To further account for the high dimensions of $\boldsymbol{z}_c$, we have altered the sample strategy to follow the recommendations by Corneut et al. (2012) described in Section 2.2.3. The sampling strategy is as follows: first, we generate $N_1 = 5000$ samples, and the proposal distribution is adapted; then, $N_t = 500$ samples are generated for $t \in (2, 21)$ epochs, which results in a total of $N = 15000$ samples.

We simulate a dataset of 300 observations from model (4.11) with parameters: $\beta_0 = 1$, $\beta_1 = 2.2$, $\alpha = 1.1$, and $\gamma = 1$. First, the covariate x is drawn from a uniform distribution between zero and one; then, the frailty terms are generated from (4.12) for each cluster i, and the linear predictor η_{ij} for each subject j in cluster i is calculated. Lastly, the censoring times are drawn from a Weibull distribution with shape α and scale $\exp(\eta)^{-1/\alpha}$.

We consider two examples, one with $M = 4$ and one with $M = 20$ clusters. Increasing the number of clusters makes the problem more difficult as the size of the set $\boldsymbol{z}_c$ grows. First, we examine the simple example of $M = 4$, and the resulting posterior estimates of the log frailty terms $\boldsymbol{w}$ are presented in Figure 4.10. In addition, the posterior marginals of the remaining parameters are shown in Figure 4.12. The approximations are relatively accurate for a $\boldsymbol{z}_c$ of dimension five.

Then, we consider a model with $M = 20$ clusters using a similar sampling strategy as $M = 4$. Figure 4.11 shows a 95% confidence interval (CI) for the frailty terms together with their true values. The CI covers the true value even if the estimates appear to be less precise than for the $M = 4$ case. The other parameters are reported in Figure 4.12, where there seems to be a slight bias in the estimation.

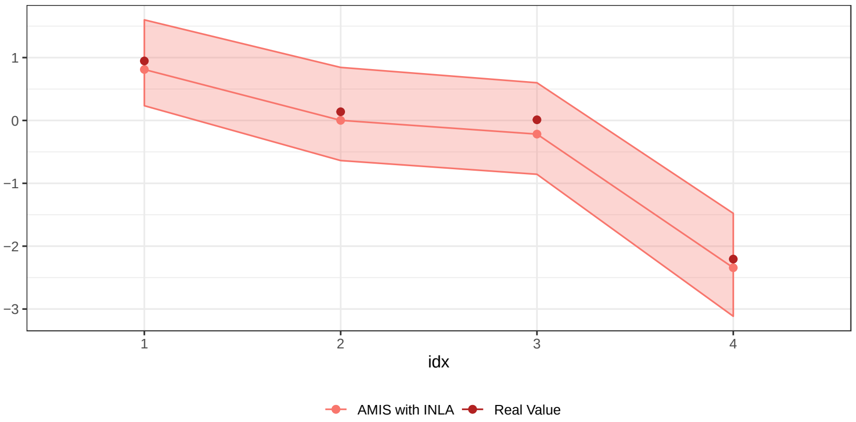

Figure 4.10: The estimated posterior mean of the log frailty w_i using AMIS with INLA on a simulated dataset of $M = 4$ clusters. The bands are the 0.025 and 0.975 quantiles of w_i, where i is the x-axis. The dark red points are the values of w_i in the simulation of data.

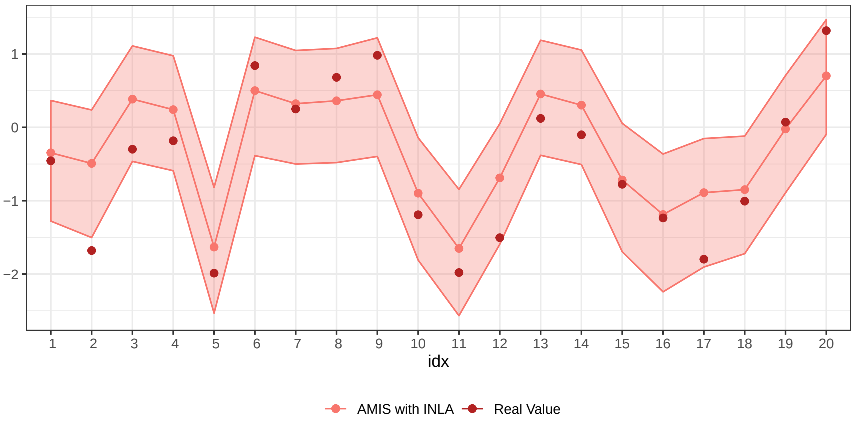

Figure 4.11: The estimated posterior mean of the log frailty w_i using AMIS with INLA on a simulated dataset of $M = 20$ clusters. The bands are the 0.025 and 0.975 quantiles of w_i, where i is the x-axis. The dark red points are the values of w_i in the simulation of data.

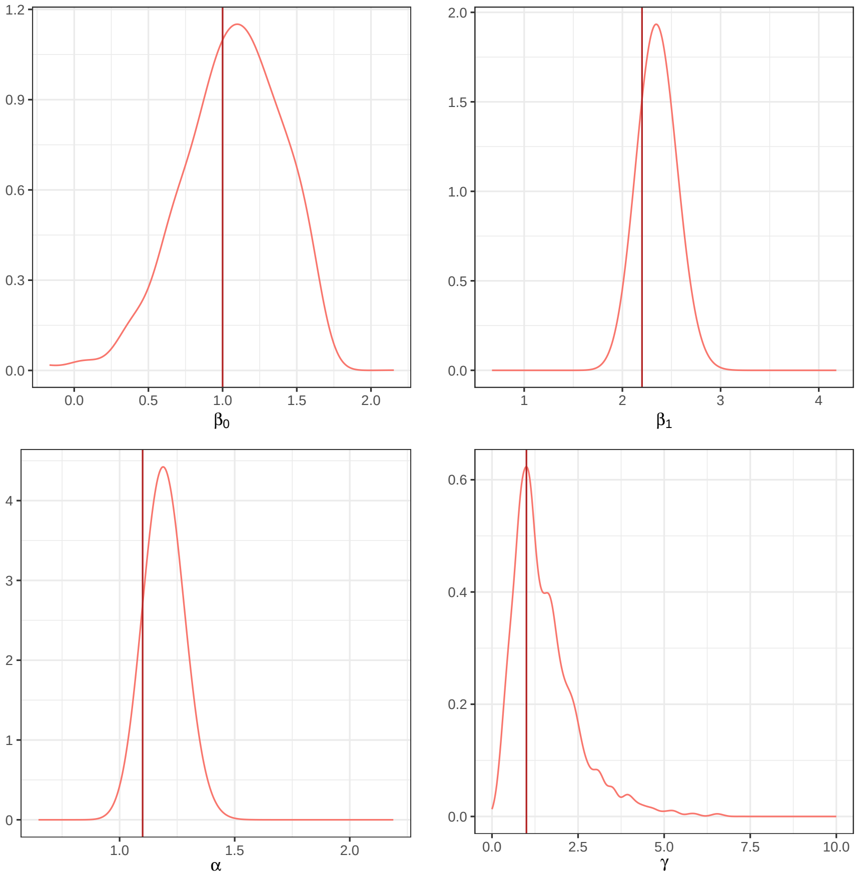

Figure 4.12: The posterior marginals of β_0 (top left), β_1 (top right), α (bottom left), and γ (bottom right) approximated using AMIS with INLA on a simulated dataset of $M = 4$ clusters. The dark red lines are the set values for the parameters in the simulation of data.

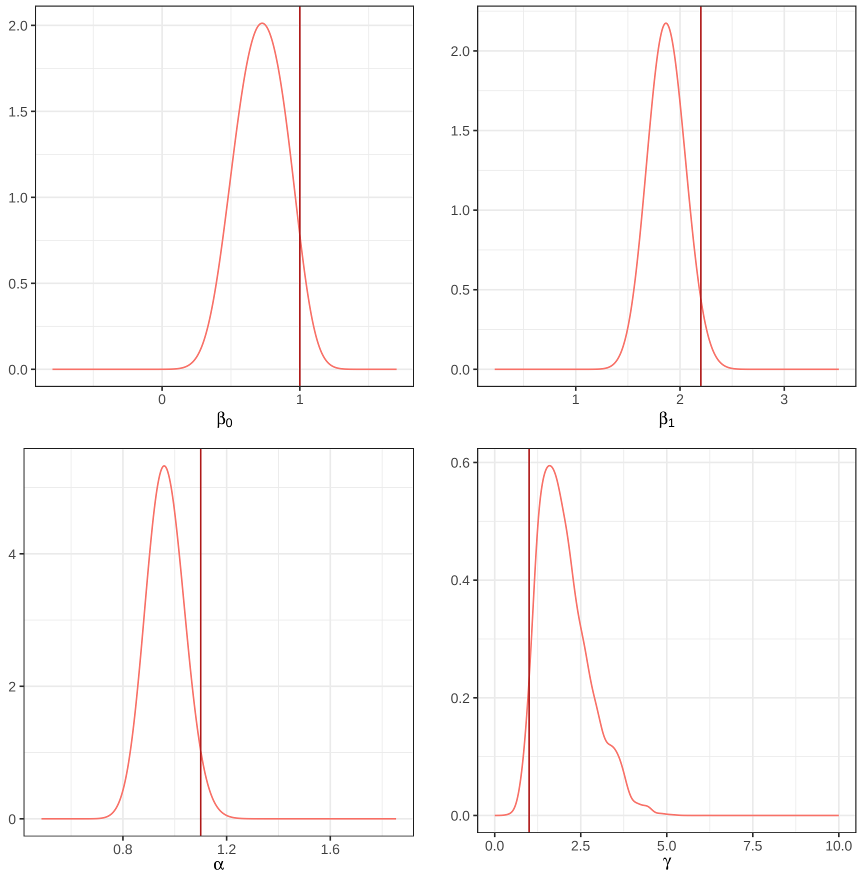

Figure 4.13: The posterior marginals of β_0 (top left), β_1 (top right), α (bottom left), and γ (bottom right) approximated using AMIS with INLA on a simulated dataset of $M = 20$ clusters. The dark red lines are the set values for the parameters in the simulation of data.

Chapter 5

Summary and Discussion

5.1 Summary

In this thesis, we have investigated the properties of three INLA within Monte Carlo methods to extend the class of models that can be fit with INLA; namely MCMC with INLA, IS with INLA, and the novel AMIS with INLA approach. We have built up the theory of Bayesian inference and model assumptions. Then, we presented the considered Monte Carlo methods, and INLA, motivating the merging of Monte Carlo techniques and INLA. Then, we described the general models compatible with these approaches and built the algorithms of the INLA within Monte Carlo methods formulating many posterior estimates.

We have evaluated the performance of these methods on four types of models. First, we validated our implementation of the methods on a bivariate linear model by comparing posteriors with the exact INLA. We also investigated the behavior of the Markov chain in MCMC with INLA and the adaptation in IS with INLA and AMIS with INLA. Next, we tested the performance of the approaches on a spatial autoregressive combined (SAC) model, comparing posteriors to MCMC results. Then, we proposed a novel model-aware Bayesian quantile regression approach using AMIS with INLA, and evaluated it on several simulations before applying it to two existing datasets; second-order random walk model on LIDAR data, and a Gamma model on immunoglobulin G concentration in children. Lastly, we approximated Gamma frailty models in a simulation study using AMIS with INLA for different numbers of clusters.

5.2 Discussion

On the simplest models, the IS with INLA algorithm has proven the most effective method for inference. This is because it can be parallelized indefinitely if one wanted. In our experience, it resembles that of grid exploration with INLA (Gómez-Rubio et al., 2019), since the choice of proposal distribution is very similar to choosing the range of the grid. In Section 4.1, we set good proposals for all methods, and IS with INLA outperforms all of them. For the SAC model, it is also simple to set good proposal distributions because the autoregressive parameters are bound to -1 and 1. However, the **R-INLA** approximation runs much slower for this model, such that adapting the proposal to better approximate the target makes AMIS with INLA the more effective method for the set amount of samples.

In the Bayesian quantile regression Section 4.3, we have purposely focused on the AMIS with INLA algorithm, and chose very vague proposal distributions. In doing so, the IS with INLA algorithm struggled to obtain good approximations unless we drastically altered the proposal. Similarly, the INLA within MCMC method needed many re-runs and tweaking before obtaining acceptable results. Nevertheless, our approach to Bayesian quantile regression shows promise and is accurate when compared to the actual quantile curves in the simulation study, or by just looking at the data. In Section 4.3.3, it is clear that it lacks some of the flexibility of non model-aware Bayesian quantile regression approaches, but it reduces the variance in the extreme quantiles and is fully parametric except for smooth effects in the linear predictor. Noufaily et al. (2013) uses a more flexible likelihood function, the generalized Gamma, which is a three-parameter distribution that includes the Gamma, Weibull, and exponential distribution. We would employ the same likelihood if it were available in **R-INLA**.

The Gamma frailty models tested the AMIS with INLA algorithm to the limits of its capability, where we deliberately increased the number of clusters to make the parameter space of the AMIS very high. On a simple model with four clusters, the method obtained promising posterior estimates when compared to their actual values. However, when the number of clusters increased to 20 or more, the AMIS algorithm struggled to adapt the proposal to fit the target accurately. In this example, we refrained from applying the IS with INLA because of the complexity of the model, and it would struggle to find proper posteriors in these many dimensions. Similarly, we forgo the application of MCMC with INLA algorithm on this example, but acknowledge that it would eventually converge if ran for a very long time.

In our opinion, the AMIS with INLA algorithm is in a practical setting the most robust algorithm, as it requires the least effort to obtain good results compared to the other methods. The posterior statistics are generally very accurate for all methods, but the AMIS with INLA really shines when approximating posterior densities. We recommend the use of IS with INLA on simple models or when a proper proposal distribution is known; however, we could say the same for the BMA with INLA method proposed by Gómez-Rubio et al. (2019). In more complex models, or when it is difficult to find a good proposal, we will encourage the application of AMIS with INLA if computation time is of interest.

We have intentionally not discussed MCMC with INLA that much. It is a computationally intensive and sequential algorithm that, in our experience, takes a too long time to provide accurate results. As observed in the SAC model, if the MCMC with INLA obtained the unfeasible 10000 effective samples, it would still have a lower effective sample size per second than the other methods. Nevertheless, the approach is theoretically sound, and convergence is promised under mild conditions and, in general, we would view MCMC with INLA as a proof of concept for the combined approaches.

Lastly, we want to acknowledge that a standalone MCMC algorithm might run faster than the INLA within Monte Carlo methods, but the development of the algorithm might be a demanding process. The INLA with Monte Carlo framework serves as a simple tool if you have some knowledge of **R-INLA**, to obtain fast inference on models that can "almost" be approximated with INLA alone.

5.3 Further work

For the INLA within Monte Carlo methods to be a simple tool for Bayesian inference, it needs to be easily available. For this, some work is still required on our implementations to create a `R` package or to include it in an existing one. It would also be interesting to investigate the optimization of effective sample size as an adaptation criterion in AMIS with INLA. However, this would require some more theoretical development of the AMIS effective samples estimator (Elvira et al., 2018). Another compelling addition to the AMIS with INLA algorithm is that the decision of whether or not the proposal should be adapted is determined by the effective sample size per sample, as described in Section 3.3. This would improve the AMIS with INLA algorithm on the simple bivariate linear model, and could in general, make the method run faster. However, this requires some investigation of the best threshold of effective sample size per sample.

It would also be interesting to explore other likelihoods in the Bayesian parametric quantile regression approach. There are many likelihoods available in the **R-INLA** package (see `inla.list.models()` in `R` for a extensive list), and many of these are compatible with this approach. Also, there are several random effects easily available in **R-INLA** that can be explored for some interesting quantile regression models.

Further testing of the Gamma frailty model is required to draw any particular conclusion about the applications of AMIS with INLA algorithm. However, it seems unlikely that the approach can handle more than 100 clusters. Future research on this approach could also involve some application on an existing Gamma frailty dataset, and compare it to inference provided by other methods.

Future development of the AMIS with INLA algorithm could involve combining more modern AMIS algorithms with INLA. For example, the *modified* AMIS (MAMIS) proposed by Marin et al. (2014), and the *effective* AMIS (EAMIS) presented by El-Laham et al. (2019). MAMIS proves the convergence of AMIS by modifying the AMIS algorithm slightly, and EAMIS employs an approximate version of the temporal mixture development in AMIS that lowers the computational complexity of the algorithm. We have not investigated the compatibility of these methods with INLA, but they might be an interesting consideration in future research.

Bibliography

Berild, Martin O. (2020). ‘Adaptive Multiple Importance Sampling with the Integrated Nested Laplace Approximation’. Project. Trondheim, Norway: Norwegian University of Science and Technology.

Bernardo, José M. and Adrian F. M. Smith (2000). *Bayesian Theory*. 1. Aufl. Wiley series in probability and mathematical statistics. GB: Wiley, John Wiley & Sons, Incorporated, Wiley-Blackwell.

Bivand, Roger S., Virgilio Gómez-Rubio and Håvard Rue (2014). ‘Approximate Bayesian inference for spatial econometrics models’. In: *Spatial Statistics*. Revealing Intricacies in Spatial and Spatio-Temporal Data: Papers from the Spatial Statistics 2013 Conference 9, pp. 146–165.

Bivand, Roger S., Edzer Pebesma and Virgilio Gómez-Rubio (2013). *Applied Spatial Data Analysis with R*. 2nd ed. 2013. Vol. 10. Use R. New York, NY: Springer New York.

Bivand, Roger, Virgilio Gómez-Rubio and Håvard Rue (2015a). ‘Spatial Data Analysis with R-INLA with Some Extensions’. In: *Journal of Statistical Software* 63.1, pp. 1–31.

Bivand, Roger and Gianfranco Piras (2015b). ‘Comparing Implementations of Estimation Methods for Spatial Econometrics’. en. In: *Journal of Statistical Software* 63.1, pp. 1–36.

Bugallo, Monica F., Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez and Petar M. Djuric (2017). ‘Adaptive Importance Sampling: The past, the present, and the future’. In: *IEEE Signal Processing Magazine* 34.4, pp. 60–79.

Cappé, O., A. Guillin, J. M. Marin and C. P. Robert (2004). ‘Population Monte Carlo’. In: *Journal of Computational and Graphical Statistics* 13.4, pp. 907–929.

Cappé, Olivier, Randal Douc, Arnaud Guillin, Jean-Michel Marin and Christian P. Robert (2008). ‘Adaptive importance sampling in general mixture classes’. In: *Statistics and Computing* 18.4, pp. 447–459.

Corneut, Jean-Marie, Jean-Michel Marin, Antonietta Mira and Christian P. Robert (2012). ‘Adaptive Multiple Importance Sampling’. In: *Scandinavian Journal of Statistics* 39.4, pp. 798–812.

Duchateau, Luc and Paul Janssen (2008). *The Frailty Model*. 1st ed. 2008. Statistics for Biology and Health. New York, NY: Springer New York : Imprint: Springer.

Elvira, Víctor, Luca Martino, David Luengo and Mónica F. Bugallo (2019). 'Generalized Multiple Importance Sampling'. In: *Statistical Science* 34.1, pp. 129–155.

Elvira, Víctor, Luca Martino and Christian P. Robert (2018). 'Rethinking the Effective Sample Size'. In: *arXiv:1809.04129 [stat]*. arXiv: 1809.04129.

George E. P. Box (1987). *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics. Applied probability and statistics. New York: Wiley.

Geweke, John (1989). 'Bayesian Inference in Econometric Models Using Monte Carlo Integration'. In: *Econometrica* 57.6, pp. 1317–1339.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Gómez-Rubio, Virgilio (2019). 'Importance Sampling with the Integrated Nested Laplace Aproximation'. In: *BISP 2019 conference* Madrid, Spain, 12-14 June, Poster.

Gómez-Rubio, Virgilio (2020). *Bayesian inference with INLA*. 1 edition. Chapman and Hall/CRC.

Gómez-Rubio, Virgilio, Roger S. Bivand and Håvard Rue (2019). 'Bayesian model averaging with the integrated nested Laplace approximation'. In: *arXiv:1911.00797 [stat]*. arXiv: 1911.00797.

Gómez-Rubio, Virgilio and Håvard Rue (2018). 'Markov chain Monte Carlo with the Integrated Nested Laplace Approximation'. In: *Statistics and Computing* 28.5, pp. 1033–1051.

Gómez-Rubio, Virgilio and Francisco Palmí-Perales (2019). 'Multivariate posterior inference for spatial models with the integrated nested Laplace approximation'. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.1, pp. 199–215.

Hastings, W. K. (1970). 'Monte Carlo Sampling Methods Using Markov Chains and Their Applications'. In: *Biometrika* 57.1, pp. 97–109.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery and Chris T. Volinsky (1999). 'Bayesian Model Averaging: A Tutorial'. In: *Statistical Science* 14.4, pp. 382–401.

Hubin, Aliaksandr and Geir Storvik (2016). 'Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA)'. In: *arXiv:1611.01450 [stat]*. arXiv: 1611.01450.

Isaacs, D., D. G. Altman, C. E. Tidmarsh, H. B. Valman and A. D. Webster (1983). 'Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for IgG, IgA, IgM.' In: *Journal of Clinical Pathology; London* 36.10, p. 1193.

Jara, A. and Timothy E. Hanson (2011). 'A class of mixtures of dependent tail-free processes'. In: *Biometrika* 98.3, pp. 553–566.

Kahn, H. (1950). 'Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems. II'. English. In: *Nucleonics (U.S.) Ceased publication* Vol: 6, No. 6.

Koenker, Roger and Gilbert Bassett (1978). 'Regression Quantiles'. In: *Econometrica* 46.1, pp. 33–50.

El-Laham, Yousef, Luca Martino, Victor Elvira and Mónica Bugallo (2019). 'Efficient Adaptive Multiple Importance Sampling'. In: pp. 1–5.

LeSage, James and Robert Kelley Pace (2009). *Introduction to Spatial Econometrics*. 1 edition. Boca Raton: Chapman and Hall/CRC.

Li, Ye, Patrick Brown, Håvard Rue, Mustafa al-Maini and Paul Fortin (2012). 'Spatial modelling of lupus incidence over 40 years with changes in census areas'. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.1, pp. 99–115.

Manski, Charles F. (1993). 'Identification of Endogenous Social Effects: The Reflection Problem'. In: *The Review of Economic Studies* 60.3, pp. 531–542.

Marin, Jean-Michel, Pierre Pudlo and Mohammed Sedki (2014). 'Consistency of the Adaptive Multiple Importance Sampling'. In: *arXiv:1211.2548 [math, stat]*. arXiv: 1211.2548.

Martino, Luca, Victor Elvira and Francisco Louzada (2017). 'Effective Sample Size for Importance Sampling based on discrepancy measures'. In: *Signal Processing* 131. arXiv: 1602.03572, pp. 386–401.

Martino, Sara and Andrea Riebler (2019). 'Integrated Nested Laplace Approximations (INLA)'. In: *arXiv:1907.01248 [stat]*. arXiv: 1907.01248.

Martins, Thiago G. and Håvard Rue (2013a). 'Extending INLA to a class of near-Gaussian latent models'. In: *arXiv:1210.1434 [stat]*. arXiv: 1210.1434.

Martins, Thiago G., Daniel Simpson, Finn Lindgren and Håvard Rue (2013b). 'Bayesian computing with INLA: New features'. en. In: *Computational Statistics & Data Analysis* 67, pp. 68–83.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller and Edward Teller (1953). 'Equation of State Calculations by Fast Computing Machines'. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Michael D. Ward (2008). *Spatial regression models*. Vol. 07-155. Quantitative applications in the social sciences. Los Angeles, Calif: Sage.

Noufaily, Angela and M. C. Jones (2013). 'Parametric quantile regression based on the generalized gamma distribution'. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 62.5, pp. 723–740.

Owen, Art and Yi Zhou (2000). 'Safe and effective importance sampling'. In: *Journal of the American Statistical Association; Alexandria* 95.449, pp. 135–143.

Padellini, Tullia and Håvard Rue (2019). 'Model-aware Quantile Regression for Discrete Data'. In: *arXiv:1804.03714 [stat]*. arXiv: 1804.03714.

Robert, Christian P. and George Casella (2004). *Monte Carlo statistical methods*. 2nd ed. Springer texts in statistics. New York: Springer.

Rodrigues, T. and Y. Fan (2017). 'Regression Adjustment for Noncrossing Bayesian Quantile Regression'. In: *Journal of Computational and Graphical Statistics* 26.2, pp. 275–284.

Rue, H., A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson and F.K. Lindgren (2017). 'Bayesian computing with INLA: A review'. In: *Annual Review of Statistics and Its Application* 4, pp. 395–421.

Rue, Håvard (2020). *The R-INLA project*. URL: http://www.r-inla.org/ (visited on 15/01/2020).

Rue, Håvard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. Vol. 104. Monographs on statistics and applied probability. Boca Raton, Fla: Chapman & Hall/CRC.

Rue, Håvard and Sara Martino (2007). 'Approximate Bayesian inference for hierarchical Gaussian Markov random field models'. en. In: *Journal of Statistical Planning and Inference*. Special Issue: Bayesian Inference for Stochastic Processes 137.10, pp. 3177–3192.

Rue, Håvard, Sara Martino and Nicolas Chopin (2009). 'Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations'. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71.2, pp. 319–392.

Sigrist, Markus W. and James D. Winefordner (1994). *Air Monitoring by Spectroscopic Techniques*. Vol. 197. Chemical Analysis Series. New York: Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London: Chapman and Hall.

Tierney, Luke and Joseph B. Kadane (1986). 'Accurate Approximations for Posterior Moments and Marginal Densities'. In: *Journal of the American Statistical Association* 81.393, pp. 82–86.

Yu, Keming and Rana A. Moyeed (2001). 'Bayesian quantile regression'. In: *Statistics & Probability Letters* 54.4, pp. 437–447.

Appendix A

Additional Figures

A.1 SAC model

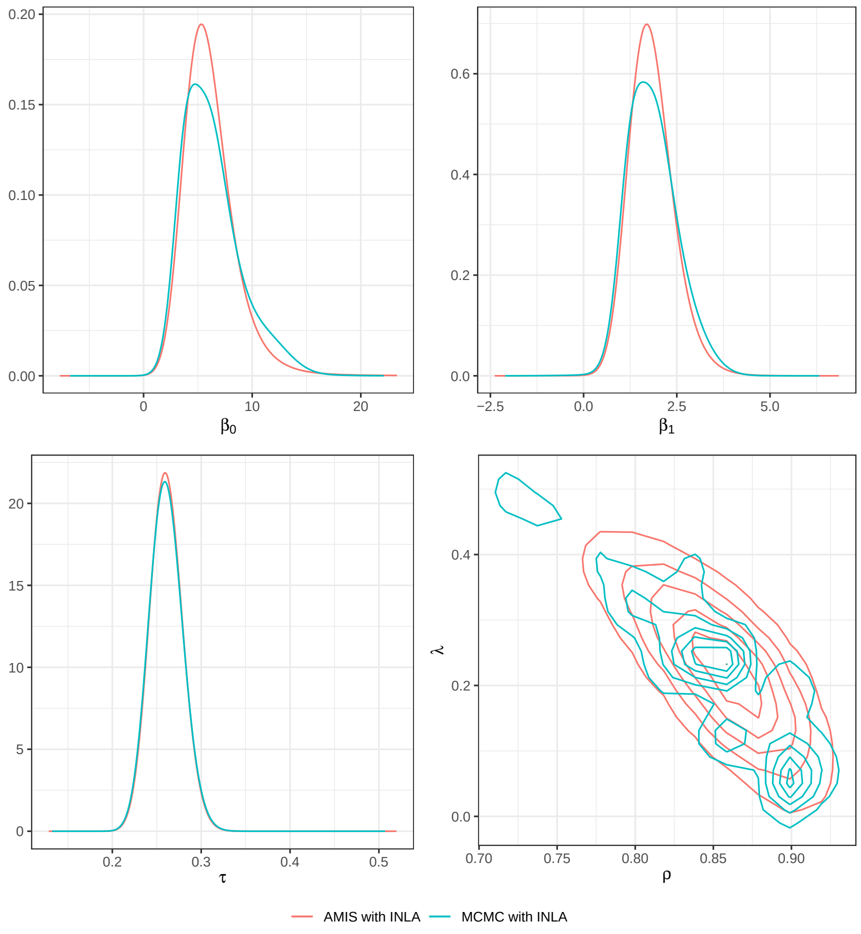

Figure A.1: Joint posterior distribution of the autoregressive terms ρ and λ (bottom right) and the posterior marginals of the intercept β_0, log GDP per capita β_1, and the precision of the noise τ in the SAC model approximated with AMIS with INLA and MCMC with INLA.

A.2 Ratio of received light in LIDAR measurements

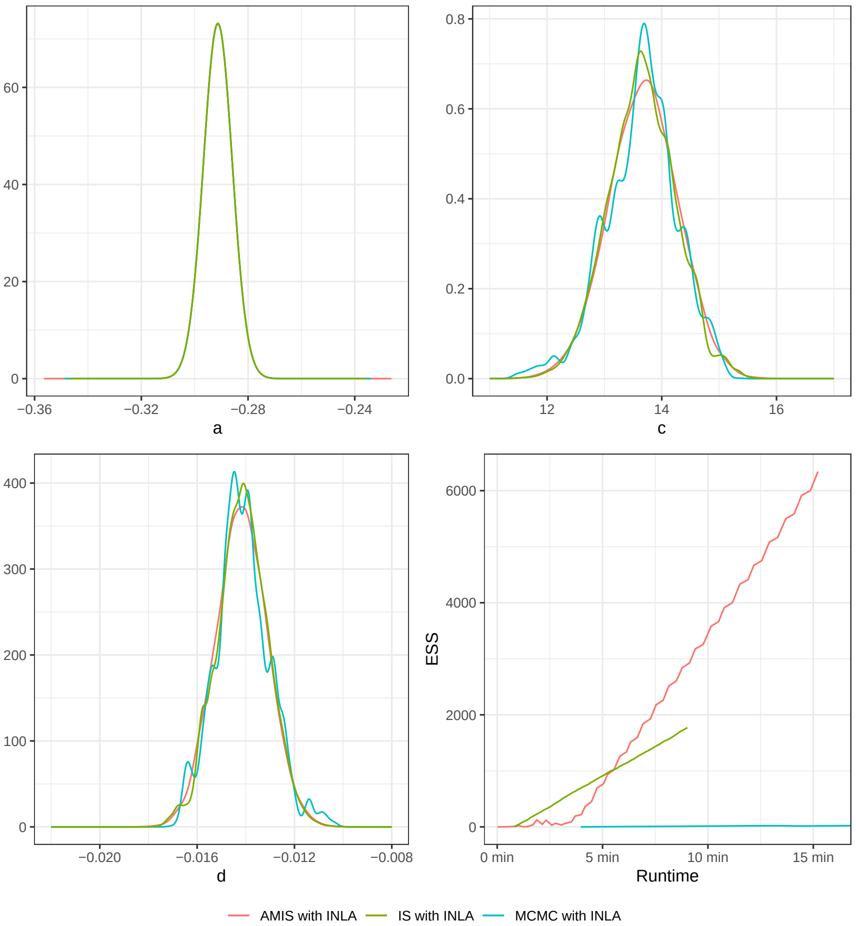

Figure A.2: Posterior marginals of $\boldsymbol{z} = (a, c, b)$, and the running effective sample size (bottom right) in the second order random walk model for LIDAR data approximated with AMIS with INLA, IS with INLA, and MCMC with INLA.

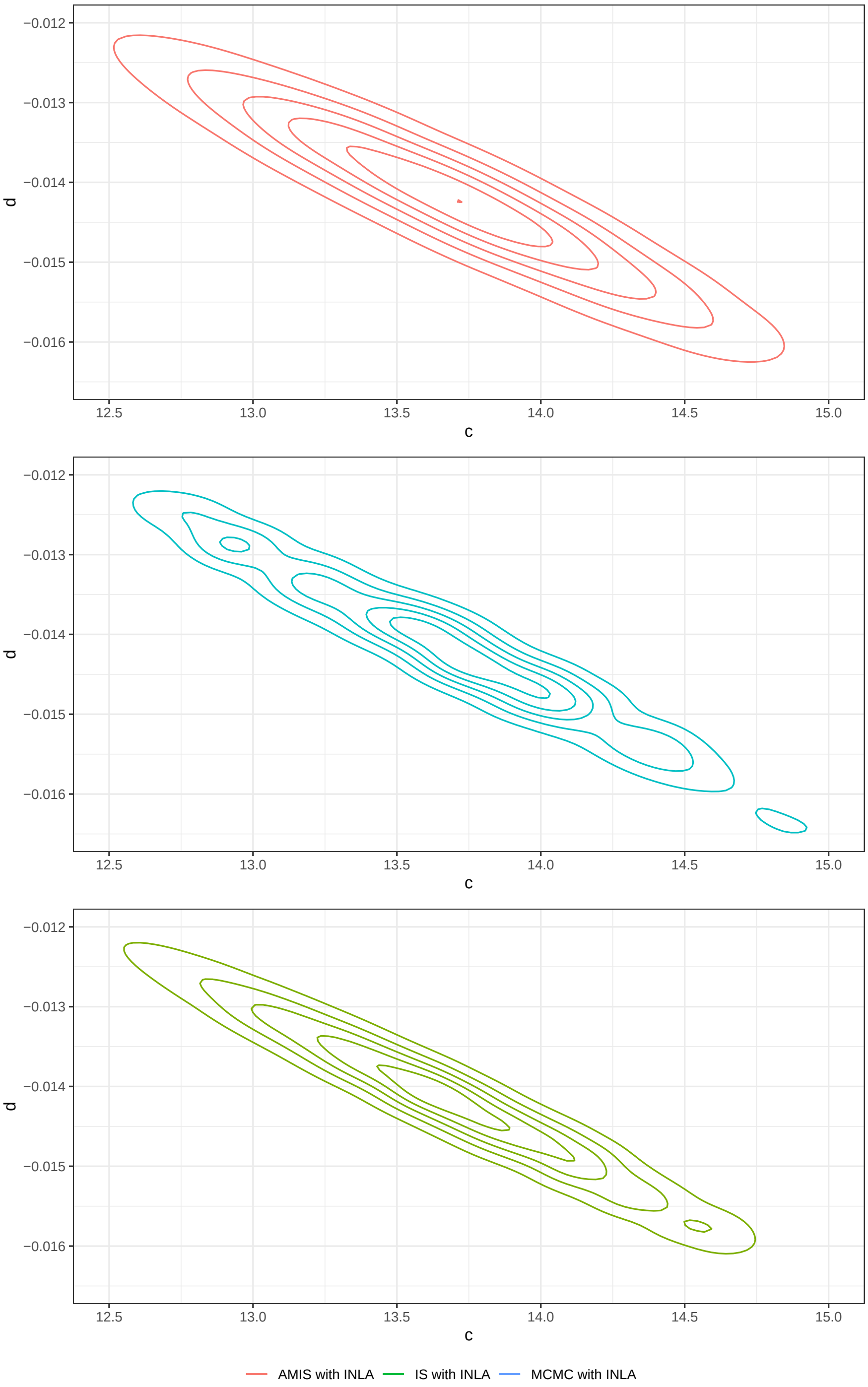

Figure A.3: Joint posterior distribution of $\boldsymbol{z}_c = (c, b)$ in the random walk model for LIDAR data approximated with AMIS with INLA, IS with INLA, and MCMC with INLA.

A.3 Serum immunoglobulin G concentrations in children

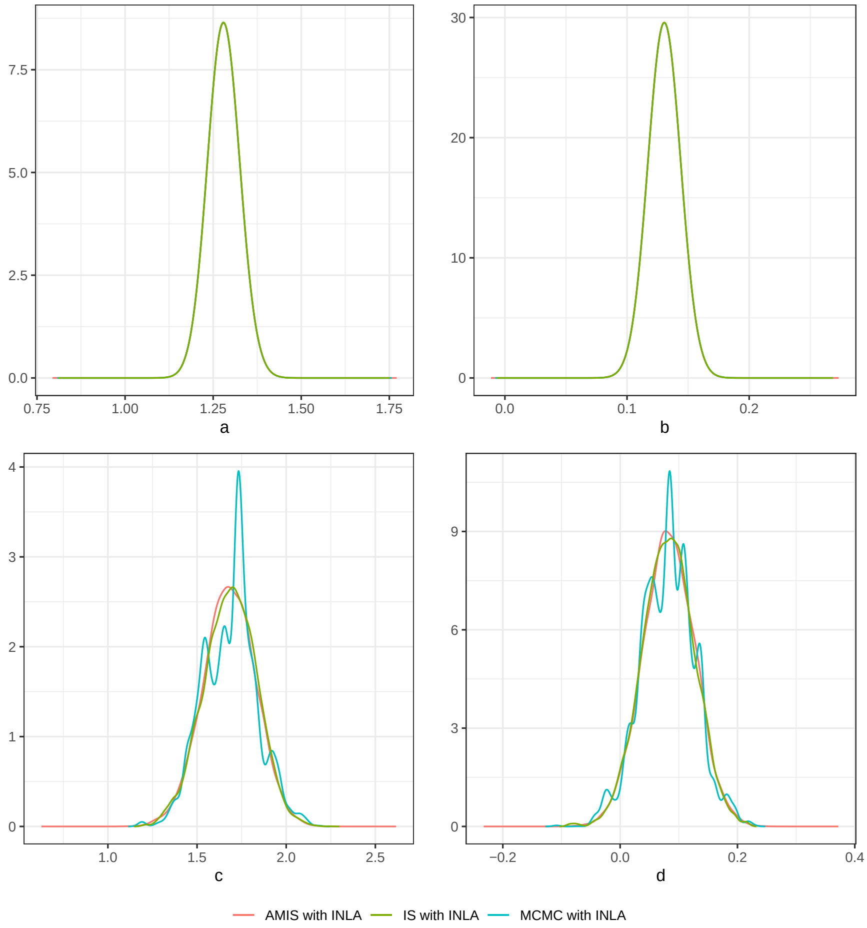

Figure A.4: Posterior marginals of $\mathbf{z} = (a, b, c, b)$ in the gamma model for Immunoglobulin G data approximated with AMIS with INLA, IS with INLA, and MCMC with INLA.

NTNU

Norwegian University of
Science and Technology