INLA for GMRFs (e.g. GLMMs, Spatial Models) with Spatial Examples of Leukemia Cases and Heavy Metal Concentrations

Frances Lin

June 2022

1. Introduction

1.1 Background and Introduction

The steps involving the Bayesian inference may appear easy and straightforward: updating prior beliefs about the unknown parameters with observed data and obtaining the posterior distribution for the parameters. However, this is much harder to do in practice since solutions in closed-form may not always be found.

The simulation-based inference through the idea of MCMC (Markov chain Monte Carlo) was introduced and represented a breakthrough in Bayesian inference (Robert & Casella, 1999) in the early 1990s. MCMC tools such as `WinBugs` (Spiegelhalter et al., 1995), `JAGS` (Plummer, 2016), and `stan` (Stan Development Team, 2015) have also been developed. Bayesian statistics has quickly gained popularity in many fields. However, while MCMC are asymptotically exact methods, based on sampling, they can not only be computationally demanding (i.e. requires a large amount of CPU), but also present convergence issues.

INLA (integrated nested Laplace approximation) is a fast alternative to MCMC for Bayesian inference for a specific class of models named LGMs (latent Gaussian models). INLA does not require sampling and is a faster but approximate method. INLA uses a nested version ("nested") of the Laplace approximation, combined with modern numerical techniques for integration ("integrated"). INLA can be applied to a very wide and flexible class of models named LGMs (latent Gaussian models), which ranges from GLMMs (generalized linear mixed models), GAMMs (generalized additive mixed models) to time-series, and spatial and spatio-temporal models. INLA also allows for faster and more accurate inference without trading speed for accuracy, and it is accessible through the **R** package `R-INLA` (Rue et al., 2017).

1.2 Applications

INLA have found applications in a wide variety of fields. In particular, INLA have found spatial or spatio-temporal applications in fields such as environment, ecology, disease mapping, medical imaging, public health, cancer research, energy, economics, risk analysis, etc.

Some selected examples include: environmental risk factors to liver fluke in cattle (Innocent et al., 2017); modelling recovering fish populations (Boudreau et al., 2017); polio-virus eradication in Pakistan (Mercer et al., 2017); cortical surface fMRI data (Mejia et al., 2017); socio-demographic and geographic impact of

HPV vaccination (Rutten et al., 2017); topsoil metals and cancer mortality (Lopez-Abente et al., 2017) with spatially misaligned data; ethanol and gasoline pricing (Laurini, 2017); applications in spatial econometrics (Bivand et al., 2014; Gomez-Rubio et al., 2015; Gomez-Rubio et al., 2014); probabilistic prediction of wind power (Lenzi et al., 2017); modeling landslides as point processes (Lombardo et al., 2018); predicting extreme rainfall events in space and time (Opitz et al., 2018), etc.

Section 2 contains key components that make up INLA. Section 3 introduces INLA. Section 4 includes discussion. Two spatial examples using the `R-INLA` package are included in Appendix.

2. Key Components

2.0. Bayesian Inference

The posterior distribution is proportional to the likelihood function multiples by the prior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta),$$

where $p(y|\theta)$ is the likelihood function, $p(\theta)$ is the prior, and $\int p(y|\theta)p(\theta)d\theta$ is the normalizing constant.

Later, we will see that the posterior distribution, structured in a hierarchical way, becomes

$$\begin{aligned} p(x, \theta|y) &\propto p(y|x, \theta)p(x, \theta) \\ &\propto p(y|x, \theta)p(x|\theta)p(\theta). \end{aligned}$$

Based on the posterior distribution, relevant statistics for the parameters of interest such as marginal distribution, means, variances, quantiles, credibility intervals, etc. can be obtained.

However, the integral is generally intractable in closed-form, thus requiring the use of numerical methods such as MCMC.

2.1. Latent Gaussian Models

The latent Gaussian models (LGMs) is a class of three-stage Bayesian hierarchical models where $p(x|\theta)$ is a Gaussian distribution. Applications of LGMs include, for example, regression models (e.g. GAMs/GLMs/GAMMs/GLMMs/++), dynamic models, and spatial (e.g. Gaussian and BYM models) and spatio-temporal models (Rue et.al, 2009). A LGM involves the following stages:

1. In the first stage, observations (or data) y is assumed to be conditionally independent, given a latent Gaussian random field x (joint distribution of all parameters in the linear predictor) and hyperparameter θ_1

$$y|x, \theta_1 \sim \prod_{i \in I} p(y_i|x_i, \theta_1). \qquad likelihood$$

2. In the second stage, the latent field $x|\theta_2$ is assumed to be a GMRF (Gaussian Markov random field) with a sparse precision matrix Q

$$x|\theta_2 \sim p(x|\theta_2) = N(\mu(\theta_2), Q^{-1}(\theta_2)), \qquad _{latent\ field}$$

where $Q = \Sigma^{-1}$ is the precision matrix and θ_2 is a hyperparameter. The versatility of the model class can be specified through the latent field, which can be non-linear but is unobserved. The latent field includes all random terms and captures the underlying dependence structure of the data. Latent field can be, for example, covariates, unstructured random effects (e.g. white noise), or structured random effects (e.g. temporal dependency, spatial dependency, smoothness terms; Bolin, 2015).

3. In the last stage, the hyperparameters of the latent field that are not necessarily Gaussian are assumed to follow a prior distribution

$$\theta = (\theta_1, \theta_2) \sim p(\theta), \qquad _{hyperpriors}$$

where $p(\cdot)$ is a known distribution. The hyperparameters control the likelihood for the data and/or the latent Gaussian field and can be used to account for variability and strength of dependence. Hyperparameters can be, for example, variance of observation noise, variance of the unstructured random field, range of a structured random effect, or autocorrelation parameter (Bolin, 2015).

Inference in hierarchical models is based on the posterior distribution, and the joint posterior distribution, structured in a hierarchical way, becomes

$$p(x, \theta|y) \propto p(\theta)p(x|\theta)\prod_{i\in I} p(y_i|x_i, \theta)$$

$$\propto p(\theta)|Q(\theta)|^{1/2}\exp(-\frac{1}{2}x^T Q(\theta)x + \sum_i \log p(y_i|x_i, \theta)).$$

For computational reasons and to ensure accurate approximations, the following assumptions hold:

1. Each observation y_i depends only on one component of the latent field x_i, and most components of x will not be observed.
2. The distribution of the latent field x is Gaussian and is close to a Gaussian Markov random field (GMRF) when the dimension of n is high (10^3 to 10^5).
3. The number of hyperparameters θ is small ($\sim$ 2 to 5 but < 20).

2.2. Additive Models

LGMs (latent Gaussian models) is an umbrella class that generalizes the large number of related variants of additive and/or generalized linear models.

Consider the Bayesian structured additive model setup, for example,

$$y \sim \prod_i^N p(y_i|x_i, \theta),$$

then the mean μ_i (for observation y_i) can be linked to the linear predictor η_i through a link function g

$$\eta_i = g(\mu_i) = \alpha + \sum_j \beta_j z_{ji} + \sum_k f_k(w_{ki}) + \varepsilon_i,$$

where α is the overall intercept, β are linear effects of fixed covariates z, $\{f_k\}$, which are used to represent specific Gaussian processes, are nonlinear/smooth effects of covariates w, and ε are iid random effects. The model components f_k are what make LGMs flexible. Examples of f_k include spatially or temporally correlated effects, smoothing and stochastic spline, measure errors, and random effects with different types of correlations. GLMs (generalized linear models) is a special case with the expression $\alpha+\sum_j \beta_j z_j$ (i.e. $f(\cdot) = 0$). GAMs (generalized additive models) is another special case with the expression $\alpha + \sum_k f_k(w_k)$.

The model formulation in the above model and LGM relate to the same class of models iff the joint distribution of the latent field x is jointly Gaussian. I.e.

$$x = (\eta, \alpha, \beta, f(\cdot)) \sim N(0, \Sigma = Q^{-1}).$$

This can be achieved by assigning Gaussian priors to all terms (the intercept and the parameter of the fixed effects) in x. If we further assume conditional independence of x, then this latent field x given θ is a Gaussian Markov random field with the distribution

$$p(x|\theta) \propto |Q(\theta)|^{1/2} \exp(-\frac{1}{2} x^T Q(\theta) x),$$

where $Q(\theta)$ is the sparse precision matrix and $|Q(\theta)|$ is its determinant.

2.3. Gaussian Markov Random Fields

GMRFs (Gaussian Markov Random Fields) provide computational benefit since calculations involving a sparse matrix are much less costly. A GMRF is a random vector that follows a multivariate normal distribution with additional conditional independence properties: for $i \neq j$, x_i and x_j are conditionally independent given the remaining elements x_{-ij}, for several $\{i, j\}$s.

Undirected graphs G are typically used to represent the conditional independence properties of the GMRF. An undirected graph $G = (V, E)$ consists of a set of nodes V and edges E, where V is a set of nodes $\{1, ..., n\}$ and E is a set of edges $\{i, j\}$, where $i \neq j \in V$ (Bolin, 2015).

Formally, a random vector $x = (x_1, ..., x_n)^T \in \mathbb{R}$ is called a GMRF with respect to a labelled graph $G = (V, E)$ with mean μ and positive definite matrix Q iff its density has the form

$$p(x) = (2\pi)^{-n/2} |Q|^{1/2} \exp(-\frac{1}{2}(x - \mu)^T Q(x - \mu))$$

and

$$Q_{ij} \neq 0 \iff \{i, j\} \in E \quad \forall i \neq j.$$

Let x be a GMRF with respect to a graph $G = (V, E)$, then it is equivalent to say that x_i and x_j are

conditionally independent, given the remaining elements x_{-ij}

$$x_i \perp x_j | x_{-ij} \quad if \;\; i,j \in E, i \neq j,$$

where $-ij$ refers to all elements other than i and j. This is referred to as the pairwise Markov property. Equivalent properties are the local Markov property and global Markov property.

The Markov assumption in the GMRFs results in a sparse precision matrix. When a matrix is sparse (with lots of elements = 0), the computational cost tends also to be lower, allowing for much faster computation. Recall that $x \sim N(0, Q = \Sigma^{-1})$ and

$$x_i \perp x_j \iff \Sigma_{ij} = 0,$$

where Σ is the covariance matrix. For Σ to be sparse requires the marginal independence assumption, which can be unreasonable. On the other hand, it can be shown that

$$x_i \perp x_j | x_{-ij} \iff Q_{ij} = 0,$$

where Q is the precision matrix (the inverse of the covariance matrix), and the conditional independence is a more reasonable assumption and their properties are encoded in the precision matrix (Rue & Held, 2005).

Although the exact computational cost depends on the actual sparsity pattern in the precision matrix, for the autoregressive (temporal) example, the mxm sparse precision matrix can be factorized from $\mathcal{O}(m^3)$ to $\mathcal{O}(m)$ with memory requirement reduced from $\mathcal{O}(m^2)$ to $\mathcal{O}(m)$. For models with a spatial structure, the cost is in $\mathcal{O}(m^{3/2})$ paired with a $\mathcal{O}(m \log(m))$ memory requirement, and for models with a spatio-temporal structure, the cost is in $\mathcal{O}(m^2)$.

2.4. Additive Models and Gaussian Markov Random Fields

One of the primary reasons why INLA approach is so efficient is that it is able to treat the joint distribution for the latent field x as a GMRF with a precision matrix that is easy to compute. The sparsity of the precision matrix also boosts computational efficiency, compared with operations on dense matrices.

2.5. Laplace Approximations

The Laplace approximation is an old technique for the approximation of integrals (Barndorff-Nielsen & Cox, 1989). The idea is to approximate the Gaussian target by matching the mode and curvature at the mode, and the mode is computed iteratively using a Newton-Raphson method (also known as the scoring algorithm or its variant, the Fisher scoring algorithm, Rue & Martino, 2009).

Let $nf(x)$ be the sum of log-likelihoods and x the unknown parameter, the goal is to approximate the integral

$$I_n = \int_x \exp(n(f(x)))dx$$

as $n \to \infty$. Let x_0 be the point in which $f(x)$ has its maximum, then

$$I_n \approx \int_x \exp(n(f(x_0) + \frac{1}{2}(x - x_0)^2 f^{''}(x_0)))dx$$

$$= \exp(nf(x_0))\sqrt{(}\frac{2\pi}{-nf''(x_0)}) = \tilde{I}_n.$$

By the central limit theorem, the Gaussian approximation will be exact as $n \to \infty$. The extension to higher-dimensional integrals is also immediate with the errors given as

$$I_n = \tilde{I}_n(1 + \mathcal{O}(n^{-1})).$$

This provides two advantages. The error rate is relative with rate $\mathcal{O}(n^{-1})$, as supposed to additive with rate $\mathcal{O}(n^{-1/2})$, which is common in simulation-based inference. It was also once a key tool for high-dimensional integration pre-MCMC times.

3. INLA

3.1. INLA

The INLA approach is designed specifically for the structure of LGMs, where (1) the likelihood is conditional independent (i.e., y_i only depends on one x_i and θ), (2) $x|\theta$ is a GMRF, and (3) $|\theta|$ is low-dimensional. For LGMs, the problem can be reformulated as series of subproblems that allows the use of Laplace approximations.

The exact joint posterior distribution of x and θ in Section 2.1. is generally difficult to obtain. The main goal of Bayesian inference is to approximate the posterior marginals for the hyperparameters and latent field respectively

$$p(\theta_j|y), j = 1, ..., |\theta|, \qquad p(x_i|y),\ i = 1, ..., n.$$

The posterior marginals of interest can be written as

$$p(\theta_j|y) = \int p(\theta|y)d\theta_{-j},$$

$$p(x_i|y) = \int p(x_i, \theta|y)d\theta = \int p(x_i|\theta, y)p(\theta|y)d\theta,$$

where $\theta_{-j} = (\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ...)$. The key feature of the INLA approach is to use the above form to construct nested approximations

$$\tilde{p}(\theta_j|y) = \int \tilde{p}(\theta|y)d\theta_{-j},$$

$$\tilde{p}(x_i|y) = \int \tilde{p}(x_i|\theta, y)\tilde{p}(\theta|y)d\theta,$$

where $\tilde{p}(\cdot|\cdot)$ is an approximated conditional density of its arguments (Rue et al., 2009). Approximation to $p(\theta_j|y)$ is computed by integrating out θ_{-j} from $\tilde{p}(\theta_j|y)$. Approximations to $p(x_i|y)$ are computed by approximating $p(x_i|\theta, y)$ and $p(\theta|y)$ and using numerical integration to integrate out θ (Rue et al., 2009).

3.1.1. Approximating the Posterior Marginals for the Hyperparameters

Applying Bayes' theorem, the conditional probability of θ given y can be rewritten as

$$p(\theta|y) = \frac{p(x, \theta|y)}{p(x|\theta, y)}.$$

Expanding the numerator and replacing the denominator with a Laplace approximation, then the above expression becomes

$$p(\theta|y) \propto \frac{p(\theta)p(x|\theta)p(y|x, \theta)}{\tilde{p}(x|\theta, y)}\bigg|_{x=x^*(\theta)} = \tilde{p}(\theta|y),$$

where $\tilde{p}(x|\theta, y)$ is the Gaussian approximation to the full conditional of x and $x^*(\theta)$ is the mode of x for a given θ (Rue et al., 2009, as cited in Morrison, 2017).

The denominator is approximated using a Gaussian approximation

$$p(x|\theta, y) \propto \exp(-\frac{1}{2}x^T Q(\theta)x + \sum_i \log(p(y_i|x_i, \theta)))$$

$$= (2\pi)^{-n/2}|P(\theta)|^{1/2} \exp(-\frac{1}{2}(x - \mu(\theta))^T P(\theta)(x - \mu(\theta))) = \tilde{p}(x|\theta, y), \quad (*)$$

where $P(\theta) = Q(\theta) + diag(c(\theta))$, $\mu(\theta)$ is the location of the mode, and $c(\theta)$ is the vector that contains the negative second derivative e of the log-likelihood at the mode.

To approximate $p(\theta_j|y)$, the following steps are involved (Rue et al., 2009):

1. locate the mode of $\tilde{p}(\theta|y)$ by optimizing $\log(\tilde{p}(\theta|y))$ with respect to θ using some quasi-Newton method and let θ^* be the modal configuration.
2. at the modal configuration θ^*, compute the negative Hessian matrix $H > 0$ using finite differences.
3. explore $\log(\tilde{p}(\theta|y))$ to locate the bulk of the probability mass using the z-parameterization.
4. approximate $p(\theta_j|y)$ by using the points that were already computed during steps 1-3 to construct an interpolant to $\log(\tilde{p}(\theta|y))$ and compute marginals using numerical integration such as Newton–Raphson method from this $\log(\tilde{p}(\theta|y))$.

Since $|\theta|$ is of low dimension, marginals for $\theta_j|y$ can be derived directly from the approximation to $\theta|y$.

The challenge lies in finding a quick and reliable approach while keeping the number of evaluation points low. Interested readers are referred to Martins et al.'s (2013, section 3.2) paper for details about the default approach that is used now.

3.1.2. Approximating the Posterior Marginals for the Latent Field

Approximate the posterior marginals for the latent field is similar but more challenging since the dimension of x can be very large.

The posterior marginals for the latent field can be expressed as

$$p(x_i|\theta, y) = \int p(x_i|\theta, y)p(\theta|y)d\theta,$$

which results in two challenges:

1. Integrating over $p(\theta|y)$ is shown to be too computationally costly in Section 3.1.1. since the cost of standard numerical integration is exponential in the dimension of θ.
2. Approximating $p(x_i|\theta, y)$ for a subset of all $i = 1, ..., n$ using the Laplace approximation can be too demanding since n can be very large (10^3 to 10^5).

For the first challenge, classical numerical integration is restricted to lower dimensions because higher-dimensional integrals can not only be difficult but also impossible. To avoid the integration step, empirical Bayes approach is used, which uses the mode. In dimensions > 2, ideas were borrowed from central composite design (Box & Wilson, 1951), which uses integration points on a sphere around the center.

For the second challenge, three approximation options are available in the `R-INLA` package: the Gaussian, the Laplace, and the simplified Laplace approximation (Rue et al., 2009, as cited in Morrison, 2017).

The default approach is to compute a Taylor's expansion up to the third order around the mode of the Laplace approximation, which provides an approximation to the standardized Gaussian approximation and appears to be highly accurate

$$\log p(x_i|\theta, y) \approx b_i(\theta)x_i - \frac{1}{2}{x_i}^2 + \frac{1}{6}c_i(\theta){x_i}^3$$

The Gaussian option is fast but the assumption is strong so results tend to be poor and the Laplace option works well but is computationally more expensive (Rue et al., 2009, as cited in Morrison, 2017).

3.1.3. Predictive Criteria

In addition to posterior marginals, estimates for predictive criteria such as the deviance information criterion (DIC, Spiegelhalter et al., 2002), Watanabe-Akaike information criterion (WAIC, Wantanabe, 2010; Gelman et al., 2014) and marginal likelihood and conditional predictive ordinates (Held et al., 2010) can also be derived.

3.2. INLA-SPDE (Stochastic Partial Differential Equations) Approach

For certain members of GFs (Gaussian fields) with the Matérn covariance function, the GMRF representation on a triangulated lattice can be constructed explicitly through the use of certain SPDE (stochastic partial differential equation, Lindgren et al., 2011). As a result, GMRFs are no longer restricted to lattice (or areal) data. The link between GFs and GMRFs also allows for more realistic spatial statistical modeling.

Let $||\cdot||$ denote the Euclidean distance in $\mathbb{R}^d$, the Matérn covariance function between locations u, $v \in \mathbb{R}^d$ is defined as

$$C(u, v) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa_\nu||v - u||)^\nu K_\nu(\kappa_\nu||v - u||)$$

$$\propto (\kappa_\nu ||v-u||)^\nu K_\nu(\kappa_\nu ||v-u||),$$

where Γ is the gamma function, $\nu > 0$ is a shape parameter, $\kappa > 0$ is a scaling parameter, and $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+d/2)(4\pi)^{d/2}\kappa^{2\nu}}$ is the marginal variance, and K_ν is the modified Bessel function of the second kind with order $\nu > 0$, $\kappa > 0$ (Lindgren et al., 2011).

The Matérn covariance function appears naturally in various scientific fields (Guttorp and Gneiting, 2006). GFs $x(u)$ with the covariance function of the form above is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} x(u) = W(u), \quad u \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \quad \nu > 0,$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudodifferential operator, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian, $W(u)$ is spatial Gaussian white noise, ν controls the smoothness, and κ controls the range (Lindgren et al., 2011; Bolin, 2015). Any solution to the SPDE are named Matérn fields. Interested readers are referred to Lindgren et al.'s (2011) work (*or my dissertation but can't promise yet!*) for details about the SPDE approach.

3.3. Deep Generative Models

Just like common GP models have previously been shown to linked to GMRFs (Lindgren et al., 2011), a formal connection between GMRFs and CNNs (convolutional neural networks) has recently been established (Sidén & Lindsten, 2020).

4. Discussion

Reference

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., & Lindgren, F. (2018). Spatial modeling with R-INLA: A review. Wiley Interdisciplinary Reviews: Computational Statistics, 10(6), e1443.

Bolin, D. (2015). *Lecture 1: Introduction Gaussian Markov random fields.* Personal Collection of D. Bolin, Chalmers University of Technology, Gothenburg Sweden.

Bolin, D. (2015). *Lecture 2: Definitions and properties Gaussian Markov random fields.* Personal Collection of D. Bolin, Chalmers University of Technology, Gothenburg Sweden.

Bolin, D. (2015). *Lecture 9: SPDEs and GMRFs (part 2) Gaussian Markov random fields.* Personal Collection of D. Bolin, Chalmers University of Technology, Gothenburg Sweden.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423-498.

Morrison, K. (2017). A gentle INLA tutorial. Precision Analytics. https://www.precision-analytics.ca/articles/a-gentle-inla-tutorial/.

Rue, H., & Held, L. (2005). Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.

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

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., Krainski, E. T., & Fuglstad, G. A. (2017). INLA: Bayesian analysis of latent Gaussian models using integrated nested laplace approximations. R package version, 17, 20.

Sidén, P., & Lindsten, F. (2020). Deep gaussian markov random fields. In International Conference on Machine Learning (pp. 8916-8926). PMLR.

Tutorial

Gómez-Rubio, V. (2019). R-bloggers. Spatial Data Analysis with INLA. https://www.r-bloggers.com/2019/11/spatial-data-analysis-with-inla/.

Gómez-Rubio, V. (2020). 7.3 Geostatistics. *Bayesian inference with INLA.* CRC Press.

Moraga, P. (2019). 8 Geostatistical data. *Geospatial health data: Modeling and visualization with R-INLA and shiny.* CRC Press.