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 determined.

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 gained popularity in many fields. While MCMC are asymptotically exact methods, based on sampling, these methods not only can 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 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.

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

$$p(x,\theta|y) \propto p(y|x,\theta)p(x,\theta)$$

$$\propto p(y|x,\theta)p(x|\theta)p(\theta).$$

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:

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

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

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

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

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 unobserved latent field. 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), structure random effects (e.g. temporal dependency, spatial dependency, smoothness terms) (Bolin, 2015).

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

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

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 (Bolin, 2015).

Then, the posterior distribution, structured in a hierarchical way, becomes

$$\begin{aligned} p(x, \theta | y) &\propto p(y|x, \theta) p(x, \theta) \\ &\propto \prod_{i \in I} p(y_i | x_i, \theta) p(x|\theta) p(\theta). \end{aligned}$$

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 (~ 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 some 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 is a LGM iff the joint distribution of the latent field

$$x = (\eta, \alpha, \beta, f(\cdot))$$

is Gaussian. I.e.

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

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 is a Gaussian Markov random field.

2.3. Gaussian Markov Random Fields

A GMRF (Gaussian Markov Random Field) 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}.

Undirected graphs G are typically used to represent the conditional independence properties of the GMRF. An undirected graph G consists of a set of nodes V and edges E

$$G = (V, E),$$

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 include 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, but this can be an unreasonable assumption. 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 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 key reasons why the 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. The sparsity of the precision matrix also boosts computational efficiency, as compared to operations on dense matrices.

2.5. Laplace Approximations

The Laplace approximation is an old technique for the approximation of integrals (Barndorff-Nielsen & Cox, 1989). 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 $n \to \infty$.

Put it simply, Taylor series expansion states that a function about a point a can be expanded into a sum of terms and a finite number of these terms can be used as an approximation (Morrison, 2017). That is,

$$f(x) = f(a) + \frac{f^{'}(a)}{1!}(x-a) + \frac{f^{''}(a)}{2!}(x-a)^2 + \frac{f^{'''}(a)}{3!}(x-a)^3 + \cdots .$$

Specifying $f(x)$ with $\log g(x)$ and using only the first three terms of the expansion around the mode $\hat{x}$, then

$$\log g(x) \approx \log g(\hat{x}) + \frac{\partial \log g(\hat{x})}{\partial x}(x-\hat{x}) + \frac{\partial^2 \log g(\hat{x})}{\partial x^2}\frac{1}{2!}(x-\hat{x})^2 = \log g(\hat{x}) + \frac{\partial^2 \log g(\hat{x})}{2\partial x^2}(x-\hat{x})^2$$

$$= \log g(\hat{x}) - \frac{1}{2\sigma^2}(x-\hat{x})^2,$$

since the first derivative at the mode $\hat{x}$ is zero and $\hat{\sigma}^2 = -1/\frac{\partial^2 \log g(\hat{x})}{2\partial x^2}$.

Exponentiating and integrating both sides, then the above expression becomes

$$\int \exp(\log g(x)dx) = \int g(x)dx \approx \int \exp(\log g(\hat{x}) - \frac{1}{2\sigma^2}(x-\hat{x})^2)dx$$

$$= c\int \exp(-\frac{(x-\hat{x})^2}{2\sigma^2})dx,$$

where $c = \int \exp(\log g(\hat{x}))dx$ is some constant. Using a Laplace approximation to approximate the distribution $f(x)$ that is approximately normal, mean $\hat{x}$ can be found by solving $f^{'}(x) = 0$ and variance $\hat{\sigma}^2 = -1/f^{''}(x)$ at the mode $\hat{x}$.

$\dots\dots$ Laplace approximation, which uses a sequence of Gaussian approximations, can do much better.

3. INLA

3.1. INLA

The main goal of Bayesian inference is to approximate the posterior marginals for the hyperparameters and latent field respectively

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

The posterior marginals of interest can be rewritten as

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

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

and the key feature of the INLA approach is to use the above form to construct nested approximations

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

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

where $\tilde{p}(\cdot|\cdot)$ is an approximated conditional density of its arguments (Rue et al., 2009).

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.

......

3.1.1. Approximating the Posterior Marginals for the Hyperparameters

......

The conditional probability of θ given y is given as

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

Expanding the numerator (such that $p(x,\theta|y) = p(y|x,\theta)p(x|\theta)p(\theta)$) and replacing the denominator with a Laplace approximation, the above expression becomes

$$\tilde{p}(\theta|y) = \left.\frac{p(y|x,\theta)p(x|\theta)p(\theta)}{\tilde{p}(x|\theta,y)}\right|_{x=x^*(\theta)} = \tilde{p}((\theta|y)),$$

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).

3.1.2. Approximating the Posterior Marginals for the Latent Field

Approximate the posterior marginals for the latent field is 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 large (10^3 to 10^5).

......

......

The third option, which is also the default setting in the `R-INLA` software, uses a simplified Laplace approximation.

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 can be constructed explicitly through the use of certain SPDE (stochastic partial differential equation, Lindgren et al., 2011).

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

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

where $\nu > 0$ 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 (Lindgren et al., 2011). Since the Matérn covariance function, along with other covariance functions such as exponential and Gaussian, is isotropic (such that the covariance function depends only on distance, and not direction, between points), $C(u,v)$ is sometimes written as $C(h)$, where h is the distance.

The Matérn covariance function appears naturally in various scientific fields (Guttorp and Gneiting, 2006). Matérn fields, which have covariance function of the form above, are the stationary solutions 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.

4. Spatial Examples Using The `R-INLA` Package

Two data sets are considered. The `NY8` data set is areal data. The `meuse` data set is point-referenced (or geostatistical) data. Since the focus of this part of the project is on the use of `R-INLA` package, the details of the areal and geostatistical models used for each data set are only briefly introduced here.

4.1. A Spatial Areal Example of Leukemia Incident Cases

The `NY8` data set contains the number of incident leukemia cases per census tract in an eight-country region of upstate New York from 1978-1982 (Waller & Gotway, 2004; Bivand et al., 2008). The `NY8` data set can be accessed from the **R** package `DClusterm`, and it is a `SpatialPolygonsDataFrame` object.

For this data set, a total of 5 models (fixed effects, random effects (iid), ICAR, BYM and Leroux et al.) are fitted, and results include criteria for model selection (marginal log-likelihood, DIC and WAIC) and plots of predicted values.

Since the number of incident leukemia cases is count, Poisson GLMs with fixed effects and random effects are fitted. Since there is spatial dependence in the data, spatial models (GLMs with spatial random effects) such as ICAR (Intrinsic Conditional autoregressive), BYM and Leroux et al. model are also considered.

Without going into details, recall that the GLMs have the following form

$$Y = X\beta + Z\alpha + \varepsilon,$$

where β is a vector of fixed effects with design matrix X, α is a vector of random effects with design matrix Z, and ε is an error term, where it is assumed that $\varepsilon_i \sim N(0, \sigma^2), i = 1, ..., n$. The vector of random effects α is modeled as MVN (it is assumed that)

$$\alpha \sim N(0, \sigma_\alpha^2 \Sigma),$$

where the covariance matrix Σ is defined such that it induces higher correlation with adjacent areas.

There are several ways to include spatial dependence in Σ, and in spatial areal model especially, it is more common to model the precision matrix Q directly, where $Q = \Sigma^{-1}$. In ICAR (Intrinsic CAR), $\Sigma^{-1} = diag(n_i) - W$, where n_i is the number of neighbors of area i. In Leroux et al.'s model (mixture of matrices), $\Sigma^{-1} = ((1-\lambda)I_n + \lambda M), \lambda \in (0, 1)$, where M is precision of intrinsic CAR specification. The BYM (Besag, York and Mollié) model includes two latent random effects: an ICAR latent effect and a Gaussian iid latent effect.

Results show that for spatially dependent data, spatial models generally perform better than GLM with fixed or random (iid) effects. It is also no surprise that the baseline model (fixed effects model) appears to be the poorest fit of all.

	Marg_logLik	DIC	WAIC
fixed	-514.4	1016	1017
iid	-512.1	979.3	983.6
ICAR	-718.9	968.3	972.2
BYM	-458.9	967.4	971.9
Leroux	-508.3	967.3	971.3

Plots of predicted values are omitted here since the data analysis focus is on Section 4.2.

4.2. A Spatial Geostatistical Example of Heavy Metal Concentrations

5. 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.

Tutorial