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

Frances Lin

June 2022

Background and Introduction

The steps involving the Bayesian inference may appear easy and straightforward:

- ▶ updating prior beliefs about the unknown parameters 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.

- ▶ MCMC (Markov chain Monte Carlo) represented a breakthrough in Bayesian inference in the early 1990s.
- ▶ Tools such as `WinBugs` (Spiegelhalter et al., 1995), `JAGS` (Plummer, 2016), and `stan` (Stan Development Team, 2015) have also been developed, and
- ▶ Bayesian statistics has gained popularity in many fields.

Background and Introduction

However, MCMC methods

- ▶ 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. INLA

- ▶ can be applied to a very flexible class of models named LGMs (latent Gaussian models), which ranges from GLMMs (generalized linear mixed models) to time-series, and spatial and spatio-temporal models.
- ▶ allows for faster and more accurate inference without trading speed for accuracy, and
- ▶ is accessible through the **R** package R-INLA.

Applications

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

Selected examples include:

▶ polio-virus eradication in Pakistan (Mercer 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);

▶ probabilistic prediction of wind power (Lenzi et al., 2017);

▶ applications in spatial econometrics (Bivand et al., 2014; Gomez-Rubio et al., 2015; Gomez-Rubio et al., 2014);

▶ predicting extreme rainfall events in space and time (Opitz et al., 2018), etc.

Outline

- ▶ Key Components
  - 0. Bayesian Inference
  - 1. Latent Gaussian Models
  - 2. Additive Models
  - 3. Gaussian Markov Random Fields
  - 4. Additive Models and Gaussian Markov Random Fields
  - 5. Laplace Approximations
- ▶ INLA
- ▶ INLA-SPDE (Stochastic Partial Differential Equations) Approach
- ▶ Discussion
- ▶ Spatial Examples using the package R-INLA
  - 1. Leukemia Incident Cases (Areal data)
  - 2. Heavy Metal Concentrations (Geostatistical data)

0. Bayesian Inference

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

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

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.

▶ Based on the posterior distribution, relevant statistics for the parameters of interest (e.g. marginal distribution, means, variances, and credibility intervals) can be obtained.

▶ However, the (multiple) integral $\int p(y|\theta)p(\theta)d\theta$ is generally intractable in closed-form, thus requiring the use of numerical methods such as MCMC.

1. Latent Gaussian Models

The latent Gaussian models (LGMs) is a class of three-stage Bayesian hierarchical models. It involves the following stages:

1. Observations y is assumed to be conditionally independent, given a latent Gaussian random field x and hyperparameter θ_1

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

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

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

where $Q = \Sigma^{-1}$ is the precision matrix and θ_2 is a hyperparameter.

3. The hyperparameters of the latent field that are not necessarily Gaussian are assumed to follow a prior distribution

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

where $p(\cdot)$ is a known distribution.

1. Latent Gaussian Models

Then, the 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 *dim* of n is high (10^3 to 10^5).
3. The number of hyperparameters θ is small (~ 2 to 5 but < 20).

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.

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.

Consider the Bayesian structured additive model setup, for example,

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

2. Additive Models

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.

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 $\alpha + \sum_k f_k(w_k)$.

2. Additive Models

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.

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 $i \neq j \in V$.

3. Gaussian Markov Random Fields

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}

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

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.

3. Gaussian Markov Random Fields

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 ($Q = \Sigma^{-1}$), and conditional independence is a more reasonable assumption and their properties are encoded in the precision matrix.

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.

5. Laplace Approximations

The Laplace approximation (Barndorff-Nielsen & Cox, 1989)

- ▶ It is an old technique for the approximation of integrals.
- ▶ 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.

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 $n \to \infty$.

5. Laplace Approximations

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

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

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.

INLA (Integrated Nested Laplace Approximation)

INLA uses a nested version of the Laplace approximation, combined with modern numerical techniques for integration.

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.

INLA (Integrated Nested Laplace Approximation)

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 (each element of) interest can be obtained by

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

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

where $\theta_{-j} = (\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ...)$.

INLA (Integrated Nested Laplace Approximation)

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

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

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

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

- ▶ Approximation to $p(\theta_j|y)$ is computed by integrating out the remaining parameters θ_{-j} from $\tilde{p}(\theta|y)$.
- ▶ 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).

Approximating the Posterior Marginals for the Hyperparameters

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

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

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.

Approximating the Posterior Marginals for the Hyperparameters

The denominator is approximated using a Gaussian approximation

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

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

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 derivatives of the log-likelihood at the mode, with respect to x_i.

Approximating the Posterior Marginals for the Hyperparameters

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 interpolant.

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

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

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).

Approximating the Posterior Marginals for the Latent Field

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.

Approximating the Posterior Marginals for the Latent Field

For the second challenge, three approximation options are available in the R-INLA package:

- ▶ the Gaussian approximation,
- ▶ the Laplace approximation, and
- ▶ the simplified Laplace approximation (Rue et al., 2009, as cited in Morrison, 2017).

Approximating the Posterior Marginals for the Latent Field

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).

Approximating the Posterior Marginals for the Latent Field

Then Newton-like methods are used to explore the joint posterior distribution of the hyperparameters to find points for the numerical integration

$$\tilde{p}(x_i|y) \approx \sum_{h=1}^{H} \tilde{p}(x_i|\theta_h^*, y)\tilde{p}(\theta_h^*|y)\Delta_h$$

(Rue et al., 2009, as cited in Morrison, 2017).

For the numerical integration here, three options are available: grid search, central composite design, and empirical Bayes (Rue et al., 2009).

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.

INLA-SPDE (Stochastic Partial Differential Equations) Approach

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).

INLA-SPDE (Stochastic Partial Differential Equations) Approach

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

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

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.

Discussion

(1) The INLA approach appears promising for a wide variety of applications because of the model abstraction of the LGMs. Due to time constraint, some of the theoretical as well as computational details are only briefly mentioned in this project. For example, key component such as GMRFs, Laplace Approximations, or SPDEs requires its own exploration and examination.

(2) The generality of the package R-INLA increases complexity for the user. For a restricted set of models, efforts have been done to create packages with a simplified interface.

Discussion

(3) For this project, the data preprocessing stage for the `meuse` (geostatistical) data is quite involved, compared to that for the `NY8` (areal) data. Extending its application to a large-scale (or big) data set may be time consuming.

(4) 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).

Spatial Areal Example Using The R-INLA Package

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.

▶ 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).

Spatial Areal Example Using The R-INLA Package

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.

Table 1: criteria for model selection

	Marg_logLik	DIC	WAIC
fixed	-514.4	1016	1017
iid	-512.1	979.2	983.6
ICAR	-718.9	968.3	972.2
BYM	-458.9	967.4	971.9
Leroux	-508.3	967.3	971.3

Spatial Geostatistical Example Using The R-INLA Package

The `meuse` data set contains locations, topsoil heavy metal concentrations and a number of soil and landscape variables at the observed locations in a flood plain of the river Meuse.

- ▶ The `meuse` data set can be accessed from **R** package gstat, and it is a `data.frame` object.
- ▶ An universal kriging model and a continuous spatial process with a Matérn covariance function using the INLA-SPDE approach (referred to as the SPDE model) are fitted, and results include summary results and predicted plots of kriging vs SPDE.
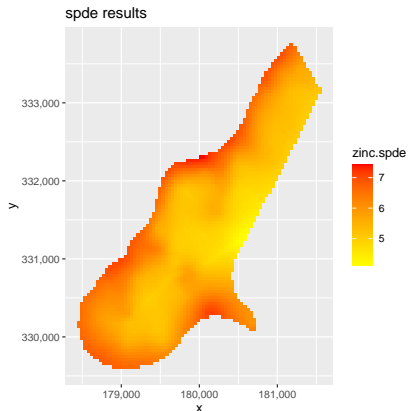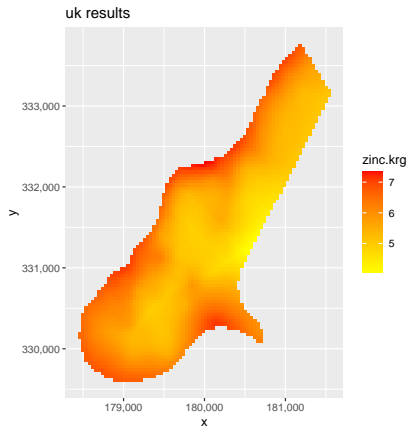
Spatial Geostatistical Example Using The R-INLA Package

Results show that the universal kriging and the SPDE model provide similar estimates.

- ► For the fitted variogram model, sill (or variance) = 0.2053, nugget (i.e., variance when distance = 0) = 0.07643, and range (i.e. the distance after which the variogram levels off) = 728.7.

- ► For the SPDE model, mean variance = 0.2351 with $Q1 = 0.183$ and $Q3 = 0.2754$, and max range = 955.8 (This value is a bit higher than expected).
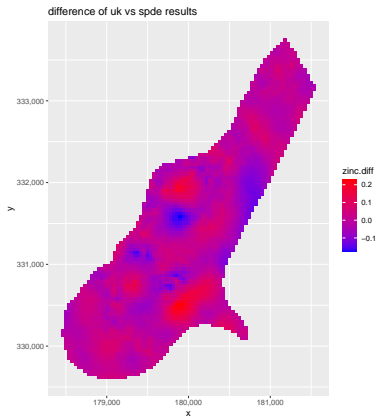
Spatial Geostatistical Example Using The R–INLA Package

Both plots show higher estimated means of log-concentrations of zinc at locations closer to the Meuse river.

Spatial Geostatistical Example Using The R-INLA Package

The differences may be due to the ways how the models were defined and how their model components were specified.

Reference

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., 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.

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.

Thank you!

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

Frances Lin

PhD student, Dept. of Statistics, Oregon State University

Full report, code, etc. are available at

GitHub: franceslinyc, INLA-with-Spatial-data-2022