Reparameterized SGLMM (Spatial Generalized Linear Mixed Model): A Review

Frances Lin

June 2022

Background and Introduction

The SGLMM (spatial generalized linear mixed model)

- is a hierarchical model that introduces spatial dependence through a GMRF (Gaussian Markov random field, Besag et al., 1991).
- was initially proposed for prediction for count data, but
- has later been applied to estimation and prediction for other types of data (e.g. binary) and
- found applications in many fields (e.g. ecology, geology, epidemiology, image analysis, and forestry).

Background and Introduction

The SGLMM has been the dominant model for areal data because of

- its flexible hierarchical specification,
- the availability of the software `WinBUGS` for (Bayesian) data analysis (Lunn et al., 2000) and
- various theoretical and computational advantages over the competitive model named automodel.

However, SGLMMs suffer from two major shortcomings:

i) variance inflation due to spatial confounding and

ii) computational challenges posed by high dimensional latent variables ("random effects").

Background and Introduction

On the other hand, while another model named RHS model

- seeks to alleviate confounding (Reich et al., 2006),
- can result in random-effects structure with negative spatial dependence (i.e. repulsion) that is not typically applied in practice.

In this paper, a reparameterized model is proposed, and it is able to

- alleviate confounding, and, at the same time,
- include patterns of positive spatial dependence (i.e. attraction) while excluding patterns of repulsion.

The proposed model is also one of the first dimension reduction techniques for spatial areal models.

Outline

- Traditional SGLMM (Spatial Generalized Linear Mixed Model)
- Spatial confounding (via the automodel)
- RHZ model
- **Sparse reparameterization of the areal SGLMM**
- Dimension reduction
- Simulated application
- Application (using the **R** package `ngspatial`)
- Discussion

Traditional SGLMM

The traditional SGLMM (sometimes referred to as the BYM model) is a hierarchical model (Besag, et al., 1991).

Let $G = (V, E)$ be an undirected, labelled graph, where $V = \{1, 2, ..., n\}$ is a set of vertices (nodes) and $E = \{i, j\}$ is a set of edges, where $i, j \in V$, $i \neq j$.

- Each vertex represents an area of interest and each edge represents the proximity of areas i and j.
- The graph G is represented using an adjacency matrix A, which is a nxn matrix with $diag(A) = 0$ and entries $A_{ij} = 1\{(i, j) \in E, i \neq j\}$, where $1(\cdot)$ is an indicator function (i.e. entry $a_{ij} = 1$ if vertex v_i and v_j are adjacent. $a_{ij} = 0$, otherwise).

Traditional SGLMM

Further let $Z = (Z_1, \ldots Z_n)^T$ be the random field of interest, where Z_i is the random variable associated with vertex i. Then, the first stage of the model is given by

$$g(E(Z_i|\beta, W_i)) = X_i\beta + \mathbf{W_i}, \tag{1}$$

where

- g is a link function,
- X_i is the ith row of the design matrix X,
- β is a p-vector of regression parameters and
- W_i **is a spatial random effect associated with vertex** i**.**

Different types of data require different canonical choices of the link function g (e.g. the logit function for spatial binary data and the logarithm function for spatial count data).

GMRF prior for the random effects

The field of random effects $W = (W_1, ..., W_n)^T$, through which spatial dependence is incorporated, is assumed to follow the intrinsic conditionally autoregressive (ICAR) or improper GMRF prior

$$p(W|\tau) \propto \tau^{rank(Q)/2} exp(-\frac{\tau}{2} W^T Q W), \qquad (2)$$

where τ is a smoothing ("precision") parameter and $Q = diag(A1) - A$ is a precision matrix. The precision matrix Q incorporates both dependence and prior uncertainty.

Note. A SGLMM

1. can be reformatted by replacing GMRF with GP (Gaussian process) for point-referenced (or geostatistical) data.
2. is restricted to a Bayesian or restricted maximum likelihood (REML) analysis since the prior (2) is improper.

Spatial confounding (via the automodel)

The automodel, SGLMM's closest competitor, is a Markov random field (MRF) model that incorporates dependence directly and is defined as

$$g(E(Z_i|\beta,\eta,Z_{-i})) = X_i\beta + \eta \sum_{(i,j)\in E} Z_j^*, \tag{3}$$

where

- g is a (canonical) link function,
- η is the dependence parameter ($\eta > 0$ implies attraction; $\eta < 0$ implies repulsion), and
- Z_{-i} is the field excluding the ith observation.

Spatial confounding (via the automodel)

- For the centered automodel, $Z_j^* = Z_j - \mu_j$, where μ_j is the independence expectation of Z_j
  (i.e. $\mu_j = E(Z_j|\beta, \eta = 0) = g^{-1}(X_j\beta)$).
- For the uncentered automodel, $Z_j^* = Z_j$.

The sum term $\sum_{(i,j)\in E} Z_j^*$ is called the autocovariate, and it is considered as a synthetic predictor.

- The centered autocovariate makes the dependence parameter easily interpretable (η captures the relativity of an observation to its neighbours, conditional on the hypothesized regression component).
- The uncentered autocovariate not only poses conceptual challenge but also shows spatial confounding.

The RHS model

The traditional SGLMM can also shown to be confounded.

Consider P be a projection onto $C(X)$

$$P = X(X^T X)^{-1} X^T$$

and let $P^{\perp}$ be the projection onto $C(X)$'s complement such that $P^{\perp} = I - P$, then equation (1) can be rewritten as

$$g(E(Z_i|\beta, W_i)) = X_i\beta + \mathbf{K_i}\gamma + \mathbf{L_i}\delta,$$

where K and L are orthogonal bases for $C(X)$ and $C(X)^{\perp}$ respectively and γ and δ are random coefficients.

K and X now share the same column space, and this is the source of the *spatial confounding*.

The RHS model

Since K has no practical meaning, setting $\gamma = 0$, the first stage of the RHS model is given by

$$g(E(Z_i|\beta, \delta)) = X_i\beta + \mathbf{L_i}\delta,$$

and, compared to equation (2), the prior for the random effects δ becomes

$$p(\delta|\tau) \propto \tau^{(n-p)/2} exp(-\frac{\tau}{2}\delta^T Q_R \delta),$$

where $Q_R = L^T Q L$.

The sparse, reparameterized areal SGLMM

However, the RHZ model does not allow parsimonious fitting of the residual clustering.

- The geometry corresponding to the projection $P^{\perp}$ fails to account for the underlying graph G, thus permitting structure of negative spatial dependence (i.e. repulsion) in the random effects.
- This is not useful in practice since neighboring observation tends to be similar, rather than dissimilar.

The proposed reparameterized model

i) considers an alternative projection that captures the geometry of the models, thus allowing models with only patterns of positive spatial dependence (i.e. attraction).

ii) utilizes the geometry of the models, which leads to dimension reduction of the random effects naturally.

The sparse, reparameterized areal SGLMM

Consider the operator $(I - 11^T/n)A(I - 11^T/n)$ that appears in the numerator of Moran's I-statistic (a commonly used for nonparametric method for spatial dependence)

$$I(A) = \frac{n}{1^T A 1} \frac{Z^T (I - 11^T/n) A (I - 11^T/n) Z}{Z^T (I - 11^T/n) Z},$$

where I is a nxn identity matrix and 1 is a n-vector of 1s.

Next replace $I - 11^T/n$ with $P^\perp$, then the resulting operator called the Moran operator for X with respect to G, $P^\perp A P^\perp$, appears in the numerator of the generalized Moran's I-statistic

$$I_x(A) = \frac{n}{1^T A 1} \frac{Z^T P^\perp A P^\perp Z}{Z^T P^\perp Z},$$

where P again is a projection onto $C(X)$ so $P^\perp$ is the projection onto $C(X)$'s complement.

The sparse, reparameterized areal SGLMM

Replacing L with M in the RHS model, the first stage of the reparameterized model is given by

$$g(E(Z_i|\beta, \delta_S)) = X_i\beta + \mathbf{M_i}\delta_S,$$

and the prior for the random effects δ_S becomes

$$p(\delta_S|\tau) \propto \tau^{q/2} exp(-\frac{\tau}{2}\delta_S^T Q_S \delta_S),$$

where $Q_S = M^T Q M$.

Note. It is assumed that

1. M is a matrix that contains the first $q \ll n$ eigenvectors of the Moran operator.
2. $\lambda_q > 0$ (since neighboring observations tend to be similar in practice). Half of the eigenvectors can be discarded as a result, making it possible to achieve a much greater dimension reduction.

Comparision of various SGLMMs

Table 1 compares and contrasts the five models.

Table 1: Comparision of various SGLMMs

Model	Confounded	Account_G
Traditional SGLMM	Yes	No
Uncentered automodel	Yes	Yes
RHZ SGLMM	No	No
Centered automodel	No	Yes
Sparse SGLMM	No	Yes

Dimension reduction for spatial models

- An areal mixed model: Metropolis-Hastings algorithm for sampling from the posterior distribution of the field of random effects W is slow when components of W exhibit strong *a posteriori* dependence. Various methods such as MCMC block sampler have been proposed.
- The random effects for the RHZ and the sparse models, in contrast, are *a posteriori* uncorrelated.
- The RHZ model: Time complexity of the operation is in $\mathcal{O}(n^2)$, which is still significant enough to discourage or prevent application of the model to large data sets.
- The sparse model: Evaluation of the quadratic form $\delta_S^T Q_S \delta_S$ can be $\mathcal{O}(1)$, which makes the model more suitable for large-scale data sets. (Recall that taking full advantage of G reduces the *dim* of random effects.)

Simulated application

The section is kept as a brief summary, and the figures are obtained from the paper.

Five models (e.g. a non-spatial ordinary linear model, the centered autologistic, the traditional SGLMM, the RHZ model and the sparse model) are fitted for the simulated binary, count and normal data.

Simulated application: Algorithms

SGLMMs are fitted using Metropolis-Hastings random walk and/or Gibbs updates.

- For the binary and count data,
  - β is updated using a random walk with normal proposals $\beta^{(j+1)} \sim N(\beta^{(j)}, \hat{V})$, where $\hat{V}$ is the estimated asymptotic covariance matrix from a classical generalized linear model fit.
  - W (for the traditional SGLMM) is updated using a univariate random walk with normal proposals.
  - δ (for the RHZ model) and δ_S (for the sparse SGLMM) are updated using a multivariate random walk with a spherical normal proposal.
- For the normal data, all parameters are updated using Gibbs updates.
- For all three types of data, τ is a Gibbs update irrespective of the response distribution.

Simulated application: Metropolis-Hastings random walk

For example, `sparse.sglmm.fit.binomial()`:

1. Initialize β, δ_S and τ (precision parameter)
2. Input Y, X, M, etc. and obtain $\hat{V}$ estimated asymptotic covariance matrix from a `glm` fit
3. Update
   - $\beta^{(j+1)} \sim N(\beta^{(j)}, \hat{V})$
   - $\delta_S^{(j)}$
   - $\eta^{(j+1)}(= g(\cdot)^{(j+1)}) = X\beta^{(j+1)} + M\delta_S^{(j)}$
   - $\eta^{(j)}(= g(\cdot)^{(j)}) = X\beta^{(j)} + M\delta_S^{(j)}$
   - Construct the log-likelihood: $y^T \log(\hat{y}) + (1-y)^T \log(1-\hat{y}) + \cdots$ for $j+1$ and j, where $\hat{y} = g^{-1}(\eta)$
   - Take difference of the log-likelihood (ratio of the likelihood) for $j+1$ and j and set it to α acceptance ratio
4. Generate $u \sim Unif(0,1)$ and if $\log(u) \leq \alpha$, then accept it as the next step $j+1$ and update β, δ_S and repeat and if. . . , then accept and update δ_S and τ. Otherwise, stay at the current step j. (Nested if-else is required.)

Simulated application: Data sets

- Binary data:
  (1) setting $\tau = 1$,
  (2) simulating random effects according to $\delta_S \sim N(0, Q_s^{-1})$, and
  (3) simulating independent observations according to $Z|\delta_S \sim Bern(p)$, where $p = \exp(x + y + M\delta_S)/(1 + \exp(x + y + M\delta_S))$ is a vector of true probabilities.
- Count data:
  (1) simulating random effects according to $\delta_S \sim N(0, (3Q_s)^{-1})$, and
  (2) simulating independent observations according to $Z|\delta_S \sim Pos(\lambda)$, where $\lambda = \exp(x + y + M\delta_S)$ is a vector of true rates.
- Normal data:
  (1) simulating random effects according to $\delta_S \sim N(0, Q_s^{-1})$, and
  (2) simulating observations according to $Z|\delta_S \sim N(\mu = x + y + M\delta_S, \sigma^2 I)$, where $\sigma^2 = 1$.

Simulated application

- Binary data: The RHZ and sparse models perform comparably and reliably. The traditional SGLMM, in contrast, not only gives poor estimate but also results in wider CIs.

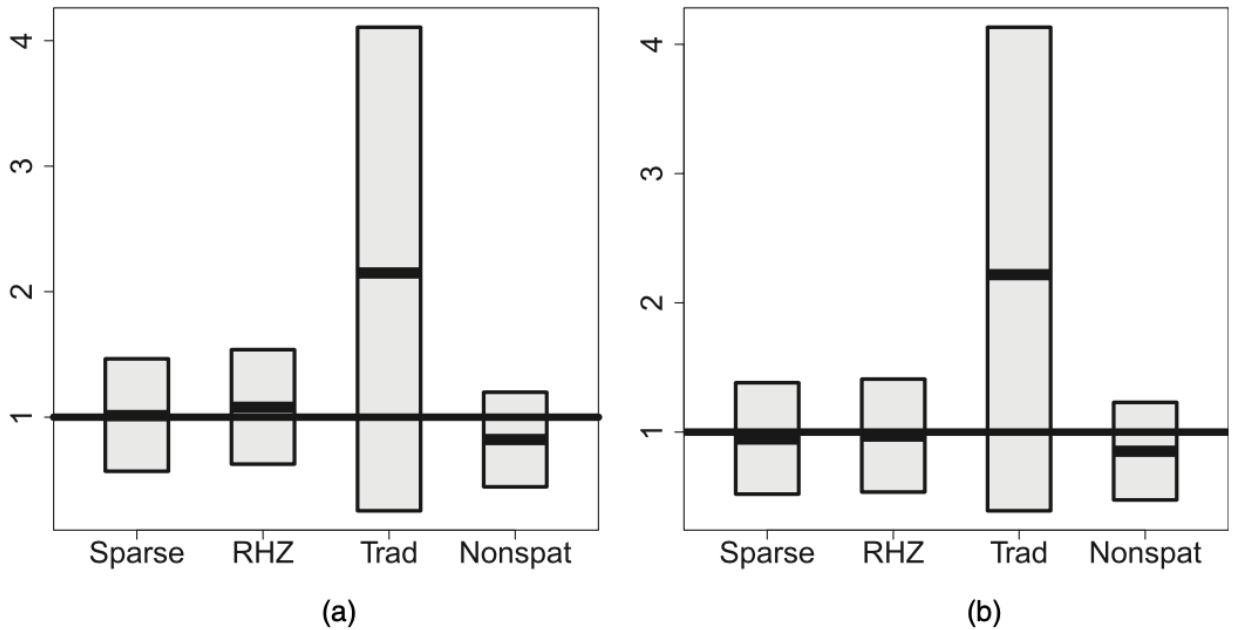

Fig. 4. Boxplots illustrating inference for β for the simulated binary data: (a) β_1; (b) β_2

Simulated application

- Count data: The traditional SGLMM gives good estimate but again results in wider CIs.

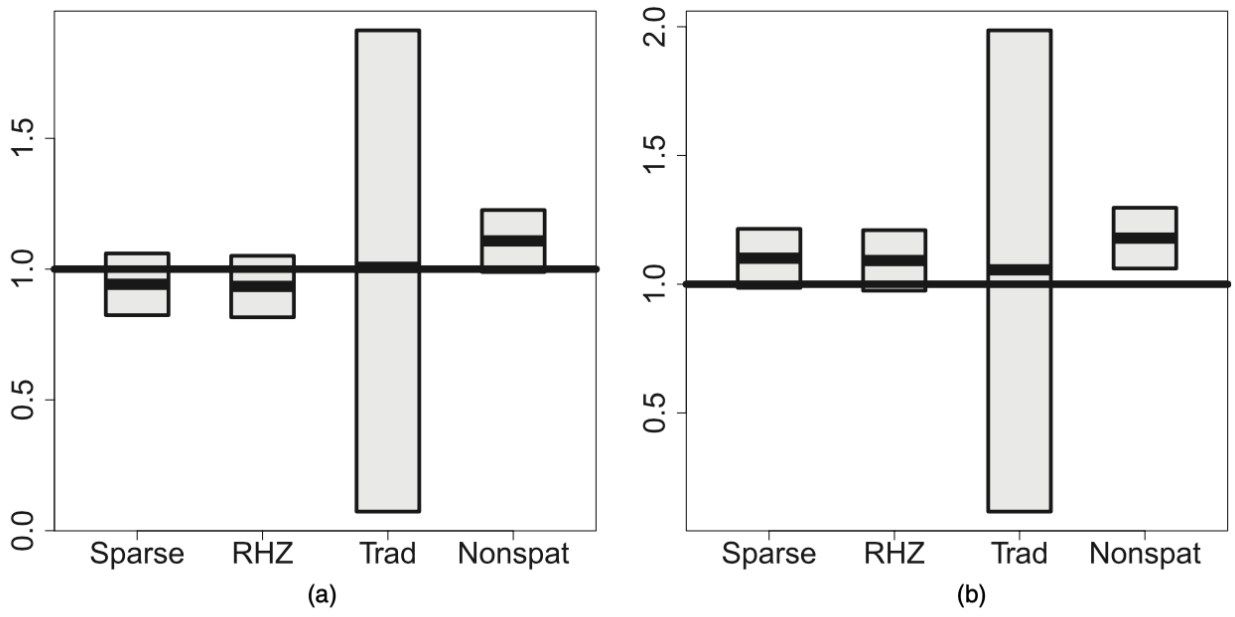

Fig. 8. Boxplots illustrating inference for β for the simulated count data: (a) β_1; (b) β_2

Simulated application

- Normal data: The RHZ and sparse models result in narrower CIs because non-spatial model overestimates σ^2. The traditional SGLMM gives poor estimate and causes variance inflation so large that it causes a type II error.

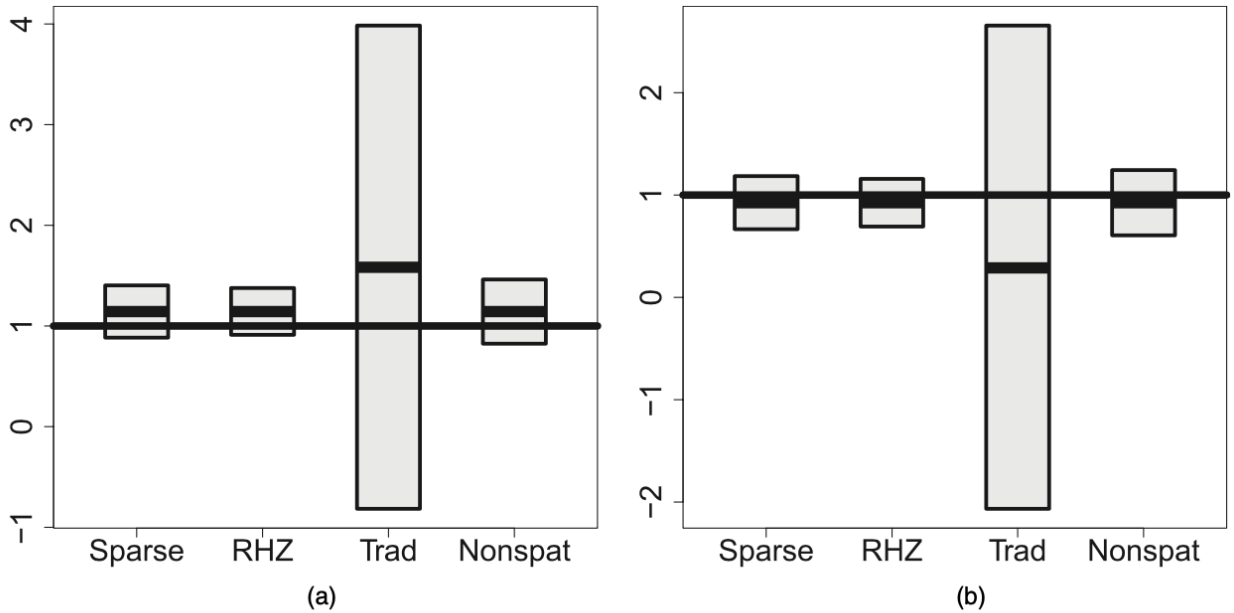

Fig. 9. Boxplots illustrating inference for β for the simulated Gaussian data: (a) β_1; (b) β_2

Application

The `infant` data set contains infant mortality rate for $n = 3071$ US counties.

- The `infant` data set and its adjacency matrix A (respectively, a `data.frame` and `matrix` object) can be accessed from the **R** package `ngspatial`.
- A sparse Poisson SGLMM is fitted

$$E(deaths_i|\beta, \delta_S) = births_i \exp(\beta_0 + \beta_1 low_i + \beta_2 black_i + \beta_3 hisp_i + \beta_4 gini_i + \beta_5 aff_i + \beta_6 stab_i + M_i\delta_S),$$

where

  - death is the number of infant deaths, births is the number of live births, low is the rate of low birth weight, black is the percentage of black residents (2000 US census), hisp is the percentage of Hispanic residents (2000 US census), gini is the Gini coefficient which measures income inequality, aff is a score of social affluence and stab is residential stability.

Application

Results show that some estimates (e.g. `low_weight`, `gini`) differ from the results from the original study. The process of fitting the model takes approximately 52.47736 minutes.

Coefficients:

	Estimate	Lower	Upper	MCSE
X	-4.987e+00	-5.166e+00	-4.807000	1.011e-03
Xlow_weight	-1.069e+02	-1.905e+02	-24.930000	3.409e-01
Xblack	9.668e-03	8.516e-03	0.010840	9.032e-06
Xhispanic	-5.416e-03	-6.578e-03	-0.004273	8.165e-06
Xgini	6.227e-02	-3.803e-01	0.491400	2.547e-03
Xaffluence	-9.476e-02	-1.070e-01	-0.082480	5.893e-05
Xstability	-1.333e-02	-2.890e-02	0.002744	9.467e-05

DIC: 10280

Number of iterations: 1000000

Discussion

(1) The sparse SGLMM not only alleviates spatial confounding but also accounts for the underlying graph while achieving dimension reduction naturally.

- However, due to time constraint, some of the computational details are not well explored in this project (e.g. the choice of operators, the choice of priors, MCMC methods and its variants as well as the C++ code used in the package).

Discussion

(2) The sparse SGLMM appear faster, compared to the traditional SGLMM or the RHZ model. However, MCMC methods may still take hours or days to run.

- INLA algorithm (a project that I am currently working on for the Bayesian Statistics course) is a fast alternative for fitting areal and more models and can be accessed through the **R** package `R-INLA`. It would be interesting to see comparisons of accuracy and speed.

Reference

Bolin, D. (2015). *Lecture 5: Intrinsic GMRFs Gaussian Markov random fields.* Personal Collection of D. Bolin, Chalmers University of Technology, Gothenburg Sweden.

Haran, M. (2011). Gaussian random field models for spatial data. Handbook of Markov Chain Monte Carlo, 449-478.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(1), 139-159.

Hughes, J. (2014). ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data. R Journal, 6(2).

Thank you!

Reparameterized SGLMM (Spatial Generalized Linear Mixed Model): A Review

Frances Lin

PhD student, Dept. of Statistics, Oregon State University