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.

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
-

Traditional SGLMM

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 $Z = (Z_1, ...Z_n)^T$ 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

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

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 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 data (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 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 centred 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 uncentred 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 centred autocovariate makes the dependence parameter easily interpretable (η captures the relativity of an observation to its neighbours, conditional on the hypothesized regression component).
- The uncentred 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

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

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

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

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 $Q_R = L^T QL$.

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 *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 (top) of Moran's I-statistic (a commonly used for nonparametric method for spatial dependence)

$$I(A) = \frac{n}{1^T A1} \frac{Z^T(I - 11^T/n)A(I - 11^T/n)Z}{Z^T(I - 11^T/n)Z},$$

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, $P^\perp AP^\perp$, appears in the numerator of the generalized Moran's I-statistic

$$I_x(A) = \frac{n}{1^T A1} \frac{Z^T P^\perp AP^\perp Z}{Z^T P^\perp Z}.$$

The sparse, reparameterized areal SGLMM

The sparse, reparameterized areal SGLMM

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

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

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 $Q_S = M^T QM$.

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
Uncentred automodel	Yes	Yes
Centred automodel	No	Yes
RHZ SGLMM	No	No
Sparse SGLMM	No	Yes

Dimension reduction for spatial models

Evaluation of the quadratic form $\delta_S^T Q_S \delta_S$ for the sparse model can be $\mathcal{O}(1)$, which makes the model more suitable (than the RHZ model) for large-scale data sets.