SGLMM (Spatial Generalized Linear Mixed Model): A Review

Frances Lin

June 2022

1. 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). The SGLMM was initially proposed for prediction for count data. Later, it has been applied to estimation and prediction for other types of data (e.g. binary) and found applications in many fields (e.g. ecology, geology, and forestry).

The SGLMM has been the dominant model for areal data because of its flexible hierarchical specification, the availability of the software `WinBUGS` for 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). On the other hand, while a new model named RHS model seeks to alleviate confounding by introducing synthetic predictors that are orthogonal to the fixed-effects predictors (Reich et al., 2006), failing to account for the underlying graph, the RHS model can result in random-effects structure with negative spatial dependence (i.e. repulsion) that is not typically applied in practice.

A new model that is able to alleviate confounding, and, at the same time, include patterns of positive spatial dependence (i.e. attraction) and exclude patterns of repulsion is proposed. While other dimension reduction methods have been proposed (Higdon, 2002; Cressie & Johannesson, 2008; Banerjee et al., 2008; Furrer et al., 2006; Rue & Tjelmeland, 2002), these methods focus on spatial point-referenced models. The proposed model is one of the first dimension reduction techniques for spatial areal models.

2. Traditional SGLMM (Spatial Generalized Linear Mixed Model)

Undirected graphs

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. 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.

The traditional SGLMM

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

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

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. For example, for binary spatial data is, it is the logit function, for count data, it is the natural logarithm function, and for normal data, it is the identity function (Nelder & Wedderburn, 1972).

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 GMRF prior

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

where τ is a smoothing parameter and $Q = diag(A1) - A$ is a precision matrix. The precision matrix Q incorporates both dependence and prior uncertainty. W_i and W_j, $i \neq j$, are independent given their neighbors $iff\ Q_{ij} = Q_{ji} = 0\ iff\ (i, j) \in E$, and uncertainty about W_i is inversely proportional to the degree of vertex i.

For spatial data over a continuous domain, i.e. point-referenced data or geostatistical data, a SGLMM can be formulated by replacing the GMRF with a GP (Gaussian process) (Diggle et al., 1998; De Oliveira, 2000; Christensen & Waagepetersen, 2002). In addition, since the GMRF prior is improper (Q is singular), the SGLMM is restricted to a Bayesian or restricted maximum likelihood (REML) analysis.

3. Spatial confounding

The automodel

Recall that SGLMM's closest competitor the automodel, a Markov random-field model that incorporates dependence directly, is given as

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

where g is a (canonical) link function, Z_{-i} is the field excluding the ith observation, and η is dependence parameter ($\eta > 0$ implies attraction; $\eta < 0$ implies repulsion). $Z_j^* = Z_j - \mu_j$ is the centred automodel, where μ_j is the independence expectation of Z_j (i.e. $\mu_j = E(Z_j|\beta, \eta = 0) = g^{-1}(X_j\beta)$).

The sum term $\sum_{(i,j)\in E} Z_j^*$ is called the autocovariate, and it is a synthetic predictor. For the centred automodel, the autocovariate is easily interpretable. However, for the uncentred automodel, the autocovariate not only poses conceptual challenge but also shows spatial confounding. I.e. the uncentred autocovariate is not easy to interpret, and β and η also tend to be strongly correlated.

The RHS model

The traditional SGLMM can also shown to be confounded. More specifically, introduction of the random effects can inflate the variance of the posterior distribution of β. This is because the traditional model implicitly contains predictors that are collinear with X, and this linearity causes the variance inflation (Reich et al., 2006).

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.

Setting $\gamma = 0$, the first stage of the RHS model is given by

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

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

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

where $Q_R = L^T Q L$.

Both of the traditional SGLMM and RHZ model share the same approach to that of the automodel. The traditional SGLMM model makes use of synthetic predictors K and L, and the RHZ model makes use of L. The traditional SGLMM is analogous to the uncentred automodel, and the RHZ model is analogous to the centred automodel.

4. Sparse reparameterization of the areal SGLMM

5. Dimension reduction for spatial models

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.

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.