Honors Thesis

September 20th, 2020

1 Introduction

Spatial data are of interest in a wide range of domains, for example, modeling animal populations in ecology or modeling disease prevalence in epidemiology, as in this study of Pertussis prevalence in Minnesota (Iroh et al, 2016). When working with such data, researchers often must take into account the spatial dependence of such data; that observations that are close in space are likely to be more similar than observations that are far apart in space. This violates the independence assumption critical to many common statistical models, such as generalized linear models. Fitting models that assume independence with dependent data leads to incorrect standard error calculations, which can, for example, provide misleading results when performing hypothesis tests. Data collected over time also often suffer from the same issue, where observations close in time are likely to be more similar than observations far apart in time. To make things more difficult, many data contexts have data collected both over time and space. This means we must not only account for spatial and temporal dependence, but may also need to account for the ways in which these dependence structures change over time, adding further complexity to modeling such data.

With this model complexity often comes computational expense. Because each location and time can be viewed as an additional variable observed about a single data point, spatiotemporal data is often very high dimensional. This causes several computational challenges that make computing models for even moderately sized data, in the range of 1000 to 10000 observations take prohibitively long. This includes large matrix operations as well as strong correlations between random effects, which slows mixing of Markov Chain Monte Carlo (MCMC) algorithms when using Bayesian inference (Guan and Haran, 2018; Bradley et al, 2015). Additionally, many methods that attempt to make such models computationally tractable do so by requiring the modeler to specify complicated ideas, that can make spatiotemporal approaches inaccessible to those with less statistical training or less computational resources. One prominent example of this in practice is the work of the United Church of Christ's seminal paper on racism in toxic waste facility siting decisions. In the original study, performed in 1987, and in their follow-up study in 2007, they were unable to appropriately account for the spatial dependence present in both facility

locations and populations that live near them due to lack of methods accessible to social scientists and activists not trained in statistics. These problems persist in more modern methods, such as the necessity of specifying the basis functions used to approximate spatial dependence in Spatio-Temporal Statistics with R book (Wikle et al, 2019?).

One recent technique that attempts to account for spatial dependence while prioritizing both computational efficiency and specification simplicity is the work of Guan and Haran (2018) in their paper "A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models." Here, the covariance structure of the data is projected onto a lower-dimensional space, allowing for speedier computation of matrix operations, while additionally decorrelating random effects, allowing for faster mixing of MCMC methods. I extend this method into the temporal domain while retaining many of the computational advantages as well as a relatively simple model specification, making the model more accessible to non-experts. I then apply this model to both simulated and real-world data to demonstrate its computational efficiency as well as its accuracy.

2 Existing Methods

2.1 Spatial Gaussian Processes and the SGLMM

One common approach to spatial modeling treats the data as a Gaussian process realized at observed locations. A Gaussian process is a random process such that every finite set of random variables drawn from it, here observed locations, follow a multivariate normal distribution. Following the discussion in (Banerjee, Carlin, and Gelfand, 2015; Zhang 2002) For any spatial location s, let $Y(s)$ be the response variable and $\boldsymbol{x(s)}$ be a $p \times 1$ vector of explanatory variables. Let $Y(s)$ be modeled by

$$Y(s) = \boldsymbol{x}(s)^T \boldsymbol{\beta}(s) + e(s)$$

Where $\boldsymbol{\beta}(s)$ is a $p \times 1$ vector of slopes, which are often assumed to be identical for each location, leaving us with $\boldsymbol{x(s)}^T \boldsymbol{\beta}$. $e(s)$ is the residual term of the model. In order to capture spatial dependence, we let

$$e(s) = w(s) + \epsilon(s)$$

where $\epsilon(s)$ is a Gaussian noise process and $w(s)$ is the spatial Gaussian process realized at location s such that, $w \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}(\phi))$. This lets w be a mean zero random effect with variance parameter σ and a $n \times n$ correlation matrix $\Sigma(\phi)$ for n observed locations. $\Sigma(\phi)$ is created from a correlation function $\rho(\cdot; \phi)$ and captures the spatial dependence. It is often based on the distance between two locations in space. Then, for locations i and j, with distance between them d_{ij},

$$\Sigma(\phi)_{ij} = \rho(d_{ij}; \phi)$$

The parameter ϕ in the correlation function controls the specific properties of the correlation function, such as the strength of correlation between two locations. If $\rho(\cdot;\phi)$ were such that the correlation between any two distinct locations were zero, ie. $\rho(d_{ij};\phi) = 0$ for $i \neq j$, our correlation matrix would be the identity, and our model would simplify to a linear model. Notice that because all of the spatial dependence is captured in the random variables $w(s)$, $Y(s)|w(s)$ is conditionally independent. For the spatial case, a common choice for $\rho(;\phi)$ is the Matérn covariance function. This function has two parameters, ϕ and ν. Only one of these parameters is identifiable (citation? its somewhere in Banerjee 2015), and because of this we often set ν to be 1.5 or 2.5, which work well in most applications.

This is sometimes referred to as a spatial linear mixed model due to the combination of fixed effects $\boldsymbol{x}(s)^T\boldsymbol{\beta}$ and random effects $w(s)$ with a Gaussian residual (Zhang 2002). This spatial model can be extended for non-Gaussian outcomes $Z(s)$ using the standard general linear model formulation, where

$$g(E[Z(s)|\beta, w(s)]) = \boldsymbol{x}(s)^T\boldsymbol{\beta} + w(s)$$
$$w \sim N(0, \sigma^2\Sigma(\phi))$$

For some link function $g(\cdot)$. This model structure is often referred to as a spatial generalized linear mixed model, or the SGLMM. One issue with the SGLMM is the high dimension of the random effects w. With the model above, we will have one random effect for each location. This leads to computational issues when fitting the model. For example, to calculate the likelihood, we must invert the covariance matrix. This corresponds to inverting a $n \times n$ matrix. Matrix inversion is a computationally expensive operation, with time complexity $O(n^3)$. For large n, this quickly becomes infeasible. Additionally, when fitting the model using Bayesian methods, the correlation among the high dimensional mixed effects lowers the efficiency of MCMC methods, meaning we will need to evaluate the likelihood more times in order to get an accurate estimate of the posterior distribution of parameters.

2.2 Existing Approaches to the SGLMM

Many approaches have been developed to get around these computational issues. In the Gaussian case, it is possible to integrate out the random effects, which is referred to as marginalization. This allows for simpler model fitting, however this approach fails for the non-Gaussian case (Wikel 2019).

One popular choice for dealing with the compuational issues is the predictive process model (Banerjee et al, 2008). The predictive process replaces $w(s)$ with $\tilde{w}(s)$, where for some small subset of locations S^* (called knots), we let $w^*(s)$ be the realization of the full spatial Gaussian process at the points in $S*$. We are then left with $w^* \sim N(0, Cov^*(\phi))$, which has a much smaller dimension covariance matrix. Letting $c(s_0;\phi)$ be the evaluation of the covariance function used to generate Cov^* at some point s_0 with all of the knots, we have

$$\tilde{w}(s_0) = c^T(s_0; w^*)C^{*-1}w^*$$

Thus, when the size of the set of knots is much smaller than the original number of locations, a much smaller dimension covariance matrix must be inverted.

TODO: Talk about drawbacks of predictive process

TODO: Talk about other low rank methods. (Basis functions, Moran's I, displaced centroids in Prates?)

2.3 Spatiotemporal Gaussian Processes

The spatial Gaussian process model can be extended into the spatiotemporal domain. Let $Y(s,t)$ be the response variable at location s and time t. Let there be n locations and m times for each location. Hence, $\boldsymbol{Y}$ is $mn \times 1$. Given a $p \times 1$ vector of explanatory variables $\boldsymbol{x}(s,t)$, we have

$$Y(s,t) = \boldsymbol{x}(s,t)\boldsymbol{\beta}(s,t) + e(s,t)$$
$$e(s,t) = w(s,t) + \epsilon(s,t)$$

As before, we often let $\boldsymbol{\beta}(s,t) = \boldsymbol{\beta}$ to simplify the model. There are several ways to handle the spatiotemporal process variable $w(s,t)$. The simplest choice is to let $w(s,t) = w(s) + \alpha(t)$. Then, our spatial effects $w(s)$ are independent of our temporal effects $\alpha(t)$. This simplifies the model fitting process. If we let both $w(s)$ and $\alpha(t)$ to be mean zero Gaussian processes, the full model can be written out hierarchically as

$$Y(s,t) = \mu(x,t) + w(s,t) + \epsilon(s,t)$$
$$w(s,t) = \alpha(t) + w(s)$$
$$w \sim N(0, \sigma^2\boldsymbol{\Sigma}(\phi))$$
$$\alpha \sim N(0, \theta^2\boldsymbol{\Theta}(\psi))$$

Where θ is the variance parameter of the temporal effects and $\boldsymbol{\Theta}$ is the $m \times m$ temporal correlation matrix created by some correlation function $\upsilon(\cdot;\psi)$, which may or may not be the same as the spatial correlation function $\rho(\cdot;\phi)$. As in the spatial case we can extend this to the general linear mixed model formulation by replacing the first equation with $g(E[Z(s,t)|\beta, w(s,t)]) = \boldsymbol{x}(s)^T\boldsymbol{\beta} + w(s,t)$ for some link function $g(\cdot)$. Other choices for $w(s,t)$ include letting $w(s,t) = (s)\alpha(t)$, which maintains independence of spatial and temporal effects. We might also let $w(s,t) = w_t(s)$ so that for each time t we model different spatial random effects $w_t(s)$, or analogously $w(s,t) = \alpha_s(t)$, so that for each location s we model different temporal random effects $\alpha_s(t)$. These two models allow either the spatial or the temporal effect to vary across time or space respectively, but still preclude spatial and temporal interaction. An attractive approach for spatiotemporal interaction is the dynamic spatiotemporal model (Banerjee 2015

or Wikel 2019 or the original paper?). However, this model is beyond the scope of this paper. All of these models suffer from similar computational difficulties as in the spatial case, where likelihood evaluation is computationally intensive, and for Bayesian fitting correlated mixed effects lead to a lower MCMC mixing efficiency.

2.4 Existing approaches to Spatiotemporal Gaussian Processes

Many approaches have been developed to circumvent these issues. For example, to ease MCMC mixing, Bradley et al. (2018) develop a model for count valued observations which uses clever choices of prior distributions and parameter distributions to allow the construction of a Gibbs sample for the model. This sidesteps difficulties that arise when picking proposal distributions for the high dimensional mixed effects that occur when fitting models using the Metropolis Hasting algorithm (Rue 2009). However, computational issues around calculating the likelihood for the high dimensional random effects remain. The model can be extended to use basis function approximations popular in spatial modeling, but this still leaves the difficulty of determining the appropriate basis functions for a given application. Other methods focus on low-rank methods analogous to those in the spatial frameworks to reduce computational load, such as Bradely et al's (2015) use of the Moran's I basis functions approach. However, the Moran's I approach only works for areal data. The predictive process can be extended to the spatiotemporal domain, allowing for modeling of point referenced data, but then the issues of selecting the appropriate knots (locations) from which to approximate the mixed effects remains (idk the citation for this). (Could also talk about INLA here).

2.5 Random Projections for the SGLMM

One approach of interest is the random projections based approach to dimension reduction of the spatial effects in an SGLMM from Guan and Haran (2018). This is a low rank approach to the SGLMM which differs from the basis function or predictive process approach in that neither basis functions nor knots need be specified. Instead, this model relies on decorrelating the random effects using the eigendecomposition of their covariance matrix. Calculating the eigendecomposition of a matrix is just as computationally expensive as a matrix inversion, so they rely on the random projections algorithm () to quickly calculate an approximation of the eigendecomposition. To see how this works, recall the SGLMM framework:

$$g(E[Z(s)|\beta, w(s)]) = \boldsymbol{x}(s)^T\boldsymbol{\beta} + w(s)$$
$$\boldsymbol{w} \sim N(0, \sigma^2\boldsymbol{\Sigma}(\phi))$$

Note that the covariance matrix of $\boldsymbol{w}$, $\sigma^2\Sigma(\phi)$, is symmetric and positive semi-definite. This is true for all covariance matrices, and means that the eigen-

decomposition $\sigma^2\Sigma(\phi) = V\Lambda V^T$ exists for real valued diagonal $n \times n$ matrix Λ consisting of the eigenvalues of $\sigma^2\Sigma(\phi)$ arranged in descending magnitude and real valued orthonormal $n \times n$ matrix V, consisting of the eigenvectors of $\sigma^2\boldsymbol{\Sigma}(\phi)$. Consider the synthetic random variable

$$\boldsymbol{\delta} = (\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2})^T\boldsymbol{w}$$

Then, leveraging the fact that because V is orthonormal, $V^{-1} = V^T$, the covariance matrix of δ is given by

$$\begin{aligned} Cov(\boldsymbol{\delta}) &= Cov((\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2})^T\boldsymbol{w}) \\ &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{V}Cov(\boldsymbol{w})\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2} \\ &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2} \\ &= \boldsymbol{I} \end{aligned}$$

Hence, $\delta \sim N(0, I)$, meaning it is easy to calculate the likelihood of δ. Then, we have transferred the computational complexity of this likelihood calculation into of the eigendecomposition for $\sigma^2\Sigma(\phi)$. This issue is resolved through the use of the random projections algorithm, which creates a low rank approximation of the eigendecomposition of $\sigma^2\Sigma(\phi)$. (Ask Alicia how much detail to go into here.)

To use the low rank approximation created by the random projections algorithm, for some $k << n$, we take the approximated leading k eigenvectors and eigenvalues of $\sigma^2\Sigma(\phi)$, which we'll call $\tilde{\boldsymbol{V}}$, which is now $n \times k$ and $\tilde{\boldsymbol{\Lambda}}$, which is now $k \times k$. If we now let $\boldsymbol{\delta} = (\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}^{-1/2})^T\boldsymbol{w}$, we can model the SGLMM using the following specification

$$\begin{aligned} g(E[Z(s)|\beta, w(s)]) &= \boldsymbol{x}(s)^T\boldsymbol{\beta} + (\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}^{-1/2})_s\boldsymbol{\delta} \\ \boldsymbol{\delta} &\sim N(0, \boldsymbol{I}) \end{aligned}$$

Notice how $(\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}^{-1/2})\boldsymbol{\delta} = (\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}^{-1/2})(\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}^{-1/2})^T\boldsymbol{w} = \boldsymbol{w}$, so we are still modeling our original spatial mixed effect $\boldsymbol{W}$, only using the decorrelated synthetic variable $\boldsymbol{\delta}$. This model has a number of advantages. Decorrelation of random effects works to increase the rate of MCMC mixing. The model interpretation is almost unchanged from the original spatial Gaussian process model, only now our random effects are approximations. And, as opposed to choosing basis functions to approximate the random effects or knot locations to use, we simply need to specify the rank of the approximation. If the rank is sufficiently small, computational issues of calculating the likelihood disappear because the random projections algorithm has time complexity on the order of $O(nk^2)$ as opposed to the original $O(n^3)$ time complexity of matrix inversion. Choosing a rank close to 50 seems to work well in most applications while maintainting computational efficiency.

3 Contributions

The Guan and Haran (2018) model can naturally be extended into the spatiotemporal case. Recall the spatiotemporal model with independent spatial and temporal effects given by

$$\begin{aligned} Y(s,t) &= \mu(x,t) + w(s,t) + \epsilon(s,t) \\ w(s,t) &= \alpha(t) + w(s) \\ w &\sim N(0, \sigma^2 \boldsymbol{\Sigma}(\phi)) \\ \alpha &\sim N(0, \theta^2 \boldsymbol{\Theta}(\psi)) \end{aligned}$$

Because the spatial and temporal effects are independent, we can apply the random projections algorithm to one or both of them, depending on the dimension of the temporal and spatial data. Then, our model would become

$$\begin{aligned} Y(s,t) &= \mu(x,t) + w(s,t) + \epsilon(s,t) \\ w(s,t) &= (\boldsymbol{U}\boldsymbol{K}^{1/2})_t \boldsymbol{\gamma} + (\boldsymbol{V}\boldsymbol{\Lambda}^{1/2})_s \boldsymbol{\delta} \\ \boldsymbol{\delta} &\sim N(0, I) \\ \boldsymbol{\gamma} &\sim N(0, I) \end{aligned}$$

Where $\boldsymbol{\delta}, \boldsymbol{V}, \boldsymbol{\Lambda}$ are the same as above. The temporal effects are decorrelated in the same way as the spatial effects using $\boldsymbol{\gamma} = (UK^{-1/2})^T \boldsymbol{\alpha}$, where the $m \times m$ orthonormal matrix U and $m \times m$ diagonal matrix K come from the eigendecompositon of the temporal covariance matrix such that $\theta^2 \boldsymbol{\Theta}(\boldsymbol{\psi}) = \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}^T$. Again, to lower the dimension of these effects, would would choose $k << n$ and $l << m$ such that $\tilde{\boldsymbol{V}}$ is $n \times k$, $\tilde{\boldsymbol{\Lambda}}$ is $k \times k$, $\tilde{\boldsymbol{U}}$ is $m \times l$, $\tilde{\boldsymbol{K}}$ is $l \times l$. Again, to model the generalized linear mixed model version, replace the first equation with $g(E[Z(s,t)|\beta, w(s,t)]) = \boldsymbol{x}(s)^T \boldsymbol{\beta} + w(s,t)$.

TODO: Discuss extension to other seperable spatiotemporal effect specifcations, and maybe mention usage in dynamic spatiotemporal model

4 Applications

5 Discussion and Future Work