Multilevel Modeling for Legacy Data

Jiehua Chen

The general model of the multilevel 3-D spatial model is stated as following:

$$Y(s,d) = X(s,d)\beta + S_{\theta,\sigma_\alpha^2}(s,d) + \epsilon(s,d), \tag{1}$$

where s is the point location, d is the depth, and $X(s,d)$ are covariates for each point. $S(s,d)$ is the 3-D spatial process, and $\epsilon(s,d)$ is the random measurement error with mean 0, and variance σ^2, for each observation.

When both $S(s,d)$ and $\epsilon(s,d)$ are assumed to be Gaussian distributed, most of the posterior distributions of the unknown parameters have closed forms. In our implementation, $S(s,d)$ are assumed to be the same for all measurements at one location, so it will be written as $S(s)$ in the following sections. $S(s)$ is also assumed to be a Gaussian spatial process, i.e.:

$$S(\mathbf{s}) \sim N(\mathbf{0}, \sigma_\alpha^2 \Sigma_\theta), \tag{2}$$

where Σ_θ is a spatial correlation matrix.

The following MCMC procedure uses the hybrid of Hastings algorithm and Gibbs sampling to estimate unknown parameters in model (1):

- the mean coefficients with uninformative Gaussian priors:

$$\beta \mid \theta, \sigma^2, \sigma_\alpha^2 \sim N\left((X^T(\sigma^2 I + \sigma_\alpha^2\Sigma_\theta)^{-1}X)^{-1}X^T(\sigma^2 I + \sigma_\alpha^2\Sigma_\theta)^{-1}Y, (X^T(\sigma^2 I + \sigma_\alpha^2\Sigma_\theta)^{-1}X)^{-1}\right);$$

- the spatial process with uninformative Gaussian priors:

$$\begin{aligned} S_{\theta,\sigma_\alpha^2} \mid \beta, \theta, \sigma_\alpha^2, \sigma^2 &\sim N(\mu_S, \Sigma_S) \\ \mu_S &= \Sigma_S(\bar{Y}_1 - \beta\bar{X}_1, \cdots, \bar{Y}_K - \beta\bar{X}_K)^T \\ \Sigma_S &= \left((\sigma_\alpha^2)^{-1}\Sigma_\theta^{-1} + \sigma^{-2}\begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_K \end{pmatrix}\right)^{-1}, \end{aligned}$$

where K is the number of unique locations, n_i is number of observations at the ith location.

- the variance parameters with prior $\sigma^2 \sim \chi^{-1}(a_\sigma, b_\sigma)$, and $\sigma_\alpha^2 \sim \chi^{-1}(a_{\sigma_\alpha}, b_{\sigma_\alpha})$:

$$\begin{aligned} \sigma^2 \mid \beta, S &\sim \chi^{-1}\left(a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}\sum (Y - X\beta - S)^2\right) \\ \sigma_\alpha^2 \mid \theta, S &\sim \chi^{-1}\left(a_{\sigma_\alpha} + \frac{K}{2}, b_{\sigma_\alpha} + \frac{1}{2} S^T \Sigma_\theta^{-1} S)\right) \end{aligned}$$

- the posterior distribution of the range parameter with uniform prior does not have a closed form, so Hastings algorithm is implemented:

$$P(\theta \mid S, \sigma_\alpha^2) \propto |\Sigma_\theta|^{-1} \exp\left(-\frac{1}{\sigma_\alpha^2} S^T \Sigma_\theta^{-1} S\right).$$

The biggest challenge for MCMC aforementioned steps is the inverse of Σ_θ, when the dataset is large. We implemented two approaches: tapered covariance model, and spatial kernel model to handle the computation issues, when short, and long range of spatial correlation exist in the data respectively. In the current computation, it seems that tapered covariance matrix models can handle the computation better.

1 Tapered Covariance Matrix

When the spatial correlations only exit within a short distance, we set Σ_θ to be a tapered covariance matrix. The tapered covariance function $V_\theta(s_1, s_2)$ between locations s_1, and s_2 is usually constructed in the following way:

$$V_\theta(s_1, s_2) = C_\theta(s_1, s_2) T(s_1, s_2), \tag{3}$$

where $C_\theta(s_1, s_2)$ can be any spatial covariance function with unknown range parameters θ, and $T(s_1, s_2)$ is a covariance function, which is zero when s_1 and s_2 are farther away than a prespecified distance. According to Schur product theorem, $V_\theta(s_1, s_2)$ is still a legitimate covariance function (i.e., the corresponding covariance matrix is still positive semidefinite). The tapered covariance function has the advantage of being able to create a sparse covariance matrix, and be flexible of modeling short-distance correlations. The

main disadvantage of the tapered covariance function is that it ignores the possible long range of correlation, because it assumes that correlation is zero for the points farther away than a prespecified distance, and the prespecified distance is large, we will not gain much computation efficiency.

In the MCMC estimation code, we first implemented DBSCAN (density-based spatial clustering of applications with noise) algorithm to cluster the profile locations, such that points of different clusters are always farther away than the prespecified distance. Then instead of having a big covariance matrix, we can rearrange the points so that the tapered covariance matrix Σ_θ is a diagonal block matrix, then that parallel computation can be implemented for each of the diagonal blocks.

The following is the Bayesian kriging prediction for model (1):

- the posterior distribution of the predicted spatial process at the prediction location s_0 is

$$S(s_0) \mid S, \theta, \sigma_\alpha^2 \sim N(c(\theta)\Sigma_\theta^{-1}S, c(\theta)\Sigma_\theta^{-1}c^T(\theta)),$$

  where $c(\theta)$ is a vector of estimated correlations between the prediction location s_0, and all the observed locations.

- the posterior distribution of the predicted property $Y(s_0, d)$ at depth d is:

$$Y(s_0, d) \mid d, S(s_0), \beta, \sigma^2 \sim N(X_d\beta + S(s_0), \sigma^2).$$

2 Spatial Kernel

When the range of spatial correlation is large, therefore tapered covariance function cannot handle the computation any more, we implemented the Bayesian spatial kernel model.

A Bayesian spatial kernel model is defined as follows: At location $\mathbf{s_i}$, and kernel grid locations $\mathbf{g_j}$, $j = 1, \cdots, k$, the data $Y(s_i)$ at that location is modeled by the equation

$$Y(\mathbf{s_i}) = X(\mathbf{s_i})\beta + \sum_{j=1}^{k} \mathcal{K}_\phi(\mathbf{s_i}, \mathbf{g_j})w(\mathbf{g_j}) + \epsilon(\mathbf{s_i}), \tag{4}$$

where

- $\mathcal{K}_\phi(\mathbf{s_i}, \mathbf{g_j})$ is the kernel function between locations $\mathbf{s_i}$, and $\mathbf{g_j}$ with the unknown parameters ϕ. This kernel function can be the matern kernel, spherical kernel, or Gaussian kernel, as determined by goodness of model fitting;
- $w_{\mathbf{g_j}} \sim \mathcal{N}(0, \sigma_w^2)$, $j = 1, \cdots, k$ are random effects; and
- $\epsilon(\mathbf{s}) \sim \mathcal{N}(0, \sigma^2)$ is the residual.

In model (4), the spatial covariance matrix of data points $Y(\mathbf{s_i})$, $i = 1, \cdots, n$, conditioned on the covariates X is

$$\sigma^2 I + \sigma_w^2 \mathbf{K}^T \mathbf{K},$$

where K is the $n \times k$ spatial kernel matrix. When $k \ll n$, model (4) provides significant savings in memory by using Sherman-Morrison-Woodbury formula to calculate the matrix inverse. In theory, the spacing of the kernel centers g_j should not be larger than the range of the spatial correlation, in order to approximate the covariance matrix well, therefore, the kernel model will not work well, when the spatial correlation only exits within a short distance.