STAT 8025

Lecture 6: Methods for Analyzing Large Spatial Datasets

Dr. Emily Lei Kang

Division of Statistics & Data Science
Department of Mathematical Sciences
University of Cincinnati

UNIVERSITY OF Cincinnati

Introduction

- Nowadays spatial data can be large, thousands of, or even millions of data
- Kriging may become infeasible computationally. Can you think about the reasons?
  - Kriging equation

$$\hat{Y}_{OK}(s_0) = \left\{ c(s_0) + 1 \frac{1 - 1'\Sigma^{-1}c(s_0)}{1'\Sigma^{-1}1} \right\}' \Sigma^{-1} Y$$

  - Likelihood

$$l(\beta, \Theta) = -\frac{1}{2}\log|\Sigma(\Theta)| - \frac{1}{2}(Y - X\beta)'\Sigma(\Theta)^{-1}(Y - X\beta) + \text{const.}$$

  - Bayesian inference

$$\begin{aligned} p(Y(s_0)|Y) &= \int p(Y(s_0), \beta, \Theta|Y) d\beta d\Theta \\ &= \int p(Y(s_0)|\beta, \Theta, Y) p(\beta, \Theta|Y) d\beta d\Theta \end{aligned}$$

- This is a research area with a lot work done in the past years.

In general, the methods proposed to handle large spatial data are based on or use some of the components below:

- Local analysis
- Dimension reduction
- Sparse matrices

Local Analysis

- Local kriging:
  - Say our objective is to make a prediction at location s_0
  - We only use data within certain distance of s_0 to estimate parameters and make the prediction.
  - Varying this ‘window’, when we move to the next prediction location
- Recent update: Local Approximate Gaussian Process (Gramacy and Apley, 2015)

- Advantages
  - Small dataset and small matrices
  - Fully parallelizable
  - No need to assume stationarity
- Potential limitations
  - Inefficient if the process is stationary (or stationary in a large subregion)
  - not resulting in a valid global model; global predictive covariance is unavailable
  - the resulting prediction surface not necessarily smooth (though this may not be visualized in practice)

- Several methods fall into the category of low-rank methods, including Fixed rank kriging (FRK, Cressie and Johanneson, 2006, 2008), predictive process (PP, Banerjee et al., 2008; Finley et al., 2009)
- The general formulation is:

$$Y(\mathrm{s}) = \mu(\mathrm{s}) + \sum_{i=1}^{r} B_i(\mathrm{s}) w_i + \xi(\mathrm{s})$$

  - $B_i(\mathrm{s})$ are basis functions
  - $\mathrm{w} = (w_1, \ldots, w_r)' \sim \mathcal{N}_r(0, \Sigma)$, random effects
  - $\xi(\mathrm{s})$ spatially independent, nugget effect with variance τ^2
- Different low-rank methods specify these components in different ways.

- Comparison
  - PP is flexible and can be put into hierarchical models for multivariate, spatially varying coefficient, non-stationary etc.
  - Basis functions in FRK are fully specified and don't need to be updated in MCMC, saving computation time compared to PP
  - FRK is semiparametric and can handle nonstationary data
- Limitations:
  - The prediction surface can be oversmoothed

Extensions of PP and FRK:

- Full scale approximation (FSA; Sang and Huang, 2012)
- Multiresolution approximation (MRA; Katzfuss, 2017)
- PP has also been used in multivariate analysis and non-Gaussian spatial data analysis
- FRK has been extended to handle data fusion
- Built upon FRK, Ma and Kang (2019) proposed a model Fused Gaussian Process (FGP) that can give robust predictive performance

LatticeKrig

- LatticeKrig (Nychka et al., 2015) follows the basis-function representation:

$$Y(s) = \mu(s) + \sum_{i=1}^{r} B_i(s) w_i$$

  - Multiresolutional basis functions, r can even be larger than n
  - w_i's from different resolutions are independent
  - w_i's from the same resolution follows a multivariate normal distribution with *sparse* precision matrix
  - Many parameters are pre-specified or constrained, leaving a single free parameter
  - R package `LatticeKrig`

Exploring Sparse Matrix Operations

- A matrix is said to be sparse if it has a lot of zeros
- There are special computational methods to handle sparse matrices. For example, a sparse matrix can be inverted much faster
- Commonly used spatial covariance function don't result in sparse matrices, e.g., Matérn
- Methods to handle large spatial data:
  - Tapering
  - Stochastic partial differential equation approach (SPDE)

Tapering

- Forcing the covariance matrix to be sparse

$$C^*(\cdot,\cdot;\theta,\gamma) = C(\cdot,\cdot;\theta)C_{taper}(\cdot,\cdot;\gamma)$$

where C_{taper} is an isotropic correlation function equal to zero when the distance is outside of a range denoted by γ

- The resulting covariance matrix is

$$\Sigma^* = \Sigma \odot \mathrm{T}$$

where $\odot$ denotes the element-wise product, "Hadamard" product.

- T and the resulting Σ^* are sparse.
- Tapering functions:
  - Wendland: $C_{taper} = (1 - \frac{|h|}{\gamma})^4(1 + 4\frac{|h|}{\gamma})\mathbb{I}(|h| < \gamma)$
  - Spatially adaptive covariance tapering (Bolin and Wallin, 2016)

- There are work investigating theoretical properties related to covariance tapering
- There are two approaches for parameter estimation, one- and two-taper approaches

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^{n} |\Sigma|^{-1/2} etr\left\{-\frac{1}{2}(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})'\Sigma^{-1}\right\}$$

  - $etr(A) = exp(trace(A))$
  - One-taper: $\Sigma \odot T$ replaces Σ
  - Two-taper: $\Sigma \odot T$ replaces Σ AND $(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})' \odot T$ replaces $(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})'$
  - Unbiased parameter estimators from two-taper but severe loss of computational efficiency
- When $n > 50,000$, computation can become slow

SPDE

- SPDE is based on the equivalence between the Matérn covariance fields and stochastic PDEs.
- Based on a triangulation of the domain (which can be dense), the resulting precision matrix Q is a sparse matrix dependent on the covariance parameters and the triangulation (adjacency matrix)
- SPDE is implemented via the R package `INLA` for Bayesian generalized linear models, but empirical Bayes estimate of the covariance parameters are used (estimated via optimization)
- For data not on a regular grid, choose local linear functions and perform interpolation.

Divide-and-Conquer Methods

- Dividing the data sets
  - Metakriging (Guhaniyogi and Banerjee, 2018): approximate the posterior predictive distribution by the subset posterior predictive distributions
- Dividing the spatial domain
  - Heaton et al. (2017) Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences
  - Konomi et al. (2014) Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations: partition + FSA
  - Konomi et al. (2019) Computationally efficient nonstationary nearest neighbor Gaussian process models using data-driven techniques: partition + NNGP

Other Methods

- Construct a model based on neighborhood structure/hierarchical spatial-resolution partition/network structure: conditional distribution

Summary

- Methods for large spatial datasets

Preview:

- FGP and two applications