Statistical Methods for Large Spatial Datasets

Ying Sun

Department of Statistics

The Ohio State University
Large Datasets Problem

- Large observational and computer-generated datasets:
  - Often have spatial and temporal aspects.
  - Nearly global coverage.
  - High resolutions.

- Examples:
  - Satellite measurements.
  - Computer model outputs.

- Goal:
  - Make inference on underlying spatial processes from observations at $n$ locations where $n$ is large.
Gaussian Processes

- Gaussian process models can be used to
  - describe the spatial variability in the process.
  - predict unobserved values of the process, and provide prediction uncertainties.
  - serve as a building block for more complex models.

- Gaussian process $Z$ on a domain $D \subset \mathbb{R}^d$ is fully specified by
  - $\mu(x) = E\{Z(x)\}$, and
  - $K(x, y) = \text{cov}\{Z(x), Z(y)\}$, for all $x, y \in D$.

- Make inferences:
  - Estimation: $\mu$ and $K$ when specified up to $\theta \in \mathbb{R}^p$.
  - Prediction: kriging.

- Methods:
  - Likelihood-based methods.
  - Bayesian approaches.
Maximum Likelihood Estimation

Suppose data $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ is observed from a Gaussian random field $Z \sim GP(0, K(h; \theta))$ at $n$ irregularly spaced locations.

- **Goal:** estimate $\theta \in \mathbb{R}^p$ by likelihood methods.

- **Loglikelihood:**

  $$\ell(\theta) = -\frac{1}{2} \mathbf{Z}^T \Sigma_{nn}^{-1}(\theta) \mathbf{Z} - \frac{1}{2} \log |\Sigma_{nn}(\theta)|.$$  

- **Score equations:**

  $$\mathbf{Z}^T \Sigma^{-1} \Sigma_i \Sigma^{-1} \mathbf{Z} - \text{tr}(\Sigma^{-1} \Sigma_i) = 0, \quad i = 1, \ldots, p,$$

  where $\Sigma_i = \partial \Sigma(\theta)/\partial \theta_i$.

- **The standard way:**

  - Cholesky decomposition of $\Sigma_{nn}$.
  - Generally requires $O(n^3)$ computations and $O(n^2)$ memory.

- **The covariance matrix $\Sigma_{nn}$ is**

  - large: $n \times n$ for $n$ locations.
  - unstructured: irregular spaced locations.
  - dense: non-negligible correlations.
Large $n$

- Options for large $n$:
  - Use models that reduce computations and/or storage.
  - Use approximate methods.
  - Both.

- Models that might allow for exact computations:
  - Compactly supported covariance functions.
  - Reduced rank covariance functions.
  - Markov models.

- Approximation methods:
  - Approximating likelihoods: obtain approximate functions to be maximized.
  - Approximating score equations: yield biased/unbiased estimating equations.

- Statistical and computational efficiency:
  - Exact computations.
  - Approximation methods.
1. Tapering
2. Low Rank Approximations
3. Low Rank + Tapering and Multi-resolution Models
4. Markov Models
5. Likelihood and Score Equation Approximations
6. Multivariate Spatial Data and Space-time Data
7. Methods in Numerical Analysis
8. Discussion
Covariance tapering:

\[ \tilde{K}(h; \theta) = K(h; \theta) \circ T(h; \gamma), \]

- \( T(h; \gamma) \): an isotropic correlation function of compact support, i.e., \( T(h; \gamma) = 0 \) for \( h \geq \gamma \).

Assumptions:

- The covariance function has compact support.
- Its range is sufficiently small.

The tapered covariance matrix \( \tilde{K} \):

- Retains the property of positive definiteness.
- Zero at large distances.
- Minimal distortion to \( K \) for nearby locations.
- Efficient sparse matrix algorithms can be used.
- Also saves storage.
How much statistical efficiency is lost?

Estimation: properties of the MLEs.

- Kaufman et al. (2008), JASA: proposed biased and unbiased estimating equations with tapered covariance matrices.
- Stein (2014) JCGS: studied the statistical properties of isotropic covariance tapers and showed numerically that independent blocks are usually better.

Prediction: spatial interpolation using kriging with known covariance functions.

- Furrer et al. (2006), JCGS: proposed covariance tapering for kriging and studied the properties of the resulting MSPE.

Open questions:

- Tapers for nonstationary processes.
- Anisotropic tapers.
- Multivariate tapers: need compact supported cross-covariance functions.
Find reduced rank covariance function representation:

- Banerjee et al. (2008), JRSSB: proposed Gaussian predictive processes $\tilde{\omega}(s)$ to replace $\omega(s)$ in

$$Z(s) = x^T(s)\beta + \omega(s) + \epsilon(s),$$

by projecting $\omega(s)$ onto a $m$-dimension (lower) subspace

$$\tilde{\omega}(s) = E(\omega(s)|\omega(x_1), \ldots, \omega(x_m)).$$

- Cressie and Johannesson (2008), JRSSB: proposed fixed rank kriging by defining a spatial random effect model:

$$\omega(s) = B^T(s)\eta,$$

where $B$ is a vector consisting of $m$ basis functions and $\text{var}(\eta) = G$.

Have computational advantages but also limitations. (Stein, 2013, Spatial Statistics).
Combinations

- **Low rank+tapering**: Sang and Huang (2011), JRSSB
  - A reduced rank process for large-scale dependence: low rank approximation.
  - A residual process for small-scale dependence: covariance tapering.

- **Multi-resolution models**: Nychka et al. (2013), Manuscript
  - The basis functions at each level of resolution are constructed using a compactly supported correlation function with the nodes arranged on a rectangular grid.
  - Numerically, it gives a good approximation to the Matérn covariance function.
Markov Models

- **Markov models**
  - The conditional distributions only depend on nearby neighbors.
  - Lead to sparseness of the precision matrix, the inverse of the covariance matrix.
  - Computational cost: $O(n^{3/2})$.

- **Gaussian Markov Random Fields:**
  - **Rue et al. (2009), JRSSB:**
    - Proposed integrated nested Laplace approximation (INLA).
    - Studies the computational gains for latent Gaussian field models in Bayesian inference.
  - **Lindgren et al. (2011), JRSSB:**
    - Represented a GRF with Matérn covariance function as the solution of a particular type of SPDE.
    - Proposed an approach to find GMRFs with local neighborhood and precision matrix to represent certain Gaussian random fields with Matérn covariance structure.
Likelihood approximation

- **Spatial domain**: Stein et al. (2004), JRSSB
  - Used the composite likelihood method (Vecchia, 1998) to approximate REML.
  - Joint density: product of conditional densities.
  - Condition on only subset of the “past” observations.

- **Spectral domain**: Fuentes (2007), JASA
  - A version of Whittle’s approximation (1954) for irregularly spaced data by introducing a lattice process.

- **Score equation approximation**: estimating equations.
  - Kaufman et al. (2008), JASA: sparse covariance matrix approximation.
  - Sun and Stein (2013), Manuscript: sparse precision matrix approximation.
Multivariate spatial data: Furrer and Genton (2011), Biometrika

- Proposed aggregation-cokriging.
- Based on a linear aggregation of the covariables.
- The secondary variables are weighted by the strength of their correlation with the location of interest.
- The prediction is then performed using a simple cokriging approach with the primary variable and the aggregated secondary variables.

Space-time Data: Genton (2007), Environmetrics

- Separable covariance structure approximation.
- To identify two small matrices that minimize the Frobenius norm of the difference between the original covariance matrix and the Kronecker product of those two matrices.
Iterative methods: solve $\Sigma x = Z$.

- $x_k \rightarrow x_{k+1}$, then check residuals.
- For positive definite $\Sigma \Leftrightarrow$ minimizing $f(x) = \frac{1}{2}x^T\Sigma x - x^TZ$.
- Can be solved by conjugate gradient method.

Matrix-free:

- Never have to store an $n \times n$ matrix.
- Computation is becoming cheap much faster than memory.

Main computation: matrix-vector multiplication.

- Requires $O(n^2)$ for dense and unstructured matrices.
- This is fast, if
  - $\Sigma$ is sparse, or
  - $\Sigma$ has some exploitable structures (e.g., Toeplitz).

Let $m$ be the number of iterations:

$$O(n^2 \times m) \quad v. \quad O(n^3)$$
Computational Difficulties

- Loglikelihood:

\[
\ell(\theta) = -\frac{1}{2}Z^T \Sigma_{n \times n}(\theta) Z - \frac{1}{2} \log |\Sigma_{n \times n}(\theta)|.
\]

- Score equations:

\[
Z^T \Sigma^{-1} \Sigma_i \Sigma^{-1} Z - \text{tr}(\Sigma^{-1} \Sigma_i) = 0, \quad i = 1, \ldots, p,
\]

where \(\Sigma_i = \frac{\partial \Sigma(\theta)}{\partial \theta_i}\).

- Computing \(\Sigma^{-1} Z\): best done by solving systems \(\Sigma x = Z\).

- Loglikelihood:

  - Main computation is due to calculating \(\log |\Sigma|\).

- Score equations:

  - Need \(n\) solves to compute \(\text{tr}(\Sigma^{-1} \Sigma_i)\).
  - May not be any easier than computing \(\log |\Sigma|\).
Comparisons

- Sparse covariance matrix approximation:
  - Covariance tapering.
  - Assume $\Sigma$ is sparse.
  - $\Sigma^{-1}$ is not generally sparse.
- Approximating $\Sigma^{-1}$ by a sparse matrix:
  - No need to assume $\Sigma^{-1}$ is sparse everywhere in the computation.
- Markov random field models:
  - Assume $\Sigma^{-1}$ is actually sparse.
Discussion

- Low Rank Approximations
  - Cannot capture local dependence well.
  - How to improve it?

- Sparse Covariance Approximations
  - Distortion of the covariance matrix.
  - Other types of tapers?

- Markov Random Field Approximations
  - Sparse precision matrix.
  - Precision matrix approximation?

- Combine methods and learn from numerical analysis community.