Geostatistical Modeling for Large Data Sets: Low-rank methods

Whitney Huang

CLEMSON UNIVERSITY

wkhuang@clemson.edu

Clemson ENVR Group, September 23, 2020

Outline

Gaussian process (GP) geostatistics

Model:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s}), \qquad \boldsymbol{s} \in \mathcal{S} \subset \mathbb{R}^d$$

where

$$\mu(\boldsymbol{s}) = \mathbb{E}\left[Y(\boldsymbol{s})\right] = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta}$$

$$\{\eta(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathcal{S}} \sim GP\left(0, C\left(\cdot,\cdot\right)\right)$$

$$\varepsilon(\boldsymbol{s}) \overset{\text{i.i.d.}}{\sim} N(0,\tau^2) \quad \forall \boldsymbol{s} \in \mathcal{S}$$

Likelihood

When the process $Y(\boldsymbol{s})$ is observed at n locations, say, $\boldsymbol{s}_1, \cdots .\boldsymbol{s}_n$ i.e.,

$$\boldsymbol{Y} = \left(Y\left(\boldsymbol{s}_1\right), \cdots, Y\left(\boldsymbol{s}_n\right)\right)^{\mathrm{T}}$$

Log-likelihood:

$$\ell_n \propto -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{Y}}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{Y}}]^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_Y &= \left[\mathrm{Cov}\left\{Y\left(\boldsymbol{s}_i\right), Y\left(\boldsymbol{s}_j\right)\right\}\right]_{i,j=1,\cdots,n} \\ &= \left[C\left(\boldsymbol{s}_i, \boldsymbol{s}_j\right) + \tau^2 \mathbb{1}_{\{\boldsymbol{s}_i = \boldsymbol{s}_j\}}\right]_{i,j=1,\cdots,n}\end{aligned}$$

"Big n Problem" in geostatistics

- Modern environmental instruments have produced a wealth of space–time data $\Rightarrow$ n is big
- Evaluation of the likelihood function involves factorizing large covariance matrices that generally requires
  - $\mathcal{O}(n^3)$ operations
  - $\mathcal{O}(n^2)$ memory
- Modeling strategies are needed to deal with large spatial data set.
  - parameter estimation $\Rightarrow$ MLE, Bayesian
  - spatial interpolation $\Rightarrow$ Kriging
  - multivariate spatial data $(np \times np)$, spatio-temporal data $(nt \times nt)$

Outline

Low–rank approximation

Goal: want to approximate $\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ into a lower-dimensional structure.

Hierarchical Representation (assume zero mean)

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$$

$$\boldsymbol{Z} \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$$

- $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$, $r \ll n$
- $\boldsymbol{A}$ is the $(n \times r)$ expansion matrix that maps $\boldsymbol{Z}$ to $\boldsymbol{\eta}$
- Σ_ε and Σ_ξ are (usually assumed to be) diagonal
- $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi} \Rightarrow \eta(\boldsymbol{s}_i) = \sum_{j=1}^{r} a_{ij} Z_j + \xi(\boldsymbol{s}_i)$

Low–rank approximation

Goal: want to approximate $\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ into a lower-dimensional structure.

Hierarchical Representation (assume zero mean)

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$$

$$\boldsymbol{Z} \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$$

- $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$, $r \ll n$
- $\boldsymbol{A}$ is the $(n \times r)$ expansion matrix that maps $\boldsymbol{Z}$ to $\boldsymbol{\eta}$
- Σ_ε and Σ_ξ are (usually assumed to be) diagonal
- $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi} \Rightarrow \eta(\boldsymbol{s}_i) = \sum_{j=1}^{r} a_{ij} Z_j + \xi(\boldsymbol{s}_i)$

Low–rank approximation

Goal: want to approximate $\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ into a lower-dimensional structure.

Hierarchical Representation (assume zero mean)

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$$

$$\boldsymbol{Z} \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$$

- $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$, $r \ll n$
- $\boldsymbol{A}$ is the $(n \times r)$ expansion matrix that maps $\boldsymbol{Z}$ to $\boldsymbol{\eta}$
- $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ are (usually assumed to be) diagonal
- $\eta = \boldsymbol{A}\boldsymbol{Z} + \xi \Rightarrow \eta(\boldsymbol{s}_i) = \sum_{j=1}^{r} a_{ij} Z_j + \xi(\boldsymbol{s}_i)$

Low–rank approximation

Goal: want to approximate $\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ into a lower-dimensional structure.

Hierarchical Representation (assume zero mean)

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$$

$$\boldsymbol{Z} \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$$

- $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$, $r \ll n$
- $\boldsymbol{A}$ is the $(n \times r)$ expansion matrix that maps $\boldsymbol{Z}$ to $\boldsymbol{\eta}$
- $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ are (usually assumed to be) diagonal
- $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi} \Rightarrow \eta(\boldsymbol{s}_i) = \sum_{j=1}^{r} a_{ij} Z_j + \xi(\boldsymbol{s}_i)$

Computational savings using low–rank approximation

To carry out parameter estimation and spatial interpolation one need to compute

$$\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{V}\right)^{-1}$$

where $\boldsymbol{V}=\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}+\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$.

Sherman–Morrison–Woodbury formula

$(A+BCD)^{-1}=A^{-1}-A^{-1}B\left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}$

In low–rank approximation, we have

$$\left(\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{V}\right)^{-1}=\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{A}\left(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}^{-1}$$

Low–rank approximation: expansion matrix A

$$\underbrace{\begin{bmatrix} \eta(\boldsymbol{s}_1) \\ \eta(\boldsymbol{s}_2) \\ \vdots \\ \vdots \\ \vdots \\ \eta(\boldsymbol{s}_n) \end{bmatrix}}_{\boldsymbol{\eta}} \approx \underbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nr} \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_r \end{bmatrix}}_{\boldsymbol{Z}}$$

Question

- How to choose $\boldsymbol{A}$?
- How to determine $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$?

  $\Rightarrow$ choose r "knots" $\boldsymbol{s}_1^*, \cdots, \boldsymbol{s}_r^* \in \mathcal{S}$ and let $\boldsymbol{Z} = \left(Z\left(\boldsymbol{s}_1^*\right), \cdots, Z\left(\boldsymbol{s}_r^*\right)\right)^{\mathrm{T}}$

Low–rank approximation: expansion matrix $\boldsymbol{A}$

$$\underbrace{\begin{bmatrix} \eta(\boldsymbol{s}_1) \\ \eta(\boldsymbol{s}_2) \\ \vdots \\ \vdots \\ \vdots \\ \eta(\boldsymbol{s}_n) \end{bmatrix}}_{\boldsymbol{\eta}} \approx \underbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nr} \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_r \end{bmatrix}}_{\boldsymbol{Z}}$$

Question

- How to choose $\boldsymbol{A}$?
- How to determine $\boldsymbol{Z} = (Z_1, \cdots, Z_r)^{\mathrm{T}}$?
  $\Rightarrow$ choose r "knots" $\boldsymbol{s}_1^*, \cdots, \boldsymbol{s}_r^* \in \mathcal{S}$ and let
  $\boldsymbol{Z} = (Z(\boldsymbol{s}_1^*), \cdots, Z(\boldsymbol{s}_r^*))^{\mathrm{T}}$

Choice of $\boldsymbol{A}$

- Orthogonal basis functions: e.g., Fourier, orthogonal polynomial, etc
  - Karhune–Loéve expansion: $\mathrm{Cov}(\eta(\boldsymbol{s}_1), \eta(\boldsymbol{s}_2)) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\boldsymbol{s}_1)\phi_j(\boldsymbol{s}_2)$ where

$$\int_{\mathcal{S}} \mathrm{Cov}(\eta(\boldsymbol{s}_1), \eta(\boldsymbol{s}_2))\phi_j(\boldsymbol{s}_1)\, d\boldsymbol{s}_1 = \lambda_j \phi_j(\boldsymbol{s}_2)$$

  - Empirical orthogonal functions (EOFs)

- Nonorthogonal Basis Functions: represent the spatial process as combination of kernel functions

$$\eta(\boldsymbol{s}) = \sum_{j=1}^{r} a(\boldsymbol{s}, \boldsymbol{s}_j^*) Z_j + \xi(\boldsymbol{s})$$

Outline

Fixed Rank Kriging (Cressie & Johannesson 08)

Fixed-rank kriging outlines how kriging (i.e., spatial best linear unbiased predictor) can be applied in the low-rank parameterization of a spatial process.

- $\boldsymbol{A}$: (multiresolutional) non orthogonal basis functions

$$a(\boldsymbol{s})^{\mathrm{T}} = \left(\{1 - (\|\boldsymbol{s} - s^*_{1(l)}\|/\rho_{1(l)})^2\}^2_+, \cdots, \{1 - (\|\boldsymbol{s} - s^*_{r(l)}\|/\rho_{r(l)})^2\}^2_+\right)$$

- $\boldsymbol{Z} = \left(Z\left(\boldsymbol{s}^*_1\right), \cdots, Z\left(\boldsymbol{s}^*_r\right)\right)^{\mathrm{T}}$, $\boldsymbol{\Sigma_Z}$ to be estimated from data

The fixed rank kriging is equivalent to the following low rank model

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{r} a(\boldsymbol{s} - \boldsymbol{s}^*_{j(l)})Z_j + \varepsilon(\boldsymbol{s}) + \xi(\boldsymbol{s})$$

Multiresolutional basis functions

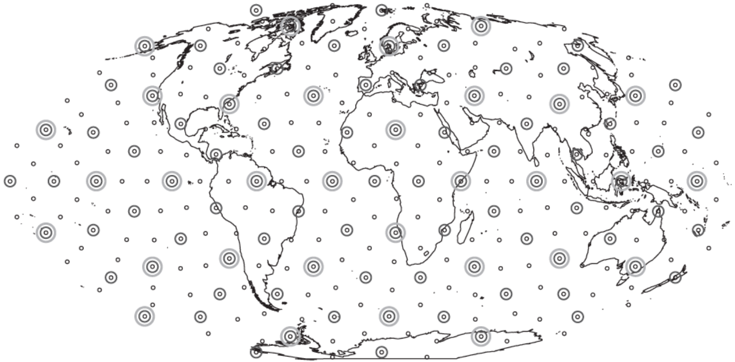

Figure: courtesy of Cressie & Johannesson 08

Nonstationary covariance function

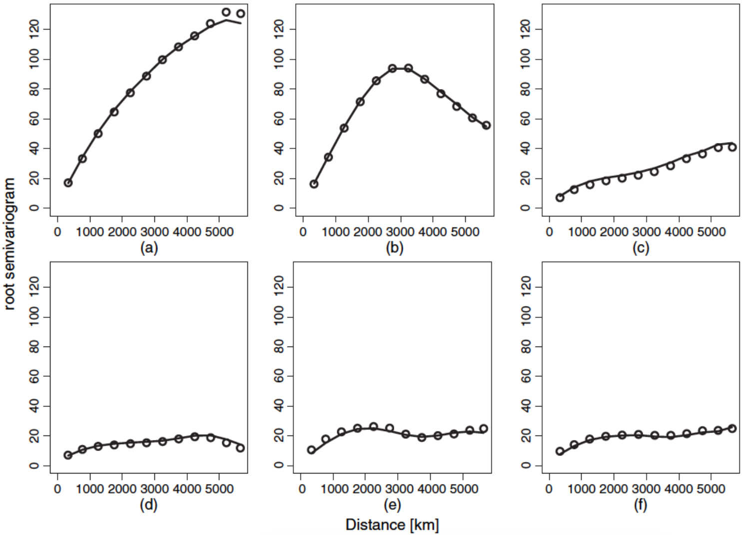

Figure: courtesy of Cressie & Johannesson 08

Gaussian Predictive Process (Banerjee et al. 08)

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \eta(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \quad \eta \sim GP(0, C(\cdot, \cdot; \boldsymbol{\theta}))$$

Reasoning: Given $\boldsymbol{Z}$, we seek a linear combination $a(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{Z}$ such that

$$\int_{\mathcal{S}} \mathbb{E}\left[\eta\left(\boldsymbol{s}\right) - a\left(\boldsymbol{s}\right)^{\mathrm{T}}\boldsymbol{Z}\right]^2 d\boldsymbol{s}$$

is minimized.

In predictive process one let $\boldsymbol{Z} = \boldsymbol{\eta}^* = \{\eta(\boldsymbol{s}_i^*)\}_{i=1}^p \sim N_p(\mathbf{0}, C^*(\boldsymbol{\theta}))$ where $C^*(\boldsymbol{\theta}) = \left[C(s_i^*, s_j^*; \boldsymbol{\theta})\right]_{i,j=1}^p$ and

$$\boldsymbol{A} = \left[C(\boldsymbol{s}_i, \boldsymbol{s}_j^*; \boldsymbol{\theta})\right]_{i=1,\cdots,n}^{j=1,\cdots,p}\left[C^*\right]^{-1}$$

Covaraince function approximation

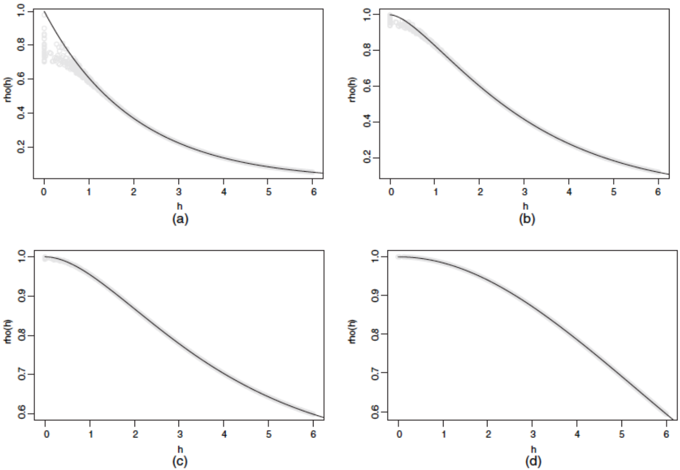

Fig. 1. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance (●): (a) smoothness parameter 0.5; (b) smoothness parameter 1; (c) smoothness parameter 1.5; (d) smoothness parameter 5

Figure: courtesy of Banerjee et al. 08

Covaraince function approximation cont'd

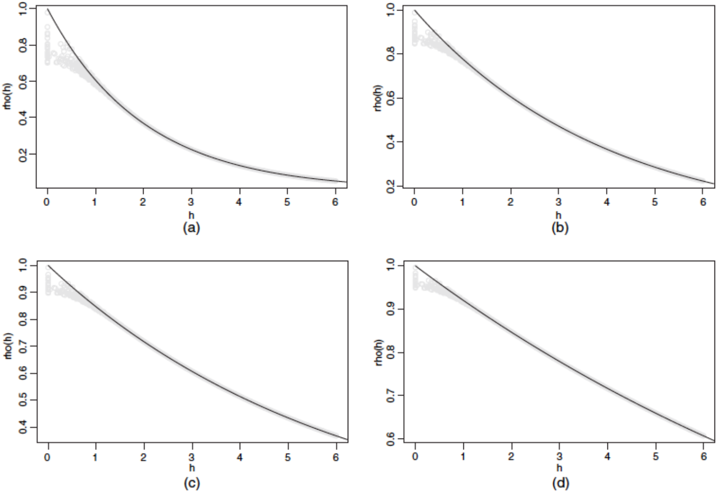

Fig. 2. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance: (a) range parameter 2; (b) range parameter 4; (c) range parameter 6; (d) range parameter 12

Figure: courtesy of Banerjee et al. 08

Discussion

References

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang
Gaussian Predictive Process Models for Large Spatial Data Sets
JRSSB, 70:825–848, 2008.

N. A. C. Cressie, and G. Johannesson
Fixed Rank Kriging for Very Large Spatial Data Sets
JRSSB, 70:209–226, 2008.

C. K. Wikle
Low-rank representations for spatial processes, Chapter 8 of Handbook of Spatial Statistics.
Chapman & Hall/CRC, 2010.