Large-Scale Spatial Data Science with ExaGeoStat

Sameh Abdulah, Marc G. Genton, Yan Song

In collaboration w/Ying Sun, Hatem Ltaief, David E. Keyes

Spatio-Temporal Statistics and Data Science (stsds.kaust.edu.sa)
Statistics Program (stat.kaust.edu.sa)
King Abdullah University of Science and Technology

February 26, 27, 28, 2024

Part I

Spatial Statistics Overview

Marc's Background

- B.S. & M.S. in Applied Mathematics from Swiss Federal Institute of Technology (EPFL), Switzerland, 1992/1994
- Ph.D. in Statistics from EPFL, Switzerland, 1996
- Professor in the USA, 1997-2012 (MIT, NCSU, TAMU)
- Al-Khawarizmi Distinguished Professor of Statistics at KAUST, Saudi Arabia (joined 2012)

Research Interests: statistical analysis, visualization, modeling, prediction, and uncertainty quantification of spatio-temporal data, skewed multivariate non-Gaussian distributions and robust statistics, with applications in environmental and climate science, and renewable energies such as wind and solar power

Types of Spatial Data

Spatial data: data indexed by locations (coordinates) in space

- Geostatistical data: index can vary continuously in space
- Regularly spaced data vs irregularly spaced data
- Point measurement vs block averages (or areal data)
- Data: multivariate, space-time, directional, on the sphere
- Other types: lattice data; point patterns

Types of Spatial Data

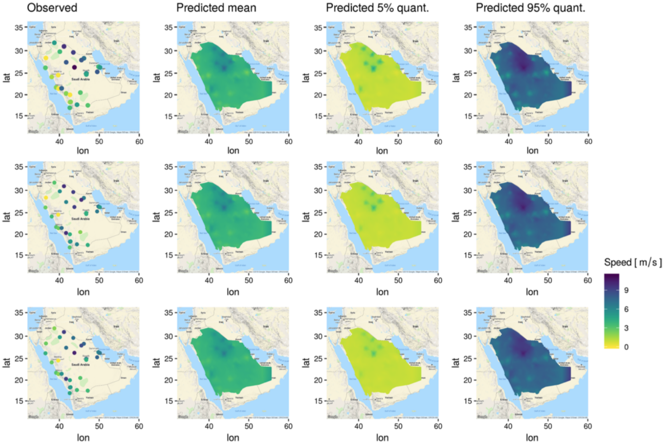
Spatial data: data indexed by locations (coordinates) in space

- Geostatistical data: index can vary continuously in space
- Regularly spaced data vs irregularly spaced data
- Point measurement vs block averages (or areal data)
- Data: multivariate, space-time, directional, on the sphere
- Other types: lattice data; point patterns

Law of Geography:
nearby things tend to be more alike than those far apart

Spatial Data Example

Wind speed (hourly) at 28 stations in Saudi Arabia in June 2010



Source: Lenzi, A., and Genton, M. G. (2020), Spatio-temporal probabilistic wind vector forecasting over Saudi Arabia, *Annals of Applied Statistics*, 14, 1359-1378.

Stochastic Processes and Random Fields

A stochastic process is a family or collection of random variables whose members can be identified or indexed according to some set

- Example: a time series $Z(t), t = t_1, ..., t_n$

Stochastic Processes and Random Fields

A stochastic process is a family or collection of random variables whose members can be identified or indexed according to some set

- Example: a time series $Z(t), t = t_1, ..., t_n$

We call a spatial process, $Z(\boldsymbol{s}), \boldsymbol{s} \in D, D \subseteq \mathbb{R}^d$, a random field

- Typically $d = 2$ but d can be greater than 2

Stochastic Processes and Random Fields

A stochastic process is a family or collection of random variables whose members can be identified or indexed according to some set

- Example: a time series $Z(t), t = t_1, ..., t_n$

We call a spatial process, $Z(\boldsymbol{s}), \boldsymbol{s} \in D, D \subseteq \mathbb{R}^d$, a random field

- Typically $d = 2$ but d can be greater than 2

A random field is Gaussian (Gaussian random field) if all finite-dimensional distributions are multivariate normal, i.e., given any $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, the vector $(Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))^\top$ is multivariate normal

- If $Z(\boldsymbol{s})$ is a Gaussian random field then it is completely determined by its mean and covariance functions

Mean and Covariance Functions of Random Fields

The mean function of $Z(\boldsymbol{s})$ is

$$\mu(\boldsymbol{s}) = \mathbb{E}\{Z(\boldsymbol{s})\}$$

Mean and Covariance Functions of Random Fields

The mean function of $Z(\boldsymbol{s})$ is

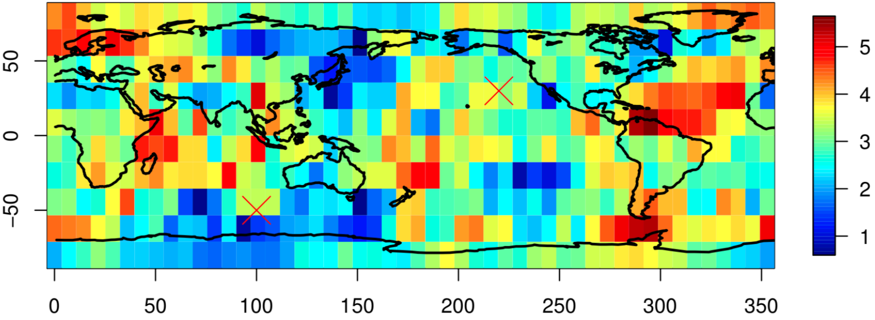
$$\mu(\boldsymbol{s}) = \mathbb{E}\{Z(\boldsymbol{s})\}$$

The covariance function of $Z(\boldsymbol{s})$ is

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{cov}\{Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2)\} = \mathbb{E}[\{Z(\boldsymbol{s}_1) - \mu(\boldsymbol{s}_1)\}\{Z(\boldsymbol{s}_2) - \mu(\boldsymbol{s}_2)\}]$$

where $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are two spatial locations

Covariance Function



- What is the domain of the spatial random field?
- How do we calculate the covariance between the random field at the two locations?

Stationarity and Isotropy

Strict stationarity:

- $\Pr\big(Z(\boldsymbol{s}_1) \leq z_1, \dots, Z(\boldsymbol{s}_n) \leq z_n\big) = \Pr\big(Z(\boldsymbol{s}_1 + \boldsymbol{h}) \leq z_1, \dots, Z(\boldsymbol{s}_n + \boldsymbol{h}) \leq z_n\big)$ for any finite $n \in \mathbb{N}$ and $\boldsymbol{h}, \boldsymbol{s}_1, \dots, \boldsymbol{s}_n \in D$

Weak stationarity:

- $\mu(\boldsymbol{s}) = \mu(\boldsymbol{s} + \boldsymbol{h})$ for all $\boldsymbol{h}, \boldsymbol{s} \in D$ and
- $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\boldsymbol{s}_1 + \boldsymbol{h}, \boldsymbol{s}_2 + \boldsymbol{h})$ for all $\boldsymbol{h}, \boldsymbol{s}_1, \boldsymbol{s}_2 \in D$

For a Gaussian process:

- Strict Stationarity $\iff$ Weak Stationarity

  Then, $\mu(\boldsymbol{s})$ is a constant and $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C_1(\boldsymbol{s}_1 - \boldsymbol{s}_2)$

Isotropy:

- $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C_2(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$

Requirements of Valid Covariance Functions

- Finite moments: $\mathbb{E}[\{Z(\boldsymbol{s})\}^2] < \infty$

Requirements of Valid Covariance Functions

- Finite moments: $\mathbb{E}[\{Z(\boldsymbol{s})\}^2] < \infty$
- Nonnegative definiteness (or positive semi-definiteness):

$$\sum_{j,k=1}^{n} c_j c_k C(\boldsymbol{s}_j, \boldsymbol{s}_k) \geq 0$$

for any finite n, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, and real numbers $c_1, \ldots, c_n$

Requirements of Valid Covariance Functions

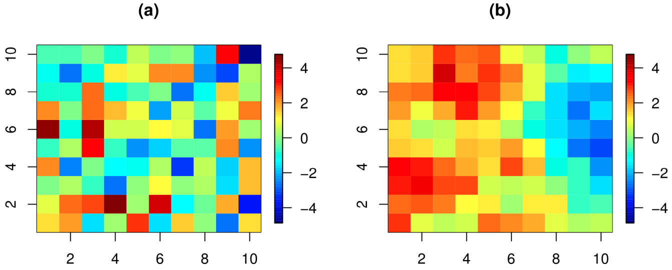
- Finite moments: $\mathbb{E}[\{Z(\boldsymbol{s})\}^2] < \infty$
- Nonnegative definiteness (or positive semi-definiteness):

$$\sum_{j,k=1}^{n} c_j c_k C(\boldsymbol{s}_j, \boldsymbol{s}_k) \geq 0$$

  for any finite n, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, and real numbers $c_1, \ldots, c_n$

- Properties:
  linear combination with positive coefficients; product; limit

Why do We Care about Covariance?



- Both (a) and (b) give zero-mean random fields on the domain $[0, 10] \times [0, 10]$ but in (a), every point is independent and in (b), nearby points have positive correlations
- In (a), if a pixel is missing, what would be the “best” guess for that missing pixel?
- How about (b)?

Why do We Care about Covariance?

- This leads to the concept of *Kriging*
- *Kriging* is another name for the Best Linear Unbiased Prediction (BLUP): the predicted value at a new location is a linear combination of the observations
  - Linear: $\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i)$
  - Unbiased: $\mathbb{E}\{\hat{Z}(\boldsymbol{s}_0)\} = \mu(\boldsymbol{s}_0)$
  - Best: $\mathrm{var}\{\hat{Z}(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\}$ is minimal among all linear unbiased predictions
- In determining the coefficients (*weights*) λ_i of the linear combination, the covariance plays an important role

Matérn Covariance Function

The popular parameterization of Matérn covariance function:

$$\mathrm{cov}\{Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\} = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta} \right)^{\nu} K_{\nu} \left(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta} \right) + \tau^2 \mathbb{1}_{\{i=j\}}$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, $\Gamma(\cdot)$ is the Gamma function, and $\mathbb{1}$ is the indicator function

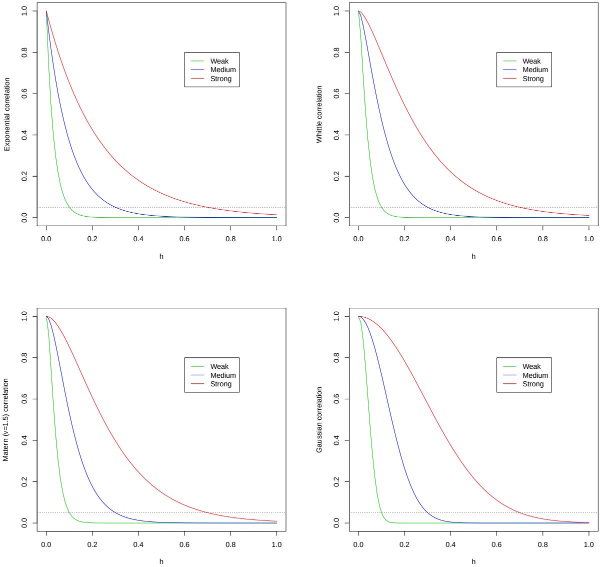
The four parameters determining the covariance structure are:
the partial sill σ^2, range $\beta > 0$, smoothness $\nu > 0$, and nugget τ^2

Effective Range

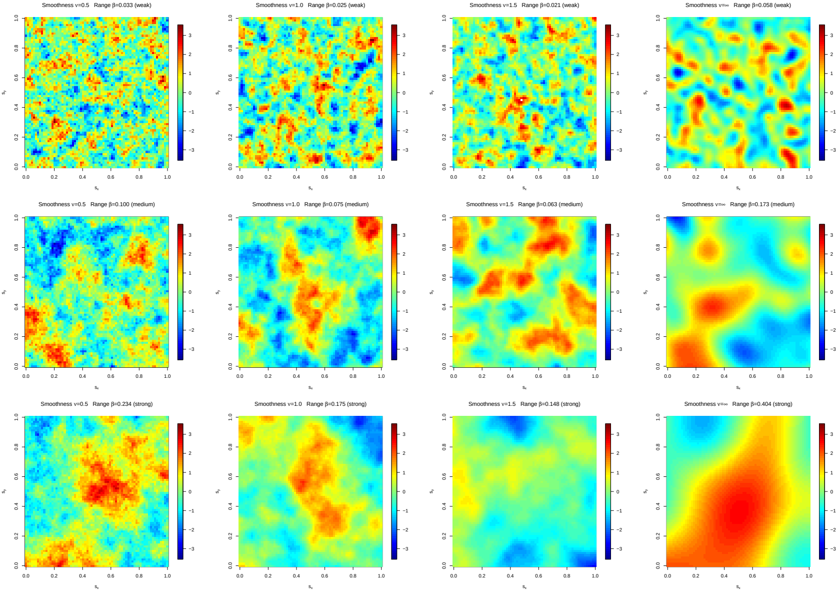
The effective range is the distance at which the correlation function reaches a small value, e.g. 5%

On the unit square, one can say the dependence is
weak / medium / strong
if the effective range is for example
0.1 / 0.3 / 0.7

Plots of Matérn Covariance Functions



Simulated Gaussian Random Fields using Matérn



Extension to Multivariate Covariance Models

For p-variate random fields

- Cross-covariance function $\mathbf{C}(\boldsymbol{h}) = \{C_{ij}(\boldsymbol{h})\}_{i,j=1}^{p}$
  - $C_{ij}(\boldsymbol{h}) = \mathrm{cov}\{Z_i(\boldsymbol{s}_1), Z_j(\boldsymbol{s}_2)\}$, where $\boldsymbol{h} = \boldsymbol{s}_1 - \boldsymbol{s}_2$
- Parsimonious multivariate Matérn cross-covariance function

$$C_{ij}(\boldsymbol{h}) = \frac{\rho_{ij}\sigma_{ii}\sigma_{jj}}{2^{\nu_{ij}}\Gamma(\nu_{ij})}\left(\frac{\|\boldsymbol{h}\|}{\beta}\right)^{\nu_{ij}} K_{\nu_{ij}}\left(\frac{\|\boldsymbol{h}\|}{\beta}\right)$$

  - σ_{ii}, σ_{jj}: marginal standard deviation
  - ν_{ii}, ν_{jj}: marginal smoothness and $\nu_{ij} = (\nu_{ii} + \nu_{jj})/2$
  - β: spatial range
  - $\rho_{ij} = r_{ij}\sqrt{\frac{\Gamma(\nu_{ii}+d/2)\Gamma(\nu_{jj}+d/2)}{\Gamma(\nu_{ii})\Gamma(\nu_{jj})}}\frac{\Gamma(\nu_{ij})}{\Gamma(\nu_{ij}+d/2)}$ controls the dependence between i-th and j-th variables, where $\{r_{ij}\}_{i,j=1}^{p}$ is a correlation matrix

References:

1. Gneiting, T., Kleiber, W., and Schlather, M. (2010), Matérn cross-covari- ance functions for multivariate random fields, *Journal of the American Statistical Association 105*(491), 1167–1177.

2. Apanasovich, T. V., Genton, M. G., and Sun, Y.(2012), A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association 107*(497), 180–193.

Extension to Spatio-Temporal Models

Interaction between space and time? Separable or Nonseparable?

- Generic separable models: $C(\boldsymbol{h}, u) = \sigma^2 \rho_S(\boldsymbol{h}) \rho_T(u)$
- Gneiting model:

$$C(\boldsymbol{h}, u) = \frac{\sigma^2}{(a|u|^{2\alpha}+1)^{\tau}} \exp\left(-\frac{c\|\boldsymbol{h}\|^{2\gamma}}{(a|u|^{2\alpha}+1)^{\beta\gamma}}\right)$$

- Cressie and Huang non-separable model:

$$C(\boldsymbol{h}, u) = \frac{\sigma^2(a|u|+1)}{\{(a|u|+1)^2 + b^2\|\boldsymbol{h}\|^2\}^{3/2}}$$

- A possible corresponding separable model by Mitchell et al.

$$C(\boldsymbol{h}, u) = \frac{\sigma^2}{(a|u|+1)^2(b^2\|\boldsymbol{h}\|^2+1)^{3/2}}$$

References:

1. Gneiting, T. (2010), Nonseparable, stationary covariance functions for space-time data, *Journal of the American Statistical Association 97*(458), 590-600.
2. Cressie N. and Huang H. (1999), Classes of nonseparable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association 94*, 1330-1340.
3. Mitchell M.W., Genton M.G. and Gumpertz M.L. (2005), *Environmetrics, 16*(8), 819-831.

Yan’s Background

- B.S. in Statistics from Beijing Institute of Technology, China, 2018
- Ph.D. in Statistics from Renmin University of China, China, 2023
- Postdoc at the Spatio-Temporal Statistics and Data Science (STSDS), KAUST, Saudi Arabia, 2023-now

Research Interests: large-scale spatio-temporal statistics, subsampling, and non-parametric regression

Covariance Parameter Estimation

- Variograms
  - Fast
- Likelihood-based methods
  - Good theoretical properties
  - Generally better performance

Variograms

Consider a stationary (or isotropic) random field Z with a covariance function C

$$\begin{aligned} \mathrm{var}\{Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2)\} &= \mathrm{var}\{Z(\boldsymbol{s}_1)\} + \mathrm{var}\{Z(\boldsymbol{s}_2)\} - 2\mathrm{cov}\{Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2)\} \\ &= 2C(\mathbf{0}) - 2C(\boldsymbol{s}_1 - \boldsymbol{s}_2) \end{aligned}$$

- Denoting $\gamma(\boldsymbol{h}) = C(\mathbf{0}) - C(\boldsymbol{h})$, we call $\gamma(\boldsymbol{h})$ a semivariogram and $2\gamma(\boldsymbol{h})$ a variogram
- Properties: $\gamma(\mathbf{0}) = 0$, $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$
- If Z is isotropic, $\gamma(\boldsymbol{h}) = \gamma_0(\|\boldsymbol{h}\|)$

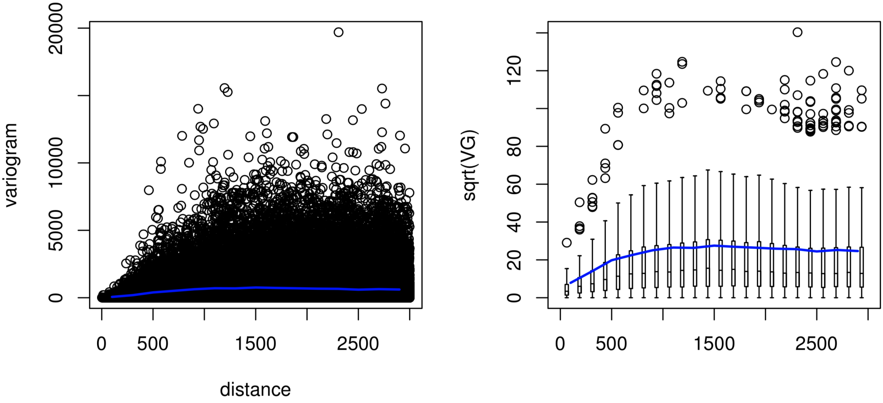
Draw and Estimate Variograms

For simplicity, we assume stationarity and isotropy

- First, for every possible location pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j)$, calculate the distance $\boldsymbol{h}_{ij} = \boldsymbol{s}_i - \boldsymbol{s}_j$ between them and the difference between the observed data on them, i.e., $\{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\}^2$. Plot $\{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\}^2$ against $\boldsymbol{h}_{ij}$ for all location pairs, we get a variogram cloud.
- Then bin the distances $H = \{\boldsymbol{h}_{ij}\} = \cup_{l=1,\ldots,L} H_l$. In each bin, take the average of distances, e.g., $\boldsymbol{h}_l = N_{\boldsymbol{h}_l}^{-1} \sum_{\boldsymbol{h}_{ij} \in H_l} \boldsymbol{h}_{ij}$ and take the average of the squared differences
- Empirical estimator (Matheron):

$$2\hat{\gamma}(\boldsymbol{h}_l) = \frac{1}{N_{\boldsymbol{h}_l}} \sum_{\boldsymbol{s}_i - \boldsymbol{s}_j \in H_l} \{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\}^2$$

Draw and Estimate Variograms



Estimate Covariance Parameters via Variograms

- Suppose $2\gamma(\boldsymbol{h};\boldsymbol{\theta})$ is the true variogram and we have empirical variogram values at L distance lags: $2\hat{\gamma}(\boldsymbol{h}_i), i = 1, \ldots, L$
- We assume $\hat{\gamma}(\boldsymbol{h}_i) = \gamma(\boldsymbol{h};\boldsymbol{\theta}) + e(\boldsymbol{h})$ and $\mathbb{E}\{e(\boldsymbol{h})\} = 0$
- $\boldsymbol{R}(\boldsymbol{\theta})_{ij} := \mathrm{cov}\{e(\boldsymbol{h}_i), e(\boldsymbol{h}_j)\}$
- $\hat{\boldsymbol{\gamma}} = \big(\hat{\gamma}(\boldsymbol{h}_1), \ldots, \hat{\gamma}(\boldsymbol{h}_L)\big)^\top$
- $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \big(\gamma(\boldsymbol{h}_1;\boldsymbol{\theta}), \ldots, \gamma(\boldsymbol{h}_L;\boldsymbol{\theta})\big)^\top$
- $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\}^\top \boldsymbol{R}(\boldsymbol{\theta})^{-1}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\}$

Various Least Squares in Variograms Estimation

- Ordinary least squares (OLS):

$$\boldsymbol{R}(\boldsymbol{\theta}) = \phi^2 \boldsymbol{I}_L$$

- Weighted least squares (WLS):

$$\boldsymbol{R}(\boldsymbol{\theta}) = \mathrm{diag}\Big(\mathrm{var}\{2\hat{\gamma}(\boldsymbol{h}_1)\}, \ldots, \mathrm{var}\{2\hat{\gamma}(\boldsymbol{h}_L)\}\Big)$$

  where $\mathrm{var}\{2\hat{\gamma}(\boldsymbol{h}_i)\} \approx 2\dfrac{\{2\gamma(\boldsymbol{h}_i)\}^2}{|N(\boldsymbol{h}_i)|}$

- Generalized least squares (GLS) considers the correlation between $e(\boldsymbol{h}_i)$ and $e(\boldsymbol{h}_j)$; details omitted here
- In practice, OLS or WLS commonly used for computational reasons

Likelihoods

For simplicity, we focus on zero-mean stationary Gaussian random fields
The log-likelihood for n locations:

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}$$

where

$$\boldsymbol{Z} = \begin{pmatrix} Z(\boldsymbol{s}_1) \\ \vdots \\ Z(\boldsymbol{s}_n) \end{pmatrix}, \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} C(\boldsymbol{s}_1, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_1, \boldsymbol{s}_n; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ C(\boldsymbol{s}_n, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_n, \boldsymbol{s}_n; \boldsymbol{\theta}) \end{pmatrix}$$

- Log determinant and linear solver require a Cholesky factorization of the given covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$ memory

Maximum Likelihood Estimation (MLE)

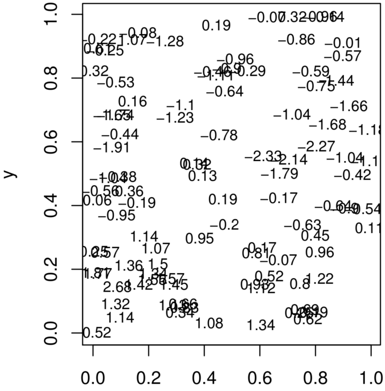
Maximum Likelihood Estimator

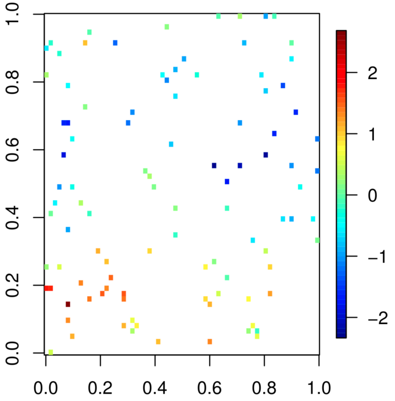
$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \{\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \boldsymbol{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{Z}\}
\end{aligned}
$$

Simulated Data Example

Zero-mean random process Z on $[0,1] \times [0,1]$ with 100 randomly chosen observation sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{100}$. Covariance:

$$C(\boldsymbol{h}) = \exp(-\|\boldsymbol{h}\|/0.2) + 0.1 \times \mathbb{1}_{\{\boldsymbol{h}=\boldsymbol{0}\}}$$





Simulated Data Results

- True model:

$$C(\boldsymbol{h}) = \exp(-\|\boldsymbol{h}\|/0.2) + 0.1 \times \mathbb{1}_{\{\boldsymbol{h}=\boldsymbol{0}\}}$$

- MLE result:

$$C(\boldsymbol{h}) = 0.86 \exp(-\|\boldsymbol{h}\|/0.26) + 0.067 \times \mathbb{1}_{\{\boldsymbol{h}=\boldsymbol{0}\}}$$

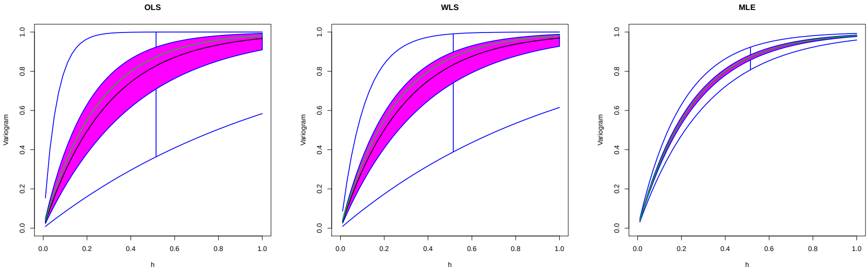
- Variogram OLS result:

$$C(\boldsymbol{h}) = 0.096 \exp(-\|\boldsymbol{h}\|/0.0026) + 0.98 \times \mathbb{1}_{\{\boldsymbol{h}=\boldsymbol{0}\}}$$

- With only 100 data points, we may not expect both methods to estimate the true parameters perfectly. However, we clearly see variogram OLS performs much worse than MLE

More Simulated Data Example

- Exponential variogram: $2\gamma(h) = 1 - \exp(-h/\theta)$ with $\theta = 0.25$
- Mean-zero GP generated at 400 random locations in unit square
- Estimate θ by OLS, WLS, MLE with 1000 replicates
- Functional boxplots (Sun and Genton, 2011)
- Note: no outlier detection
  (the factor is set to be large in order to see the variability better)
- More in Yan and Genton (2018)



Prediction

For Gaussian random fields, kriging coincides with the conditional mean

$$\begin{pmatrix} \boldsymbol{Z} \\ Z(\boldsymbol{s}_0) \end{pmatrix} \sim N_{n+1}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{\theta}) & \boldsymbol{k}(\boldsymbol{\theta}) \\ \boldsymbol{k}(\boldsymbol{\theta})^\top & C(\boldsymbol{s}_0, \boldsymbol{s}_0; \boldsymbol{\theta}) \end{pmatrix}\right)$$

where $\boldsymbol{k}(\boldsymbol{\theta}) = \left(C(\boldsymbol{s}_1, \boldsymbol{s}_0; \boldsymbol{\theta}), \ldots, C(\boldsymbol{s}_n, \boldsymbol{s}_0; \boldsymbol{\theta})\right)^\top$
The conditional distribution is

$$Z(\boldsymbol{s}_0)|\boldsymbol{Z} \sim N\Big(\boldsymbol{k}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}, C(\boldsymbol{s}_0, \boldsymbol{s}_0; \boldsymbol{\theta}) - \boldsymbol{k}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{k}(\boldsymbol{\theta})\Big)$$

- Solution of system of linear equation $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}$ also needs a Cholesky factorization of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

When Size n of Datasets Becomes Large

- $O(n^3)$ floating point operations and $O(n^2)$ memory for **exact computations** of Cholesky factorization
  - High-Performance Computing (HPC) can help when n is large
  - *ExaGeoStat* software: https://github.com/ecrc/exageostat

  Note: $n = 1'000'000$ then $n^3 = 10^{18}$= 1 billion billions

- *ExaGeoStat* for:
  1. Likelihood inference/learning for Matérn covariance function (+others)
  2. Spatial kriging (interpolation)
  3. Random field simulations

- Various **approximation methods** have been proposed in literature to ease computation & memory burden

- **2021/2022/2023 KAUST Competitions on Spatial Statistics for Large Datasets** investigate the performance of different approximation methods with large synthetic data generated by *ExaGeoStat*

Composite Likelihood Methods

Composite likelihood:

$$\mathcal{L}_C\Big(\boldsymbol{\theta}; Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\Big) = \prod_{k=1}^{K} \mathcal{L}\Big(\boldsymbol{\theta}; Z(\boldsymbol{s}_{k_1}), \ldots, Z(\boldsymbol{s}_{k_{j_k}})\Big)^{w_k}$$

Composite conditional likelihood under Vecchia's approximation:

$$\mathcal{L}_C\Big(\boldsymbol{\theta}; Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\Big) = \prod_{k=1}^{n} f\Big(Z(\boldsymbol{s}_k) \mid \{Z(\boldsymbol{s}_i) : i < k, \boldsymbol{s}_i \in N(\boldsymbol{s}_k)\}; \boldsymbol{\theta}\Big)$$

Composite pairwise likelihood:

$$\mathcal{L}_C\Big(\boldsymbol{\theta}; Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\Big) = \prod_{k=1}^{n-1} \prod_{i=k+1}^{n} f\Big(Z(\boldsymbol{s}_k), Z(\boldsymbol{s}_i); \boldsymbol{\theta}\Big)$$

References:

1. Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological) 50*(2), 297–312.
2. Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis 92*(1), 1–28.

Low-rank Methods

Low-rank methods

- predictive process
  - a small set of knots $\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*$
  - the predictive process $\tilde{Z}(\boldsymbol{s})$ approximates the original process $Z(\boldsymbol{s})$ as

$$\begin{aligned} \tilde{Z}(\boldsymbol{s}) &= \mathbb{E}\{Z(\boldsymbol{s}) \mid Z(\boldsymbol{s}_1^*), \ldots, Z(\boldsymbol{s}_{n^*}^*)\} \\ &= \boldsymbol{c}^\top {\Sigma^*}^{-1} \boldsymbol{Z}^* \end{aligned}$$

  where $\boldsymbol{Z}^* = \big(Z(\boldsymbol{s}_1^*), \ldots, Z(\boldsymbol{s}_{n^*}^*)\big)^\top$, $\boldsymbol{c} = \mathrm{cov}(Z(\boldsymbol{s}), \boldsymbol{Z}^*)^\top$, $\Sigma^* = \mathrm{var}(\boldsymbol{Z}^*)$

- fixed rank kriging, etc.

Reference: Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Sparse Matrix Methods

Induced sparse covariance or precision matrix

- covariance tapering
  - $\tilde{C}(\cdot,\cdot) = C(\cdot,\cdot)C_{\mathrm{taper}}(\cdot,\cdot)$, where $C_{\mathrm{taper}}(\cdot,\cdot)$ is a covariance function with compact support
  - Sparsity is induced in the approximate covariance matrix
- Gaussian Markov Random Field (GMRF)
  - Assumed to be conditional dependent on neighbors only
  - Sparsity is induced in the approximate precision matrix

References:
1. Furrer, R., Genton, M. G., and Nychka, D. (2006), Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics 15*(3), 502-523.
2. Kaufman, C. G., M. J. Schervish, and D. W. Nychka (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association 103*(484), 1545-1555.
3. Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Competitions on Spatial Statistics for Large Datasets

- **Goal:** investigate the **performance of different approximation methods** with large synthetic datasets generated by *ExaGeoStat*
- Through the competition, we can better understand when each approximation method is adequate
- The full datasets with one million locations are publicly available and act as **benchmarking data** for future research
- The exact MLEs and lowest RMSEs achieved by researchers worldwide are released so that other/new methods can be easily compared

In 2021: Gaussian and non-Gaussian

- Launched November 23, 2020; Ended February 1, 2021
- 29 research teams worldwide registered and 21 teams successfully submitted results
- Competition consists of four parts:

	Task	Data model	Data size
1a	GP estimation	GP	90,000
1b	prediction	GP	predict 10,000 conditional on 90,000
2a	prediction	Tukey g-and-h	predict 10,000 conditional on 90,000
2b	prediction	GP & Tukey g-and-h	predict 100,000 conditional on 900,000

- **Metric for GP estimation:**
  Mean Loss of Efficiency (MLOE) and
  Mean Misspecification of the Mean Square Error (MMOM)
- **Metric for prediction:** RMSE

In 2021: Gaussian estimation/prediction results

Sub-competition	Submission	Score	Rank
1a	ExaGeoStat(estimated-model)	154	0
	SpatStat-Fans	156	1
	GpGp	186	2
	RESSTE(CL/krig)	229	3
1b	ExaGeoStat(true-model)	72	0
	RESSTE(CL/krig)	78	1
	ExaGeoStat(estimated-model)	79	1.5
	HCHISS	93	2
	Chile-Team	113	3

Reference:
Huang, H., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2021), Competition on spatial statistics for large datasets (with discussion), *Journal of Agricultural, Biological, and Environmental Statistics*, 1-16.

In 2022: Nonstationary, space-time, multivariate

- Launched March 1, 2022; Ended May 1, 2022
- 20 research teams worldwide registered
- Hosted the competition on the **Kaggle** machine learning and data science platform

Sub-comp	Setting	True Data Model	# of Datasets	Training Data Size	Testing Data Size
1a	Univariate **Nonstationary** Spatial	GP with Nonstationary Mean or Cov	2	90K	10K
1b	Univariate **Nonstationary** Spatial	GP with Nonstationary Mean or Cov	2	900K	100K
2a	Univariate Stat. **ST**	GP with Non-Separable Cov	9	90K	10K
2b	Univariate Stat. **ST**	GP with Non-Separable Cov	9	900K	100K
3a	**Bivariate** Stationary Spatial	GP with Parsimonious/Flexible Matérn Cross-Cov	3	45K	5K
3b	**Bivariate** Stationary Spatial	GP with Parsimonious/Flexible Matérn Cross-Cov	3	450K	50K

Reference:
Abdulah, S., Alamri, F., Nag, P., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2022), The second competition on spatial statistics for large data sets, *Journal of Data Science*, 20, 439-460.

In 2023: Irregular locations, confidence/prediction intervals

- Launched February 1, 2023; Ended May 1, 2023
- 11 research teams worldwide registered
- Five different **designs** considered for the locations of the observations: 1. Chessboard; 2. Left-bottom; 3. Satellite; 4. Clusters; 5. Regular

Sub-comp	Model	Target	# designs	Training	Testing
1a	Gaussian Matérn	Estimation (95% conf interval)	5	90K	–
1b	Gaussian Matérn	Estimation (95% conf interval)	5	900K	–
2a	Gaussian Matérn	Prediction (95% pred interval)	5	90K	10K
2b	Gaussian Matérn	Prediction (95% pred interval)	5	900K	100K

Reference:
Hong, Y., Song, Y., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2023), The third competition on spatial statistics for large datasets, *Journal of Agricultural, Biological, and Environmental Statistics*, 28, 618-635.