Geostatistical estimation - Part I

Applied Spatial Statistics

Estimation strategies

- We now have several possible models for spatial processes
- In this lecture we discuss methods for fitting models to data
- One task is model selection:
  - Which covariates to include in $\mathbf{X}$?
  - Exponential or Matern correlation?
  - Should we include a nugget?
  - Is the covariance stationary?
- Another is parameter estimation:
  - Mean parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$
  - Covariance parameters $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi, \nu)$

Variogram

- The variogram is a common exploratory analysis tool
- It is used as a quick visual check to suggest an appropriate covariance model
- It is often applied to the least squares residuals

$$\hat{\varepsilon}_i = Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$$

- The expressions below use Y_i instead of $\hat{\varepsilon}_i$ to match notation used in books/web

Variogram - Definition

- The true variogram is a function of the parameters; the empirical variogram is a function of the data
- The true variogram is

$$2\gamma(\mathbf{s}_i, \mathbf{s}_h) = \text{Var}(Y_i - Y_j)^2$$

- $\gamma(\mathbf{s}_i, \mathbf{s}_j)$ is the semi-variogram
- Assuming Y_i and Y_j have the same mean, then the variogram is related to the covariance as

$$2\gamma(\mathbf{s}_i, \mathbf{s}_i) = \text{Var}(Y_i) + \text{Var}(Y_j) - \text{Cov}(Y_i, Y_j)$$

- The variogram increases with distance

Variogram - Understanding the variogram

- If the mean is smooth over space, the variogram removes it by local differencing
- Assuming Y_i and Y_j have the same mean, the variogram is

$$\mathrm{E}(Y_i - Y_j)^2$$

- If the observations are spatially correlated, the variogram is small for small distances
- The magnitude of local differences, and thus the variogram, increase with distance

Variogram - Understanding the variogram

Assume the isotropic model $Y_i = Z_i + \varepsilon_i$

- $\mathrm{V}(Z_i) = \sigma^2$
- $\mathrm{V}(\varepsilon_i) = \tau^2$
- $\mathrm{Cor}(Z_i, Z_j) = \rho(d_{ij})$
- d_{ij} is the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$
- $\rho(0) = 1$ and decreases to $\rho(\infty) = 0$

Variogram - Understanding the variogram

Under this isotropic mean-zero model we have

- $\text{Var}(Y_i) = \text{Var}(Z_i + \varepsilon_i) = \text{Var}(Z_i) + \text{Var}(\varepsilon_i) = \sigma^2 + \tau^2$

- The spatial covariance is

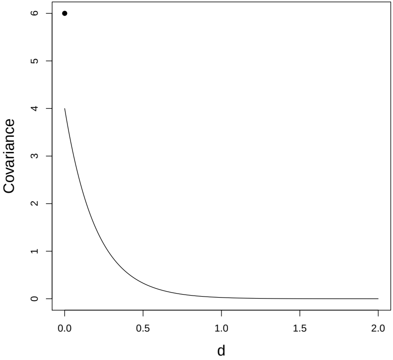
$$\begin{aligned}
\text{Cov}(Y_i, Y_j) &= \text{Cov}(Z_i + \varepsilon_i, Z_j + \varepsilon_j) \\
&= \text{Cov}(Z_i, Z_j) + \text{Cov}(Z_i, \varepsilon_j) + \text{Cov}(\varepsilon_i, Z_j) + \text{Cov}(\varepsilon_i, \varepsilon_j) \\
&= \text{Cov}(Z_i, Z_j) \\
&= \sigma^2 \rho(d_{ij})
\end{aligned}$$

- The correlation is

$$\text{Cor}(Y_i, Y_j) = \frac{\sigma^2}{\sigma^2 + \tau^2} \rho(d_{ij})$$

Exponential covariance plot

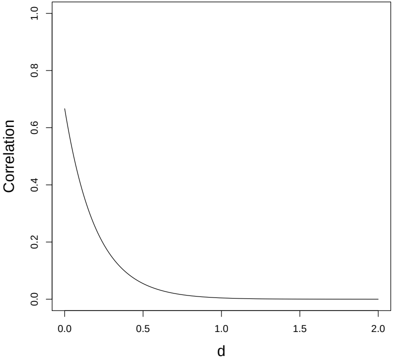
The exponential model is $\mathrm{V}(Y_i) = \sigma^2 + \tau^2$ and $\mathrm{Cov}(Y_j, Y_j) = \sigma^2 \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 2$ and $\phi = 0.2$

Exponential correlation plot

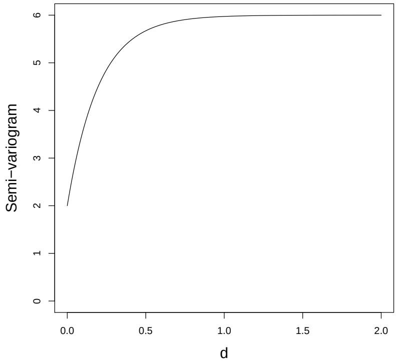
The exponential model $\mathrm{Cor}(Y_j, Y_j) = \frac{\sigma^2}{\sigma^2+\tau^2}\exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 2$ and $\phi = 0.2$

Exponential semi-variogram plot

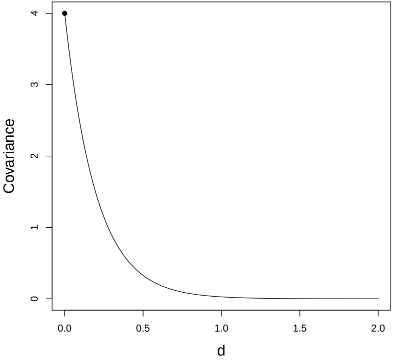
The exponential model is $\gamma(d) = \sigma^2 + \tau^2 - \sigma^2 \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 2$ and $\phi = 0.2$

Exponential covariance plot

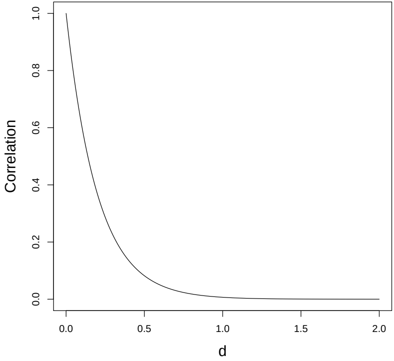
The exponential model is $\mathrm{V}(Y_i) = \sigma^2 + \tau^2$ and $\mathrm{Cov}(Y_j, Y_j) = \sigma^2 \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 0$ and $\phi = 0.2$

Exponential correlation plot

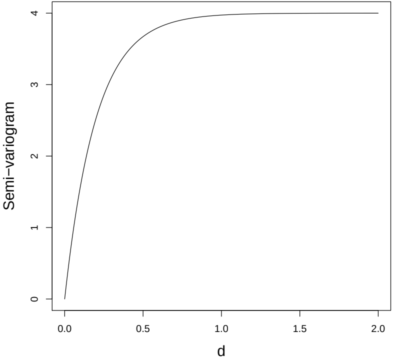
The exponential model $\text{Cor}(Y_j, Y_j) = \frac{\sigma^2}{\sigma^2+\tau^2} \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 0$ and $\phi = 0.2$

Exponential semi-variogram plot

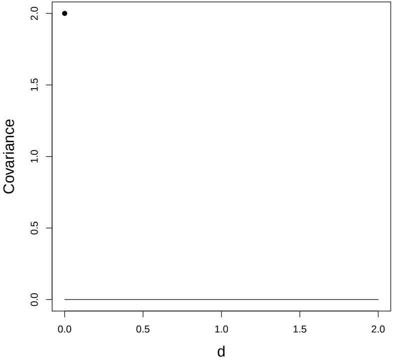
The exponential model is $\gamma(d) = \sigma^2 + \tau^2 - \sigma^2 \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 4$, $\tau^2 = 0$ and $\phi = 0.2$

Exponential covariance plot

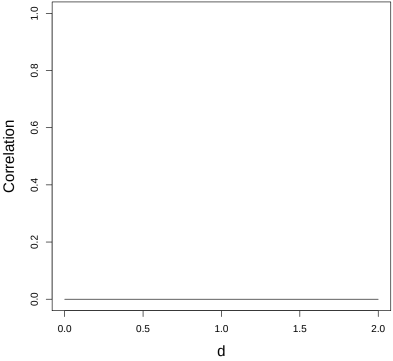
The exponential model is $V(Y_i) = \sigma^2 + \tau^2$ and $\text{Cov}(Y_j, Y_j) = \sigma^2 \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 0$, $\tau^2 = 2$ and $\phi = 0.2$

Exponential correlation plot

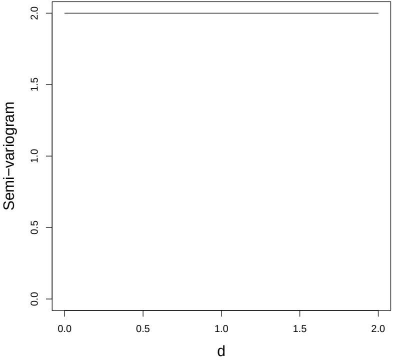
The exponential model $\text{Cor}(Y_j, Y_j) = \frac{\sigma^2}{\sigma^2+\tau^2} \exp(-d_{ij}/\phi)$



This plot assumes $\sigma^2 = 0$, $\tau^2 = 2$ and $\phi = 0.2$

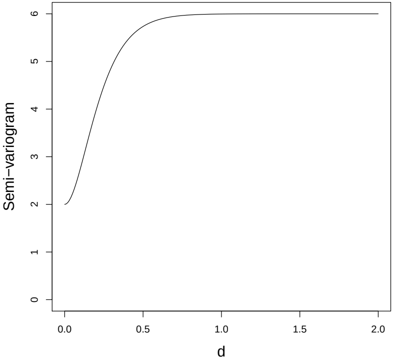
Exponential semi-variogram plot

The exponential model is $\gamma(d) = \sigma^2 + \tau^2 - \sigma^2 \exp(-d_{ij}/\phi)$



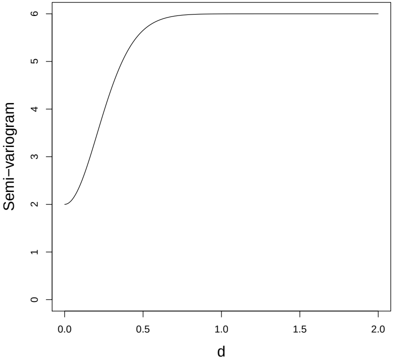
This plot assumes $\sigma^2 = 0$, $\tau^2 = 2$ and $\phi = 0.2$

Matern semi-variogram plot



This plot assumes $\sigma^2 = 4$, $\tau^2 = 2$, $\nu = 2$ and $\phi = 0.1$

Matern semi-variogram plot

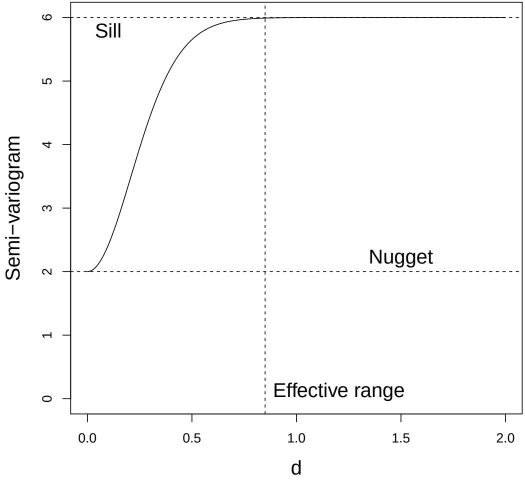


This plot assumes $\sigma^2 = 4$, $\tau^2 = 2$, $\nu = 10$ and $\phi = 0.05$

Variogram - Terminology

- The nugget variance, $\text{Var}(\varepsilon_i) = \tau^2$, is the semi-variogram at distance 0
- The spatial variance is the partial sill, $\text{Var}(Z_i) = \sigma^2$
- The semi-variogram plateaus at the sill, $\text{Var}(Y_i) = \sigma^2 + \tau^2$
- The effective range is the distance at which the variogram hits the sill

Variogram - Terminology



Variogram - Empirical variogram

- The empirical variogram uses data to approximate the true variogram
- The idea is to group pairs of observations by their distance and approximate the variance for each group
- Let $w_{ij}(d) = 1$ if $d_{ij} \in (d - h, d + h)$ and $w_{ij} = 0$ otherwise
- The empirical variogram is at distance d is

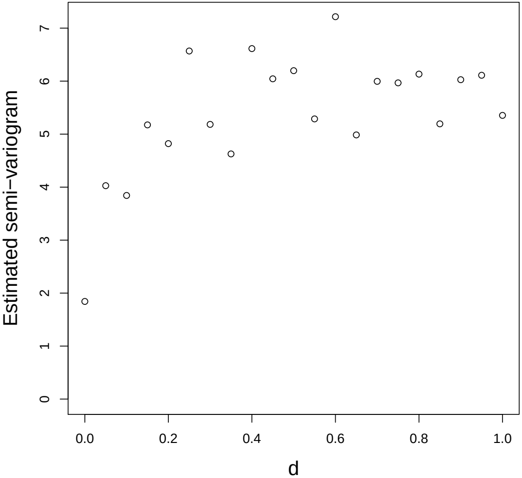
$$\hat{\gamma}(d) = \frac{1}{2N(d)} \sum_{i=1}^{n} \sum_{j=1}^{i} w_{ij}(d)(Y_i - Y_j)^2$$

where $N(d)$ as the number of pairs with $w_{ij}(d) = 1$

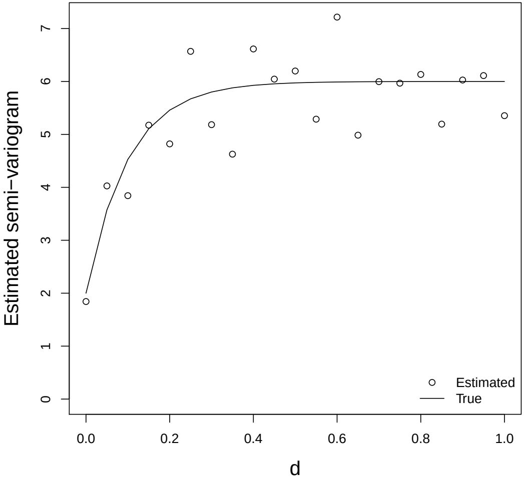
Variogram - tuning the empirical variogram

- The emprical variogram is computed for L distances, $d_1, ..., d_L$
- The width h is set to $(d_2 - d_1)/2$
- We need to pick L and the maximum distance d_L
- Rule of thumb: Set d_L to twice the effective range (larger will give nonsense!)
- Rule of thumb: Set L so that the number of pairs for each bin is at least 30
- Since we do not know the effective range at the beginning, this takes some iteration

Variogram - Examples



Variogram - Examples



The curve is $\sigma^2 + \tau^2 - \sigma^2 \exp(-d/\rho)$ for $\sigma^2 = 4$, $\tau^2 = 4$ and $\rho = 0.1$

Variogram - What to look for (questions)

1. Is there a nugget?
2. What is the effective range?
3. Does an exponential fit well or do I need a Matern?
4. Is the covariance isotropic?
5. Is the covariance stationary?

Variogram - What to look for (answers)

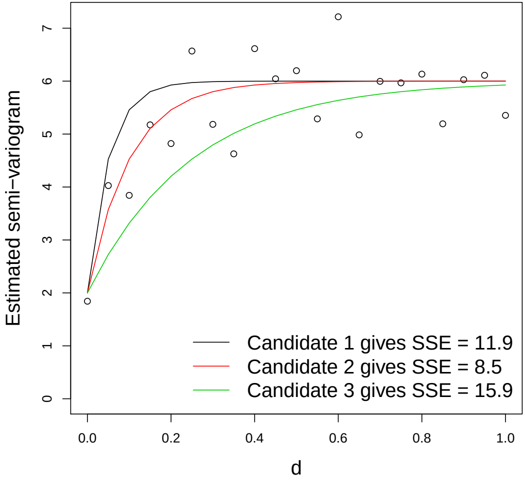
1. Check if the variogram goes through the origin
2. Find the distance at which the varogram plateaus
3. Plot the best fitting exponential model (see next slide)
4. Plot the variogram for pairs separated by different angles (N/S v E/W pairs), see if they are similar
5. Compute the variogram separately for different subregions, see if they are similar

Variogram - Least squares fitting

- Variograms can be used for parameter estimation
- Data: $\hat{\gamma}(d_1), ..., \hat{\gamma}(d_L)$
- Model: $\gamma(d;\boldsymbol{\theta})$, e.g., $\gamma(d;\boldsymbol{\theta}) = \tau^2 + \sigma^2 - \sigma^2 \exp(-d/\phi)$
- Estimate θ to minimize

$$\sum_{l=1}^{L} \{\hat{\gamma}(d_l) - \gamma(d_l)\}^2$$

Variogram - Examples



All take $\sigma^2 = 4$ and $\tau^2 = 2$ but vary by $\rho \in \{0.05, 0.10, 0.20\}$