Sample and Population Semivariogram

Setting:
Set of \( N \) measurements \( \{ y(\mathbf{s}_n), n = 1, \ldots, N \} \) of attribute \( Y \) which varies continuously in space; \( y(\mathbf{s}_n) \) denotes \( n \)-th measurement obtained at \( n \)-th sample location with coordinate vector \( \mathbf{s}_n \).

Inference objective:
Go beyond sample semivariogram to infer a model of spatial association for the population, i.e., the entire \( Y \)-attribute field; most often, one is talking about parametric semivariogram models which are expressed as functions of Euclidean distance.

Objectives of this handout:
- to pinpoint the limitations of a sample semivariogram/covariogram/correlogram
- to highlight some conditions that can be used to check whether an arbitrary function of distance is as valid semivariogram model
- to survey some of the most frequently used semivariogram models in practical applications

Sample Semivariogram

- Consider a set of \( L \) lag distance classes; let \( \{ h_l, l = 1, \ldots, L \} \) denote the set of average distances between data pairs in each class
- Compute sample semivariance \( \hat{\gamma}(h_l) \), or moment of inertia of \( h_l \)-specific scatter-plot of lagged \( y \)-attribute values, for each distance class \( h_l \):
  \[
  \hat{\gamma}(h_l) = \frac{1}{2N(h_l)} \sum_{n=1}^{N(h_l)} [y(\mathbf{s}_n) - y(\mathbf{s}_n + h_l)]^2
  \]
  where \( N(h_l) \) denotes \# of data pairs whose inter-distances fall in the \( l \)-th distance class \( h_l \); \( h_l \) is the separation vector with magnitude \( h_l = ||h_l|| \)
- Plot average distances \( \{ h_l, l = 1, \ldots, L \} \) versus corresponding sample semivariance values \( \{ \hat{\gamma}(h_l), l = 1, \ldots, L \} \); such a plot is called a sample (empirical) semivariogram
Problems With Sample Semivariograms/Covariograms

Semivariogram interpretation: (i) any vector \( h \) is associated with a \( h \)-specific scatter plot of lagged (tail and head) attribute pairs, (ii) these attribute pairs can be regarded as realizations from any two random variables (RVs) \( Y(s_n) \) and \( Y(s_n + h) \) defined at any two sample locations \( s_n \) and \( s_n + h \) separated by vector \( h \), (iii) the sample semivariance \( \hat{\gamma}(h) \) and covariance \( \hat{\sigma}(h) \) are estimates of linear association between lagged attribute pairs in these \( h \)-specific scatter plots.

Example: Consider three RVs \( Y(s_1), Y(s_2), \) and \( Y(s_3) \), and assume that their corresponding distances \( h_{12} = ||s_2 - s_1||, h_{23} = ||s_3 - s_2|| \) and \( h_{13} = ||s_3 - s_1|| \) are three of the distance values in the abscissa of the covariogram plot, i.e., \( h_{12}, h_{23}, h_{13} \) are three of the \( h \) values in the set \( \{h_l, l = 1, \ldots, L\} \). Based on the above interpretation, the estimated (co)variance matrix between these three RVs is:

\[
\hat{\Sigma} = \begin{bmatrix}
\hat{\sigma}(0) & \hat{\sigma}(h_{12}) & \hat{\sigma}(h_{13}) \\
\hat{\sigma}(h_{12}) & \hat{\sigma}(0) & \hat{\sigma}(h_{23}) \\
\hat{\sigma}(h_{13}) & \hat{\sigma}(h_{23}) & \hat{\sigma}(0)
\end{bmatrix}
\]

where \( \hat{\sigma}(0) \) is an estimate of the overall \( Y \)-variance (corresponding to distance \( h_{11} = h_{22} = h_{33} = 0 \))

Caveats:

- entries of sample covariance matrix \( \hat{\Sigma} \) have been computed from different \# of data pairs; for example, \( N(h_{12}) \neq N(h_{23}) \) → no guarantee that \( \hat{\Sigma} \) is a valid covariance matrix
- sample semivariogram values computed only for small number \( L \) of lag vectors; however, a continuum of semivariogram values for any arbitrary distance \( h \) between any two locations (sample-to-sample) and (sample-to-prediction) is needed for spatial interpolation

Model Semivariogram

Definition:

- a (typically parametric) function of distance \( \gamma_\theta(h) \) or \( \gamma(h; \theta) \), often fitted to \( L \) sample semivariance values \( \{\hat{\gamma}(h_l), l = 1, \ldots, L\} \)
- \( \theta = \) parameter vector: range and sill, for a given semivariogram model type (functional form)

Issues to consider:

- semivariogram modeling = curve fitting exercise? Note: fitting = parameter estimation
- cannot use any curve as a semivariogram model!!!
  Sometimes, physics can guide the selection of an appropriate model
- consider only those functions (typically of Euclidean distance) known to be valid models
Semivariogram / Covariogram / Correlogram Model

Slide 5

**Conversion between models** under 2nd-order stationarity, with \( \sigma(0) = \gamma(\infty) \) being the sill of the semivariogram model:

- **Semivariogram \( \rightarrow \) covariogram:** \[ \sigma(h) = \sigma(0) - \gamma(h) \]
- **Covariogram \( \rightarrow \) correlogram:** \[ \rho(h) = \frac{\sigma(h)}{\sigma(0)} \]
- **Semivariogram \( \rightarrow \) correlogram:** \[ \rho(h) = 1 - \frac{\gamma(h)}{\sigma(0)} \]
- **Covariogram \( \rightarrow \) semivariogram:** \[ \gamma(h) = \sigma(0) - \sigma(h) \]

Linear Combinations of RVs (1)

**Setting:** Set of \( N \) RVs arranged in a \((N \times 1)\) random vector \( y = [Y_n, n = 1, \ldots, N]^T \). Let \( \mu = [\mu_n, n = 1, \ldots, N]^T \) denote their \((N \times 1)\) mean vector with \( \mu_n = E\{Y_n\} \), and \( \Sigma = [\sigma_{nn'}, n = 1, \ldots, N, n' = 1, \ldots, N] \) denote their \((N \times N)\) (co)variance matrix with \( \sigma_{nn'} = C\{Y_n, Y_{n'}\} \).

**Linear combination of RVs:** Consider another RV \( Z \) defined as a weighted sum of the \( N \) \( Y \)-RVs:

\[
Z = \sum_{n=1}^{N} \lambda_n Y_n = \lambda^T y
\]

where \( \lambda = [\lambda_n, n = 1, \ldots, N]^T \) denotes a \((N \times 1)\) vector of arbitrary weights

**Expected value:** (or mean) of RV \( Z \):

\[
E\{Z\} = E\left\{ \sum_{n=1}^{N} \lambda_n Y_n \right\} = \sum_{n=1}^{N} \lambda_n E\{Y_n\} = \sum_{n=1}^{N} \lambda_n \mu_n = \lambda^T \mu
\]

The mean \( E\{Z\} \) of any RV \( Z \) defined as a weighted linear combination (a linear function) of \( N \) RVs \( \{Y_n, n = 1, \ldots, N\} \) can be expressed as a weighted linear combination of the means of these \( N \) RVs.
**Linear Combinations of RVs (2)**

**Variance of linear combination:**

\[ \mathbb{V}\{Z\} = \sum_{n=1}^{N} \sum_{n'=1}^{N} \lambda_n \sigma_{nn'} \lambda_{n'} = \mathbf{\lambda}^T \Sigma \mathbf{\lambda} \]

**Example:** \( Z = \lambda_1 Y_1 + \lambda_2 Y_2 \):

\[
\begin{align*}
\mathbb{V}\{Z\} &= \lambda_1^2 \mathbb{V}\{Y_1\} + \lambda_2^2 \mathbb{V}\{Y_2\} + 2 \lambda_1 \lambda_2 \mathbb{C}\{Y_1, Y_2\} \\
&= \lambda_1 \sigma_{11} \lambda_1 + \lambda_1 \sigma_{12} \lambda_2 + \lambda_2 \sigma_{21} \lambda_1 + \lambda_2 \sigma_{22} \lambda_2 \\
&= [\lambda_1 \lambda_2] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}
\end{align*}
\]

The variance \( \mathbb{V}\{Z\} \) of any RV \( Z \) defined as a weighted sum of \( N \) RVs \( \{Y_n, n = 1, \ldots, N\} \) can be expressed as a weighted sum of the \((N^2)\) pair-wise (co)variances of these \( N \) RVs.

**In a spatial context:** \( Y_n = Y(s_n) \), i.e., \( n\)-th RV \( Y_n \) is defined at a location \( s_n \). Under 2nd-order stationarity, the mean of all RVs is the same, i.e., \( \mu_Y(s_n) = \mu_Y(s_{n'}) = \mu_Y \), and the covariance \( \sigma_Y(s_n, s_{n'}) \) between any two RVs is a function of the distance \( h_{nn'} \) between their respective locations \( s_n \) and \( s_{n'} \), i.e., \( \sigma_Y(s_n, s_{n'}) = \sigma_Y(s_n - s_{n'}) = \sigma_Y(h_{nn'}) \); this latter statement implies a constant variance for each RV: \( \sigma_Y(s_n, s_n) = \sigma_Y(s_n - s_n) = \sigma_Y(0) \)

**Criteria for Valid Covariogram/Semivariogram Models**

**Positive definiteness:** A (parametric or not) function of distance \( \sigma(h) \), defined in \( \mathcal{R}^d \), is called positive semi-definite, in \( \mathcal{R}^d \), if the variance of any weighted linear combination satisfies the condition:

\[ \mathbb{V}\left\{ \sum_{n=1}^{N} \lambda_n Y(s_n) \right\} = \sum_{n=1}^{N} \sum_{n'=1}^{N} \lambda_n \sigma_Y(h_{nn'}) \lambda_{n'} \geq 0, \quad \forall \lambda_n, \lambda_{n'}, h_{nn'} \]

The above statement \( \forall h_{nn'} \) implies for any configuration of sample locations; a function satisfying the above equation under the constraint: \( \sum_{n=1}^{N} \lambda_n = 0 \); is said to be conditionally positive semi-definite.

**Domain of validity:** If a function is positive definite in \( \mathcal{R}^d \), it is also in \( \mathcal{R}^{d'}, \forall d' \leq d \); the reverse is not always true.

**Valid models:** A covariogram model \( \sigma(h) \) is valid only if it is positive definite (in appropriate \( \mathcal{R}^d \)); a semivariogram model \( \gamma(h) \) is valid only if \( -\gamma(h) \) is conditionally positive definite.

Not all functions of distance are positive definite (especially in \( \mathcal{R}^{d>1} \)); but some functions are known to be, hence can serve as valid covariogram/semivariogram models.
Valid Semivariogram Models (1)

Pure nugget effect (valid in $\mathcal{R}^d$):

$$\gamma(h; \theta) = \sigma(0) [1 - \delta_{h0}] = \begin{cases} 0, & \text{if } h = 0 \\ \sigma(0), & \text{if } h > 0 \end{cases} \quad \theta = [\sigma(0)]$$

Kronecker delta function $\delta_{ij} = 1$ if $i = j$, 0 otherwise;

$$\sigma(0) = \text{sill} \approx \text{Y-variance } \text{Var}\{Y(s)\}$$

Slide 9

- indicates complete absence of spatial correlation
- could correspond to measurement error and/or microstructure, i.e., features occurring at scales smaller than sampling interval

Valid Semivariogram Models (2)

Spherical (valid up to $\mathcal{R}^3$):

$$\gamma(h; \theta) = \begin{cases} \sigma(0) \left[\frac{3}{2} \left(\frac{h}{r}\right) - \frac{1}{2} \left(\frac{h}{r}\right)^3\right], & \text{if } h < r \\ \sigma(0), & \text{if } h \geq r \end{cases} \quad \theta = [\sigma(0) \ r]$$

Slide 10

- linear behavior at origin
- range parameter $r$
Valid Semivariogram Models (3)

**Exponential** (valid in $\mathcal{R}^d$):

$$\gamma(h; \theta) = \sigma(0) \left[ 1 - \exp\left( -\frac{3h}{r} \right) \right] \quad \theta = [\sigma(0) \ r]$$

- linear behavior at origin; rises faster than spherical; reaches sill asymptotically
- effective range parameter $r$; distance at which 95% of sill reached

Valid Semivariogram Models (4)

**Gaussian** (valid in $\mathcal{R}^d$):

$$\gamma(h; \theta) = \sigma(0) \left[ 1 - \exp\left( -\frac{3h^2}{r^2} \right) \right] \quad \theta = [\sigma(0) \ r]$$

- quadratic behavior at origin; reaches sill asymptotically
- effective range parameter $r$; distance at which 95% of sill reached
Valid Semivariogram Models (5)

**Stable** (valid in $\mathcal{R}^d$):

$$\gamma(h; \theta) = \sigma(0) \left[ 1 - \exp \left( -\frac{3h^\omega}{r\omega} \right) \right] \quad 0 < \omega \leq 2 \quad \theta = [\sigma(0) \ r \ \omega]$$

- special cases: (i) $\omega = 1 \rightarrow$ exponential model, (ii) $\omega = 2 \rightarrow$ Gaussian model
- effective range parameter $r$; distance at which 95% of sill reached

Valid Semivariogram Models (6)

**Linear** (valid in $\mathcal{R}^d$):

$$\gamma(h; \theta) = \sigma(0) \cdot h \quad \theta = [\sigma(0)]$$

- unbounded (without sill) semivariogram model; implies self-similarity
- link to Brownian motion in 1D
### Valid Semivariogram Models (7)

**Power** (valid in $\mathbb{R}^d$):

$$\gamma(h; \theta) = \sigma(0) h^\omega \quad 0 < \omega < 2 \quad \theta = [\sigma(0) \ \omega]$$

- unbounded (without sill) semivariogram model; implies self-similarity
- for $\omega = 0$: $\gamma(h) = \sigma(0) \rightarrow$ pure nugget effect;
- for $\omega = 1$: $\gamma(h) = \sigma(0) h \rightarrow$ linear semivariogram;
- for $\omega > 2$: non-stationary random field
- link to self-affine random fractals: $D = E + 1 - \omega/2$, where $D =$ fractal dimension, and $E =$ topological dimension of space ($E = 1, 2, 3$)

### Valid Semivariogram Models (8)

**Hole-effect** of cosine (valid in $\mathbb{R}^1$):

$$\gamma(h; \theta) = \sigma(0) \left[ 1 - \cos \left( \frac{2\pi h}{r} \right) \right] \quad \theta = [\sigma(0) \ \ r]$$

- indicates periodic spatial variability
- distance from origin to first peak = size of underlying cyclic features

$sill =$ amplitude, $range =$ wavelength
Valid Semivariogram Models (9)

**Cardinal sine** (valid up to $\mathcal{R}^3$):

$$\gamma(h; \theta) = \sigma(0) \left[ 1 - \left( \frac{r}{2\pi h} \right) \sin \left( \frac{2\pi h}{r} \right) \right]$$

- indicates periodic spatial variability
- distance from origin to first peak = size of underlying cyclic features

Combinations of Semivariogram Models (1)

One can build complex, yet positive definite, covariogram/semivariogram models by combining simpler positive definite models, each possibly having its own anisotropy.

Sums of valid models: Any (positive) weighted linear combination of $K$ valid correlogram models is a valid correlogram model; same holds true for standardized (to unit sill) semivariogram models:

$$\sigma(h; \theta) = \sum_{k=1}^{K} a_k \rho_k(h; \theta_k) \quad \text{or} \quad \gamma(h; \theta) = \sum_{k=1}^{K} a_k g_k(h; \theta_k)$$

$\text{a}_k > 0$ denotes partial sill of $k$-th model; $g_k(h; \theta)$ denotes $k$-th, unit-sill, semivariogram model

Products of valid models:

$$\sigma(h; \theta) = \prod_{k=1}^{K} a_k \rho_k(h; \theta_k)$$

product of valid semivariogram models does not always yield a valid semivariogram model

Domain of validity: A function derived from $K$ positive definite functions, each valid in $\mathcal{R}^{d_k}$, is generally valid (positive definite) in $\mathcal{R}^{d'}$, where $d' = \min\{d_k, k = 1, \ldots, K\}$
Combinations of Semivariogram Models (2)

Nugget + Exponential (valid in $\mathbb{R}^d$):

$$\gamma(h; \theta) = a \left[ 1 - \delta_{h0} \right] + \left[ \sigma(0) - a \right] \left[ 1 - \exp \left( \frac{3h}{r} \right) \right] \quad \theta = \left[ \sigma(0), a, r \right]$$

Characteristics:

- discontinuous at origin; reaches sill asymptotically
- practical range parameter $r$; distance at which 95% of sill reached
- $a/\sigma(0)$ = relative nugget contribution = proportion (to total sill) of purely random spatial variability

Combinations of Semivariogram Models (3)

Nugget + two spherical models (valid in $\mathbb{R}^3$):

$$\gamma(h) = 0.1 + 0.4Sph \left( \frac{h}{5} \right) + 0.5Sph \left( \frac{h}{15} \right) \quad \theta = [0.1, 0.4, 0.5, 5, 15]$$

Characteristics:

- discontinuous at origin due to nugget effect
- two range parameters: $r_1 = 5$ due to small-scale (short-distance) spatial variability, and $r_2 = 15$ due to larger-scale (longer-distance) spatial variability
Combinations of Semivariogram Models (4)

Dampened hole-effect model (conditionally valid in $\mathcal{R}^3$):

$$\gamma(h; \theta) = \sigma(0) \cdot \left[ 1 - \cos \left( \frac{\pi h}{r_1} \right) \exp \left( -\frac{h}{r_2} \right) \right]$$

Characteristics:
- product of two valid semivariogram models
- intensity (amplitude) of periodic spatial variability levels off with increasing distance
- valid in $\mathcal{R}^2$ iff $r_1 \geq r_2$; valid in $\mathcal{R}^3$ iff $r_1 \geq r_2 \sqrt{3}$

Recap

Sample semivariogram:
- an exploratory analysis tool used to quantify spatial association in a set of measurements sampled from a continuous (in this handout) population
- a set of semivariance values plotted on a sample semivariogram do not constitute a semivariogram model for the population

Model semivariogram:
- goes beyond the sample semivariogram and provides a model of spatial association for the continuous population itself
- semivariogram models are typically expressed as parametric functions of distance. Not all functions of distance, however, can serve as valid semivariogram models; some are known to be valid, so use those or their combinations
- A useful analogy: Think of a model semivariance value $\gamma(h)$ for an arbitrary distance $h$ as the moment of inertia of an imaginary scatter plot of lagged attribute pairs for that distance

The role of prior information: Physical or other spatial process models can be useful in guiding the selection of an appropriate semivariogram model, and can sometime compensate for the lack of abundant data