**Model Semivariogram**

**Definition:** A (typically parametric) function of distance $\gamma(h; \theta)$ often “fitted” to $L$ sample semivariance values $\{\hat{\gamma}(h_l), l = 1, \ldots, L\}$; here, $h$ denotes an arbitrary distance value, $h_l$ denotes one of the $L$ lag distances used in the construction of the empirical (sample) semivariogram, and $\theta$ is a parameter vector typically containing the range and sill for a given semivariogram model type or functional form.

**Objective of this handout:**
- to highlight the basics of (non-linear) parameter estimation for models of spatial structure
- **Notes:** semivariogram model parameter estimation is also known as “fitting” a semivariogram model to sample data; this often amounts to fitting a semivariogram model to an empirical semivariogram plot obtained from sample data; there is one parameter estimation approach (maximum likelihood) which does not involve the sample semivariogram plot, but the covariogram cloud.

**Model Proposals (1)**

**Nugget model:**

$$\gamma(h; \beta) = \beta, \forall h > 0$$

**Cases shown:** $\gamma(h_l; 5) = 5$ and $\gamma(h_l; 8) = 8$, $\forall h_l > 0$, with $h_l$ denoting the discrete lag distance used to compute the sample semivariogram $\hat{\gamma}(h_l)$.
Model Proposals (2)

Linear model:

\[ \gamma(h; \beta) = \beta h \]

Slide 3

Cases shown: \( \gamma(h_i; 8) = 8h_i \) and \( \gamma(h_i; 12) = 12h_i \), with \( h_i \) denoting the discrete lag distance used to compute the sample semivariogram \( \hat{\gamma}(h_i) \)

Model Proposals (3)

Exponential model:

\[ \gamma(h; \beta, \theta) = \beta \left[ 1 - \exp\left(\frac{-3h}{\theta}\right) \right] \]

Slide 4

Cases shown: \( \gamma(h_i; 5, 1) = 5[1 - \exp(-3h_i/1)] \) and \( \gamma(h_i; 10, 1) = 10[1 - \exp(-3h_i/1)] \), with \( h_i \) denoting the discrete lag distance used to compute the sample semivariogram \( \hat{\gamma}(h_i) \)
Combinations of Variogram Models (1)

One can build complex, yet valid, semivariogram or covariogram models by combining simpler valid models, each possibly having its own anisotropy.

**Theorem I:** any weighted sum of valid semivariogram models is also a valid model:

\[
\gamma(h; \theta) = \sum_{m=0}^{M} \beta_m \gamma_m(h; \theta_m), \quad \beta_m \geq 0
\]

\(M\) denotes \# of nested models, \(\beta_m\) denotes partial sill of \(m\)-th model \(\gamma_m(h, \theta_m)\) with range \(\theta_m\);

by convention, \(\gamma_0(h; \theta_0)\) corresponds to the pure nugget effect model, hence \(\theta_0 = \epsilon\)

\(\theta = [\beta_m, m = 0, \ldots, M; \theta_m, m = 0, \ldots, M]^T\)

**Theorem II:** any weighted product of valid semivariogram models is also a valid model:

\[
\gamma(h; \theta) = \prod_{m=0}^{M} \beta_m \gamma_m(h; \theta_m), \quad \beta_m \geq 0
\]

actually, this is not always true for some models

A model derived from \(M\) valid models, each valid in \(\mathcal{R}^{d_m}\), is generally valid in \(\mathcal{R}^{d'}\), where \(d' = \min\{d_m, m = 1, \ldots, M\}\)

Combinations of Variogram Models (2)

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

\[
\gamma(h; \theta) = \beta_0 [1 - \delta_{h0}] + \beta_1 \left[ 1 - \exp \left( \frac{3h}{\theta_1} \right) \right]
\]

\(\delta_{h0} = 1, \text{ if } h = 0, 0 \text{ if not}, \quad \theta = [\beta_0, \beta_1, \epsilon, \theta_1]^T\)

**Characteristics:**
- discontinuous at origin; reaches sill asymptotically
- practical range parameter \(\theta_1\); distance at which 95% of sill reached
- \(\beta_0/(\beta_0 + \beta_1)\) = relative nugget contribution = proportion (to total sill) of purely random spatial variability
### Model Proposals (4)

**Exponential + Nugget model:**

\[
\gamma(h; \beta_1, \beta_2, \theta_1) = \beta_0 + \beta_1 \left[ 1 - \exp\left( -\frac{3h}{\theta_1} \right) \right], \quad \forall h > 0
\]

Cases shown: \(\gamma(h_1; 1.5, 1) = 1 + 5[1 - \exp(-3h_1/1)]\) and \(\gamma(h_1; 2, 7, 1) = 2 + 7[1 - \exp(-3h_1/1)]\), with \(h_1\) denoting the discrete lag distance used to compute the sample semivariogram \(\hat{\gamma}(h_1)\)

### Parameter Estimation (“Fitting”) Alternatives

**Manual fitting:**

- select number of semivariograms, their types, sills, and ranges
- in anisotropic cases, ensure consistency between models for directional semivariograms, i.e., determine single model which applies (differently) in any direction
- model behavior at origin, i.e., nugget effect, shape of semivariogram at distances smaller than first lag, using prior knowledge about the smoothness of the spatial attribute

**Automatic fitting:**

- least squares fit (ordinary, generalized, weighted): choose semivariogram model parameters (typically iteratively) so as to minimize discrepancy between model and sample semivariogram values over all lags
- maximum likelihood, and Bayesian methods also available
- treat with caution, especially with sparse data and outliers

**Leave-one-out cross-validation:**

- (i) hide one sample, (ii) estimate its value from remaining data, (iii) repeat for all samples
- repeat for different semivariogram models, and look at statistics of cross-validation errors for selecting “optimal” model
Objective Function To be Minimized (1)

Weighted sum of squared mismatches or errors (WSSE) between observed and model semivariogram values

**Isotropic case:**

\[ O(\theta) = \sum_{l=1}^{L} \omega_l \left[ \hat{\gamma}(h_l) - \gamma(h_l; \theta) \right]^2 \]

\( \hat{\gamma}(h_l) \) = sample semivariogram at lag distance \( h_l = \|h_l\| \),
\( \gamma(h_l; \theta) \) = model semivariogram at lag distance \( h_l \),
\( \omega_l \) = weight for \( l \)-th squared mismatch value \( [\hat{\gamma}(h_l) - \gamma(h_l; \theta)]^2 \) at lag distance \( h_l \)

**Anisotropic case:**

\[ O(\theta) = \sum_{k=1}^{K} \sum_{l=1}^{L} \omega^k_l \left[ \hat{\gamma}^k(h^k_l) - \gamma^k(h^k_l; \theta) \right]^2 \]

\( \hat{\gamma}^k(h^k_l) \) = sample semivariogram at lag \( h^k_l \) for \( k \)-th direction,
\( \gamma^k(h^k_l; \theta) \) = model semivariogram at \( h^k_l \) for \( k \)-th direction
\( \omega^k_l \) = weight for \( l \)-th squared mismatch value \( [\hat{\gamma}^k(h^k_l) - \gamma^k(h^k_l; \theta)]^2 \) at lag distance \( h^k_l \) along \( k \)-th direction

**Goal:**

Find parameter vector \( \theta \) that minimizes \( O(\theta) \)

*Note:* Not that straightforward, because for most models \( O(\theta) \) is non-linear in \( \theta \)

Need to: (i) use iterative algorithms or linearize the models and (ii) account for non-negativity constraints on entries of \( \theta \)

*Note:* sample semivariograms derived from bad choices of lag distances and tolerances, will inevitably lead to poor semivariogram models (garbage in = garbage out)

Objective Function To be Minimized (2)

\[ O(\theta) = \sum_{l=1}^{L} \omega_l \left[ \hat{\gamma}(h_l) - \gamma(h_l; \theta) \right]^2 \]

\( \hat{\gamma}(h_l) \) = sample semivariogram at lag distance \( h_l = \|h_l\| \), \( \gamma(h_l; \theta) \) = model semivariogram at lag distance \( h_l \),
\( \omega_l \) = weight for \( l \)-th squared mismatch value \( [\hat{\gamma}(h_l) - \gamma(h_l; \theta)]^2 \)

**Weighting schemes,** with \( N(h_l) \) denoting \# of sample pairs separated by lag distance \( h_l \):

1. \( \omega_l = 1, \forall l \) \( \rightarrow \) all mismatches receive equal weighting
2. \( \omega_l = N(h_l) \) \( \rightarrow \) mismatches calculated using more data pairs \( N(h_l) \) receive more weight
3. \( \omega_l = 1/\gamma(h_l; \theta)^2 \) \( \rightarrow \) mismatches at small lags \( h_l \) receive more weight, since \( \gamma(h_l; \theta) \) is small for small \( h_l \)
4. \( \omega_l = N(h_l)/\gamma(h_l; \theta)^2 \) \( \rightarrow \) weights inversely proportional to uncertainty of \( \hat{\gamma}(h_l) \):
   \( \text{Var}\{\hat{\gamma}(h_l)\} \simeq \gamma(h_l; \theta)^2/N(h_l) \); would be optimal if sample semivariogram values were independent
5. \( \omega_l = N(h_l)/\tilde{h}_l^2 \), where \( \tilde{h}_l \) = average distance between sample pairs contributing to \( \hat{\gamma}(h_l) \) \( \rightarrow \) mismatches computed at smaller lags using more sample pairs receive more weight
Matrix Expressions

\[ O(\theta) = \sum_{l=1}^{L} \omega_l [\hat{\gamma}(h_l) - \gamma(h_l; \theta)]^2 \]

\( \hat{\gamma}(h_l) \) = sample semivariogram at lag distance \( h_l = |h_l| \), \( \gamma(h_l; \theta) \) = model semivariogram at lag distance \( h_l \), \( \omega_l \) = weight for \( l \)-th mismatch value

More general formulation:

\[ O(\theta) = [\hat{\gamma} - \gamma(\theta)]^T \Omega [\hat{\gamma} - \gamma(\theta)] \]

\( \hat{\gamma} = [\hat{\gamma}(h_l), l = 1, \ldots, L]^T \): vector of sample variogram values for \( L \) lags, \( \gamma(\theta) = [\gamma(h_l; \theta), l = 1, \ldots, L]^T \): vector of model values at \( L \) lags, \( \Omega = [\omega_{ll'}, l = 1, \ldots, L, l' = 1, \ldots, L] \): matrix of weights

Alternatives for \( \Omega \):

1. \( \Omega = I \): \((L \times L)\) identity matrix \(\rightarrow\) Ordinary Least Squares (OLS); not recommended because mismatches are not independent, and their uncertainty is not the same across all \( L \) lags

2. \( \Omega \) with zero off-diagonal entries, and non-zero but different diagonal entries \(\rightarrow\) Weighted Least Squares (WLS) using iterative procedures with easy-to-invert matrices (most widely used)

3. \( \Omega = \) inverse variance-covariance matrix of sample semivariogram, which can be calculated from a hypothesized semivariogram model \(\rightarrow\) Generalized Least Squares (GLS) using iterative procedures that require full matrix inversion

Minimization of Objective Function

Problem statement: find \( \theta \) such that: \( O(\theta) = f(\theta) \to \text{min} \)

Caveats:

- more entries in \( \theta \) make \( O(\theta) \) more complex, and hence render minimization more difficult and slower
- some minimization algorithms require good initial starting values for \( \theta \) to even give meaningful results
- other algorithms require derivative information (to help them find the direction of the minimum); these derivatives can be numerically evaluated (slow) or can be analytically provided (difficult if \( O(\theta) \) is complex)
**Goodness-of-Fit Measures**

Need to balance over-fitting against parsimony:

A model with many parameters might provide an excellent fit to the data, but need not be physically justified.

**Akaike’s Information Criterion (AIC):**

\[
\hat{AIC} = L \ln \left( \frac{O(\hat{\theta})}{L} \right) + 2P \quad \text{or} \quad \hat{AIC} = L + L \ln(2\pi) + L \ln(O(\hat{\theta})/L) + 2(P + 1)
\]

\(L = \# \text{ of lags considered in sample variogram} \) (this plays the role of the \# of data);
\( O(\hat{\theta}) = \text{minimum objective function} \);
\( P = \text{number of parameters (entries of vector } \theta \text{) that lead to } O(\hat{\theta}) \)

Models with larger \( O(\hat{\theta}) \), i.e., with larger overall mismatch as measured by \( WSSE(\hat{\theta}) \), and larger \( P \), i.e., more complex, are penalized more equivalent to an \( F \) test for adding more nested structures in a model.

**Bayesian Information Criterion (BIC):**

\[
\hat{BIC} = L \ln \left( \frac{O(\hat{\theta})}{L} \right) + \ln(L)P \quad \text{or} \quad \hat{BIC} = L + L \ln(2\pi) + L \ln(O(\hat{\theta})/L) + \ln(L)(P + 1)
\]

Similar to AIC, but with more importance on \( P \) as \( P \) becomes larger \( \hat{BIC} > \hat{AIC} \) (more penalty to complex models).

Choose models whose parameters yield smallest \( \hat{AIC} \) or \( \hat{BIC} \)

**(Multi-Gaussian) Maximum Likelihood Fitting (1)**

**Likelihood of data given a set of parameters:**

\( f(y|\theta) \)

- \( y = [y(s_n), n = 1, \ldots, N]^T = (N \times 1) \) vector of sample attribute values obtained at \( N \) supports; these \( N \) attribute values are regarded as one joint outcome of \( N \) correlated RVs \( \{Y(s_n), n = 1, \ldots, N\} \) constituting a random vector; \( \theta \) vector of covariogram model parameters

\[
f(y|\theta) = \prod_{i=1}^{N} f(y(s_i)|\theta) = \prod_{i=1}^{N} \Phi(Y(s_i) - y(s_i)|\theta)
\]

- \( \mu = [\mu_n, n = 1, \ldots, N]^T = (N \times 1) \) vector of known mean values: \( \mu_n = E\{Y(s_n)\} \); under 2nd-order stationarity, all \( N \) entries of \( \mu \) are assumed equal to a constant, i.e., \( \mu_n = \mu, \forall n \)

\( \Sigma(\theta) = \{\sigma(s_n, s_{n'}; \theta), n = 1, \ldots, N, n' = 1, \ldots, N\} = (N \times N) \) matrix of model covariance values \( \sigma(s_n, s_{n'}; \theta) = \text{Cov}\{Y(s_n), Y(s_{n'})\} \) between any support pair \( s_n \) and \( s_{n'} \); under 2nd-order stationarity, any entry \( \sigma(s_n, s_{n'}; \theta) \) is assumed to be a function only of the separation vector \( h_{nn'} \) between \( s_n \) and \( s_{n'} \), i.e., \( \sigma(s_n, s_{n'}; \theta) = \sigma(h_{nn'}; \theta) \)

**Objective:** For point data, determine parameter vector \( \theta \) of covariogram model \( \sigma(h; \theta) \) used to populate \( \Sigma(\theta) \), such that above likelihood is maximum.
Multi-Gaussian) Maximum Likelihood Fitting (2)

\( N \)-variate Gaussian likelihood function:

\[
f(y|\theta) = (2\pi)^{-N/2} \det[\Sigma(\theta)]^{-1/2} \exp\left(-\frac{1}{2}[y - \mu]^T \Sigma(\theta)^{-1} [y - \mu]\right)
\]

Negative log-likelihood:

\[
L(\theta, y) = \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det[\Sigma(\theta)]) + \frac{1}{2}[y - \mu]^T \Sigma(\theta)^{-1} [y - \mu]
\]

Instead of maximizing \( f(y|\theta) \), minimize \( L(\theta, y) \),
which leads to the same solution vector \( \theta \) and is faster.

Issues and extensions:

- negative log-likelihood is non-linear in the parameter vector \( \theta \) → iterative optimization algorithms must be used
- above estimation procedure calls for multiple inversions of a possibly large \( (N \times N) \) covariance matrix \( \Sigma(\theta) \) → slow and prohibitive for large data sets
- extensions, such as Restricted Maximum Likelihood, and shortcuts, such as Approximate Maximum Likelihood → all rely on the multi-variate Gaussian assumption
- BUT, there is no need for deciding on (admittedly arbitrary) lag distances, directions, tolerances, and such . . .

Recap

Fitting spatial association models to data:

- once a particular parametric variogram/covariogram model (or a combination of models) is selected, parameter estimation is the procedure of finding those model parameters that minimize some goodness-of-fit criterion; this typically involves non-linear minimization

Least squares fitting methods:

- in semivariogram model fitting via weighted least squares, for example, the observations from which that goodness-of-fit criterion is computed are not the actual attribute data, but the sample semivariogram values \( \{\hat{\gamma}(h_l), l = 1, \ldots, L\} \) obtained at a set of \( L \) lag distances

Maximum likelihood fitting methods:

- in semivariogram model fitting via Gaussian maximum likelihood, for example, the goodness-of-fit criterion involves maximizing the probability of observing the particular attribute data given the model parameters

All methods have their pros and cons;
if you do not suspect the presence of trends (first-order effects) in your model,
choose least-squares methods . . .