Suppose our spatial process has a mean, $\mu(s) = E(Y(s))$, and that the variance of $Y(s)$ exists for all $s \in D$.

- **Strong stationarity**: If for any given set of sites, and any displacement $h$, the distribution of $(Y(s_1), ..., Y(s_n))$ is the same as, $(Y(s_1 + h), ..., Y(s_n + h))$.
- **Weak stationarity**: Constant mean $\mu(s)$, and $Cov(Y(s), Y(s+h)) = C(h)$: the covariance depends only upon the displacement (or separation) vector.
- **Strong stationarity implies weak stationarity**
- Process is defined through finite dimensional distributions
- The process is Gaussian (a GP) if $Y = (Y(s_1), ..., Y(s_n))$ has a multivariate normal distribution.
- For GPs, strong and weak stationarity equivalent.

### Variograms

- **Suppose we assume** $E[Y(s + h) - Y(s)] = 0$ and define $E[(Y(s + h) - Y(s))^2] = Var(Y(s + h) - Y(s)) = 2\gamma(h)$.

  This is sensible if the left hand side depends only upon $h$. Then we say the process is **intrinsically stationary**.

- $\gamma(h)$ is called the **semivariogram** and $2\gamma(h)$ is called the variogram.

- **Where did $\gamma(h)$ come from?**

  Note that intrinsic stationarity defines only the first and second moments of the differences $Y(s + h) - Y(s)$. It says nothing about the joint distribution of a collection of variables $Y(s_1), ..., Y(s_n)$, and thus provides no likelihood.

### Intrinsic Stationarity and Ergodicity

- **Relationship between $\gamma(h)$ and $C(h)$**:

  $2\gamma(h) = Var(Y(s + h)) + Var(Y(s)) - 2Cov(Y(s + h), Y(s)) = C(0) + C(0) - 2C(h) = 2[C(0) - C(h)]$.

- **Easy to recover $\gamma$ from $C$**. The converse needs the additional assumption of ergodicity: $\lim_{|u| \to \infty} C(u) = 0$.

- **So $\lim_{|u| \to \infty} \gamma(u) = C(0)$, and we can recover $C$ from $\gamma$ as long as this limit exists**.

  $$C(h) = \lim_{|u| \to \infty} \gamma(u) - \gamma(h).$$

- **When $\gamma(h)$ or $C(h)$ depends upon the separation vector only through the distance $|h|$, we say that the variogram or the covariance function is isotropic, respectively. In that case, we write $\gamma(|h|)$ or $C(|h|)$**. Otherwise we say that they are anisotropic.

- **If the process is intrinsically stationary with an isotropic covariance function, it is also called homogeneous**.

Isotropic models are popular because of their simplicity, interpretability, and because a number of relatively simple parametric forms are available as candidates for $C$ (and $\gamma$). Denoting $|h|$ by $t$ for notational simplicity, the next two tables provide a few examples.
## Stationary Gaussian Processes

### Isotropy

#### Examples: Spherical Variogram

![Spherical Variogram Graph](image)

### Model Covariance Function, $C(t)$

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariance Function, $C(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$C(t) = \begin{cases} \tau^2 + \sigma^2 &amp; \text{if } t &gt; 0 \ \sigma^2 \exp(-\phi t) &amp; \text{if } t &gt; 0 \end{cases}$</td>
</tr>
<tr>
<td>Spherical</td>
<td>$C(t) = \begin{cases} \tau^2 + \sigma^2 &amp; \text{if } t &gt; 0 \ \frac{1}{2} \phi t - \frac{1}{2} (\phi t)^2 &amp; \text{if } 0 &lt; t \leq 1/\phi \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$C(t) = \begin{cases} \tau^2 + \sigma^2 (1 - \exp(-\phi t)) &amp; \text{if } t &gt; 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>Powered exponential</td>
<td>$C(t) = \begin{cases} \tau^2 + \sigma^2 (1 - \exp(-</td>
</tr>
<tr>
<td>Matérn at $\nu = 3/2$</td>
<td>$C(t) = \begin{cases} \tau^2 + \sigma^2 [1 - (1 + \phi t) e^{-\phi t}] &amp; \text{if } t &gt; 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
</tbody>
</table>

### Notes on exponential model

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases}$$

- We define the effective range, $t_0$, as the distance at which this correlation has dropped to only 0.05. Setting $\exp(-\phi t_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$.
- Finally, the form of $C(t)$ shows why the nugget $\tau^2$ is often viewed as a "nonspatial effect variance," and the partial sill ($\sigma^2$) is viewed as a "spatial effect variance."
- Note discontinuity at 0. Intentional!

### Valid covariance functions

- Need care in selecting $C(t)$
- Needs to be valid. For all $n$ and all $\{s_1, s_2, ..., s_n\}$, the resulting covariance matrix for $\{Y(s_1), Y(s_2), ..., Y(s_n)\}$ must be positive definite
- So, $C(t)$ needs to be a positive definite function
- Much theory, characterization, construction, exemplification in the literature
- Why GPs are attractive - only need a mean function and a valid covariance function
The Matérn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms
- The Matérn is a very versatile family:
  \[ C(t) = \begin{cases} \frac{\sigma^2}{\gamma(\nu)} \left( 2 \sqrt{\pi t} \right)^{\nu} \frac{K_{\nu}(2 \sqrt{t \gamma(\nu)})}{\Gamma(\nu)} & \text{if } t > 0 \\ \sigma^2 & \text{if } t = 0 \end{cases} \]
  where \( K_{\nu} \) is the modified Bessel function of order \( \nu \) (computationally tractable)
- \( \nu \) is a smoothness parameter (a fractal) controlling process smoothness. Remarkable!

Variogram model fitting

- How do we select a variogram? Can the data really distinguish between variograms?
- Empirical Variogram:
  \[ \gamma(t) = \frac{1}{2N(t)} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2 \]
  where \( N(t) \) is the number of points such that \( \|\mathbf{s}_i - \mathbf{s}_j\| = t \) and \( |N(t)| \) is the number of points in \( N(t) \).
- Grid up the space into intervals \( I_1 = (0, t_1), I_2 = (t_1, t_2), \) and so forth, up to \( I_K = (t_{K-1}, t_K) \). Representing \( t \) values in each interval by its midpoint, we define:
  \[ N(tk) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k \}, k = 1, \ldots, K. \]

Empirical variogram: scallops data

- Should we fit a variogram?
- \( N(t) \) varies with \( t \)
- Squared differences are not independent
- Should we average squared differences?
- Weighted variogram estimators?
- No inference, no uncertainty; EDA at best

Anisotropy

- Here association depends upon the separation vector between locations (i.e., direction and distance).
- Example: geometric anisotropy, where
  \[ c(\mathbf{s} - \mathbf{s}') = \sigma^2 \rho((\mathbf{s} - \mathbf{s}')^T B (\mathbf{s} - \mathbf{s}')). \]
  \( B \) is positive definite with \( \rho \) a valid correlation function in \( \mathbb{R}^r \) (e.g., exponential).
  - We omit the range/decay parameter \( \phi \) since we now use \( B \);
  - when \( r = 2 \) this means three "range parameters" instead of one.
  - Contours of constant association arising from \( c \) above are elliptical. In particular, the contour corresponding to \( \rho = .05 \) provides the effective range in each spatial direction.
First step in analyzing data
First Law of Geography: Mean + Error
Mean: first-order behavior
Error: second-order behavior (covariance function)
EDA tools examine both first and second order behavior
Preliminary displays: Simple locations to surface displays

for any function $g$

Conditional expectation of $Y$ given $x$, conditioning on $x$.

Just the ordinary kriging estimate. We seek the function $g(y)$ that minimizes the mean-squared prediction error, $E[(Y(s_0) - g(y))^2 | y]$.

But this expression equals $E((Y(s_0) - E[Y(s_0) | y])^2 | y) + (E[Y(s_0) | y] - g(y))^2$.

Since the second term is nonnegative, we have $E[(Y(s_0) - g(y))^2 | y] \geq E((Y(s_0) - E[Y(s_0)] | y)^2 | y)$ for any function $g(y)$.

We seek the function $g(y)$ that minimizes the mean-squared prediction error, $E[(Y(s_0) - g(y))^2 | y]$.

But this expression equals $E((Y(s_0) - E[Y(s_0) | y])^2 | y) + (E[Y(s_0) | y] - g(y))^2$.

Since the expectation of the cross-product term is 0.

Equality holds above $\iff g(y) = E[Y(s_0) | y]$, so it must be that the predictor $g(y)$ that minimizes the error is the conditional expectation of $Y(s_0)$ given the data.

Intuitive from a Bayesian point of view, since this $f(y)$ is just the posterior mean of $Y(s_0)$.

Using standard conditional normal distribution calculations, we obtain $E[Y(s_0) | y] = \beta^T y + \sum \Sigma^{-1} (y - X \beta)$, and $\text{Var}[Y(s_0) | y] = \sigma^2 + \tau^2 - \gamma^2 \Sigma^{-1} \gamma$.

For a model having a nugget effect, we instead set $\gamma = 0$.

With an estimate of $\rho$ and $\sigma$.

Under squared error loss, the best linear prediction is $\text{arg min}_\gamma \{ E[(Y(s_0) - E[Y(s_0) | y])^2 | y] \}$.

Since the second term is nonnegative, we have $E[(Y(s_0) - g(y))^2 | y] \geq E((Y(s_0) - E[Y(s_0)] | y)^2 | y)$ for any function $g(y)$.

Other than intrinsic stationarity, no distributional assumptions are required for the $Y(s_i)$.

For a model having a nugget effect, we instead set $\gamma = 0$.

With an estimate of $\rho$ and $\sigma$.

Under squared error loss, the best linear prediction is $\text{arg min}_\gamma \{ E[(Y(s_0) - E[Y(s_0) | y])^2 | y] \}$.

Since the second term is nonnegative, we have $E[(Y(s_0) - g(y))^2 | y] \geq E((Y(s_0) - E[Y(s_0)] | y)^2 | y)$ for any function $g(y)$.

For a model having a nugget effect, we instead set $\gamma = 0$.

With an estimate of $\rho$ and $\sigma$.

Under squared error loss, the best linear prediction is $\text{arg min}_\gamma \{ E[(Y(s_0) - E[Y(s_0) | y])^2 | y] \}$.

Since the second term is nonnegative, we have $E[(Y(s_0) - g(y))^2 | y] \geq E((Y(s_0) - E[Y(s_0)] | y)^2 | y)$ for any function $g(y)$.

For a model having a nugget effect, we instead set $\gamma = 0$.

With an estimate of $\rho$ and $\sigma$.

Under squared error loss, the best linear prediction is $\text{arg min}_\gamma \{ E[(Y(s_0) - E[Y(s_0) | y])^2 | y] \}$.

Since the second term is nonnegative, we have $E[(Y(s_0) - g(y))^2 | y] \geq E((Y(s_0) - E[Y(s_0)] | y)^2 | y)$ for any function $g(y)$.

For a model having a nugget effect, we instead set $\gamma = 0$.
Classical spatial prediction (Kriging)

- These are not estimators; they are really $E[Y(s_0) | y, \theta]$ and $\text{Var}[Y(s_0) | y, \theta]$. Parameters are unknown.

- Plug in estimates of the parameter?

- Estimators are no longer linear, no longer unbiased, don’t account for the uncertainty in the parameter estimates

- Can we do satisfying inference with these estimators?

- A cleaner way: the posterior predictive distribution of $Y(s_0) | y$