Elements of Point Level Modeling

Sudipto Banerjee\textsuperscript{1}, Bradley P. Carlin\textsuperscript{1} and Alan E. Gelfand\textsuperscript{2}

\textsuperscript{1}Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

\textsuperscript{2}Department of Statistical Sciences, Duke University, Durham, North Carolina.

August 02, 2014
Point-level modelling refers to modelling of spatial data collected at locations referenced by coordinates (e.g., lat-long, Easting-Northing).

**Fundamental concept:** Data from a spatial process \( \{Y(s) : s \in D\} \), \( D \) is a subset in Euclidean space.

**Example:** \( Y(s) \) is a pollutant level at site \( s \)

**Conceptually:** Pollutant level exists at all possible sites

**Practically:** Data will be a partial realization of a spatial process – observed at \( \{s_1, \ldots, s_n\} \)

**Statistical objectives:** Inference about the process \( Y(s) \); predict at new locations.

**Remarkable:** Can learn about entire \( Y(s) \) surface. The key: Structured dependence
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 $(Y(s_1), \ldots, Y(s_n))$ is the same as, $(Y(s_1 + h), \ldots, Y(s_n + h))$.

- **Weak stationarity:** Constant mean $\mu(s) = \mu$, 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 $Y = (Y(s_1), \ldots, Y(s_n))$ 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 $Y(s_1), \ldots, Y(s_n)$, and thus provides no likelihood.
Intrinsic Stationarity and Ergodicity

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

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

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

\[
\gamma(t) = \begin{cases} 
\tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\
\tau^2 + \sigma^2 \left[ \frac{3}{2} \phi t - \frac{1}{2} (\phi t)^3 \right] & \text{if } 0 < t \leq 1/\phi \\
0 & \text{if } t = 0.
\end{cases}
\]

- While \( \gamma(0) = 0 \) by definition, \( \gamma(0^+) \equiv \lim_{t \to 0^+} \gamma(t) = \tau^2 \); this quantity is the \textit{nugget}.
- \( \lim_{t \to \infty} \gamma(t) = \tau^2 + \sigma^2 \); this asymptotic value of the semivariogram is called the \textit{sill}. (The sill minus the nugget, \( \sigma^2 \) in this case, is called the \textit{partial sill}.)
- Finally, the value \( t = 1/\phi \) at which \( \gamma(t) \) first reaches its ultimate level (the sill) is called the \textit{range}, \( R \equiv 1/\phi \).
Examples: Spherical Variogram

b) spherical; $a_0 = 0.2, a_1 = 1, R = 1$
Some common isotropic covariograms

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariance function, $C(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$C(t)$ does not exist</td>
</tr>
</tbody>
</table>
| Spherical              | $C(t) = \begin{cases} 
0 & \text{if } t \geq 1/\phi \\
\sigma^2 \left(1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right) & \text{if } 0 < t \leq 1/\phi \\
\frac{\tau^2}{\tau^2 + \sigma^2} & \text{otherwise} 
\end{cases}$ |
| Exponential            | $C(t) = \begin{cases} 
\sigma^2 \exp(-\phi t) & \text{if } t > 0 \\
\frac{\tau^2}{\tau^2 + \sigma^2} & \text{otherwise} 
\end{cases}$ |
| Powered exponential    | $C(t) = \begin{cases} 
\sigma^2 \exp(-|\phi t|^p) & \text{if } t > 0 \\
\frac{\tau^2}{\tau^2 + \sigma^2} & \text{otherwise} 
\end{cases}$ |
| Matérn at $\nu = 3/2$ | $C(t) = \begin{cases} 
\sigma^2 \left(1 + \phi t\right) \exp(-\phi t) & \text{if } t > 0 \\
\frac{\tau^2}{\tau^2 + \sigma^2} & \text{otherwise} 
\end{cases}$ |
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 Matèrn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms
- The Matèrn is a very versatile family:

\[
C(t) = \begin{cases} 
\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu t})^\nu K_\nu(2\sqrt{\nu t}) & \text{if } t > 0 \\
\tau^2 + \sigma^2 & \text{if } t = 0
\end{cases}
\]

- \(K_\nu\) 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:

\[
\gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2
\]

where \( N(t) \) is the number of points such that \( \|s_i - s_j\| = t \) and \( |N(t)| \) is the number of points in \( N(t) \).

- Grid up the \( t \) 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:

\[
N(t_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, \ k = 1, \ldots, K.
\]
Empirical variogram: scallops data
Empirical variogram: scallops data

Parametric Semivariograms

Bessel Mixtures - Random Weights

Bessel Mixtures - Random Phi's
Should we fit a variogram?

\[ N(t) \text{ varies with } t \]

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

\[ c(s - s') = \sigma^2 \rho((s - s')^T B(s - s')) \]

- \( B \) 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
First Law of Geography

\[ \text{data} = \text{mean} + \text{error} \]
Classical spatial prediction or “Kriging”

- Named by Matheron (1963) in honor of D.G. Krige, a South African mining engineer whose seminal work on empirical methods for geostatistical data inspired the general approach.

- **Optimal spatial prediction**: given observations of a random field \( \mathbf{Y} = (Y(s_1), \ldots, Y(s_n))' \), predict the variable \( Y \) at a site \( s_0 \) where it has not been observed.

- Under squared error loss, the best linear prediction minimizes \( \mathbb{E}[Y(s_0) - (\sum \ell_i Y(s_i) + \delta_0)]^2 \) over \( \delta_0 \) and the \( \ell_i \).

- With an estimate of \( \gamma \), one immediately obtains the ordinary kriging estimate.

- Other than intrinsic stationarity, no distributional assumptions are required for the \( Y(s_i) \).
Kriging with Gaussian processes

- Given covariate values \( x(s_i), i = 0, 1, \ldots, n \), suppose

  \[
  Y = X\beta + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \Sigma).
  \]

- For a spatial covariance structure having no nugget effect, we specify \( \Sigma \) as

  \[
  \Sigma = \sigma^2 H(\phi) \quad \text{where} \quad (H(\phi))_{ij} = \rho(\phi; d_{ij}),
  \]

  where \( d_{ij} = ||s_i - s_j|| \), the distance between \( s_i \) and \( s_j \), and \( \rho \) is a valid correlation function on \( \mathbb{R}^r \).

- For a model having a nugget effect, we instead set

  \[
  \Sigma = \sigma^2 H(\phi) + \tau^2 I.
  \]
We seek the function $g(y)$ that minimizes the mean-squared prediction error,

$$E \left[ (Y(s_0) - g(y))^2 \mid y \right].$$

But this expression equals

$$E \left\{ (Y(s_0) - E[Y(s_0) \mid y])^2 \mid y \right\} + \{E[Y(s_0) \mid y] - g(y)\}^2,$$

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

Since the second term is nonnegative, we have

$$E \left[ (Y(s_0) - g(y))^2 \mid y \right] \geq E \left\{ (Y(s_0) - E[Y(s_0) \mid y])^2 \mid y \right\}$$

for any function $g(y)$. 


Classical spatial prediction (Kriging)

Equality holds above \( \iff \quad g(y) = E[Y(s_0)|y] \), 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

\[
E[Y(s_0)|y] = x_0^T \beta + \gamma^T \Sigma^{-1} (y - X \beta),
\]

and

\[
\text{Var}[Y(s_0)|y] = \sigma^2 + \tau^2 - \gamma^T \Sigma^{-1} \gamma.
\]
These are not estimators; They are really $E [Y(s_0)|\mathbf{y}, \theta]$ and $Var[Y(s_0)|\mathbf{y}, \theta]$. 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 $Y(s_0) | \mathbf{y}$