Variograms, covariance functions and kriging

Abhi Datta

04.07.2017

Review of last lecture

- Types of spatial data – point referenced, areal and point pattern data
- Exploratory data analysis for point referenced data – plotting, empirical variograms

$$\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$$

- Applicability of of ordinary linear regression
- How to incorporate spatial information in linear regression?

$$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$$

Review of last lecture

- Types of spatial data – point referenced, areal and point pattern data
- Exploratory data analysis for point referenced data – plotting, empirical variograms

$$\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$$

- Applicability of of ordinary linear regression
- How to incorporate spatial information in linear regression?

$$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$$

- One approach: Model $w(s)$ as a stochastic process instead of a deterministic function

Stochastic Processes

- Collection of random variables indexed by locations in continuous domain D: $\{Y(s) \,|\, s \in D\}$
- For any $s_1, s_2, \ldots, s_n$, $(Y(s_1), Y(s_2), \ldots, Y(s_n))'$ is a multivariate rv
- Kolmogorov's Consistency conditions for a well-defined stochastic process:
    - $(Y(s_1), Y(s_2), \ldots, Y(s_n)) \overset{d}{\sim} (Y(s_{\pi(1)}), Y(s_{\pi(2)}), \ldots, Y(s_{\pi(n)}))$ for any permutation π
    - $\int dF(Y(s_0), Y(s_1), Y(s_2), \ldots, Y(s_n)) ds_0 = dF(Y(s_1), Y(s_2), \ldots, Y(s_n))$

Intrinsic stationarity and semivariograms

- Strong stationarity: If for any given set of sites, and any displacement $\mathbf{h}$, the distribution of $(Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$ is the same as, $(Y(\mathbf{s}_1 + \mathbf{h}), ..., Y(\mathbf{s}_n + \mathbf{h}))$.
- Recall: Empirical variogram: $\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$
- Under strong stationarity: $E(Y(s+h) - Y(s))^2 = 2\gamma(h)$

Intrinsic stationarity and semivariograms

- Strong stationarity: If for any given set of sites, and any displacement $\mathbf{h}$, the distribution of $(Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$ is the same as, $(Y(\mathbf{s}_1 + \mathbf{h}), ..., Y(\mathbf{s}_n + \mathbf{h}))$.
- Recall: Empirical variogram: $\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$
- Under strong stationarity: $E(Y(s+h) - Y(s))^2 = 2\gamma(h)$
- Weaker condition
    - Intrinsic stationarity: $E(Y(s+h) - Y(s)) = 0$ and $E(Y(s+h) - Y(s))^2 = 2\gamma(h)$
    - $\gamma(h)$ is referred to as the semivariogram

Intrinsic stationarity and semivariograms

- Strong stationarity: If for any given set of sites, and any displacement $\mathbf{h}$, the distribution of $(Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$ is the same as, $(Y(\mathbf{s}_1 + \mathbf{h}), ..., Y(\mathbf{s}_n + \mathbf{h}))$.
- Recall: Empirical variogram:
  $\gamma(t_k) = \frac{1}{|N(t_k)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t_k)} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$
- Under strong stationarity: $E(Y(s+h) - Y(s))^2 = 2\gamma(h)$
- Weaker condition
    - Intrinsic stationarity: $E(Y(s+h) - Y(s)) = 0$ and $E(Y(s+h) - Y(s))^2 = 2\gamma(h)$
    - $\gamma(h)$ is referred to as the semivariogram
    - Can we characterize the stochastic process by modeling $\gamma(h)$?
    - What are some necessary assumptions on the function $\gamma(h)$?

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 $\mathbf{Y} = \left(Y\left(\mathbf{s}_1\right), \ldots, Y\left(\mathbf{s}_n\right)\right)'$, predict the variable Y at a site $\mathbf{s}_0$ where it has not been observed
- Under squared error loss, the best linear prediction minimizes $E[Y(\mathbf{s}_0) - \sum \ell_i Y(\mathbf{s}_i)]^2$ over ℓ_i
- Unbiased $\iff \sum \ell_i = 1$

Kriging

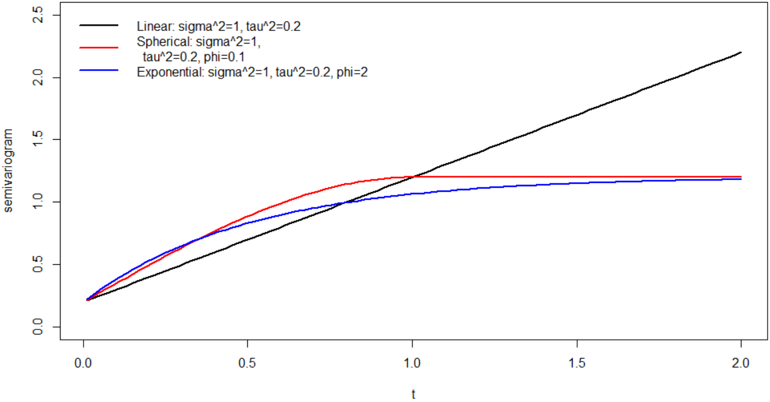
- Let $\Gamma = (\gamma(s_i - s_j))_{1\leq i,j\leq n}$ and $\gamma_0 = (\gamma(s_0 - s_1)), \gamma(s_0 - s_2)), \ldots, \gamma(s_0 - s_n)))'$
- Optimal $\ell = \Gamma^{-1}\left(\gamma_0 + \frac{(1-1'\Gamma^{-1}\gamma_0)}{1'\Gamma^{-1}1}1\right)$
- Minimum $E[Y(\mathbf{s}_0) - \sum \ell_i Y(\mathbf{s}_i)]^2 = \gamma_0'\Gamma^{-1}\gamma_0 - \frac{(1'\Gamma^{-1}\gamma_0-1)^2}{1'\Gamma^{-1}1}$
- With a model/estimate for γ, one immediately obtains the ordinary kriging estimate
- Other than intrinsic stationarity, no distributional assumptions are required for the $Y(\mathbf{s}_i)$
- Under intrinsic stationarity: For any $a_0, a_1, \ldots, a_n$, such that $\sum a_i = 0$, $E(\sum a_i Y(s_i))^2 = -\sum_{i,j} a_i a_j \gamma(s_i - s_j)$

Variogram models

- Necessary condition for valid variogram – Conditional negative definiteness: For any a with $a'1 = 0$, $-\sum_{i,j} a_i a_j \gamma(s_i - s_j) \geq 0$

model	Variogram, $\gamma(h) = \gamma(t)$ where $t = \|\|h\|\|$
Linear	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Spherical	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{otherwise} \end{cases}$
Exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Powered exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\|\phi t\|^p)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$

Variogram models

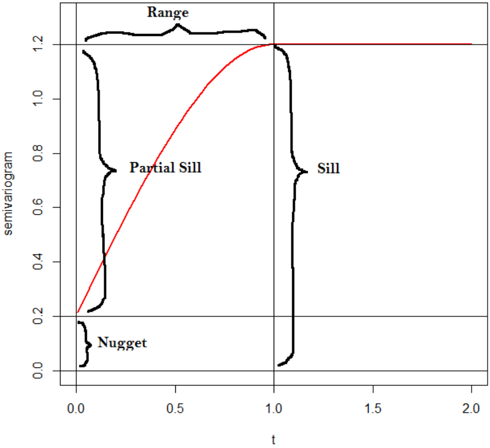


Examples: Spherical Variogram

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

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

Examples: Spherical Variogram



Variogram fitting using least squares

- Fit a parametric model to the empirical estimates $\hat{\gamma}(t_k)$ using least squares

$$\hat{\theta} = \arg\min \sum_k (\hat{\gamma}_(t_k) - \gamma(t_k \mid \theta))^2$$

Variogram fitting using least squares

- Fit a parametric model to the empirical estimates $\hat{\gamma}(t_k)$ using least squares

$$\hat{\theta} = \arg\min \sum_k (\hat{\gamma}_(t_k) - \gamma(t_k \,|\, \theta))^2$$

- Problems:
  - $(Y(s+h) - Y(s))^2$ is a highly skewed rv
  - $\gamma(t_k)$'s are not independent
- Weighted least squares with weights w_k
  - Default in geoR package: $w_k = |N(t_k)|$
  - Cressie 1985: $w_k \propto 1/Var(\hat{\gamma}(t_k)) \approx |N(t_k)|/\gamma(t_k \,|\, \theta)^2$
- Use the variogram $\gamma(\cdot \,|\, \hat{\theta})$ to krige at a new location

Beyond variograms

- Process depends on choice of binning (Thumb rule: at least 30 observations per bin)
- Squared differences are not independent
- No inference, no uncertainty
- Does not lead to a unique stochastic process $Y(s)$
- 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

Weak Stationarity

- $E(Y(s)) = E(Y(s+h)),\ Cov(Y(s), Y(s+h)) = C(h)$
- WS $\Rightarrow$ IS

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var(Y(\mathbf{s}+\mathbf{h})) + Var(Y(\mathbf{s})) - 2Cov(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})]. \end{aligned}$$

- Does IS $\Rightarrow$ WS?

Weak Stationarity

- $E(Y(s)) = E(Y(s+h))$, $Cov(Y(s), Y(s+h)) = C(h)$
- WS $\Rightarrow$ IS

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var(Y(\mathbf{s}+\mathbf{h})) + Var(Y(\mathbf{s})) - 2Cov(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})]. \end{aligned}$$

- Does IS $\Rightarrow$ WS?
- Needs the additional assumption $\lim_{\|\mathbf{u}\|\to\infty} C(\mathbf{u}) = 0$.
- Then we can recover the covariance function C from γ as

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

Gaussian Processes (GPs)

- $\{Y(s) \,|\, s \in D\}$ is a Gaussian Process if all finite dimensional densities $\{Y(s_1), \ldots, Y(s_n)\}$ follow multivariate Gaussian distribution
- A Gaussian Process is completely characterized by a mean function $m(s)$ and a covariance function $C(\cdot, \cdot)$
- Advantage: Likelihood based inference.
  $Y = (Y(s_1), \ldots, Y(s_n))' \sim N(m, C)$ where
  $m = (m(s_1), \ldots, m(s_n))'$ and $C = C(s_i, s_j)$

Valid covariance functions and isotropy

- $C(\cdot,\cdot)$ needs to be valid. For all n and all $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, the resulting covariance matrix $C(s_i, s_j)$ for $(Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n))$ must be positive definite
- So, $C(\cdot,\cdot)$ needs to be a positive definite function
- When $\gamma(\mathbf{h})$ or $C(\mathbf{h})$ depends only on $\|\mathbf{h}\|$, we say that the variogram or the covariance function is *isotropic*, respectively. In that case, we write $\gamma(\|\mathbf{h}\|)$ or $C(\|\mathbf{h}\|)$. Otherwise we say that they are *anisotropic*.
- 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 γ).

Some common isotropic covariograms

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

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 τ^2 is often viewed as a "nonspatial effect variance," and the partial sill (σ^2) is viewed as a "spatial effect variance."
- Note discontinuity at 0 due to the nugget. Intentional! To account for measurement error or micro-sale variability.

The Matèrn Correlation Function

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

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

  K_ν is the modified Bessel function of order ν (computationally tractable)

- ν is a smoothness parameter controlling process smoothness. Remarkable!
- $\nu = 1/2$ gives the exponential covariance function

Another look at kriging

- If $\{Y(s)\}$ is a zero-mean GP then $E(Y(s_0) \mid Y(s_1), \ldots, Y(s_n))$ minimizes $E(Y(s_0) - f(Y(s_1), \ldots, Y(s_n))^2$
- $E(Y(s_0) \mid Y(s_1), \ldots, Y(s_n)) = c_0' C^{-1} Y$ where
  $Y = (Y(s_1), \ldots, Y(s_n))'$,
  $c_0 = (C(||s_1 - s_0||), \ldots, C(||s_n - s_0||))'$ and
  $C = (C(||s_i - s_j))_{1 \leq i,j \leq n}$
- Same as the variogram based kriging estimator
- Full predictive distribution:
  $Y(s_0) \mid Y \sim N(c_0' C^{-1} Y, C(s_0, s_0) - c_0' C^{-1} c_0)$

Modeling: estimation and prediction

$$y(\mathbf{s}) = \beta_0 + x(\mathbf{s})\beta_1 + w(\mathbf{s}) + \epsilon(\mathbf{s})$$

- $w(s)$ modeled as $GP(0, C(\cdot \,|\, \theta))$ (usually without a nugget)
- $\epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2)$ contributes to the nugget
- Under isotropy: $C(s+h, s) = \sigma^2 R(||h||\,; \phi)$
- $w = (w(s_1), \ldots, w(s_n))' \sim N(0, \sigma^2 R(\phi))$ where $R(\phi) = \sigma^2 R(||s_i - s_j||\,; \phi)$
- $y = (y(s_1), \ldots, y(s_n))' \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I)$
- Kriging: $y(s_0) \,|\, Y \sim N(x'\beta + c_0'(\sigma^2 R(\phi) + \tau^2 I)^{-1}(Y - X\beta), \sigma^2 + \tau^2 - c_0'(\sigma^2 R(\phi) + \tau^2 I)^{-1} c_0)$

References

- BCG book chapters 2.1 and 2.4
- Cressie, N. 1985. *Fitting Variogram Models by Weighted Least Squares.* Mathematical Geology, 17: 565–586
- Lichtenstern, A. 2013. *Kriging Methods in Spatial Statistics*, Bachelor's thesis, Munchen University.