A review of traditional stationary geostatistical models

Peter F. Craigmile

http://www.stat.osu.edu/~pfc/

Stat 8750.06: Research Group in Spatial Statistics & Environmental Statistics

Department of Statistics, The Ohio State University

Tuesday 2 September, 2014.

Introduction

- A **geostatistical process** is the stochastic process

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\},$$

  where the spatial domain $D \subset \mathbb{R}^p$.

- A **geostatistical model** is a specification or summary of the probabilistic distribution of a collection of random variables (RVs) $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$.

  (The observed data is a **realization** of these RVs).

- Important to model dependence in spatial data correctly:
  - If we **ignore** the dependencies or get them wrong, then we can be led to incorrect statistical inferences.

Useful references

- Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2014), Hierarchical modeling and analysis for spatial data (2nd Edition), Chapman and Hall/CRC Press, Boca Raton, FL.

- Cressie, N. (1993), Statistics for Spatial Data (Revised edition). Wiley, NY.

- Stein M. L. (1999), Interpolation of spatial data: some theory for kriging, Springer, NY.

- Gelfand, A.E., Diggle, P., Guttorp, P., and Fuentes, M. (2010), Handbook of Spatial Statistics, Chapman and Hall/CRC Press, Boca Raton, FL.

- Waller, L.A. and Gotway, C.A. (2004), Applied Spatial Statistics for Public Health Data, Wiley, NY.

A simplification: stationary models

- **Stationarity** means that some characteristic of the distribution of a spatial process does not depend on the spatial location, only the displacement between the locations.

- If you shift the spatial process, that characteristic of the distribution will not change.

- While in most spatial data are **not stationary**,

  – there are **often** ways to either remove or model the non-stationary parts (the components that depend on the spatial location),

  so that we are only left with a stationary component.

A geostatistical model

- The simplest geostatistical model is **IID noise**, where $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ are independent and identically distributed random variables.
  - There is **no dependence** in this model.
  - We **cannot predict** at other spatial locations with IID noise as there are no dependencies between the random variables at different locations.

The mean function

- The **mean function** of $\{Z(\boldsymbol{s})\}$ is

$$\mu_Z(\boldsymbol{s}) \;=\; E(Z(\boldsymbol{s})), \quad \boldsymbol{s} \in D.$$

Think of $\mu_Z(\boldsymbol{s})$ are being the theoretical mean/expectation at location $\boldsymbol{s}$, taken over the possible values that could have generated $Z(\boldsymbol{s})$.

The covariance function

- The **covariance function** of $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is

$$\begin{aligned} C_Z(\boldsymbol{s}, \boldsymbol{t}) &= \operatorname{cov}(Z(\boldsymbol{s}), Z(\boldsymbol{t})) \\ &= E\{[Z(\boldsymbol{s}) - \mu_Z(\boldsymbol{s})][Z(\boldsymbol{t}) - \mu_Z(\boldsymbol{t})]\}. \end{aligned}$$

- The covariance measures the strength of linear dependence between the two RVs $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$.

- Properties:

  1. $C_Z(\boldsymbol{s}, \boldsymbol{t}) = C_Z(\boldsymbol{t}, \boldsymbol{s})$ for each $\boldsymbol{s}$, $\boldsymbol{t} \in D$.
  2. When $\boldsymbol{s} = \boldsymbol{t}$ we obtain $C_Z(\boldsymbol{s}, \boldsymbol{s}) = \operatorname{var}(Z(\boldsymbol{s})) = \sigma_Z^2(\boldsymbol{s})$, the **variance function** of $\{Z(\boldsymbol{s})\}$.
  3. $C_Z(s, t)$ is a **nonnegative definite** function.

Why nonnegative definite functions?

- Consider the following weighted average of the geostatistical process $\{Z(\boldsymbol{s})\}$ at $n \geq 1$ locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$:

$$Y = \sum_{j=1}^{n} a_j Z(\boldsymbol{s}_j),$$

where $a_1, \ldots, a_n$ are real constants.

- The variance of Y is

$$\begin{aligned} \operatorname{var}(Y) &= \operatorname{cov}(Y, Y) = \operatorname{cov}\left(\sum_{j=1}^{n} a_j Z(\boldsymbol{s}_j), \sum_{k=1}^{n} a_k Z(\boldsymbol{s}_k) \right) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} a_j \, a_k \operatorname{cov}(Z(\boldsymbol{s}_j), Z(\boldsymbol{s}_k)) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} a_j \, a_k \, C_Z(\boldsymbol{s}_j, \boldsymbol{s}_k). \end{aligned}$$

Why nonnegative definite functions?, cont

- A function $f(\cdot, \cdot)$ is **nonnegative definite** if

$$\sum_{j=1}^{n} \sum_{k=1}^{n} a_j \, f(\boldsymbol{s}_j, \boldsymbol{s}_k) \, a_k \; \geq \; 0,$$

  for **all** positive integers n and real-valued constants $a_1, \ldots, a_n$.

- Thus $C_Z(\boldsymbol{s}, \boldsymbol{t})$ must be a nonnegative definite function.

  Why?

The correlation function

- The **correlation function** of $\{Z(\boldsymbol{s})\}$ is

$$\begin{aligned}\rho_Z(\boldsymbol{s},\boldsymbol{t}) &= \mathrm{corr}(Z(\boldsymbol{s}), Z(\boldsymbol{t})) \\ &= \frac{C_Z(\boldsymbol{s},\boldsymbol{t})}{\sqrt{C_Z(\boldsymbol{s},\boldsymbol{s})C_Z(\boldsymbol{t},\boldsymbol{t})}}.\end{aligned}$$

- The correlation measures the strength of linear association between the two RVs $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$.

- Properties:

  1. $-1 \le \rho_Z(\boldsymbol{s},\boldsymbol{t}) \le 1$ for each $\boldsymbol{s}$, $\boldsymbol{t} \in D$
  2. $\rho_Z(\boldsymbol{s},\boldsymbol{t}) = \rho_Z(\boldsymbol{t},\boldsymbol{s})$ for each $\boldsymbol{s}$, $\boldsymbol{t} \in D$
  3. $\rho_Z(\boldsymbol{t},\boldsymbol{t}) = 1$ for each $\boldsymbol{t} \in D$.
  4. $\rho_Z(\boldsymbol{s},\boldsymbol{t})$ is a **nonnegative definite** function.

Strictly stationary processes

- In **strict stationarity** the joint distribution of a set of RVs are unaffected by spatial shifts.

- A geostatistical process, $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, is **strictly stationary** if

$$(Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)) \;=_d\; (Z(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, Z(\boldsymbol{s}_n + \boldsymbol{h}))$$

  for all $n \geq 1$, spatial locations $\{\boldsymbol{s}_j : j = 1, \ldots, n\}$ and a **displacement** or **spatial lag** $\boldsymbol{h}$.

Properties of a strictly stationary $\{Z(\boldsymbol{s})\}$

1. $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is identically distributed

    - Not necessarily independent!

2. $(Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h})) =_d (Z(\mathbf{0}), Z(\boldsymbol{h}))$ for all $\boldsymbol{s}$ and $\boldsymbol{h}$;

3. When μ_Z is finite, $\mu_Z(\boldsymbol{s}) = \mu_Z$ is independent of spatial location.

4. When the variance function exists,

    $C_Z(\boldsymbol{s}, \boldsymbol{t}) = C_Z(\boldsymbol{s}+\boldsymbol{h}, \boldsymbol{t}+\boldsymbol{h})$ for any $\boldsymbol{s}$, $\boldsymbol{t}$ and $\boldsymbol{h}$.

(Weakly) stationary processes

- $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is **(weakly) stationary** if

  1. $E(Z(\boldsymbol{s})) = \mu_Z(\boldsymbol{s}) = \mu_Z$ for some constant μ_Z which does not depend on $\boldsymbol{s}$.

  2. $\text{cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h})) = C_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = C_Z(\boldsymbol{h})$, a finite constant that can depend on $\boldsymbol{h}$ but **not** on $\boldsymbol{s}$.

- The quantity $\boldsymbol{h}$ is called the **spatial lag** or **displacement**.

- Other terms for this type of stationarity include **second-order**, **covariance**, **wide sense**.

Relating weak and strict stationarity:

1. A strictly stationary process $\{Z(\boldsymbol{s})\}$ is also weakly stationary as long as $\sigma_Z^2(\boldsymbol{s})$ is finite for all $\boldsymbol{s}$.

2. **Weak** stationarity does **not** imply **strict** stationarity! (unless we have a Gaussian process – see later).

What suffices to be a valid covariance or correlation function?

Bochner's theorem:

A real-valued function $C_Z(\cdot)$ defined on $\mathbb{R}^p$ is the covariance function of a

stationary process

if and only if

it is ***even*** *and* ***nonnegative definite***, OR

A real-valued function $\rho_Z(\cdot)$ defined on $\mathbb{R}^p$ is the correlation function of a

stationary process

if and only if

it is ***even*** *and* ***nonnegative definite AND*** $\rho_Z(\mathbf{0}) = 1$.

- We can use Bochner's theorem to construct stationary processes, using the literature of what is known about nonnegative definite functions.

Norms

- Let $||\cdot||$ be some **norm**.
  - The most common norm is the **Euclidean norm** in $\mathbb{R}^p$,

$$||\boldsymbol{x}|| = \sqrt{\sum_{j=1}^{p} x_j^2},$$

    where $\boldsymbol{x} = (x_1, \ldots, x_p)^T$.
  - See Banerjee [2005] for a discussion of distances in different coordinate systems.

Isotropic and anisotropic processes

- A stationary process for which $C_Z(\boldsymbol{s}, \boldsymbol{t})$ only depends on the distance between the locations, $||\boldsymbol{s} - \boldsymbol{t}||$, is called an **isotropic process**.
  - With a spatial lag $\boldsymbol{h}$, the covariance function of an isotropic process can be written in terms of its length $||\boldsymbol{h}||$:

$$C_Z(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}) = C_Z(||\boldsymbol{h}||).$$

- If the covariance of a stationary process depends on the **direction and distance** between the locations, then the process is called **anisotropic**.

Gaussian processes

- $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is a **Gaussian process** if the joint distribution of any collection of the RVs has a **multivariate normal (Gaussian)** distribution; the distribution is completely characterized by $\mu_Z(\cdot)$ and $C_Z(\cdot, \cdot)$.

- The joint probability density function of $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))^T$ at a finite set of locations is

$$f_Z(\boldsymbol{z}) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \right),$$

where $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))^T$ and the (j,k) element of the covariance matrix Σ is $C_Z(\boldsymbol{s}_j, \boldsymbol{s}_k)$.

- If the Gaussian process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is (weakly) stationary then the process is also strictly stationary with mean μ_Z and covariance $C_Z(\cdot)$.

The variogram

- Another commonly-used measure of spatial dependence is the **variogram**, which measures the **variance** of the **difference** of a geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ at two spatial locations $\boldsymbol{s}$ and $\boldsymbol{t}$:

$$2\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) = \mathrm{var}(Z(\boldsymbol{s}) - Z(\boldsymbol{t})).$$

- When $E[Z(\boldsymbol{s}) - Z(\boldsymbol{t})] = 0$, we have that

$$2\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) = E([Z(\boldsymbol{s}) - Z(\boldsymbol{t})]^2).$$

- Terminology: $2\gamma_Z(\boldsymbol{s}, \boldsymbol{t})$ is called the **variogram**; $\gamma_Z(\boldsymbol{s}, \boldsymbol{t})$ is called the **semi-variogram**.

The variogram

- Interpretation:
  - When there is little variability in the difference, $Z(\boldsymbol{s}) - Z(\boldsymbol{t})$, then $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$ are more similar (more dependent);
  - When there is greater variability in $Z(\boldsymbol{s}) - Z(\boldsymbol{t})$, then $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$ are less similar (less dependent).

Intrinsic stationarity

- A geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is **intrinsic (stationary)** when $2\gamma_Z(\boldsymbol{s}+\boldsymbol{h}, \boldsymbol{s}) = \text{var}(Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s}))$ only depends on the displacement $\boldsymbol{h}$ for all $\boldsymbol{s}$.

- When the process is intrinsic stationary we can denote the variogram by $2\gamma_Z(\boldsymbol{h})$.

- As with stationary processes we can have intrinsic stationary processes that are isotropic. Such processes are called **homogeneous**, and we can denote the variogram by $\gamma_Z(||\boldsymbol{h}||)$ for some norm $||\cdot||$.

- Weakly stationary implies intrinsic stationary (but not vice versa).

Writing the variogram in terms of covariances

- Using rules for covariances we can relate the variogram to the covariance function:

$$\begin{aligned} 2\gamma_Z(\boldsymbol{s}, \boldsymbol{t}) &= \mathrm{var}(Z(\boldsymbol{s}) - Z(\boldsymbol{t})) \\ &= \mathrm{cov}(Z(\boldsymbol{s}) - Z(\boldsymbol{t}), Z(\boldsymbol{s}) - Z(\boldsymbol{t})) \\ &= C_Z(\boldsymbol{s}, \boldsymbol{s}) + C_Z(\boldsymbol{t}, \boldsymbol{t}) - 2C_Z(\boldsymbol{s}, \boldsymbol{t}). \end{aligned}$$

- When the process is weakly stationary we can simplify to

$$\gamma_Z(\boldsymbol{h}) = C_Z(\boldsymbol{0}) - C_Z(\boldsymbol{h}).$$

- Thus, given the covariance function (when it exists), we can calculate the variogram.

Writing the covariance in terms of the variogram?

- This is possible when a stationary geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is **ergodic**: the mean of a finite area of a geostatistical process tends to the mean of the entire area as the finite area gets larger and larger.

- In that case $C_Z(\boldsymbol{h}) \to 0$, as $||\boldsymbol{h}|| \to \infty$.

- Then

$$\begin{aligned} \lim_{||\boldsymbol{h}|| \to \infty} \gamma_Z(\boldsymbol{h}) &= \lim_{||\boldsymbol{h}|| \to \infty} [C_Z(\mathbf{0}) - C_Z(\boldsymbol{h})] \\ &= C_Z(\mathbf{0}) - 0 \;=\; C_Z(\mathbf{0}). \end{aligned}$$

- With this result we get that

$$C_Z(\boldsymbol{h}) \;=\; \lim_{||\boldsymbol{u}|| \to \infty} \gamma_Z(\boldsymbol{u}) - \gamma_Z(\boldsymbol{h}).$$

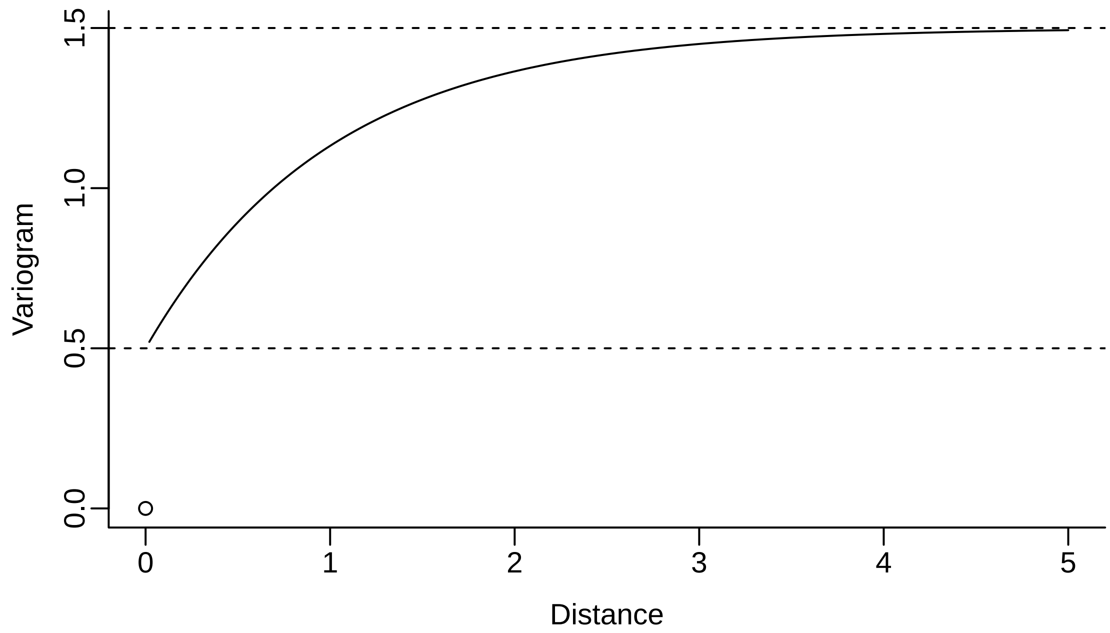
Properties of variograms

- Properties of the variogram for an intrinsic stationary process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$:

  1. $\gamma_Z(\mathbf{0}) = 0$.
  2. $\gamma_Z(\boldsymbol{h}) \geq 0$ for all $\boldsymbol{h}$.
  3. $\gamma_Z(-\boldsymbol{h}) = \gamma_Z(\boldsymbol{h})$ for each $\boldsymbol{h}$ ($\gamma_Z(\cdot)$ is an **even** function).
  4. $\gamma_Z(\boldsymbol{h})$ is **conditional negative definite**:
     - For any $n \geq 1$ consider any spatial locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ and constants $a_1, \ldots, a_n$ such that $\sum_{j=1}^{n} a_j = 0$.

       Then
$$\sum_{j=1}^{n} \sum_{k=1}^{n} a_j \gamma_Z(\boldsymbol{s}_j - \boldsymbol{s}_k) a_k \leq 0.$$

Visualizing and describing variograms



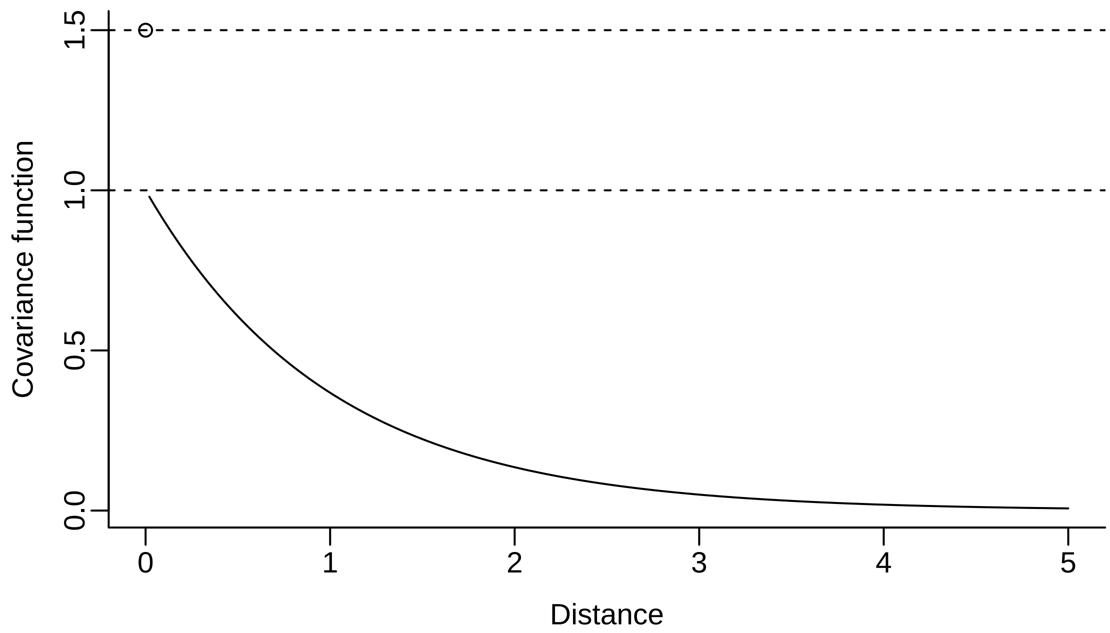
The **sill** is the limiting value of the variogram at $t \to \infty$.

The **range** is the dist. at which the variogram reaches the sill (could be ∞).

The **nugget** is limiting value of the variogram as $t \to 0$, from the right.

The **partial sill** = sill − nugget.

Visualizing and describing covariances



The **sill** is the covariance at zero distance.

The **range** is the distance at which the covariance reaches zero (could be ∞).

The **partial sill** is the limit of the covariance as $t \to 0$, from the right.

The **nugget** $=$ sill $-$ partial sill.

Parametric models for spatial covariances/variograms

- There are many parametric models for the variogram and covariance function that are used for geostatistical modelling.

- Suppose that the geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is stationary and isotropic.

- In these models:

  – We rewrite the distance $||\boldsymbol{h}||$ as just t.

  – The parameter $\tau^2 > 0$ is the **nugget**.

  – The parameter $\sigma^2 > 0$ is the **partial sill**.

  – The **range parameter** ...

The range parameter

- The parameter $\phi > 0$ is a **range parameter** (not the range, but measures how quickly the covariance decays to zero). Fixing the distances:
  - With a smaller value of ϕ the covariance function decays to zero quicker.
  - With a larger value of ϕ the covariance function decays to zero slower.

In the upcoming plots we set $\tau^2 = 1/2$ and $\sigma^2 = 1$. In all examples except the wave $\phi = 1$; for the wave $\phi = 1/4$.

1. The exponential covariance/variogram

- Certainly the most commonly used parametric model.
- The **exponential covariance** function is

$$C_Z(t) = \begin{cases} \sigma^2 \exp(-t/\phi), & t > 0; \\ \tau^2 + \sigma^2, & t = 0, \end{cases}$$

and the **exponential variogram** is

$$\gamma_Z(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-t/\phi)), & t > 0; \\ 0, & t = 0. \end{cases}$$

2. The Gaussian covariance/variogram

- The **Gaussian covariance** function is

$$C_Z(t) \;=\; \begin{cases} \sigma^2 \exp(-(t/\phi)^2), & t > 0; \\ \tau^2 + \sigma^2, & t = 0, \end{cases}$$

and the associated variogram is

$$\gamma_Z(t) \;=\; \begin{cases} \tau^2 + \sigma^2(1 - \exp(-(t/\phi)^2)), & t > 0; \\ 0, & t = 0. \end{cases}$$

3. The power exponential covariance/variogram

- This model is valid for $0 < r \leq 2$.

- The **power exponential covariance function** is

$$C_Z(t) = \begin{cases} \sigma^2 \exp(-|t/\phi|^r), & t > 0; \\ \tau^2 + \sigma^2, & t = 0, \end{cases}$$

and the associated variogram is

$$\gamma_Z(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-|t/\phi|^r)), & t > 0; \\ 0, & t = 0. \end{cases}$$

- With $r = 1$ we get the exponential, and with $r = 2$ we get the Gaussian.

4. The Matèrn covariance function

- We say that the mean zero isotropic geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ has a **Matèrn covariance function** with smoothness parameter $\nu > 0$ if,

$$C_Z(t) \;=\; \begin{cases} \sigma^2 + \tau^2 & t = 0, \\ \sigma^2 \dfrac{(t/\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(t/\phi), & t > 0, \end{cases}$$

  at distance t.

- In the above expression for the covariance, $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is a modified Bessel function [Abramowitz and Stegun, 1965, Chapter 9].

- As usual, ϕ is the range parameter, σ^2 is the partial sill, and τ^2 is the nugget.

4. The Matèrn covariance function, cont.

- Special cases of ν:

  1. With $\nu = 1/2$ we get the exponential covariance function.
  2. As $\nu \to \infty$, we approach the Gaussian covariance function, which is infinitely differentiable.
  3. $\nu = 3/2$ is another common choice.

Smoothness

- What is wrong about having **too much** smoothness?
  - As ν increases we can better predict a geostatistical process at any point in the domain of interest, based on knowledge from smaller and smaller regions.
- Not a reasonable assumption in many applications.

  (a similar story can be said for having **too little** smoothness).
- The problem with introducing a smoothness parameter: it can be hard to estimate based on data. Also, the estimator of the smoothness parameter is strongly correlated with other parameters in the model for the covariance.

Nonisotropic and nonstationary covariance modeling

- There is a large interest in modeling nonisotropic and nonstationary geostatistical processes.

- A popular method to introduce **nonisotropy**:

  – Start with an isotropic process, and scale and rotate the coordinate axes to introduce anisotropy.

Nonisotropic and nonstationary covariance modeling, cont.

- Some popular methods of introducing **nonstationarity** (there are others!)

1. Start with a stationary process, and then transform the distances in some smooth way (we **deform the space**) to construct a nonstationary process [Sampson and Guttorp, 1992]
2. Start with a white noise process, and average the process using local weights that vary spatially – this is called the **process convolution approach** [Higdon et al., 1999].
3. Write down covariance functions that **include covariates** [e.g., Schmidt et al., 2011].

References

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* 1965. URL http://people.maths.ox.ac.uk/~macdonald/aands/index.html.

S. Banerjee. On geodetic distance computations in spatial modeling. *Biometrics*, 61:617–625, 2005. URL http://www.jstor.org/stable/3695985.

D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 761–768. Oxford University Press, Oxford, UK, 1999.

P. D. Sampson and P. Guttorp. Nonparametric-estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119, 1992. URL www.jstor.org/stable/2290458.

A. M. Schmidt, P. Guttorp, and A. O'Hagan. Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 5:2425–2447, 2011. URL http://onlinelibrary.wiley.com/doi/10.1002/env.1101/abstract.