Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

Lecture notes part 2: Point referenced data modeling / Geostatistics

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce point referenced data modeling (geostatistics) with particular focus on concepts spatial variables, random fields, semi-variogram, kriging, change of support, multivariate geostatistics, for Bayesian and classical inference.

Reading list & references:

[1] Cressie, N. (2015; Part I). Statistics for spatial data. John Wiley & Sons.
[2] Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media. (on Geostatistics)
[3] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley \& Sons. (on Spatial analysis)
[4] Gaetan, C., & Guyon, X. (2010; Ch 2 & 5.1). Spatial statistics and modeling (Vol. 90). New York: Springer.

Part 1. Basic stochastic models & related concepts for model building

Note 1. We discuss basic stochastic models and concepts for modeling point referenced data in the Geostatistics framework.

1. Random fields (or Stochastic processes)

Definition 2. A random field (or stochastic process, or random function) $Z = (Z(s); s \in \mathcal{S})$ taking values in $\mathcal{Z} \subseteq \mathbb{R}^q$, $q \geq 1$ is a family of random variables $\{Z(s) := Z(s;\omega); s \in \mathcal{S}, \omega \in \Omega\}$ defined on the same probability space $(\Omega, \mathfrak{F}, \mathrm{pr})$ and taking values in $\mathcal{Z}$. The label $s \in \mathcal{S}$ is called site, the set $\mathcal{S} \subseteq \mathbb{R}^d$ is called the (spatial) set of sites at which the random field is defined, and $\mathcal{Z}$ is called the state space of the field.

Note 3. Given a set of sites $\{s_1, ..., s_n\}$, with $s_i \in S$ and $n \in \mathbb{N}$, the random vector $(Z(s_1), ..., Z(s_n))^\top$ has a well-defined probability distribution that is completely determined by its joint CDF

$$\mathrm{F}_{s_1,...,s_n}(z_1, ..., z_n) := \mathrm{pr}(Z(s_1) \leq z_1, ..., Z(s_n) \leq z_n)$$

The family of all finite-dimensional distributions (or fidi's) of Z is called the spatial distribution of the process .

Note 4. According to Kolmogorov Theorem 5, to define a random field model, one must specify the joint distribution of $(Z(s_1), ..., Z(s_n))^\top$ for all of n and all $\{s_i \in S; i = 1, ..., n\}$ in a consistent way.

Proposition 5. *(Kolmogorov consistency theorem) Let $pr_{s_1,...,s_n}$ be a probability on $\mathbb{R}^n$ with join CDF $F_{s_1,...,s_n}$ for every finite collection of points $s_1, ..., s_n$. If $F_{s_1,...,s_n}$ is symmetric w.r.t. any permutation $\mathfrak{p}$*

$$F_{\mathfrak{p}(s_1),...,\mathfrak{p}(s_n)}\left(z_{\mathfrak{p}(1)}, ..., z_{\mathfrak{p}(n)}\right) = F_{s_1,...,s_n}(z_1, ..., z_n)$$

for all $n \in \mathbb{N}$, $\{s_i \in S\}$, and $\{z \in \mathbb{R}\}$, and all if all permulations $\mathfrak{p}$ are consistent in the sense

$$\lim_{z_n \to \infty} F_{s_1,...,s_n}(z_1, ..., z_n) = F_{s_1,...,s_{n-1}}(z_1, ..., z_{n-1})$$

or all $n \in \mathbb{N}$, $\{s_i \in S\}$, and $\{z_i \in \mathbb{R}\}$, then there exists a random field Z whose fidi's coincide with those in F.

Example 6. Let $n \in \mathbb{N}$, let $\{X_i : T \to \mathbb{R};\ i = 1, ..., n\}$ be a set of constant functions, and let $\{Z_i \sim \mathrm{N}(0, 1)\}_{i=1}^n$ be a set of independent random variables. Then

$$\tilde{Z}(s) = \sum_{i=1}^n Z_i X_i(s), \quad s \in S \tag{1.1}$$

is a well defined random field as it satisfies Theorem 5.

1.1. **Mean and covariance functions.**

Definition 7. The mean function $\mu(\cdot)$ and covariance function $c(\cdot, \cdot)$ of a random field $(Z(s); s \in \mathcal{S})$ are defined as

$$\mu(s) = \mathrm{E}(Z(s)), \qquad \forall s \in S \tag{1.2}$$

$$c(s, s') = \mathrm{Cov}(Z(s), Z(s')) = \mathrm{E}\left((Z(s) - \mu(s))(Z(s') - \mu(s'))^\top\right), \quad \forall s, s' \in S \tag{1.3}$$

Example 8. For (1.1), the mean function is $\mu(s) = \mathrm{E}\left(\tilde{Z}_s\right) = 0$ and covariance function is

$$\begin{aligned} c(s, s') &= \mathrm{Cov}(Z(s), Z(s')) = \mathrm{Cov}\left(\sum_{i=1}^n Z_i X_i(s), \sum_{j=1}^n Z_j X_j(s')\right) \\ &= \sum_{i=1}^n X_i(s) \sum_{j=1}^n X_j(s') \cancel{\mathrm{Cov}(Z_i, Z_j)}^{1(i=j)} = \sum_{i=1}^n X_i(s) X_i(s') \end{aligned}$$

1.1.1. *Construction of covariance functions.*

Note 9. What follows provides the means for checking and constructing covariance functions.

Proposition 10. *The function* $c : S \times S \to \mathbb{R}$, $S \subseteq \mathbb{R}^n$ is a covariance function iff $c(\cdot,\cdot)$ is semi-positive definite; i.e.

$$\forall n \in \mathbb{N}\,,\ \forall a \in \mathbb{R}^n \text{ and } \forall\,(s_1,...,s_n) \in \mathbb{R}^n : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j c(s_i,s_j) \geq 0$$

or in other words, the Gram matrix $(c(s_i,s_j))_{i,j=1}^{n}$ is non-negative definite for any $\{s_i\}_{i=1}^{n}$, $n \in \mathbb{N}$.

Example 11. $c(s,s') = 1(\{s = s'\})$ is a proper covariance function because

$$\sum_i \sum_j a_i a_j c(s_i,s_j) = \sum_i a_i^2 \geq 0,\ \ \forall a$$

Note 12. One way to construct a c.f c is to set $c(s,s') = \psi(s)^\top \psi(s')$, for a given vector of basis functions $\psi(\cdot) = (\psi_1(\cdot),...,\psi_n(\cdot))$.

Proof. From Proposition 10, as

$$\sum_i \sum_j a_i a_j c(s_i,s_j) = (\psi a)^\top (\psi a) \geq 0,\ \ \forall a \in \mathbb{R}^n$$

□

2. Second order random fields (or Second order processes)

Note 13. We introduce a particular class of random fields whose mean and covariance functions exist and which can be used for spatial data modeling.

Definition 14. Second order random field (or second order process) $(Z(s)\,; s \in \mathcal{S})$ is called the random field where $\mathrm{E}\left((Z(s))^2\right) < \infty$ for all $s \in S$.

Example 15. In second order random field $(Z(s)\,; s \in \mathcal{S})$ the associated mean function $\mu(\cdot)$ and covariance function $c(\cdot,\cdot)$ exist, because $c(s,t) = \mathrm{E}(Z(s)Z(t)) - \mathrm{E}(Z(s))\mathrm{E}(Z(t))$ for $s,t \in \mathcal{S}$.

3. Gaussian random field (or Gaussian process)

Note 16. Gaussian random field is a particular class of second order random field which is widely used in spatial data modeling due to its computational tractability.

Definition 17. $(Z(s)\,; s \in \mathcal{S})$ is a Gaussian random field (GRF) or Gaussian process (GP) on $\mathcal{S}$ if for any $n \in \mathbb{N}$ and for any finite set $\{s_1,...,s_n; s_i \in \mathcal{S}\}$, the random vector $(Z(s_1),...,Z(s_n))^\top$ follows a multivariate normal distribution.

Also Example of Proposition

Proposition 18. *A GP* $(Z(s)\,; s \in \mathcal{S})$ *is fully characterized by its mean function* $\mu : S \to \mathbb{R}$ *with* $\mu(s) = E(Z(s))$*, and its covariance function with* $c(s,s') = Cov(Z(s), Z(s'))$.

Notation 19. Hence, we denote the GP as $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$.

Note 20. When using GP for spatial modeling we just need to specify its functional parameters i.e. the mean and covariance functions.

Note 21. Popular forms of mean functions are polynomial expansions, such as $\mu(s) = \sum_{j=0}^{p-1} \beta_j s^j$ for tunable unknown parameter β. A popular form of covariance functions (c.f.), for tunable unknown parameters $\phi > 0$, and $\sigma^2 > 0$, are

(1) Exponential c.f. $c(s,s') = \sigma^2 \exp(-\phi \|s - s'\|_1)$
(2) Gaussian c.f. $c(s,s') = \sigma^2 \exp\left(-\phi \|s - s'\|_2^2\right)$
(3) Nugget c.f. $c(s,s') = \sigma^2 1(s = s')$

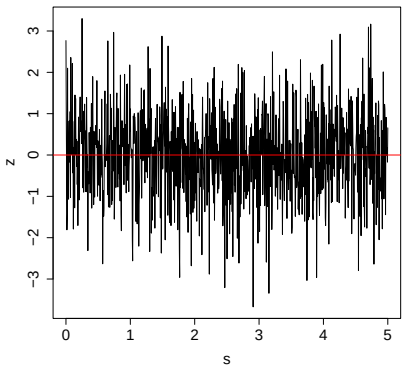
Example 22. Recall your linear regression lessons where you specified the sampling distribution to be $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta, \sigma^2\right)$, $\forall x \in \mathbb{R}^d$. Well that can be considered as a GP $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(x) = x^\top\beta$ and $c(x,x') = \sigma^2 1(x = x')$ in (3).

Example 23. Figures 3.1 & 3.2 presents realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(s) = 0$ and differently parameterized covariance functions in 1D and 2D. In 1D the code to simulate the GP is given in Algorithm 1. Note that we actually discretize it and simulate it from the fidi.

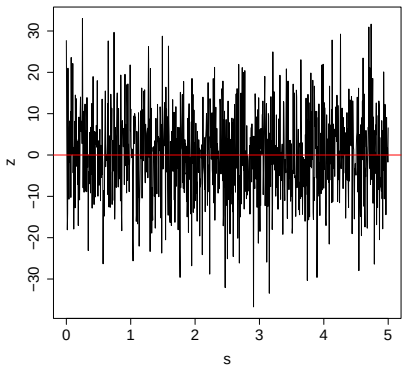
Algorithm 1 R script for simulating from a GP $(Z(s); s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\phi \|s-t\|_2^2\right)$

```
# set the GP parameterized mean and covariance function
mu_fun <- function(s) { return (0) }
cov_fun_gauss <- function(s,t,sig2,phi) { return (
sig2*exp(-phi*norm(c(s-t),type="2")**2) ) }
# discretize the problem in n = 100 spatial points
n <- 100
s_vec <- seq(from = 0, to = 5, length = n)
mu_vec <- matrix(nrow = n, ncol = 1)
Cov_mat <- matrix(nrow = n, ncol = n)
# compute the associated mean vector and covariance matrix of the n=100 dimensional
Normal r.v.
sig2_val <- 1.0 ;
phi_val <- 5
for (i in 1:n) {
    mu_vec[i] <- mu_fun(s_vec[i])
    for (j in 1:n) {
        Cov_mat[i,j] <- cov_fun_gauss(s_vec[i],s_vec[j],sig2_val,phi_val)
    }
}
# simulate from the associated distribution
z_vec <- mu_vec + t(chol(Cov_mat))%*%rnorm(n, mean=0, sd=1)
# plot the path (R produces a line plot)
plot(s_vec, z_vec, type="l")
abline(h=0,col="red")
```

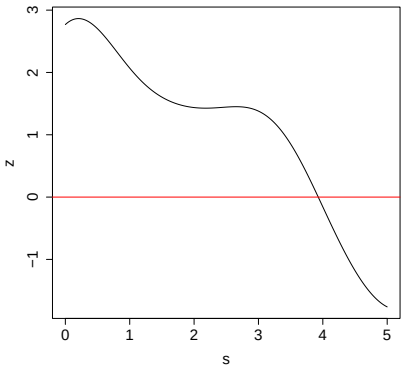
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency of the ups and downs is controlled by β (Figures 3.1d & 3.1e ; Figures 3.2d & 3.2e). Realizations with different c.f. have different behavior (Figures 3.1a, 3.1d & 3.1e ; Figures 3.2a, 3.2d & 3.2e)



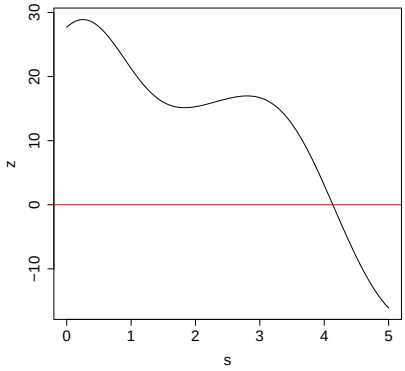
(A) Nugget c.f
($\sigma^2 = 1$)



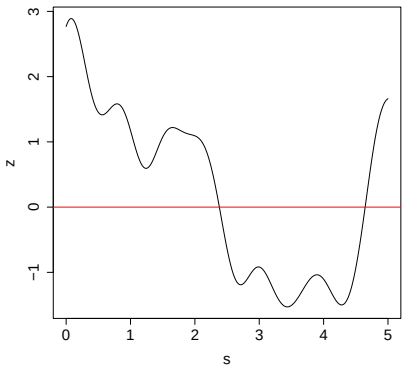
(B) Nugget c.f
($\sigma^2 = 100$)



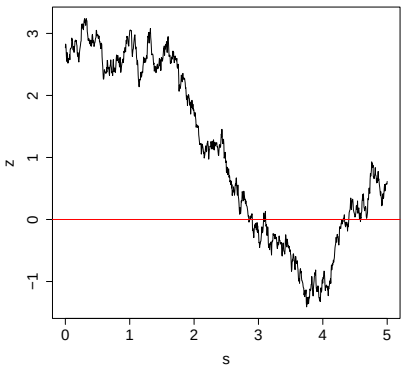
(C) Gauss c.f
($\sigma^2 = 1$,$\phi = 0.5$)



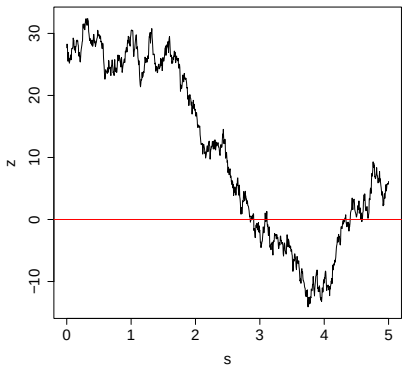
(D) Gauss c.f
($\sigma^2 = 100$,$\phi = 0.5$)



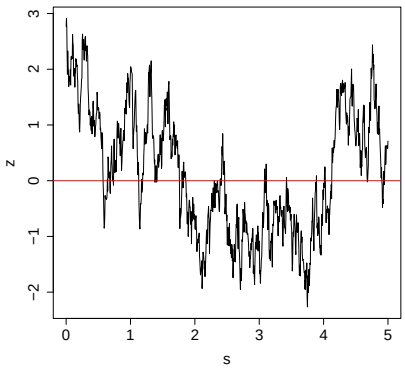
(E) Gauss c.f
($\sigma^2 = 1$,$\phi = 5$)



(F) Exp c.f
($\sigma^2 = 1$,$\phi = 0.5$)

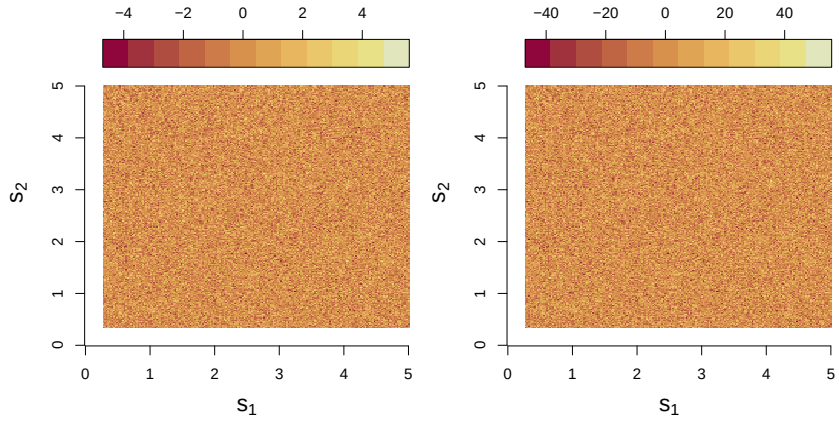


(G) Exp c.f
($\sigma^2 = 100$,$\phi = 0.5$)



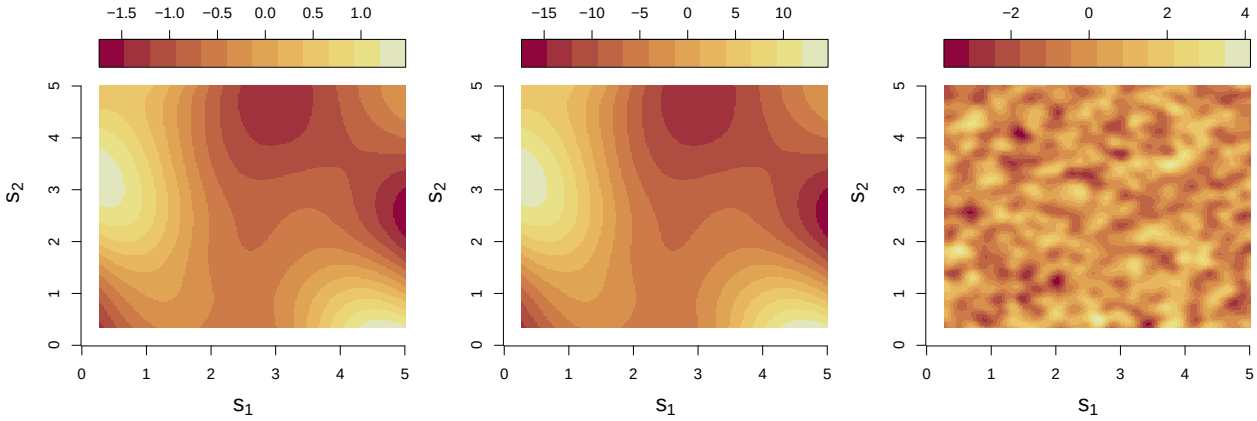
(H) Exp c.f
($\sigma^2 = 1$,$\phi = 5$)

FIGURE 3.1. Realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ when $s \in [0,5]$ (using same seed)



(A) Nugget c.f
($\sigma^2 = 1$)

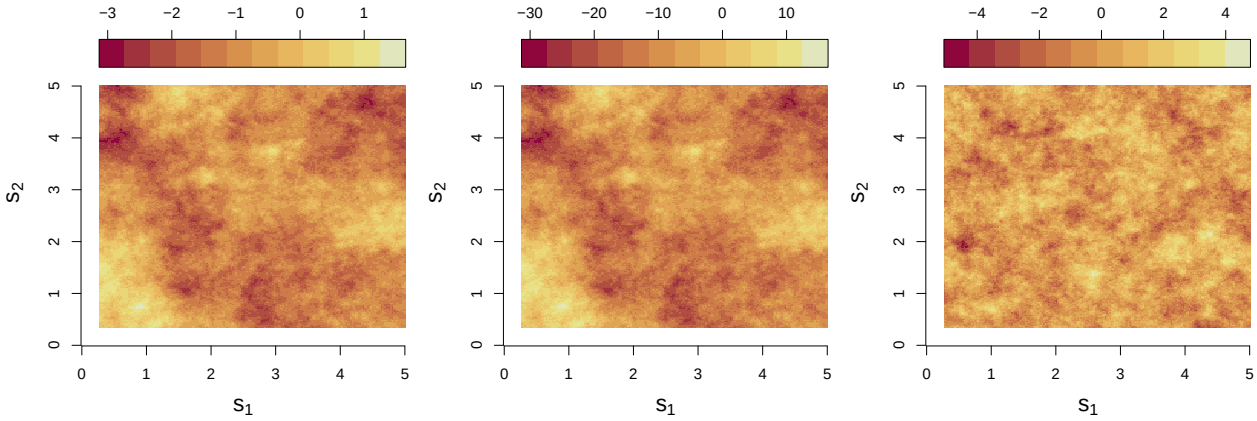
(B) Nugget c.f
($\sigma^2 = 100$)



(C) Gauss c.f
($\sigma^2 = 1$,$\phi = 0.5$)

(D) Gauss c.f
($\sigma^2 = 100$,$\phi = 0.5$)

(E) Gauss c.f
($\sigma^2 = 1$,$\phi = 5$)



(F) Exp c.f
($\sigma^2 = 1$,$\phi = 0.5$)

(G) Exp c.f
($\sigma^2 = 100$,$\phi = 0.5$)

(H) Exp c.f
($\sigma^2 = 1$,$\phi = 5$)

FIGURE 3.2. Realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ when $s \in [0,5]^2$ (using same seed)

4. Strong stationarity

Note 24. We introduce a specific behavior of random field to build our models.

Note 25. Assume $\mathcal{S} = \mathbb{R}^d$ for simplicity. [1]

Definition 26. A random field $(Z(s); s \in \mathcal{S})$ is strongly stationary on $\mathcal{S}$ if for all finite sets consisting of $s_1, ..., s_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, for all $k_1, ..., k_n \in \mathbb{R}$, and for all $h \in \mathbb{R}^d$

$$\text{pr}(Z(s_1 + h) \leq k_1, ..., Z(s_n + h) \leq k_n) = \text{pr}(Z(s_1) \leq k_1, ..., Z(s_n) \leq k_n)$$

Note 27. Yuh... strong stationary may represent a behavior being too "restrictive" to be used for spatial data modeling as it is able to represent only limiting number of spatial dependencies.

5. Weak stationarity (or second order stationarity)

Note 28. We introduce another weaker random field behavior which represents a larger class of spatial dependencies.

Note 29. Instead of working with the "restrictive" strong stationarity, we could just properly specify the behavior of the first two moments only; notice that Definition 26 implies that, given $\text{E}(Z_s^2) < \infty$, it is $\text{E}(Z(s)) = \text{E}(Z(s+h)) = \text{contst...}$ and $\text{Cov}(Z(s), Z(s')) = \text{Cov}(Z(s+h), Z(s'+h)) \overset{h=-s'}{=} \text{Cov}(Z(s-s'), Z(0)) = \text{funct of lag...}$

Definition 30. A random field $(Z(s); s \in \mathcal{S})$ is weakly stationary (or second order stationary) if it has constant mean and translation invariant covariance; i.e. for all $s, s' \in \mathbb{R}^d$,

(1) $\text{E}\left((Z(s))^2\right) < \infty$ (finite)
(2) $\text{E}(Z(s)) = \mu$ (constant)
(3) $\text{Cov}(Z(s), Z(s')) = c(s' - s)$ for some even function $c : \mathbb{R}^d \to \mathbb{R}$ (lag dependency)

Definition 31. Weakly (or second order) stationary covariance function is called the c.f. of a weakly stationary random field.

6. Covariogram

Note 32. We introduce the covariogram function able to express many aspects of the behavior of a weakly stationary random field and hence be used as statistical descriptive tool.

Definition 33. The covariogram function of a weakly stationary random field $\left(Z(s); s \in \mathbb{R}^d\right)$ is defined by $c : \mathbb{R}^d \to \mathbb{R}$ with

$$c(h) = \text{Cov}(Z(s), Z(s+h)), \ \forall s \in \mathbb{R}^d.$$

[1] Otherwise, we should set $s, s' \in \mathcal{S}$, $h \in \mathcal{H}$, such as $\mathcal{H} = \left\{h \in \mathbb{R}^d : s + h \in \mathcal{S}\right\}$.

Example 34. For the Gaussian c.f. $c(s,t) = \sigma^2 \exp\left(-\phi \|s-t\|_2^2\right)$ in (Ex. 20(2)), we may denote just

$$c(h) = c(s, s+h) = \sigma^2 \exp\left(-\phi \|h\|_2^2\right) \tag{6.1}$$

Observe that, in Figures 3.1 &3.2, the smaller the ϕ, the smoother the realization (aka slower changes). One way to justify this observation is to think that smaller values of ϕ essentially bring the points closer by re-scaling spatial lags h in the c.f.

Proposition 35. *If $c : \mathbb{R}^d \to \mathbb{R}$ is the covariogram of a weakly stationary random field $\left(Z(s); s \in \mathbb{R}^d\right)$ then:*

(1) $c(h) = c(-h)$ for all $h \in \mathbb{R}^d$

(2) $|c(h)| \leq c(0) = Var(Z(s))$ for all $h \in \mathbb{R}^d$

(3) $c(\cdot)$ is semi-positive definite; i.e. for all $n \in \mathbb{N}$, $a \in \mathbb{R}^n$, and $\{s_1, ..., s_n\} \subseteq S$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j c(s_i - s_j) \geq 0$$

Note 36. Given there is some knowledge of the characteristic functions of a suitable distribution, the following spectral representation theorem helps in the specification of a suitable covariogram.

Theorem 37. *(Bochner's theorem) Let $c : \mathbb{R}^d \to \mathbb{R}$ be a continuous even real-valued function. Then $c(\cdot)$ is positive semi-definete (hence a covariogram of a stationary random field) if and only if it can be represented as*

$$c(h) = \int_{\mathbb{R}^d} \exp\left(i\omega^\top h\right) dF(\omega)$$

where F is a *symmetric positive finite measure on $\mathbb{R}^d$ called spectral measure.*

Note 38. In our course, we focus on cases where F has a density $f(\cdot)$ i.e. $\mathrm{d}F(\omega) = f(\omega)\,\mathrm{d}\omega$. $f(\cdot)$ is called spectral density of $c(\cdot)$, it is

$$c(h) = \int_{\mathbb{R}^d} \exp\left(i\omega^\top h\right) f(\omega)\,\mathrm{d}\omega,$$

and it dies as $\lim_{|h|\to\infty} c(h) = 0$

Theorem 39. *If $c(\cdot)$ is integrable, the spectral density $f(\cdot)$ can be computed by inverse Fast Fourier transformation*

$$f(\omega) = \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp\left(-i\omega^\top h\right) c(h)\, dh$$

Example 40. Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\phi \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Then, by using Theorem 37, the spectral density is

$$\begin{aligned} f(\omega) &= \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp\left(-i\omega^\top h\right) \sigma^2 \exp\left(-\phi \|h\|_2^2\right) \mathrm{d}h \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^d \int_{\mathbb{R}} \exp\left(-i\omega_j h_j - \phi h_j^2\right) \mathrm{d}h \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^d \int_{\mathbb{R}} \exp\left(-\phi \left(h_j - \left(-i\omega / (2\phi)\right)\right)^2\right) \mathrm{d}h_j \\ &= \sigma^2 \left(\frac{1}{4\pi\phi}\right)^{d/2} \exp\left(-\|\omega\|_2^2 / (4\phi)\right) \end{aligned}$$

i.e. it has a Gaussian form.

Definition 41. Let $\left(Z(s); s \in \mathbb{R}^d\right)$ be a weakly stationary random field with covariogram function $c : \mathbb{R}^d \to \mathbb{R}$ and $c(h) = \mathrm{Cov}\left(Z(s), Z(s+h)\right)$. The correlogram function $\rho : \mathbb{R}^d \to [-1, 1]$ is defined as

$$\rho(h) = \frac{c(h)}{c(0)}.$$