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. Stochastic processes (or Random fields)

**Definition 2.** A stochastic process (or random field) $Z = (Z_s; s \in \mathcal{S})$ 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 process is defined, and $\mathcal{Z}$ is called the state space of the process.

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

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

Finite dimensional distributions (or fidi's) of $Z$ is called the ensemble of all such joint CDF's with $n \in \mathbb{N}$ and $\{s_i \in S\}$.

*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 $\mathfrak{p}$*

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

*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*

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

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

**Example 6.** Let $n \in \mathbb{N}$, let $\{f_i : T \to \mathbb{R};\ i = 1, ..., n\}$ 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

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

is a well defined stochastic process 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 $Z = (Z_s)_{s \in S}$ are defined as

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

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

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

$$\begin{aligned} c(s, s') = \mathrm{Cov}(Z_s, Z_{s'}) &= \mathrm{Cov}\left(\sum_{i=1}^n Z_i f_i(s), \sum_{j=1}^n Z_j f_j(s')\right) \\ &= \sum_{i=1}^n f_i(s) \sum_{j=1}^n f_j(s') \cancelto{1(i=j)}{\mathrm{Cov}(Z_i, Z_j)} = \sum_{i=1}^n f_i(s) f_i(s') \end{aligned}$$

### 1.1.1. *Construction of covariance functions.*

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

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

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

*Note* 12. Proposition 13 uses the experience from basis functions, while Theorem **??** uses experience from characteristic functions to be incorporated int he process for modeling reasons.

*Remark* 13. 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 processes (or Second order random fields)

*Note* 14. We introduce a particular class of stochastic processes whose mean and covariance functions exist and which can be used of spatial data modeling.

**Definition 15.** Second order process (or second order random field) $Z = (Z_s; s \in \mathcal{S})$ is called the stochastic process where $\mathrm{E}(Z_s^2) < \infty$ for all $s \in S$.

**Example 16.** In second order processes $(Z_s)_{s \in \mathcal{S}}$, then 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 process

*Note* 17. We introduce a particular class of second order stochastic processes with specific joint distribution which can be used of spatial data modeling.

Also Example of Proposition

**Definition 18.** $Z = (Z_s; s \in S)$ indexed by $S \subseteq \mathbb{R}^d$ is a Gaussian process (GP) or random field (GRF) 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.

**Proposition 19.** *A GP $Z = (Z_s; s \in S)$ 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* 20. Hence, we denote the GP as $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$.

*Note* 21. When using GP for spatial modeling we may need to specify its functional parameters i.e. the mean and covariance functions.

*Note* 22. An popular form of mean functions are polynomial expansions, such as $\mu(s) = \sum_{j=0}^{p-1} \beta_j s^j$ for some tunable unknown parameter $\beta$. An popular form of covariance functions (c.f.), for some tunable unknown parameter $\beta$,$\sigma^2$, are

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

**Example 23.** 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).

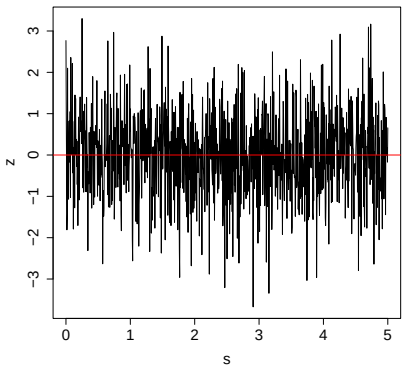
**Example 24.** Figures 3.1 & 3.2 presents realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(s) = 0$ and differently parameterized covariance functions in 1D and 2D. In 1D the code to simulate the GP is given in Algorithm 1. Note that we actually discretize it and 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 $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

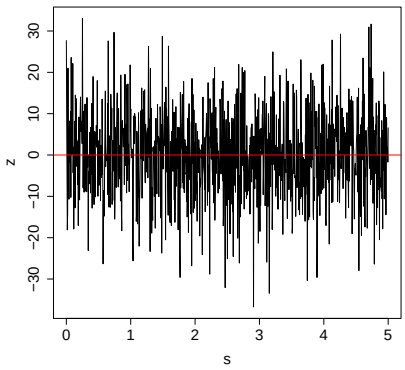
```
# set the GP parameterized mean and covariance function
mu_fun <- function(s) { return (0) }
cov_fun_gauss <- function(s,t,sig2,beta) { return (
sig2*exp(-beta*norm(c(s-t),type="2")**2) ) }
# discretize the problem in n = 100 spatial points
n <- 100
s_vec <- seq(from = 0, to = 5, length = n)
mu_vec <- 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 ;
beta_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,beta_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 $\sigma^2$ (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and

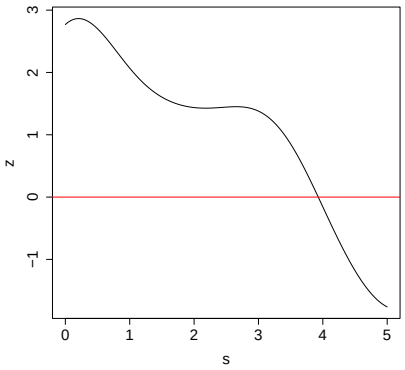
downs are random and controlled by $\sigma^2$ (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency of the ups and downs is controlled by $\beta$ (Figures 3.1d & 3.1e ; Figures 3.2d & 3.2e). Realizations with different c.f. have different behavior (Figures 3.1a, 3.1d & 3.1e ; Figures 3.2a, 3.2d & 3.2e)



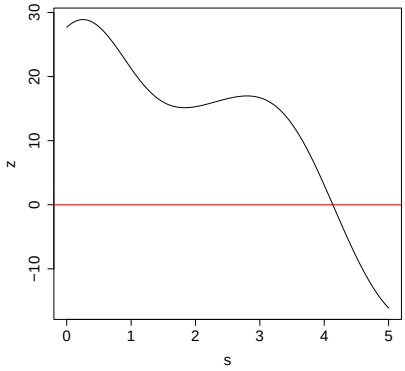

(A) Nugget c.f
($\sigma^2 = 1$)



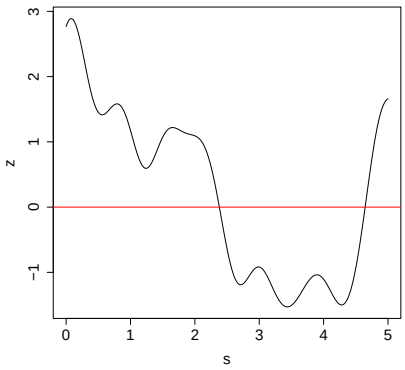

(B) Nugget c.f
($\sigma^2 = 100$)



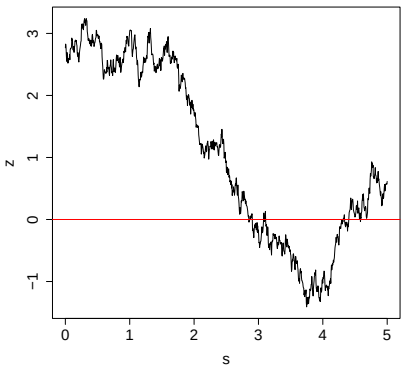

(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



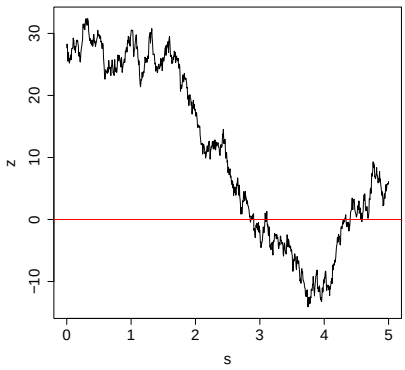

(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



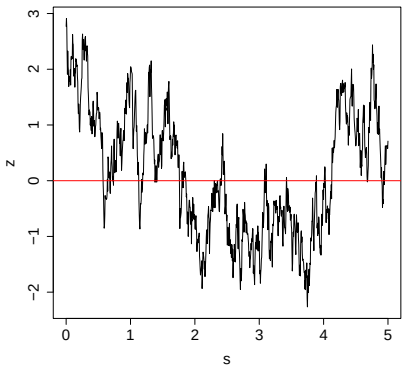

(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)




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

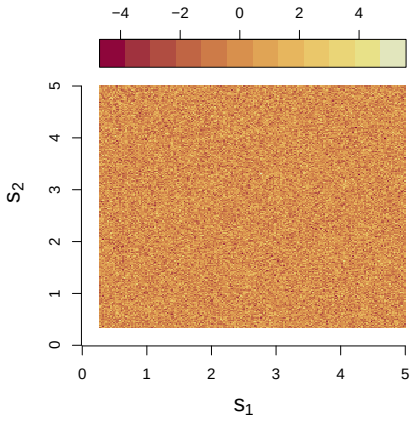



(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)

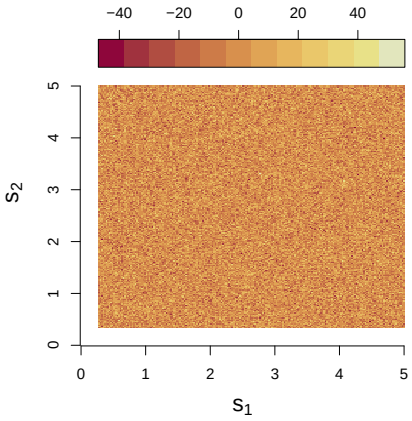



(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

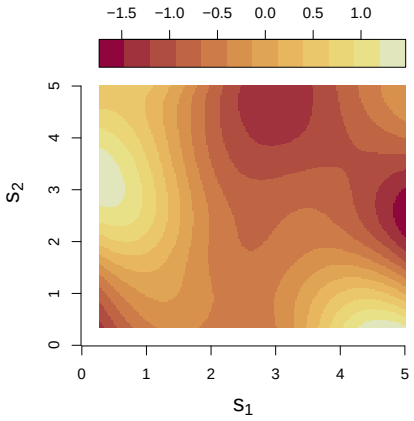
FIGURE 3.1. Realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ when $s \in [0,5]$ (using same seed)



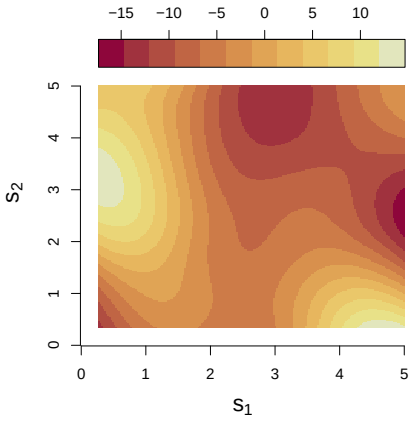

(A) Nugget c.f
($\sigma^2 = 1$)



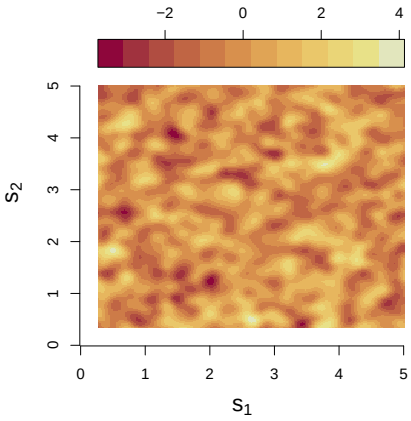

(B) Nugget c.f
($\sigma^2 = 100$)



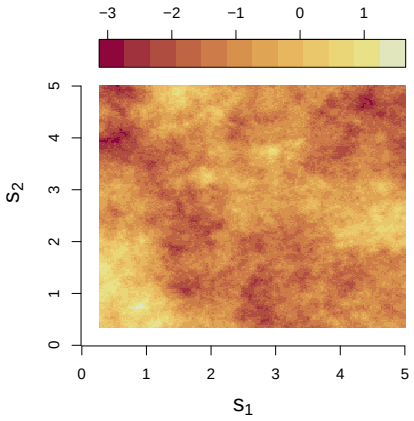

(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



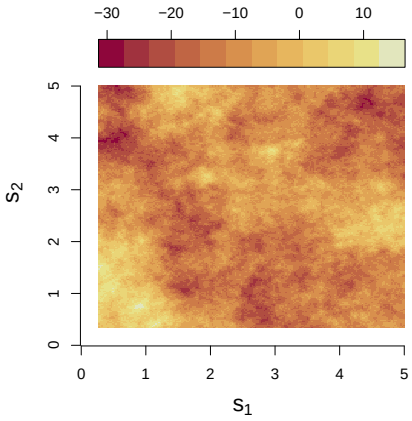

(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



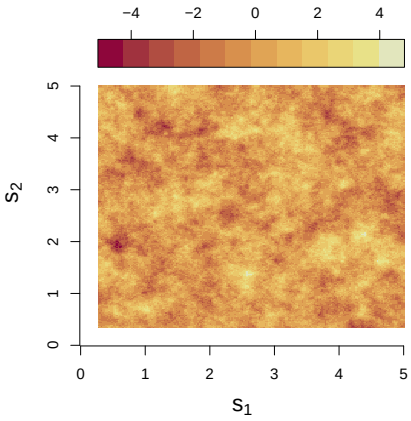

(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)




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




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




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

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* 25. We introduce a specific behavior of stochastic process.

*Note* 26. Assume $\mathcal{S} = \mathbb{R}^d$ for simplicity. [1]

**Definition 27.** A random field $Z = (Z_s)_{s \in \mathbb{R}^d}$ is strongly stationary 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 $h \in \mathbb{R}^d$

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

## 5. Weak stationarity (or second order stationarity)

*Note* 28. We introduce another specific behavior of stochastic process.

*Note* 29. Yuh... strong stationary may represent a very "restricting" behavior to be used for spatial data modeling; it may be able to represent limiting number of spatial dependences. Instead, we could just properly specify the behavior of the first two moments only; notice that Definition 27 implies that, given $\text{E}\left(Z_s^2\right) < \infty$, it is $\text{E}\left(Z_s\right) = \text{E}\left(Z_{s+h}\right) = \text{contst...}$ and $\text{Cov}\left(Z_s, Z_{s'}\right) = \text{Cov}\left(Z_{s+h}, Z_{s'+h}\right) \overset{h=-s'}{=} \text{Cov}\left(Z_{s-s'}, Z_0\right) = \text{funct of lag...}$

**Definition 30.** A random field $Z = (Z_s)_{s \in \mathbb{R}^d}$ is weakly stationary (or second order stationary) if, for all $s, s' \in \mathbb{R}^d$,

(1) $\text{E}\left(Z_s^2\right) < \infty$ (finite)
(2) $\text{E}\left(Z_s\right) = m$ (constant)
(3) $\text{Cov}\left(Z_s, Z_{s'}\right) = c\left(s' - s\right)$ 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 stochastic process.

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