Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 3: Point referenced data modeling / Geostatistics

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

Aim. To introduce Point referenced data modeling / Geostatistics: regional variables, random field, variogram,

Reading list & references:

[1] Cressie, N. (2015; Part I). Statistics for spatial data. John Wiley & Sons.
[2] Gaetan, C., & Guyon, X. (2010; Ch 2 & 5.1). Spatial statistics and modeling (Vol. 90). New York: Springer.

Specialized reading.

[3] Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media. (on Geostatistics)
[4] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons. (on Spatial analysis)

Part 1. Intro to building stochastic models & concepts

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 Thm 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 Thm 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

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

1.1.1. *Construction of covariance functions.* (The following provides the means for checking and constructing covariance functions.)

Proposition 9. *The function $c : S \times S \to \mathbb{R}$, $S \subseteq \mathbb{R}^d$ is the 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 10. $c(s,s') = 1(s = s')$ 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 11. Prop 12 uses the experience from basis functions, while Theorem 30 uses experience from characteristic functions to be incorporated int he process for modeling reasons.

Remark 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 Prop 9, 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 Random fields)

Definition 13. Second order process (or random field) $Z = (Z_s; s \in \mathcal{S})$ is called the stochastic process where $\mathrm{E}(Z_s^2) < \infty$ for all $s \in S$. Then the associated mean function $\mu(\cdot)$ and covariance function $c(\cdot,\cdot)$ exist.

3. Gaussian process

Definition 14. $Z = (Z_s; s \in S)$ 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 $(Z_{s_1}, ..., Z_{s_n})^\top$ has a multivariate normal distribution.

Also Example of Proposition

Proposition 15. *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 16. Hence, we denote the GP as $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$.

Example 17. When using the GP as a model we may need to parameterize its parameters. An example of mean functions are polynomial expansions, such as $\mu(s) = \sum_{j=0}^{p-1} \beta_j s^j$ for some tunable unknown parameter β. Some examples of covariance functions (c.f.), for some tunable unknown parameter β,σ^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. $c(s,s') = \sigma^2 1(s = s')$

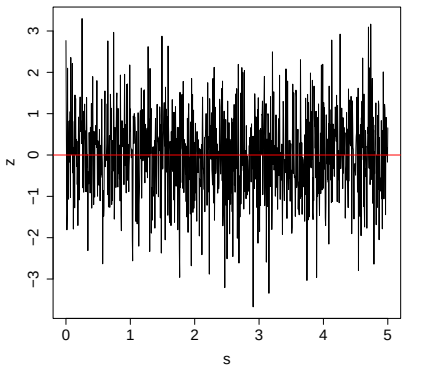
Example 18. Recall your linear regression lessons where you specified a sampling distribution $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta, \sigma^2\right)$, $\forall x \in \mathbb{R}^d$; well that can be considered as a GP with $\mu_x = x^\top\beta$ and $c(x,x') = \sigma^2 1(x = x')$ in (3).

Example 19. Figs. 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.

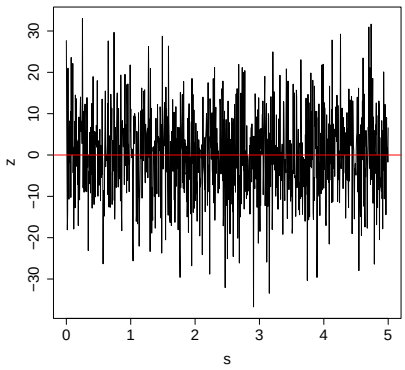
Algorithm 1 R script for simulating from a GP $(Z_s; s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

```
# 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")
```

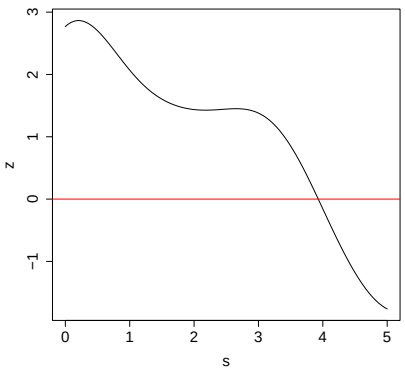
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Fig. 3.1a & 3.1b ; Fig. 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Fig. 3.2c & 3.2d), and the spatial dependence / frequency of the ups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



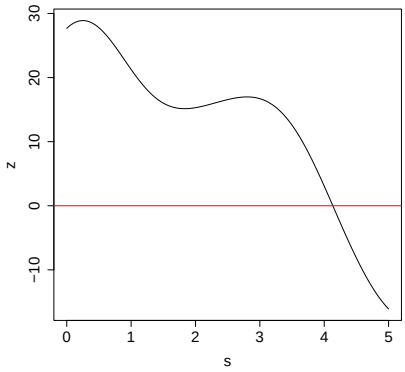
(A) Nugget c.f
($\sigma^2 = 1$)



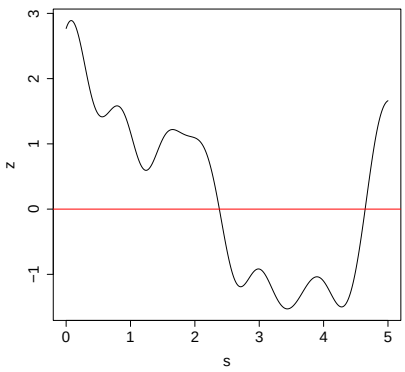
(B) Nugget c.f
($\sigma^2 = 100$)



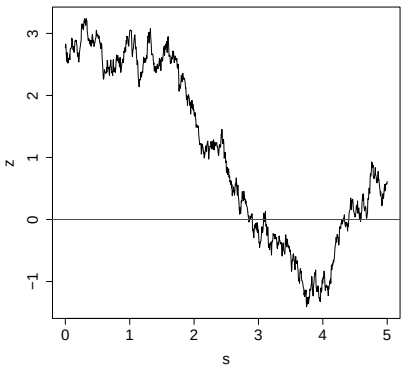
(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



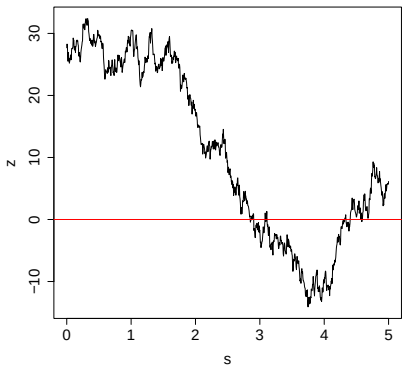
(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



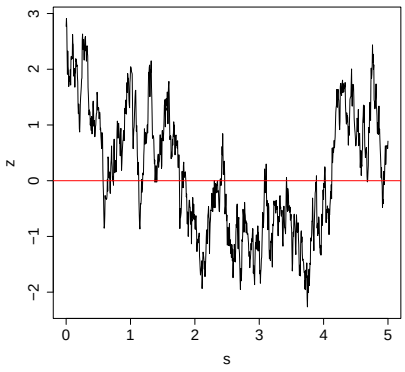
(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)



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

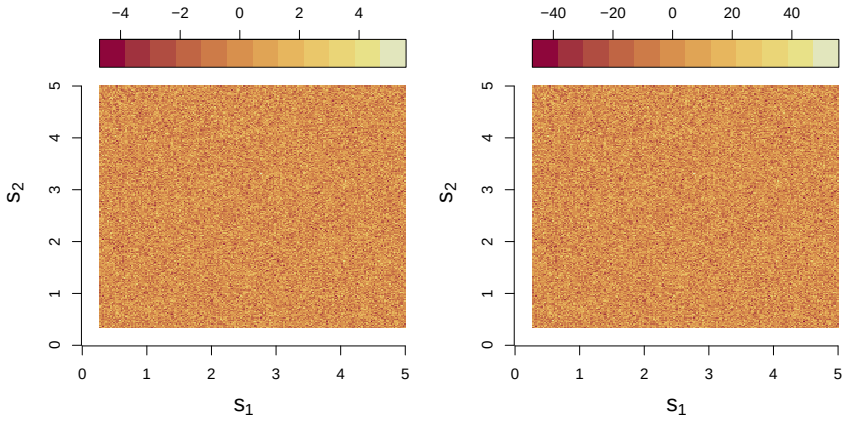


(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)



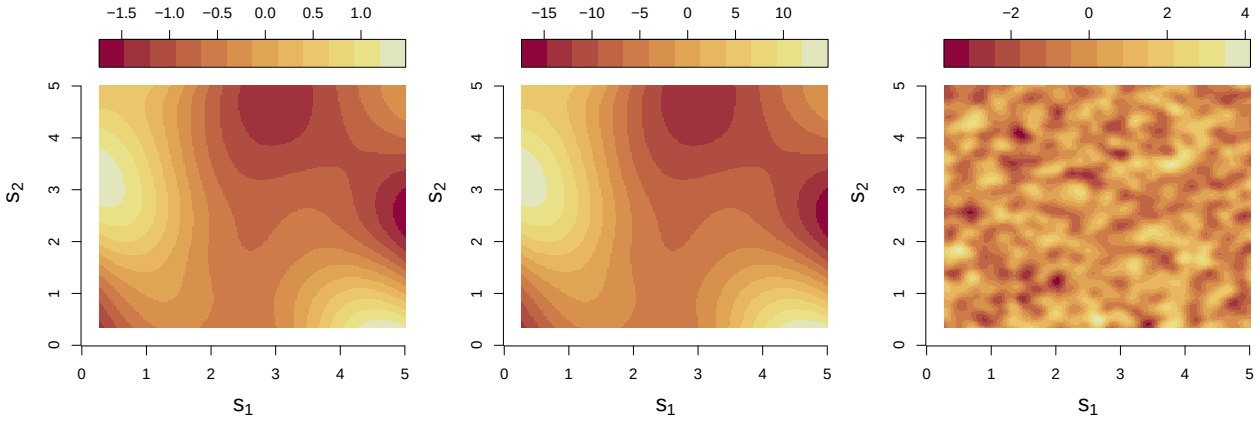
(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

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
($\sigma^2 = 1$)

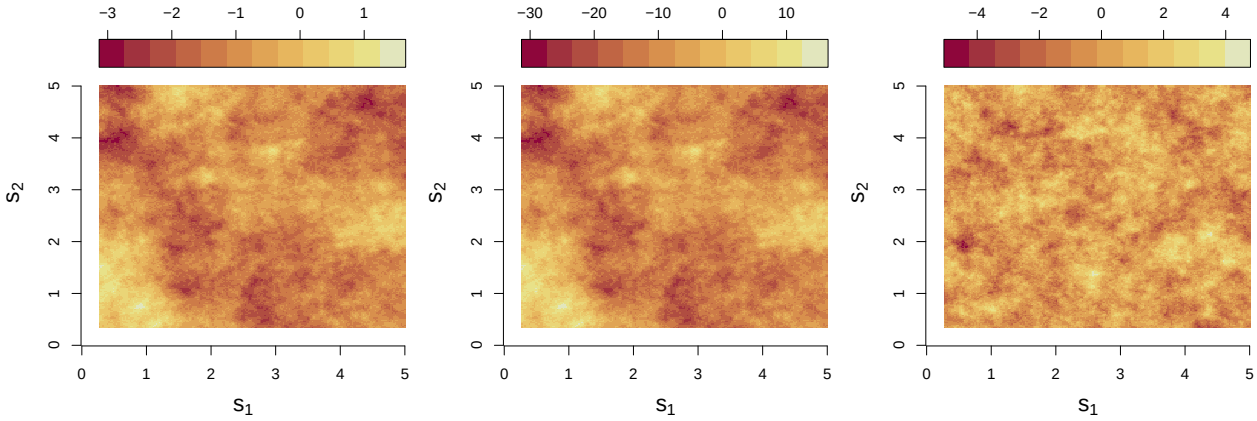
(B) Nugget c.f
($\sigma^2 = 100$)



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

Definition 21. 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 $h \in \mathbb{R}^d$

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

5. Weak stationarity (or second order stationarity)

Note 22. Yuh... strong stationary may be a too "restricting" a characteristic for our modeling... Perhaps, we could only restrict the first two moments them properly; notice Def. 21 implies that, given $\mathrm{E}\left(Z_s^2\right) < \infty$, it is $\mathrm{E}\left(Z_s\right) = \mathrm{E}\left(Z_{s+h}\right) = \text{contst...}$ and $\mathrm{Cov}\left(Z_s, Z_{s'}\right) = \mathrm{Cov}\left(Z_{s+h}, Z_{s'+h}\right) \overset{h=-s'}{=} \mathrm{Cov}\left(Z_{s-s'}, Z_0\right) = \text{funct of lag...}$

Definition 23. 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) $\mathrm{E}\left(Z_s^2\right) < \infty$ (finite)
(2) $\mathrm{E}\left(Z_s\right) = m$ (constant)
(3) $\mathrm{Cov}\left(Z_s, Z_{s'}\right) = c\left(s'-s\right)$ for some even function $c : \mathbb{R}^d \to \mathbb{R}$ (lag dependency)

Definition 24. Weakly (or second order) stationary covariance function is called the c.f. of a weakly stationary stochastic process.

6. Covariogram

Note 25. The definition of the covariogram function requires the random field to be weakly stationary.

Definition 26. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be a weakly stationary random field. The covariogram function of $Z = (Z_s)_{s\in\mathbb{R}^d}$ is defined by $c : \mathbb{R}^d \to \mathbb{R}$ with

$$c\left(h\right) = \mathrm{Cov}\left(Z_s, Z_{s+h}\right), \ \forall s \in \mathbb{R}^d.$$

Example 27. For the Gaussian c.f. $c\left(s,t\right) = \sigma^2 \exp\left(-\beta \left\|s-t\right\|_2^2\right)$ in (Ex. 17(2)), we may denote just

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

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

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

Proposition 28. *If $c : \mathbb{R}^d \to \mathbb{R}$ is the covariogram of a weakly stationary random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ then:*

(1) $c(0) \geq 0$

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

(3) $|c(h)| \leq c(0) = Var(Z_s)$ *for all* $h \in \mathbb{R}^d$

(4) $c(\cdot)$ *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 29. The following helps in the specification of cavariograms by considering properties of characteristic functions.

Theorem 30. *(Bochner's theorem) A continuous even real-valued function $c : \mathbb{R}^d \to \mathbb{R}$ is a covariance function of a weakly stationary random process 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 $\mathrm{d}F(\omega)$ is a *symmetric positive finite measure on* $\mathbb{R}^d$.

- *Here, we will focus on cases of the form* $dF(\omega) = f(\omega)\, d\omega$ *where* $f(\cdot)$ *is called spectral density of* $c(\cdot)$ *i.e.*

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

In this case, $\lim_{h\to\infty} c(h) = 0$

Theorem 31. *If $c(\cdot)$ is integrable, $F(\cdot)$ is absolutely continuous with spectral density $f(\cdot)$ of $Z = (Z_s; s \in \mathcal{S})$ then by 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 32. Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Then the spectral density from Thm 30 is

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

i.e. of a Gaussian form.

7. Intrinsic stationarity

Note 33. Getting greedier, we can further weaken the weak stationarity by considering lag dependent variance in the increments with purpose to be able to use more inclusive models; Def 23 implies that $\mathrm{Var}(Z_{s+h} - Z_s) = \mathrm{Var}(Z_{s+h}) + \mathrm{Var}(Z_s) - 2\mathrm{Cov}(Z_{s+h}, Z_s) = 2c(0) - 2c(h)$.

Definition 34. A random field $Z = (Z_s)_{s \in \mathbb{R}^d}$ is intrinsically stationary if, for all $h \in \mathbb{R}^d$, $(Z_{s+h} - Z_s)_{s \in \mathbb{R}^d}$ is weakly stationary; i.e.

(1) $\mathrm{E}(Z_{s+h} - Z_s)^2 < \infty$
(2) $\mathrm{E}(Z_{s+h} - Z_s) = m$ (constant)
(3) $\mathrm{Var}(Z_{s+h} - Z_s) = 2\gamma(h)$ for some function $\gamma : \mathbb{R}^d \to \mathbb{R}$ (lag dependent)

Definition 35. Intrinsically stationary covariance function is called the c.f. of an intrinsically stationary stochastic process.

Example 36. The following covariance function is not weakly but intrinsically stationary

$$c(s,t) = \frac{1}{2}\left(\|s\|^{2H} + \|t\|^{2H} - \|t-s\|^{2H}\right), \ H \in (0,1)$$

because for $h \in \mathbb{R}^d$

$$c(s, s+h) = \frac{1}{2}\left(\|s\|^{2H} + \|s+h\|^{2H} - \|h\|^{2H}\right)$$

and

$$\frac{1}{2}\mathrm{Var}(Z_s - Z_{s+h}) = \frac{1}{2}\left(\mathrm{Var}(Z_s) + \mathrm{Var}(Z_{s+h}) - 2\mathrm{Cov}(Z_s, Z_{s+h})\right) = \frac{1}{2}\|h\|^{2H}$$

8. (Semi) variogram

Note 37. The definition of the semi-variogram function requires the random field to be intrinsic stationarity; which is weaker assumption than weak stationary required by covariogram.

Definition 38. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be intrinsically stationary. The semi-variogram of Z is defined by $\gamma : \mathbb{R}^d \to \mathbb{R}$ with

$$\gamma(h) = \frac{1}{2}\mathrm{Var}(Z_{s+h} - Z_s), \ \ \forall s \in \mathbb{R}^d$$

Definition 39. Variogram of an intrinsically stationary random field is called the quantity $2\gamma(h)$.

Note 40. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be weakly stationary with covariogram $c(\cdot)$. Then Z is intrinsic stationary with semi-variogram

$$\gamma(h) = c(0) - c(h), \ \ \forall h \in \mathbb{R}^d \tag{8.1}$$

Example 41. For the Gaussian covariance function (Ex. 27) the semi-variogram is

$$\gamma(h) = c(0) - c(h) = \sigma^2\left(1 - \exp\left(-\beta\|h\|_2^2\right)\right)$$

Proposition 42. *Properties of semi-variogrames. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be an intrinsically stationary process.*

1. *It is $\gamma(h) = \gamma(-h)$, $\gamma(h) \geq 0$, and $\gamma(0) = 0$*
2. *Semi-variogram is conditionally negative definite (c.n.d.): for all $a \in \mathbb{R}^n$ s.t. $\sum_{i=1}^n a_i = 0$, and for all $\forall\{s_1, ..., s_n\} \subseteq S$*
$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(s_i - s_j) \leq 0$$
3. *If $\gamma(h)$ is a semi-variogram, and A is a linear transformation in $\mathbb{R}^d$ then $\tilde{\gamma}(h) = \gamma(Ah)$ is a semi-variogram too.*
4. *The following functions are semi-variograms*
   (a) *$\gamma(\cdot) = \sum_{i=1}^n a_i\gamma_i(\cdot)$, if $a_i \geq 0$, and $\{\gamma_i(\cdot)\}$ are semi-variograms*
   (b) *$\gamma(\cdot) = \int \gamma_u(\cdot)\, dF(u)$, if $\gamma_u(\cdot)$ is a semi-variogram parametrized by $u \sim F$*
   (c) *$\gamma(\cdot) = \lim_{n\to\infty}\gamma_n(\cdot)$ if $\gamma_n(\cdot)$ is semi-variogram and the limit exists*
5. *Consider intrinsically stationary stochastic processes $Y = (Y_s)_{s\in\mathbb{R}^d}$ and $E = (E_s)_{s\in\mathbb{R}^d}$ where Y and E are independent each other. Let $Z_s = Y_s + E_s$. Then*
$$\gamma_Z(h) = \gamma_Y(h) + \gamma_E(h)$$

8.1. **Behavior of variogram (Nugget effect, Sill, Range).** The variogram $\gamma(h)$ is very informative when plotted against the lag h, below we discuss some of the characteristics of it, using Fig. 8.1

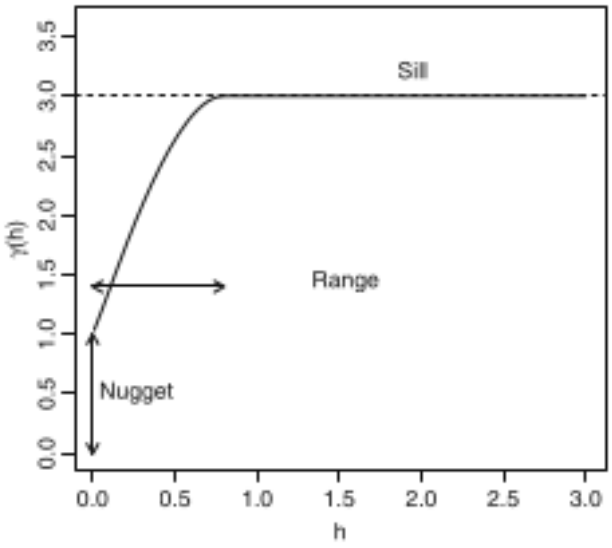


FIGURE 8.1. Variogram's characteristics

Note 43. A semivariogram tends to be an increasing function of the lag $\|h\|$. Recall in weakly stationary processes, $\gamma(h) = c(0) - c(h)$ where common logic suggests that $c(h)$ is decreases with $\|h\|$.

Note 44. If $\gamma(h)$ is a positive constant for all lags $h \neq 0$, then $Z(s_1)$ and $Z(s_2)$ are uncorrelated regardless of how close $s_1$and s_2 are; and $Z = (Z_s)_{s\in\mathbb{R}^d}$ is often called white noise.

Note 45. Conversely, a non zero slope of the variogram indicates structure.

Nugget Effect.

Note 46. Nugget effect is the semivariagram's limiting value

$$\sigma_\varepsilon^2 = \lim_{\|h\|\to 0} \gamma(h)$$

In particular when $\sigma_\varepsilon^2 \neq 0$.

Note 47. Nugget effect $\sigma_\varepsilon^2 \neq 0$ may expected or assumed to appear due to (1) measurement errors (e.g., if we collect repeated measurements at the same location s) or (2) due to some microscale variation causing discontinuity in the origin that cannot be detected from the data i.e. the spatial gaps because we collect a finite set of measurements at spatial locations. Hence theoretically, we could consider a more detailed decomposition $\sigma_\varepsilon^2 = \sigma_{\text{MS}}^2 + \sigma_{\text{MS}}^2$ where σ_{MS}^2 refers to the microscale and σ_{MS}^2 refers to the measurement error; however (my experience) this is non-identifiable.

Note 48. For a continuous processes $Z = (Z_s)_{s\in\mathbb{R}^d}$, it is expected

$$\lim_{\|h\|\to 0} \mathrm{E}(Z_{s+h} - Z_s)^2 = 0$$

which is equivalent to a continuous semivariagram $\gamma(h)$ for all h, and in particular, $\lim_{\|h\|\to 0}\gamma(h) = \gamma(0) = 0$, because $\gamma(0) = 0$. However, when modeling a real problem we may need to consider (or it may appear from the data) that $\gamma(h)$ should have a discontinuity $\lim_{\|h\|\to 0}\gamma(h) = \sigma_\varepsilon^2 \neq 0$.

Note 49. Nugget effect is often mathematically described by considering a decomposition ;

$$Z(s) = Y(s) + \varepsilon(s) \tag{8.2}$$

where Y can be a continuous stationary process with $\gamma_Y(\cdot)$, and ε can be a process (called errors-in-variables model) with (nugget) semivariogram $\gamma_\varepsilon(h) = \sigma_\varepsilon^2 1(h \neq 0)$. In this case,

$$\gamma_Z(h) = \gamma_Y(h) + \gamma_\varepsilon(h) \overset{\|h\|\to 0}{\to} \sigma_\varepsilon^2$$

Sill.

Definition 50. Sill is the variogram's limiting value $\lim_{\|h\|\to\infty}\gamma(h)$.

Note 51. For weakly stationary processes the sill is always finite. However, for intrinsic processes, the sill may be infinite.

Partial sill.

Definition 52. Partial sill is $\lim_{\|h\|\to\infty}\gamma(h) - \lim_{\|h\|\to 0}\gamma(h)$ which takes into account the nugget.

Range. Range is the distance at which the semivariogram reaches the Sill; it can be infinite. Other.

Note 53. An abrupt change in slope indicates the passage to a different structuration of the values in space. This is often modeled via decompoditions of processes with different semivariograms as in (8.2).

9. Isotropy

Note 54. Isotropy as a notion imposes the assumption of "rotation invariance" in the stochastic process.

Definition 55. An intrinsic stochastic process $(Z_s)_{s\in\mathbb{R}^d}$ is isotropic iff

$$\forall s,t \in \mathcal{S},\ \frac{1}{2}\mathrm{Var}(Z_s - Z_t) = \gamma(\|t-s\|),\ \text{for some function } \gamma{:}\mathbb{R}^+ \to\mathbb{R}. \tag{9.1}$$

Definition 56. Isotropic semi-variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$ is the semi-variogram of the isotropic stochastic process. (sometimes for simplicity of notation we use $\gamma : \mathbb{R}^+ \to \mathbb{R}$ with $\gamma(\|h\|) = \frac{1}{2}\mathrm{Var}(Z_s - Z_{s-h})$.

Definition 57. Isotropic covariance function $C : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called the covariance function satisfying (9.1).

Definition 58. Isotropic covariogram $c : \mathbb{R}^d \to \mathbb{R}$ of a weakly stationary process is the covariogram associated to an isotropic semi-variogram (sometimes for simplicity of notation we use $c : \mathbb{R}^+ \to \mathbb{R}$ with $c(\|h\|)$ from (9.1)).

9.1. **Parametric forms of frequently used isotropic covariance functions.**

Note 59. Given the covariogram $c(\cdot)$, and the semi-variogram can be computed from $\gamma(h) = c(0) - c(h)$ for any h.

9.1.1. *Nugget-effect.* For $\sigma^2 > 0$,

$$c(h) = \sigma^2 1_{\{0\}}(\|h\|).$$

It is associate to white noise. It is used to model a discontinuity in the origin of the covariogram / sem-variogram.

9.1.2. *Matern c.f.* For $\sigma^2 > 0$, $\phi > 0$, and $\nu \geq 0$

$$c(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|h\|}{\phi} \right)^{\nu} K_\nu \left(\frac{\|h\|}{\phi} \right)$$

Parameter ν controls the variogram's regularity at 0 which in turn controls the quadratic mean (q.m.) regularity of the associated process. For $\nu = 1/2$, we get the exponential c.f.,

$$c(h) = \sigma^2 \exp\left(-\frac{1}{\phi} \|h\|_1 \right)$$

which is not differentiable at $h = 0$, while for $\nu \to \infty$, we get the Gaussian c.f.

$$c(h) = \sigma^2 \exp\left(-\frac{1}{\phi} \|h\|_2^2 \right)$$

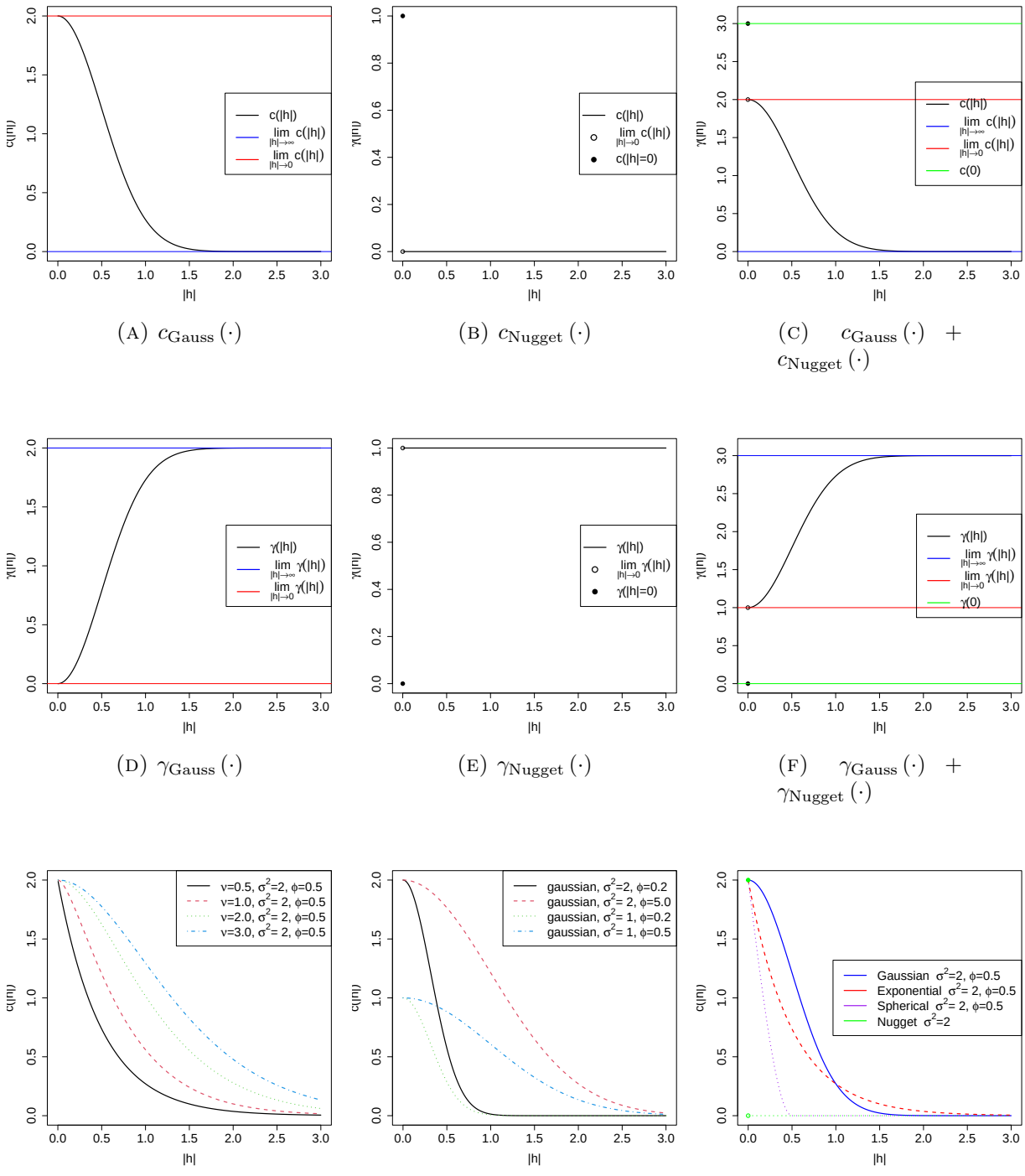
which is infinite differentiable. ϕ is a range parameter, and σ^2 is the (partial) sill parameter.

9.1.3. *Spherical c.f.* [2]For $\sigma^2 > 0$ and $\phi > 0$

$$c(h) = \begin{cases} \sigma^2 \left(1 - \frac{3}{2} \frac{\|h\|_1}{\phi} + \frac{1}{2} \left(\frac{\|h\|_1}{\phi} \right)^3 \right) & \|h\|_1 \leq \phi \\ 0 & \|h\|_1 > \phi \end{cases}, \ h \in \mathbb{R}^3. \tag{9.2}$$

The c.f. starts from its maximum value σ^2 at the origin, then steadily decreases, and finally vanishes when its range ϕ is reached. ϕ is a range parameter, and σ^2 is the (partial) sill parameter.

[2]For it's derivation see Ch 8 in [3]

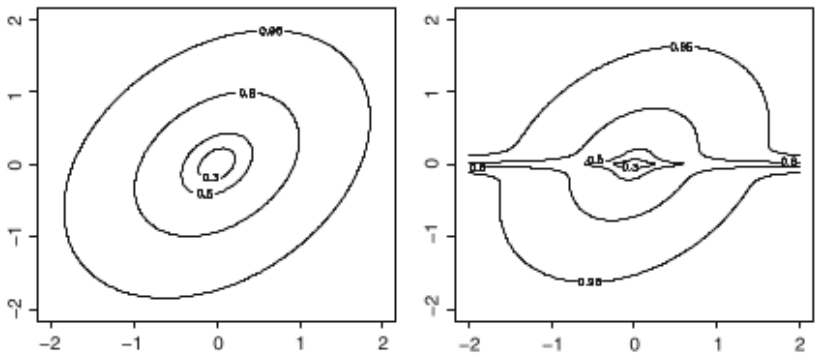


(A) $c_{\text{Gauss}}(\cdot)$ (B) $c_{\text{Nugget}}(\cdot)$ (C) $c_{\text{Gauss}}(\cdot) + c_{\text{Nugget}}(\cdot)$

(D) $\gamma_{\text{Gauss}}(\cdot)$ (E) $\gamma_{\text{Nugget}}(\cdot)$ (F) $\gamma_{\text{Gauss}}(\cdot) + \gamma_{\text{Nugget}}(\cdot)$

(G) c_{Matern} (H) $c_{\text{Gauss}}(\cdot; \sigma^2, \phi)$ (I) $c.(\cdot)$

FIGURE 9.1. Covariograms $c(\cdot)$ and semivariograms $\gamma(\cdot)$



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 10.1. Isotropy vs Anisotropy

10. ANISOTROPY

Note 60. Dependence between $Z(s)$ and $Z(s+h)$ is a function of both the magnitude and the direction of separation h. This can be caused by the underlying physical process evolving differently in space (e.g., vertical and horizontal axes).

Definition 61. The variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$ is anisotropic if there are h_1 and h_2 with same length $\|h_1\| = \|h_2\|$ but different direction $h_1/\|h_1\| \neq h_2/\|h_2\|$ that produce different variograms $\gamma(h_1) \neq \gamma(h_2)$.

Definition 62. The intrinsically stationary process $(Z_s)_{s\in\mathbb{R}^d}$ is anisotropic if its variogram is anisotropic.

Definition 63. The covariogram $c : \mathbb{R}^d \to \mathbb{R}$ is anisotropic if there are h_1 and h_2 with same length $\|h_1\| = \|h_2\|$ but different direction $h_1/\|h_1\| \neq h_2/\|h_2\|$ that produce different covariogram $c(h_1) \neq c(h_2)$.

Definition 64. The weakly stationary process $(Z_s)_{s\in\mathbb{R}^d}$ is anisotropic if its covariogram is anisotropic.

Note 65. For brevity, below we discuss about intrinsically stationary process and variograms, however the concepts/definitions apply to weakly stationary process and covariograms when defined, as in Defs 61 & 63.

10.1. **Geometric anisotropy.**

Definition 66. The semi-variogram $\gamma_{\text{g.a.}} : \mathbb{R}^d \to \mathbb{R}$ exhibits geometric anisotropy if it results from an A-linear deformation of an isotropic semi-variogram with function $\gamma_{\text{iso}}(\cdot)$; i.e.

$$\gamma_{\text{g.a.}}(h) = \gamma_{\text{iso}}(\|Ah\|_2)$$

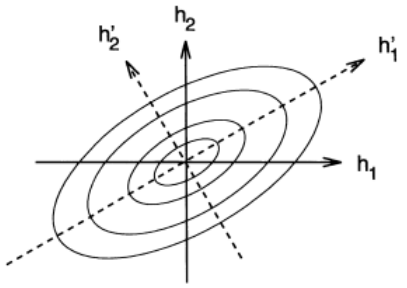


FIGURE 10.2. Rotation of the 2D coordinate system

Note 67. Such variograms have the same sill in all directions but with ranges that vary depending on the direction. See Fig 10.1a.

Example 68. For instance, if $\gamma_{\text{g.a.}}(h) = \gamma_{\text{iso}}\left(\sqrt{h^\top Q h}\right)$, where $Q = A^\top A$.

Example 69. [Rotating and dilating an ellipsoid in 2D] Consider a coordinate system for $h = (h_1, ..., h_n)^\top$. We wish to find a new coordinate system for h in which the iso-variogram lines are spherical.

(1) [Rotate] Apply rotation matrix R to h such as $h' = Qh$. In 2D, it is

$$R = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \text{ for } \theta \in (0, 2\pi), \text{ is the rotation angle.}$$

(2) [Dilate] Apply a dilation of the principal axes of the ellipsoid using a diagonal matrix $\sqrt{\Lambda} = \text{diag}\left(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}\right)$, as $\tilde{h} = \sqrt{\Lambda} h'$.

Now the ellipsoids become spheres with radius $r = \left\|\tilde{h}\right\|_2 = \sqrt{\tilde{h}^\top \tilde{h}}$. This yields the equation of an ellipsoid in the h coordinate system

$$h^\top \left(R^\top \Lambda R\right) h = r^2$$

where the diameters d_j (principal axes) of the ellipsoid along the principal directions are

$$d_j = 2r/\sqrt{\lambda_j}$$

and the principal direction is the j-th column of the rotation matrix $R_{:,j}$.

Hence the anisotropic semivariogram is $\gamma_{\text{g.a.}}(h) = \gamma_{\text{iso}}\left(\sqrt{h^\top Q h}\right)$ with $Q = R^\top \Lambda R$. This derivation extends to d dimensions.

10.2. **Zonal (or stratified) anisotropy.**

Definition 70. Support anisotropy is called the type of anysotropy when the semi-variogram $\gamma(h)$ of the process depends only on certain coordinates of h.

Example 71. If it is $\gamma(h = (h_1, h_2)) = \gamma(h_1)$, then I ve support anisotropy

Definition 72. Zonal anisotropy occurs when the semi-variogram $\gamma(h)$ is the sum of several components each with a support anisotropy.

Example 73. Let γ' and γ'' be semi-variograms. If it is $\gamma(h=(h_1,h_2)) = \gamma'(\|h_1\|) + \gamma''\left(\sqrt{\|h_1\|+\|h_2\|}\right)$, then I 've Zonal anisotrotropy.

Note 74. We have Zonal anisotropy then the variograms calculated in different directions suggest a different value for the sill (and possibly the range).

Note 75. If in 2D case, the sill in h_1 is larger than that in h_2, we can model zonal anysotropy of stochastic process (Z_s) by assuming $Z(s) = I(s) + A(s)$, where $I(s)$ is an isotropic process with isotropic semi-variogram γ_I along dimension of h_1 and $A(s)$ is an process with anisotropic semi-variogram γ_I without effect on dimension h_1; i.e. $\gamma_Z(h) = \gamma_I(h) + \gamma_A(h)$.

10.3. **Non-linear deformations.**

Note 76. A (rather too general) non-stationary model can be specified by considering semi-variogram $2\mathrm{Var}(Z(s) - Z(t)) = 2\gamma_o(\|G(s) - G(t)\|)$ where we have performed a bijective non-linear (function) deformation $G(\cdot)$ of space $\mathcal{S}$ and applied on the isotropic semi-variogram γ_o. For instance, $\gamma_o(h) = \sigma^2 \exp(-\|h\|/\phi)$ and $G(s) = s^2$ as a deterministic function. Now, if function $G(\cdot)$ is considered as unknown, one can model it as a stochastic process $(G_s)_{s\in\mathcal{S}}$, and then we will be talking about deep learning modeling stuff.

11. Geometrical properties

(!): We discuss basic geometric properties of the basic models we will use for modeling, as it can give us a deeper intuition on how to design appropriate spatial statistical models.

Definition 77. (Continuity in quadratic mean (q.m.)) Second-order process $Z = (Z_s)_{s\in S}$ is q.m. continuous at $s \in \mathcal{S}$ if

$$\lim_{h\to 0} \mathrm{E}(Z(s+h) - Z(s))^2 = 0.$$

Proposition 78. *For $Z = (Z_s)_{s\in S}$ it is*

$$E(Z(s+h) - Z(s))^2 = (E(Z(s+h)) - E(Z(s)))^2 + Var(Z(s+h) - Z(s))$$

- *If Z is intrinsically stationary, then*

$$E(Z(s+h) - Z(s))^2 = \frac{1}{2}\gamma(h)$$

  and hence q.m. continuous iff $\lim_{h\to 0}\gamma(h) = \gamma(0)$.
- *If Z is weakly stationary, then*

$$E(Z(s+h) - Z(s))^2 = \frac{1}{2}(c(0) - c(h))$$

and hence q.m. continuous iff $\lim_{h\to 0} c(h) = c(0)$ *(i.e. ,c is continuous).*

Note 79. It has been shown that if a random field $Z = (Z_s)_{s\in S}$ has a variogram which is everywhere continuous apart from the origin i.e. $\lim_{s\to 0}\gamma(s) \neq \gamma(0)$ then Z it can be represented as $Z_s = Y_s + \varepsilon_s$ where (Y_s) has everywhere a continuous variogram and (ε_s) has a nugget effect, and Y_s, ε_s are independent. [2; Ch 1.4.1]

Definition 80. Differentiable in quadratic mean (q.m.)) Second-order process $Z = (Z_s)_{s\in\mathbb{R}}$ is q.m. differentiable at $s \in \mathbb{R}$ there exist

$$\dot{Z}(s) = \lim_{h\to 0} \frac{Z(s+h) - Z(s)}{h}. \text{ in } q.m. \tag{11.1}$$

Proposition 81. *Let $c(s,t)$ be the covariance function of $Z = (Z_s)_{s\in S}$. Then Z is everywhere differentiable if $\frac{\partial^2}{\partial s\partial t}c(s,t)$ exists and it is finite. Also, $\frac{\partial^2}{\partial s\partial t}c(s,t)$ is the covariance function of (11.1).*

Example 82. The process with Gaussian c.f. $c(h) = \sigma^2 \exp(-|h|/\phi)$ is continuous because $\lim_{h\to 0} c(h) = \sigma^2 = c(0)$ but not differentiable because $\frac{\partial^2}{\partial h^2}c(h)$ does not exist at $h = 0$.

Part 2. Model building

12. The Geostatistical model

12.1. **Linear Model of Regionalization.** A spatial phenomenon can be thought as being the sum of several independent subphenomena acting at different characteristic scales. A linear model can be set up to splits the stochastic process representing the phenomenon into several uncorrelated stochastic processes, each with a different variogram or covariance function and characterizing different aspect of the overall phenomenon under investigation. Decomposition of the stochastic process.

Note 83. The linear model of regionalization consider the decomposition of the stochastic process of interest $Z(s)$ as a summation of m independent zero-mean stochastic processes $\{Z_j(s)\}_{j=0}^m$ each of them characterizing different spatial scales, as

$$Z(s) = \mu(s) + Z_1(s) + ... + Z_m(s) \tag{12.1}$$

with $\mu(s) = \mathrm{E}(Z(s))$ be a deterministic function.

Note 84. In (12.1), let $Z_j(\cdot)$ be intrinsically stationary with semi-variogram $\gamma_j(\cdot)$, then the semi-variogram of $Z(\cdot)$ is $\gamma(\cdot) = \sum_{j=1}^m \gamma_j(\cdot)$.

Example 85. For instance consider (12.1) with $\mu(s) = 0$, $m = 3$, $Z_1(s)$ with a spherical semi-variogram (9.2) with range $\phi_1 = 3.5$, $Z_2(s)$ with a spherical semi-variogram (9.2) with

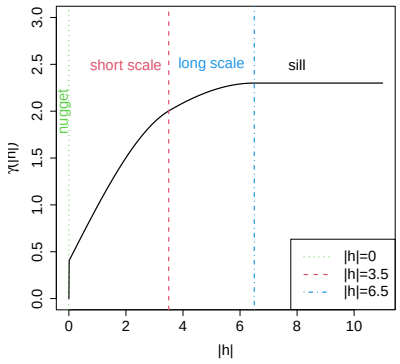


FIGURE 12.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|;\sigma^2 = 0.4\right)$.

range $\phi_2 = 6.5$, and $Z_3(s)$ with a nugget semi-variogram. See the "sudden" changes of the line in Fig. 12.1 representing change of spatial behavior.

12.2. **Scale of variation.**

Note 86. Cressi [1] Consider the following intuitive decomposition

$$Z(s) = \mu(s) + W(s) + \eta(s) + \varepsilon(s), \ s \in S \tag{12.2}$$

where

$\mu(s) = \mathbf{E}(Z(s))$**:** is the deterministic mean structure. It aims to represent the "large scale variation".

$W(s)$**:** is a zero mean second order continuous intrinsically stationary process whose range is larger than gaps between the sites (sampling grid). It aims to represent "smooth small scale variation".

$\eta(s)$**:** is a zero mean intrinsically stationary process whose variogram range exists and is smaller than the gaps between the sites. It aims to represent "microscale variation"

$\varepsilon(s)$**:** is a zero-mean white-noise process (modeled as nugget effect). It aims to represent "measurement error or noise"

$W(s)$, $\eta(s)$, $\varepsilon(s)$ are mutually independent.

Note 87. Reasonably, larger scale components, such as $\mu(s)$, $W(s)$ can represented in the variogram if the diameter of the sampling domain is large $\mathcal{S}$ is large enough.

Note 88. Clearly, smaller scale components, such as $\eta(s)$, $\varepsilon(s)$ could be identified if the sampling grid is sufficiently fine.

Note 89. Decomposition 86 is not unique and the components are not clearly identifiable from the data when modeled; e.g. one may find two pairs of $\mu(s), W(s)$ doign the same thing; yet, separating $\eta(s)$ and $\varepsilon(s)$ is difficult as they often describe changes with range smaller than that of the sites (!)

Note 90. The geostatistical model (decomposition) is often presented (with reference to (12.2)) as

$$Z(s) = \mu(s) + w(s) + \varepsilon(s), \ s \in S$$

where $w(s) = W(s) + \eta(s)$ contains all the spatial variation.

Note 91. Alternatively, the hierarchical statistical model (Handout 1, 3.5) is used

$$Z(s) = Y(s) + \varepsilon(s), \ s \in S \tag{12.3}$$

where $Y(s) = \mu(s) + W(s) + \eta(s)$ is the spatial process model, or latent process or signal process or noiseless process.

Note 92. A simpler decomposition is

$$Z(s) = \mu(s) + \delta(s), \ s \in S$$

where $\delta(s) = W(s) + \eta(s) + \varepsilon(s)$ is the called the correlated process.

13. Training & inference

Note 93. Suppose that the intrinsic stationary random field $(Z_s)_{s \in \mathcal{S}}$, $\mathcal{S} \in \mathbb{R}^d$ is observed at n sites $S = \{s_1, ..., s_n\}$, and we get n observed dataset $\{(s_i, Z(s_i))\}_{i=1}^n$.

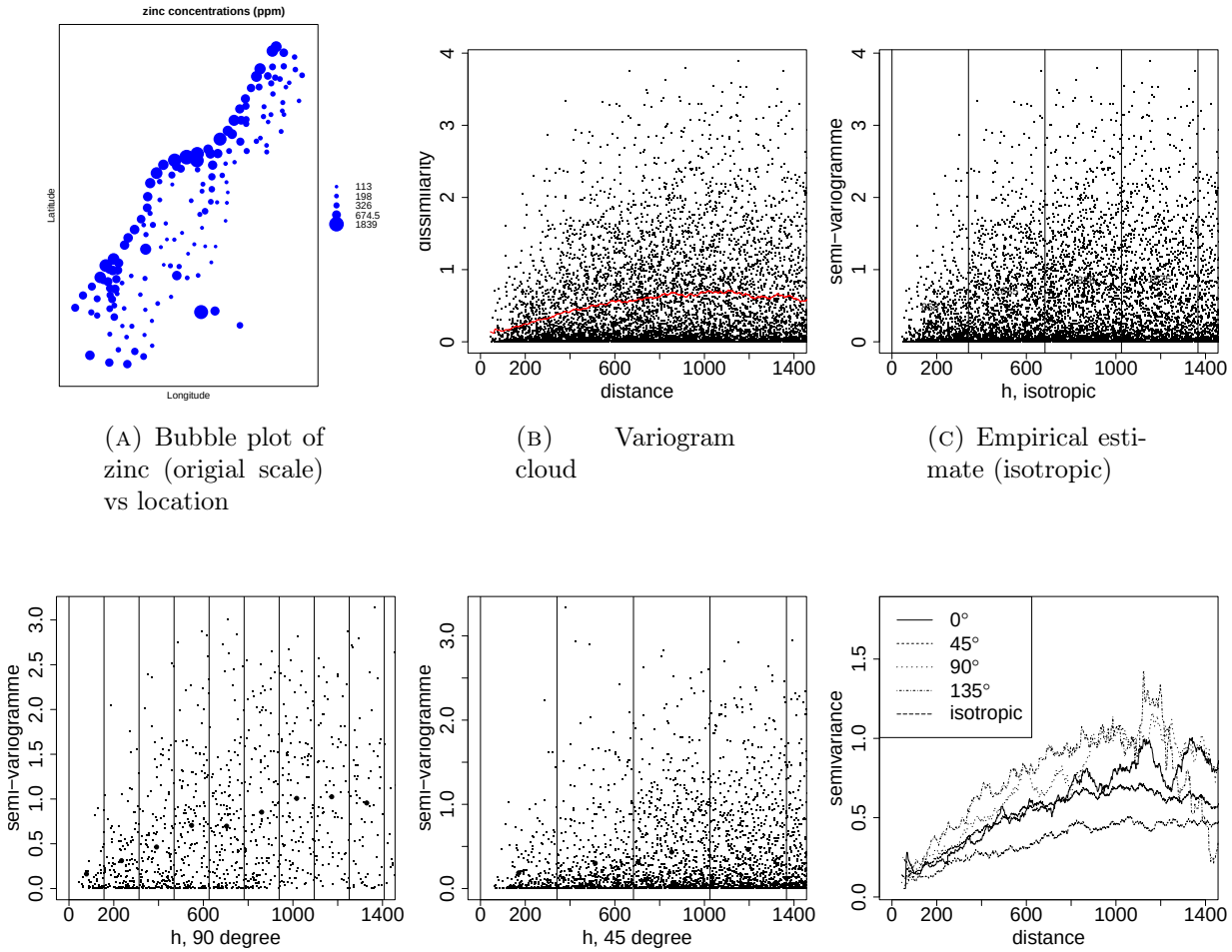
Example 94. (Meuse river data set) The Meuse river dataset set, used as a running example gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Here, we use the topsoil zinc concentration, mg kg-1 soil ("ppm") as quantity of interest (Z). Heavy metal concentrations are from composite samples of an area of approximately 15m × 15m. See Fig. 13.1a. This is the R dataset `meuse{sp}`.

Example 95. (Wolfcamp-aquifer dataset) We also consider the Wolfcamp-aquifer dataset in Ex 5 in the Exercise sheet. See Fig. 13.2a

13.1. **The variogram cloud.**

Definition 96. Dissimilarity between pairs of data values $Z(s_a)$ and $Z(s_b)$ is called the measure

$$\gamma^\star(s_a, s_b) = \frac{1}{2}(Z(s_b) - Z(s_a))^2 \tag{13.1}$$



(A) Bubble plot of zinc (origial scale) vs location

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 13.1. Meuse dataset variogram estimations (Zinc in log scale)

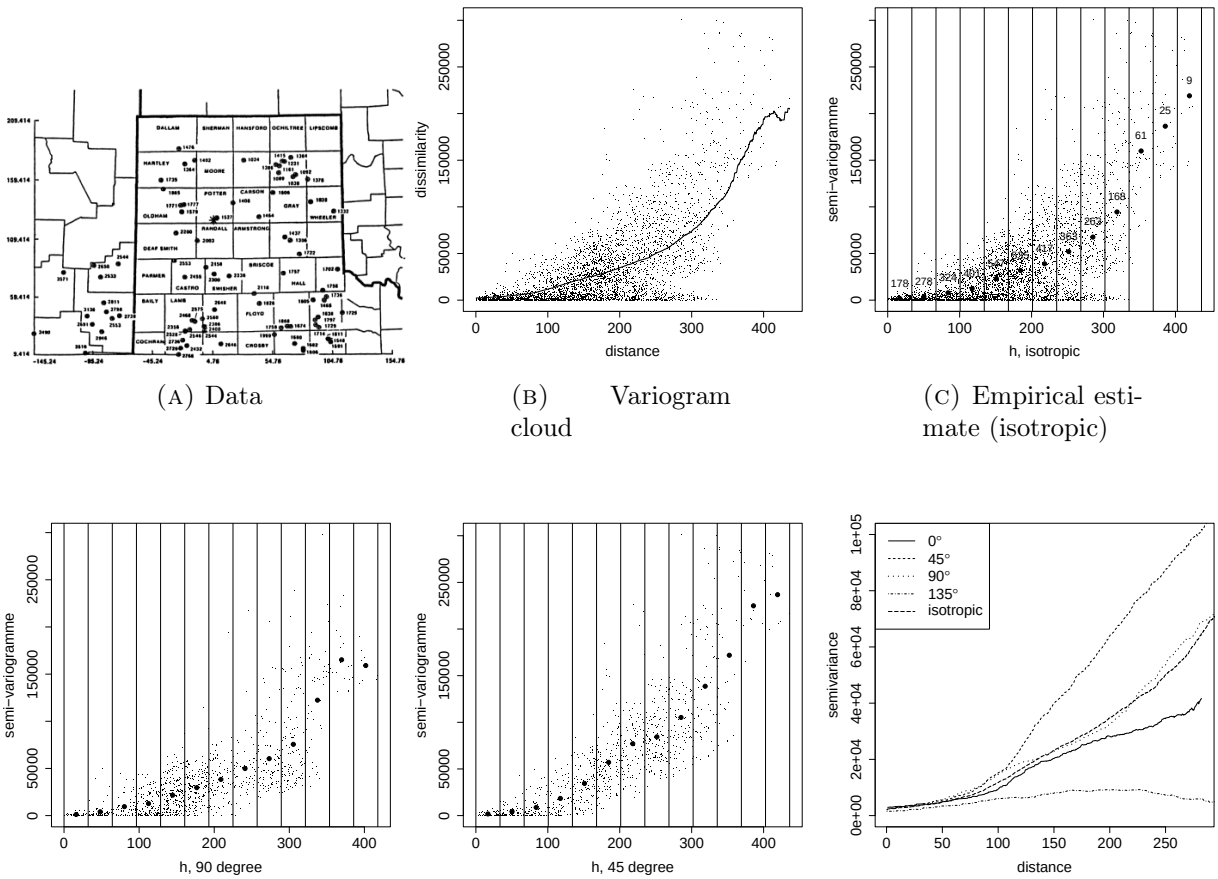
Definition 97. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (distance and orientation) then we get

$$\gamma^{\star}(h) = \frac{1}{2}\left(Z(s+h) - Z(s)\right)^2.$$

Definition 98. The variogram cloud is the set of $n(n-1)/2$ points

$$\mathfrak{C}_S = \{(\|s_i - s_j\|, \gamma^{\star}(s_j, s_i)),\ i, j = 1, ..., n, \text{ and } s_i \neq s_j\}$$

Note 99. Note that (13.1) is an unbiased estimator of the variogram and hence the variogram cloud is too.



(A) Data

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 13.2. Wolfcamp-aquifer dataset variogram estimations

Note 100. Often there is a smoothing of the cloud is superimposed onto the cloud itself to help us see variogram's characteristics (e.g., sill, nugget, range) which may be "hidden" due to potential outliers in the plot.

Example 101. Fig. 13.1b and Fig. 13.2b show the variogram cloud plots (that is a point plot of the dissimilarities vs the distances) for the datasets Meuse and Wolfcamp-aquifer dataset. The red line is a smoother line of the cloud.

13.2. **Non-parametric estimation of variogram.**

Proposition 102. *Smoothed Matheron estimator* $\hat{\gamma}(\cdot)$ *of semi-variogram* $\gamma(\cdot)$ *is*

$$\hat{\gamma}_M(h) = \frac{1}{2\left|N_{r_1,r_2}(h)\right|} \sum_{\forall (s_i,s_j)\in N_{r_1,r_2}(h)} \left(Z(s_i) - Z(s_j)\right)^2 \tag{13.2}$$

where

$$N_{r_1,r_2}(h) = \{(s_i, s_j) \in \mathcal{S} : s_i - s_j \in B_{r_1,r_2}(h)\}$$

contains all the pairs of spatial points points whole difference is in a ball

$$B_{r_1,r_2}(h) = \left\{ x : |\|x\| - \|h\|| < r_1, \text{ and } \left\| \frac{x}{\|x\|_2} - \frac{h}{\|h\|_2} \right\|_2 < r_2 \right\} \tag{13.3}$$

centered at h *with radius* $r_1 > 0$ *and* $r_2 > 0$.

Note 103. Estimator 13.2 can be writen in matrix form as $\hat{\gamma}_{\mathrm{M}}(h) = Z^{\top} A(h) Z$, where $[A(h)]_{i,j} = 1(i \neq j) - 1/|N_{r_1,r_2}(h)|$ is a positive definite matrix.

Note 104. If we consider isotropic semi-variogram $\gamma(\cdot)$ then the ball may just considerate only the length of the distance as

$$B_{r_1}(h) = \{x : |\|x\| - \|h\|| < r_1\} \tag{13.4}$$

because the direction does not have any effect.

Note 105. The choice of r_1 , r_2 is an art, and a trade-off between variance and bias, similar to the bin length in histograms.

Note 106. In practice, we consider a finite number of k separations $\mathscr{H} = \{h_1, ..., h_k\}$, we estimate in such a way that each class contains at least 30 pairs of points. Then compute $\{\hat{\gamma}_{\mathrm{M}}(h); h \in \mathscr{H}\}$, and plot $\{(h_j, \hat{\gamma}_{\mathrm{M}}(h_j)); j = 1, ..., k\}$.

Example 107. Figs 13.1c and 13.2c, show the nonparametric estimator ignoring the direction for the datasets Meuse and Wolfcamp-aquifer dataset. The estimator is calculated by using the ball in (13.4).

Example 108. Figs 13.2dand 13.1e show the nonparametric estimator considering directions 90° and 45° for the dataset Meuse. Figs 13.2d and 13.2e do the same for the Wolfcamp-aquifer dataset. The estimator is calculated by using the ball (13.3).

Note 109. In practice anisotropies are detected by inspecting experimental variograms in different directions and are induced into the model by tuning predefined anisotropy parameters.

Example 110. Fig 13.1f and 13.2a show the nonparametric variogram estimator for different directions for the two datasets. We observe possible anisotropy due to the differences in the lines.

Proposition 111. *Assume a stationary Gaussian process* $(Z_s \sim GP(0, c(\cdot,\cdot)))_{s\in\mathcal{S}}$ *with semi-variogram* $\gamma(\cdot) = c(0) - c(\cdot)$. *The empirical semi-variogram* $\hat{\gamma}_{\mathrm{M}}$ in *(13.2) is*

$$\hat{\gamma}_M(h) \sim \sum_{i=1}^{\left|N_{r_1,r_2}(h)\right|} \lambda_i \xi_i$$

where $\xi_i \overset{iid}{\sim} \chi^2_1$ *and* $\{\lambda_i\}$ *are the non-zero eigen-values of* $A(h)\,C$, $[C]_{i,j} = c(s_i, s_j)$.

Note 112. Estimation of the covariogram is done by

$$\hat{c}(h) = \frac{1}{2\left|N_{r_1,r_2}(h)\right|} \sum_{\forall (s_i,s_j)\in N_{r_1,r_2}(h)} \left(Z(s_i) - \bar{Z}\right)\left(Z(s_j) - \bar{Z}\right) \tag{13.5}$$

where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$. It's sampling distribution etc. can be computed in a similar manner.

13.3. **Parametric estimation.**

Note 113. Smoothed Matheron estimator (13.2) does not necessarily satisfies semi-variogram properties, such as negative definiteness. To address this we use a parametric family of appropriate semi-variogram functions and tune them against data.

Note 114. Popular parametrized isotopic semi-variogrames/covariogrames are those Sec 9.1. Anisotropic semi-variogrames/covariogrames can be specified by using isotropic ones and applying a rotation and dilation as in Ex 68.

Proposition 115. *(Criteria checking variogram's validity.) A continuous function* $2\gamma(\cdot)$ *with* $\gamma(0) = 0$ *is a valid variogram iff: any of the following is satisfied:*

(1) $\lim_{\|h\|\to\infty} \frac{\gamma(h)}{\|h\|^2} = 0$ *, or*

(2) $\exp(-a\gamma(\cdot))$ *is positive definite for any* $a > 0$.

Example 116. Gaussian semi-variogram in Ex 41, it is

$$\lim_{\|h\|\to\infty} \frac{\gamma(h)}{\|h\|^2} = \lim_{\|h\|\to\infty} \frac{\sigma^2\left(1 - \exp\left(-\beta\|h\|_2^2\right)\right)}{\|h\|^2} = -\lim_{\|h\|\to\infty} \frac{\exp\left(-\beta\|h\|_2^2\right)}{\|h\|^2} = 0.$$

Yet $\gamma(h) = \|h\|^2$ is variogram as well because $\exp\left(-\beta\|h\|_2^2\right)$ is a c.f. and hence positive definite.

13.3.1. *Least Square Errors training methods for semi-variogram.*

Proposition 117. *(Least Square Errors) Consider that the empirical semivariogram* $\hat{\gamma}$ *(e.g., Matheron (13.2)) of* γ *have been computed at* k *classes, i.e. it is available* $\{h_j, \hat{\gamma}(h_j)\}_{j=1}^{k}$.

The Least Square Errors (LSE) estimator of $\gamma_\theta(h)$ parametrised by the unknown θ for all h is $\hat{\gamma}_{LSE}(h) = \gamma\left(h; \hat{\theta}_{LSE}\right)$, where

$$\hat{\theta}_{LSE} = \arg\min_\theta \left(\hat{\gamma} - \gamma(h;\theta)\right)^\top V(\theta) \left(\hat{\gamma} - \gamma(h;\theta)\right), \tag{13.6}$$

$V(\theta)$ is a user specific positive definite matrix $V(\theta)$ serving as a weight, $\hat{\gamma} = \left(\hat{\gamma}(h_1), ..., \hat{\gamma}(h_k)\right)^\top$, and $\gamma(h;\theta) = \left(\gamma(h_1;\theta), ..., \gamma(h_k;\theta)\right)^\top$.

Proposition 118. *(Ordinary least squares) If use $V(\theta) = I$ in (13.6), we get the OLS $\hat{\gamma}_{OLS}(h) = \gamma\left(h; \hat{\theta}_{OLS}\right)$*

$$\hat{\theta}_{OLS} = \arg\min_\theta \left(\sum_j \left(\hat{\gamma}(h_j) - (h;\theta)\right)^2\right) \tag{13.7}$$

Proposition 119. *(Weighted least squares) If use $V(\theta) = diag\left(\varpi_1(\theta), ..., \varpi_k(\theta)\right)$ for some weight function $\{\varpi_j(\theta)\}$, we get the WLE $\hat{\gamma}_{WLE}(h) = \gamma\left(h; \hat{\theta}_{WLE}\right)$*

$$\hat{\theta}_{WLE} = \arg\min_\theta \left(\sum_j \varpi_j(\theta) \left(\hat{\gamma}(h_j) - \gamma_\theta(h_j)\right)^2\right) \tag{13.8}$$

For instance $\varpi_j(\theta) = |N_r(h_j)|$ or $\varpi_j(\theta) = |N_r(h_j)| / \left(\gamma_\theta(h_j)\right)^2$.

Example 120. Figs 13.3a and 13.3b show the OLE and WLE estimates (13.7) and (13.8) of the exponential and spherical semi-variogram for the Meuse dataset. Fig 13.3c shows the OLE and WLE estimates (13.7) and (13.8) of the exponential semi-variogram for the Wolfcamp dataset. The parametric semi-variograms were tuned against the non-parametric estimator (13.2) presented in dots, as discussed in Proposition 117.

13.3.2. *Training methods for semi-variogram with trend.*

Note 121. Assume a stochastic process model (Z_s) decomposed as

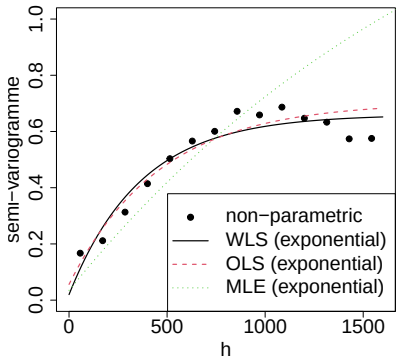
$$Z(s) = \mu(s;\beta) + \delta(s;\theta)$$

where the trend $\mu(s;\beta)$ is parameterized by unknown β (e.g. $\mu(s;\beta) = s^\top\beta$), and the zero mean intrinsic process $\delta(s;\theta)$ has a semi-variogram $\gamma(h;\theta)$ parameterised by unknown θ.

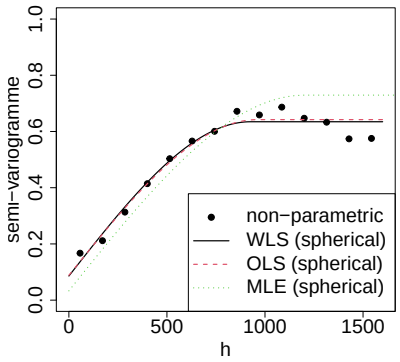
Proposition 122. *(Least square errors with trend) Do the following:*

(1) Compute estimates $\hat{\beta}$ via LSE (or equivalent)

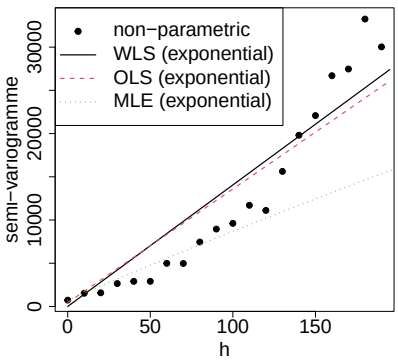
$$\hat{\beta}_{LSE} = \arg\min_\theta \left(\sum_i \left(Z(s_i) - \mu(s_i;\beta)\right)^2\right)$$



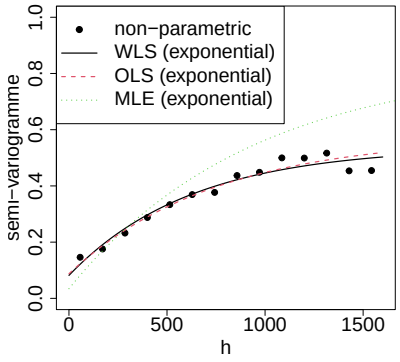
(A) Exponential c.f. with constant mean (Meuse dataset)



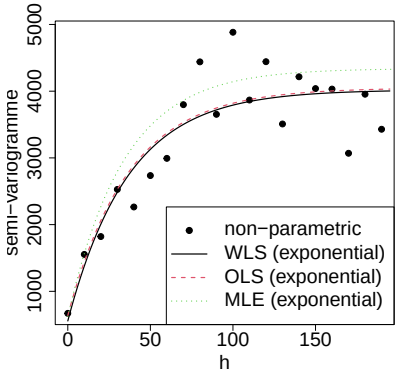
(B) Spherical c.f. with constant mean (Meuse dataset)



(C) Exponential c.f. with constant mean (Wolfcamp dataset)



(D) Exponential c.f. with linear trend (Meuse dataset)



(E) Exponential c.f. with linear trend (Wolfcamp dataset)

FIGURE 13.3. Parametric training

(2) Compute the residuals $\hat{\delta} := \hat{\delta}(s_i)$ *from*

$$\hat{\delta}(s_i) = Z(s_i) - \mu\left(s_i; \hat{\beta}_{LSE}\right)$$

(3) Estimate the empirical variogram for $\hat{\delta}$ *on* $\mathscr{H}$ *according to Prop 102, and estimate* θ *according to Prop 117.*

Example 123. Fig 13.3a and 13.3b fit an exponential c.f. and a spherical c.f. in the data of Meuse dataset (assuming constant mean); we cannot eyeball any big difference. Fig 13.3c fit an exponential c.f. in the data of Wolfcamp dataset (assuming constant mean); the fit looks really bad, possibly we should consider a non-constant mean and remove the trend.

Example 124. Fig 13.3d fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-42.8, -9.5 \cdot 10^{-4}, -6.6 \cdot 10^{-4})^{\top}$in Meuse dataset. Possibly inference would suggest a constant mean function. Fig 13.3e fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-607, -1.12, -1.13)^{\top}$ in Wolfcamp dataset; we see an improvement in fit compared to Fig 13.3c.

13.4. **Training via Maximum likelihood estimation.**

Note 125. Given that a probability distribution has been specified for the stochastic process $(Z_s)_{s\in\mathcal{S}}$, the MLE involves (1) the derivation of the associated pdf $\text{pr}(Z_1, ..., Z_n|\beta, \theta)$ of the n-dimensional sampling distribution, (2) the computation of the associated likelihood function $L(z_1, ..., z_n|\beta, \theta)$ given the data $\{(s_i, Z_i)\}_{i=1}^n$, and finally (3) the computation of the MLE estimates $\left(\hat{\beta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}\right)$ of $\left(\hat{\beta}, \hat{\theta}\right)$ as

$$\left(\hat{\beta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}\right) = \arg\min_{\beta,\theta} \left(-2\log\left(L(z_1, ..., z_n|\beta, \theta)\right)\right)$$

Example 126. If $(Z_s)_{s\in\mathcal{S}}$ is specified as $Z(\cdot) \sim \text{GP}(\mu(\cdot;\beta), c(\cdot,\cdot;\theta))$, with $\mu(s;\beta) = \beta_0 + s_1\beta_1 + s_2\beta_2$ then MLE of (β, θ) is

$$\left(\hat{\beta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}\right) = \arg\min_{\beta,\theta} \left(-2\log\left(\text{N}(Z|\mu_\beta, C_\theta)\right)\right)$$

where $\text{N}(Z|\mu_\beta, C_\theta)$ is the Gaussian pdf at $Z = (Z(s_1), ..., Z(s_n))^{\top}$, with mean $[\mu_\beta]_i = \mu(s_i;\beta) = \beta_0 + s_{1,i}\beta_1 + s_{2,i}\beta_2$ and covariance matrix $[C_\theta]_{i,j} = c_\theta(s_i, s_j)$.

13.5. **Training via Bayesian statistics.**

Note 127. Given that a probability distribution has been specified for the stochastic process $(Z_s)_{s\in\mathcal{S}}$, the Bayesian training involves (1) the derivation of the pdf $\text{pr}(Z_1, ..., Z_n|\beta, \theta)$ of the n-dimensional sampling distribution, (2) the computation of the associated likelihood function $L(z_1, ..., z_n|\beta, \theta)$ given the data $\{(s_i, Z_i)\}_{i=1}^n$; and (3) the specification of the prior model $(\beta, \theta) \sim \text{pr}(\beta, \theta)$, leading to the Bayesian hierarchical model

$$\begin{cases} Z|\beta, \theta \sim \text{pr}(Z|\beta, \theta) & \text{sampling distr.} \\ \beta, \theta \sim \text{pr}(\beta, \theta) & \text{priors} \end{cases}$$

Posterior moments can be derived from the posterior distribution of β, θ given is given the data by using the Bayes theorem as

$$\text{pr}(\beta, \theta|Z) = \frac{\text{pr}(Z|\beta, \theta)\,\text{pr}(\beta, \theta)}{\int \text{pr}(Z|\beta, \theta)\,\text{pr}(\beta, \theta)\,\text{d}(\beta, \theta)}.$$

Note 128. If the stochastic model is $Z(\cdot) \sim \mathrm{GP}(\mu(\cdot;\beta), c(\cdot,\cdot;\theta))$, and specify priors $(\beta,\theta) \sim \mathrm{pr}(\beta,\theta)$, the Bayesian hierarchical model is

$$\begin{cases} Z|\beta,\theta \sim \mathrm{N}(Z|\mu_\beta, C_\theta) & \text{sampling distr.} \\ \beta,\theta \sim \mathrm{pr}(\beta,\theta) & \text{priors} \end{cases}$$

and the posterior is given by the Bayes theorem as

$$\mathrm{pr}(\beta,\theta|Z) = \frac{\mathrm{N}(Z|\mu_\beta, C_\theta)\,\mathrm{pr}(\beta,\theta)}{\int \mathrm{N}(Z|\mu_\beta, C_\theta)\,\mathrm{pr}(\beta,\theta)\,\mathrm{d}(\beta,\theta)}.$$

The parametric variogram can be estimated via

$$\hat{\gamma}(h) = \mathrm{E}_{\mathrm{pr}(\theta|Z)}(\gamma(h;\theta)) = \int \gamma(h;\theta)\,\mathrm{pr}(\theta|Z)\,\mathrm{d}\theta$$

14. The (traditional) Kriging paradigm

Note 129. "Kriging" (UK) is a general technique for deriving an estimator / predictor of $Z(\cdot)$ (or a function of it) at a location (such as a spatial point s_0, or a block of points $\{s_j^*\}$ or a subregion $\mathcal{S}_0$) of a spatial region $\mathcal{S}$ by properly averaging out data in the neighborhood around the location of interest.

14.1. **Universal Kriging.**

Note 130. Consider we have specified the statistical model as a stochastic process $(Z_s)_{s\in\mathcal{S}}$ with

$$Z(s) = \mu(s) + \delta(s) \tag{14.1}$$

where $\mu(s)$ is a deterministic linear expansion of known basis functions $\{\psi_j(\cdot)\}_{j=0}^p$ and unknown coefficients $\{\beta_j\}_{j=0}^p$ such as

$$\mu(s) = \sum_{j=0}^{p} \psi_j(s)\,\beta_j = (\psi(s))^\top \beta$$

with $\beta = (\beta_0, ..., \beta_p)^\top$ and $\psi(s) = (\psi_0(s), ..., \psi_p(s))^\top$. Also , $\delta(s)$ is a zero mean process, and for this derivation, assume that $\delta(s)$ is an intrinsic stationary sprocess with a (presumably known) semi-variogram $\gamma(\cdot)$[3]

Note 131. Consider there is available a dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $Z_i := Z(s_i)$ being a realization of $(Z_s)_{s\in\mathcal{S}}$ at site s_i. Then one can consider matrix form for (14.1) as

$$Z = \mu + \delta = \Psi\beta + \delta$$

[3]As mentioned in Note 144, stationarity and hence existence of the semivariogram are not necessary in general.

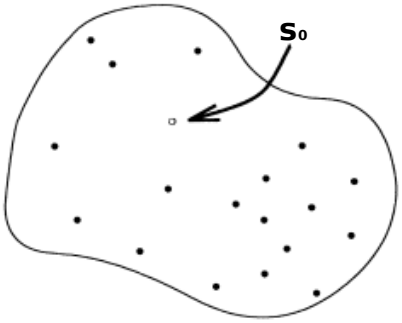


FIGURE 14.1. Kriging area

with vector $Z=\left(Z\left(s_1\right),...,Z\left(s_n\right)\right)^\top$ vector $\delta=\left(\delta\left(s_1\right),...,\delta\left(s_n\right)\right)^\top$, vector $\mu=\left(\mu\left(s_1\right),...,\mu\left(s_n\right)\right)^\top$, and (design) matrix Ψ with $\left[\Psi\right]_{i,j}=\psi_j\left(s_i\right)$.

Note 132. We are interested in learning/predicting $Z\left(s_0\right)$ at an unseen spatial location s_0 (Fig 14.1) .

Note 133. "Universal Kriging" (UK) is the technique for producing a BLUE predictor for $Z_0:=Z\left(s_0\right)$ at spatial location $s_0\in\mathcal{S}$ by using data in the neighborhood of the location of interest.

Note 134. The Universal Kriging (UK) predictor $Z_{\text{UK}}\left(s_0\right)$ of $Z\left(s_0\right)$ at location $s_0\in\mathcal{S}$ is the Best Linear Unbiased Estimator (BLUE) of $Z\left(s_0\right)$ given the data $\left\{\left(s_i,Z_i\right)\right\}_{i=1}^n$.

Note 135. The UK predictor $Z_{\text{UK}}\left(s_0\right)$ of $Z\left(s_0\right)$ at s_0 has the following linear form weighted by a set of tunable unknown weights $\left\{w_i\right\}$

$$\begin{aligned} Z_{\text{UK}}\left(s_0\right) =& w_{n+1}+\sum_{i=1}^n w_i Z\left(s_i\right) \\ =& w_{n+1}+w^\top Z \end{aligned} \tag{14.2}$$

where $Z=\left(Z_1,...,Z_n\right)^\top$ and $w=\left(w_1,...,w_n\right)^\top$.

Note 136. For (14.2), to satisfy unbiasness (that is zero systematic error"), we get

$$\begin{aligned} &\mathrm{E}\left(Z_{\text{UK}}\left(s_0\right)\right)=w_{n+1}+\sum_{i=1}^n w_i\mathrm{E}\left(Z\left(s_i\right)\right)\Leftrightarrow\mathrm{E}\left(Z_{\text{UK}}\left(s_0\right)\right)=w_{n+1}+\sum_{i=1}^n w_i\mathrm{E}\left(Z\left(s_i\right)\right) \\ &\Leftrightarrow\mu\left(s_0\right)=w_{n+1}+\sum_{i=1}^n w_i\mu\left(s_i\right)\Leftrightarrow\left(\psi\left(s_0\right)\right)^\top\beta=w_{n+1}+\sum_{i=1}^n w_i\left(\psi\left(s_i\right)\right)^\top\beta \\ &\Leftrightarrow\Psi_0\beta=w_{n+1}+w^\top\Psi\beta \end{aligned} \tag{14.3}$$

where matrix Ψ with $\left[\Psi\right]_{i,j}=\psi_j\left(s_i\right)$ and (column) vector Ψ_0 with $\left[\Psi_0\right]_j=\psi_j\left(s_0\right)$. Because in (14.3) both sides are polynomial w.r.t β all coefficients must be equal; hence sufficient

conditions for unbiaseness are $w_{n+1} = 0$ and

I write it in matrix and form FYI

$$\text{ASSUMPTION:} \quad \Psi_0 = \sum_{i=1}^{n} w_i \psi(s_i) \Leftrightarrow \psi_0 = w^\top \Psi \tag{14.4}$$

Note 137. The MSE of $Z_{\text{UK}}(s_0)$, given the Assumption (14.4) is

(14.5)

$$\begin{aligned}
\text{MSE}(Z_{\text{UK}}(s_0)) =& \text{E}(Z_{\text{UK}}(s_0) - Z(s_0))^2 \\
=& \text{E}\left(\Psi\beta + \delta(s_0) - w^\top \Psi\beta - w^\top\delta\right)^2; \qquad \left\{\text{let } \delta = (\delta(s_1), ..., \delta(s_n))^\top\right\} \\
=& \text{E}\left(\sum_{i=1}^{n} w_i\delta(s_i) - \delta(s_0)\right)^2 \overset{w_0=-1}{=} \text{E}\left(\sum_{i=0}^{n} w_i\delta(s_i)\right)^2 \qquad (14.6) \\
=& -\text{E}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (\delta(s_i) - \delta(s_j))^2 - 2\sum_{i=1}^{n} w_i (\delta(s_i) - \delta(s_0))^2\right) \qquad (14.7) \\
=& -\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j \frac{1}{2}\text{E}(\delta(s_i) - \delta(s_j))^2 + 2\sum_{i=1}^{n} w_i \frac{1}{2}\text{E}(\delta(s_i) - \delta(s_0))^2 \qquad (14.8)
\end{aligned}$$

Note 138. Now, since we have assumed that (δ_s) is intrinsic stationary, we can exprese it w.r.t. the the semivariogram as

$$\begin{aligned}
\text{E}\left(\sum_{i=1}^{n} w_i\delta(s_i) - \delta(s_0)\right)^2 =& -\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j\gamma(s_i - s_j) + 2\sum_{i=1}^{n} w_i\gamma(s_i - s_0) \qquad (14.9) \\
=& -w^\top \mathbf{\Gamma} w + 2w^\top \boldsymbol{\gamma}_0 \quad = \text{MSE}(Z_{\text{OK}}(s_0))
\end{aligned}$$

where $w = (w_1, ..., w_n)^\top$, $\boldsymbol{\gamma}_0 = (\gamma(s_0 - s_i), ..., \gamma(s_0 - s_n))^\top$, and $[\mathbf{\Gamma}]_{i,j} = \gamma(s_i - s_j)$.

Note 139. The Lagrange function for minimizing the MSE (14.9) under (14.3) is

$$\begin{aligned}
\mathfrak{L}(w, \lambda) =& -\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(s_i - s_j) + 2\sum_{i=1}^{n} w_i\gamma(s_0 - s_i) - \sum_{j=0}^{p} \lambda_j \left(\sum_{i=1}^{n} w_i\psi_j(s_i) - \Psi_{0,j}\right) \\
=& -w^\top \mathbf{\Gamma} w + 2w^\top \boldsymbol{\gamma}_0 - \left(w^\top \Psi - \Psi_0\right)\lambda
\end{aligned}$$

Note 140. The UK system of equations is

$$0 = \nabla_{(w,\lambda)} \mathfrak{L}(w,\lambda)\big|_{(w_{\text{UK}},\lambda_{\text{UK}})} \iff$$

(14.10)

$$\begin{cases} 0 = -2\sum_{j=1}^{n} w_{\text{UK},j}\gamma(s_i - s_j) + 2\gamma(s_0 - s_i) - \sum_{j=0}^{p} \lambda_{\text{UK},j}\psi_j(s_i), & i = 1,...,n \\ \psi_j(s_0) = \sum_{i=1}^{n} w_{\text{UK},i}\psi_j(s_i), & j = 0,...,p \end{cases} \iff$$

(14.11)

$$\begin{cases} 0 = -2\boldsymbol{\Gamma} w + 2\boldsymbol{\gamma}_0 - \Psi\lambda_{\text{UK}} \\ \Psi_0 = w_{\text{UK}}^{\top}\Psi \end{cases}$$

Then by multipling both sides by $\Psi^{\top}\Gamma^{-1}$ I get

$$0 = -2\Psi^{\top}\Gamma^{-1}\boldsymbol{\Gamma} w_{\text{UK}} + 2\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi^{\top}\Gamma^{-1}\Psi\lambda_{\text{UK}} \iff$$

$$\lambda_{\text{UK}} = 2\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right) \tag{14.12}$$

and then by substituting (14.12) in (14.10), I get the UK weights as

$$w_{\text{UK}} = \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\gamma}_0 - \Psi\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)\right) \tag{14.13}$$

Note 141. Hence the UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

$$Z_{\text{UK}}(s_0) = \left(\boldsymbol{\gamma}_0 - \Psi\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)\right)^{\top}\boldsymbol{\Gamma}^{-1}Z \tag{14.14}$$

with standard error

$$\sigma_{\text{UK}}(s_0) = \sqrt{-w_{\text{UK}}^{\top}\boldsymbol{\Gamma} w_{\text{UK}} + 2w_{\text{UK}}^{\top}\boldsymbol{\gamma}_0} \tag{14.15}$$

$$= \sqrt{\boldsymbol{\gamma}_0\Gamma^{-1}\boldsymbol{\gamma}_0 - \left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)^{\top}\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)} \tag{14.16}$$

Note 142. $(1-\alpha)\,100\%$ Prediction interval of UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

$$\left(Z_{\text{UK}}(s_0) - q_{\alpha/2}\sqrt{\sigma_{\text{UK}}^2(s_0)},\ Z_{\text{UK}}(s_0) + q_{1-\alpha/2}\sqrt{\sigma_{\text{UK}}^2(s_0)}\right) \tag{14.17}$$

where $q_{\cdot}$ are suitable quantiles of the distribution of Z_s . E.g. if $Z_s \sim \text{GP}(\mu(\cdot), c(\cdot,\cdot))$ then $q_{0.05/2} = -1.96$ and $q_{0.95/2} = 1.96$ at $\alpha = 0.05$.

Note 143. Note that we have not assumed a particular distribution of Z_s or δ_s, but only stationarity assumptions.

Note 144. It was not necessary to consider the stationarity assumption in order to derive the Universal Kriging predictor; we could have derived its formulas (14.14) & (14.15) with respect to the covariance function $c(\cdot,\cdot)$ of (Z_s) instead of its semivariogram $\gamma(\cdot)$. Here,

intrinsic stationarity was assumed for practical reasons, i.e. we have already discussed how to estimate the semi-variogram in Sec 13.

Note 145. To use (14.14), (14.15), and (14.17), we need to learn the unknown coefficients $\{\beta_j\}$ and the semi-variogram $\gamma(\cdot)$, or "equivalently" the unknown hyper-parameter θ of the parametric semivariogram $\gamma_\theta(\cdot)$ used to cast $\gamma(\cdot)$. In practice, we use the same dataset used to compute (14.13), however in principle a fresh training dataset $\{(s_i', Z_i')\}_{i=1}^n$ is required (never use the same training data 2 times). A training procedure can be the following.

(1) Compute estimates $\hat{\beta}$ via LSE (or equivalent)

$$\hat{\beta}_{\text{LSE}} = \arg\min_{\beta} \left(\sum_i \left(Z(s_i) - \underbrace{\psi(s_i)^\top \beta}_{=\mu(s_i)} \right)^2 \right) \tag{14.18}$$

(2) Compute the residuals

$$\hat{\delta}_i := Z(s_i) - \psi(s_i)^\top \hat{\beta}_{\text{LSE}} \tag{14.19}$$

(3) Compute the empirical variogram $\hat{\gamma}$ for $\hat{\delta}$ on $\mathscr{H}$ according to Prop 102,

(4) Compute the estimate $\hat{\theta}$ of θ of the parameterized semivariogram γ_θ, according to Prop 117, and hence compute $\gamma_{\hat{\theta}}(\cdot)$.

Example 146. [4] Consider the example with the Meuse dataset. Fig 14.2b presents the UK prediction $Z_{\text{UK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (14.1) for when the spatial mean has a linear form $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$. Following Note 145, we computed the $\hat{\beta}_{\text{LSE}}$ of β by (14.18), then we removed the linear trend by 14.19 and computed the residual process $\left\{\hat{\delta}_i\right\}$, then we computed the semi-variogram $\hat{\gamma}$ (13.2) of δ as in Prop 102; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ of δ where we computed the OLS $\hat{\theta}_{\text{OLS}} = \left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (13.7) (see Fig. 13.3d); and then we pluged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)}$ in (14.14) to compute the UK weights w_{UK} for the UK predictor $Z_{\text{UK}}(s_0) = w_{\text{UK}} Z$ for any $s_0 \in \mathcal{S}$. The reason that we do not see much difference between OK in Fig 14.2a and UK in Fig 14.2b is reather because the slops int eh linear trend (mean) of UK are rather small and insignificant (See Example 124).

Example 147. Consider the example with the Meuse dataset. The dataset has another measurement (a potential regressor in the deterministic mean $\mu(s)$), the "distance to the Meuse river bed" $\{d_i\}$ at the associated locations $\{s_i\}$, let's denote it by d. Fig 14.2c shows a rather linear relationship between Z and $\sqrt{d}$, hence we can consider a UK predictor with

[4] https://github.com/georgios-stats/Spatio-Temporal_Statistics_Michaelmas_2023/blob/main/Lecture_handouts/R_scripts/03.Geostatistical_data_meuse_gstats.R

deterministic mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d_s}$. We follow the same procedure as in Example 146 and we get the UK predictor in Figure 14.2d.

14.2. **Ordinary Kriging.**

Note 148. Ordinary Kriging (OK) addresses spatial prediction in cases that the specified statistical model on $(Z_s)_{s\in\mathcal{S}}$ has the form

$$Z(s) = \beta_0 + \delta(s) \tag{14.20}$$

with unknown $\beta_0 \neq 0$ and intrinsically stationary process (δ_s). OK can be derived as a special case of the Universal Kriging by setting $p = 0$ and constant spatial mean $\mu(s) = \beta_0$.

Example 149. The derivation is in (Exercise 19 Exercise sheet). As a supplementary and for demonstration, we mention that the OK assumption is $\sum_{i=1}^{n} w_i = 1$; the OK system of equations is $0 = \nabla_{(\{w_i\},\lambda)} L(w,\lambda)\big|_{(w,\lambda)}$ producing

$$\begin{cases} 0 = -2\boldsymbol{\Gamma} w_{\text{OK}} + 2\boldsymbol{\gamma}_0 - 1\lambda \\ w_{\text{OK}}^{\top} 1 = 1 \end{cases} \tag{14.21}$$

the weights are

$$w_{\text{OK}} = \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\gamma}_0 + \frac{1 - 1^{\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{1^{\top}\boldsymbol{\Gamma}^{-1}1} 1\right) \tag{14.22}$$

the Kriging standard error of $Z_{\text{OK}}(s_0)$ at s_0 is

$$\sigma^2_{\text{OK}}(s_0) = \boldsymbol{\gamma}_0^{\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \frac{\left(1 - 1^{\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0\right)^2}{1^{\top}\boldsymbol{\Gamma}^{-1}1}. \tag{14.23}$$

14.3. **Simple Kriging.**

Note 150. Simple Kriging (SK) addresses spatial prediction in cases that the specified statistical model on $(Z_s)_{s\in\mathcal{S}}$ has the form

$$Z(s) = \mu(s) + \delta(s) \tag{14.24}$$

where the deterministic mean $\mu(s)$ is known, and (δ_s) is a weakly stationary process with covariogram $c(\cdot)$.

Example 151. The derivation is in (Exercise 17 Exercise sheet). It does not require any assumption in the weights such as 14.4 or (14.21).As a supplementary and for demonstration, we mention the SK predictor at s_0 and standard error:

$$Z_{\text{SK}}(s_0) = \mu(s_0) + C_0^{\top} C^{-1}[Z - \mu]$$

$$\sigma_{\text{SK}} = \sqrt{c(s_0,s_0) - C_0^{\top} C^{-1} C_0}$$

with $\mu = (\mu(s_1), ..., \mu(s_n))^\top$, $C_0 = (c(s_0 - s_i), ..., \gamma(s_0 - s_n))^\top$, and $[C]_{i,j} = c(s_i - s_j)$.

Example 152. Consider the example with the Meuse dataset. Fig 14.2a presents the OK prediction $Z_{\text{OK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (14.20) that is the UK case (14.1) for when $\mu(s) = \beta_0$. First we computed the non-parametric semivariogram $\hat{\gamma}$ (13.2) as in Prop 102; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ where we computed the OLS $\hat{\theta}_{\text{OLS}} = \left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (13.7) (see Fig. 13.3a); and then we plugged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)}$ in (14.22) to compute the OK weights w_{OK} for the OK predictor $Z_{\text{OK}}(s_0) = w_{\text{OK}} Z$ for any $s_0 \in \mathcal{S}$.

15. The Bayesian Kriging paradigm (Hierarchical modeling)

15.1. **A general framework (The hierarchical modeling).**

Note 153. Consider the geostatistical model of (Z_s) with a scale decomposition such as in (12.3)

$$Z(s) = Y(s) + \varepsilon(s), \quad s \in \mathcal{S} \tag{15.1}$$

where (Y_s) is a stochastic process, and (ε_s) is a nugget process. (Z_s) may be labeled by parameters $\vartheta \in \Theta$ when (Y_s) and (ε_s) are parameterized as probabilistic models.

Note 154. Consider a dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $Z_i = Z(s_i)$ being a realization of (15.1) at site $s_i \in \mathcal{S}$. Let $Z = (Z_1, ..., Z_n)^\top$, and $Y = (Y_1, ..., Y_n)^\top$.

Recall

Note 155. Uncertainty can be decomposed according to the Hierarchical spatial model

$$\begin{cases} Z|Y, \vartheta & \text{data model} \\ Y|\vartheta & \text{spatial process model} \end{cases} \tag{15.2}$$

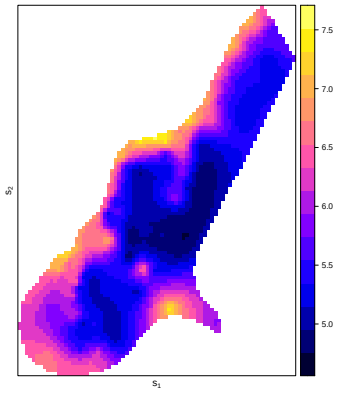
with

$$\text{pr}(Z, Y|\vartheta) = \text{pr}(Z|Y, \vartheta)\,\text{pr}(Y|\vartheta)$$

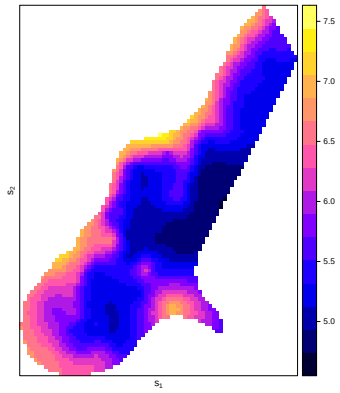
Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\text{pr}(Y|\vartheta)$ possibly labeled by some parameter ϑ.

Data model: expresses the measurement uncertainty (e.g., that coming from (ε_s)) as it is quantified via the distribution $\text{pr}(Z|Y, \vartheta)$ possibly labeled by some parameter ϑ.

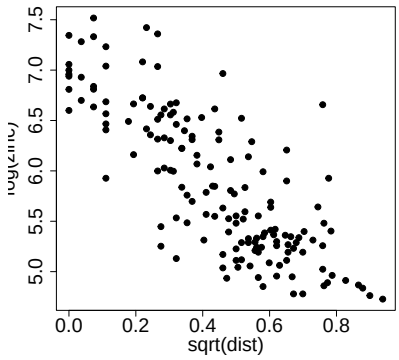
Note 156. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \text{pr}(\vartheta_1|\vartheta_2)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior; it can be considered sometimes as known and sometimes as unknown in what follows. (!)



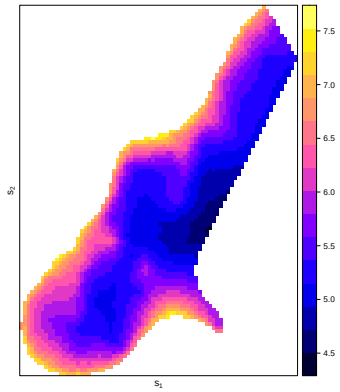
(A) Ordinary Kriging, constant mean, exponential covariogram



(B) Universial Kriging, mean $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$, exponential covariogram



(C) Scatter plot of Z vs $\log(d)$ where d is a third variable representing the distance to the riven bed.



(D) Universial Kriging, mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d}$, exponential covariogram

FIGURE 14.2. Kriging Meuse dataset.

Note 157. Then the Bayesian spatial hierarchical model becomes

$$\begin{cases} Z|Y,\vartheta_1,\vartheta_2 & \text{data model} \\ Y|\vartheta_1,\vartheta_2 & \text{spatial process model} \\ \vartheta_1|\vartheta_2 & \text{hyper-parameter prior model} \end{cases} \tag{15.3}$$

where uncertainty is described by

$$\mathrm{pr}\left(Z, Y, \vartheta_1|\vartheta_2\right) = \mathrm{pr}\left(Z|Y, \vartheta_1|\vartheta_2\right)\mathrm{pr}\left(Y|\vartheta_1, \vartheta_2\right)\mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$$

Note 158. Under Bayesian model (15.3), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

$$\hat{\vartheta}_2 = \arg\min_{\vartheta_2}\left(-2\log\left(\mathrm{pr}\left(Z|\vartheta_2\right)\right)\right)$$

by maximizing the marginal likelihood

$$\mathrm{pr}\left(Z|\vartheta_2\right) = \int \mathrm{pr}\left(Z, Y, \vartheta_1|\vartheta_2\right)\mathrm{d}Y\mathrm{d}\vartheta_1$$

Under Bayesian model (15.3), when ϑ_1 is considered as unknown (but random), namely, the a prior $\vartheta_1 \sim \mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$ has been specified, uncertainty about unknown ϑ_1 given Y and ϑ_2 can be represented by the posterior distribution

$$\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \frac{\mathrm{pr}\left(Z|\vartheta_1, \vartheta_2\right)\mathrm{pr}\left(\vartheta_1|\vartheta_2 = \hat{\vartheta}_2\right)}{\mathrm{pr}\left(Z|\vartheta_2 = \hat{\vartheta}_2\right)}$$

where the value $\hat{\vartheta}_2$ is plugged in.

Note 159. General interest lies in computing the posterior predictive distributions of the spatial process model (Y_s), (or latent process, or noiseless process) given the data Z

$$\mathrm{pr}\left(Y\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Y\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}\vartheta_1$$

and / or the marginal process (Z_s) given the data

$$\mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}\vartheta_1$$

$$\mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Z\left(s_0\right), Y\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}Y\left(s_0\right)$$

for any $s_0 \in \mathcal{S}$.

Note 160. The above statistical problem is naturally addressed in the (either full or empirical) Bayesian statistical framework. It is often called Bayesian Kriging.

15.2. **Bayesian Kriging (Gaussian process regression).**

Inventory of useful formulas.

Fact 161. *Let $X \sim N(\mu_X, \Sigma_X)$ $Y \sim N(\mu_Y, \Sigma_Y)$ and Y, X independent. Let fixed matrices A and B and vector c of appropriate sizes. Then*

$$AX + BY + c \sim N\left(A\mu_X + B\mu_Y + c, A\Sigma_X A^\top + B\Sigma_Y B^\top\right) \tag{15.4}$$

Fact 162. *Let $N(\beta|b, B)$ be the Gaussian pdf with mean b and covariance B at β. It is*

$$\int N(Z|\Psi\beta, C)\, N(\beta|b, B)\, d\beta = N\left(Z|\Psi b, C + \Psi B \Psi^\top\right)$$

Fact 163. *[Marginalization & conditioning] Let $x_1 \in \mathbb{R}^{d_1}$, and $x_2 \in \mathbb{R}^{d_2}$. If*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathrm{N}_{d_1+d_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix}\right)$$

then it is

$$x_2|x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, \Sigma_{2|1}\right)$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_1 - \mu_1) \text{ and } \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top$$

Note 164. To demonstrate how to work in the "Bayesian Kriging" framework e.g., with the spatial hierarchical models (15.2) and (15.3), we are going through a particular example of the Bayesian Gaussian process regression (or Bayesian Kriging).

A possible narrative - a story.

Note 165. Assume there is available a dataset $\{(s_i, Z_i)\}_{i=1}^n$ where $Z_i = Z(s_i)$ is a realization of a stochastic process (Z_s) with $\{Z_i \in \mathbb{R}\}$. In particular, assume that data are instances of an unknown function $Y(\cdot)$ at s_i but contaminated by additive random noise $\{\varepsilon_i \sim \mathrm{N}(0, \tau^2); i = 1, ..., n\}$ with scale $\tau > 0$; i.e. $Z_i = Y(s_i) + \varepsilon_i$.

Note 166. Assume we are interested in recovering $Z(\cdot)$.

Specifying the hierarchical model.

Note 167. A natural model to cast this problem is the geostatistical model

$$Z(s) = Y(s) + \varepsilon(s), \quad s \in \mathcal{S} \tag{15.5}$$

- we specify a zero-mean Gaussian process $\varepsilon(\cdot) \sim \mathrm{GP}(0, c_\varepsilon(\cdot, \cdot|\tau))$ with nugget covariogram $c_\varepsilon(s, s'|\tau) = \tau^2 1_{\{0\}}(\|s - s'\|)$ to represent the noise. Hence

$$Z(\cdot)|Y(\cdot), \tau \sim \mathrm{GP}(Y(\cdot), c_\varepsilon(\cdot, \cdot|\tau)). \tag{15.6}$$

- To quantify uncertainty of the unknown $Y(\cdot)$, we specify a GP prior on $Y(\cdot)$

$$Y(\cdot)|\beta,\theta \sim \text{GP}\left(\mu(\cdot|\beta), c_Y\left(\cdot,\cdot|\sigma^2,\phi\right)\right) \tag{15.7}$$

  with mean function $\mu(\cdot|\beta)$ labeled by unknown parameter β and covariance function $c_Y(\cdot,\cdot|\sigma^2,\phi)$, labeled by unknown parameter $(\sigma^2,\phi)^\top$.
- we assume ε_s and Y_s to be independent.

Note 168. Given (15.6) and (15.7), the Bayesian model (15.2) is

$$\begin{cases} Z_i|Y_i,\tau^2 \overset{\text{ind}}{\sim} \text{N}(Y_i,\tau^2),\ i=1,...,n & \text{data model} \\ Y|\beta,\sigma^2,\phi \sim \text{N}(\mu(S|\beta), c_Y(S,S|\sigma^2,\phi)) & \text{spatial process model} \end{cases} \tag{15.8}$$

where $\vartheta = (\beta,\sigma^2,\phi)^\top$, $[\mu(S|\beta)]_i = \mu(s_i|\beta)$, and $[c_Y(S,S|\sigma^2,\phi)]_{i,j} = c_Y(s_i,s_j|\sigma^2,\phi)$.

Computing the marginal process $Z(\cdot)|\beta,\theta$.

Note 169. The marginal process (Z_s) given parameters β, $\theta = (\sigma^2,\phi,\tau)^\top$ (in (15.8)) is

$$Z(\cdot)|\beta,\theta \sim \text{GP}(\mu(\cdot|\beta), c(\cdot,\cdot|\theta)) \tag{15.9}$$

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, and covariance function parameters $\theta = (\sigma^2,\phi,\tau)^\top$. *[We used the additive property of Gaussian random variables in Fact* 161 *].*

Computing the predictive distribution $Z(\cdot)|Z,\beta,\theta$.

Note 170. Assume a vector of "unseen" sites $S_* = (s_{*,1},...,s_{*,q})^\top$ for any $q \in \mathbb{N}_0$. Let convenient notation $Z := Z(S)$, and $Z_* := Z(S_*)$. The join marginal distribution of $(Z_*,Z)^\top$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is

$$\begin{pmatrix} Z_* \\ Z \end{pmatrix}|\beta,\theta \sim \text{N}\left(\begin{pmatrix} \mu(S_*;\beta) \\ \mu(S;\beta) \end{pmatrix}, \begin{pmatrix} C(S_*,S_*|\theta) & (C(S_*,S|\theta))^\top \\ C(S_*,S|\theta) & C(S,S|\theta) \end{pmatrix}\right)$$

by using convenient notation $[C(S_*,S|\theta)]_{i,j} = s(s_{*,i},s_j|\theta)$ and $[\mu(S;\beta)]_i = \mu(s_i;\beta)$.

Note 171. Given that vector Z is observed/known, the (posterior) predictive distribution of $Z_*|Z$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is the conditional distribution

$$Z_*|Z,\beta,\theta \sim \text{N}(\mu_*(S_*|\beta,\theta), C_*(S_*,S_*|\theta)) \tag{15.10}$$

where

$$\begin{aligned} C_*(S_*,S_*|\theta) =& C(S_*,S_*|\theta) + (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} C(S,S_*|\theta) \\ \mu_*(S_*|\beta,\theta) =& \mu(S_*|\beta) - (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} (\mu(S|\beta) - Z) \end{aligned}$$

[We used the formula for computing the conditional Gaussian distribution in Fact 163].

Note 172. Since the derivation of (15.10) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (15.10) can be extended to a Gaussian Process

$$Z(\cdot)|Z,\beta,\theta \sim \mathrm{GP}(\mu_1(\cdot|\beta,\theta), c_1(\cdot,\cdot|\theta)) \tag{15.11}$$

with

$$\begin{aligned} c_1(s,s'|\theta) =& c(s,s|\theta) + (C(S,s|\theta))^\top (C(S,S|\theta))^{-1} C(S,s'|\theta) \\ \mu_1(s|\beta,\theta) =& \mu(s|\beta) - (C(S,s|\theta))^\top (C(S,S|\theta))^{-1} (\mu(S|\beta) - Z) \end{aligned}$$

for any $s, s' \in \mathcal{S}$. This is the predictive process of $Z(s)$ at any $s \in \mathcal{S}$ given Z, β, θ. *[Here we used the definition of GP (Def 14) given Note* 171*].*

Note 173. Assume that the parameters (β, θ) are unknown but fixed (i.e. no prior is specified). Training can be performed by maximizing the marginal likelihood of Z given β, θ

$$\mathrm{pr}(Z|\beta,\theta) = \mathrm{N}(Z|\mu(S|\beta), C(S,S|\theta)) \tag{15.12}$$

derived from (15.9) by solving

$$\left(\hat{\beta},\hat{\theta}\right)^\top = \arg\min_{\beta,\theta} \left(-2\log\left(\mathrm{N}(Z|\mu(S|\beta), C(S,S|\theta))\right)\right)$$

Note 174. The estimated "Kriking predictor" results by plugging $\left(\hat{\beta},\hat{\theta}\right)^\top$ in (15.11), as

$$Z(\cdot)|Z,\hat{\beta},\hat{\theta} \sim \mathrm{GP}\left(\mu_1\left(\cdot|\hat{\beta},\hat{\theta}\right), c_1\left(\cdot,\cdot|\hat{\theta}\right)\right).$$

Computing the predictive distribution $Z(\cdot)|Z,\theta$.

Note 175. Assume β is an unknown random hyper-parameter in the sense we assign a prior distribution on it to account for uncertainty. Hence, we will specify a conjugate distribution on β, and compute the produced predictive distribution.

Note 176. Like in Universal Kriging, assume that the spatial mean is parameterized as an expansion of basis functions $\psi(\cdot) = (\psi_1(\cdot), ..., \psi_p(\cdot))^\top$ with unknown coefficients β, i.e.

$$\mu(s|\beta) = \psi(s)^\top \beta$$

Note 177. The marginal process (Z_s) given parameters β, and θ can be re-writen as

$$Z(\cdot)|\beta,\theta \sim \mathrm{GP}\left(\psi(s)^\top \beta, c(\cdot,\cdot|\theta)\right)$$

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, $\theta = (\sigma^2,\phi,\tau)^\top$

Note 178. We specify a conjugate prior $\beta|\sigma^2 \sim \mathrm{N}(b, \sigma^2 B)$ on β, for some user-specified fixed hyper-parameters b and $B > 0$.

Note 179. The marginal Bayesian model is now extended to

$$\begin{cases} Z|\beta,\theta \sim \mathrm{N}\left(\Psi\beta, C\left(S,S|\theta\right)\right) \\ \beta|\sigma^2 \sim \mathrm{N}\left(b,\sigma^2 B\right) \end{cases} \tag{15.13}$$

with matrix Ψ such as $\left[\Psi\right]_{i,j} = \psi_j\left(s_i\right)$.

Note 180. The posterior of β given data Z and θ is computed via the Bayes theorem

$$\begin{aligned} \mathrm{pr}\left(\beta|Z,\theta\right) &\propto \mathrm{pr}\left(Z|\beta,\theta\right)\mathrm{pr}\left(\beta|\theta\right) \\ &\propto \mathrm{N}\left(Z|\Psi\beta, C\left(S,S|\theta\right)\right)\mathrm{N}\left(\beta|b,\sigma^2 B\right) \end{aligned}$$

and results as

$$\beta|Z,\theta \sim \mathrm{N}\left(b_n,\sigma^2 B_n\right) \tag{15.14}$$

with

$$\begin{aligned} B_n =& \left(B^{-1} + \Psi^\top\left(C\left(S,S|\theta\right)/\sigma^2\right)^{-1}\Psi\right)^{-1} \\ b_n =& B_n\left(B^{-1}b + \Psi^\top\left(C\left(S,S|\theta\right)/\sigma^2\right)^{-1}Z\right) \end{aligned}$$

[The derivation is the same as in Bayesian linear regression and will be given as a Homework]

Note 181. The posterior predictive distribution of $Z\left(\cdot\right)$ given the data Z and θ , results by integrating (15.11) with respect to (15.14) i.e.

$$\begin{aligned} \mathrm{pr}\left(Z_*|Z,\theta\right) &= \int \mathrm{pr}\left(Z_*|Z,\beta,\theta\right)\mathrm{pr}\left(\beta|Z,\theta\right)\mathrm{d}\beta \\ &= \int \mathrm{N}\left(Z_*|\mu_*\left(S_*|\beta,\theta\right), C_*\left(S_*,S_*|\theta\right)\right)\mathrm{N}\left(\beta|b_n,\sigma^2 B_n\right)\mathrm{d}\beta \end{aligned}$$

and it is again a GP

$$Z\left(\cdot\right)|Z,\theta \sim \mathrm{GP}\left(\mu_2\left(\cdot|\theta\right), c_2\left(\cdot,\cdot|\theta\right)\right) \tag{15.15}$$

with

$$\begin{aligned} \mu_2\left(s|\theta\right) =& \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1} + \Psi^\top C^{-1}\Psi\right)^{-1}B^{-1}b \\ &+ \left[\left(C\left(s\right)\right)^\top + \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1}/\sigma^2 + \Psi^\top C^{-1}\Psi\right)^{-1}\Psi\right]C^{-1}Z \end{aligned} \tag{15.16}$$

(15.17)

$$\begin{aligned} c_2\left(s,s'|\theta\right) =& c\left(s,s'|\theta\right) - \left(C\left(s\right)\right)^\top C^{-1}C\left(s'\right) \\ &+ \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1}/\sigma^2 + \Psi^\top C^{-1}\Psi\right)^{-1}\left(\Psi C^{-1}\left(C\left(s'\right)\right)^\top - \psi\left(s'\right)\right) \end{aligned}$$

with column vector $C(s) = (c(s, s_1), ..., c(s, s_n))^\top$, and matrix $C = C(S, S|\theta)$. *[The derivation will be given as a Homework]*

Note 182. If we consider non-informative priors in (15.13) such as $\text{pr}(\beta|\sigma^2) \propto 1$, for instance, by allowing $B \to 0$, and $b < \infty$ then (15.17) produces the Universal Kriging predictor (check with (14.14)).

Note 183. Assume that $\theta = (\sigma^2, \phi, \tau)^\top$ is an unknown fixed hyper-parameter without a prior distribution being specified. Training can be performed by maximizing the marginal likelihood of Z given θ

$$\text{pr}(Z|\theta) = \int \text{pr}(Z|\beta, \theta)\, \text{pr}(\beta|\theta)\, \text{d}\beta \tag{15.18}$$

$$= \int \text{pr}(Z|\Psi\beta, C(S, S|\theta))\, \text{N}(\beta|b, B)\, \text{d}\beta \tag{15.19}$$

$$= \text{N}\left(Z|\Psi b, C(S, S|\theta) + \sigma^2 \Psi B \Psi^\top\right) \tag{15.20}$$

[from Fact 162] by computing

$$\hat{\theta} = \arg\min_{\theta} \left(-2 \log\left(\text{N}\left(Z|\Psi b, C(S, S|\theta) + \sigma^2 \Psi B \Psi^\top\right)\right)\right)$$

Note 184. The estimated "Kriging predictor" results by plugging $\hat{\theta}$ in (15.15)

$$Z(\cdot)\,|Z, \hat{\theta} \sim \text{GP}\left(\mu_2\left(\cdot|\hat{\theta}\right), c_2\left(\cdot, \cdot|\hat{\theta}\right)\right) \tag{15.21}$$

Computing the predictive distribution $Z(\cdot)\,|Z, \phi, \tau$.

FYI: we specify a conjugate prior on σ^2 and then we follow the same routine as above... and we can get a Students-T process...