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) 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] 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. 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 $\mathfrak{p}$*

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

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

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

or all $n \in \mathbb{N}$, $\{s_i \in S\}$, and $\{z \in \mathbb{R}\}$, 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

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

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

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 37 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.

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.

Also Example of Proposition

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 β. An popular form of covariance functions (c.f.), for some tunable unknown parameter β,σ^2, are

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

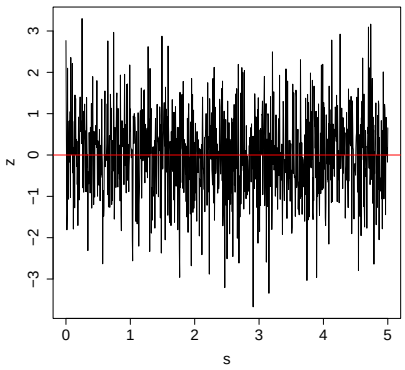
Example 23. Recall your linear regression lessons where you specified the sampling distribution to be $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta,\sigma^2\right)$, $\forall x \in \mathbb{R}^d$. Well that can be considered as a GP $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ 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.

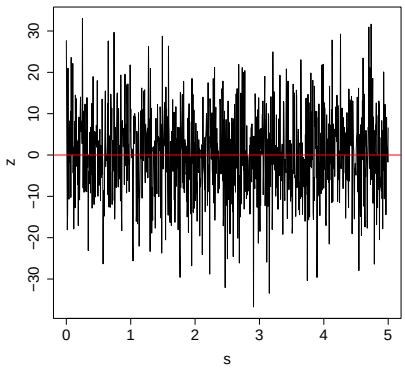
Algorithm 1 R script for simulating from a GP $(Z_s; s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

```
# 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")
```

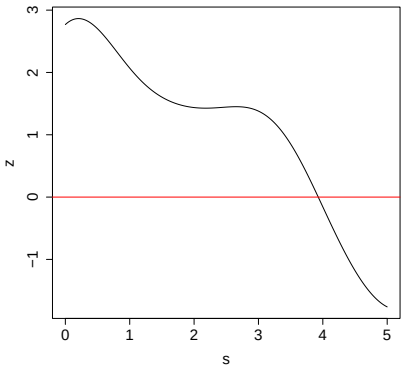
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency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



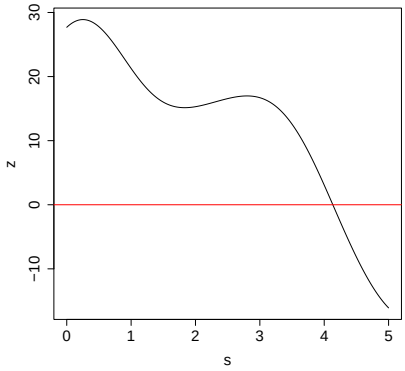
(A) Nugget c.f
($\sigma^2 = 1$)



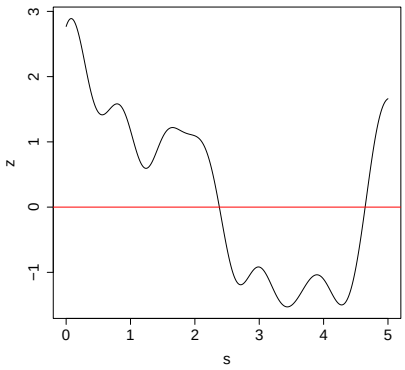
(B) Nugget c.f
($\sigma^2 = 100$)



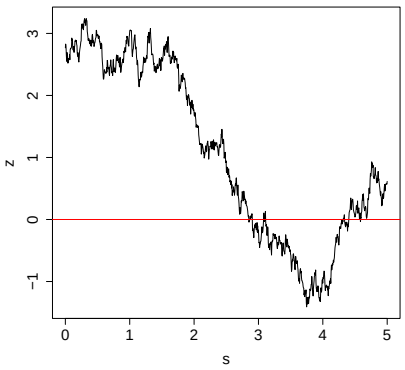
(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



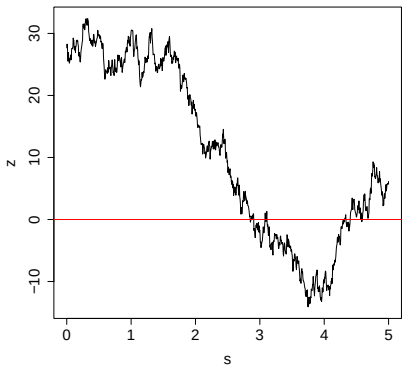
(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



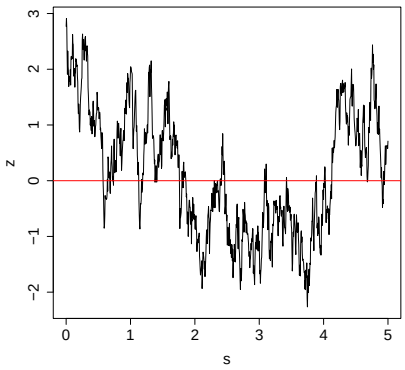
(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)



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

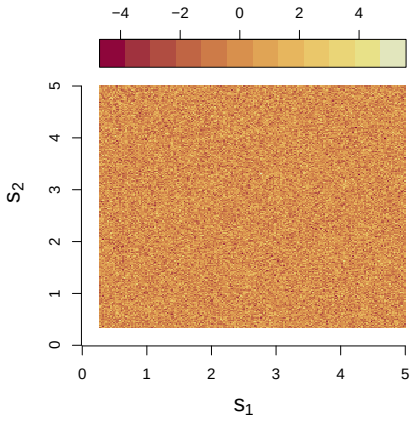


(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)

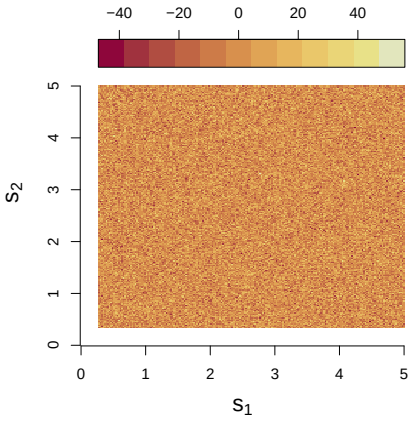


(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

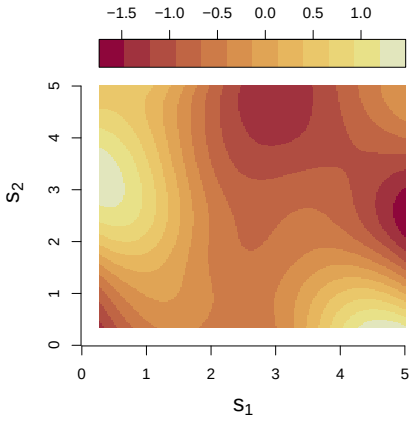
FIGURE 3.1. Realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ when $s \in [0,5]$ (using same seed)



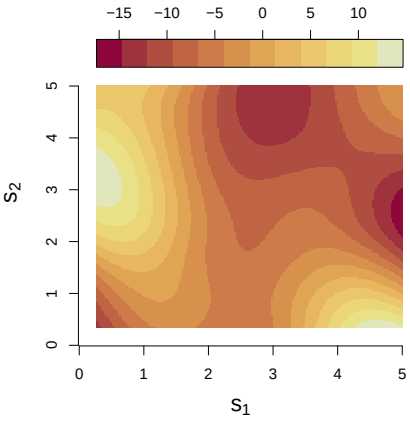
(A) Nugget c.f
($\sigma^2 = 1$)



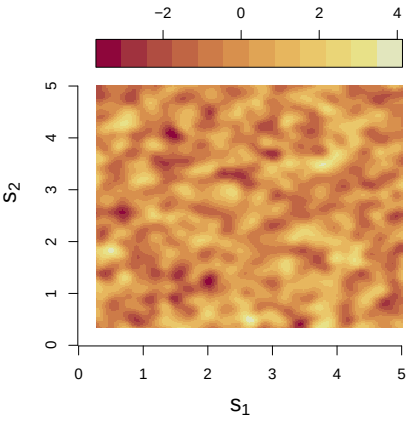
(B) Nugget c.f
($\sigma^2 = 100$)



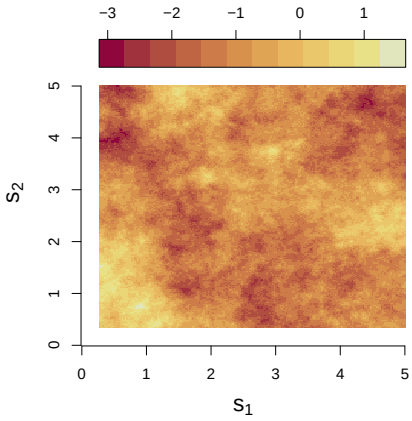
(C) Gauss c.f
($\sigma^2 = 1, \beta = 0.5$)



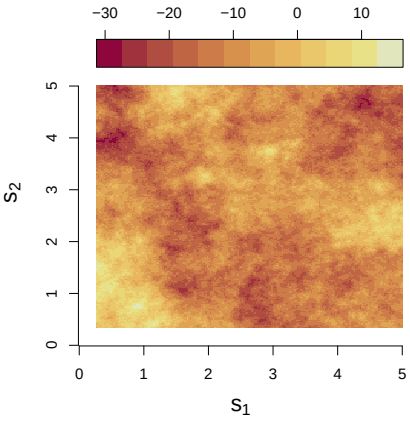
(D) Gauss c.f
($\sigma^2 = 100, \beta = 0.5$)



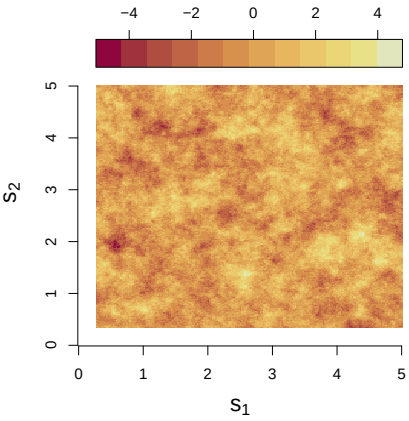
(E) Gauss c.f
($\sigma^2 = 1, \beta = 5$)



(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 $h \in \mathbb{R}^d$

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

5. Weak stationarity (or second order stationarity)

Note 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 $\mathrm{E}\left(Z_s^2\right) < \infty$, 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 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) $\mathrm{E}\left(Z_s^2\right) < \infty$ (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 31. Weakly (or second order) stationary covariance function is called the c.f. of a weakly stationary stochastic process.

6. Covariogram

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

Definition 33. 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 34. For the Gaussian c.f. $c\left(s,t\right) = \sigma^2 \exp\left(-\beta \left\|s-t\right\|_2^2\right)$ in (Ex. 21(2)), we may denote just

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

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

Observe that, in Figures 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.

Proposition 35. *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 36. Given there is some knowledge of the characteristic functions of a suitable distribution, the following Theorem helps in the specification of a suitable covariogram.

Theorem 37. *(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 38. *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 39. 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, by using Theorem 37, the spectral density is

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

i.e. it has a Gaussian form.

7. Intrinsic stationarity

Note 40. Getting greedier, we introduce an even weaker stationarity than the weak stationarity by considering lag dependent variance in the increments of the process with purpose to be able to use more inclusive models; Definition 30 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 41. 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 42. Intrinsically stationary covariance function is called the c.f. of an intrinsically stationary stochastic process.

Example 43. 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 and Variogram

Note 44. 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 45. 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 46. Variogram of an intrinsically stationary random field is called the quantity $2\gamma(h)$.

Remark 47. 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 48. For the Gaussian covariance function (Ex. 34) the semi-variogram is

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

Proposition 49. *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).

Note 50. The variogram $\gamma(h)$ is very informative when plotted against the lag h, below we discuss some of the characteristics of it, using Figure 8.1

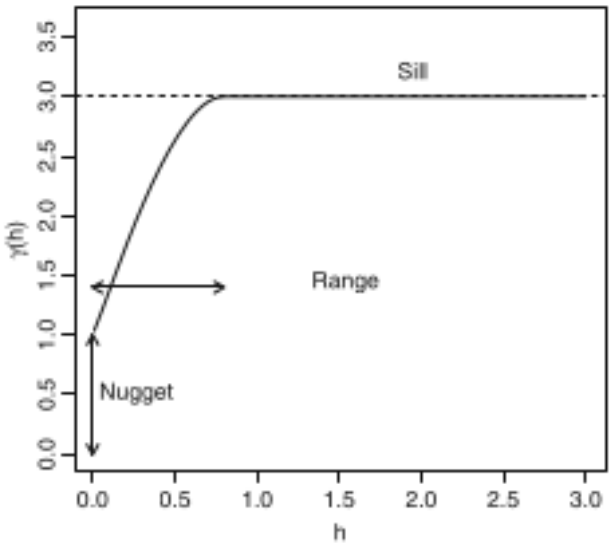


FIGURE 8.1. Semi Variogram's characteristics

Note 51. 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 52. 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 53. Conversely, a non zero slope of the variogram indicates structure.

Nugget Effect.

Note 54. 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 55. 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 56. For a continuous processes $Z = (Z_s)_{s\in\mathbb{R}^d}$, it is expected

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

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

Note 59. For weakly stationary processes the sill is always finite. However, for intrinsic processes, the sill may be infinite.

Partial sill.

Definition 60. 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 61. 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 62. Isotropy as a concept imposes the assumption of "rotation invariance" in the stochastic process.

Note 63. Isotropy applies to both intrinsic stationary and weakly stationary processes.

Definition 64. An intrinsic stochastic process $(Z_s)_{s\in\mathbb{R}^d}$ is isotropic iff

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

Definition 65. 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 66. Isotropic covariance function $C : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called the covariance function satisfying (9.1).

Definition 67. 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. **Popular isotropic covariance functions.**

Note 68. 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.*

Note 69. For $\sigma^2 > 0$,

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

is the nugget-effect covariogram. 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.*

No need to memorize (9.2)

Note 70. For $\sigma^2 > 0$, $\phi > 0$, and $\nu \geq 0$

$$c(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|h\|}{\phi} \right)^{\nu} K_\nu \left(\frac{\|h\|}{\phi} \right) \tag{9.2}$$

is the Matern covariogram. 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.,

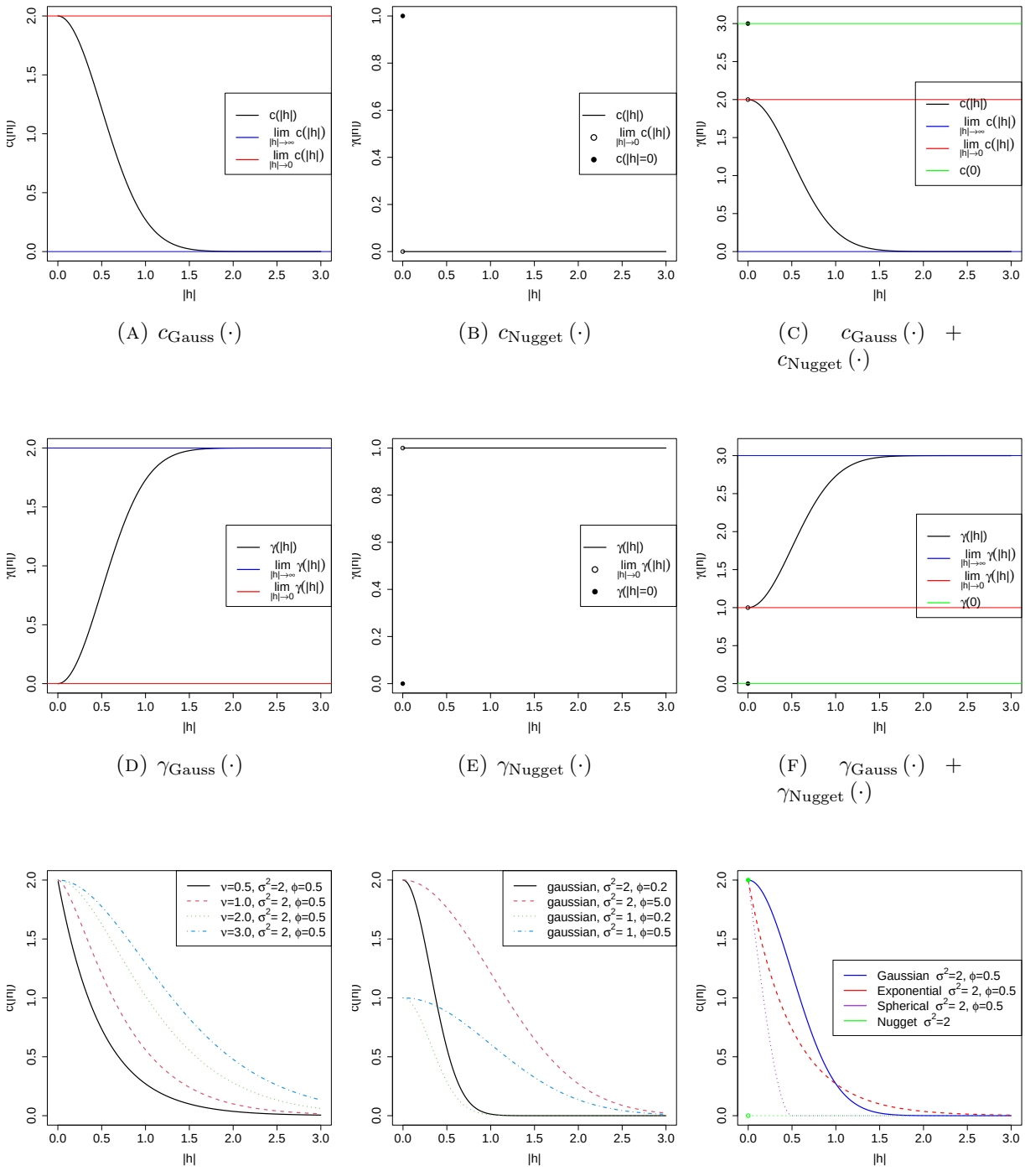
$$c(h) = \sigma^2 \exp\left(-\frac{1}{\phi} \|h\|_1 \right)$$

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.*



(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 $\gamma(\cdot)$

Note 71. [2]For $\sigma^2 > 0$ and $\phi > 0$ No need to memorize (9.3)

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

$$c(h) = \begin{cases} \sigma^2 \left(1 - \frac{3}{2}\frac{\|h\|_1}{\phi} + \frac{1}{2}\left(\frac{\|h\|_1}{\phi}\right)^3\right) & \|h\|_1 \leq \phi \\ 0 & \|h\|_1 > \phi \end{cases}, \; h \in \mathbb{R}^3. \tag{9.3}$$

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.

10. Anisotropy

Note 72. 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 73. 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 74. The intrinsically stationary process $(Z_s)_{s\in\mathbb{R}^d}$ is anisotropic if its variogram is anisotropic.

Definition 75. 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 76. The weakly stationary process $(Z_s)_{s\in\mathbb{R}^d}$ is anisotropic if its covariogram is anisotropic.

Note 77. 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 73 & 75.

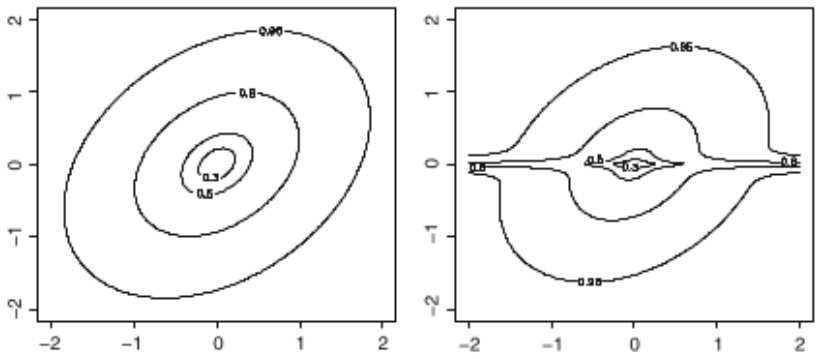
10.1. **Geometric anisotropy.**

Definition 78. 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.

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

Note 79. Such variograms have the same sill in all directions but with ranges that vary depending on the direction. See Figure 10.1a.

Example 80. For instance, if $\gamma_{\text{g.a.}}(h) = \gamma_{\text{iso}}\left(\sqrt{h^\top Q h}\right)$, where $Q = A^\top A$.



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 10.1. Isotropy vs Anisotropy

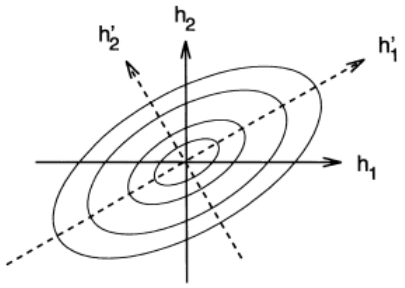


FIGURE 10.2. Rotation of the 2D coordinate system

Example 81. [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 82. 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 83. If it is $\gamma(h = (h_1, h_2)) = \gamma(h_1)$, then I ve support anisotropy

Definition 84. Zonal anisotropy occurs when the semi-variogram $\gamma(h)$ is the sum of several components each with a support anisotropy.

Example 85. 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 86. We have Zonal anisotropy then the variograms calculated in different directions suggest a different value for the sill (and possibly the range).

Note 87. 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 88. A (rather too general) non-stationary model can be specified by considering semi-variogram $2\text{Var}(Z(s) - Z(t)) = 2\gamma_o(\|G(s) - G(t)\|)$ 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

Note 89. 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 90. (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

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

Proposition 91. *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 92. 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 93. 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 94. *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 95. 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 & related parametric inference

12. The Geostatistical model

12.1. Linear Model of Regionalization.

Note 96. 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 split 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.

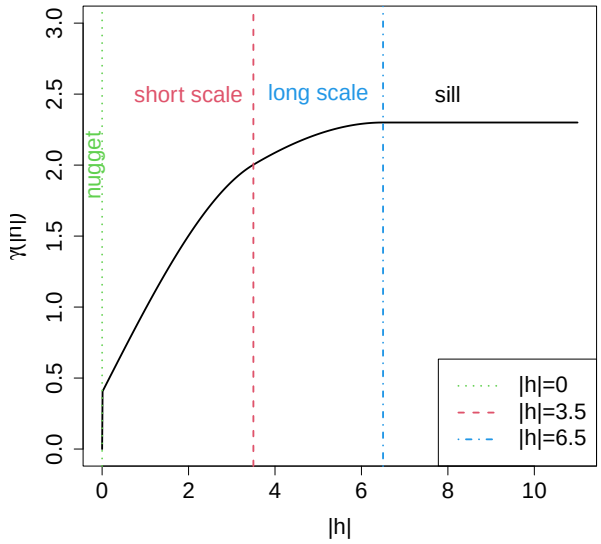


FIGURE 12.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|; \sigma^2 = 0.4\right)$.

12.1.1. *Decomposition of the stochastic process.*

Note 97. 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.

Remark 98. 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 99. For instance consider (12.1) with $\mu(s) = 0$, $m = 3$, $Z_1(s)$ with a spherical semi-variogram (9.3) with range $\phi_1 = 3.5$, $Z_2(s)$ with a spherical semi-variogram (9.3) with range $\phi_2 = 6.5$, and $Z_3(s)$ with a nugget semi-variogram. See the "sudden" changes of the line in Figure 12.1 representing change of spatial behavior.

12.1.2. *Scale of variation.*

Note 100. Cressi [1] Considers 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 101. 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 102. Clearly, smaller scale components, such as $\eta(s)$, $\varepsilon(s)$ could be identified if the sampling grid is sufficiently fine.

Note 103. Decomposition (12.2) is not unique and the components are not clearly identifiable from the data when modeled; e.g. one may find two pairs of $\mu(s), W(s)$ doing 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 104. The geostatistical model is often presented (with reference to (12.2)) is a form

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

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

Note 105. 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 106. 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.

Note 107. In several problems, additional covariates may be considered. The available dataset is of the form $\{(x_i, s_i, Z_i)\}_{i \in \mathcal{S}}$ where $Z_i := Z(s_i, x_i)$ is the observed response at

location s_i, associated with the p-dimensional covariate $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$. The popular scale decomposition in (12.2) is

$$Z(s,x) = \mu(s,x) + W(s) + \eta(s) + \varepsilon(s), \ s \in S, x \in \mathcal{X}. \tag{12.4}$$

where the dependence of $Z(\cdot)$ on x is usually propagated via the deterministic mean structure $\mu(s,x) = \mathrm{E}(Z(s,x))$ via a linear expansion of basis function. Here, to simplify the presentation, we suppress dependence on possible covariates $x \in \mathcal{X}$.

13. Training & inference

Note 108. 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 109. (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 Figure 13.1a. This is the R dataset `meuse{sp}`.

Example 110. (Wolfcamp-aquifer dataset) We also consider the Wolfcamp-aquifer dataset in Ex 5 in the Exercise sheet. See Figure 13.2a

13.1. The variogram cloud.

Definition 111. Dissimilarity between pairs of data values $Z(s_a)$ and $Z(s_b)$ is called the measure

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

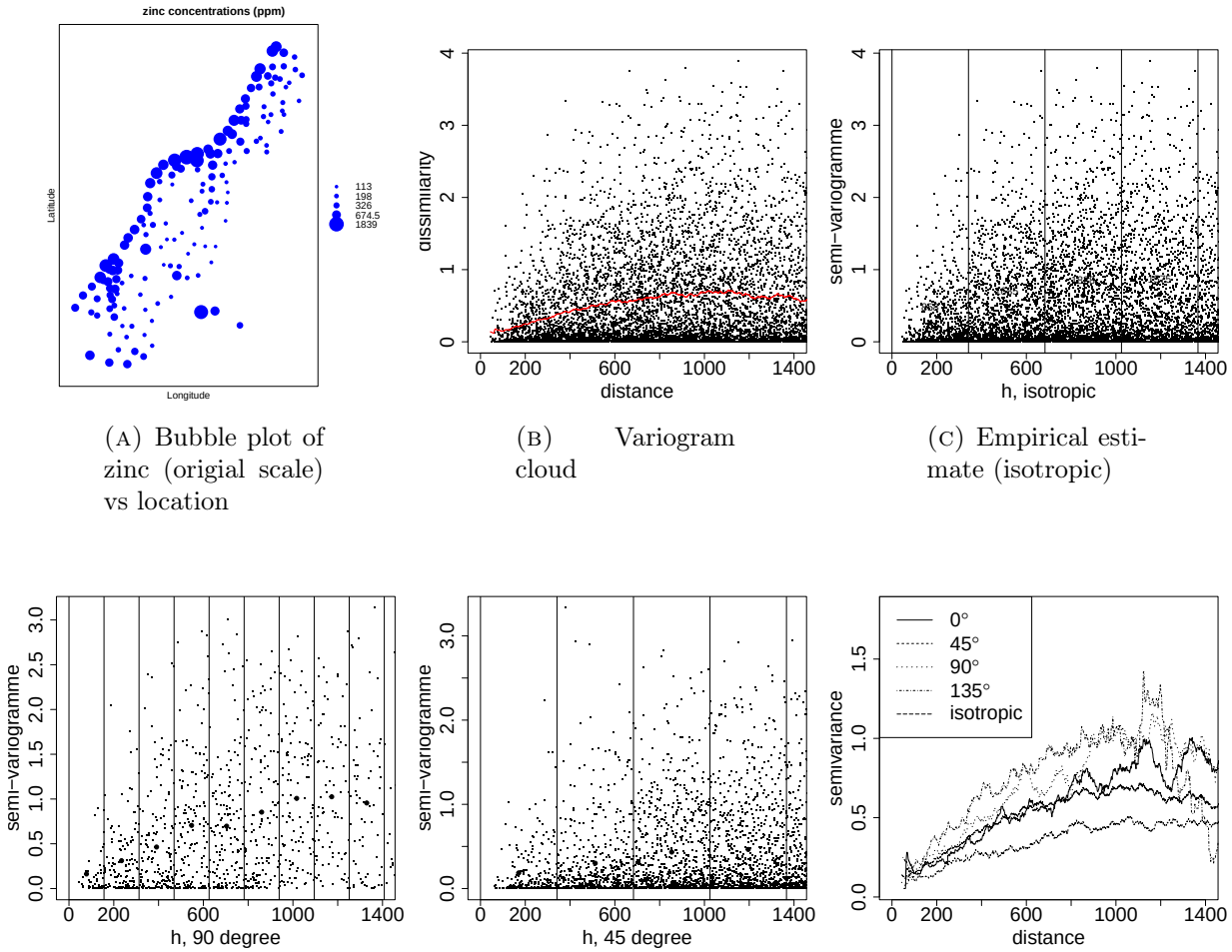
Definition 112. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (distance and orientation) then we get

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

Definition 113. 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 114. Note that (13.1) is an unbiased estimator of the variogram and hence the variogram cloud is too.



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

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

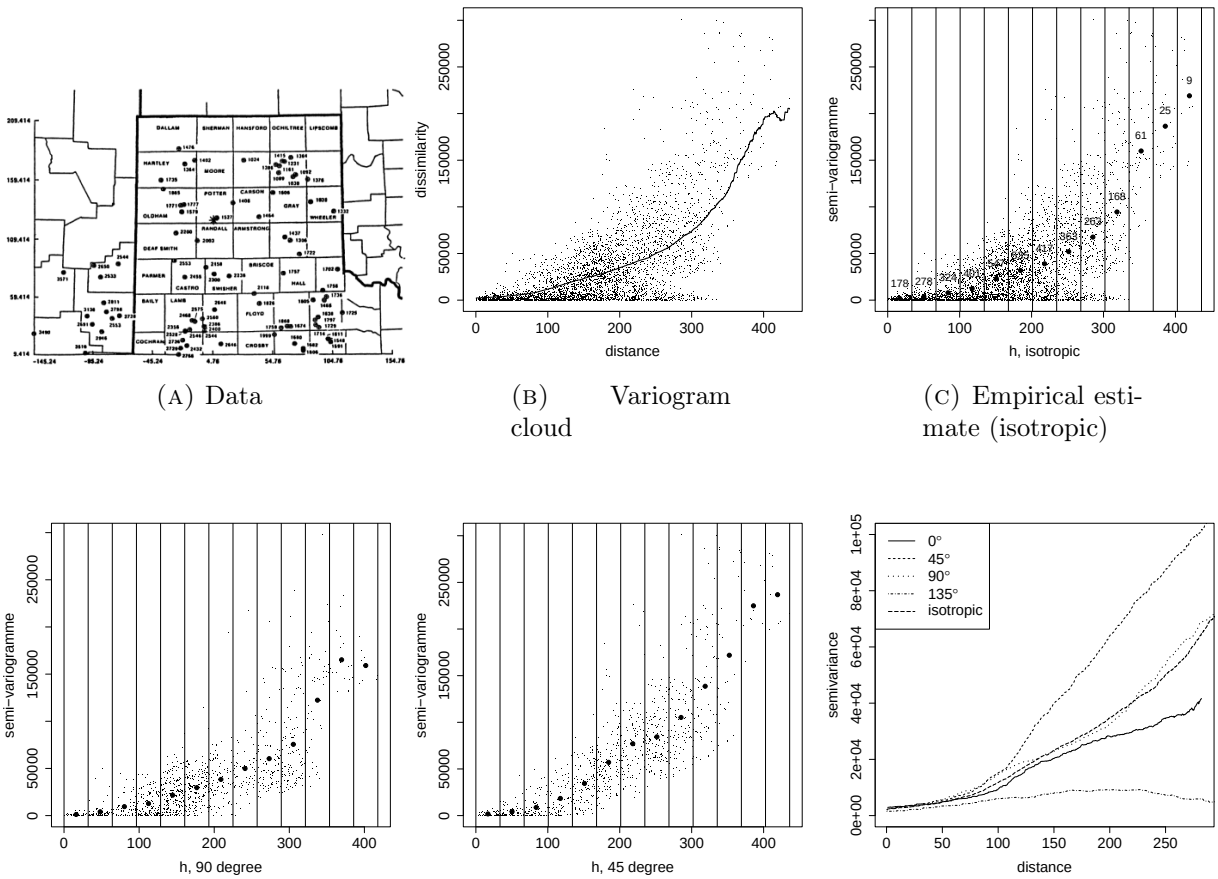
(F) Empirical estimate (different directions)

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

Note 115. 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 116. Figure 13.1b and Figure 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.**



(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

Proposition 117. *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

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

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 118. Estimator 13.2 can be written 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 119. 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 120. 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 121. 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 122. Figures 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 123. Figures 13.2dand 13.1e show the nonparametric estimator considering directions 90° and 45° for the dataset Meuse. Figures 13.2d and 13.2e do the same for the Wolfcamp-aquifer dataset. The estimator is calculated by using the ball (13.3).

Note 124. In practice anisotropies are detected by inspecting experimental variograms in different directions and are induced into the model by tuning predefined anisotropy parameters.

Example 125. Figure 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 126. *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*

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

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 127. 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. **Classic parametric estimation.**

Note 128. 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 129. Popular parametrized isotopic semi-variogrames/covariogrames are those Section 9.1. Anisotropic semi-variogrames/covariogrames can be specified by using isotropic ones and applying a rotation and dilation as in Ex 80.

Proposition 130. *(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 131. Gaussian semi-variogram in Ex 48, 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 132. *(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 133. *(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 134. *(Weighted least squares) If use $V(\theta) = diag(\varpi_1(\theta), ..., \varpi_k(\theta))$ 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 $\varpi_j(\theta) = |N_r(h_j)| / (\gamma_\theta(h_j))^2$.

Example 135. Figures 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. Figure 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 132.

13.3.2. *Least Square Errors training methods for semi-variogram with trend.*

Note 136. 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 137. *(Least square errors with trend) Do the following:*

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

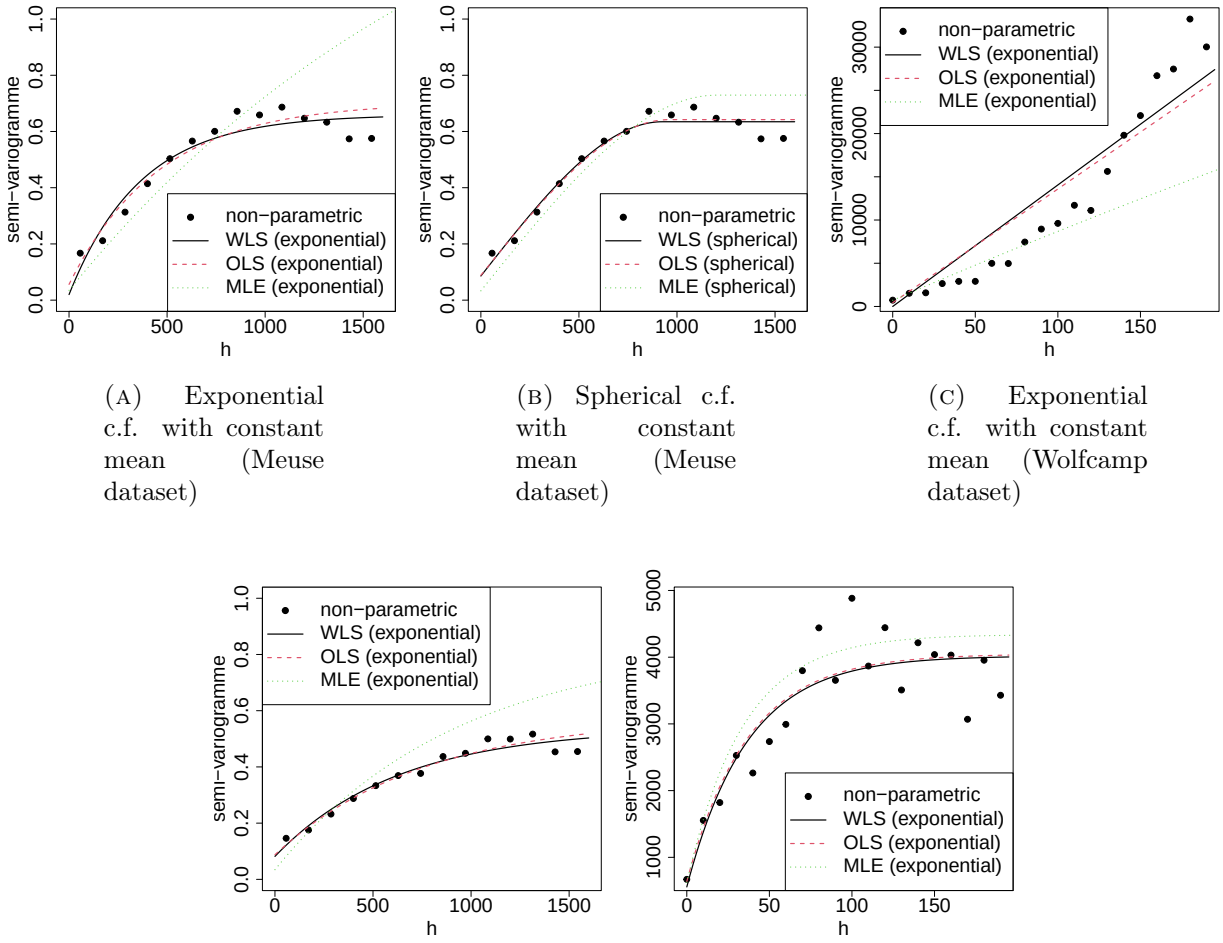
$$\hat{\beta}_{LSE} = \arg\min_{\theta} \left(\sum_i \left(Z(s_i) - \mu(s_i;\beta) \right)^2 \right)$$

(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 Proposition 117, and estimate θ according to Proposition 132.

Example 138. Figure 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. Figure



(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

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 139. Figure 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. Figure 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 Figure 13.3c.

13.4. **Maximum likelihood estimation training methods (regardless the trend).**

Note 140. 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

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

Example 141. 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

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

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. **Bayesian statistics training methods (regardless the trend).**

Note 142. 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

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

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

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

(See Handout 1, Section 3)

Note 143. If the stochastic model is $Z(\cdot)\sim\text{GP}(\mu(\cdot;\beta),c(\cdot,\cdot;\theta))$, and specify priors $(\beta,\theta)\sim\text{pr}(\beta,\theta)$, the Bayesian hierarchical model is

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

and the posterior is given by the Bayes theorem as

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

The parametric variogram can be estimated via

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

where $\mathrm{pr}\left(\theta|Z\right)=\int\mathrm{pr}\left(\beta,\theta|Z\right)\mathrm{d}\beta$.

Part 3. Prediction in geostatistics

14. The (traditional) Kriging paradigm

Note 144. "Kriging" is a general technique for deriving an estimator / predictor of $Z\left(\cdot\right)$ (or a function of it) at a location (such as a spatial point s_0, or a block of points $\left\{s_j^*\right\}$ or a subregion v_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 145. Consider we have specified the statistical model as a stochastic process $\left(Z_s\right)_{s\in\mathcal{S}}$ with

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

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

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

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

Note 146. Consider there is available a dataset $\left\{\left(s_i,Z_i\right)\right\}_{i=1}^{n}$ with $Z_i:=Z\left(s_i\right)$ being a realization of $\left(Z_s\right)_{s\in\mathcal{S}}$ at site s_i. Then one can consider the matrix form for (14.1) as

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

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 $\left[\Psi\right]_{i,j}=\psi_j\left(s_i\right)$.

[3]As mentioned in Note 159, stationarity and hence existence of the semi-variogram are not necessary in general, but they are convenient for training via the semi-variogram estimation.

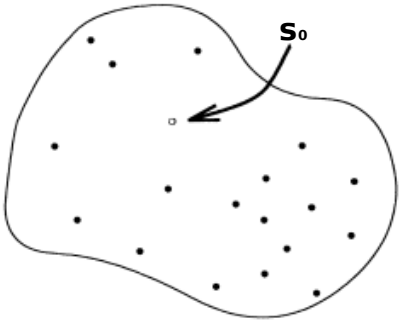


FIGURE 14.1. Kriging area

Note 147. We are interested in learning/predicting $Z(s_0)$ at an unseen spatial location s_0 (Figure 14.1).

Note 148. "Universal Kriging" (UK) is the technique for producing a Best Linear Unbiased Estimator (BLUE) predictor for $Z_0 := Z(s_0)$ at spatial location $s_0 \in \mathcal{S}$ by using data in the neighborhood of the location of interest.

Definition 149. The Universal Kriging (UK) predictor $Z_{\text{UK}}(s_0)$ of $Z(s_0)$ at location $s_0 \in \mathcal{S}$ is the Best Linear Unbiased Estimator (BLUE) of $Z(s_0)$ given the data $\{(s_i, Z_i)\}_{i=1}^n$.

Note 150. The UK predictor $Z_{\text{UK}}(s_0)$ of $Z(s_0)$ at s_0 has the following linear form weighted by a set of tunable unknown weights $\{w_i\}$

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

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

Note 151. For (14.2), to satisfy unbiasness (that is zero systematic error"), we get

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

where matrix Ψ with $[\Psi]_{i,j} = \psi_j(s_i)$ and (column) vector Ψ_0 with $[\Psi_0]_j = \psi_j(s_0)$. 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

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

Note 152. The MSE of $Z_{\text{UK}}\left(s_0\right)$, given the Assumption (14.4) is

(14.5)

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

Note 153. Now, since we have assumed that (δ_s) is intrinsic stationary, we can express $\text{E}\left(Z_{\text{UK}}\left(s_0\right)\right)$ w.r.t. the semi-variogram as

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

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

Note 154. The Lagrange function for minimizing the MSE (14.9) under (14.3) is

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

Note 155. 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 multiplying 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 156. 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 157. $(1-\alpha)\,100\%$ 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 158. Note that we have not assumed a particular distribution of Z_s or δ_s, but only stationarity assumptions.

Note 159. It was not necessary to consider the intrinsic stationarity assumption in Note 145 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: it allowed us to express

14.14 and (14.15) as functions of the semi-variogram which is discussed how to be estimated in Section 13.

Note 160. 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 Proposition 117,

(4) Compute the estimate $\hat{\theta}$ of θ of the parameterized semivariogram γ_θ, according to Proposition 132, and hence compute $\gamma_{\hat{\theta}}(\cdot)$.

Example 161. [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 160, 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 Proposition 117; 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 Figure 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 Figure 14.2a and UK in Figure 14.2b is possibly because the slops in the linear trend (mean) of UK are rather small and insignificant (See Example 139).

Example 162. (Cont. Examples 109, 125) 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. Figure 14.2c shows a rather linear relationship between Z and $\sqrt{d}$, hence we can consider

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

a UK predictor with deterministic mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d_s}$. We follow the same procedure as in Example 161 and we get the UK predictor in Figure 14.2d.

14.2. **Ordinary Kriging.**

Note 163. 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).

Note 164. 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 165. [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

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

the weights are

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

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

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

14.3. **Simple Kriging.**

Note 166. 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 167. [The derivation is in (Exercise 17 in the 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 168. 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 Proposition 117; 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 Figure 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 169. 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 170. 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 $Y = (Y_1, ..., Y_n)^\top$.

Note 171. Unlike in the traditional kriging framework, in Bayesian kriging, we have to specify a certain probabilistic model on the spatial process.

Recall

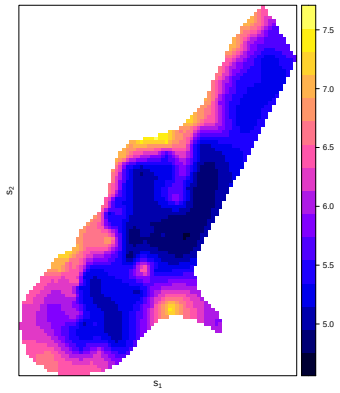
Note 172. Uncertainty can be decomposed according to the Hierarchical spatial model

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

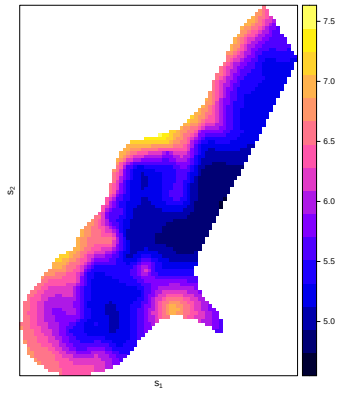
with

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

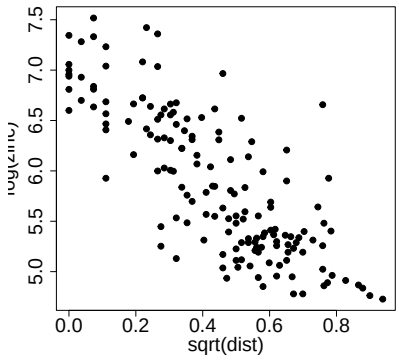
Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\text{pr}(Y|\vartheta)$ possibly labeled by some parameter ϑ.



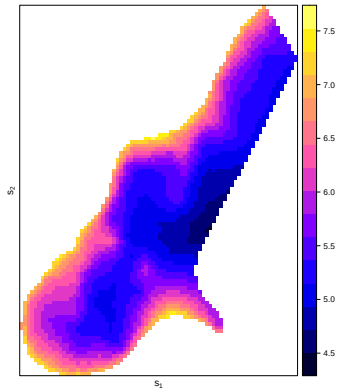
(A) Ordinary Kriging, constant mean, exponential covariogram



(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.

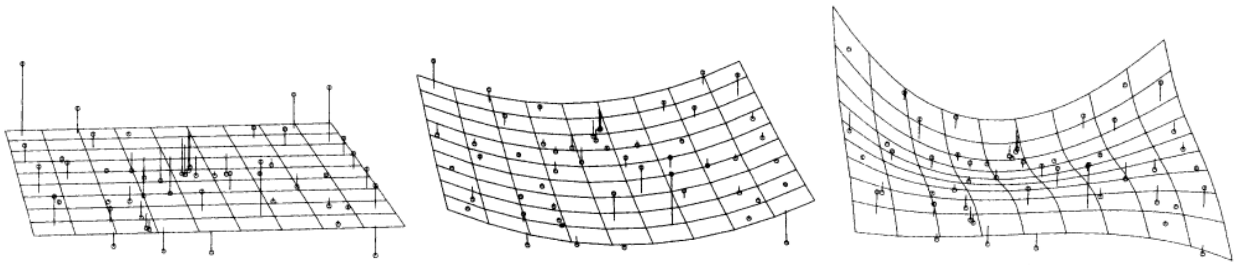


(D) Universial Kriging, mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d}$, exponential covariogram

FIGURE 14.2. Kriging Meuse dataset.

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 173. 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; in certain problems, ϑ_2 can be considered as known and sometimes as unknown in what follows.



(A) (B) (C)

FIGURE 15.1. Examples representing the hierarchical spatial model 15.2 for different values of ϑ

Note 174. Then the hierarhical model (15.2) extends to the Bayesian spatial hierarchical model

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

where uncertainty is described by

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

Note 175. Figure 15.1 presents a visualization of the hierarchical model in Notes 172 and 174. The surfaces can be considered as a realization of the spatial process model, and the dotes can be considered as realizations of the data model at specific sites given the spatial process.

Note 176. 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.

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

by maximizing the marginal likelihood

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

Note 177. Under Bayesian model (15.3), when ϑ_1 is considered as unknown (but random), namely, the a prior $\vartheta_1\sim\text{pr}\left(\vartheta_1|\vartheta_2\right)$ has been specified, uncertainty about unknown ϑ_1 given Y and ϑ_2 can be represented by the posterior distribution

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

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

Note 178. 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 179. 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 180. *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*

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

Fact 181. *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 182. *[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

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

Note 183. We are going through a particular example of the Bayesian Gaussian process regression (or Bayesian Kriging) to demonstrate how to work in the "Bayesian Kriging" framework e.g., with the spatial hierarchical models (15.2) and (15.3).

A possible narrative - a story.

Note 184. Consider 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 $\{Z_i \in \mathbb{R}\}$.

Note 185. 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 186. Consider we are interested in recovering $Z(\cdot)$

Specifying the hierarchical model.

Note 187. A natural model to setup for 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 $Y(\cdot)$

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

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

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

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

Computing the marginal process $Z(\cdot)|\beta, \theta$ for $\theta = (\sigma^2, \phi, \tau)^\top$.

Note 189. The marginal process (Z_s) given parameters β, $\theta = (\sigma^2, \phi, \tau)^\top$ (in (15.8)) is

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

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* 180 *].*

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

Note 190. 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

$$\begin{pmatrix} Z_* \\ Z \end{pmatrix} |\beta,\theta \sim \mathrm{N}\left(\begin{pmatrix} \mu(S_*;\beta) \\ \mu(S;\beta) \end{pmatrix}, \begin{pmatrix} C(S_*,S_*|\theta) & (C(S_*,S|\theta))^\top \\ C(S_*,S|\theta) & C(S,S|\theta) \end{pmatrix} \right) \tag{15.10}$$

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 191. 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

$$Z_*|Z,\beta,\theta \sim \mathrm{N}(\mu_1(S_*|\beta,\theta), C_1(S_*,S_*|\theta)) \tag{15.11}$$

where

$$\begin{aligned} C_1(S_*,S_*|\theta) =& C(S_*,S_*|\theta) + (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} C(S,S_*|\theta) \\ \mu_1(S_*|\beta,\theta) =& \mu(S_*|\beta) - (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} (\mu(S|\beta) - Z) \end{aligned}$$

[We used the formula for computing the conditional Gaussian distribution in Fact 182].

Note 192. Since the derivation of (15.11) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (15.11) can be extended to a Gaussian Process

$$Z(\cdot)|Z,\beta,\theta \sim \mathrm{GP}(\mu_1(\cdot|\beta,\theta), c_1(\cdot,\cdot|\theta)) \tag{15.12}$$

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 (Definition 18) given Note* 191*].*

Note 193. 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

$$\mathrm{pr}(Z|\beta,\theta) = \mathrm{N}(Z|\mu(S|\beta), C(S,S|\theta)) \tag{15.13}$$

derived from (15.9) by solving

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

Note 194. The estimated "Kriking predictor" results by plugging $\left(\hat{\beta},\hat{\theta}\right)^{\top}$ in (15.12), as

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

Computing the predictive distribution $Z\left(\cdot\right)|Z,\theta$ *for* $\theta=\left(\sigma^2,\phi,\tau\right)^{\top}$.

Note 195. Now, we consider that β is an unknown random hyper-parameter. To account for uncertainty, we will assign a prior distribution on β. Our aim is to compute the predictive distribution $Z\left(\cdot\right)|Z,\theta$ by integrating out β in $Z\left(\cdot\right)|Z,\beta,\theta$ wrt its posterior $\mathrm{pr}\left(\beta|Z,\theta\right)$. To facilitate this integration, below, we aim to specify a conjugate distribution on β for computational convenience.

Note 196. Like in Universal Kriging, assume that the spatial mean is parameterized as an expansion of basis functions $\psi\left(\cdot\right)=\left(\psi_1\left(\cdot\right),...,\psi_p\left(\cdot\right)\right)^{\top}$ with unknown coefficients β, i.e.

$$\mu\left(s|\beta\right)=\psi\left(s\right)^{\top}\beta$$

Note 197. The marginal process $\left(Z_s\right)$ given parameters β, and θ can be re-writen as

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

where $c\left(s,s'|\theta\right)=c_Y\left(s,s'|\sigma^2,\phi\right)+c_\varepsilon\left(s,s'|\tau\right)$, $\theta=\left(\sigma^2,\phi,\tau\right)^{\top}$ (See Note 15.9)

Note 198. We specify a conjugate prior $\beta\sim\mathrm{N}\left(b,B\right)$ on β, for some user-specified fixed hyper-parameters b and $B>0$.

Note 199. The marginal Bayesian model is now extended to

$$\begin{cases}Z|\beta,\theta\sim\mathrm{N}\left(\Psi\beta,C\left(S,S|\theta\right)\right)\\ \beta\sim\mathrm{N}\left(b,B\right)\end{cases}\tag{15.14}$$

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

Note 200. The posterior of β given data Z and θ is computed via the Bayes theorem

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

and results as

$$\beta|Z,\theta\sim\mathrm{N}\left(b_n\left(\theta\right),B_n\left(\theta\right)\right)\tag{15.15}$$

with

$$B_n(\theta) = \left(B^{-1} + \Psi^\top \left(C(S,S|\theta)\right)^{-1}\Psi\right)^{-1}$$
$$b_n(\theta) = B_n(\theta)\ \left(B^{-1}b + \Psi^\top \left(C(S,S|\theta)\right)^{-1} Z\right)$$

[For the detailed derivation see Exercise 20 in *the Exercise sheet.]*

Note 201. The posterior predictive distribution of $Z(\cdot)$ given the data Z and θ , results by integrating (15.12) with respect to (15.15) i.e.

$$\begin{aligned}\mathrm{pr}(Z_*|Z,\theta) &= \int \mathrm{pr}(Z_*|Z,\beta,\theta)\,\mathrm{pr}(\beta|Z,\theta)\,\mathrm{d}\beta \\ &= \int \mathrm{N}\left(Z_*|\mu_1(S_*|\beta,\theta), C_1(S_*,S_*|\theta)\right)\mathrm{N}(\beta|b_n,B_n)\,\mathrm{d}\beta\end{aligned}$$

and it is again a GP

$$Z(\cdot)\,|Z,\theta \sim \mathrm{GP}\left(\mu_2(\cdot|\theta), c_2(\cdot,\cdot|\theta)\right) \tag{15.16}$$

with

$$\begin{aligned}\mu_2(s|\theta) &= \left(\psi(s) - \left(C(s)\right)^\top C^{-1}\Psi\right)\left(B^{-1} + \Psi^\top C^{-1}\Psi\right)^{-1} B^{-1}b \\ &\quad + \left[\left(\psi(s) - \left(C(s)\right)^\top C^{-1}\Psi\right)\left(B^{-1} + \Psi^\top C^{-1}\Psi\right)^{-1}\Psi^\top + \left(C(s)\right)^\top\right] C^{-1}Z \qquad (15.17)\end{aligned}$$

(15.18)

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

with column vector $C(s) = \left(c(s,s_1),...,c(s,s_n)\right)^\top$, and matrix $C = C(S,S|\theta)$. *[For the detailed derivation see Exercise* 20 in *the Exercise sheet.]*

Note 202. If we consider non-informative priors in (15.14) such as $\mathrm{pr}(\beta) \propto 1$, for instance, by allowing $B^{-1} \to 0$, and $b \to 0$ then (15.18) produces the Universal Kriging predictor (check with (14.14)).

Note 203. 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

$$\mathrm{pr}(Z|\theta) = \int \mathrm{pr}(Z|\beta,\theta)\,\mathrm{pr}(\beta)\,\mathrm{d}\beta \tag{15.19}$$

$$= \int \mathrm{pr}\left(Z|\Psi\beta, C(S,S|\theta)\right)\mathrm{N}(\beta|b,B)\,\mathrm{d}\beta \tag{15.20}$$

$$= \mathrm{N}\left(Z|\Psi b, C(S,S|\theta) + \Psi B\Psi^\top\right) \tag{15.21}$$

[from Fact 181] by computing

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

Note 204. The estimated "Kriging predictor" results by plugging $\hat{\theta}$ in (15.16)

$$Z\left(\cdot\right)|Z,\hat{\theta} \sim \mathrm{GP}\left(\mu_2\left(\cdot|\hat{\theta}\right), c_2\left(\cdot,\cdot|\hat{\theta}\right)\right) \tag{15.22}$$

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

FYI: we re-parameterize the model by replacing $\tau^2 = \sigma^2\xi^2$, we consider prior $\beta|\sigma^2 \sim \mathrm{N}\left(b,\sigma^2 B\right)$ for β (essentially we replace B with $\sigma^2 B$ in the above formulas), and we specify a conjugate prior $\sigma^2 \sim \chi^2_\nu$ for $\nu > 0$ on σ^2. Then we follow the same routine as above... we can get a Students-T predictive process...

Part 4. Spatial misalignment (special topic)

16. Intro to spatial misalignment

Note 205. Consider a stochastic process $\left(Z_s\right)_{s\in\mathcal{S}}$ where $\mathcal{S} \in \mathbb{R}^d$ with $\mathrm{Var}\left(s\right) < \infty$ for all $s \in \mathcal{S}$.

Definition 206. We define the block average $Z\left(B\right)$ as

$$Z\left(B\right) = \begin{cases} \frac{1}{|B|}\int_B Z\left(x\right)\mathrm{d}x & |B| > 0 \\ \mathrm{average}\left(Z\left(x\right) : x \in B\right) & |B| = 0 \end{cases} \tag{16.1}$$

where $|B| = \int 1_B\left(x\right)\mathrm{d}x$.

Definition 207. The integral in (16.1) can be defined by Riemann sums. E.g. in 2D if $B = [a_1,a_2]\times[b_1,b_2]$, $a_1 < u_0 < ... < u_n < a_2$, $b_1 < v_0 < ... < v_n < b_2$, $u'_j \in [u_{j-1},u_j]$, and $v'_j \in [v_{j-1},v_j]$, then

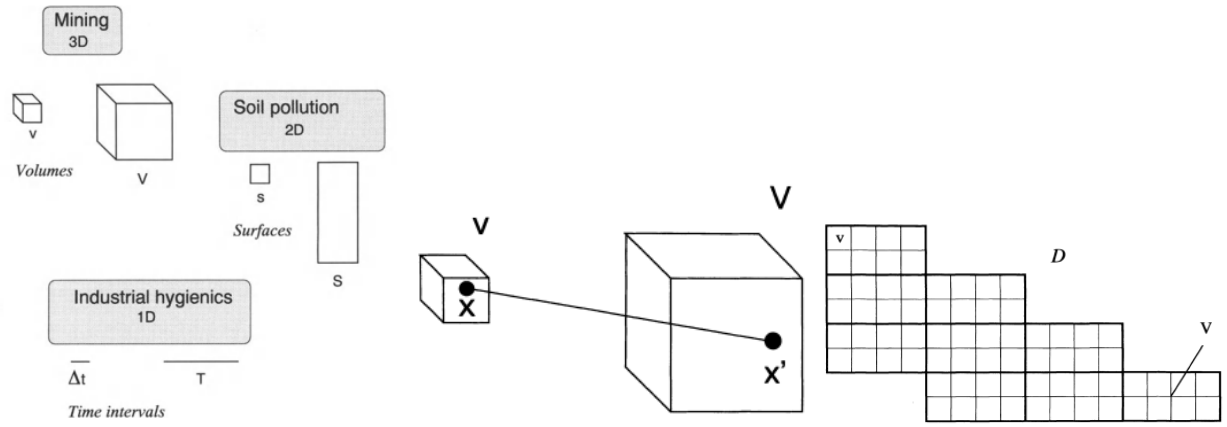
$$\int_B Z\left(x\right)\mathrm{d}x = \lim_{n\to\infty,m\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(v_j - v_{j-1}\right)\left(u_j - u_{j-1}\right)Z\left(u'_j,v'_j\right) \tag{16.2}$$

Note 208. Notice that the integral in (16.1) is a linear operator, hence for A and B it is

$$\mathrm{E}\left(Z\left(A\right)\right) = \mathrm{E}\left(\frac{1}{|A|}\int_A Z\left(x\right)\mathrm{d}x\right) = \frac{1}{|A|}\int_A \mathrm{E}\left(Z\left(u\right)\right)\mathrm{d}u$$

$$\mathrm{Cov}\left(Z\left(A\right),Z\left(B\right)\right) = \frac{1}{|A|}\frac{1}{|B|}\int_A\int_B \mathrm{Cov}\left(Z\left(u\right),Z\left(v\right)\right)\mathrm{d}u\mathrm{d}v$$

Note 209. A common problem is to predict the block average $Z\left(B\right)$ of a process $\left(Z_s\right)_{s\in\mathcal{S}}$ over a block B whose location and geometry are known and whose d-dimensional volume is $|B|$. (See Figure 16.2)



(A) Different supports (B) Extension of variance (C) Dispersion of variance

FIGURE 16.1. Change of support

Definition 210. The support of the block average $Z(B)$ in (16.1) is B and involves the geometry, size, and spatial orientation of the line, area, or volume of the input.

Change of support problem.

Note 211. Changing the support of a variable creates a new variable related to the original one but with different statistical characteristics: mean, co-variance, dependencies, etc...

Definition 212. Change of support problem refers to making inference on block of averages whose supports are different than from those of the data. $Z = (Z(B_1), ..., Z(B_n))^\top$. Often points have a point support.

16.1. **Extension and Dispersion Variance.**

Note 213. With spatial variables it is necessary to take account the spatial disposal of points, surfaces or volumes for which the variance of a quantity should be computed.

Definition 214. Extension variance $\sigma_E^2(s, s')$ of a point s with respect to another point s' is defined as

$$\sigma_E^2(s, s') = \text{Var}(Z(s) - Z(s')) = 2\gamma(s - s')$$

Notation 215. Let v be a small volume v and let V be a larger volume. Then we denote a semivariogram integral

$$\bar{\gamma}(v, V) = \frac{1}{|v||V|} \int_{s \in v} \int_{s \in V} \gamma(s - s') \, \mathrm{d}s\mathrm{d}s'$$

Proposition 216. *Extension variance $\sigma_E^2(v,V)$ of a small volume v to a larger volume V is obtained by*

$$\sigma_E^2(v,V) = 2\bar{\gamma}(v,V) - \bar{\gamma}(v,v) - \bar{\gamma}(V,V)$$

Proof. ...see Exercise 21 in the Exercise sheet. □

Definition 217. The dispersion variance of the values $\{z_j\}$ of the small volumes v_j building up V is

$$\sigma^2(v|V) = \frac{1}{n}\sum_{j=1}^{n}\sigma_E^2(v_j,V)$$

Note 218. Now, suppose a large volume V is partitioned into n smaller units $\{v_j\}_{j=1}^n$ of equal size and geometry (Figure 16.1b). We get the following two intuitive results.

Proposition 219. *Suppose a large volume V is partitioned into n smaller units $\{v_j\}_{j=1}^n$ of equal size and geometry. The dispersion variance $\sigma^2(v|V)$ can be written in term of variogram integrals as*

$$\sigma^2(v|V) = \bar{\gamma}(V,V) - \bar{\gamma}(v,v) \tag{16.3}$$

Proof. ...see Exercise 23 in the Exercise sheet. □

Proposition 220. *[Krige's relation] Consider that the domain S is partitioned into volumes V which are partitioned into smaller volumes v. Then the relation between the three supports is*

$$\sigma^2(v|\mathcal{S}) = \sigma^2(v|V) + \sigma^2(V|\mathcal{S}) \tag{16.4}$$

Proof. (Sketch of the proof) For just a point s, (16.3) becomes

$$\sigma^2(s|V) = \bar{\gamma}(V,V) - \cancel{\bar{\gamma}(s,s)} = 0$$

similar

$$\sigma^2(s|v) = \bar{\gamma}(v,v) - \cancel{\bar{\gamma}(s,s)} = 0$$

so (16.3) gives

$$\sigma^2(v|V) = \sigma^2(s|V) - \sigma^2(s|v) \Leftrightarrow$$
$$\sigma^2(s|V) = \sigma^2(s|v) + \sigma^2(v|V) \tag{16.5}$$

In greater scale, the above can be extended to

$$\sigma^2(v|\mathcal{S}) = \sigma^2(v|V) + \sigma^2(V|\mathcal{S})$$

□

Note 221. The knowledge of the semi-variogram makes the computation of $\sigma^2(v|\mathcal{S})$, $\sigma^2(v|V)$, and $\sigma^2(V|\mathcal{S})$ possible.

Change of support effect.

Note 222. Consider the case that the domain S is partitioned into volumes V which are partitioned into smaller volumes v. Assume there are available samples at "point" locations s each of them lies to the center of one of the smaller volumes v. Making the assumption that the sampled value at each point location s is extended to each area of influence v implies that the distribution of average values of the blocks is the same as the distribution of the values at the sample points. However from (16.5), we see that this is not true; in fact the distribution of the values for a support v is narrower than the distribution of point values because the variance $\sigma^2(s|v)$ of the points in v generally is not negligible; i.e. $\sigma^2(s|V) - \sigma^2(v|V) = \sigma^2(s|v) > 0$.

Change of support: affine model.

Note 223. Consider a stationary process $Z(s)$ for $s \in \mathcal{S}$, and consider a block process $Z_v(s)$ on a block v. The affine model assumes that the standardized point variable $Z(s)$ follows the same distribution $Z(v)$ as the standardized block variable.

Example 224. An example of the use of affine models is the Gaussian process case, where $Z(s) \sim \mathrm{N}(\mu, \sigma^2)$ and $Z(v) \sim \mathrm{N}(\mu, \sigma_v^2)$, –same mean but different variances– it is

$$\frac{Z(s) - \mu}{\sqrt{\sigma^2}} \overset{\text{distr.}}{\sim} \frac{Z(v) - \mu}{\sqrt{\sigma_v^2}} \sim \mathrm{N}(0, 1)$$

which implies the relation

$$Z(v) \overset{\text{distr.}}{\sim} \mu + \sqrt{\frac{\sigma_v^2}{\sigma^2}} (Z(s) - \mu) \sim \mathrm{N}\left(\mu, \sigma_v^2\right)$$

16.2. **Block (Universal) Kriging.**

Note 225. Block Kriging aims to predict a block value $Z(v_0)$ at block v_0 instead of at a point value s_0; see Figure 16.2. It can be used within the framework of Universal, Ordinary, Simple, and Bayesian Kriging cases we saw in Section 14.1.

Note 226. Assume we want the estimate a block value $Z(v_0)$ at block v_0 with some volume $|v_0|$ given that my data $\{(s_i, Z_i)\}_{i=1}^n$ are realizations $Z_i = Z(s_i)$ at point values s_i (Figure 16.2).

Note 227. Here, we present the Block Kriging in the (traditional) Universal Kriging framework (Section 14.1). We will refer to the UK in Section 14.1 as point-to-point UK.

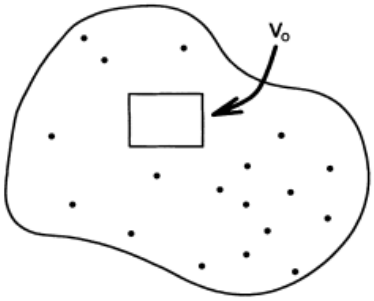


FIGURE 16.2. Block Kriging cartoon

Note 228. Consider that the statistical model is the stochastic process $(Z_s)_{s\in\mathcal{S}}$ with

$$Z(s) = \mu(s) + \delta(s) \tag{16.6}$$

Assume

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

with vector of unknown coefficients $\beta = (\beta_0, ..., \beta_p)^\top$ and vector of known basis functions $\psi(s) = (\psi_0(s), ..., \psi_p(s))^\top$. Assume $\delta(s)$ is a zero mean process. Assume $\delta(s)$ is an intrinsic stationary process with a semi-variogram $\gamma(\cdot)$ –as in UK in Section 14.1, intrinsic stationarity is not a necessary assumption if one can estimate the covariance function directly.

Note 229. The Block UK predictor $Z_{\text{UK}}(v_0)$ of $Z(v_0)$ at block v_0 with support $|v_0| > 0$ has the following linear form weighted by a set of tunable unknown weights

$$Z^*_{\text{BK}}(v_0) = w_{n+1} + \sum_{i=1}^{n} w_i Z(s_i) = w_{n+1} + w^\top Z \tag{16.7}$$

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

Note 230. Following the steps in (point-to-point) UK (Note 151), unbiasness implies conditions $w_{n+1} = 0$ and

$$\text{ASSUMPTION:} \quad \Psi_0 = \sum_{i=1}^{n} w_i \psi(s_i) \Leftrightarrow \psi_0 = w^\top \Psi \tag{16.8}$$

where $[\Psi_0]_j = \psi_j(v_0)$, and $\psi_j(v_0) = \frac{1}{|v_0|}\int \psi_j(s)\,\mathrm{d}s$ for $j = 0, ..., p$.

Note 231. Following the steps in (point-to-point) UK (Note 153), I get

$$\begin{aligned} \text{MSE}\left(Z_{\text{BK}}\left(v_0\right)\right) - \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j \gamma\left(s_i - s_j\right) + 2\sum_{i=1}^{n} w_i \bar{\gamma}\left(s_i, v_0\right) \\ = -w^\top \mathbf{\Gamma} w + 2w^\top \bar{\boldsymbol{\gamma}}_0 \end{aligned} \tag{16.9}$$

where $\bar{\boldsymbol{\gamma}}_0 = \left(\bar{\gamma}\left(s_1, v_0\right), ..., \bar{\gamma}\left(s_n, v_0\right)\right)^\top$, and $\bar{\gamma}\left(s_i, v_0\right)$ be the average variogram of each sample point with the block of interest. This is the same as that of point-to-point UK in (14.5) where the point $\gamma\left(s_i, s_0\right)$ is substituted by the integral $\bar{\gamma}\left(s_i, v_0\right)$.

Note 232. The Block Universal Kriging equations then are

$$\begin{cases} 0 = -2\mathbf{\Gamma} w + 2\bar{\boldsymbol{\gamma}}_0 - \Psi \lambda_{\text{UK}} \\ \Psi_0 = w_{\text{BK}}^\top \Psi \end{cases} \tag{16.10}$$

which essentially produce the same weights as the point-to-point Universal Kriging but averaged out in the block

$$w_{\text{BK}} = \mathbf{\Gamma}^{-1}\left(\bar{\boldsymbol{\gamma}}_0 - \Psi\left(\Psi^\top \Gamma^{-1} \Psi\right)^{-1}\left(\Psi^\top \Gamma^{-1} \bar{\boldsymbol{\gamma}}_0 - \Psi_0^\top\right)\right) \tag{16.11}$$

$$\lambda_{\text{BK}} = 2\left(\Psi^\top \Gamma^{-1} \Psi\right)^{-1}\left(\Psi^\top \Gamma^{-1} \boldsymbol{\gamma}_0 - \Psi_0^\top\right) \tag{16.12}$$

Note 233. Hence the UK predictor $Z_{\text{UK}}\left(s_0\right)$ at s_0 is

$$Z_{\text{BK}}\left(s_0\right) = \left(\bar{\boldsymbol{\gamma}}_0 - \Psi\left(\Psi^\top \Gamma^{-1} \Psi\right)^{-1}\left(\Psi^\top \Gamma^{-1} \bar{\boldsymbol{\gamma}}_0 - \Psi_0^\top\right)\right)^\top \mathbf{\Gamma}^{-1} Z \tag{16.13}$$

with standard error (by substituting (16.10) in(16.9))

$$\sigma_{\text{BK}}\left(v_0\right) = \sqrt{w^\top \Psi \lambda_{\text{UK}} + w^\top \bar{\boldsymbol{\gamma}}_0} \tag{16.14}$$

Note 234. Block Kriging as a concept can be implemented even when s_i are not points but have some volume $|s_i| > 0$. Then we call the case as aggregation if $|s_i| < |v_0|$, or disaggregation if $|s_i| > |v_0|$.

16.3. **Block (Bayesian) Kriging.**

Note 235. If $Z\left(s\right)$ is a Gaussian process defined on points $s \in \mathcal{S}$, then the block average $Z\left(v\right)$ with $v \subset \mathcal{S}$ is a Gaussian process as well. This is because integration (or averaging) in (16.1) is a linear operation as seen in (16.2), and linear combinations of Gaussians is Gaussian as well.

Note 236. Block (Bayesian) Kriging is produced in the same lines as the Bayesian Kriging procedure (Section 15): (1.) compute the joint distribution in (15.10) i.e.

$$\begin{pmatrix} Z_{v_0} \\ Z \end{pmatrix} | \beta, \theta \sim \text{N}\left(\begin{pmatrix} \mu\left(v_0; \beta\right) \\ \mu\left(S; \beta\right) \end{pmatrix}, \begin{pmatrix} c\left(v_0, v_0 | \theta\right) & \left(C\left(v_0, S | \theta\right)\right)^\top \\ C\left(v_0, S | \theta\right) & C\left(S, S | \theta\right) \end{pmatrix}\right)$$

with

$$\mu(v_0;\beta) = \frac{1}{|v_0|}\int_{x\in v_0}\mu(x;\beta)\,\mathrm{d}x$$
$$c(v_0,s_i|\theta) = \frac{1}{|v_0|}\int_{x\in v_0}c(x,s_i|\theta)\,\mathrm{d}x$$
$$c(v_0,v_0'|\theta) = \frac{1}{|v_0|\,|v_0'|}\int_{x\in v_0}\int_{y\in v_0'}c(x,y|\theta)\,\mathrm{d}x\mathrm{d}y$$

(2.) compute the the predictive distribution as the conditional Normal distribution $\text{pr}(Z_{v_0}|Z,\beta,\theta)$ (Note 191), and (3.) recognize the corresponding Gaussian process as in Note 192. The derivation is identical to that Section 15.2.

Part 5. Extensions to multivariate Geostatistics (special topic)

17. Extensions to Multivariate Geostatistics

17.1. Cross-variance functions.

Definition 237. Let $Z_1(s),...,Z_k(s)$ be k stochastic processes on $s\in\mathcal{S}$. The cross-covariance function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

$$C_{i,j}(s,t) = \text{Cov}(Z_i(s),Z_j(t)) = \text{E}((Z_i(s)-\text{E}Z_i(s))(Z_j(t)-\text{E}Z_j(t)))$$

for $i,j=1,...,k$ and $s,t\in\mathcal{S}$.

Definition 238. Let $Z_1(s),...,Z_k(s)$ be k weakly stationary stochastic processes on $s\in\mathcal{S}$. The cross-covariogram function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

$$C_{i,j}(h) = \text{Cov}(Z_i(s),Z_j(s+h)) = \text{E}((Z_i(s)-\text{E}(Z_i(s)))(Z_j(s+h)-\text{E}(Z_j(s+h))))$$

for $i,j=1,...,k$ and $s,s+h\in\mathcal{S}$.

Example 239. Cross-covariograms have the following properties

(1) $C_{i,j}(h) = C_{j,i}(-h)$ and $C_{i,j}(h)\neq C_{j,i}(h)$
(2) $C_{i,j}(h)$ is semi-positive definite

Solution. Well, part 1 is easy to check. Now for Part 2, $\forall w_{j,i}\in\mathbb{R}$, I get

$$0\leq\text{Var}\left(\sum_j\sum_i w_{j,i}Z_j(s_i)\right) = \sum_j\sum_{j'}\sum_i\sum_{i'}w_{j,i}w_{j',i'}C_{j,j'}(s_i-s_{i'})$$

Definition 240. Let $Z_1(s),...,Z_k(s)$ be k intrinsically stationary stochastic processes on $s\in\mathcal{S}$. The cross-variogram function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

$$\gamma_{i,j}(h) = \frac{1}{2}\text{Cov}((Z_i(s+h)-Z_i(s)),(Z_j(s+h)-Z_j(s)))$$

for $i,j=1,...,k$ and $s,s+h\in\mathcal{S}$.

Example 241. Let $Z_i(s)$ and $Z_j(s)$ be weakly stationary stochastic processes on $s\in\mathcal{S}$, with $\mathrm{E}Z_i(s)=\mathrm{E}Z_i(s+h)$. Then

$$C_{i,j}(h)=\underbrace{\frac{1}{2}\left(C_{i,j}(+h)-C_{i,j}(-h)\right)}_{\text{odd term}}+\underbrace{\frac{1}{2}\left(C_{i,j}(+h)+C_{i,j}(-h)\right)}_{\text{even term}}$$

$$\begin{aligned}\gamma_{i,j}(h)&=\frac{1}{2}\mathrm{E}\left(\left(Z_i(s+h)-Z_i(s)\right)\left(Z_j(s+h)-Z_j(s)\right)\right)\\&=\frac{1}{2}\left(\mathrm{E}\left(Z_i(s+h)Z_j(s+h)\right)-\mathrm{E}\left(Z_i(s)Z_j(s+h)\right)\right.\\&\qquad\left.-\mathrm{E}\left(Z_i(s)Z_j(s+h)\right)+\mathrm{E}\left(Z_i(s)Z_j(s)\right)\right)\\&=C_{i,j}(0)-\underbrace{\frac{1}{2}\left(C_{i,j}(+h)+C_{i,j}(-h)\right)}_{\text{even term}}\end{aligned}$$

Note 242. Ex 241 implies that the cross-variogram is not adequate for modeling as it covers only the even term of the cross-covariance function.

Definition 243. Let $Z_1(s),...,Z_k(s)$ be k intrinsically stationary stochastic processes on $s\in\mathcal{S}$. The pseudo-cross-variogram function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

$$\tilde{\gamma}_{i,j}(h)=\frac{1}{2}\mathrm{Var}\left(Z_i(s+h)-Z_i(s)\right)$$

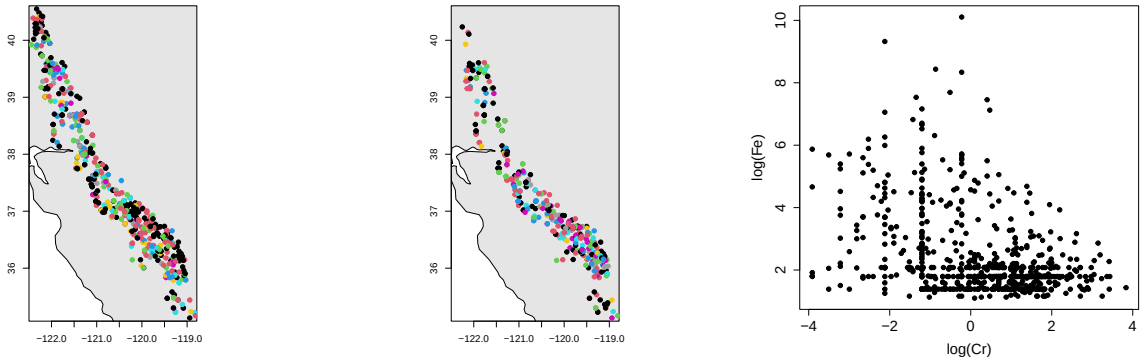
for $i,j=1,...,k$ and $s,s+h\in\mathcal{S}$.

Note 244. Pseudo-cross-variogram function has the advantage that it is not necessarily even, hence it can model more cases. Its disadvantages involve that (1) it is positive, hence it cannot model negative cross-dependencies, and that (2) stationarity across increments is unrealistic as it may consider differences of variables with different unites.

17.2. **Co-Kriging.**

Note 245. The cokriging procedure is a natural extension of kriging when the cross-covariance function/model is available. A variable of interest is cokriged at a specific location from data about itself and about auxiliary variables in the neighborhood. Examples:

(1) Different variables correspond to a different characteristics. The variable of interest can be chromium (Cr) and the auxiliary variables can be $\{Z_1(s_i)\}_{i=1}^{n_1}$ the precipitation measured by a weather station , and $\{Z_1(s_i)\}_{i=1}^{n_1}$ Chromium (Cr), and $\{Z_2(s_i)\}_{i=1}^{n_2}$ Iron (Fe) in Ex. 15 of (Handout 1). (Figure 17.1) Interest lies in the predictive process $Z_1(\cdot)\,|\,\{Z_{1,i}\},\{Z_{2,i}\}$ of $Z_1(s)$ at any point s given all the available data $\{Z_1(s_i)\}_{i=1}^{n_1}$ and $\{Z_2(s_i)\}_{i=1}^{n_1}$ (i.e. given the combined data).



(A) Cr concentration

(B) Fe concentration

(C) Scatterplot Cr vs Fe

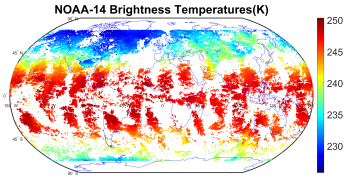
FIGURE 17.1. Central valley Groundware example data

(2) Different variables correspond to a different accuracy level or support. The variable of interest could be the precipitation in a location, and the auxiliary variables could be: $\{Z_1(s_i)\}_{i=1}^{n_1}$ the precipitation measured by a weather station , and $\{Z_2(s_i)\}_{i=1}^{n_2}$ the precipitation measured by a satellite. The weather station measurements are much more accurate than those from the satellite however they are taken at a smaller number of locations $n_1 \ll n_2$. Interest lies in the predictive process $Z_1(\cdot) \mid \{Z_{1,i}\}, \{Z_{2,i}\}$ of $Z_1(s)$ at any point s given all the available data $\{Z_1(s_i)\}_{i=1}^{n_1}$ and $\{Z_2(s_i)\}_{i=1}^{n_1}$ (i.e. given the combined data).

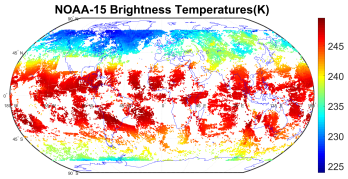
(3) Different variables correspond to a different accuracy level or support. The variable of interest could be the temperature reading in a location, and the auxiliary variables could be $\{Z_1(s_i)\}_{i=1}^{n_1}$ the temperature readings by an old technology (less accurate) satellite, and $\{Z_2(s_i)\}_{i=1}^{n_1}$ the temperature readings by a new technology (more accurate) satellite. Interest lies in the predictive process $Z_2(\cdot) \mid \{Z_{1,i}\}, \{Z_{2,i}\}$ of $Z_2(s)$ at any point s given all the available data $\{Z_1(s_i)\}_{i=1}^{n_1}$ and $\{Z_2(s_i)\}_{i=1}^{n_1}$ (i.e. given the combined data). (Figure 17.2)

Note 246. We present the concept in the ordinary Kriging framework.

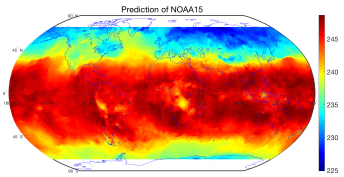
Note 247. Consider k stochastic processes $Z_1(s), ..., Z_k(s)$, $s \in \mathcal{S}$. Consider data at n sites $\{s_i\}_{i=1}^{n}$. Let $\boldsymbol{Z}(\boldsymbol{s})$ be a $n \times k$ matrix $\boldsymbol{Z}(\boldsymbol{s}) = Z_j(s_i)$ for $i = 1, ..., n_j$, and $j = 1, ..., k$. It is desired to predict the j_0-th variable $Z_{j_0}(s_0)$ for some $j_0 \in \{1, ..., k\}$ at location $s_0 \in \mathcal{S}$.



(A) N14 (less accurate)



(B) N15 (more accurate)



(C) N15 (co-kriging prediction)

FIGURE 17.2. Satellite temperature readings data

Note 248. Assume

$$\mathrm{E}\left(Z_j\left(s\right)\right)=\mu_j,\text{forall }\ j=1,...,k,\ \text{ and }\ s\in\mathcal{S}$$
$$\mathrm{Cov}\left(Z_i\left(s\right),Z_j\left(t\right)\right)=C_{i,j}\left(s,t\right),\text{forall }\ i,j=1,...,k,\ \text{ and }\ s\in\mathcal{S}$$

Note 249. Co-Kriging predictor $Z_{\mathrm{CK},j_0}\left(s_0\right)$ is the BLUE predictor $Z_{\mathrm{CK},j_0}\left(s_0\right)$ of $Z_{j_0}\left(\cdot\right)$ at s_0.

Note 250. The Co-Kriging predictor has the linear form

$$Z_{\mathrm{CK},j_0}\left(s_0\right)=w_{0,0}+\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)=w_{0,0}+\sum_{j=1}^{k}w_j^{\top}Z_j \tag{17.1}$$

weighted by a set of tunable unknown weights $\{w_{j,i}\}$, $Z_j=\left(Z_j\left(s_i\right),...,Z_j\left(s_i\right)\right)^{\top}$ and $w_j=\left(w_{j,1},...,w_{j,n_j}\right)^{\top}$.

Note 251. Parametrization (17.1) requires that all $Z_j\left(\cdot\right)$ components are observed at each site s_i. However the concept of co-kriging can also be adjusted to consider more general cases such as those where different processes $Z_j\left(\cdot\right)$ are observed at different sets of sites from each other.

Note 252. To enforce unbiassness, we find sufficient conditions for $\{w_{j,i}\}$

$$\begin{aligned}\mathrm{E}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)-Z_{j_0}\left(s_0\right)\right)=&\mathrm{E}\left(\overbrace{w_{0,0}}^{\overset{\text{ass}}{=}0}+\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)\right.\\&\left.-\underbrace{\sum_{i=1}^{n_{j_0}}w_{j_0,i}}_{\underset{\text{ass}}{=1}}Z_{j_0}\left(s_i\right)-\sum_{j\neq j_0}\underbrace{\sum_{i=1}^{n_j}w_{j,i}}_{\underset{\text{ass}}{=0}}Z_j\left(s_i\right)\right)\\=&\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}\left(Z_j\left(s_i\right)-Z_j\left(s_i\right)\right)\right)=0\end{aligned}$$

so sufficient conditions for $\{w_{j,i}\}$ are $w_{0,0}=0$ and for $j=1,...,k$,

$$\sum_{i=1}^{n_j}w_{j,i}=\begin{cases}1 & , \quad j=j_0\\ 0 & , \quad j\neq j_0\end{cases},\qquad \text{for } \ j=1,...,k$$

Notation 253. Set convenient notation for the calculations below as

$$w_{j,0}=\begin{cases}-1 & , \quad j=j_0\\ 0 & , \quad j\neq j_0\end{cases},\qquad \text{for } \ j=1,...,k$$

Note 254. The MSE (or Variance) is

$$\begin{aligned}
\operatorname{MSE}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)\right) =& \mathrm{E}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)-Z_{j_0}\left(s_0\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)-Z_{j_0}\left(s_0\right)\right)^2 \\
=& \mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)-Z_{j_0}\left(s_0\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=0}^{n_j}w_{j,i}Z_j\left(s_i\right)\right)^2 \\
=& \mathrm{E}\left(\sum_{j=1}^{k}\left(\sum_{i=0}^{n_j}w_{j,i}Z_j\left(s_i\right)-\sum_{i=1}^{n_j}w_{j,i}\mu_j\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=0}^{n_j}\left(Z_j\left(s_i\right)-\mu_j\right)\right)^2 \\
=& \sum_{j=1}^{k}\sum_{i=0}^{n_j}\sum_{j'=1}^{k}\sum_{i'=0}^{n_{j'}}w_{j,i}w_{j',i'}\mathrm{E}\left(Z_j\left(s_i\right)-\mu_j\right)\left(Z_{j'}\left(s_{i'}\right)-\mu_{j'}\right) \\
=& \sum_{j=1}^{k}\sum_{i=0}^{n_j}\sum_{j'=1}^{k}\sum_{i'=0}^{n_{j'}}w_{j,i}w_{j',i'}C_{j,j'}\left(s_i,s_{i'}\right) \\
=& \sum_{j=1}^{k}\sum_{i=1}^{n_j}\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j,i}w_{j',i'}C_{j,j'}\left(s_i,s_{i'}\right) \\
& -2\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j',i'}C_{j_0,j'}\left(s_0,s_{i'}\right)+C_{j_0,j_0}\left(s_0,s_0\right) \qquad (17.2)
\end{aligned}$$

Note 255. The Lagrange function is

$$\begin{aligned}
\mathfrak{L}\left(w,\lambda\right) =& \sum_{j=1}^{k}\sum_{i=1}^{n_j}\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j,i}w_{j',i'}C_{j,j'}\left(s_i,s_{i'}\right)-2\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j',i'}C_{j_0,j'}\left(s_0,s_{i'}\right)+C_{j_0,j_0}\left(s_0,s_0\right) \\
& -2\sum_{j'\neq j_0}\lambda_{j'}\left(\sum_{i=1}^{n_{j'}}w_{j',i}-0\right)-2\lambda_{j_0}\left(\sum_{i=1}^{n_{j_0}}w_{j_0,i}-1\right)
\end{aligned}$$

Note 256. The CK system of equations produced by $0=\nabla_{(w,\lambda)}\mathfrak{L}\left(w,\lambda\right)\big|_{(w_{\mathrm{CK}},\lambda_{\mathrm{CK}})}$ is

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}C_{j,j'}\left(s_i,s_{i'}\right)-\lambda_{j'}=C_{j_0,j'}\left(s_0,s_{i'}\right),\quad \forall j',i' \qquad (17.3)\\
\sum_{i=1}^{n_{j_0}}w_{j_0,i}=1,\quad \sum_{i=1}^{n_{j'}}w_{j',i}=0,\ \forall j'
\end{aligned}$$

Note 257. Plugin (17.3) in (17.2), I can get the co-Kriging variance

$$\sigma_{\mathrm{CK}}^2:=\operatorname{MSE}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)\right)=C_{j_0,j_0}\left(s_0,s_0\right)+\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}C_{j_0,j}\left(s_0,s_i\right)+\lambda_{j_0}$$

Note 258. The above derivation can be done wrt the cross-variogram as in UK, OK (by making extra assumptions). I choose to presented wrt the cross-covariance as more general.

Note 259. Regarding the Bayesian framework. Consider the paradigm that $Z_j(\cdot)$ are GP, where $\mu_j(\cdot) = \mathrm{E}(Z_j(\cdot))$ and $c_{j,j'}(\cdot) = \mathrm{Cov}(Z_j(\cdot), Z_{j'}(\cdot))$. Let set of sites $S_j = \{s_{j,1}, ..., s_{j,n_j}\}$ and assume there is an available dataset $\{(Z_{j,i}, s_{j,i})\}_{i=1}^{n_j}$ for $j = 1, ..., k$. The procedure is the same as discussed in Section 15, with the only difference that the predictive Gaussian process will be $Z_{j_0}(\cdot) \mid \{Z_{1,i}\}, ..., \{Z_{k,i}\}$, for $j_0 \in \{1, ..., k\}$ and resulted after conditioning the following joint distribution on Z_1,..., Z_k

(17.4)

$$\begin{bmatrix} [Z_{j_0}(S_*)] \\ \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} [\mu_{j_0}(S_*)] \\ \begin{bmatrix} \mu_1(S_1) \\ \vdots \\ \mu_k(S_k) \end{bmatrix} \end{bmatrix}, \begin{bmatrix} [C_{j_0,j_0}(S_*, S_*)] & \begin{bmatrix} C_{j_0,1}(S_*, S_1) & \cdots & C_{j_0,k}(S_*, S_k) \end{bmatrix} \\ \begin{bmatrix} C_{1,j_0}(S_1, S_*) \\ \vdots \\ C_{k,j_0}(S_k, S_*) \end{bmatrix} & \begin{bmatrix} C_{j_1,j_1}(S_1, S_1) & \cdots & C_{1,k}(S_1 S_k) \\ \vdots & \ddots & \vdots \\ C_{k,1}(S_k, S_1) & \cdots & C_{k,k}(S_k, S_k) \end{bmatrix} \end{bmatrix} \right)$$

Note 260. If k is large with moderate large n_j for each (or some) j's, the calculations in (17.3) and 17.4 can be too computationally challenging and have unrealistic computational requirements for a standard PC. E.g., we will have to solve a huge system of equations in (17.3), while we will have to do operations with a huge covariance matrix in 17.4. In 90's your computer (particularly its CPU and its RAM) would complain with a blue screen...

Note 261. For instance some tricks to mitigate challenges with large k and n_j involve specifying restricted forms of cross-covariance functions $C_{i,j}(s,t)$ with special structure often introducing conditional independences (hence restricting the model), as well as using suitable experimental designs.

Definition 262. Intrinsic Multivariate Correlation (co-Kriking) model is the CK model which assumes that the multivariate correlation structure is independent of the spatial correlation; i.e. the following correlation

$$\frac{C_{i,j}(s,t)}{\sqrt{C_{i,i}(s,t)\, C_{j,j}(s,t)}}$$

between any two $Z_i(\cdot)$ $Z_j(\cdot)$ at (s,t) does not depend upon spatial scale $|t-s|$ or (s,t) for any pair of sites (s,t).

Example 263. Consider a set of processes $\{Z_j(\cdot)\}$ where the cross-covariance is modeled as $C_{i,j}(s,t) = \sigma_{i,j}\varrho(|s-t|)$ where $\sigma_{i,j}$ is the (i,j) element of a semi-positive matrix Σ and $\varrho(h) = c(h)/c(0)$ is the correlogram of some isotropic covariogram function $c(\cdot)$. Show that this co-kriging model is Intrinsic Multivariate Correlation model.

Solution. The correlation between two variables For any pair of spatial points (s,t) the

$$\frac{C_{i,j}(s,t)}{\sqrt{C_{i,i}(s,t)\,C_{j,j}(s,t)}} = \frac{\sigma_{i,j}\varrho(|s-t|)}{\sqrt{\sigma_{i,j}\varrho(|s-t|)\,\sigma_{i,j}\varrho(|s-t|)}} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,j}\sigma_{i,j}}}$$

Note 264. In Co-Kriging, using the Intrinsic Multivariate Correlation model for the cross-covariance, and using sites in a grid allows the use of Kronecker product operations in cases (17.3) and 17.4 for mitigating the computational requirements.

17.3. **Linear model of coregionalisation (LMC).**

- Perhaps we will not introduce this special concept this year due to lack of time.
- The linear model of coregionalisation, essentially introduces a special structure in the cross-covariance functions $\{C_{i,j}(s,t)\}$, and performs co-ckriging.
- Although a rather computational concept, the imposed cross-covariance structure is well justified in a reasonable/intuitive manner and covers a large set of applications/problems.
- One of the purposes of LMC is for instance for co-kriging to require more convenient calculations with regards to the cross-covariance matrices as resulted by the cross-covariance functions $\{C_{i,j}(s,t)\}$. For instance, imagine how upset your computer (particularly its CPU and its RAM) would be if you try to just implement co-kriging when your data sources are a lot (i.e. large k), and/or when for each data source the individual data-set was large (i.e. large n_j)... See the covariance matrix in (17.4) or the system of equations in 17.3.

17.3.1. *Intrinsic multivariate correlation.*

17.3.2. *The Linear model of coregionalisation.*