Spatio-temporal statistics (MATH4341)

Michaelmas term

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

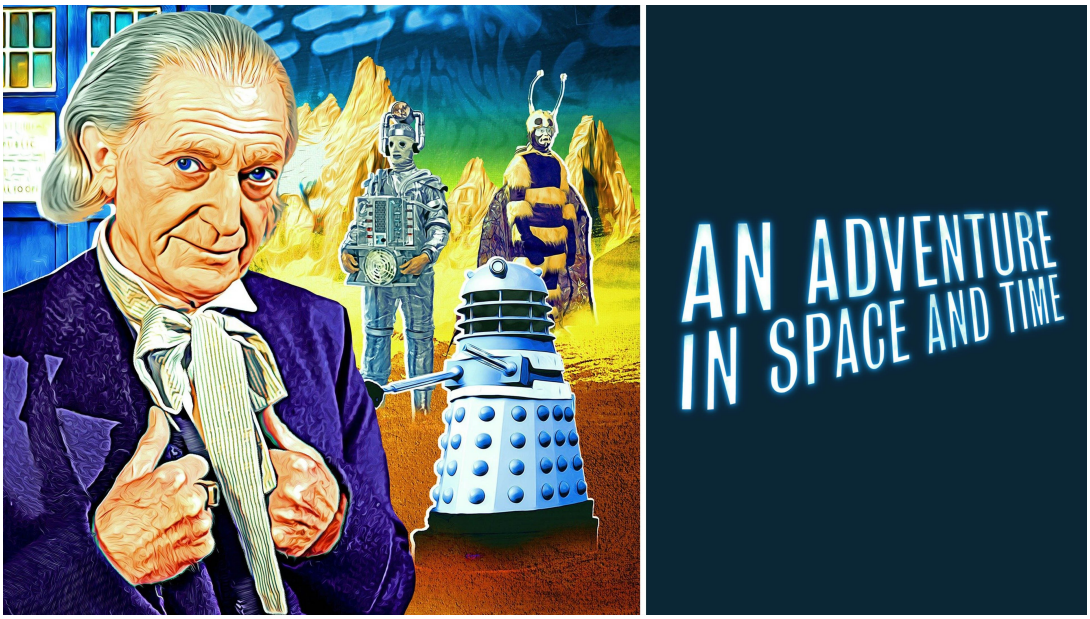
Department of Mathematical Sciences (Office MCS3088)
Durham University
Stockton Road Durham DH1 3LE UK

2023/11/08 at 18:44:31

Concepts

An introduction to spatial statistics:

- Reginalised statistical concepts
- Aerial unit data analysis
- Point referenced data analysis
- Point pattern data analysis
- Computational statistics (INLA)
- Implementation in R



Handouts

1. Handout 1: Types of spatial data
2. Handout 2: Computational methods
3. Handout 3: Point referenced data modeling / Geostatistics

Reading list

These lecture Handouts have been derived based on the above reading list.

Main texts:

- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
  - Our main reference book throughout the course. It covers all the three the spatial stats concepts we will introduce. Classic book in spatial statistics, but a bit outdated. Also very badly written.
- Gaetan, C., & Guyon, X. (2010). Spatial statistics and modeling (Vol. 90). New York: Springer.
  - Covers all the three the spatial stats concepts we will introduce but not all the details. Shorter and better written than than Cressie, N. (2015).

Supplementary textbooks for various types of spatial data:

- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. CRC press.
  - Covers all the three the spatial stats concepts we will introduce in a Bayesian manner. It requires some knowledge from multivatiare statistics, e.g. multivariate Normal distribution.
- Ripley, B. D. (2005). Spatial statistics. John Wiley & Sons.
  - Covers all the three the spatial stats concepts we will introduce. Classic book in spatial statistics, and perhaps one of the first, if not the first, textbook in the area, so outdated. It shows a good intuition in the concepts.
- Schabenberger, O., & Gotway, C. A. (2005). Statistical methods for spatial data analysis. CRC press.
  - I have not checked it yet... but I have heard that it is OK. Sorry.

Supplementary textbooks for Point reference data / Geostatistics:

- Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.
  - It covers the geostatistics /point referenced data part we will cover in advanced level, however it is easy to follow.

Supplementary textbooks for Areal data:

- TBD

Supplementary textbooks for Point pattern data:

- Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC press.
  - Major focus on [S5] –Notice that this concept may not be introduced due to the time restrictions

Supplementary textbooks for Software:

- Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.
  - It describes R packages for presentation, and visualization of spatial data sets, as well as related basic statistical inference. It does not discuss INLA.
- Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.
  - It demonstrates how to implement Integrated Nested Laplace Approximation methods for the three types of spatial stat we will introduce. It is easy to read and it has a good intro in general INLA method.
- Gómez-Rubio, V. (2020). Bayesian inference with INLA. CRC Press.
  - It demonstrate how to implement Integrated Nested Laplace Approximation methods in statistics in general (eg, regression, glmm, spatial & spatio temporal models).

Supplementary textbooks for Theory:

- Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
  - Covers theory / probabilities all the three the spatial stats concepts we will introduce.
- van Lieshout, M. N. M. (2019). Theory of spatial statistics: a concise introduction. CRC Press.
  - Covers theory / probabilities related all the three the spatial stats concepts we will introduce (however some theorems may not be included). It contains a subset of the material in Kent, J. T., & Mardia, K. V. (2022).

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 1: Types of spatial data

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

Aim. To introduce the types of spatial statistical data. To get a general idea about spatial statistics modeling.

Reading list & references:

- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
  - Chapter 1: pp 1- 28
- Datasets are available from:
  `https://github.com/georgios-stats/Spatio-Temporal_Statistics_Michaelmas_2023/tree/main/Datasets/`

1. Motivations

Note 1. Researchers in diverse areas such as geology, soil science, image processing, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science, climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are geographically referenced, and often presented as maps.

Note 2. In several problems, the data have a space (and time) label associated with them; this gives the motivation to the development and analysis of (not necessarily statistical) models that indicate when there is dependence between measurements at different locations.

Note 3. In an epidemiological investigation, for instance, one might wish to analyze lung, breast, colorectal, and cervical cancer rates by county and year in a particular state, with smoking, mammography, and other important screening and staging information also available at some level. Public health professionals who collect such data are charged not only with surveillance, but also statistical inference tasks, such as modeling of trends and correlation structures, estimation of underlying model parameters, hypothesis testing (or comparison of competing models), and prediction of observations at unobserved locations (and times).

Note 4. Spatial statistics is a branch of statistics that focuses on the analysis and modeling of data with inherent spatial relationships, by accounting for spatial dependencies and patterns to derive meaningful insights and make informed decisions.

Shall I ignore spatial dependence? –No!

Note 5. From your experimental design lectures, recall R. A. Fisher's principles of randomization, blocking and replication to neutralize (not remove) spatial dependence. In his

agricultural studies, he noticed that "After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart." To avoid the "confounding" of treatment effect Fisher properly introduced randomization, namely the controlled introduction of uncertainty.

Note 6. The First Law of Geography, according to Waldo Tobler, is "*everything is related to everything else, but near things are more related than distant things.*" Perhaps, we can paraphrase it by using stats terms to "nearby attribute values are more statistically dependent than distant attribute values".

Spatial data and spatial process.

Note 7. In spatial statistics, the basic components are data $\{Z_{s_1}, ..., Z_{s_n}\}$ observed at locations spatial locations $\{s_1, ..., s_n\}$. Classically, the locations are 2D, $s \in S \subset \mathbb{R}^2$, however it can be $S \subset \mathbb{R}^1$ (such as in chromatography applications), or $S \subset \mathbb{R}^3$ (such as in earth science, 3D imaging, etc) depending on the application. The locations $s_i \in S$ can be considered either (i.) fixed and hence used for training or (ii.) random and hence a quantity for inference. Yet, $\{s_i\}$ can be arranged irregularly in the space or regularly in a grid. Data $Z_{s_i} = Z(s_i)$ are random vectors.

Note 8. Let $s \in \mathbb{R}^d$ be a generic data location, and suppose the datum $Z(s)$ at spatial location s is an uncertain and hence random vector. Considering s to vary over index set $S \subset \mathbb{R}^d$ imposes a spatial random process (or multivariate random field)

$$\{Z(s); s \in S\}$$

which can be modeled as a stochastic process (to be defined later.).

Note 9. In spatial problems, spatial data $\{Z_{s_i}\}_{i=1}^n$ at locations $\{s_i\}_{i=1}^n$ are assumed to be realizations of a spatial process (or a multivariate random field)

$$\{Z(s); s \in S\}, \tag{1.1}$$

indexed by a spatial set $S \subset \mathbb{R}^d$.

2. Principal spatial statistics areas

Note 10. We can characterize the spatial statistical problems according to the type of measurement, their specified (assumed) stochastic generating mechanism, and the choice of the spatial locations. In principle, each of them is associated to different motivations, statistical/scientific problems, statistical tools, however, modern applications/problem may involve

characteristics from a combination of them. Here, we will study three of spatial statistical areas.

2.1. **Point referenced data (Geostatistics).**

Note 11. Climate or environmental data are often presented in the form of a map, for example the maximum temperatures on a given day in a country, the concentrations of some pollutant in a city or the mineral content in soil. In mathematical terms, such maps can be described as realisations from a random field, that is, an ensemble of random quantities indexed by points in a region of interest. The aim is usually interpolation, and the associated statistical inference.

Note 12. Such data were first analyzed in geological sciences. Hence, for historical reasons, this area of spatial statistics is often called Geostatistics and the point referenced data are also called geocoded or geostatistical data.

Note 13. Mathematically speaking, the spatial domain S is a continuous fixed subset of $\mathbb{R}^d$ that contains a d-dimensional rectangle of positive volume. The datum $Z(s)$ is a random vector (outcome) at specific location $s \in S$ which can vary continuously over domain S. In practice, the actual data are observations $\{Z(s_i)\}_{i=1}^n$ at n (finite number) fixed locations $\{s_i\}_{i=1}^n \subset S$. The locations $\{s_i\}$ are fixed and can be arranged irregularly in the space or regularly as a grid.

Note 14. Geostatistics aims to answer questions about modeling, identification and separation of small and large scale variations, prediction at unobserved locations and reconstruction of the spatial process $Z(s)$ across the whole space S.

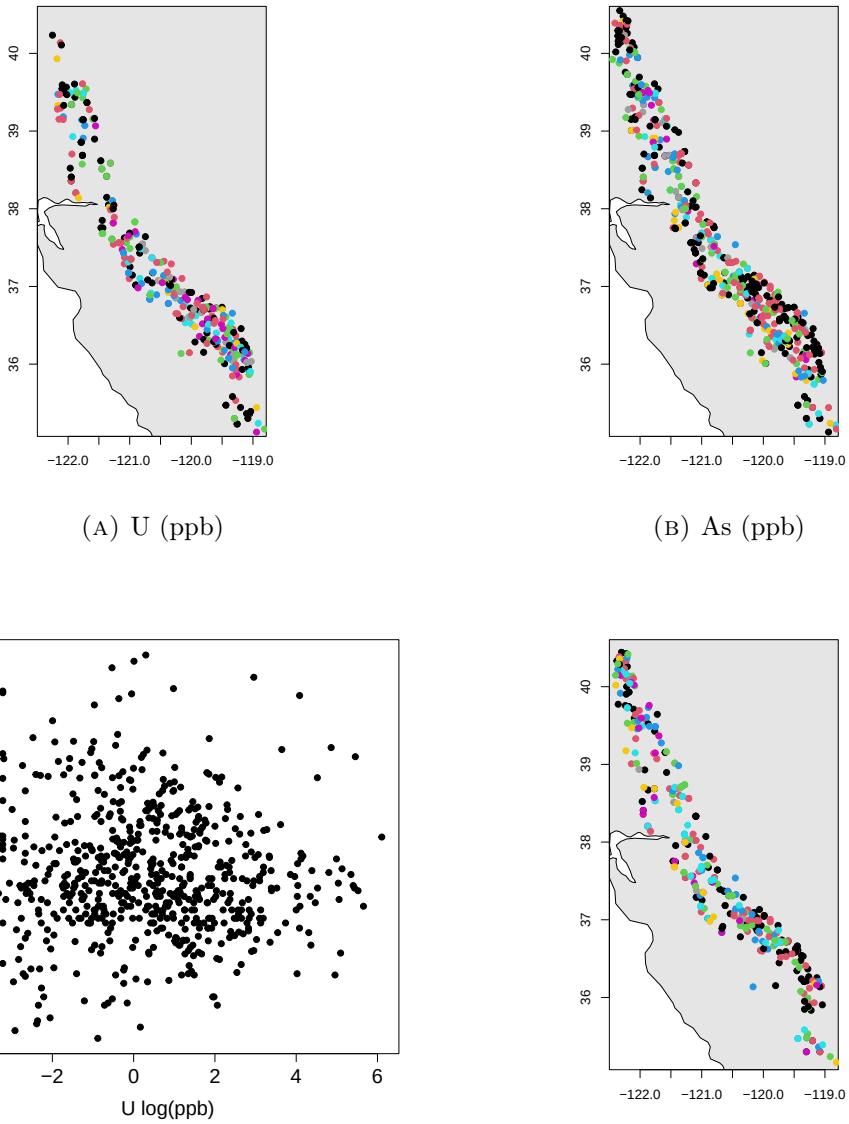
Example 15. (Ground water pollution in the Central Valley of California[1]) California's Central Valley is one of the most productive agricultural regions in the world. With an increase in population, groundwater consumption is expected to increase. Agricultural irrigation heavily draws on the groundwater system. Pumping from increasingly deeper parts of the aquifer has increased the rate of downward groundwater flow, which have been linked to the release of, for example, uranium. The question therefore concerns how we can maintain groundwater quality while dealing with this increased need for it. Understanding this trade-off is key to sustainable groundwater management. Simply increasing groundwater by supply, for example through a process of recharge (e.g., flooding a field), may affect its quality. It may lead to an increased introduction of contaminants such as pesticides. In

[1]Fakhreddine, S., Babbitt, C., Sherris, A., et al. (2019). Protecting Groundwater Quality in California, Management Considerations for Avoiding Naturally Occurring and Emerging Contaminants. Environmental Defense Fund. `www.edf.org/sites/default/files/documents/groundwater-contaminants-report.pdf`

other words, groundwater management actions may have unintended consequences. It is important to understand how groundwater quality is affected by any management actions on, for example, seawater intrusion, land subsidence, or declining water levels. A substantial amount of geochemical analysis has been collected from existing wells. Interest lies in understanding the important processes, either natural or anthropogenic, that cause variation in these data. We may find a certain "signature" or "patterns" in the data that can determine the process of contamination in a particular area. For example, the Central Valley has what are termed "geogenic" contaminants, which means it has arsenic (As), chromium (Cr), uranium (U), which you don't want to drink, naturally occurring. A simple analysis could look for high levels of these elements, indicating possible anthropogenic contamination, although they may be naturally occurring. The point we wish to make is that a signature of a feature is more than a single high value. What we are looking for is a combination of elements and of a certain composition. Some scientific questions involve:

(1) What combination of elements are indicative of a human impact in water quality versus a natural occurrence?
(2) What caused this impact? Agriculture? Pollution?
(3) Where in the Central Valley can we find these combinations of elements, thereby informing mitigation action?

Figure 2.1c presents the scatter plot of As and U in a naive manner as it ignores spatial dependency. Figures 2.1d, 2.1b, and 2.1a show the Groundwater concentration (parts per billion [ppb]) of chromium (Cr), Arsenic (As), Uranium (U) from January 2018 to January 2019. The point coordinates are the geographical locations, and the color denotes the value of the corresponding values of Cr, As, U in ppb. The locations are 2D, fixed/known and hence part of the training observations. The locations are irregularly scattered/spaced and hence not on a regular grid of points. As s are coordinates, they vary continuously over the spatial domain which is Central Valley in CA/USA. The quantity of interest $\{Z(s_i)\}$ is a random vector whose elements are the concentrations of As, U, Cr, etc. labeled by the coordinates s (locations). The spatial statistician's task may involve producing statistical models able to provide predictive inference for quantities Cr, As, U (and others) at unseen/unobserved locations. Obviously, special dependencies should be taken into account in the model, e.g. the concentration of U in two neighboring cities is expected to be more similar than two far distant cities. As seen latter this gives rise to a 'regionalized statistical analysis'. Along with the spatial dependency, and in the same model, it would be wise to take into account the dependency (e.g. correlation) between different variables, such as As and U. As seen latter this gives rise to a 'co-regionalized statistical analysis'.



(A) U (ppb)

(B) As (ppb)

(C) U vs As (log ppb)

(D) Cr (ppb)

FIGURE 2.1. Ground water pollution in the Central Valley of California

Example 16. (Coal ash dataset in Pennsylvania) Figure 2.2 shows 208 coal ash core measurements/samples collected on a regular grid of points in the Robena Mine in Greene County, Pennsylvania. The percentage of coal ash at the sampled locations is denoted by the colorbar. The sampled locations $\{s_i\}$ are fixed, and regularly spaced in a grid. As s are coordinates, they vary continuously over the spatial domain which is Robena Mine. The quantity of interest is the percentage of ash coal at these locations $\{Z(s_i)\}$. A mining

engineer could be interested in predicting the ash distributions and the washability characteristics of coal along a seam in advance of mining. A spatial statistician would be able to produce a statistical model to predict ash concentrations between sampled points as well as quantify related uncertainties. Once a reasonable model that accounts for both the global trends and the local dependencies in the data is found and validated, the mining engineer could proceed to try and fill in the gaps, in other words, to estimate the percentage of coal ash at missing grid points based on the sampled percentages.

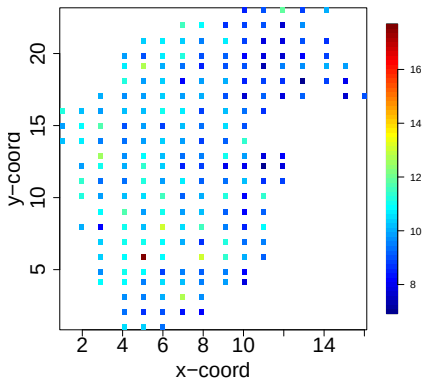


FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

Example 17. (Air pollution in Piemonte.) Figure 2.3 presents the average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte region (Northern Italy). The data (measurements) are at fixed locations at irregular grid points. PM10 is one of the most troublesome pollutants in the area. Environmental agencies need models to predict PM10 at unmonitored sites in order to assess PM10 concentration over an entire region. A geostatistician can build a model which is satisfactory in terms of goodness of fit, interpretability, parsimony, prediction capability and computational costs with purpose to build reliable PM10 concentration maps, equipped with the corresponding uncertainty measure.

2.2. **Aerial unit data / spatial data on lattices.**

Note 18. Sometimes observations are collected over areal units such as pixels, census districts, or tomographic bins. In such cases, the random field models $\{Z(s); s \in S\}$ have a discrete index set S. The aims are usually, noise removal from an image and smoothing rather than interpolation.

Note 19. Mathematically speaking, the index set S of the data $\{Z(s)\}$ is a fixed (not random) and finite collection of points (locations) $s \in S$. The locations $s \in S$ can be irregular or

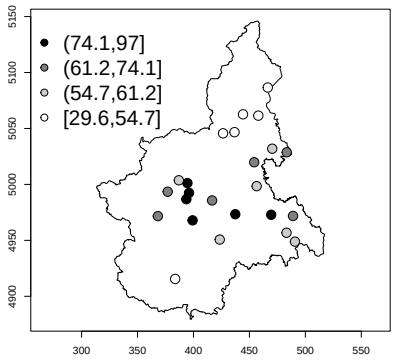


FIGURE 2.3. (Air pollution data) Average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte

arranged in a regular grid. Often, there is a natural adjacency relation or neighborhood structure. Often, datum $Z(s)$ is a random vector at location $s \in S$ and it represents an integral or average of the quantity of interest over some region represented by $s \in S$.

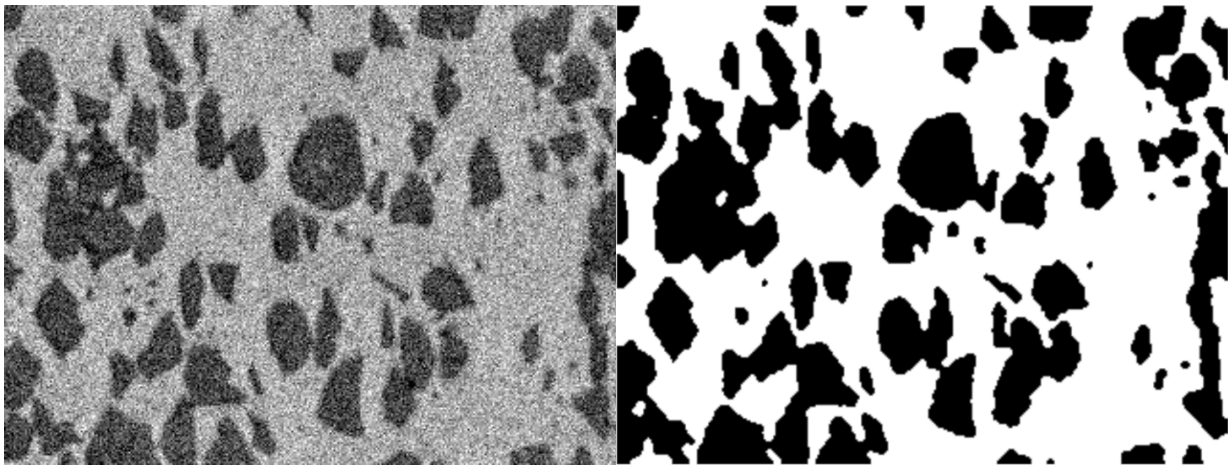
Example 20. In image processing, S may be a grid of pixels (locations are fixed and regular).

Example 21. In a UK epidemological study, S may be the centroids of the UK counties, and $Z(s)$ may represent the average value of a characteristic in county s.

Example 22. In statistical physics, S may be a collection of atoms and genuinely finite (locations are fixed and regular).

Example 23. (Image restoration data) Figure 2.4a shows an (observed) image from a gray-scale photo-micrograph of the micro-structure of the Ferrite-Pearlite steel obtained by PNNL's project supported by DoE. The lighter part is ferrite while the darker part is pearlite. We focus our analysis on the first quarter fragment of size 240×320 pixels (red frame). This image is contaminated by noise due to the instrument errors. Interest lies in removing the noise (denoising) and recovering the real image. Figure 2.4b shows the restored image after appropriate statistical processing. Here the locations are pixels arranged in a fixed regular grid (hence discrete and not continues). The each observation $Z(s)$ is the color of a pixel s; here it is scalar as the observed pixels are in tones of grey, however it could be a 3D if the pixels were colored.

Example 24. (North Carolina SIDS data set) Figure 2.5a shows the total number of deaths from Sudden Infant Death Syndrome (SIDS) in 1974 for each of the 100 counties in North Carolina. Figure 2.5b shows the corresponding live births in each county and same period. This is the R data set `nc{spdep}`. The centroids of the counties do not lie on a regular grid.



(A) Noised image (B) Reconstructed image

FIGURE 2.4. Ferrite-Pearlite steel image (Image restoration)

The sizes and shapes of the counties vary and can be quite irregular. The recorded counts are not tied to a precise location but tallied up county-wise. This kind of accumulation over administrative units is usual for privacy-sensitive data in, for instance, the crime or public health domains. A public health official could be interested in spatial patterns; e.g., whether or not there are clusters of counties with a high incidence of SIDS, or areas where the SIDS counts are higher than what would be expected based on the number of live births in the area. Perhaps, we can eyeball the figures and see that there is a higher SIDS rate in the north-east areas compared to the north-west with similar birth numbers. A statistician can develop a statistical model providing inference about such questions.

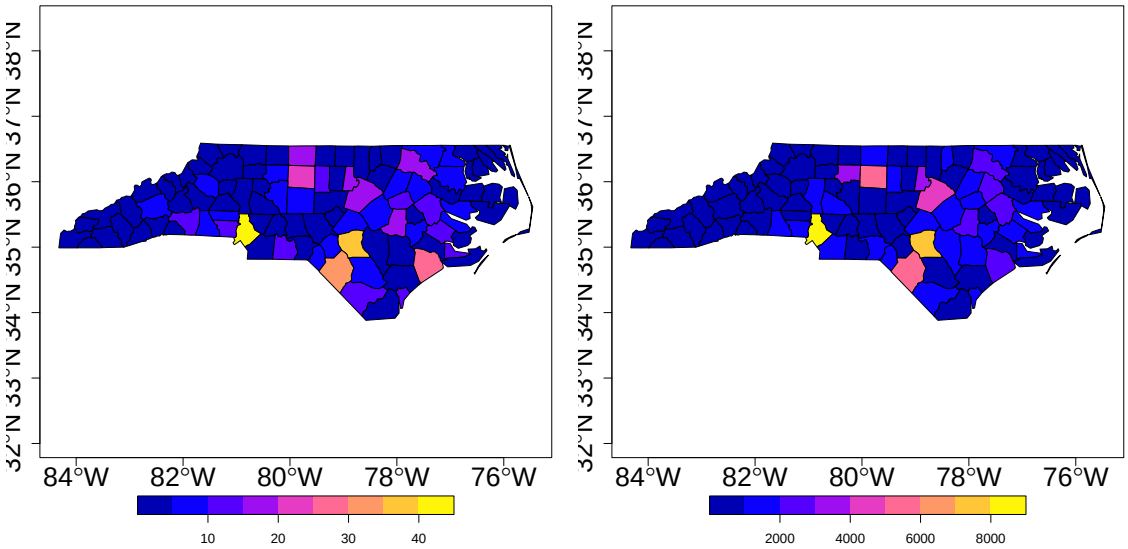
2.3. **Spatial point pattern data.**

Note 25. Sometimes the locations at which events occur are random. Typical examples include outbreaks of forest fires, or the epicentres of earthquakes. Such random patterns of locations are said to form a point process.

Note 26. Rigorously, the spatial domain S is a random set of points; specifically a point process, in $\mathbb{R}^d$ at which some events happened.

Note 27. In the most general case, $Z(s)$ is a random vector at location $s \in S$ (eg other covariates are associated to the location s); these covariates are called marked variables. We will refer to it as a Marked spatial point process.

Note 28. In the simplest case, no covariate for Z is specified, and hence $Z(s)$ represents only the occurrences of an even at location s, one could think of the data taking scalar values



(A) Numbers of SIDS (B) Live birth counts

FIGURE 2.5. North Carolina SIDS data (Aerial unit data)

$Z(s) = 1$ or $Z(s) = 0$ when the event has occurred or not for all $s \in S$. We will refer to it as a spatial point process

Note 29. Questions in the spatial point pattern problems are mainly whether the pattern of locations is exhibiting complete spatial randomness, clustering, or regularity. In the marked spatial point process where additional covariates are measured, we could possibly investigate the factors/variables associated to this behavior as well. A statistical approach to address such questions is needed as different observers may disagree on the amount of clustering or randomness. Usually patters from a completely random process may appear to be wrongly clustered when just eyeballed by an individual.

Example 30. (Tropical rain forest trees in Barro/Colorado) Figure 2.6 shows the positions (dots) of 3605 Beilschmiedia trees in a 1000×5000 meter rectangular stand in a tropical rain forest at Barro Colorado Island, Panama. All spatial coordinates are in the Cartesian coordinate system and in meters. Dataset is available from the R package `bei{spatstat}`. The scientific question may be if the trees are distributed over the area in a uniform way, they form clusters, or they are arrange in a specific pattern. Here, the locations of the dots/trees are not fixed but random/uncertain and of course they are matter of inference. This is a point process as each location is associated to an occurrence only and not any other covariate. The statistician's task is to design models able to test and quantify heterogeneity/homogeneity.

Example 31. (Longleaf Pines Point Pattern) Figure 2.7 shows locations (as Cartesian coordinates) and relative diameters at breast height in dbh (as the size of the dot) of all longleaf

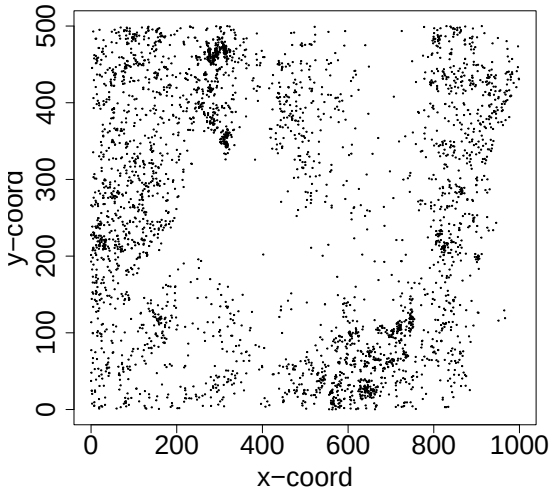


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

pine trees in the 24ha region of the Wade Tract, an old-growth forest in Thomas County, Georgia in 1979. Dataset is available from R package `bei{spatstat}`. Longleaf pine is a fire-adapted species of trees. The domain scientist is interested in knowing whether the spatial locations are spatially random, or clustered, if large (small) trees cluster and how do large and small trees interact. A statistician can design models able to quantify such notions and provide inference. Here, the locations are random (not fixed) and in fact an object of inference. The diameter at breast height recorded along with the tree's location is the marked variable, and hence, the whole process is a marked process.

3. Uncertainty quantification and modeling

To be defined rigorously later

Note 32. In spatial problems, uncertainty is expressed probabilistically through a spatial stochastic process (or a multivariate random field), which can be written most generally as

$$\{Y(s); s \in \mathscr{S}\}, \tag{3.1}$$

Here $Y(s)$ is the random attribute value at location s, $\mathscr{S} \subset \mathbb{R}^d$ is a subset of $\mathbb{R}^d$ ($d = 1, 2, 3$), contained in $\mathscr{S}$ is a possibly random fixed or random set S that indexes those parts of $\mathscr{S}$ relevant to the scientific study.

Spatial process model.

Note 33. The scientific uncertainty (i.e. the (known) uncertainty about the scientific problem) is expressed via the spatial process model. E.g., uncertainty about the real picture in Fig. 2.4a.

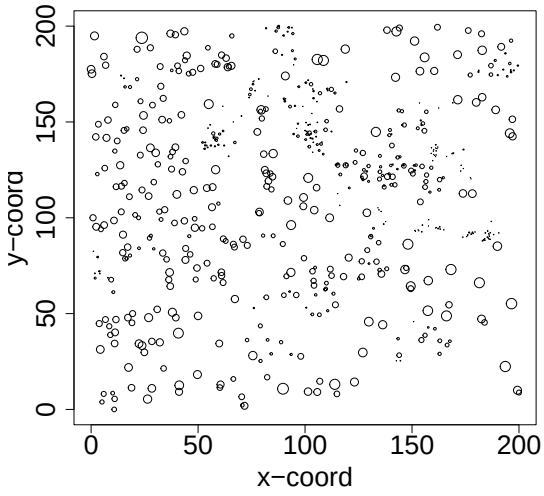


FIGURE 2.7. Longleaf Pines Point Pattern (Spatial point data)

Note 34. This spatial stochastic process can be a: geostatistical process, lattice process, or point process depending on the principal spatial statistical area (Section 2) the application is associated with.

Note 35. The joint probability model defined by the random $\{Y(s); s \in S\}$ is

$$\text{pr}(Y, S) = \text{pr}(Y|S)\,\text{pr}(S) \tag{3.2}$$

Note 36. The specification of $\text{pr}(S)$ represents the three principal spatial statistical areas. E.g., for spatial data on lattices or point referenced data problems were the locations are fixed and not uncertain, we can consider $\text{pr}(Y, S) = \text{pr}(Y|S)$ with $\text{pr}(S) = 1_{\{S\}}(S)$ and hence ignore $S$and $\text{pr}(S)$ from the notation.

Data model.

Note 37. The measurement uncertainty is quantified via the data model. E.g. the "noisy image" in Fig. 2.4a.

Note 38. The data model is specified to be the conditional distribution of the data Z given the spatial stochastic process Y and the S, namely

$$\text{pr}(Z|Y, S) \tag{3.3}$$

Note 39. If the data are assumed to be conditionally independent, such as $Z(s) \perp Z(s')|Y, S$ then

$$\text{pr}(Z|Y, S) = \prod_{i=1}^{n} \text{pr}(Z(s_i)|Y, S) \tag{3.4}$$

Note 40. The spatial statistical dependence of in Z, articulated by the First Law of Geography, follows by

$$\operatorname{pr}(Z|S) = \int \operatorname{pr}(Z|Y,S)\operatorname{pr}(Y|S)\,\mathrm{d}Y$$

The hierarchical statistical model.

Note 41. To sum up the (known) uncertainty in spatial the statistics problem is expressed via the so called Hierarchical spatial model

$$\begin{cases} Z|Y,S & \text{data model} \\ Y,S & \text{spatial process model} \end{cases} \tag{3.5}$$

with

$$\operatorname{pr}(Z,Y,S) = \operatorname{pr}(Z|Y,S)\operatorname{pr}(Y|S)\operatorname{pr}(S)$$

The Empirical (Bayes) hierarchical model.

Note 42. Often the decomposition (3.5) is parametrized with respect to unknown parameters $\theta \in \Theta$ we wish to learn given the observables; this is often called the Empirical hierarchical model i.e.

$$\begin{cases} Z|Y,S,\theta & \text{data model} \\ Y,S|\theta & \text{spatial process model} \end{cases} \tag{3.6}$$

with

$$\operatorname{pr}(Z,Y,S|\theta) = \operatorname{pr}(Z|Y,S,\theta)\operatorname{pr}(Y|S,\theta)\operatorname{pr}(S|\theta)$$

...more details in the next lecture.

The Bayesian hierarchical model.

Note 43. In Bayesian statistics, the hierarchical model in (3.5) is completed by the $\theta \sim \operatorname{pr}(\cdot)$ adding a third layer leading to the Bayesian hierarchical model

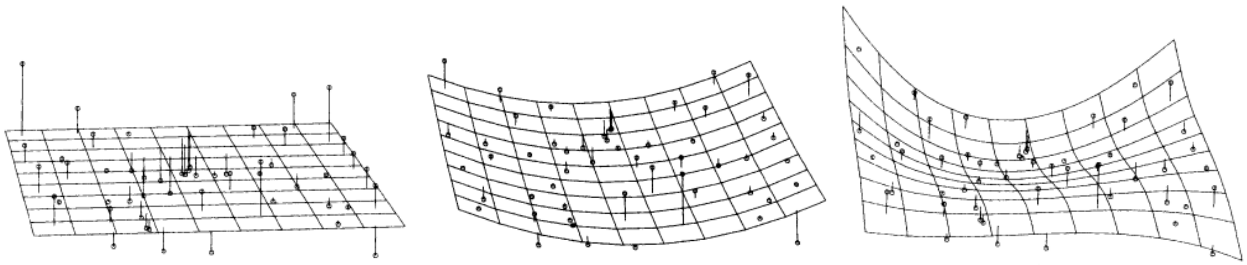
$$\begin{cases} Z|Y,S,\theta & \text{data model} \\ Y,S|\theta & \text{spatial process model} \\ \theta & \text{hyper-parameter prior model} \end{cases} \tag{3.7}$$

with

$$\operatorname{pr}(Z,Y,S,\theta) = \operatorname{pr}(Z|Y,S,\theta)\operatorname{pr}(Y|S,\theta)\operatorname{pr}(S|\theta)\operatorname{pr}(\theta)$$

Exercise 44. (A naive example) Consider Example 15, and observations $\{(Z_i, s_i)\}_{i=1}^n$ where Z_i is the Cr measurement in ppb at the i-th location $s_i \in \mathbb{R}^2$. Perhaps one may consider that the real Cr, lets denoted as Y, may follow a Normal distribution with a mean $\mu = S\beta$

That's a snapsot for what follows.



(A) (B) (C)

FIGURE 3.1. Examples representing the hierarchical spatial model (3.8) for different values of $\theta = \left(\sigma^2, \beta, \phi\right)$

parametrized as $[\mu]_i = \beta_0 + s_{(1),i}\beta_1 + s_{(2),i}\beta_2 + s_{(1),i}s_{(2),i}\beta_{12} + ...$ at a location s (to consider spatial dependence) with some unknown parameter β, and covariance matrix parametrized as $[C]_{i,j} = c\left(s_i, s_j\right)$ with $c\left(s, s'\right) = \sigma^2 \exp\left(-\phi \left\|s - s'\right\|\right)$; here β, ϕ, and σ^2 are unknown parameters. One may consider that the measurements Z at each location are the result of observing Y_i (the real Cr) but contaminated by additive random noise, as $Z_i = Y_i + \epsilon_i$, $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$. To sum up, we have build the hierarchical model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}\left(Y, I\sigma^2\right) \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}\left(S\beta, C\right) \end{cases} \tag{3.8}$$

Figure 3.1 shows the hierarchical spatial model (3.8) for different values of $\theta = (\sigma^2, \beta, \phi)$; the surface corresponds to the spatial process $\left\{Y\left(s\right); s \in \mathbb{R}^2\right\}$ and is presented at three different instances each of them with different values for (β, ϕ), while the dots correspond to the observations $\left\{\left(Z\left(s_i\right), s_i\right)\right\}_{i=1}^n$ and their deviation from the spatial process is controlled by σ^2 . If we work on the fully Bayesian framework (!!!), we can complete the model with priors on $\theta = (\sigma^2, \beta, \phi)$ as $\sigma^2 \sim \mathrm{IG}\left(\kappa_\sigma, \lambda_\sigma\right)$, $\phi \sim \mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right)$, and $\beta \sim \mathrm{N}\left(b, Iv\right)$, with some known hyper-parameters $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$. To sum up, we have build the Bayesian model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}\left(Y, I\sigma^2\right) \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}\left(S\beta, C\right) \\ \beta \sim & \mathrm{N}\left(b, Iv\right) \\ \sigma^2 \sim & \mathrm{IG}\left(\kappa_\sigma, \lambda_\sigma\right) \\ \phi \sim & \mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right) \end{cases}$$

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 2: Introduction to INLA & R-INLA

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

Aim. To introduce Laplace approximation, and Integrated Laplace Approximation computational methods. To introduce

Reading list & references:

(1) Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.
- Ch. 4.6-4.9; pp.104-126

(2) Turkman, M. A. A., Paulino, C. D., & Müller, P. (2019). Computational Bayesian statistics: an introduction (Vol. 11). Cambridge University Press.
- Ch. 8

(3) Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 71(2), 319-392.

1. Laplace Approximation (LA)

Proposition 1. *Consider integral*

$$I = \int \exp\left(nL\left(\theta\right)\right) d\theta$$

where $\theta \in \mathbb{R}^d$. *Laplace approximation (LA) method produces approximation* $I \approx \hat{I}$

$$\hat{I} = \left(2\pi\right)^{\frac{d}{2}} \left(n\right)^{-\frac{d}{2}} \left(\det\left(\Sigma\right)\right)^{\frac{1}{2}} \exp\left(nL\left(\hat{\theta}\right)\right)$$

where $\hat{\theta}$ *is the maximum of* $L\left(\cdot\right)$ *and* $\Sigma = -\left(H\left(\hat{\theta}\right)\right)^{-1}$ *with Hessian* $H\left(\hat{\theta}\right) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(L\left(\theta\right)\right)\Big|_{\theta=\hat{\theta}}$.

Proof. Sketch of the proof. Take 2nd order Taylor expansion of $L\left(\theta\right)$ around $\hat{\theta}$ i.e.

$$L\left(\theta\right) \approx L\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right) \underbrace{\nabla L\left(\hat{\theta}\right)}_{=0} + \frac{1}{2}\left(\theta - \hat{\theta}\right)^{\top} H\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right) \tag{1.1}$$

then

$$\begin{aligned} I &\approx \int \exp\left(nL\left(\hat{\theta}\right) + n\left(\theta - \hat{\theta}\right)^{\top} H\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right)\right) \mathrm{d}\theta \\ &= \exp\left(nL\left(\hat{\theta}\right)\right) \int \exp\left(-\frac{1}{2}\left(\theta - \hat{\theta}\right)^{\top} \left(\left(-nH\left(\hat{\theta}\right)\right)^{-1}\right)^{-1} \left(\theta - \hat{\theta}\right)\right) \mathrm{d}\theta \\ &= \exp\left(nL\left(\hat{\theta}\right)\right) \left(2\pi\right)^{\frac{d}{2}} \left(\det\left(\left(-nH\left(\hat{\theta}\right)\right)^{-1}\right)\right)^{\frac{1}{2}} \end{aligned}$$

Given regularity conditions related to the Taylor expansions (1.1), it can be shown that $I = \hat{I}\left(1 + O\left(n^{-1}\right)\right)$ (not discussed here). □

Example 2. Consider posterior expectation

$$\mathrm{E}\left(g\left(\theta\right)|z\right) = \int g\left(\theta\right) \mathrm{pr}\left(\theta|z\right) \mathrm{d}\theta \tag{1.2}$$

of a function $g\left(\cdot\right)$ of the parameter $\theta \in \mathbb{R}^d$ given observables z. Laplace method can produce approximation$\mathrm{E}\left(g\left(\theta\right)|z\right) \approx \widehat{\mathrm{E}\left(g\left(\theta\right)|z\right)}$

$$\widehat{\mathrm{E}\left(g\left(\theta\right)|z\right)} = \left(\frac{\det\left(\Sigma^*\right)}{\det\left(\Sigma\right)}\right)^{\frac{1}{2}} \exp\left(n\left(L^*\left(\hat{\theta}^*\right) - L\left(\hat{\theta}\right)\right)\right) \tag{1.3}$$

where $\hat{\theta}$ and Σ are the mode and minus the inverse Hessian of $L\left(\theta\right) = \log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)/n$ while $\hat{\theta}^*$ and Σ^* are the mode and minus the inverse Hessian of $L^*\left(\theta\right) = \log\left(g\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)/n$.

Solution. (Sketch of the solution) It is

$$\mathrm{E}\left(g\left(\theta\right)|z\right) = \frac{\int g\left(\theta\right)\mathrm{pr}\left(z|\theta\right)\mathrm{pr}\left(\theta\right)\mathrm{d}\theta}{\int \mathrm{pr}\left(z|\theta\right)\mathrm{pr}\left(\theta\right)\mathrm{d}\theta} = \frac{\int \exp\left(nL^*\left(\theta\right)\right)\mathrm{d}\theta}{\int \exp\left(nL\ \ \left(\theta\right)\right)\mathrm{d}\theta} \overset{(\star)}{\approx} \frac{\left(2\pi n\right)^{d/2}\sqrt{\det\left(\Sigma^*\right)}\exp\left(nL^*\left(\hat{\theta}^*\right)\right)}{\left(2\pi n\right)^{d/2}\sqrt{\det\left(\Sigma\ \right)}\exp\left(nL\left(\hat{\theta}\ \right)\right)}$$

where $(\star)$ is by applying Proposition 1 once at the top and once at the bottom of the fraction. Then the result follows.

Under regularity conditions related to Taylor expansion (not discussed here), it is $\mathrm{pr}\left(\theta_1|z\right) = \widehat{\mathrm{pr}\left(\theta_1|z\right)}\left(1 + O_{\theta_1}\left(n^{-1}\right)\right)$ where the lower index indicates the dependence of the constant on θ_1.

Example 3. Consider the marginal posterior density of $\theta_1 \in \mathbb{R}$

$$\mathrm{pr}\left(\theta_1|z\right) = \int \mathrm{pr}\left(\theta_1, \theta_2|z\right)\mathrm{d}\theta_2 \tag{1.4}$$

under a Bayesian model with observable $z \sim \mathrm{pr}\left(z|\theta\right)$ and unknown parameter $\theta = \left(\theta_1, \theta_2\right) \in \mathbb{R}^d$ with $\theta \sim \mathrm{pr}\left(\theta\right)$. Laplace method can produce approximation

$$\widehat{\mathrm{pr}\left(\theta_1|z\right)} = \left(\frac{\det\left(\Sigma^*\left(\theta_1\right)\right)}{2\pi n \det\left(\Sigma\right)}\right)^{\frac{1}{2}} \frac{\mathrm{pr}\left(\theta_1, \hat{\theta}_2\left(\theta_1\right)\right)\exp\left(\log\left(\mathrm{pr}\left(z|\theta_1, \hat{\theta}_2\left(\theta_1\right)\right)\right)\right)}{\mathrm{pr}\left(\hat{\theta}\right)\exp\left(\log\left(\mathrm{pr}\left(z|\hat{\theta}\right)\right)\right)} \tag{1.5}$$

where $\hat{\theta}$ is the maximizer of $\log\left(\mathrm{pr}\left(\theta_1, \theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1, \theta_2\right)\right)$,

Σ is the minus Hessian of $n^{-1}\left(\log\left(\mathrm{pr}\left(\theta_1, \theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1, \theta_2\right)\right)\right)$,

$\hat{\theta}_2\left(\theta_1\right)$ is the maximizer of $\log\left(\mathrm{pr}\left(z|\theta_1, \cdot\right)\right) + \log\left(\mathrm{pr}\left(\theta_1, \cdot\right)\right)$,

$\Sigma^*\left(\theta_1\right)$ is the minus Hessian of $n^{-1}\left(\log\left(\mathrm{pr}\left(z|\theta_1, \cdot\right)\right) + \log\left(\mathrm{pr}\left(\theta_1, \cdot\right)\right)\right)$

Solution. (Sketch of the solution) It is

$$\mathrm{pr}\left(\theta_1|z\right) = \frac{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta_2}{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta} = \frac{\int \exp\left(nL^*_{\theta_1}\left(\theta_2\right)\right)\mathrm{d}\theta_2}{\int \exp\left(nL\ \ \left(\theta\right)\right)\mathrm{d}\theta}$$

$$\overset{(\star)}{\approx} \left(\frac{\det\left(\Sigma^*\left(\theta_1\right)\right)}{2\pi n \det\left(\Sigma\right)}\right)^{\frac{1}{2}} \frac{\mathrm{pr}\left(\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\exp\left(\log\left(\mathrm{pr}\left(z|\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\right)\right)}{\mathrm{pr}\left(\hat{\theta}\right)\exp\left(\log\left(\mathrm{pr}\left(z|\hat{\theta}\right)\right)\right)}$$

where $L^*_{\theta_1}\left(\theta_2\right) = n^{-1}\left(\log\left(\mathrm{pr}\left(\theta_1,\theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1,\theta_2\right)\right)\right)$ and $L\left(\theta\right) = n^{-1}\left(\log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)\right)$. Here $(\star)$ results by applying Proposition 1 once at the top and once at the bottom of the fraction. Then the result is implied.

Under regularity conditions related to Taylor expansion (not discussed here), it is $\mathrm{pr}\left(\theta_1|z\right) = \widehat{\mathrm{pr}\left(\theta_1|z\right)}\left(1 + O_{\theta_1}\left(n^{-1}\right)\right)$ where the lower index indicates the dependence of the constant on θ_1.

2. Integrated Nested Laplace Approximation (INLA)

2.1. **Motivations.**

Note 4. Integrated Nested Laplace Approximation (INLA) can directly compute very accurate approximations to posterior marginals and summary statistics of statistical models with a specific type (such as those discussed in the module) even if they are high-dimensional or involve large datasets. In such models, MCMC methods may need hours or days to run, which INLA can provide more precise estimates in seconds or minutes for a certain type of models we will discuss.

2.2. **Where it can be applied; implementations.**

Note 5. INLA is suitable to facilitate Bayesian inference in spatial statistical problems related to Latent Gaussian Models (LGM).

Note 6. The class of Latent Gaussian Models (LGM) can be represented in a three level hierarchical model structure. The first level is the sampling model where the observations $z = \left(z_1,...,z_n\right)^\top$ can be assumed to be conditionally independent, given a latent random field $y = \left(y_1,...,y_n\right)^\top$ and hyper-parameter θ_1, i.e.

$$z|y,\theta_1 \sim \mathrm{pr}\left(z|y,\theta_1\right) = \prod_{i=1}^{n}\mathrm{pr}\left(z_i|y_i,\theta_1\right). \tag{2.1}$$

The second level assumes that y follows a multivariate Gaussian distribution (Essentially a Gaussian random field) given hyper-parameter θ_2 , i.e.

$$y|\theta \sim \mathrm{N}\left(\mu\left(\theta_2\right),\left(Q\left(\theta_2\right)\right)^{-1}\right) \tag{2.2}$$

The third level (relevant only to fully Bayesian statistical models) specifies a prior on the unknown parameter $\theta = (\theta_1, \theta_2)^\top$, i.e.

$$\theta \sim \text{pr}(\theta)$$

Assumption 7. *For the computational purposes of INLA, we make assumption that (2.2) is defined wrt an undirected graph* $\mathcal{G} = (\mathcal{V} = \{1, ...n\}, \mathcal{E})$ *such that*

$$y_l \perp y_m | y_{-\{l,m\}}, \quad \forall \{l, m\} \notin \mathcal{E} \tag{2.3}$$

This leads to sparse precision matrix $Q(\theta_2)$ *because*

$$y_l \perp y_m | y_{-\{l,m\}} \Leftrightarrow [Q(\theta_2)]_{l,m} = 0$$

This makes (2.2) be a Gaussian Markov Random Field (GMRF).

Note 8. The LGM (under consideration) is summarized to

$$\begin{aligned} z|y, \theta &\sim \text{pr}(z|y, \theta) = \prod_{i=1}^{n} \text{pr}(z_i|y_i, \theta) && \text{(sampling model for } z\text{)} \\ y|\theta &\sim \text{pr}_{\mathcal{G}}(y|\theta) && \text{(GMRF prior for } y\text{)} \\ \theta &\sim \text{pr}(\theta) && \text{(hyperprior for } \theta\text{)} \end{aligned} \tag{2.4}$$

Note 9. The joint posterior probability model is

$$\begin{aligned} \text{pr}(y, \theta|z) &\propto \prod_{i=1}^{n} \text{pr}(z_i|y_i, \theta) \, \text{pr}(y|\theta) \, \text{pr}(\theta) \\ &\propto \exp\left(-\frac{1}{2} (y - \mu(\theta))^\top Q(\theta) (y - \mu(\theta)) + \sum_{i=1}^{n} \log(\text{pr}(z_i|y_i, \theta)) \right) \text{pr}(\theta) \end{aligned} \tag{2.5}$$

and hence there is interest in computing the marginal densities and expectations of $y_i|z$, and $\theta_i|z$ as well as predictions of unseen y's.

Assumption 10. *For INLA to perform most efficiently (fast) and accurately (due to approximations), we make the following critical assumptions:*

1. *The number of hyperparameters* θ *is small, typically* 2 *to* 5*, but not exceeding* 20*.*
2. *pr*$(y|\theta)$ *is required to be a GMRF (or close to one) when the dimension* n *is high* (103–105)*.*
3. *The data* $\{z_i\}$ *are mutually conditionally independent of* y *and* θ*, implying that each observation* z_i *only depends on one component of the latent field, for example,* y_i*. Most components of* y_i *will not be observed.*

Note 11. LGM in (2.4) can be specified as a special case of a regression model whose response z_i are assumed to follow an exponential family distribution with mean $\mu_i = \text{E}(z_i|y_i, \theta)$ linked

to a Gaussian linear predictor η_i via a known link function $g(\cdot)$, as $g(\mu_i) = \eta_i$ and

$$\eta_i = \alpha + \sum_j \beta_j x_{j,i} + \sum_k f_k(u_{ki}) + \epsilon_i \tag{2.6}$$

where α is the intercept, $\{\beta_j\}$ are coefficients (fixed effects) of covariates $\{x_{j,i}\}$, and $f_k(\cdot)$ are unknown functions of covariates u, and ϵ_i is a random error. Casting it as an LGM, we can set

$$y = (\alpha, \{\beta_j\}, \{f_k(u_{ki})\}, \{\eta_i\})$$

is the latent field in (2.4) (for conveniency, we consider η_i instead of ϵ), and the rest hyperparameters (to be learned) constitute θ.

Note 12. Consequently the class LGM involves many computationally challenging models, such as the spatial models (geostatistical, latent, point process), the associated spatiotemporal models, and the mixed effect GLM.

2.3. **The general idea.**

Note 13. We are interested in computing the following marginals of (2.5)

$$\mathrm{pr}(\theta_j|z) = \int\int \mathrm{pr}(y,\theta|z)\,\mathrm{d}y\mathrm{d}\theta_{-j} = \int \mathrm{pr}(\theta|z)\,\mathrm{d}\theta_{-j} \tag{2.7}$$

$$\mathrm{pr}(y_i|z) = \int\int \mathrm{pr}(y,\theta|z)\,\mathrm{d}y_{-i}\mathrm{d}\theta = \int \mathrm{pr}(y_i|z,\theta)\,\mathrm{pr}(\theta|z)\,\mathrm{d}\theta \tag{2.8}$$

where integrals (2.7) and (2.8) can be of high dimentionality wrt y.

Note 14. For the approximation of (2.7) and (2.8), INLA involves three steps: evaluation of $\mathrm{pr}(y_i|z,\theta)$ via Laplace approx, evaluation of $\mathrm{pr}(\theta|z)$ via Laplace approx, and finally numerical integration.

Note 15. To compute an approximate for $\mathrm{pr}(\theta|z)$, notice that at any point y it is

$$\mathrm{pr}(\theta|z) = \frac{\mathrm{pr}(y,\theta|z)}{\mathrm{pr}(y|z,\theta)} \propto \frac{\mathrm{pr}(z|y,\theta)\,\mathrm{pr}(y|\theta)\,\mathrm{pr}(\theta)}{\mathrm{pr}(y|z,\theta)} \tag{2.9}$$

Unlike the numerator, the denumerator is not available in closed form and is hard to compute. INLA employs the approximation of $\mathrm{pr}(y|z,\theta)$ by a multivariate Gaussian distribution $\widetilde{\mathrm{pr}}_{\mathrm{G}}(y|z,\theta)$ whose mean is the mode $y^*(\theta)$ of $\mathrm{pr}(y|z,\theta)$ and covariance matrix is the minus inverse Hessian at that mode. Essentially, the approximation of (2.9) at a specific value of θ is

$$\widetilde{\mathrm{pr}}(\theta|z) \propto \left.\frac{\mathrm{pr}(z|y,\theta)\,\mathrm{pr}(y|\theta)\,\mathrm{pr}(\theta)}{\widetilde{\mathrm{pr}}_{\mathrm{G}}(y|z,\theta)}\right|_{y=y^*(\theta)} \tag{2.10}$$

which is equivalent to the Laplace approximation method for marginal densities.

Note 16. To compute an approximate for $\text{pr}(y_i|z,\theta)$ at each y_i there are three main approaches:

Gaussian approximation approach.: Compute the marginal from the Gaussian approximation $\widetilde{\text{pr}}_{\text{G}}(y|z,\theta)$ of $\text{pr}(y|z,\theta)$ in Note 15. This is fast but not generally accurate.

Laplace approximation: Similar to Note 15, compute

$$\widetilde{\text{pr}}(y_i|z,\theta) \propto \left.\frac{\text{pr}(z|y,\theta)\,\text{pr}(y|\theta)\,\text{pr}(\theta)}{\widetilde{\text{pr}}_{\text{G}}(y_{-i}|y_i,z,\theta)}\right|_{y=y^*(\theta)} \tag{2.11}$$

where $\widetilde{\text{pr}}_{\text{G}}(y_{-i}|y_i,z,\theta)$ is a multivariate Gaussian distribution whose mean is the mode $y^*_{-i}(y_i,\theta)$ and covariance matrix is the minus inverse Hessian at that mode. It is more accurate than the previous one but computational demanding because it requires the the re-calculation of the precision matrix for each y_i.

Simplified Laplace approximation: It builds on third order Taylor series expansions both in numerator and denumerator of (2.11), which improves the approximation wrt asymmetry. We skip the mathematical details here. It has improved accuracy.

2.4. **The schematic of the procedure.**

Algorithm 17. *Summing up, the INLA method proceeds as follows:*

(1) Explore the space of θ.
   (a) Locate a collection of points $\left\{\theta^{(k)}; k=1,...,K\right\}$ in the area of high density of $\widetilde{\text{pr}}(\theta|z)$.
   (b) Find the mode of $\widetilde{\text{pr}}(\theta|z)$.

(2) Compute approximation $\widetilde{\text{pr}}(\theta|z)$ at points $\left\{\theta^{(k)}; k=1,...,K\right\}$ by using (2.10).

(3) Compute approximation $\widetilde{\text{pr}}(y_i|z,\theta)$ at points $\left\{\theta^{(k)}; k=1,...,K\right\}$ of θ by using the Laplace approximation in (2.11) or the simplified Laplace approximation, or the Gaussian approximation, as said in Note 16.

(4) Compute the approximation $\widetilde{\text{pr}}(y_i|z)$ of (2.8) via standard numerical approximation as

$$\widetilde{\text{pr}}(y_i|z) = \sum_{k=1}^{K} \widetilde{\text{pr}}\left(y_i|z,\theta^{(k)}\right) \widetilde{\text{pr}}\left(\theta^{(k)}|z\right) \Delta^{(k)} \tag{2.12}$$

where $\Delta^{(k)}$ as weights depending on the locations $\left\{\theta^{(k)}\right\}$ and the numerical integration scheme. If $\left\{\theta^{(k)}\right\}$ are equal-distant then $\Delta^{(k)} = 1$.

(5) Compute the approximation $\widetilde{\text{pr}}(y_i|z)$ of (2.8) via standard numerical approximation as

$$\widetilde{\text{pr}}(\theta_j|z) = \sum_{k=1}^{K} \widetilde{\text{pr}}\left(\theta_{-j}, \theta_{-j}^{(k)}|z\right) \Delta^{(k)} \tag{2.13}$$

where $\Delta^{(k)}$ as weights depending on the locations $\left\{\theta_{-j}^{(k)}\right\}$ and the numerical integration scheme. If $\left\{\theta^{(k)}\right\}$ are equal-distant then $\Delta^{(k)} = 1$.

Note 18. The error in (2.12) comes from the Laplace approximations in $\widetilde{\text{pr}}\left(\theta^{(k)}|z\right)$ and $\widetilde{\text{pr}}\left(y_i|z,\theta^{(k)}\right)$, as well as the numerical integration and the choice of locations $\left\{\theta^{(k)}\right\}$. When the likelihood $\text{pr}\left(y|z,\theta^{(k)}\right)$ is Gaussian then its marginals are Gaussian and hence this error is eliminated.

2.5. **Byproducts.**

Note 19. Marginal likelihood $\text{pr}(z)$ is often used in Bayesian model comparison, and model averaging. A natural approximation for the marginal likelihood $\text{pr}(z)$ is

$$\widetilde{\text{pr}}(z) = \int \left.\frac{\text{pr}(z|y,\theta)\,\text{pr}(y|\theta)\,\text{pr}(\theta)}{\widetilde{\text{pr}}_{\text{G}}(y|z,\theta)}\right|_{y=y^*(\theta)} \text{d}\theta$$

The approx can fail when $\text{pr}(\theta|z)$ is multimodal, however LGM generate unimodal posteriors in most cases.

Note 20. Deviance Information Criterion (DIC) can be used in Bayesian model comparison. Analogously to AIC, the deviance of the model is

$$D(\theta) = -2\log\left(\text{pr}(z|\theta)\right),$$

the model complexity here is measured via effective number of parameters

$$p_D = \text{E}\left(D(\theta)|z\right) - D\left(\text{E}(\theta|z)\right)$$

and hence DIC is defined as

$$\text{DIC} = \text{E}\left(D(\theta)|z\right) + p_D.$$

Models with smaller DIC are better supported by the data. INLA approximates integrals/expectations numerically after (2.10) has been approximated.

Note 21. Predictive distribution of an unseen value z^{new} (includes missing data) given the observables z and model (2.4) is

$$\text{pr}\left(z^{\text{new}}|z\right) = \int \text{pr}\left(z^{\text{new}}|y^{\text{new}}\right)\text{pr}\left(y^{\text{new}}|z\right)\text{d}y^{\text{new}} \tag{2.14}$$

$$\text{pr}\left(y^{\text{new}}|z\right) = \int \text{pr}\left(y^{\text{new}}|\theta\right)\text{pr}\left(\theta|z\right)\text{d}\theta \tag{2.15}$$

due to the conditional independence in (2.1). Given that (2.10) has been approximated, INLA employs numerical integration for the integral (2.15) firstly and 2.14 secondly.

3. The R-INLA software (An empirical introduction)

Note 22. All the info is int he website of the software `https://www.r-inla.org`

3.1. **How to install R-INLA.**

Note 23. To install R-INLA do the following from `https://www.r-inla.org/download-install`.

```
# install the stable version, do
install.packages("INLA",repos=c(getOption("repos"),
    INLA="https://inla.r-inla-download.org/R/stable"),
    dep=TRUE)
install.packages("INLA",repos=c(getOption("repos"),
    INLA="https://inla.r-inla-download.org/R/testing"),
    dep=TRUE)
# update the stable version the package
inla.upgrade()
# install dependency fmesh R package
options(repos=c( inlabruorg = "https://inlabru-org.r-universe.dev",
     INLA = "https://inla.r-inla-download.org/R/testing",
     CRAN = "https://cran.rstudio.com")
     )
install.packages("fmesher")
```

3.2. **How to use R-INLA.**

Note 24. There are two essential steps:

(1) Define the linear predictor (2.6) through a formula object
(2) Complete the model definition and fit the model using the R function `inla{INLA}`.

The fitted model is returned as an inla object.

Example 25. We analyze the R dataset `Salm{INLA}`.

- Bayesian model

$$\begin{cases} z_{i,j}|\lambda_{i,j} \sim \text{Poi}\left(\lambda_{i,j}\right) & i = 1, ..., 6 \text{ and } j = 1, 2, 3 \\ \log\left(\lambda_{i,j}\right) = \beta_0 + \beta_1 \log\left(x_i + 10\right) + \beta_2 x_i + u_{i,j} & i = 1, ..., 6 \text{ and } j = 1, 2, 3 \end{cases}$$

  where $\{z_{i,j}\}$ (the observables) are number of colonies found on plate j for dose i and x_i indicate the ith dose. Let $u_{ij}|\tau \sim \text{N}\left(0, \sigma^2\right)$ be the so-called random effects, while $\{\beta_i\}$ are unknown parameters called fixed effects.
- In terms of model (2.4), the GMRF is $y = \left(\{\lambda_{i,j}\}, \{\beta_i\}, \{u_{i,j}\}\right)$.
- We consider prior on σ^2 such that

$$\tau = -\log\left(\sigma^2\right) \sim \text{type-2 Gimbel}(1/2, -\log\left(a\right)/u)$$

  This is because R-INLA specifies prior on $\tau = -\log\left(\sigma^2\right)$.

Data loading.

- Load R-INLA

```
# load the data set
library(‘‘INLA’’)
```

- We import the R data set `Salm{INLA}` as follows

```
# load the data set
data(Salm)
# get info about the R dataset
?Salm
# rename the columns to fit the notation
names(Salm) = c("z", "x", "u")
```

Training via R-INLA.

- Code the model in R-INLA language, and produce the inla object

```
# specify the prior for the log precision parameter
my.hyper <- list(theta = list(prior="pc.prec", param=c(1,0.01)))
# specify the linear predictor
formula <- z ~ log(x + 10) + x + f(u, model = "iid", hyper = my.hyper)
# run R-INLA and get the result object
result <- inla(formula=formula, data=Salm, family="Poisson",
    control.inla = list(strategy=’’laplace’’))
```

- The ‘`formula`’ is as in `lm{stats}` command.
- Function ‘`inla.list.models()`’ provides a list of available distributions for the different parts of the model, such as the "prior" (available priors for the hyper-parameters), "likelihood" (all implemented likelihoods) and "latent" (available models for the latent field).

- Function `f()` is used to specify the latent Gaussian model for the non-linear terms and random effect u_{ij}; here an independent noise model (hence the use of model = "iid"), and the hyperprior for its corresponding hyperparameters (here σ^2).
- R function `inla{INLA}` (given the input above) generates an inla object similar to that of `lm{stats}`. The data object should be data.frame or list. The likelihood is specified in form of a string. `strategy=''laplace''` refers to the approximation strategy in Note 16 and has options "gaussian", "simplified.laplace", "laplace".

Parametric inference.

- Post-processing the results from `inla` object.

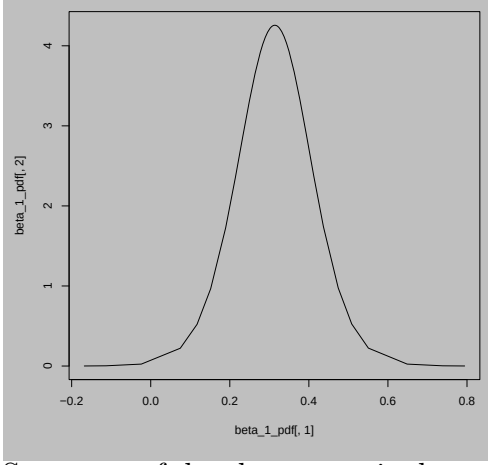
```
summary(result)
```

```
Time used:
    Pre = 0.343, Running = 0.156, Post = 0.0147, Total = 0.514
Fixed effects:
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept)  2.165 0.362      1.445    2.166      2.880  2.167   0
log(x + 10)  0.313 0.099      0.117    0.313      0.508  0.314   0
x           -0.001 0.000     -0.002   -0.001      0.000 -0.001   0

Random effects:
  Name    Model
    u IID model

Model hyperparameters:
                  mean    sd 0.025quant 0.5quant 0.975quant  mode
Precision for u 20.64 16.52       5.72    16.44      59.79 11.91

Marginal log-Likelihood:  -83.69
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

It provides summary statistics of the posterior of the fixed effect, random effect, and precision parameters, as well as the marginal log-likelihood $\log(\text{pr}(z))$.

- Marginal posteriors for the fixed effect, random effect, and hyperparameters are stored in `result$marginals.fixed`, `result$marginals.random`, `result$marginals.hyperpar`. E.g., one can plot the posterior of β_1 as

```
beta_1_pdf <- result$marginals.fixed$`log(x + 10)`
plot(beta_1_pdf[,1], beta_1_pdf[,2], type="l")
```



- Summary of the above marginal posteriors can be obtained by using `result$summary.fixed`, `result$summary.random`, `result$summary.hyperpar`

result$marginals.fixed

```
> result$summary.fixed
                    mean           sd   0.025quant      0.5quant    0.975quant          mode
(Intercept)  2.1647643605 0.3620126799  1.444666455  2.1655831923  2.879995e+00  2.1669703669
log(x + 10)  0.3132991434 0.0985605383  0.117201855  0.3134878885  5.084337e-01  0.3139144159
x           -0.0009656845 0.0004357064 -0.001827388 -0.0009671395 -9.635679e-05 -0.0009702587
                    kld
(Intercept) 1.419280e-08
log(x + 10) 2.901292e-08
x           4.525820e-08
```

result$summary.hyperpar

```
> result$summary.hyperpar
                      mean       sd 0.025quant 0.5quant 0.975quant     mode
Precision for u 20.64402 16.51935    5.72236 16.44435   59.78984 11.90988
```

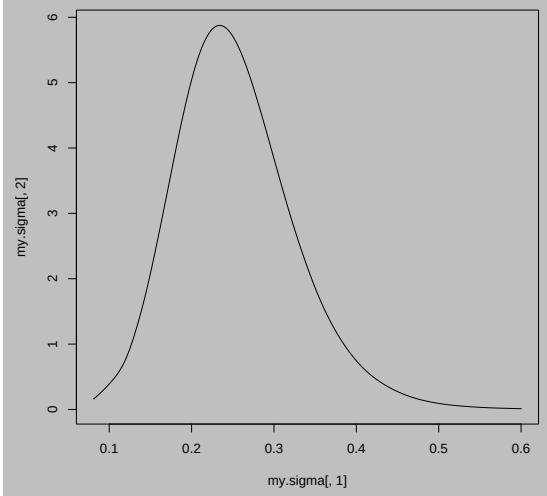
- To get the posterior summary of a function of the parameters, e.g. the posterior mean and standard deviation of $\sigma^2 = \exp(\tau)$

```
# Select the right hyperparameter marginal
tau <- result$marginals.hyperpar[[1]]
# Compute the expected value for 1/sqrt(τ) and 1/sqrt(τ)^2
E = inla.emarginal(function(x) c(1/sqrt(x),(1/sqrt(x))^2), tau)
# From this we computed the posterior standard deviation as
mysd = sqrt(E[2] - E[1]^2)
# so that we obtain the posterior mean and standard deviation
print(c(mean=E[1], sd=mysd))
```

```
      mean         sd
0.25353753 0.07325247
```

- To compute the marginal posterior distribution of $\sigma^2 = \exp(\tau)$ use the `inla.tmarginal()`

```
# Select the right hyperparameter marginal
tau <- result$marginals.hyperpar[[1]]
# Do the transformation
my.sigma <- inla.tmarginal(function(x){1/sqrt(x)}, tau)
# plot
plot(my.sigma[,1], my.sigma[,2], type="l")
```



- Other R-INLA functions providing operations on posterior marginals can be found in R help documentation,

```
?inla.marginal
```

Predictive inference.

- In R-INLA there is no function `predict{stats}` as for `glm{stats}` or `lm{stats}`. Predictions must be done as a part of the model fitting itself. Prediction can be regarded as fitting a model with missing data, hence we can simply set `y[i]=NA` for those "locations" we want to predict. Predictive distributions, which are often of interest, are however not returned directly, and the user needs to some extra "hacks". There are two reasonable "hacks".
- For illustration, pretend 7th observation is unknown, by removing it from the training data, and try to predict it.

```
## set observation 7 to NA
Salm.predict = Salm
Salm.predict[7, "y"] <- NA
# re-run the model
res.predict = inla(formula=formula, data=Salm,     family="Poisson",
    control.predictor = list(compute = TRUE),
    control.family = list(control.link=list(model="log")) )
```

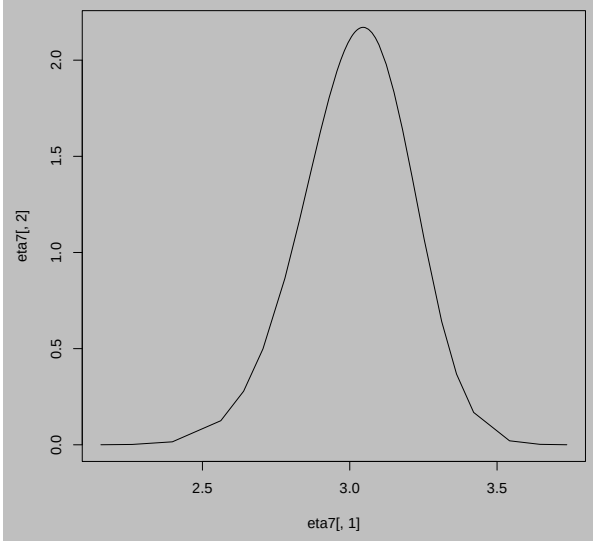
- Using the same settings as before, train the model by function `inla(INLA)`.

```
# re-run the model
res.predict = inla(formula=formula, data=Salm,     family="Poisson",
    control.predictor = list(compute = TRUE),
    control.compute=list(return.marginals.predictor=TRUE),
    control.family = list(control.link=list(model="log")) )
```

By specifying `control.predictor=list(compute=TRUE)` the posterior marginals will be included in the results object. We also need to explicitly specify the link function g connecting $g(\lambda_i) = \eta_i$, where $\lambda_i = \mathrm{E}(z_i)$, using the `control.family` object in order for `inla()` to compute the linear predictor η_i. Note that here $\lambda_i = \exp(\eta_i)$. By specifying `control.compute=list(return.marginals.predictor=TRUE)`, we ask function `inla(INLA)` to compute and return the marginal pdf of the linear predictor, which be default are not due to computational cost.

- We can compute $\mathrm{pr}(\eta_7|z_{-7})$ by

```
# marginal posterior for the linear predictor
eta7 = res.predict$marginals.linear.predictor[[7]]
```

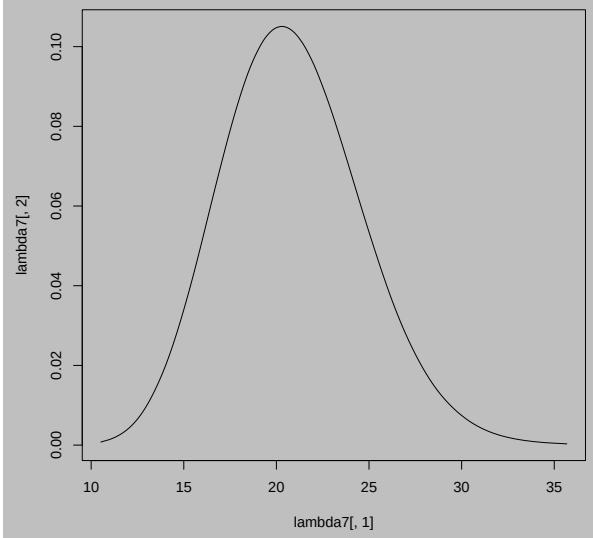


- Summary about is $\mathrm{pr}(\eta_7|z_{-7})$ taken by

```
# some summary statistics round(res.predict$summary.linear.predictor[7,], 3)
> res.predict$summary.linear.predictor[7,]
                  mean        sd 0.025quant 0.5quant 0.975quant     mode          kld
Predictor.07 3.021652 0.1847223   2.639797 3.029161   3.362469 3.045581 1.224947e-07
```

- We can compute $\text{pr}\,(\lambda_7|z_{-7})$ by

```
# marginal posterior for lambda
eta7 = res.predict$marginals.linear.predictor[[7]]
lambda7 = inla.tmarginal(function(x){exp(x)}, eta7)
## or alternatively
# lambda7_bis = inla.tmarginal(function(x){exp(x)},eta7)
## or alternatively
# lambda7_bis = inla.tmarginal(function(x){exp(x)},eta7)
# plot
plot(lambda7[,1], lambda7[,2], type="l")
```

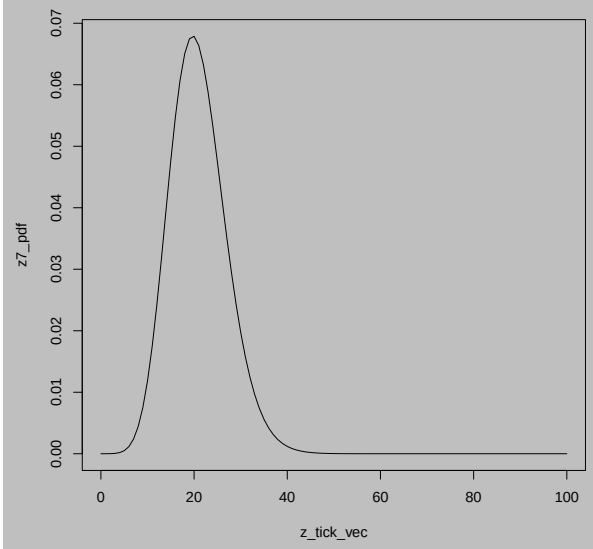


- To compute $\text{pr}\,(z_7|z_{-7})$ i.e. the predictive distribution (in this case) or the posterior distribution of the missing value (in principle), we can consider the following integration

$$\begin{aligned} \text{pr}\,(z_7|z_{-7}) &= \int \text{pr}\,(z_7|\lambda_7)\,\text{pr}\,(\lambda_7|z_{-7})\,\text{d}\lambda_7 \\ &\approx \int \widetilde{\text{pr}}\,(z_7|\lambda_7)\,\widetilde{\text{pr}}\,(\lambda_7|z_{-7})\,\text{d}\lambda_7 \end{aligned} \tag{3.1}$$

and either approximated by using numerical integration, e.g. trapezoid rule with R function `trapz{caTools}`

```
# library supporting trapezoid rule integration.
library(caTools)
# specify the support at which we want to compute the density
z_tick_vec = 0:100
z7_pdf = rep(0,101)
# go over the posterior marginal of the fitted value
for(j in 1:(length(lambda7[ ,1])-1)) {
    z7_pdf <- z7_pdf + dpois(z_tick_vec,
        lambda = ((lambda7[j,1]+ lambda7[j+1,1])/2))
            * trapz(lambda7[j:(j+1), 1], lambda7[j:(j+1), 2])
}
# plot
plot(z_tick_vec,z7_pdf, type="l")
```

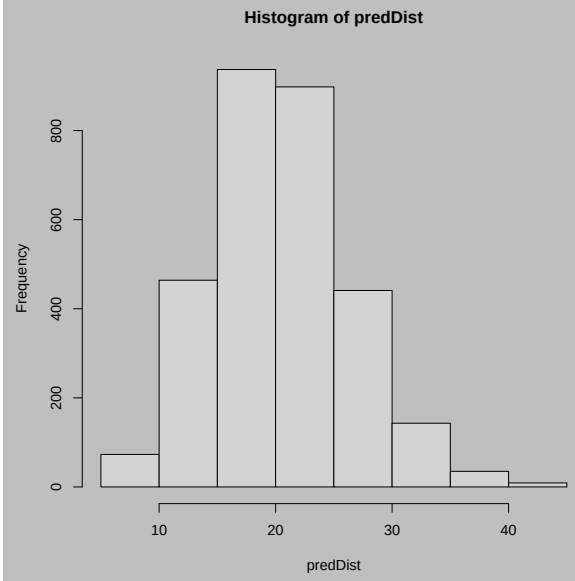


- alternatively one approximate (3.1) by Monte Carlo integration

$$\begin{aligned} \text{pr}\left(z_7|z_{-7}\right) &\approx \text{E}_{\widetilde{\text{pr}}(\lambda_7|z_{-7})}\left(\widetilde{\text{pr}}\left(z_7|\lambda_7\right)\right) \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\widetilde{\text{pr}}\left(z_7|\lambda_7^{(t)}\right) \end{aligned} \tag{3.2}$$

where $\left\{\lambda_7^{(t)}\right\}_{t=1}^{T}$ is a sample drawn from $\widetilde{\text{pr}}\left(\lambda_7|z_{-7}\right)$ by using function inla.rmarginal{INLA} as follows.

```
# set the number of samples (T)
n.samples = 3000
# sample from the marginal latent distribution
samples_lambda = inla.rmarginal(n.samples, lambda7)
# sample from the likelihood model
predDist = rpois(n.samples, lambda = samples_lambda)
```



Introduced in ASM3, MLNN3

Note 26. Assume we wish to address the minimization problem

$$\hat{\theta} = \arg\min_{\theta} \left(C\left(\theta\right) \right) \tag{A.1}$$

for some cost function $C\left(\cdot\right)$.

Note 27. For instance, Proposition 1, it is $C\left(\theta\right) = -2\log\left(L\left(\theta\right)\right)$.

Note 28. Newton algorithm and Gradient descent algorithms are two optimization algorithms aiming to address the minimization problem (A.1). Each of them generate a convergence sequence $\left\{\theta^{(t)}\right\}$ to $\hat{\theta}$ as $\theta^{(t)} \to \hat{\theta}$ under regularity conditions (omitted here).

Algorithm 29. *Newton algorithm consist of the recursion*

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla_{\theta}^{2} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}\right]^{-1} \nabla_{\theta} C\left(\theta\right)|_{\theta=\theta^{(t)}}$$

where $\nabla_{\theta} C\left(\theta\right)|_{\theta=\theta^{(t)}}$ *is the gradient of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$, $\nabla_{\theta}^{2} C\left(\theta\right)|_{\theta=\theta^{(t)}}$ *is the Hessian matrix of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$. *It requires a user specified seed* $\theta^{(0)}$. *The recursion stops when a termination criterion such as* $t \geq T_{\max}$, *for some user specified* $T_{\max} > 0$, *is satisfied.*

Algorithm 30. *Gradient descent algorithm consist of the recursion*

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \, \nabla_{\theta} C\left(\theta\right)|_{\theta=\theta^{(t)}}$$

where $\nabla_{\theta} C\left(\theta\right)|_{\theta=\theta^{(t)}}$ *is the gradient of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$. *It requires a user specified positive non-increasing sequence* $\{\eta_t\}$ *such as* $\eta_t = \sqrt{1/t}$, *and a user specified seed* $\theta^{(0)}$. *The recursion stops when a termination criterion such as* $t \geq T_{\max}$ *for some user-specified* $T_{\max} > 0$, *is satisfied.*

Example 31. Consider the marginal likelihood

$$f(x|a,b) = \left(\frac{1}{\Gamma(a) b^{a}}\right)^{n} \prod_{i=1}^{n} x_i^{a} e^{-n\bar{x}\frac{1}{b}}$$

where $a > 0$, $b > 0$. Write the Newton alg., and Gradient descent alg. recursions for to find $\theta^{*} = \arg\min_{\theta} \left(-\ell_n\left(\theta\right)\right)$ where $\ell_n\left(\theta\right) = \log f(x|\theta)$ and $\theta = (a, b)$.

Hint-1: Digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log\Gamma(x)$

Hint-2: Trigamma function $\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \log\Gamma(x)$

Hint-3: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Proof. Gradient descent's recursion is

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \eta_t \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

for $\eta_t = \sqrt{1/t}$, where

$$\ell_n(\theta) = -n\log\Gamma(a) - na\log(b) - \frac{1}{b}\sum_{i=1}^n x_i + (a-1)\sum_{i=1}^n \log(x_i)$$

$$\nabla_\theta \ell_n(\theta) = \begin{bmatrix} -n\psi(a) - n\log(b) + \sum_{i=1}^n \log(x_i) \\ -n\frac{a}{b} + n\frac{1}{b^2}\bar{x} \end{bmatrix}, \text{ and } \nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix}$$

$$\nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix} = -\frac{1}{n}\frac{1}{\psi_1(a)\frac{2\bar{x}-ab}{b} - 1}\begin{bmatrix} \frac{2\bar{x}-ab}{b} & -b \\ -b & b^2\psi_1(a) \end{bmatrix}$$

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \eta_t \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

Newton algorithm's recursion is

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \left[\left.\nabla^2_\theta C(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}\right]^{-1} \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

where additionally

$$\nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix};\ \text{hence}\ \left[\nabla^2_\theta \ell_n(\theta)\right]^{-1} = -\frac{1}{n}\frac{1}{\psi_1(a)\frac{2\bar{x}-ab}{b} - 1}\begin{bmatrix} \frac{2\bar{x}-ab}{b} & -b \\ -b & b^2\psi_1(a) \end{bmatrix}$$

□

Appendix B. Gaussian approximation of a (posterior) distribution

Introduced in SI2

Note 32. A well known approximation of the posterior distribution is the Gaussian posterior approximation.

Theorem 33. *The posterior density $pr(\theta|z_{1:n})$ of θ given n observables $z_{1:n}$ can be approximated by a multivariate Gaussian distribution density $pr_G(\theta|\mu_n, \Sigma_n)$ with mean μ_n being the mode i.e. $\frac{\partial}{\partial\theta_i}\log\left(pr(\theta|z_{1:n})\right)\Big|_{\theta=\mu_n} = 0$, and with covariance matrix $\Sigma_n > 0$* being the inverse Hessian at the mode i.e. $\Sigma_n = \left(H_{pr}(\mu_n)\right)^{-1}$ *where* $\left[H_{pr}(\mu_n)\right]_{i,j} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(pr(\theta|z_{1:n})\right)\Big|_{\theta=\mu_n}$.

Example 34. Consider a Bayesian model with sampling distribution $x_i|\theta \overset{\text{iid}}{\sim} \text{pr}(x_i|\theta) \propto \theta^{x_i}(1-\theta)^{x_i-1}$ and prior $\theta \sim \text{pr}(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. Find the Gaussian approximation of the posterior $\text{pr}(\theta|x)$ of θ given $x = (x_1, ..., x_n)$.

Solution. The log posterior density is

$$\log\left(\text{pr}\left(\theta|x\right)\right) = \left(a_n - 1\right)\log(\theta) + \left(b_n - 1\right)\log(1-\theta)$$

where $a_n = a + n\bar{x}$, and $b_n = b + n - n\bar{x}$. So

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\log\left(\text{pr}\left(\theta|x\right)\right)\Big|_{\theta=\mu_n} = \frac{a_n - 1}{\theta} - \frac{b_n - 1}{1-\theta}\Big|_{\theta=\mu_n} \implies \mu_n = \frac{a_n - 1}{a_n + b_n - 2}$$

$$\Sigma_n = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log\left(\text{pr}\left(\theta|x\right)\right)\Big|_{\theta=\mu_n} = \frac{a_n - 1}{\theta^2} - \frac{b_n - 1}{(1-\theta)^2}\Big|_{\theta=\mu_n} \implies \Sigma_n = \frac{(a_n - 1)(b_n - 1)}{(a_n + b_n - 2)^3}$$

Therefore, θ has asymptotic posterior density is that of $\text{N}\left(\mu_n, \Sigma_n\right)$; i.e. $\text{pr}\left(\theta|x\right) \approx \text{N}\left(\theta|\mu_n, \Sigma_n\right)$.

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 3: Point referenced data modeling / Geostatistics

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

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

Reading list & references:

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

Specialized reading.

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

Part 1. Intro to building stochastic models & concepts

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

1. Stochastic processes (or Random fields)

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

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

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

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

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

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

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

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

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

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

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

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

is a well defined stochastic process as it satisfies Thm 5.

1.1. **Mean and covariance functions.**

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

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

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

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

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

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

Proposition 9. *The function $c : S \times S \to \mathbb{R}$, $S \subseteq \mathbb{R}^d$* is the covariance function iff $c(\cdot,\cdot)$ is semi-positive definite; i.e. the Gram matrix $(c(s_i, s_j))_{i,j=1}^n$ is non-negative definite for any $\{s_i\}_{i=1}^n$, $n \in \mathbb{N}$.

Example 10. $c(s,s') = 1(s = s')$ is a proper covariance function as

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

Note 11. Prop 12 uses the experience from basis functions, while Theorem 30 uses experience from characteristic functions to be incorporated int he process for modeling reasons.

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

Proof. From Prop 9, as

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

□

2. Second order processes (or Random fields)

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

3. Gaussian process

Definition 14. $Z = (Z_s; s \in S)$ indexed by $S \subseteq \mathbb{R}^d$ is a Gaussian process (GP) or random field (GRF) if for any $n \in \mathbb{N}$ and for any finite set $\{s_1, ..., s_n; s_i \in \mathcal{S}\}$, the random vector $(Z_{s_1}, ..., Z_{s_n})^\top$ has a multivariate normal distribution.

Also Example of Proposition

Proposition 15. *A GP $Z = (Z_s; s \in S)$ is fully characterized by its mean function $\mu : S \to \mathbb{R}$ with $\mu(s) = E(Z_s)$, and its covariance function with $c(s,s') = Cov(Z_s, Z_{s'})$.*

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

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

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

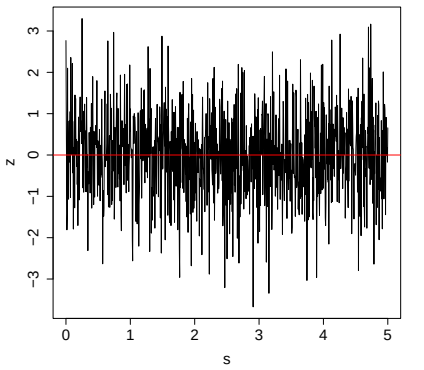
Example 18. Recall your linear regression lessons where you specified a sampling distribution $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta, \sigma^2\right)$, $\forall x \in \mathbb{R}^d$; well that can be considered as a GP with $\mu_x = x^\top\beta$ and $c(x,x') = \sigma^2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

Example 19. Figs. 3.1 & 3.2 presents realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(s) = 0$ and differently parameterized covariance functions in 1D and 2D. In 1D the code to simulate the GP is given in Algorithm 1. Note that we actually discretize it and simulate it from the fidi.

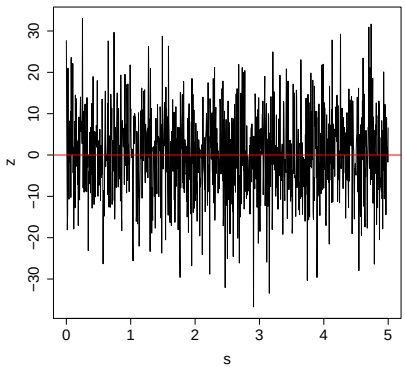
Algorithm 1 R script for simulating from a GP $(Z_s; s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

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

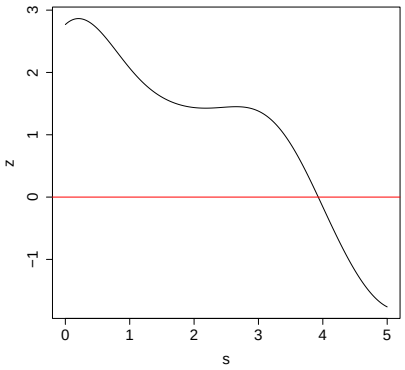
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Fig. 3.1a & 3.1b ; Fig. 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Fig. 3.2c & 3.2d), and the spatial dependence / frequency of the ups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



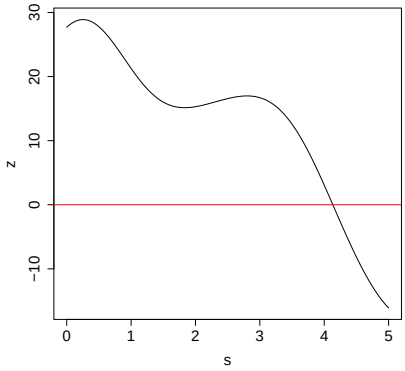
(A) Nugget c.f
($\sigma^2 = 1$)



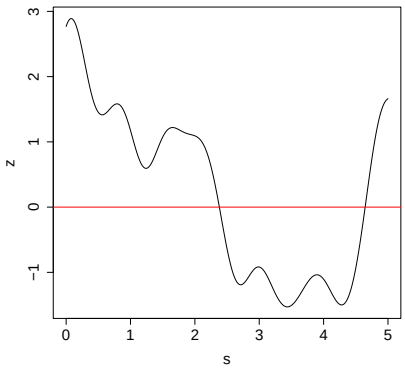
(B) Nugget c.f
($\sigma^2 = 100$)



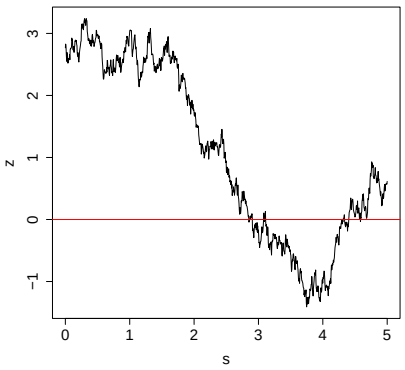
(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



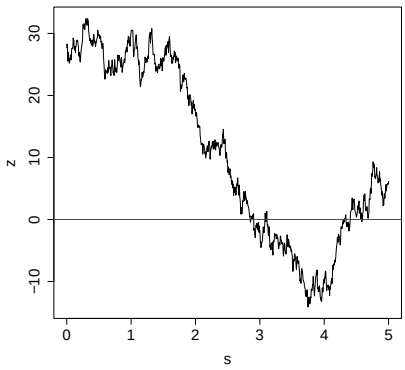
(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



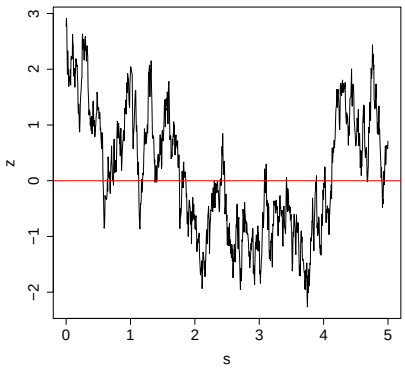
(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)



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

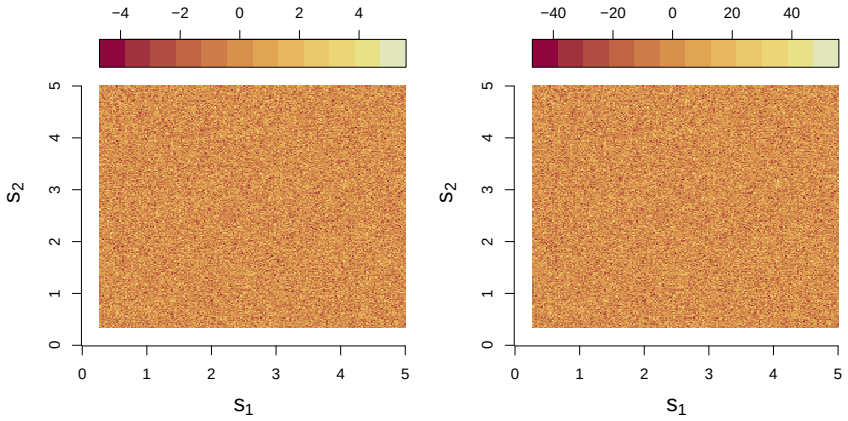


(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)



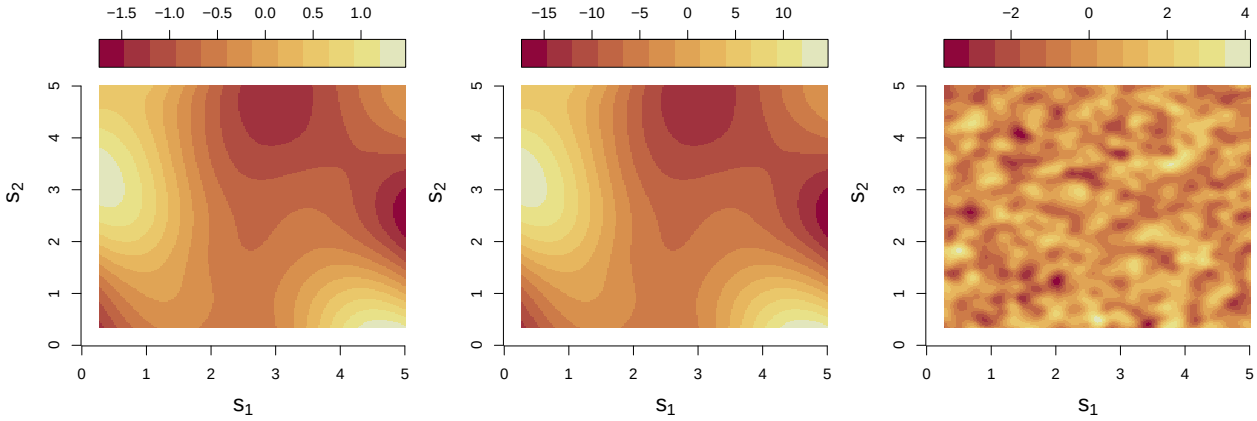
(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

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



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

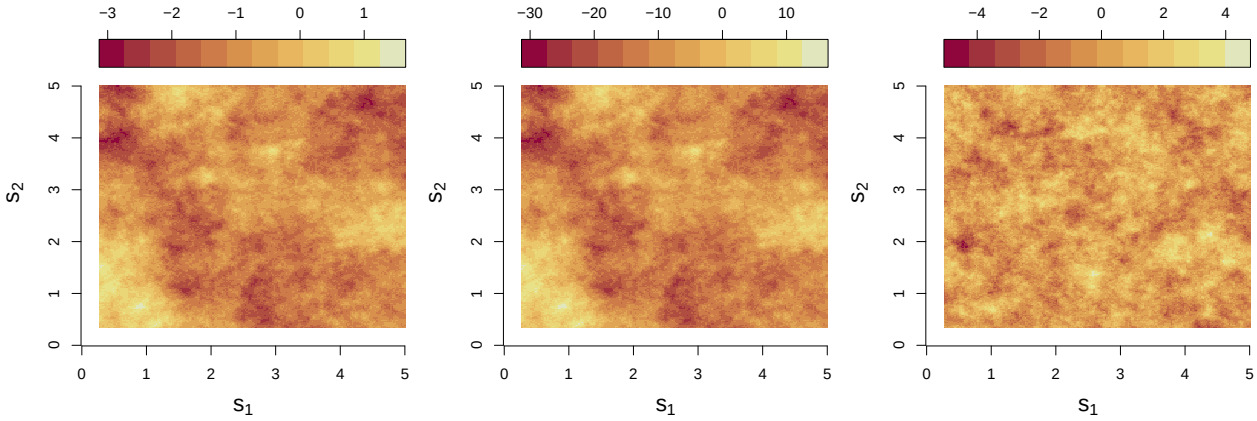
(B) Nugget c.f
($\sigma^2 = 100$)



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

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

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



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

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

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

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

4. Strong stationarity

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

Definition 21. A random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ is strongly stationary if for all finite sets consisting of $s_1, ..., s_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, for all $k_1, ..., k_n \in \mathbb{R}$, and for all $h \in \mathbb{R}^d$

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

5. Weak stationarity (or second order stationarity)

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

Definition 23. A random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ is weakly stationary (or second order stationary) if, for all $s, s' \in \mathbb{R}^d$,

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

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

6. Covariogram

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

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

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

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

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

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

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

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

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

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

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

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

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

Note 29. The following helps in the specification of cavariograms by considering properties of characteristic functions.

Theorem 30. *(Bochner's theorem) A continuous even real-valued function $c : \mathbb{R}^d \to \mathbb{R}$ is a covariance function of a weakly stationary random process if and only if it can be represented as*

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

where $\mathrm{d}F(\omega)$ is a *symmetric positive finite measure on* $\mathbb{R}^d$.

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

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

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

Theorem 31. *If $c(\cdot)$ is integrable, $F(\cdot)$ is absolutely continuous with spectral density $f(\cdot)$ of $Z = (Z_s; s \in \mathcal{S})$ then by Fast Fourier transformation*

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

Example 32. Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Then the spectral density from Thm 30 is

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

i.e. of a Gaussian form.

7. Intrinsic stationarity

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

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

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

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

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

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

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

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

and

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

8. (Semi) variogram

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

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

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

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

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

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

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

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

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

(1) It is $\gamma(h) = \gamma(-h)$, $\gamma(h) \geq 0$, and $\gamma(0) = 0$

(2) Semi-variogram is conditionally negative definite (c.n.d.): for all $a \in \mathbb{R}^n$ s.t. $\sum_{i=1}^n a_i = 0$, and for all $\forall\{s_1, ..., s_n\} \subseteq S$

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

(3) If $\gamma(h)$ is a semi-variogram, and A is a linear transformation in $\mathbb{R}^d$ then $\tilde{\gamma}(h) = \gamma(Ah)$ is a semi-variogram too.

(4) The following functions are semi-variograms

(a) $\gamma(\cdot) = \sum_{i=1}^n a_i\gamma_i(\cdot)$, if $a_i \geq 0$, and $\{\gamma_i(\cdot)\}$ are semi-variograms

(b) $\gamma(\cdot) = \int \gamma_u(\cdot)\, dF(u)$, if $\gamma_u(\cdot)$ is a semi-variogram parametrized by $u \sim F$

(c) $\gamma(\cdot) = \lim_{n\to\infty}\gamma_n(\cdot)$ if $\gamma_n(\cdot)$ is semi-variogram and the limit exists

(5) Consider intrinsically stationary stochastic processes $Y = (Y_s)_{s\in\mathbb{R}^d}$ and $E = (E_s)_{s\in\mathbb{R}^d}$ where Y and E are independent each other. Let $Z_s = Y_s + E_s$. Then

$$\gamma_Z(h) = \gamma_Y(h) + \gamma_E(h)$$

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

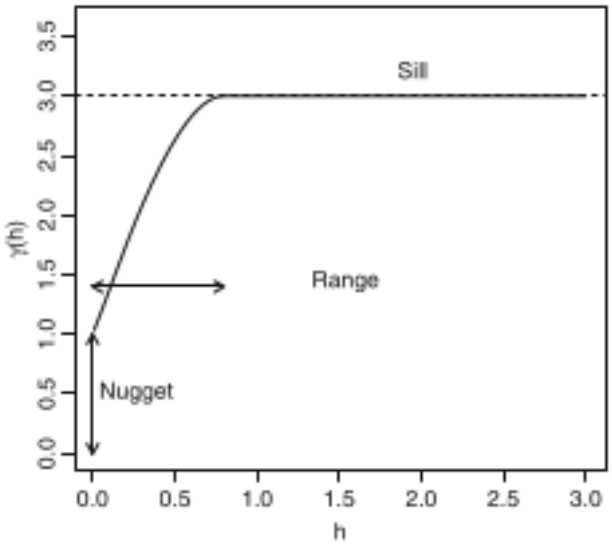


FIGURE 8.1. Variogram's characteristics

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

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

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

Nugget Effect.

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

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

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

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

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

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

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

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

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

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

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

Sill.

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

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

Partial sill.

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

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

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

9. Isotropy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

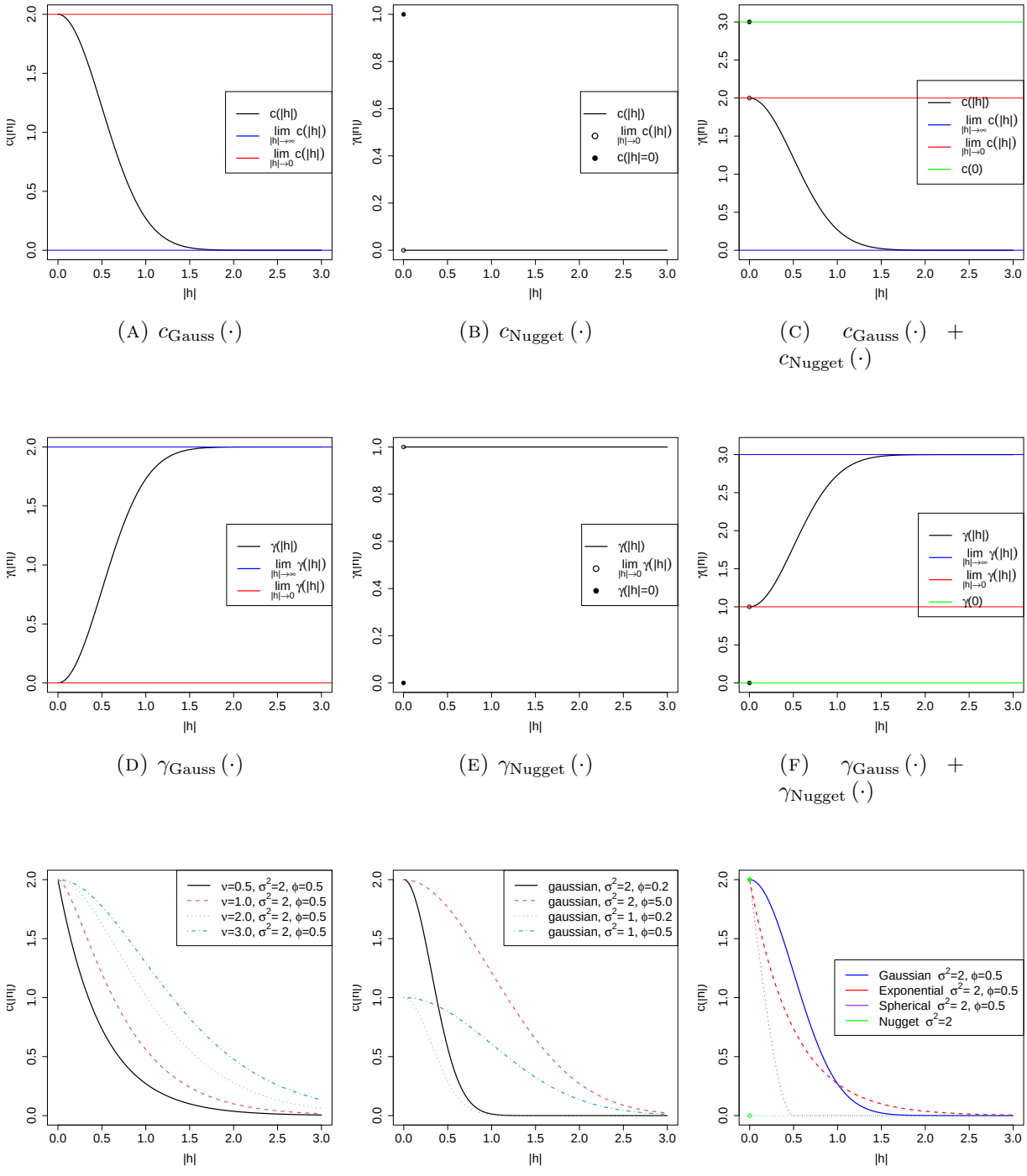
which is infinite differentiable. ϕ is a range parameter, and σ^2 is the (partial) sill parameter.

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

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

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

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

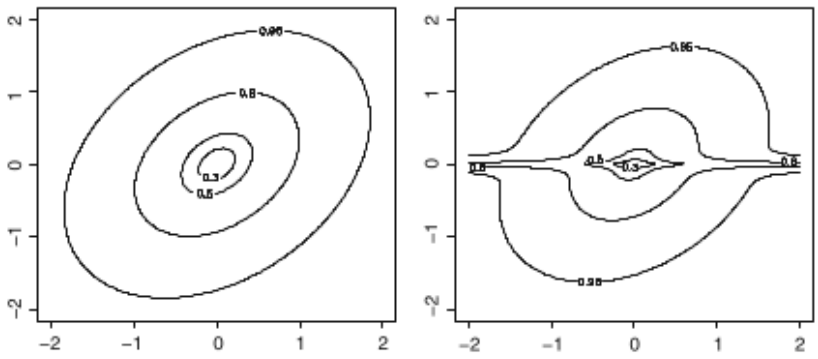


(A) $c_{\text{Gauss}}(\cdot)$ (B) $c_{\text{Nugget}}(\cdot)$ (C) $c_{\text{Gauss}}(\cdot) + c_{\text{Nugget}}(\cdot)$

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

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

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



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 10.1. Isotropy vs Anisotropy

10. ANISOTROPY

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

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

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

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

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

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

10.1. **Geometric anisotropy.**

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

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

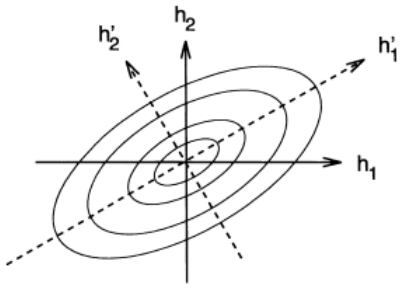


FIGURE 10.2. Rotation of the 2D coordinate system

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

10.3. **Non-linear deformations.**

Note 76. A (rather too general) non-stationary model can be specified by considering semi-variogram $2\text{Var}(Z(s) - Z(t)) = 2\gamma_o(\|G(s) - G(t)\|)$ 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

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

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

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

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

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

- *If Z is intrinsically stationary, then*

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

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

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

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

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

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

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

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

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

Part 2. Model building

12. The Geostatistical model

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

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

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

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

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

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

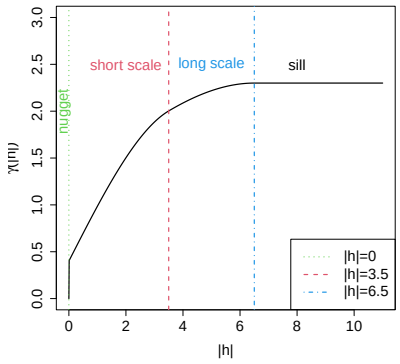


FIGURE 12.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|;\sigma^2 = 0.4\right)$.

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

12.2. **Scale of variation.**

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note 92. A simpler decomposition is

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

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

13. Training & inference

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

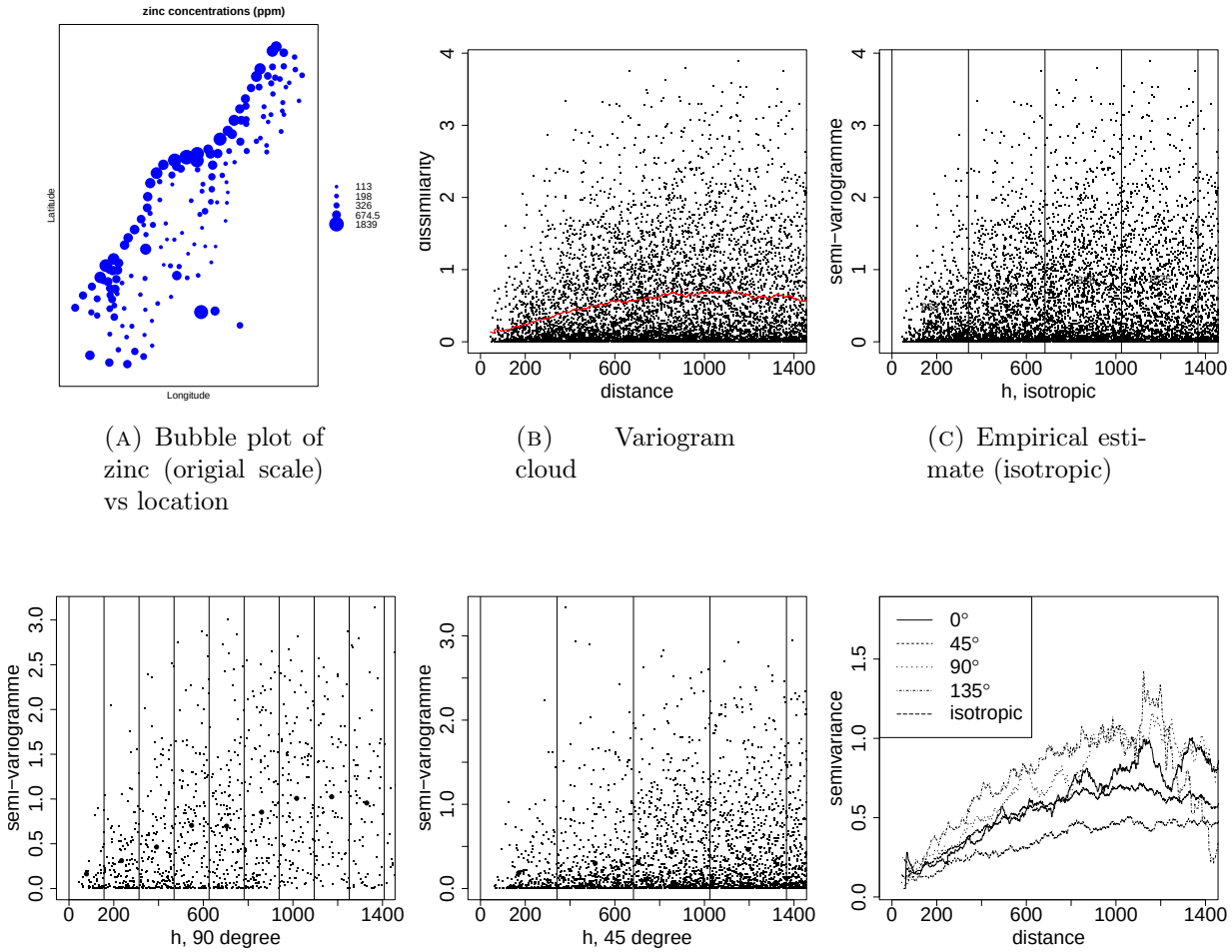
Example 94. (Meuse river data set) The Meuse river dataset set, used as a running example gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein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

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

13.1. **The variogram cloud.**

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

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



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

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

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

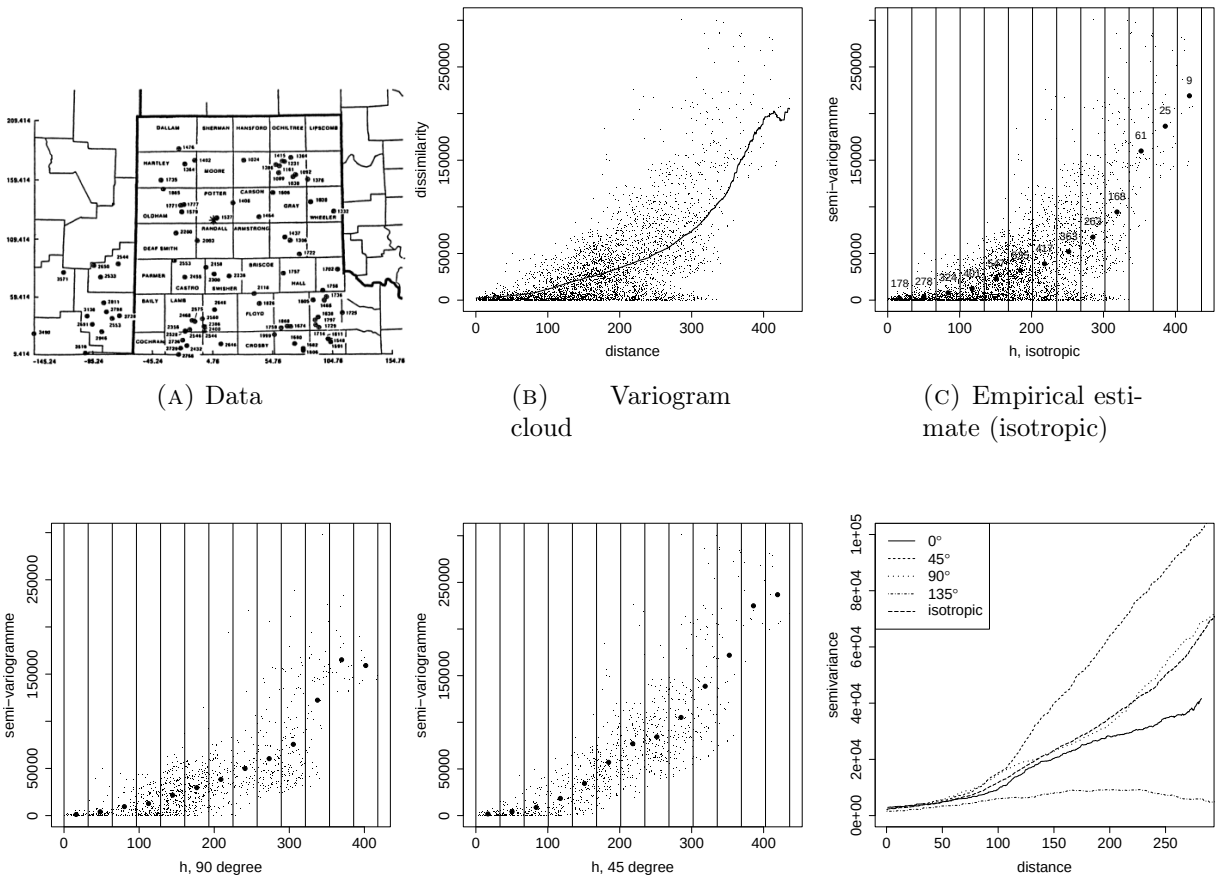
Definition 97. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (distance and orientation) then we get

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

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

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

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



(A) Data

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 13.2. Wolfcamp-aquifer dataset variogram estimations

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

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

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

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

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

where

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

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

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

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

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

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

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

because the direction does not have any effect.

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

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

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

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

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

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

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

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

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

Note 112. Estimation of the covariogram is done by

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

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

13.3. **Parametric estimation.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

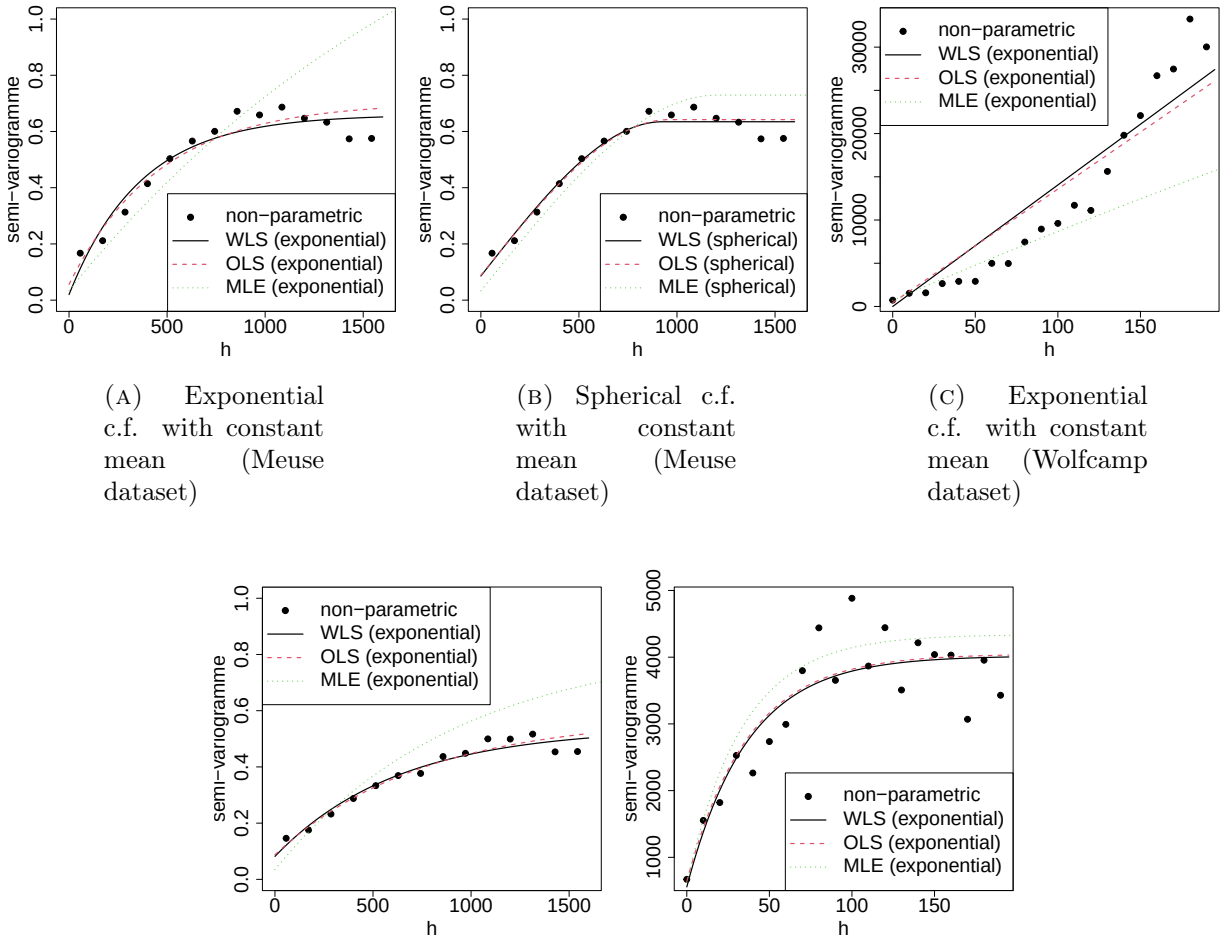
$$Z(s) = \mu(s;\beta) + \delta(s;\theta)$$

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

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

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

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



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

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

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

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

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

FIGURE 13.3. Parametric training

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

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

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

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

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

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

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

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

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

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

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

13.5. **Training via Bayesian statistics.**

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

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

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

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

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

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

and the posterior is given by the Bayes theorem as

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

The parametric variogram can be estimated via

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

14. The (traditional) Kriging paradigm

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

14.1. **Universal Kriging.**

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

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

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

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

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

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

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

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

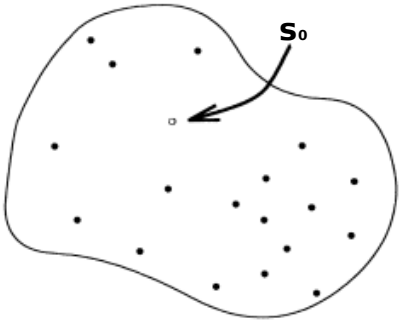


FIGURE 14.1. Kriging area

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

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

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

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

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

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

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

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

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

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

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

I write it in matrix and form FYI

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

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

(14.5)

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

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

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

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

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

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

Note 140. The UK system of equations is

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

(14.10)

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

(14.11)

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

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

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

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

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

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

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

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

with standard error

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

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

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

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

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

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

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

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

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

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

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

(2) Compute the residuals

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

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

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

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

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

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

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

14.2. **Ordinary Kriging.**

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

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

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

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

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

the weights are

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

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

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

14.3. **Simple Kriging.**

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

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

where the deterministic mean $\mu(s)$ is known, and (δ_s) is a weakly stationary process with covariogram $c(\cdot)$. The derivation is in (Exercise 17 Exercise sheet). It does not require any assumption in the weights such as 14.4 or (14.21). The SK predictor at s_0 and standard error are

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

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 151. Consider the example with the Meuse dataset. Fig 14.2a presents the OK prediction $Z_{\text{OK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (14.20) that is the UK case (14.1) for when $\mu(s) = \beta_0$. First we computed the non-parametric semivariogram $\hat{\gamma}$ (13.2) as in Prop 102; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ where we computed the OLS $\hat{\theta}_{\text{OLS}} = \left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (13.7) (see Fig. 13.3a); and then we pluged 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 152. 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 153. Consider a dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $Z_i = Z(s_i)$ being a realization of (15.1) at site $s_i \in \mathcal{S}$. Let $Z = (Z_1, ..., Z_n)^\top$, and $Y = (Y_1, ..., Y_n)^\top$.

Recall

Note 154. Uncertainty can be decomposed according to the Hierarchical spatial model

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

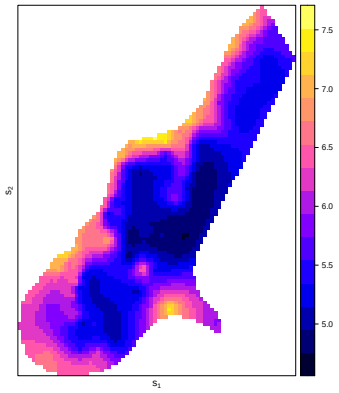
with

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

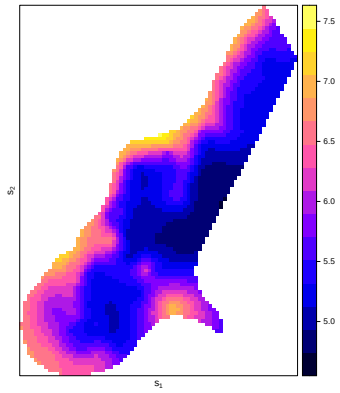
Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\text{pr}(Y|\vartheta)$ possibly labeled by some parameter ϑ.

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

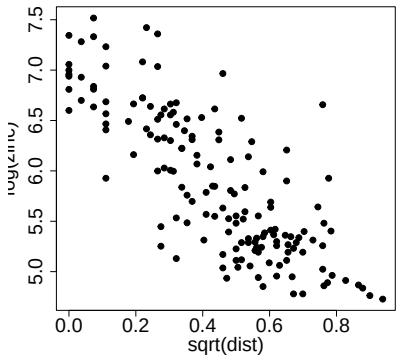
Note 155. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \text{pr}(\vartheta_1|\vartheta_2)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior; it can be considered sometimes as known and sometimes as unknown in what follows. (!)



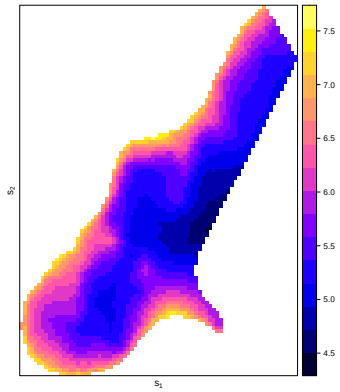
(A) Ordinary Kriging, constant mean, exponential covariogram



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



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



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

FIGURE 14.2. Kriging Meuse dataset.

Note 156. Then the Bayesian spatial hierarchical model becomes

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

where uncertainty is described by

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

Note 157. Under Bayesian model (15.3), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

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

by maximizing the marginal likelihood

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

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

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

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

Note 158. 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 159. 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 160. *Let $X \sim N(\mu_X, \Sigma_X)$ $Y \sim N(\mu_Y, \Sigma_Y)$ and Y, X independent. Let fixed matrices A and B and vector c of appropriate sizes. Then*

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

Fact 161. *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 162. *[Marginalization & conditioning] Let $x_1 \in \mathbb{R}^{d_1}$,* and $x_2 \in \mathbb{R}^{d_2}$. If

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

then it is

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

where

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

Note 163. To demonstrate how to work in the "Bayesian Kriging" framework e.g., with the spatial hierarchical models (15.2) and (15.3), we are going through a particular example of the Bayesian Gaussian process regression (or Bayesian Kriging).

A possible narrative - a story.

Note 164. Assume there is available a dataset $\{(s_i, Z_i)\}_{i=1}^n$ where $Z_i = Z(s_i)$ is a realization of a stochastic process (Z_s) with $\{Z_i \in \mathbb{R}\}$. In particular, assume that data are instances of an unknown function $Y(\cdot)$ at s_i but contaminated by additive random noise $\{\varepsilon_i \sim \mathrm{N}(0, \tau^2); i = 1, ..., n\}$ with scale $\tau > 0$; i.e. $Z_i = Y(s_i) + \varepsilon_i$.

Note 165. Assume we are interested in recovering $Z(\cdot)$.

Specifying the hierarchical model.

Note 166. A natural model to cast this problem is the geostatistical model

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

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

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

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

$$Y(\cdot)|\beta,\theta \sim \text{GP}(\mu(\cdot|\beta), c_Y(\cdot,\cdot|\theta)) \tag{15.7}$$

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

Note 167. Given (15.6) and (15.7), the Bayesian model (15.2) is

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

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

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

Note 168. The marginal process (Z_s) given parameters β, $\theta = (\sigma^2,\phi,\tau)^\top$ (in (15.8)) is

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

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, $\theta = (\sigma^2,\phi,\tau)^\top$. *[We used the additive property of Gaussian random variables in Fact* 160 *].*

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

Note 169. Assume a vector of "unseen" sites $S_* = (s_{*,1},...,s_{*,q})^\top$ for any $q \in \mathbb{N}_0$. Let convenient notation $Z := Z(S)$, and $Z_* := Z(S_*)$. The join marginal distribution of $(Z_*,Z)^\top$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is

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

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

Note 170. Given that vector Z is observed/known, the (posterior) predictive distribution of $Z_*|Z$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is the conditional distribution

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

where

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

[We used the formula for computing the conditional Gaussian distribution in Fact 162].

Note 171. Since the derivation of (15.10) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (15.10) can be extended to a Gaussian Process

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

with

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

for any $s, s' \in \mathcal{S}$. This is the predictive process of $Z(s)$ at any $s \in \mathcal{S}$ given Z, β, $\theta = (\sigma^2, \phi, \tau)^\top$. *[Here we used the definition of GP (Def 14) given Note* 170*].*

Note 172. Assume that the parameters (β, θ) are unknown but fixed (i.e. no prior is specified). Training can be performed by maximizing the marginal likelihood of Z given β, θ

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

derived from (15.9) by solving

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

Note 173. The estimated "Kriking predictor" results by plugging $\left(\hat{\beta},\hat{\theta}\right)^\top$ in (15.11), as

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

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

Note 174. Assume β is an unknown random hyper-parameter in the sense we assign a prior distribution on it to account for uncertainty. Hence, we will specify a conjugate distribution on β, and compute the produced predictive distribution.

Note 175. Like in Universal Kriging, assume that the spatial mean is parameterized as an expansion of basis functions $\psi(\cdot) = (\psi_1(\cdot), ..., \psi_p(\cdot))^\top$ with unknown coefficients β, i.e.

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

Note 176. The marginal process (Z_s) given parameters β, $\theta = (\sigma^2, \phi, \tau)^\top$ can be re-writen as

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

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

Note 177. We specify a conjugate prior $\beta|\sigma^2 \sim \mathrm{N}(b, \sigma^2 B)$ on β, for some user-specified fixed hyper-parameters b and B.

Note 178. The marginal Bayesian model is now extended to

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

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

Note 179. The posterior of β given data Z and θ is computed via the Bayes theorem

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

and results as

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

with

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

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

Note 180. The posterior predictive distribution of $Z\left(\cdot\right)$ given the data Z and θ , results by integrating (15.11) with respect to (15.14) i.e.

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

and it is again a GP

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

with

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

(15.17)

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

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

Note 181. If we consider non-informative priors in (15.13) such as $\text{pr}(\beta|\sigma^2) \propto 1$, for instance, by allowing $B \to 0$, and $b < \infty$ then (15.17) produces the Universal Kriging predictor (check with (14.14)).

Note 182. Assume that $\theta = (\sigma^2, \phi, \tau)^\top$ is an unknown fixed hyper-parameter without a prior distribution being specified. Training can be performed by maximizing the marginal likelihood of Z given θ

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

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

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

[from Fact 161] by computing

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

Note 183. The estimated "Kriging predictor" results by plugging $\hat{\theta}$ in (15.15)

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

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

we specify a conjugate prior on σ^2 and then we follow the same routine. *[It will be given as a homework.]*

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 3: Point referenced data modeling / Geostatistics

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

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

Reading list & references:

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

Specialized reading.

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

Part 1. Intro to building stochastic models & concepts

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

1. Stochastic processes (or Random fields)

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

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

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

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

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

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

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

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

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

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

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

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

is a well defined stochastic process as it satisfies Thm 5.

1.1. **Mean and covariance functions.**

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

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

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

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

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

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

Proposition 9. *The function $c : S \times S \to \mathbb{R}$, $S \subseteq \mathbb{R}^d$ is the covariance function iff $c(\cdot,\cdot)$ is semi-positive definite; i.e. the Gram matrix $(c(s_i, s_j))_{i,j=1}^n$ is non-negative definite for any $\{s_i\}_{i=1}^n$, $n \in \mathbb{N}$.*

Example 10. $c(s,s') = 1(s = s')$ is a proper covariance function as

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

Note 11. Prop 12 uses the experience from basis functions, while Theorem 30 uses experience from characteristic functions to be incorporated int he process for modeling reasons.

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

Proof. From Prop 9, as

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

□

2. Second order processes (or Random fields)

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

3. Gaussian process

Definition 14. $Z = (Z_s; s \in S)$ indexed by $S \subseteq \mathbb{R}^d$ is a Gaussian process (GP) or random field (GRF) if for any $n \in \mathbb{N}$ and for any finite set $\{s_1, ..., s_n; s_i \in \mathcal{S}\}$, the random vector $(Z_{s_1}, ..., Z_{s_n})^\top$ has a multivariate normal distribution.

Also Example of Proposition

Proposition 15. *A GP $Z = (Z_s; s \in S)$ is fully characterized by its mean function $\mu : S \to \mathbb{R}$ with $\mu(s) = E(Z_s)$, and its covariance function with $c(s,s') = Cov(Z_s, Z_{s'})$.*

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

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

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

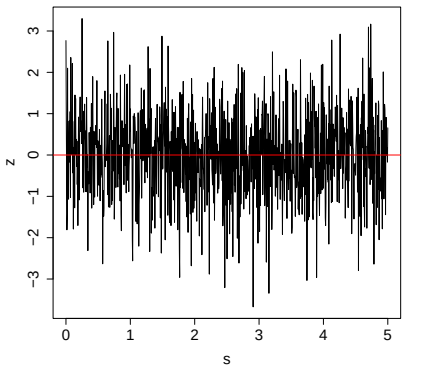
Example 18. Recall your linear regression lessons where you specified a sampling distribution $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta, \sigma^2\right)$, $\forall x \in \mathbb{R}^d$; well that can be considered as a GP with $\mu_x = x^\top\beta$ and $c(x,x') = \sigma^2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

Example 19. Figs. 3.1 & 3.2 presents realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(s) = 0$ and differently parameterized covariance functions in 1D and 2D. In 1D the code to simulate the GP is given in Algorithm 1. Note that we actually discretize it and simulate it from the fidi.

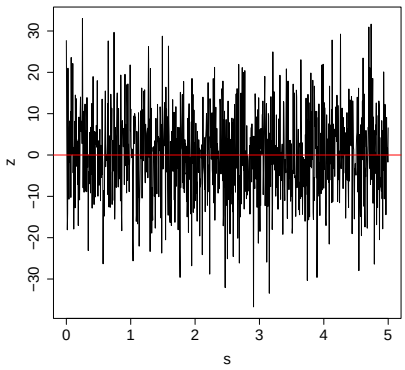
Algorithm 1 R script for simulating from a GP $(Z_s; s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

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

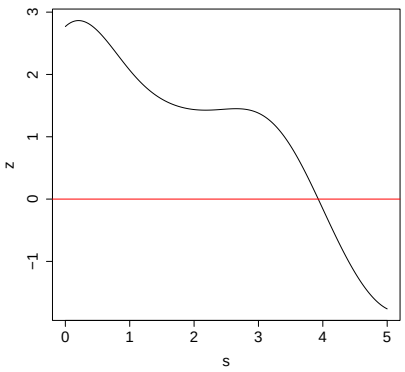
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Fig. 3.1a & 3.1b ; Fig. 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Fig. 3.2c & 3.2d), and the spatial dependence / frequency of the ups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



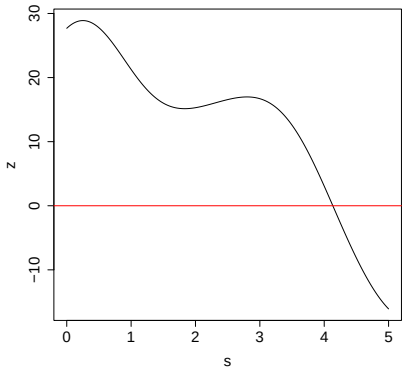
(A) Nugget c.f
($\sigma^2 = 1$)



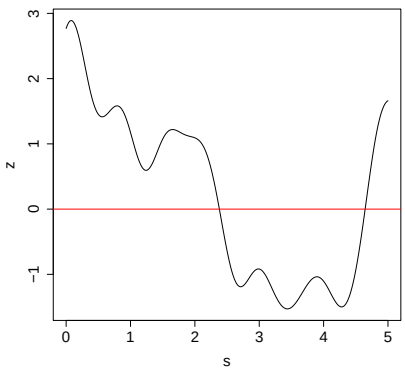
(B) Nugget c.f
($\sigma^2 = 100$)



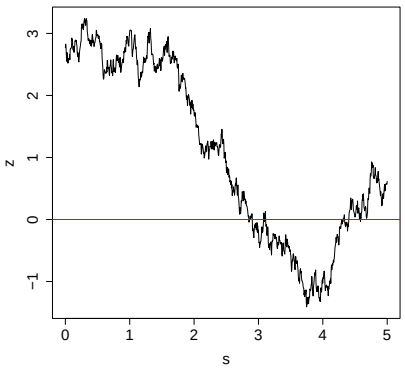
(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



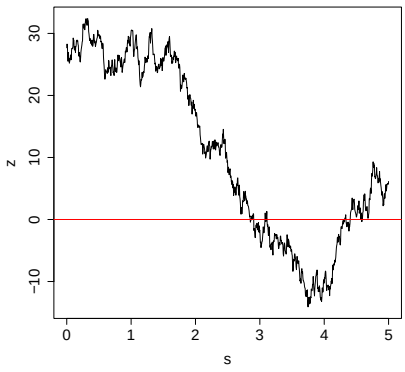
(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



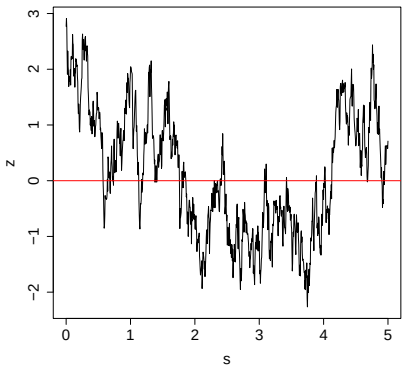
(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)



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

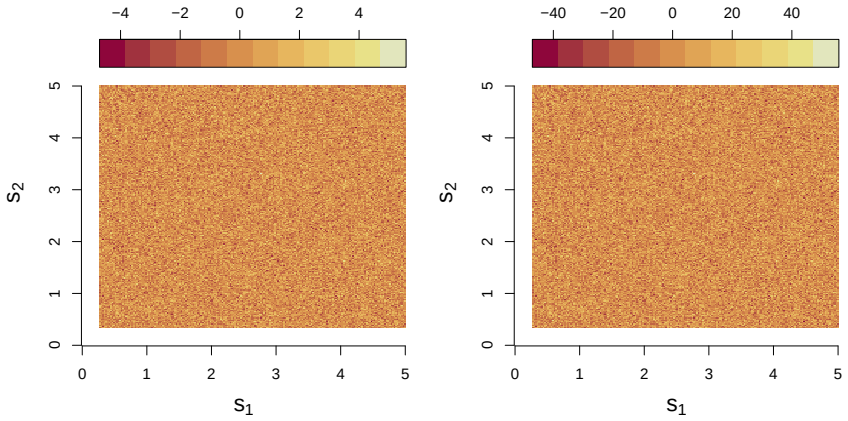


(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)



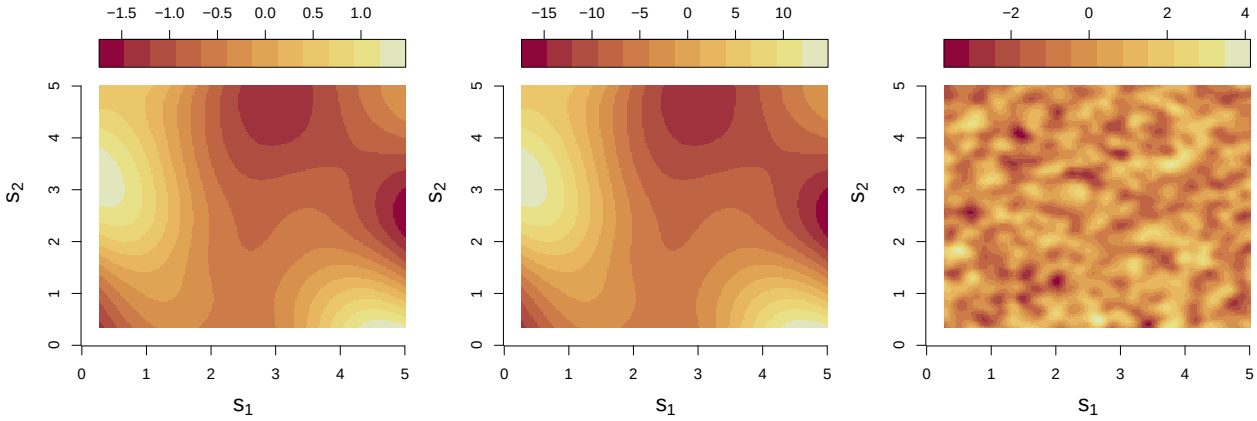
(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

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



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

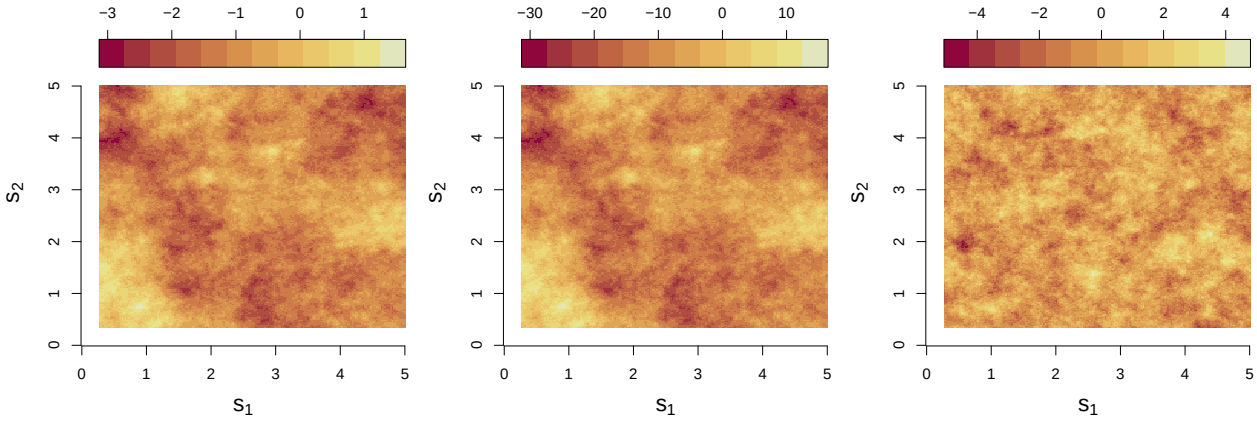
(B) Nugget c.f
($\sigma^2 = 100$)



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

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

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



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

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

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

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

4. Strong stationarity

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

Definition 21. A random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ is strongly stationary if for all finite sets consisting of $s_1, ..., s_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, for all $k_1, ..., k_n \in \mathbb{R}$, and for all $h \in \mathbb{R}^d$

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

5. Weak stationarity (or second order stationarity)

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

Definition 23. A random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ is weakly stationary (or second order stationary) if, for all $s, s' \in \mathbb{R}^d$,

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

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

6. Covariogram

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

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

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

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

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

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

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

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

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

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

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

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

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

Note 29. The following helps in the specification of cavariograms by considering properties of characteristic functions.

Theorem 30. *(Bochner's theorem) A continuous even real-valued function $c : \mathbb{R}^d \to \mathbb{R}$ is a covariance function of a weakly stationary random process if and only if it can be represented as*

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

where $\mathrm{d}F(\omega)$ is a *symmetric positive finite measure on* $\mathbb{R}^d$.

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

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

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

Theorem 31. *If $c(\cdot)$ is integrable, $F(\cdot)$ is absolutely continuous with spectral density $f(\cdot)$ of $Z = (Z_s; s \in \mathcal{S})$ then by Fast Fourier transformation*

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

Example 32. Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Then the spectral density from Thm 30 is

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

i.e. of a Gaussian form.

7. Intrinsic stationarity

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

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

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

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

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

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

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

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

and

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

8. (Semi) variogram

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

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

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

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

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

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

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

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

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

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

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

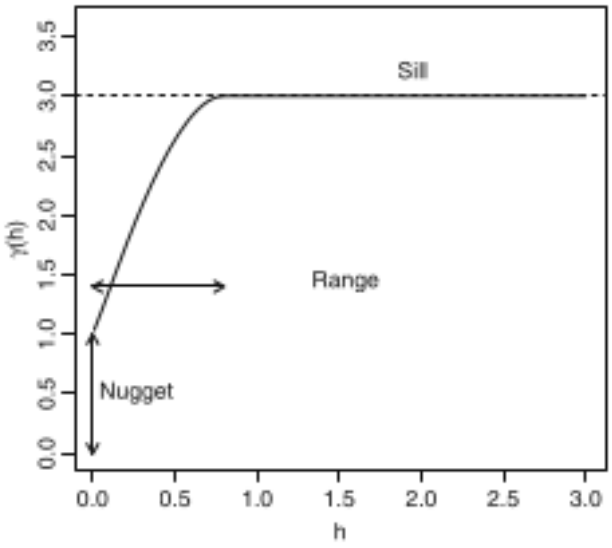


FIGURE 8.1. Variogram's characteristics

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

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

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

Nugget Effect.

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

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

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

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

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

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

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

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

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

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

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

Sill.

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

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

Partial sill.

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

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

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

9. Isotropy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

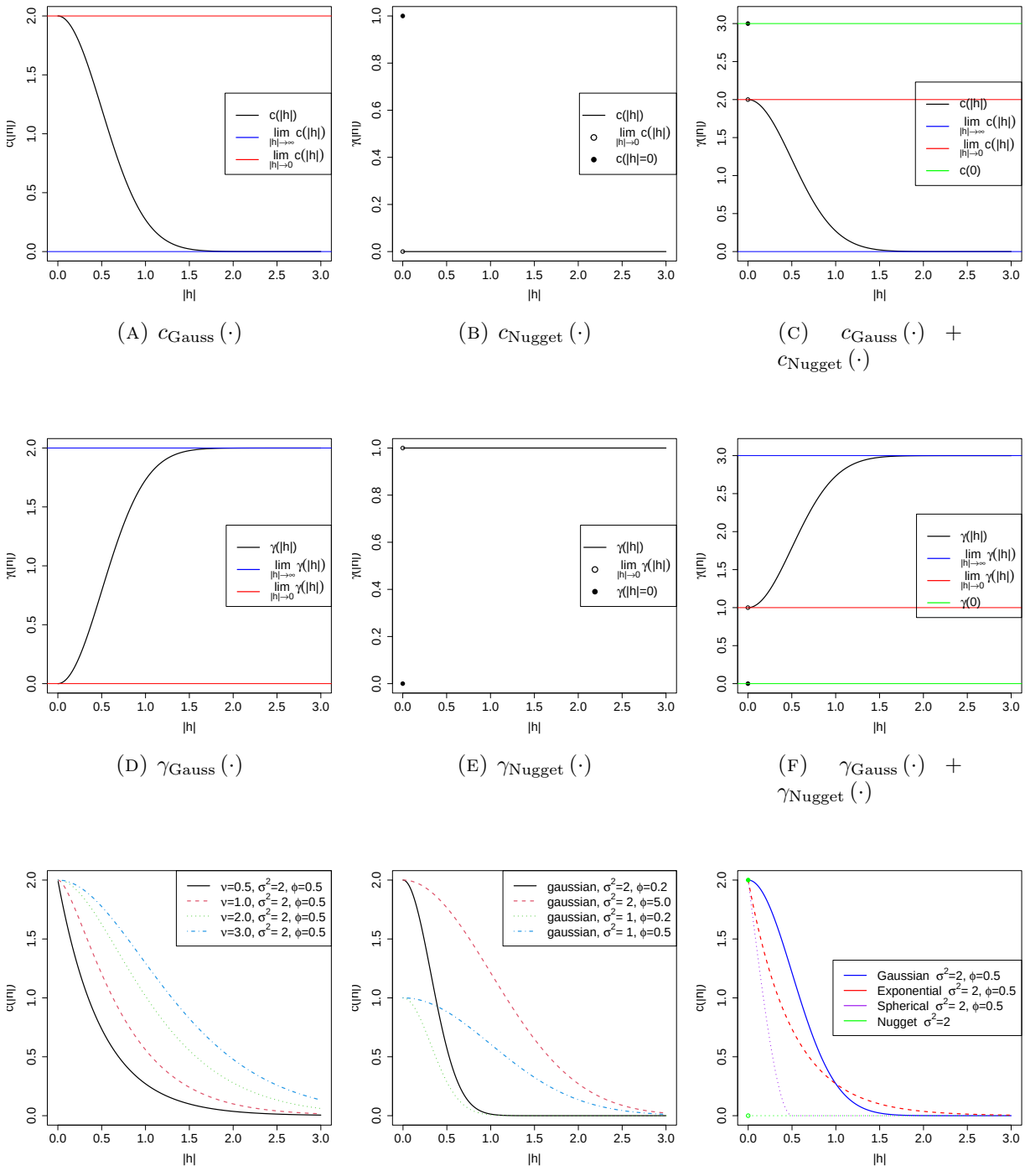
which is infinite differentiable. ϕ is a range parameter, and σ^2 is the (partial) sill parameter.

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

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

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

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

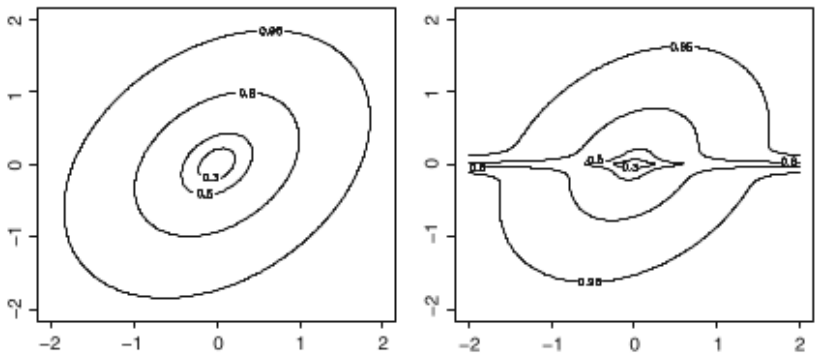


(A) $c_{\text{Gauss}}(\cdot)$ (B) $c_{\text{Nugget}}(\cdot)$ (C) $c_{\text{Gauss}}(\cdot) + c_{\text{Nugget}}(\cdot)$

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

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

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



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 10.1. Isotropy vs Anisotropy

10. ANISOTROPY

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

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

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

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

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

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

10.1. **Geometric anisotropy.**

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

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

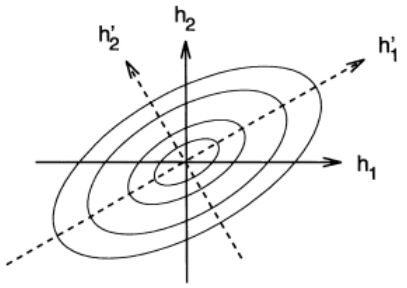


FIGURE 10.2. Rotation of the 2D coordinate system

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

10.3. **Non-linear deformations.**

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

11. Geometrical properties

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

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

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

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

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

- *If Z is intrinsically stationary, then*

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

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

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

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

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

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

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

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

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

Part 2. Model building

12. The Geostatistical model

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

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

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

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

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

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

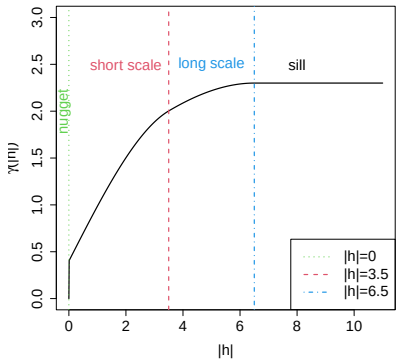


FIGURE 12.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|;\sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|;\sigma^2 = 0.4\right)$.

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

12.2. **Scale of variation.**

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note 92. A simpler decomposition is

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

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

13. Training & inference

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

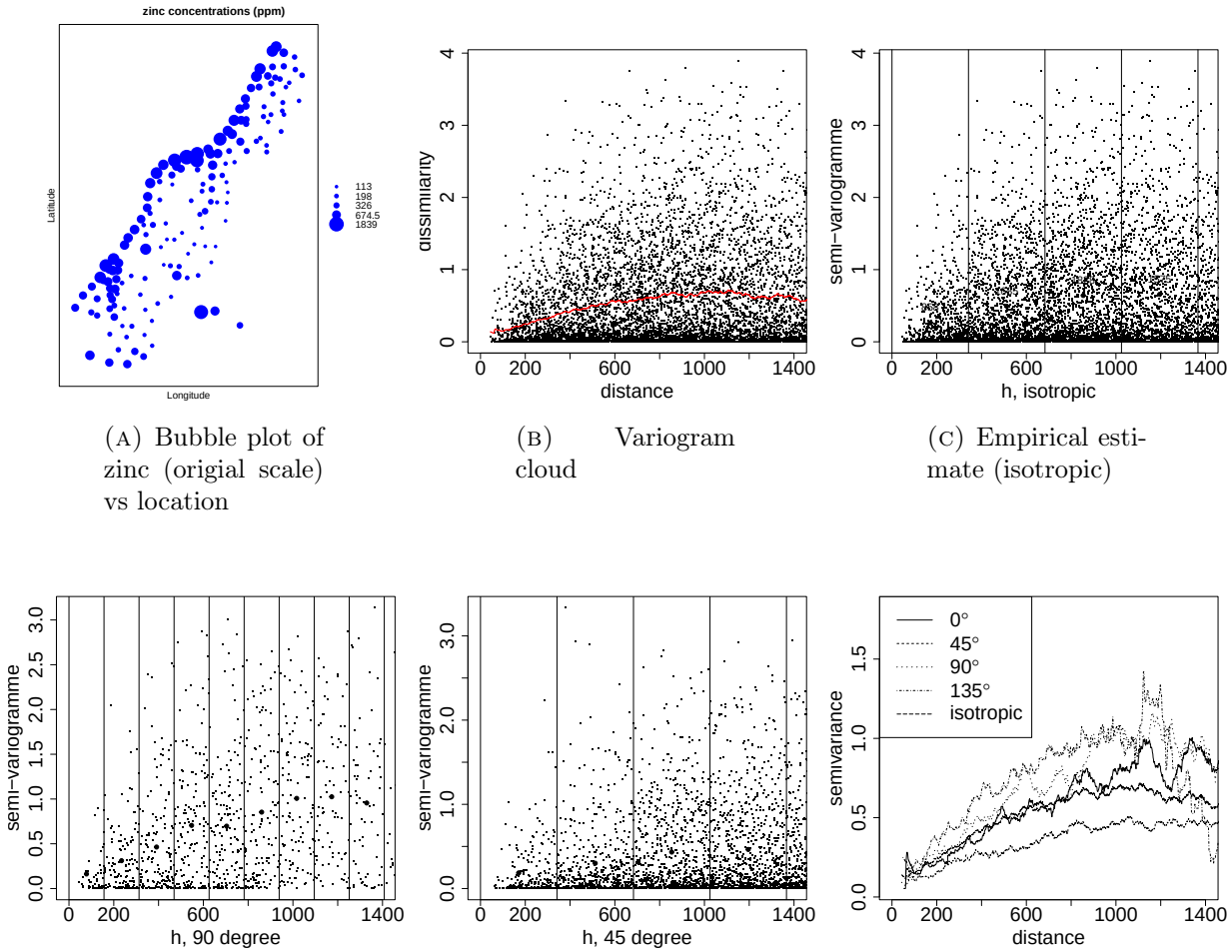
Example 94. (Meuse river data set) The Meuse river dataset set, used as a running example gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein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

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

13.1. **The variogram cloud.**

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

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



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

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

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

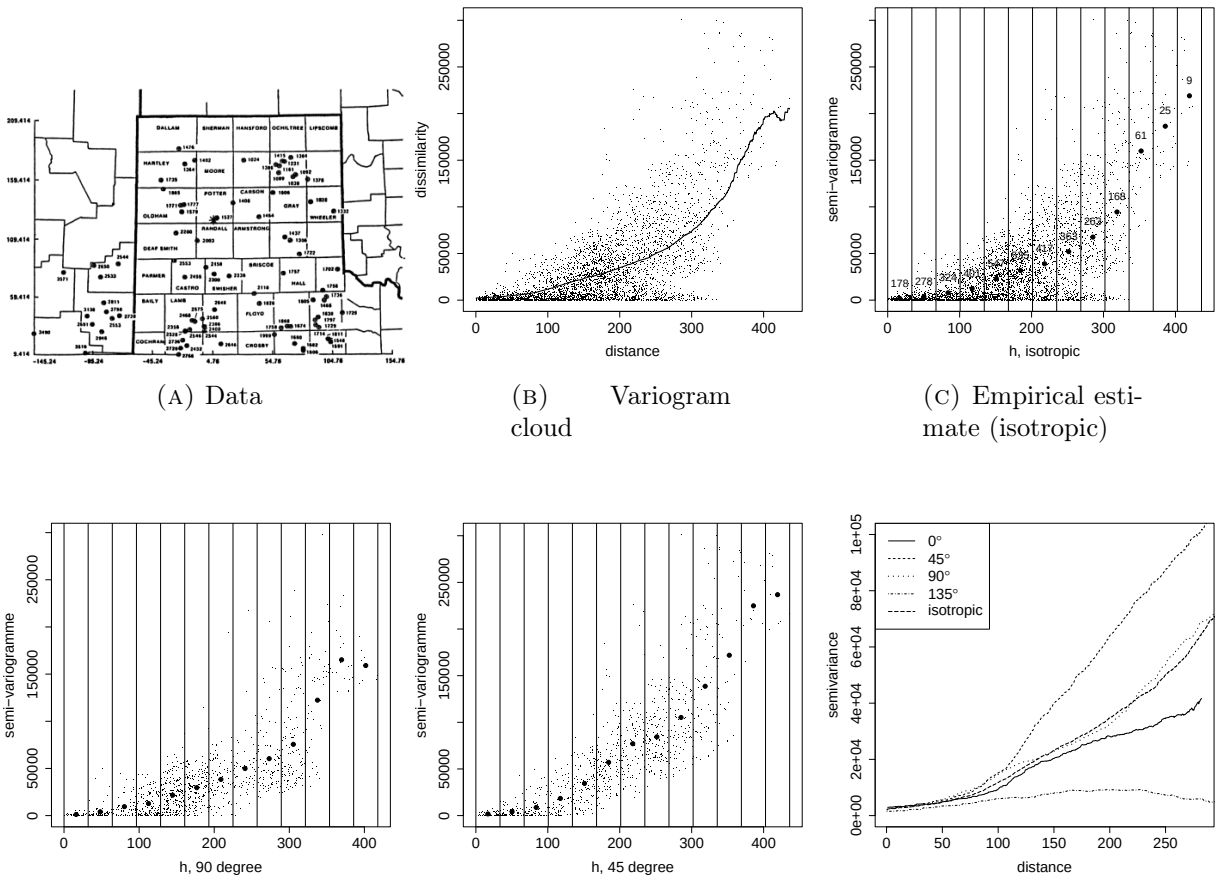
Definition 97. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (distance and orientation) then we get

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

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

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

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



(A) Data

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 13.2. Wolfcamp-aquifer dataset variogram estimations

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

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

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

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

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

where

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

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

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

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

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

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

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

because the direction does not have any effect.

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

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

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

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

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

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

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

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

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

Note 112. Estimation of the covariogram is done by

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

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

13.3. **Parametric estimation.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

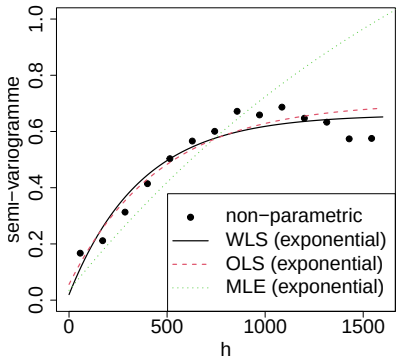
$$Z(s) = \mu(s;\beta) + \delta(s;\theta)$$

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

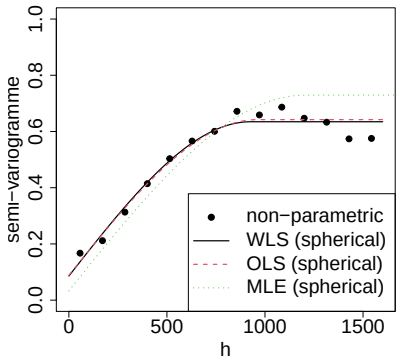
Proposition 122. *(Least square errors with trend) Do the following:*

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

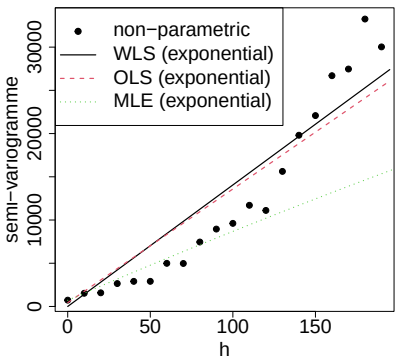
$$\hat{\beta}_{LSE} = \arg\min_\theta \left(\sum_i \left(Z(s_i) - \mu(s_i;\beta)\right)^2\right)$$



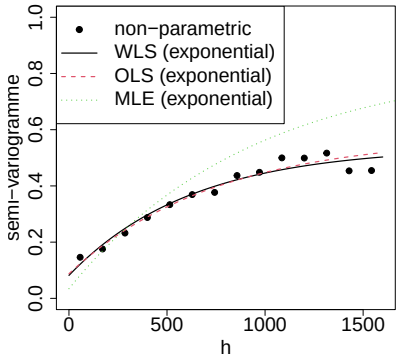
(A) Exponential c.f. with constant mean (Meuse dataset)



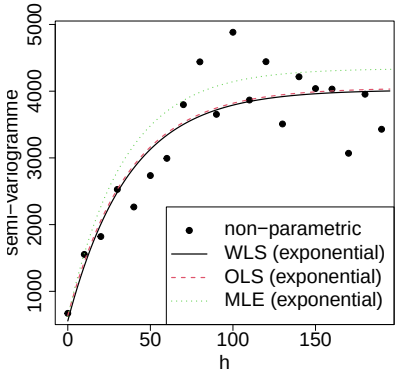
(B) Spherical c.f. with constant mean (Meuse dataset)



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



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



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

FIGURE 13.3. Parametric training

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

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

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

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

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

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

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

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

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

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

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

13.5. **Training via Bayesian statistics.**

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

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

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

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

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

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

and the posterior is given by the Bayes theorem as

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

The parametric variogram can be estimated via

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

14. The (traditional) Kriging paradigm

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

14.1. **Universal Kriging.**

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

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

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

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

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

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

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

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

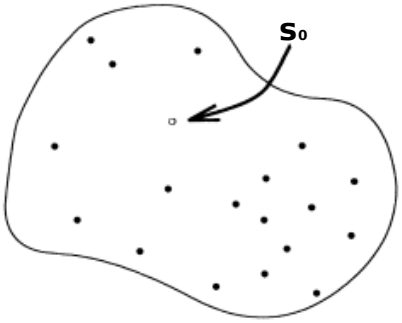


FIGURE 14.1. Kriging area

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

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

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

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

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

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

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

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

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

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

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

I write it in matrix and form FYI

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

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

(14.5)

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

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

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

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

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

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

Note 140. The UK system of equations is

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

(14.10)

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

(14.11)

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

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

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

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

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

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

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

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

with standard error

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

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

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

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

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

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

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

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

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

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

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

(2) Compute the residuals

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

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

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

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

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

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

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

14.2. **Ordinary Kriging.**

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

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

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

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

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

the weights are

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

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

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

14.3. **Simple Kriging.**

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

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

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

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

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

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

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

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

15. The Bayesian Kriging paradigm (Hierarchical modeling)

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

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

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

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

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

Recall

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

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

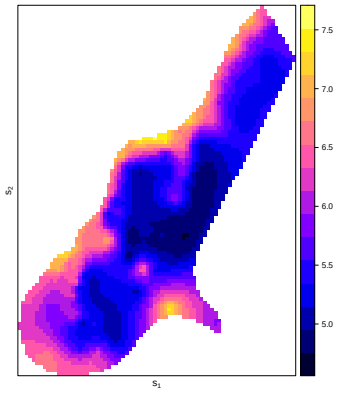
with

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

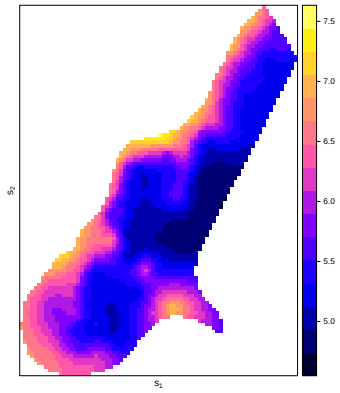
Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\text{pr}(Y|\vartheta)$ possibly labeled by some parameter ϑ.

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

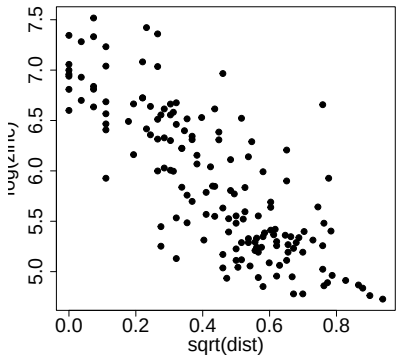
Note 156. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \text{pr}(\vartheta_1|\vartheta_2)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior;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



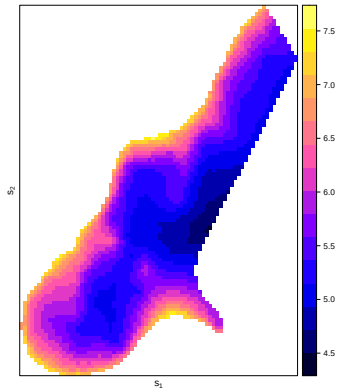
(A) Ordinary Kriging, constant mean, exponential covariogram



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



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



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

FIGURE 14.2. Kriging Meuse dataset.

Note 157. Then the Bayesian spatial hierarchical model becomes

$$\begin{cases} Z|Y,\vartheta_1,\vartheta_2 & \text{data model} \\ Y|\vartheta_1,\vartheta_2 & \text{spatial process model} \\ \vartheta_1|\vartheta_2 & \text{hyper-parameter prior model} \end{cases} \tag{15.3}$$

where uncertainty is described by

$$\mathrm{pr}\left(Z, Y, \vartheta_1|\vartheta_2\right) = \mathrm{pr}\left(Z|Y, \vartheta_1|\vartheta_2\right)\mathrm{pr}\left(Y|\vartheta_1, \vartheta_2\right)\mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$$

Note 158. Under Bayesian model (15.3), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

$$\hat{\vartheta}_2 = \arg\min_{\vartheta_2}\left(-2\log\left(\mathrm{pr}\left(Z|\vartheta_2\right)\right)\right)$$

by maximizing the marginal likelihood

$$\mathrm{pr}\left(Z|\vartheta_2\right) = \int \mathrm{pr}\left(Z, Y, \vartheta_1|\vartheta_2\right)\mathrm{d}Y\mathrm{d}\vartheta_1$$

Under Bayesian model (15.3), when ϑ_1 is considered as unknown (but random), namely, the a prior $\vartheta_1 \sim \mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$ has been specified, uncertainty about unknown ϑ_1 given Y and ϑ_2 can be represented by the posterior distribution

$$\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \frac{\mathrm{pr}\left(Z|\vartheta_1, \vartheta_2\right)\mathrm{pr}\left(\vartheta_1|\vartheta_2 = \hat{\vartheta}_2\right)}{\mathrm{pr}\left(Z|\vartheta_2 = \hat{\vartheta}_2\right)}$$

where the value $\hat{\vartheta}_2$ is plugged in.

Note 159. General interest lies in computing the posterior predictive distributions of the spatial process model (Y_s), (or latent process, or noiseless process) given the data Z

$$\mathrm{pr}\left(Y\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Y\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}\vartheta_1$$

and / or the marginal process (Z_s) given the data

$$\mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{pr}\left(\vartheta_1|Z, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}\vartheta_1$$

$$\mathrm{pr}\left(Z\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right) = \int \mathrm{pr}\left(Z\left(s_0\right), Y\left(s_0\right)|Z, \vartheta_1, \vartheta_2 = \hat{\vartheta}_2\right)\mathrm{d}Y\left(s_0\right)$$

for any $s_0 \in \mathcal{S}$.

Note 160. The above statistical problem is naturally addressed in the (either full or empirical) Bayesian statistical framework. It is often called Bayesian Kriging.

15.2. **Bayesian Kriging (Gaussian process regression).**

Inventory of useful formulas.

Fact 161. *Let $X \sim N(\mu_X, \Sigma_X)$ $Y \sim N(\mu_Y, \Sigma_Y)$ and Y, X independent. Let fixed matrices A and B and vector c of appropriate sizes. Then*

$$AX + BY + c \sim N\left(A\mu_X + B\mu_Y + c, A\Sigma_X A^\top + B\Sigma_Y B^\top\right) \tag{15.4}$$

Fact 162. *Let $N(\beta|b, B)$ be the Gaussian pdf with mean b and covariance B at β. It is*

$$\int N(Z|\Psi\beta, C)\, N(\beta|b, B)\, d\beta = N\left(Z|\Psi b, C + \Psi B \Psi^\top\right)$$

Fact 163. *[Marginalization & conditioning] Let $x_1 \in \mathbb{R}^{d_1}$, and $x_2 \in \mathbb{R}^{d_2}$. If*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathrm{N}_{d_1+d_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix}\right)$$

then it is

$$x_2|x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, \Sigma_{2|1}\right)$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_1 - \mu_1) \text{ and } \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top$$

Note 164. To demonstrate how to work in the "Bayesian Kriging" framework e.g., with the spatial hierarchical models (15.2) and (15.3), we are going through a particular example of the Bayesian Gaussian process regression (or Bayesian Kriging).

A possible narrative - a story.

Note 165. Assume there is available a dataset $\{(s_i, Z_i)\}_{i=1}^n$ where $Z_i = Z(s_i)$ is a realization of a stochastic process (Z_s) with $\{Z_i \in \mathbb{R}\}$. In particular, assume that data are instances of an unknown function $Y(\cdot)$ at s_i but contaminated by additive random noise $\{\varepsilon_i \sim \mathrm{N}(0, \tau^2); i = 1, ..., n\}$ with scale $\tau > 0$; i.e. $Z_i = Y(s_i) + \varepsilon_i$.

Note 166. Assume we are interested in recovering $Z(\cdot)$.

Specifying the hierarchical model.

Note 167. A natural model to cast this problem is the geostatistical model

$$Z(s) = Y(s) + \varepsilon(s), \quad s \in \mathcal{S} \tag{15.5}$$

- we specify a zero-mean Gaussian process $\varepsilon(\cdot) \sim \mathrm{GP}(0, c_\varepsilon(\cdot, \cdot|\tau))$ with nugget covariogram $c_\varepsilon(s, s'|\tau) = \tau^2 1_{\{0\}}(\|s - s'\|)$ to represent the noise. Hence

$$Z(\cdot)|Y(\cdot), \tau \sim \mathrm{GP}(Y(\cdot), c_\varepsilon(\cdot, \cdot|\tau)). \tag{15.6}$$

- To quantify uncertainty of the unknown $Y(\cdot)$, we specify a GP prior on $Y(\cdot)$

$$Y(\cdot)|\beta,\theta \sim \text{GP}\left(\mu(\cdot|\beta), c_Y\left(\cdot,\cdot|\sigma^2,\phi\right)\right) \tag{15.7}$$

  with mean function $\mu(\cdot|\beta)$ labeled by unknown parameter β and covariance function $c_Y(\cdot,\cdot|\sigma^2,\phi)$, labeled by unknown parameter $(\sigma^2,\phi)^\top$.
- we assume ε_s and Y_s to be independent.

Note 168. Given (15.6) and (15.7), the Bayesian model (15.2) is

$$\begin{cases} Z_i|Y_i,\tau^2 \overset{\text{ind}}{\sim} \text{N}(Y_i,\tau^2),\ i=1,...,n & \text{data model} \\ Y|\beta,\sigma^2,\phi \sim \text{N}(\mu(S|\beta), c_Y(S,S|\sigma^2,\phi)) & \text{spatial process model} \end{cases} \tag{15.8}$$

where $\vartheta = (\beta,\sigma^2,\phi)^\top$, $[\mu(S|\beta)]_i = \mu(s_i|\beta)$, and $[c_Y(S,S|\sigma^2,\phi)]_{i,j} = c_Y(s_i,s_j|\sigma^2,\phi)$.

Computing the marginal process $Z(\cdot)|\beta,\theta$.

Note 169. The marginal process (Z_s) given parameters β, $\theta = (\sigma^2,\phi,\tau)^\top$ (in (15.8)) is

$$Z(\cdot)|\beta,\theta \sim \text{GP}(\mu(\cdot|\beta), c(\cdot,\cdot|\theta)) \tag{15.9}$$

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, and covariance function parameters $\theta = (\sigma^2,\phi,\tau)^\top$. *[We used the additive property of Gaussian random variables in Fact* 161 *].*

Computing the predictive distribution $Z(\cdot)|Z,\beta,\theta$.

Note 170. Assume a vector of "unseen" sites $S_* = (s_{*,1},...,s_{*,q})^\top$ for any $q \in \mathbb{N}_0$. Let convenient notation $Z := Z(S)$, and $Z_* := Z(S_*)$. The join marginal distribution of $(Z_*,Z)^\top$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is

$$\begin{pmatrix} Z_* \\ Z \end{pmatrix}|\beta,\theta \sim \text{N}\left(\begin{pmatrix} \mu(S_*;\beta) \\ \mu(S;\beta) \end{pmatrix}, \begin{pmatrix} C(S_*,S_*|\theta) & (C(S_*,S|\theta))^\top \\ C(S_*,S|\theta) & C(S,S|\theta) \end{pmatrix}\right)$$

by using convenient notation $[C(S_*,S|\theta)]_{i,j} = s(s_{*,i},s_j|\theta)$ and $[\mu(S;\beta)]_i = \mu(s_i;\beta)$.

Note 171. Given that vector Z is observed/known, the (posterior) predictive distribution of $Z_*|Z$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is the conditional distribution

$$Z_*|Z,\beta,\theta \sim \text{N}(\mu_*(S_*|\beta,\theta), C_*(S_*,S_*|\theta)) \tag{15.10}$$

where

$$\begin{aligned} C_*(S_*,S_*|\theta) =& C(S_*,S_*|\theta) + (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} C(S,S_*|\theta) \\ \mu_*(S_*|\beta,\theta) =& \mu(S_*|\beta) - (C(S,S_*|\theta))^\top (C(S,S|\theta))^{-1} (\mu(S|\beta) - Z) \end{aligned}$$

[We used the formula for computing the conditional Gaussian distribution in Fact 163].

Note 172. Since the derivation of (15.10) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (15.10) can be extended to a Gaussian Process

$$Z(\cdot)|Z,\beta,\theta \sim \mathrm{GP}(\mu_1(\cdot|\beta,\theta), c_1(\cdot,\cdot|\theta)) \tag{15.11}$$

with

$$\begin{aligned} c_1(s,s'|\theta) =& c(s,s|\theta) + (C(S,s|\theta))^\top (C(S,S|\theta))^{-1} C(S,s'|\theta) \\ \mu_1(s|\beta,\theta) =& \mu(s|\beta) - (C(S,s|\theta))^\top (C(S,S|\theta))^{-1} (\mu(S|\beta) - Z) \end{aligned}$$

for any $s, s' \in \mathcal{S}$. This is the predictive process of $Z(s)$ at any $s \in \mathcal{S}$ given Z, β, θ. *[Here we used the definition of GP (Def 14) given Note* 171*].*

Note 173. Assume that the parameters (β, θ) are unknown but fixed (i.e. no prior is specified). Training can be performed by maximizing the marginal likelihood of Z given β, θ

$$\mathrm{pr}(Z|\beta,\theta) = \mathrm{N}(Z|\mu(S|\beta), C(S,S|\theta)) \tag{15.12}$$

derived from (15.9) by solving

$$\left(\hat{\beta},\hat{\theta}\right)^\top = \arg\min_{\beta,\theta} \left(-2\log\left(\mathrm{N}(Z|\mu(S|\beta), C(S,S|\theta))\right)\right)$$

Note 174. The estimated "Kriking predictor" results by plugging $\left(\hat{\beta},\hat{\theta}\right)^\top$ in (15.11), as

$$Z(\cdot)|Z,\hat{\beta},\hat{\theta} \sim \mathrm{GP}\left(\mu_1\left(\cdot|\hat{\beta},\hat{\theta}\right), c_1\left(\cdot,\cdot|\hat{\theta}\right)\right).$$

Computing the predictive distribution $Z(\cdot)|Z,\theta$.

Note 175. Assume β is an unknown random hyper-parameter in the sense we assign a prior distribution on it to account for uncertainty. Hence, we will specify a conjugate distribution on β, and compute the produced predictive distribution.

Note 176. Like in Universal Kriging, assume that the spatial mean is parameterized as an expansion of basis functions $\psi(\cdot) = (\psi_1(\cdot), ..., \psi_p(\cdot))^\top$ with unknown coefficients β, i.e.

$$\mu(s|\beta) = \psi(s)^\top \beta$$

Note 177. The marginal process (Z_s) given parameters β, and θ can be re-writen as

$$Z(\cdot)|\beta,\theta \sim \mathrm{GP}\left(\psi(s)^\top \beta, c(\cdot,\cdot|\theta)\right)$$

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, $\theta = (\sigma^2,\phi,\tau)^\top$

Note 178. We specify a conjugate prior $\beta|\sigma^2 \sim \mathrm{N}(b, \sigma^2 B)$ on β, for some user-specified fixed hyper-parameters b and $B > 0$.

Note 179. The marginal Bayesian model is now extended to

$$\begin{cases} Z|\beta,\theta \sim \mathrm{N}\left(\Psi\beta, C\left(S,S|\theta\right)\right) \\ \beta|\sigma^2 \sim \mathrm{N}\left(b,\sigma^2 B\right) \end{cases} \tag{15.13}$$

with matrix Ψ such as $\left[\Psi\right]_{i,j} = \psi_j\left(s_i\right)$.

Note 180. The posterior of β given data Z and θ is computed via the Bayes theorem

$$\begin{aligned} \mathrm{pr}\left(\beta|Z,\theta\right) &\propto \mathrm{pr}\left(Z|\beta,\theta\right)\mathrm{pr}\left(\beta|\theta\right) \\ &\propto \mathrm{N}\left(Z|\Psi\beta, C\left(S,S|\theta\right)\right)\mathrm{N}\left(\beta|b,\sigma^2 B\right) \end{aligned}$$

and results as

$$\beta|Z,\theta \sim \mathrm{N}\left(b_n,\sigma^2 B_n\right) \tag{15.14}$$

with

$$\begin{aligned} B_n =& \left(B^{-1} + \Psi^\top\left(C\left(S,S|\theta\right)/\sigma^2\right)^{-1}\Psi\right)^{-1} \\ b_n =& B_n\left(B^{-1}b + \Psi^\top\left(C\left(S,S|\theta\right)/\sigma^2\right)^{-1}Z\right) \end{aligned}$$

[The derivation is the same as in Bayesian linear regression and will be given as a Homework]

Note 181. The posterior predictive distribution of $Z\left(\cdot\right)$ given the data Z and θ , results by integrating (15.11) with respect to (15.14) i.e.

$$\begin{aligned} \mathrm{pr}\left(Z_*|Z,\theta\right) &= \int \mathrm{pr}\left(Z_*|Z,\beta,\theta\right)\mathrm{pr}\left(\beta|Z,\theta\right)\mathrm{d}\beta \\ &= \int \mathrm{N}\left(Z_*|\mu_*\left(S_*|\beta,\theta\right), C_*\left(S_*,S_*|\theta\right)\right)\mathrm{N}\left(\beta|b_n,\sigma^2 B_n\right)\mathrm{d}\beta \end{aligned}$$

and it is again a GP

$$Z\left(\cdot\right)|Z,\theta \sim \mathrm{GP}\left(\mu_2\left(\cdot|\theta\right), c_2\left(\cdot,\cdot|\theta\right)\right) \tag{15.15}$$

with

$$\begin{aligned} \mu_2\left(s|\theta\right) =& \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1} + \Psi^\top C^{-1}\Psi\right)^{-1}B^{-1}b \\ &+ \left[\left(C\left(s\right)\right)^\top + \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1}/\sigma^2 + \Psi^\top C^{-1}\Psi\right)^{-1}\Psi\right]C^{-1}Z \end{aligned} \tag{15.16}$$

(15.17)

$$\begin{aligned} c_2\left(s,s'|\theta\right) =& c\left(s,s'|\theta\right) - \left(C\left(s\right)\right)^\top C^{-1}C\left(s'\right) \\ &+ \left(\Psi C^{-1}\left(C\left(s\right)\right)^\top - \psi\left(s\right)\right)^\top\left(B^{-1}/\sigma^2 + \Psi^\top C^{-1}\Psi\right)^{-1}\left(\Psi C^{-1}\left(C\left(s'\right)\right)^\top - \psi\left(s'\right)\right) \end{aligned}$$

with column vector $C(s) = (c(s, s_1), ..., c(s, s_n))^\top$, and matrix $C = C(S, S|\theta)$. *[The derivation will be given as a Homework]*

Note 182. If we consider non-informative priors in (15.13) such as $\text{pr}(\beta|\sigma^2) \propto 1$, for instance, by allowing $B \to 0$, and $b < \infty$ then (15.17) produces the Universal Kriging predictor (check with (14.14)).

Note 183. Assume that $\theta = (\sigma^2, \phi, \tau)^\top$ is an unknown fixed hyper-parameter without a prior distribution being specified. Training can be performed by maximizing the marginal likelihood of Z given θ

$$\text{pr}(Z|\theta) = \int \text{pr}(Z|\beta, \theta)\, \text{pr}(\beta|\theta)\, \text{d}\beta \tag{15.18}$$

$$= \int \text{pr}(Z|\Psi\beta, C(S, S|\theta))\, \text{N}(\beta|b, B)\, \text{d}\beta \tag{15.19}$$

$$= \text{N}\left(Z|\Psi b, C(S, S|\theta) + \sigma^2 \Psi B \Psi^\top\right) \tag{15.20}$$

[from Fact 162] by computing

$$\hat{\theta} = \arg\min_{\theta} \left(-2 \log\left(\text{N}\left(Z|\Psi b, C(S, S|\theta) + \sigma^2 \Psi B \Psi^\top\right)\right)\right)$$

Note 184. The estimated "Kriging predictor" results by plugging $\hat{\theta}$ in (15.15)

$$Z(\cdot)\,|Z, \hat{\theta} \sim \text{GP}\left(\mu_2\left(\cdot|\hat{\theta}\right), c_2\left(\cdot, \cdot|\hat{\theta}\right)\right) \tag{15.21}$$

Computing the predictive distribution $Z(\cdot)\,|Z, \phi, \tau$.

FYI: we specify a conjugate prior on σ^2 and then we follow the same routine as above... and we can get a Students-T process...