Spatio-temporal statistics (MATH4341)

Michaelmas term

'Spatial statistics'

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

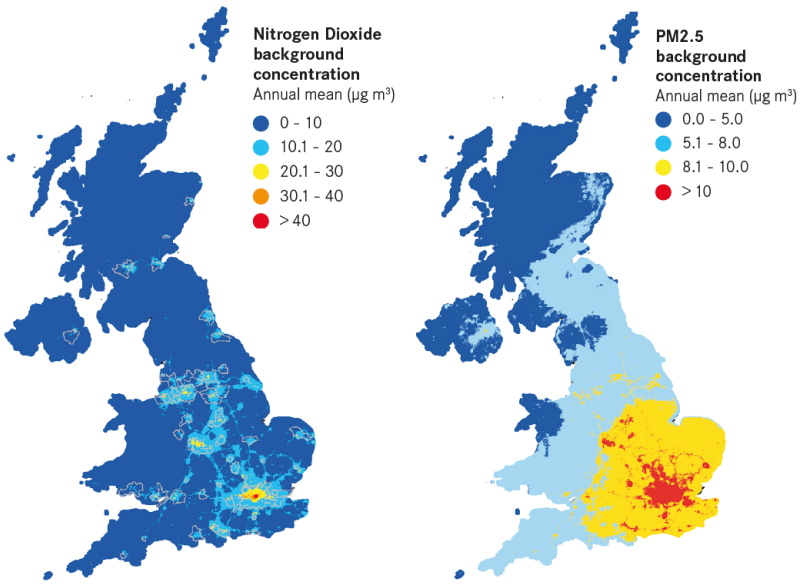
Department of Mathematical Sciences (Office MCS3088)
Durham University
Stockton Road Durham DH1 3LE UK

2024/11/28 at 13:09:45

Concepts

An introduction to spatial statistics:

- Reginalised statistical concepts
- Point referenced data analysis
- Aerial unit data analysis
- Point pattern data analysis
- Implementation in R



Lecture notes

1. Lecture notes part 1: Types of spatial data and modelling
2. Lecture notes part 2: Point referenced data modeling / Geostatistics
3. Lecture notes part 3: Aerial unit data / spatial data on lattices
4. Lecture notes part 4: Spatial point pattern modeling

Reading list

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

Main textbooks (methods, implementation, and theory):

- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
- Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
- Gaetan, C., & Guyon, X. (2010). Spatial statistics and modeling (Vol. 90). New York: Springer.
- Moller, J., & Waagepetersen, R. P. (2003). Statistical inference and simulation for spatial point processes. CRC press.

Main textbooks (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.
- Moraga, P. (2023). Spatial statistics for data science: theory and practice with R. CRC Press.

Supplementary textbooks (methods, implementation, and theory):

- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. CRC press.
- Ripley, B. D. (2005). Spatial statistics. John Wiley & Sons.
- Schabenberger, O., & Gotway, C. A. (2005). Statistical methods for spatial data analysis. CRC press
- Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.
- Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC press.
- van Lieshout, M. N. M. (2019). Theory of spatial statistics: a concise introduction. CRC Press.

Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

Lecture notes part 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

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 for the development and analysis of (not necessarily statistical) models that indicate whether 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 the 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. Galton's problem (1888) arises in spatial statistics and cross-cultural research when observations are not statistically independent due to external dependencies like borrowing or common descent. For example, if two neighboring cultures share similar traits, it might

be due to cultural borrowing rather than independent development. This autocorrelation can lead to misleading conclusions if not properly accounted.

Note 6. 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 7. 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".

Example 8. [1]Consider a random sample $\{Z_i \in \mathbb{R}; i = 1, ..., n\}$ jointly following a Normal distribution with unknown $\mathrm{E}(Z_i) = \mu$ and known $\mathrm{Var}(Z_i) = \sigma^2$.

Iid assumption: Assumption that the observables $\{Z_i\}$ are independent. This is equivalent to iid model $Z_i = \mu + \epsilon_i$ with $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

AR assumption: Assumption that the observables $\{Z_i\}$ are positively autocorrelated as $\mathrm{Cov}(Z_i, Z_j) = \sigma^2 \rho^{|i-j|}$ with known $\rho \in (0, 1)$. This is equivalent to AR model $Z_i = \rho Z_{i-1} + \epsilon_i$ with $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2(1 - \rho^2))$.

Consider the case that the observables $\{Z_i\}$ are positively autocorrelated. Assume You fail to realize the presence of positive correlation in the data and You use the i.i.d model instead.

(1) Effect on parametric estimation of μ

The BLUE of μ is

$$\hat{Z} = \bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i \tag{1.1}$$

[1]No need to worry how the equations are produced.

If the observables were iid then $\text{Var}\left(\hat{Z}|\text{iid obs}\right) = \frac{\sigma^2}{n}$ Since the observables are positively autocorrelated

$$\begin{aligned}\text{Var}\left(\hat{Z}|\text{AR obs}\right) =& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}\left(Z_i, Z_j\right) \\ =& \frac{\sigma^2}{n}\overbrace{\left(1 + 2\frac{\rho}{1-\rho}\left(1-\frac{1}{n}\right) - \frac{2}{n}\left(\frac{\rho}{1-\rho}\right)^2\frac{1-\rho^{n-1}}{2}\right)}^{\tau_{n,\rho}=}\end{aligned} \tag{1.2}$$

If $n = 10$, and $\rho = 0.26$, the $1 - \alpha = 95\%$ confidence interval for μ is

$$\left\{\bar{Z} \pm q_{\frac{\alpha}{2}}\sqrt{\text{Var}\left(\hat{Z}|\text{assump.}\right)}\right\} \implies \begin{cases} \bar{Z} \pm \overbrace{1.96}^{q_{\frac{0.95}{2}}=}\frac{\sigma}{\sqrt{10}} & \text{in i.i.d assum.} \\ \bar{Z} \pm \underbrace{2.485}_{=q_{\frac{0.95}{2}}\sqrt{\tau_{10,0.26}}}\frac{\sigma}{\sqrt{10}} & \text{in AR assum.} \end{cases}$$

as $\sqrt{\tau_{10,0.26}} = 1.608$. Hence, if we fail to realize the presence of positive autocorrelation in the observables and wrongly the i.i.d model, the resulted confidence interval would be too narrow since the actual coverage probability is $\Phi\left(2.485\right) - \Phi\left(-2.485\right) = 87.5\%$ instead of 95%.

By re-writing the variance (1.2) as $\text{Var}\left(\hat{Z}\right) = \frac{\sigma^2}{n'}$ with

$$n' = n \Bigg/ \left(1 + 2\frac{\rho}{1-\rho}\left(1-\frac{1}{n}\right) - \frac{2}{n}\left(\frac{\rho}{1-\rho}\right)^2\frac{1-\rho^{n-1}}{2}\right),$$

we intuitively understand that the effect of the positive spatial correlation is that the equivalent independent observations in the above dataset with size n is n' (i.e. $n' \approx n\frac{1-\rho}{1+\rho}$, for large n). If $n = 10$, and $\rho = 0.26$, then $n' = 6.2$.

(2) Effect on predictive estimation of Z_{n+1}

Under the i.i.d. model, the predictor minimizing the MSPE for the next outcome Z_{n+1} is $\hat{Z}_{n+1}^{\text{iid}} = \bar{Z}$ and has

$$\text{MSPE}\left(\hat{Z}_{n+1}^{\text{iid}}|\text{dep obs}\right) = \sigma^2\left(1 + 2\frac{\rho}{1-\rho}\left(\rho^n - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\frac{1-\rho^{n-1}}{2}\right)$$

Under the AR model the predictor minimizing the mse of the next outcome Z_{n+1} is

$$\hat{Z}_{n+1}^{\text{AR}} = \rho Z_n + (1-\rho)\frac{Z_1 + (1-\rho)\sum_{i=2}^{n-1} Z_i + Z_n}{n - (n-2)\rho}$$

with MSPE

$$\text{MSPE}\left(\hat{Z}_{n+1}^{\text{AR}}|\text{dep obs}\right)=\sigma^2\left(1-\rho^2\frac{(1+\rho)(1-\rho)^2}{n-(n-2)\rho}\right)$$

Note that the relative efficiency of this mistake is not too bad, e.g. for $n=10$ and $\rho=0.26$

$$\text{RE}=\frac{\text{MSPE}\left(\hat{Z}_{10+1}^{\text{AR}}|\text{dep obs}\right)}{\text{MSPE}\left(\hat{Z}_{10+1}^{\text{iid}}|\text{ind obs}\right)}=\frac{1.01952}{1.1}\approx 1.$$

However, if I consider large datasize $n\to\infty$ the asymptotic relative efficiency is

$$\text{ARE}=\lim_{n\to\infty}\frac{\text{MSPE}\left(\hat{Z}_{n+1}^{\text{AR}}|\text{dep obs}\right)}{\text{MSPE}\left(\hat{Z}_{n+1}^{\text{iid}}|\text{ind obs}\right)}=...=1-\rho^2\approx\begin{cases}93\% & ,\rho=0.26\\ 75\% & ,\rho=0.5\end{cases}$$

Spatial data and spatial process.

Note 9. In spatial statistics, the basic components are data $\{Z_1,...,Z_n\}$ observed at spatial locations $\{s_1,...,s_n\}$ correspondingly. 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.

Note 10. Even more exotically, the spatial domain S does not necessarily need to be a Euclidean space (which our course will focus) such as $S\subset\mathbb{R}^d$, $d=1,2,3,...$ but any topological space, eg sphere (recall Earth is round...).

Note 11. The locations $s_i\in S$ can be considered either (i.) fixed and hence used for training or (ii.) uncertain/random and hence a quantity for inference. Yet, $\{s_i\}$ can be arranged irregularly in the space or regularly in a grid. Data $Z_i=Z(s_i)$ are random vectors.

Note 12. 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 field (or multivariate random process)

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

which can be modeled as a random field / random function / stochastic process (to be defined later.).

Note 13. 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 random field (or a stochastic processes)

$$\{Z(s)\,;s\in S\}\,, \tag{1.3}$$

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

Note 14. 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/problems may involve characteristics from any combination of them.

Note 15. Here, we will study three of spatial statistical areas corresponding to point referenced data, aerial unit data, and point patterns.

2.1. **Point referenced data / Geostatistics.**

Note 16. 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 realizations of a random function (random field/stochastic process); 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 17. 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 18. 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 $\{Z_i\}_{i=1}^n$ at n (finite number) fixed locations $\{s_i\}_{i=1}^n \subset S$ such as $Z_i = Z(s_i)$. The locations $\{s_i\}$ are fixed and can be arranged irregularly in the space or regularly as a grid.

Note 19. 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 $Z(\cdot)$ across the whole space S.

Example 20. (Meuse river data set) Due to pollution of the Meuse river over many years, considerable amounts of heavy metals have accumulated in the overbank sediments of the embanked floodplains of their lower river reaches. The spatial variability of metal pollution of floodplain soils, which is controlled primarily by deposition of contaminated overbank sediments during flood events is under consideration. The process governing heavy metal distribution seems that polluted sediment is carried by the river, and mostly deposited close to the river bank, and areas with low elevation.

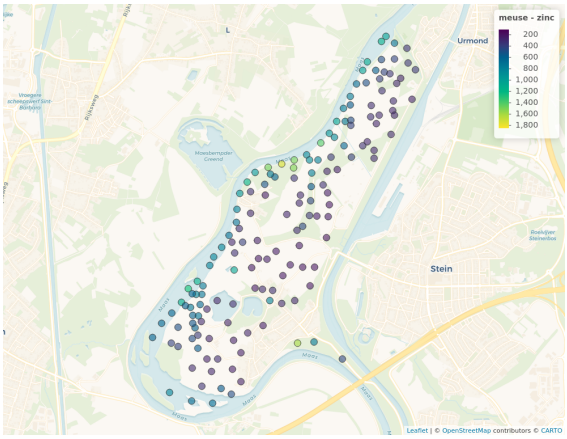
The Meuse river R dataset `meuse{sp}` contains locations and topsoil heavy metal concentrations (such as zinc, lead, copper, and cadmium), along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Figure 2.1 shows 155 measurements of topsoil concentrations of the above heavy metals cadmium, copper, lead, and zinc collected in a flood plain of the river Meuse. Here, the locations $\{s_i\}$ are fixed and arranged irregularly as chosen by experimental design. The quantities of interest (QoI) $\left\{\left(Z_i^{\text{zinc}}, Z_i^{\text{lead}}, Z_i^{\text{cooper}}, Z_i^{\text{carmium}}\right)\right\}$ are the concentrations of zinc, lead, copper, and cadmium, at these locations. Interest lie in prediction of QoI at unobserved locations (i.e. interpolation), quantification of the joint distribution of the QoI (evaluate the distribution of a random function $Z(s) = \left(Z^{\text{zinc}}(s), Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s)\right)$, for all $s \in S$), and how each of QoI depends each other along the flood plain of the river Meuse is of interest (Figure 2.1f). If You ignore spatial dependency and implement standard multivariate statistical techniques (e.g. Fig 2.1e) to analyze the dependency of the QoI's you may obtain misleading results due to confunding space.

Example 21. (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.

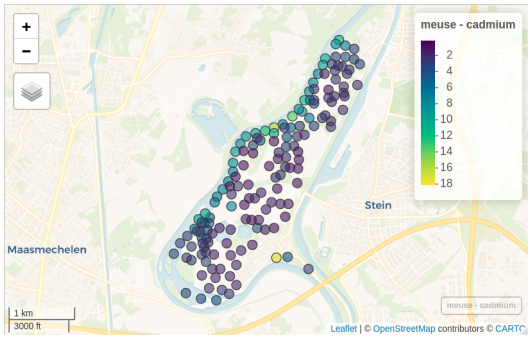
2.2. **Aerial unit data / spatial data on lattices.**

Note 22. 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.

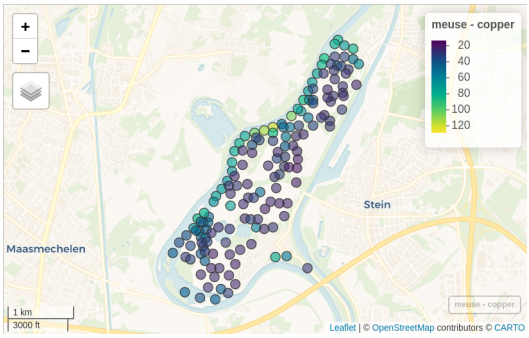
Note 23. Mathematically speaking, the index set S of the data $\{Z(s)\}$ is a fixed (not random) and finite collection of points (locations) $s \in S$. The locations $s \in S$ can be irregular or arranged in a regular grid. Often, there is a natural adjacency relation or neighborhood



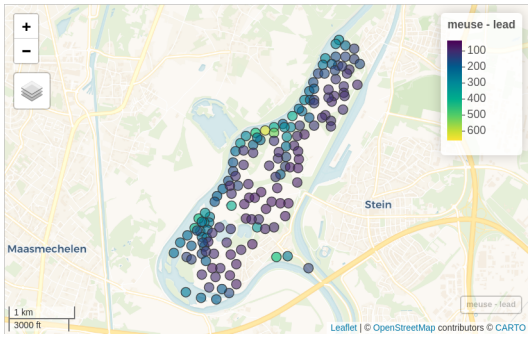
(A) zinc concentration (obs)



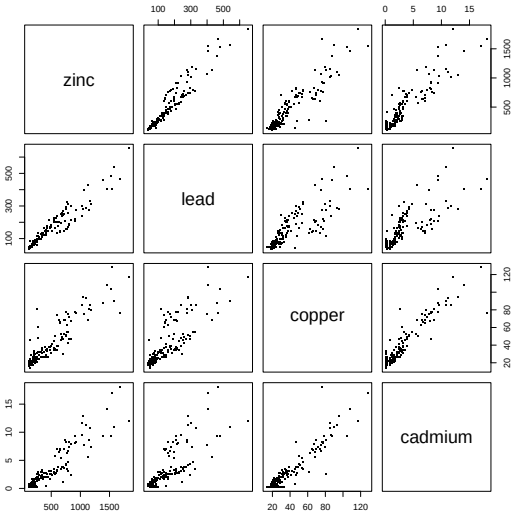
(B) cadmium concentration (obs)



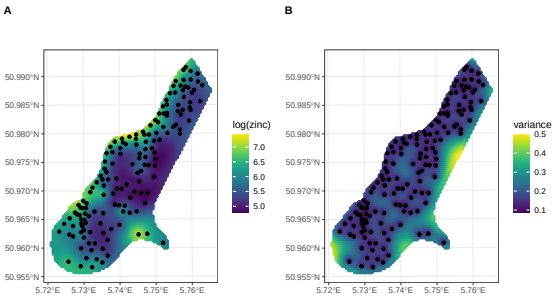
(C) copper concentration (obs)



(D) lead concentration (obs)



(E) scatter plot (obs)



(F) zinc (log scale) pred

FIGURE 2.1. Map of the meuse dataset

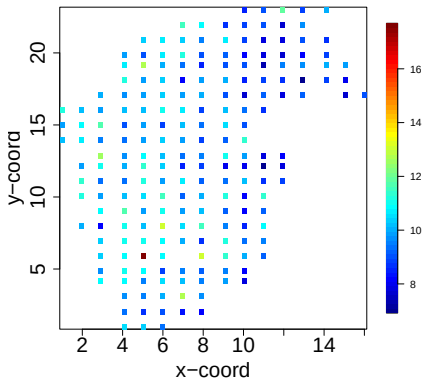


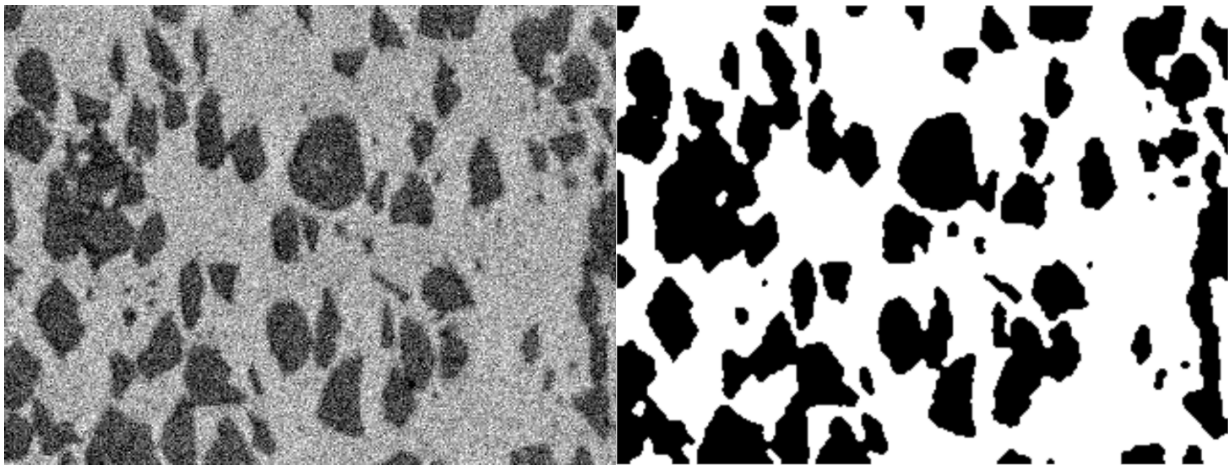
FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

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 24. 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. In image processing, S may be a grid of pixels (locations are fixed and regular). In statistical physics, S may be a collection of atoms and genuinely finite (locations are fixed and regular).

Example 25. (Image restoration data) Figure 2.3a 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 2.3b 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.

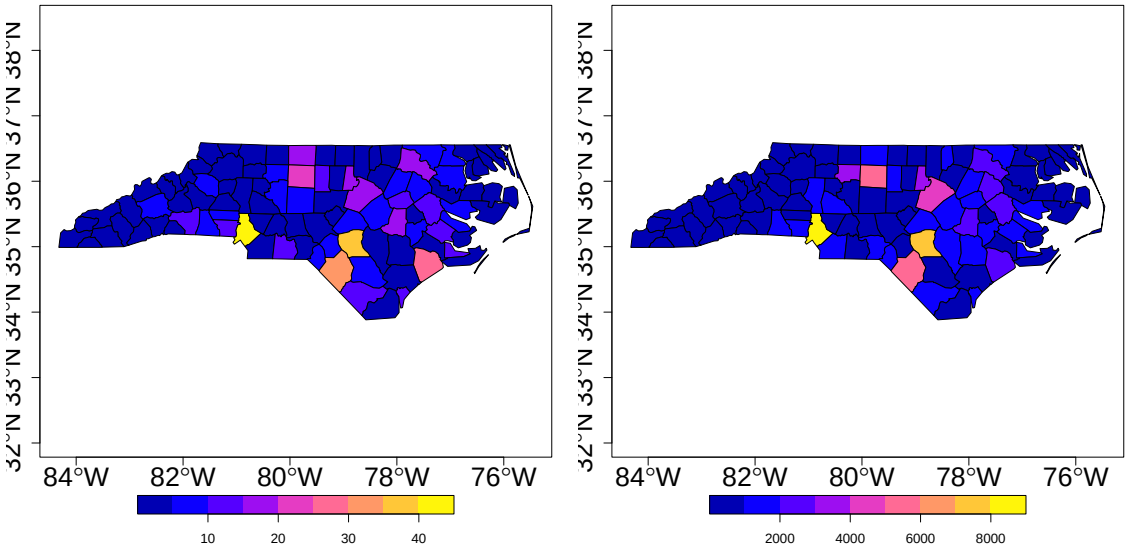
Example 26. (North Carolina SIDS data set) Figure 2.4a shows the total number of deaths from Sudden Infant Death Syndrome (SIDS) in 1974 for each of the 100 counties in North Carolina. Figure 2.4b shows the corresponding live births in each county and same period. This is the R data set `nc{spdep}`. The centroids of the counties do not lie on a regular grid. 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



(A) Noised image (B) Reconstructed image

FIGURE 2.3. Ferrite-Pearlite steel image (Image restoration)

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.

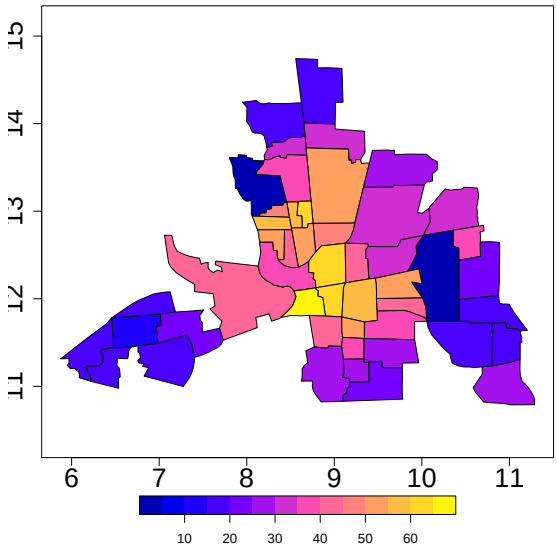


(A) Numbers of SIDS (B) Live birth counts

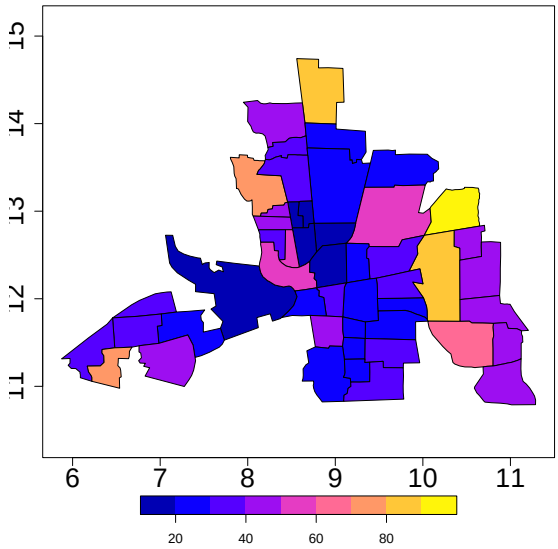
FIGURE 2.4. North Carolina SIDS data (Aerial unit data)

Example 27. (Columbus Columbus OH data set) Figures 2.5a, 2.5b, and 2.5c show the Property crime (number per thousand households) in 49 districts in Columbus in 1980, as

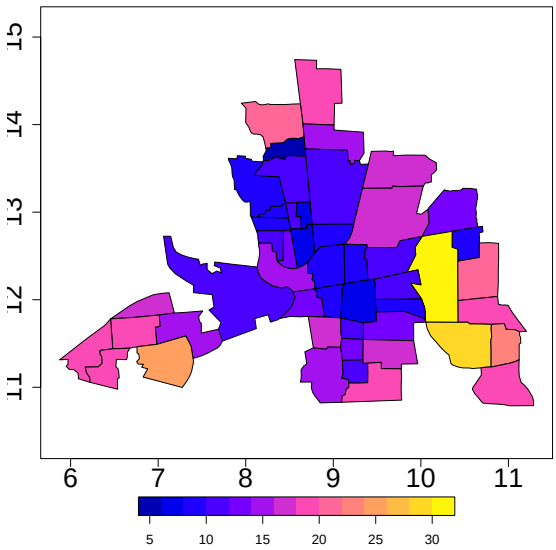
well as the average value of the house (in 1,000 USD), and the average household income (in 1,000 USD). This is the R dataset `columbus{spdep}`. The recorded counts are not tied to a precise household but tallied up county-wise (aerial unit) for privacy reasons. The locations $\{s_i\}$ are these areal unites; we observe they are not in a grid, and they do not have the same size or shape. Interest may lie to find whether high rates of crime are clustered in a particular areas, and if yes, perhaps what is the association of it with the value of the houses in the area.



(A) Property crime, i.e. residential burglaries and vehicle thefts per thousand households in the neighborhood



(B) Average housing value (in 1,000 USD)



(C) Average household income (in 1,000 USD)

FIGURE 2.5. Columbus Columbus OH spatial analysis dataset

2.3. **Spatial point patterns.**

Note 28. Sometimes the locations at which events occur are random. Typical examples include locations of trees in forests, outbreaks of forest fires, or epicentres of earthquakes. Such random patterns of locations are said to form a point pattern.

Note 29. Rigorously, the spatial domain S is a random set of points; specifically a point random field, in $\mathbb{R}^d$ at which some events happened.

Note 30. 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 $Z(s) = 1$ 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 random field

Note 31. 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 random field.

Note 32. Questions in the spatial point pattern problems are mainly whether the pattern of locations is exhibiting complete spatial randomness, clustering (aggregation), or regularity (repulsiveness). In the marked spatial point random field 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 field may appear to be wrongly clustered when just eyeballed by an individual.

Example 33. (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 random field 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 34. The domain scientist is interested in knowing whether the spatial locations are completely spatially random, clustered, or regularly distributed.

(1) (Japanese pine trees) The Japanese pine trees dataset in R `japanesepines{spatstat}` represents locations of 65 saplings of Japanese black pine (Pinus thunbergii) in a 5.7×5.7 square meter sampling region in a natural forest.
(2) (California Redwoods) The data represent the locations of 62 seedlings and saplings of California Giant Redwood (Sequoiadendron giganteum) recorded in a square sampling region.

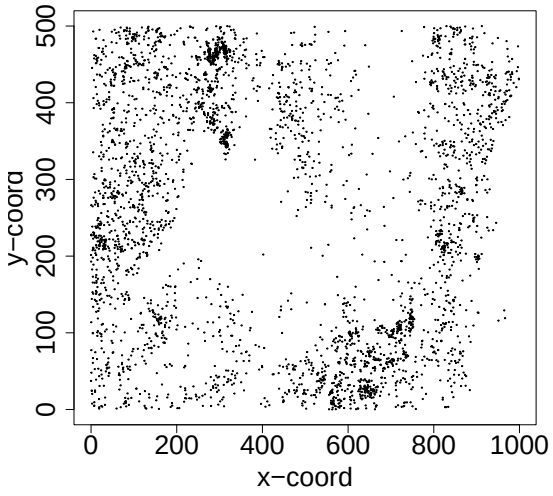
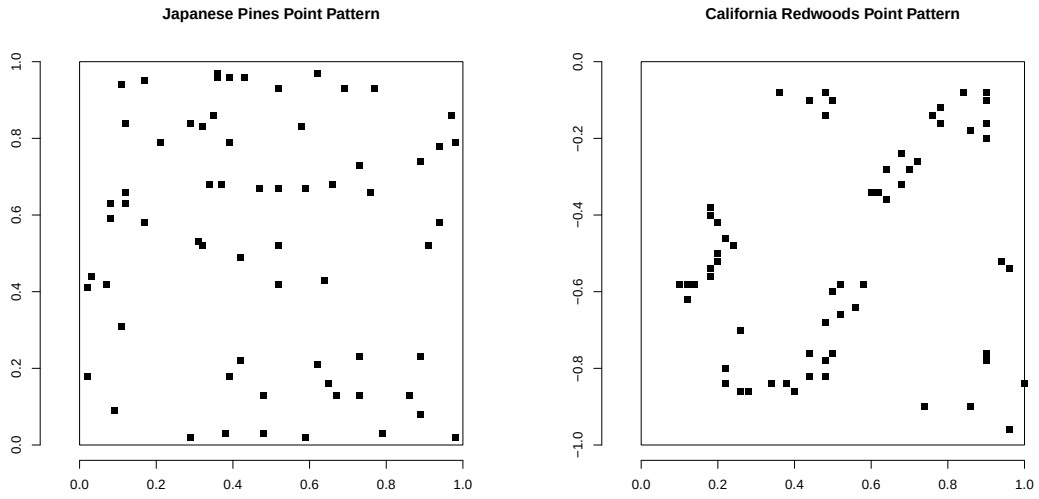


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

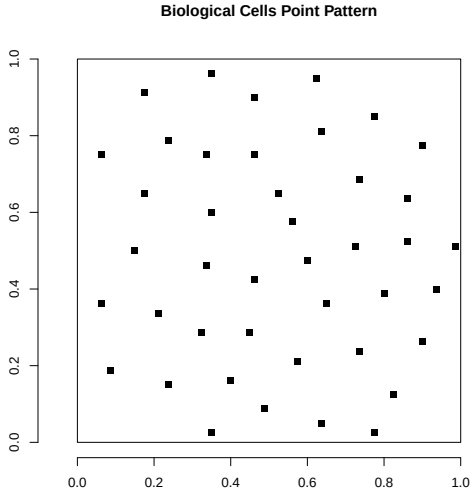
(3) (Biological Cells) The data record the locations of the centres of 42 biological cells observed under optical microscopy in a histological section. The microscope field-of-view has been rescaled to the unit square.

Figure 2.7 presents the aforesaid spatial point patters. It is difficult to eyeball, perhaps Figure 2.7a of the Japanese pine trees seems neither clustered nor regularly distributed but rather completely randomly distributed; Figure 2.7b of the California redwoods shows a clustered pattern; and Figure 2.7c shows a a regular pattern.

Example 35. (Longleaf Pines Point Pattern) Figure 2.8 shows locations (as Cartesian coordinates) and relative diameters at breast height in dbh (as the size of the dot) of all longleaf 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 random field is a marked point random field.



(A) Japanese Pines (B) California Redwoods



(C) Biological Cells

FIGURE 2.7. Spatial Point Patterns

3. Uncertainty quantification and modeling

To be defined rigorously later

Note 36. In spatial problems, uncertainty is expressed probabilistically through a spatial random field (or a stochastic process), which can be written most generally as

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

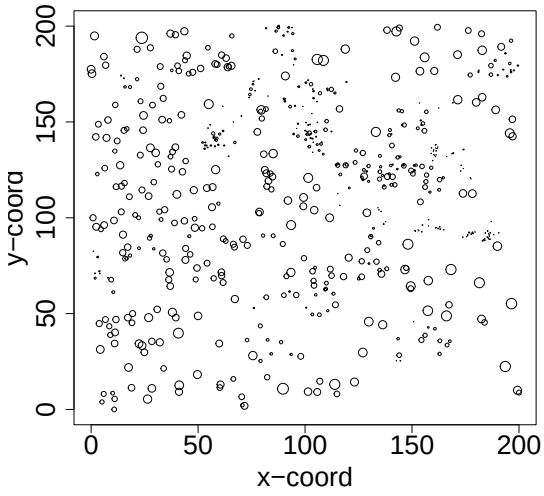


FIGURE 2.8. Longleaf Pines Point Pattern (Spatial point data)

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 37. 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. 2.3a.

Note 38. This spatial random field can be a: geostatistical random field, lattice random field, or point random field depending on the principal spatial statistical area (Section 2) the application is associated with.

Note 39. 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 40. 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 41. The measurement uncertainty is quantified via the data model. E.g. the "noisy image" in Fig. 2.3a.

Note 42. The data model is specified to be the conditional distribution of the data Z given the spatial random field Y and the S, namely

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

Note 43. If the data are assumed to be conditionally independent, such as $Z\left(s\right) \perp Z\left(s'\right)|Y,S$ then

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

Note 44. The spatial statistical dependence of in Z, articulated by the First Law of Geography, follows by

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

The hierarchical statistical model.

Note 45. 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

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

Our interest it to learn $\text{pr}\left(Z^*|Z\right)$ or $\text{pr}\left(Y^*|Z\right)$ at any unseen locations s^*.

The Empirical (Bayes) hierarchical model.

Note 46. 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

$$\text{pr}\left(Z,Y,S|\theta\right) = \text{pr}\left(Z|Y,S,\theta\right)\text{pr}\left(Y|S,\theta\right)\text{pr}\left(S|\theta\right)$$

Our additional additional interest may be to learn the unknown θ via MLE

$$\hat{\theta} = \arg\min_{\theta}\left(\text{pr}\left(Z|\theta\right)\right)$$

The Bayesian hierarchical model.

Note 47. In Bayesian statistics, the hierarchical model in (3.5) is completed by the $\theta \sim \text{pr}(\cdot)$ 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

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

Our additional additional interest may be to learn the posterior of θ $\text{pr}(\theta|Z)$.

That's a snapsot for what follows.

Example 48. (A naive example: NOAA weather data) As dataset we consider a fraction from daily data originated from the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center and can be obtained from the IRI/LDEO Climate Data Library at Columbia University[2]. The dataset is available from the R package in GitHub "andrewzm/STRbook". In this example we focus on maximum temperature (Tmax) in degrees Fahrenheit (∘F) at 138 weather stations in the central USA (between 32∘N-46∘N and 80∘W-100∘W), recorded in 1st of May 1993. The data are not complete, in the sense that there are missing measurements at various stations and at various time points, and the stations themselves are obviously not located everywhere in the central USA. (Figure Figure 3.1)

This data set contains $n = 138$ observations $\{(Z_i, s_i)\}_{i=1}^n$ where the i-th observation contains the maximum temperature (Tmax) in degrees Fahrenheit (∘F) Z_i at location specified by coordinates $s_i = (s_{1,i}, s_{2,i})^\top$ that is the latitude degrees north of the Equator $s_{2,i}$, and longitude degrees west of Greenwich $s_{1,i}$. See Figure 3.1.

This is definitely a geostatistics problem. Here, we will present a naive way to model it by reflecting what we discussed earlier.

Data model: One may consider that the observations $\{Z_i\}$ at each location are the result of observing the real (hence unknown) maximum temperature Y_i but contaminated by "additive random noise" with unknown scale $\sigma > 0$ due to instrumental error, i.e.

$$Z_i = Y_i + \epsilon_i,\ \ \epsilon_i \sim \text{N}\left(0,\sigma^2\right),\quad i = 1,...,n \implies Z_i|Y_i,\sigma^2 \overset{\text{ind}}{\sim} \text{N}\left(Y_i,\sigma^2\right),\quad i = 1,...,n$$

Spatial process model: One may consider that the real maximum temperature $Y(s)$ (over the spatial domain $s \in S$) is a function where at each finite set of locations $\{s_i\}_{i=1}^n$ follows a Normal distribution with a mean μ with $[\mu]_i = \mu(s_i)$ parameterized

[2] http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.DAILY/.FSOD/

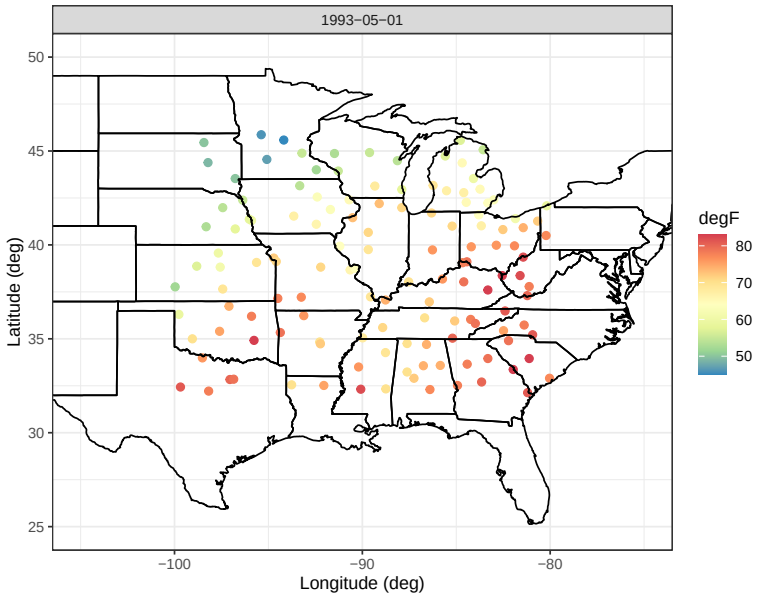


FIGURE 3.1. NOAA weather data (1 May 1993)

as

$$\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{12} s_1 s_2, \text{ at a location s}=(s_1, s_2)^\top$$

with unknown parameter β, and covariance matrix $[C]_{i,j} = c(s_i, s_j)$ parameterized with covariance function

$$c(s, s') = \sigma^2 \exp(-\phi \|s - s'\|)$$

to impose that nearer locations cause stronger dependences in the model. Here β, ϕ, and σ^2 are unknown parameters.

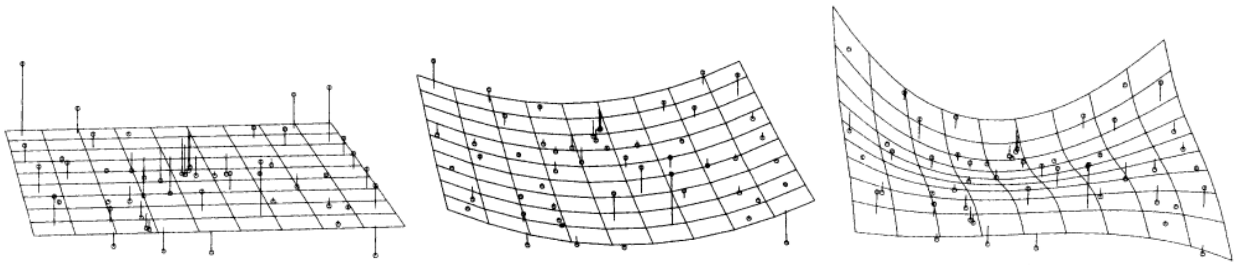
Hierarchical model: To sum up, we have build the hierarchical model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}_n(Y, I\sigma^2), \text{ data model} \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}_n(S\beta, C), \quad \text{spatial process model} \end{cases} \tag{3.8}$$

Figure 3.2 shows the hierarchical spatial model (3.8) for different values of $\theta = (\sigma^2, \beta, \phi)$; the surface corresponds to the spatial process $\{Y(s); s \in \mathbb{R}^2\}$ and is presented at three different instances each of them with different values for (β, ϕ), while the dots correspond to the observations $\{(Z(s_i), s_i)\}_{i=1}^n$ and their deviation from the spatial process is controlled by σ^2 . Note that marginally

$$Z|\sigma^2, \beta, \phi \sim \mathrm{N}_n\left(S\beta, \sigma^2(I + C)\right).$$

Bayesian hierarchical model: If we work on the fully Bayesian framework, we can complete the model with priors on $\theta = (\sigma^2, \beta, \phi)$ for instance $\sigma^2 \sim \mathrm{IG}(\kappa_\sigma, \lambda_\sigma)$, $\phi \sim$



(A) (B) (C)

FIGURE 3.2. Examples representing the hierarchical spatial model (3.8) for different values of $\theta = \left(\sigma^2, \beta, \phi\right)$

$\mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right)$, and $\beta \sim \mathrm{N}_4\left(b, Iv\right)$, with some known hyper-parameters $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$. To sum up, we have build the Bayesian model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}_n\left(Y, I\sigma^2\right), \text{ data model} \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}_n\left(S\beta, C\right), \text{ spatial process model} \\ \beta \sim & \mathrm{N}_4\left(b, Iv\right), \text{ hyper-parameter prior model} \\ \sigma^2 \sim & \mathrm{IG}\left(\kappa_\sigma, \lambda_\sigma\right), \text{ hyper-parameter prior model} \\ \phi \sim & \mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right), \text{ hyper-parameter prior model} \end{cases}$$

Note 49. Spatio-temporal data arise when information is both spatially and temporally referenced. Any of the aforesaid spatial statistics cases can be extended and become temporal. Such methods will be discussed in the Epiphany term. This is where we aim at the end of the module.

Example 50. We extend the dataset in the Spatial statistics Example 48 by considering time.

The dataset is available from the R package in GitHub "andrewzm/STRbook". These daily data originated from the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center and can be obtained from the IRI/LDEO Climate Data Library at Columbia University. The data set we consider consists the daily maximum temperature (Tmax) in degrees Fahrenheit (∘F) at We extend the dataset in the Spatial statistics Example 48 by considering time. weather stations in the central USA (between 32∘N–46∘N and 80∘W–100∘W), recorded between May-July in year 1993. These data are considered to be discrete and regular in time (daily) and geostatistical and irregular in space. However, the data are not complete, in that there are missing measurements at various stations and at various time points, and the stations themselves are obviously not located everywhere in the central USA. See Figure 4.1.

Interest lies on not only how the maximum temperature is only distributed over the spatial domain but also how it evolves during the time.

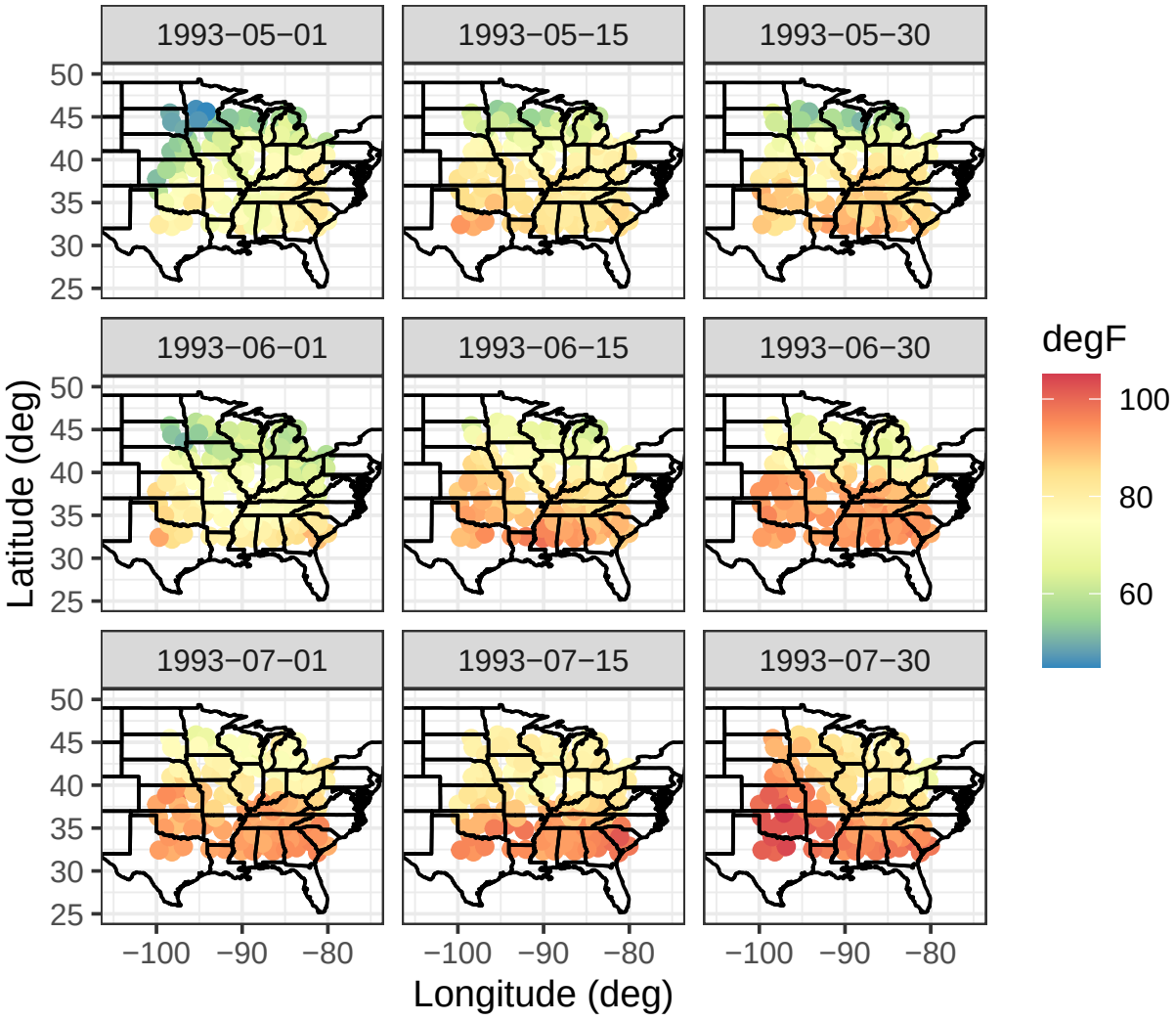


FIGURE 4.1. NOAA daily weather data

Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

Lecture notes part 2: Point referenced data modeling / Geostatistics

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

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

Reading list & references:

[1] Cressie, N. (2015; Part I). Statistics for spatial data. John Wiley & Sons.
[2] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
[3] Chiles, J. P., & Delfiner, P. (2012). Geostatistics: modeling spatial uncertainty (Vol. 713). John Wiley & Sons.
[4] Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.
[5] Gaetan, C., & Guyon, X. (2010; Ch 2 & 5.1). Spatial statistics and modeling (Vol. 90). New York: Springer.

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

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

1. Random fields (or Stochastic processes)

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

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

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

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

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

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

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

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

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

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

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

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

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

1.1. **Mean and covariance functions.**

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

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

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

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

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

1.1.1. *Construction of covariance functions.*

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

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

$$\forall n \in \mathbb{N} - \{0\}\ ,\ \forall (a_1, ..., a_n) \in \mathbb{R}^n \text{ and } \forall (s_1, ..., s_n) \in \mathcal{S}^n : \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j c(s_i, s_j) \geq 0$$

or in other words, the Gram matrix $(c(s_i, s_j))_{i,j=1}^{n}$ is non-negative definite for any $\{s_i\}_{i=1}^{n}$, $n \in \mathbb{N} - \{0\}$.

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

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

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

Proof. From Proposition 10, as

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

□

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

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

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

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

3. Gaussian random field (or Gaussian process)

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

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

Also Example of Proposition

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

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

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

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

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

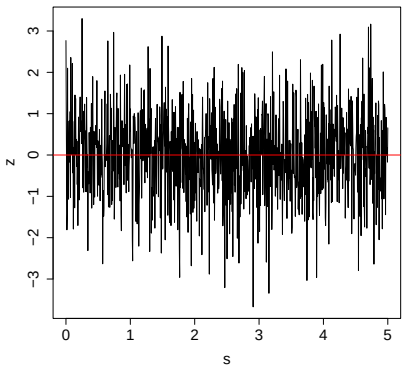
Example 22. Recall your linear regression lessons where you specified the sampling distribution to be $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta, \sigma^2\right)$, $\forall x \in \mathbb{R}^d$. Well that can be considered as a GP $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot),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

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

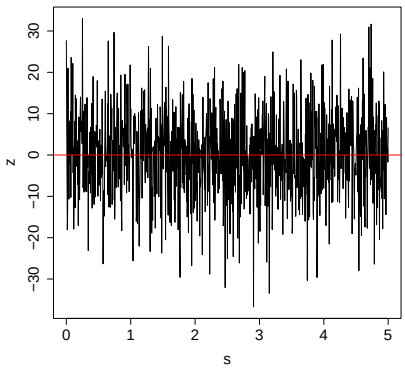
Algorithm 1 R script for simulating from a GP $(Z(s); s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\phi \|s-t\|_2^2\right)$

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

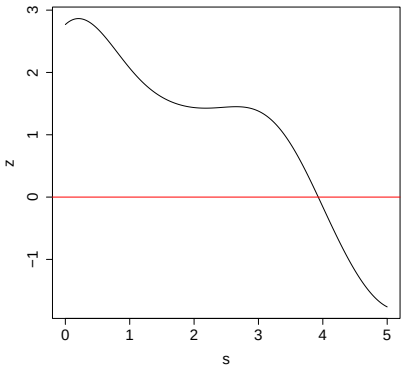
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency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



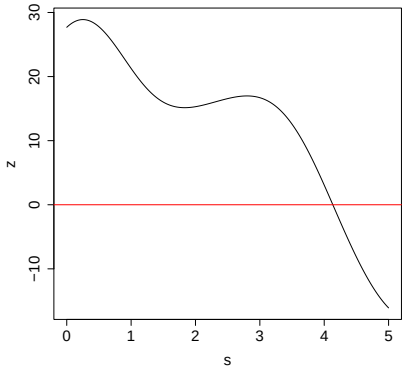
(A) Nugget c.f
($\sigma^2 = 1$)



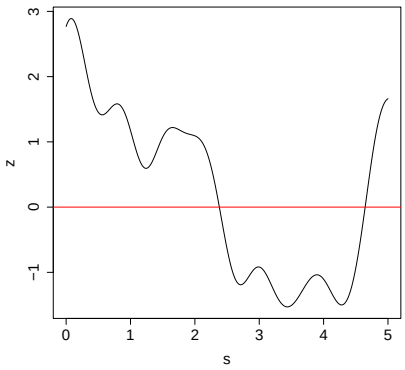
(B) Nugget c.f
($\sigma^2 = 100$)



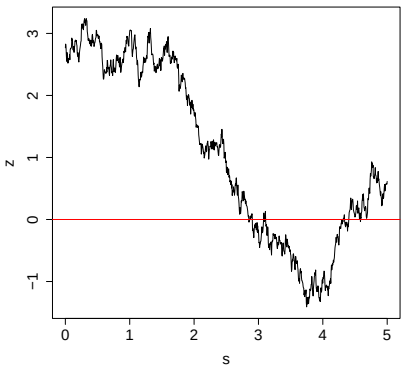
(C) Gauss c.f
($\sigma^2 = 1$,$\phi = 0.5$)



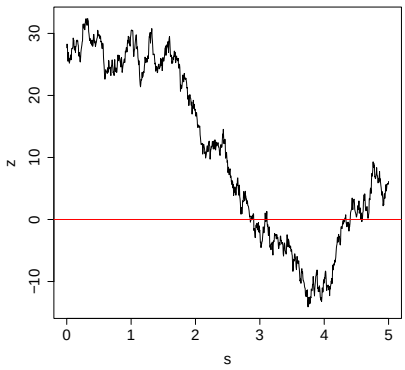
(D) Gauss c.f
($\sigma^2 = 100$,$\phi = 0.5$)



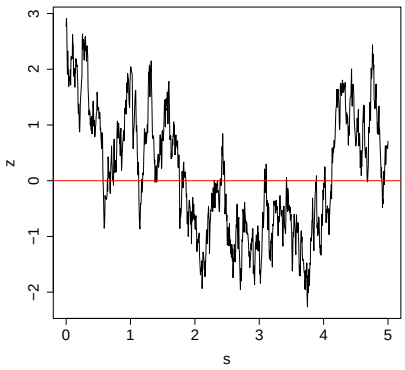
(E) Gauss c.f
($\sigma^2 = 1$,$\phi = 5$)



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

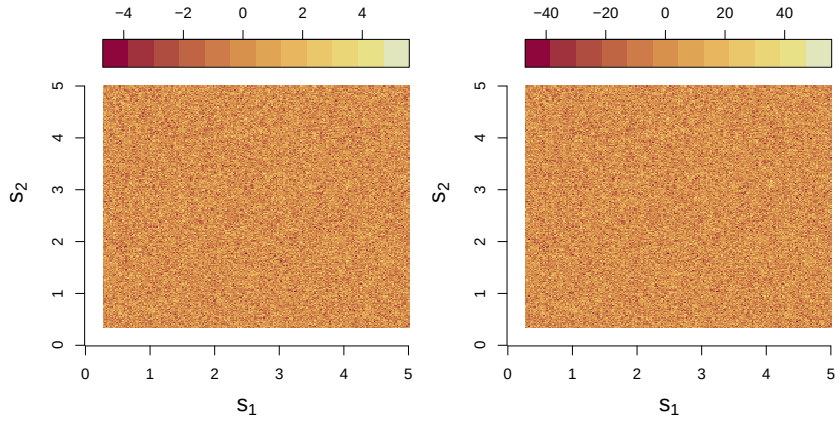


(G) Exp c.f
($\sigma^2 = 100$,$\phi = 0.5$)



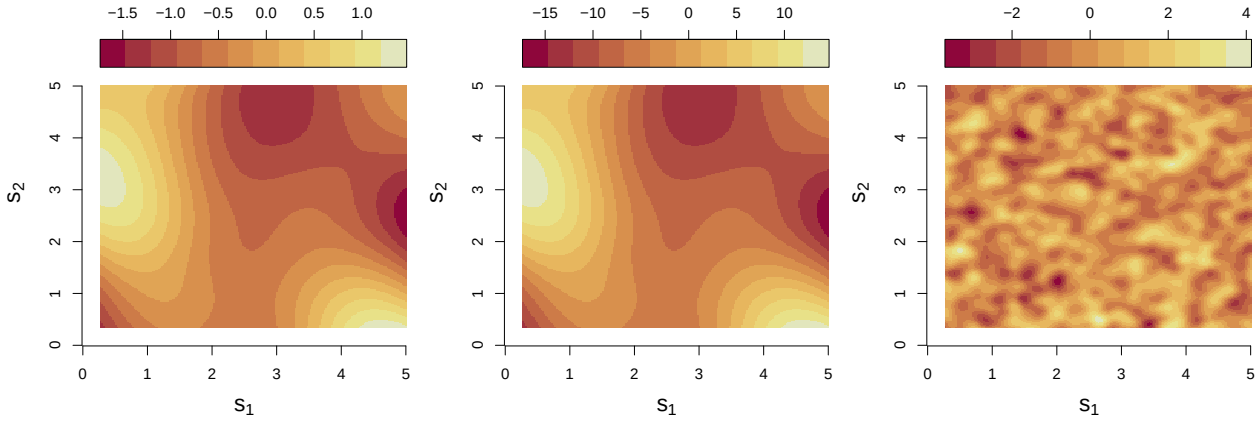
(H) Exp c.f
($\sigma^2 = 1$,$\phi = 5$)

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



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

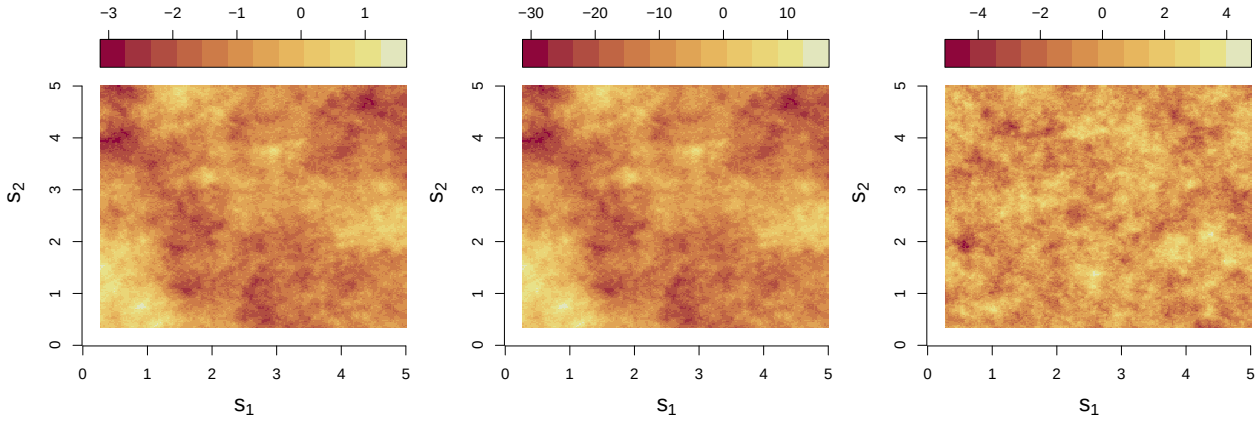
(B) Nugget c.f
($\sigma^2 = 100$)



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

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

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



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

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

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

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

4. Strong stationarity

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

Notation 25. Formally, we define the separation (or lag) set as $\mathcal{H} = \left\{h \in \mathbb{R}^d : s \in \mathcal{S}, s+h \in \mathcal{S}\right\}$ where $\mathcal{S} \subseteq \mathbb{R}^d$ is the spatial domain for $d = 1, 2, 3, ...$. However, we will consider cases where $\mathcal{S} = \mathbb{R}^d$ and $\mathcal{H} = \mathbb{R}^d$ for $d = 1, 2, 3, ...$ in Euclidean spaces.

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

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

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

5. Weak stationarity (or second order stationarity)

Note 28. We introduce another weaker behavior of random field able to represent a larger class of spatial dependencies.

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

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

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

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

6. Covariogram

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

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

$$c(h) = \text{Cov}(Z(s), Z(s+h)), \ \forall s \in \mathcal{S}, \forall h \in \mathcal{H}.$$

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

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

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

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

(1) $c(h) = c(-h)$ for all $h \in \mathcal{H}$

(2) $|c(h)| \leq c(0) = Var(Z(s))$ for all $h \in \mathcal{H}$

(3) $c(\cdot)$ is semi-positive definite; i.e. for all $n \in \mathbb{N} - \{0\}$, $(a_1, ..., a_n) \in \mathbb{R}^n$, and $(s_1, ..., s_n) \subseteq \mathcal{S}^n$

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

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

Theorem 37. *(Bochner's theorem) Let $c : \mathbb{R}^d \to \mathbb{R}$ be a continuous even real-valued function for $d \geq 1$. Then $c(\cdot)$ is positive semidefinete (hence a covariogram of a stationary random field) if and only if it can be represented as*

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

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

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

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

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

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

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

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

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

i.e. it has a Gaussian form.

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

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

7. Intrinsic stationarity (of order zero)

Note 42. The class of (weakly) stationary random fields may not be sufficiently general enough to model a large number of important applications. E.g., in certain applications, it has been noticed that the "underline process" we wish to model presents increments whose variance

$$\mathrm{Var}(Z(s+h) - Z(s)) = \mathrm{Var}(Z(s+h)) + \mathrm{Var}(Z(s)) - 2\mathrm{Cov}(Z(s+h), Z(s))$$

increases indefinitely with $|h|$; this "process" cannot be modeled within the class of (weakly) stationary random fields whose increments have bounded variance $\mathrm{Var}(Z(s+h) - Z(s)) = 2(c(0) - c(h)) < 2c(0)$. Intrinsic stationary is a weaker assumption extending the class of models we can use.

Definition 43. A random field $(Z(s) : s \in \mathcal{S})$ is called intrinsic random field (i.r.f.) (or intrinsic stationary r.f.) if, for all $h \in \mathcal{H}$, ,

(1) $\mathrm{E}\left(Z\left(s+h\right)-Z\left(s\right)\right)^{2}<\infty$
(2) $\mathrm{E}\left(Z\left(s+h\right)-Z\left(s\right)\right)=\mu\left(h\right)$ for some function $\mu:\mathcal{H}\rightarrow\mathbb{R}$ (lag dependent)
(3) $\mathrm{Var}\left(Z\left(s+h\right)-Z\left(s\right)\right)=2\gamma\left(h\right)$ for some function $\gamma:\mathcal{H}\rightarrow\mathbb{R}$ (lag dependent)

Example 44. The random field with covariance function

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

is not stationary r.f. because

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

for $h\in\mathcal{H}$ but it intrinsic r.f. because

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

Example 45. For an i.r.f. $\left(Z\left(s\right):s\in\mathcal{S}\right)$ with $\mu\left(h\right)=0$, it can be shown that

$$\mathrm{Cov}\left(Z\left(t\right)-Z\left(s\right),Z\left(v\right)-Z\left(u\right)\right)=\gamma\left(t-u\right)+\gamma\left(s-v\right)-\gamma\left(s-u\right)-\gamma\left(t-v\right) \tag{7.1}$$

by taking expectations from

$$2\left(a-b\right)\left(c-e\right)=\left(a-e\right)^{2}+\left(a-b\right)^{2}-\left(b-c\right)^{2}-\left(a-c\right)^{2}$$

Note 46. The price to be paid for i.r.f. offering a larger class of models by setting the assumptions on the increments only, is involve an indeterminacy regarding the actual r.f. $Z\left(s\right)$; E.g. i.r.f. . $\left(Z\left(s\right):s\in\mathcal{S}\right)$ and $\left(Z\left(s\right)+U:s\in\mathcal{S}\right)$ where U a single variable leave (2) and (3) in Def 43 unchanged. When this causes problems, usual trick are: (a) "registration" (Example 47), i.e. consider an additional non-used specific site $s_{0}\in\mathcal{S}$ (called origin) at which a value is known $Z\left(s_{0}\right)=z_{0}$ and try to work out (b) impose restrictions int eh increments.

Example 47. To specify the moments of an i.r.f. $Z\left(s\right)$. Consider an origin $s_{0}\in\mathcal{S}$ with known $Z\left(s_{0}\right)=z_{0}$, and specify the "registered" r.f. $\tilde{Z}\left(s\right)=z_{0}+\left(Z\left(s\right)-Z\left(s_{0}\right)\right)$. Then $\mathrm{E}\left(\tilde{Z}(s)\right)=z_{0}+\mu\left(s-s_{0}\right)$ and $\mathrm{Cov}\left(\tilde{Z}\left(s\right),\tilde{Z}\left(t\right)\right)$ computed from (7.1).

Example 48. Only the covariance of allowed linear combinations can be represented w.r.t. $\gamma\left(\cdot\right)$. I.e.

$$\mathrm{Cov}\left(\sum_{i=1}^{n}a_{i}Z\left(s_{i}\right),\sum_{j=1}^{n}b_{j}Z\left(s_{j}\right)\right)=\sum_{i=1}^{n}a_{i}\sum_{j=1}^{n}b_{j}\mathrm{Cov}\left(Z\left(s_{i}\right),Z\left(s_{j}\right)\right)$$

assuming $Z$is i.r.f. hence covariogram may not be defined, we consider origin $s_0 \in \mathcal{S}$ with $Z(s_0)$ and we restrict to the sum-to-zero linear combinations. Hence, by (7.1)

$$\begin{aligned}\mathrm{Cov}\left(\sum_{i=1}^{n} a_i Z(s_i), \sum_{j=1}^{n} b_j Z(s_j)\right) &= \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j \mathrm{Cov}\left(Z(s_i) - Z(s_0), Z(s_j) - Z(s_0)\right) \\ &= \sum_{i=1}^{n} a_i \sum_{j=1}^{n} b_j \left(\gamma(s_i - s_0) + \gamma(s_j - s_0) - \gamma(s_j - s_i)\right) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \gamma(s_j - s_i)\end{aligned}$$

8. Incremental mean function

Definition 49. Incremental mean function (or drift) of the intrinsic random field $(Z(s) : s \in \mathcal{S})$ is defined as $\mu : \mathcal{H} \to \mathbb{R}$ with $\mu(h) = \mathrm{E}(Z(s+h) - Z(s))$.

Example 50. Let $(Z(s) : s \in \mathcal{S})$ be an intrinsic random field, the incremental drift is linear

$$\mu(h) = h^\top \beta \tag{8.1}$$

for some $\beta \in \mathbb{R}^d$. Indeed, it is

$$\begin{aligned}\mu(h+h') &= \mathrm{E}(Z(s+h+h') - Z(s)) = \mathrm{E}(Z(s+h) - Z(s)) + \mathrm{E}(Z(s+h+h') - Z(s+h)) \\ &= \mu(h) + \mu(h'), \ \forall h, h'.\end{aligned}$$

Since, $\mu(\cdot)$ is continuous and $\mu(0) = 0$, than $\mu(h)$ is linear wrt h.

9. Semivariogram

Note 51. A very informative tool about the behavior of the intrinsic random field is the semivariogram function defined below.

Definition 52. The semivariogram of an intrinsic random field $(Z(s) : s \in \mathcal{S})$ is defined as $\gamma : \mathcal{H} \to \mathbb{R}$, with

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

Definition 53. Variogram of an intrinsic random field $(Z(s) : s \in \mathcal{S})$ is called the quantity $2\gamma(h)$.

Note 54. A stationary random field with covariogram $c(\cdot)$ and mean μ is intrinsic stationary as well with semivariogram

$$\gamma(h) = c(0) - c(h), \tag{9.1}$$

and constant incremental mean $\mu(h) = \mu$.

Example 55. For the Gaussian covariance function (Ex. 34) the semivariogram is

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

Proposition 56. *Properties of semivariogram. Let $(Z(s) : s \in \mathcal{S})$ be an intrinsic random field, then*

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

(2) Semivariogram is conditionally negative definite (c.n.d.): if for all $n \in \mathbb{N}$, $(a_1, ..., a_n) \subseteq \mathbb{R}^n$ s.t. $\sum_{i=1}^n a_i = 0$, and for all $(s_1, ..., s_n) \subseteq S^n$, it is

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

10. Behavior of semivariogram of intrinsic random fields

Note 57. The semivariogram $\gamma(h)$ is very informative when plotted against the lag h. Below we discuss some of the characteristics of it, using Figure 10.1

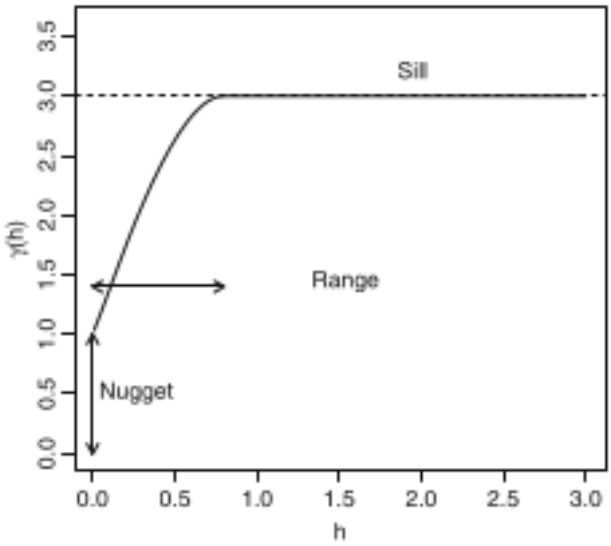


Figure 10.1. Semi Variogram's characteristics

Note 58. A semivariogram tends to be an increasing function of the lag $\|h\|$. Recall that for weakly stationary random fields with c.f. $c(\cdot)$, it is $\gamma(h) = c(0) - c(h)$ where common logic suggests that $c(h)$ is decreases with $\|h\|$.

Note 59. If $\gamma(h)$ is a positive constant for all non-zero lags $h \neq 0$, then $Z(s_1)$ and $Z(s_2)$ are uncorrelated regardless of how close $s_1$and s_2 are. Then $Z(\cdot)$ is called white noise.

Note 60. Conversely, a non zero slope of the variogram indicates some structure.

Nugget Effect.

Note 61. Nugget effect is the semivariagram limiting value

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

when $\sigma_{\varepsilon}^{2} \neq 0$.

Note 62. When used for modeling, nugget effect $\sigma_{\varepsilon}^{2} \neq 0$ may expresses (1) measurement errors (e.g., if we collect repeated measurements at the same location s) or (2) 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. Ideally, a more detailed decomposition $\sigma_{\varepsilon}^{2} = \sigma_{\mathrm{MS}}^{2} + \sigma_{\mathrm{MS}}^{2}$ can be considered where σ_{MS}^{2} refers to the microscale and σ_{MS}^{2} refers to the measurement error. However this may lead to non-identifiability, without any obvious tweak to address it.

Sill.

Definition 63. Sill is the semivariagram limiting value $\lim_{\|h\| \to \infty} \gamma(h)$.

Note 64. For intrinsic processes, the sill may be infinite or finite. For weakly random field, the sill is always finite.

Partial sill .

Definition 65. Partial sill is $\lim_{\|h\| \to \infty} \gamma(h) - \lim_{\|h\| \to 0} \gamma(h)$ which takes into account the nugget.

Range .

Note 66. Range is the distance at which the semivariogram reaches the Sill. It can be infinite or finite.

Other.

Note 67. An abrupt change in slope indicates the passage to a different structuration of the values in space. This is often modeled via decomposition of processes with different semivariograms. E.g., let independent random fields $Y(\cdot)$ and $X(\cdot)$ with different semivariograms γ_Y and γ_X,then random field $Z(\cdot)$ with $Z(s) = Y(s) + X(s)$ has semivariogram $\gamma_Z(h) = \gamma_Y(h) + \gamma_X(h)$ which may present such a behavior.

11. Isotropy

Note 68. Isotropy introduses the assumption of "rotation invariance".

Note 69. Isotropy applies to both intrinsic and (weakly) stationary random fields.

Definition 70. An intrinsic random field $(Z(s) : s \in \mathcal{S})$ is isotropic iff

$$\forall s,t \in \mathcal{S},\ \frac{1}{2}\mathrm{Var}\left(Z(s) - Z(t)\right) = \gamma\left(\|t-s\|\right),\ \text{for some function } \gamma{:}\mathbb{R}_{\geq 0} \to \mathbb{R}. \tag{11.1}$$

Definition 71. Isotropic semivariogram $\gamma : \mathcal{H} \to \mathbb{R}$ is the semivariogram of the isotropic random field (sometimes for simplicity of notation we use $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with $\gamma(\|h\|) = \frac{1}{2}\mathrm{Var}\left(Z(s) - Z(s-h)\right)$.

Definition 72. Isotropic covariance function $C : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called the covariance function satisfying (11.1).

Definition 73. Isotropic covariogram $c : \mathcal{H} \to \mathbb{R}$ of a weakly stationary process is the covariogram associated to an isotropic semivariogram. Sometimes for simplicity of notation we use $c : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with $c(\|h\|)$ from (11.1).

11.1. **Popular isotropic covariance functions.**

Note 74. Isotropic semivariograms can be computed from $\gamma(h) = c(0) - c(h)$ given covariogram $c(\cdot)$ for any h.

11.1.1. *Nugget-effect.*

Note 75. Nugget-effect covariogram takes the form

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

for $\sigma^2 > 0$. It is associate to white noise. It is used to model a discontinuity in the origin of the covariogram / sem-variogram.

11.1.2. *Matern c.f.*

No need to memorize (11.2)

Note 76. Matern covariogram takes the form

$$c(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|h\|}{\phi}\right)^{\nu} K_\nu\left(\frac{\|h\|}{\phi}\right) \tag{11.2}$$

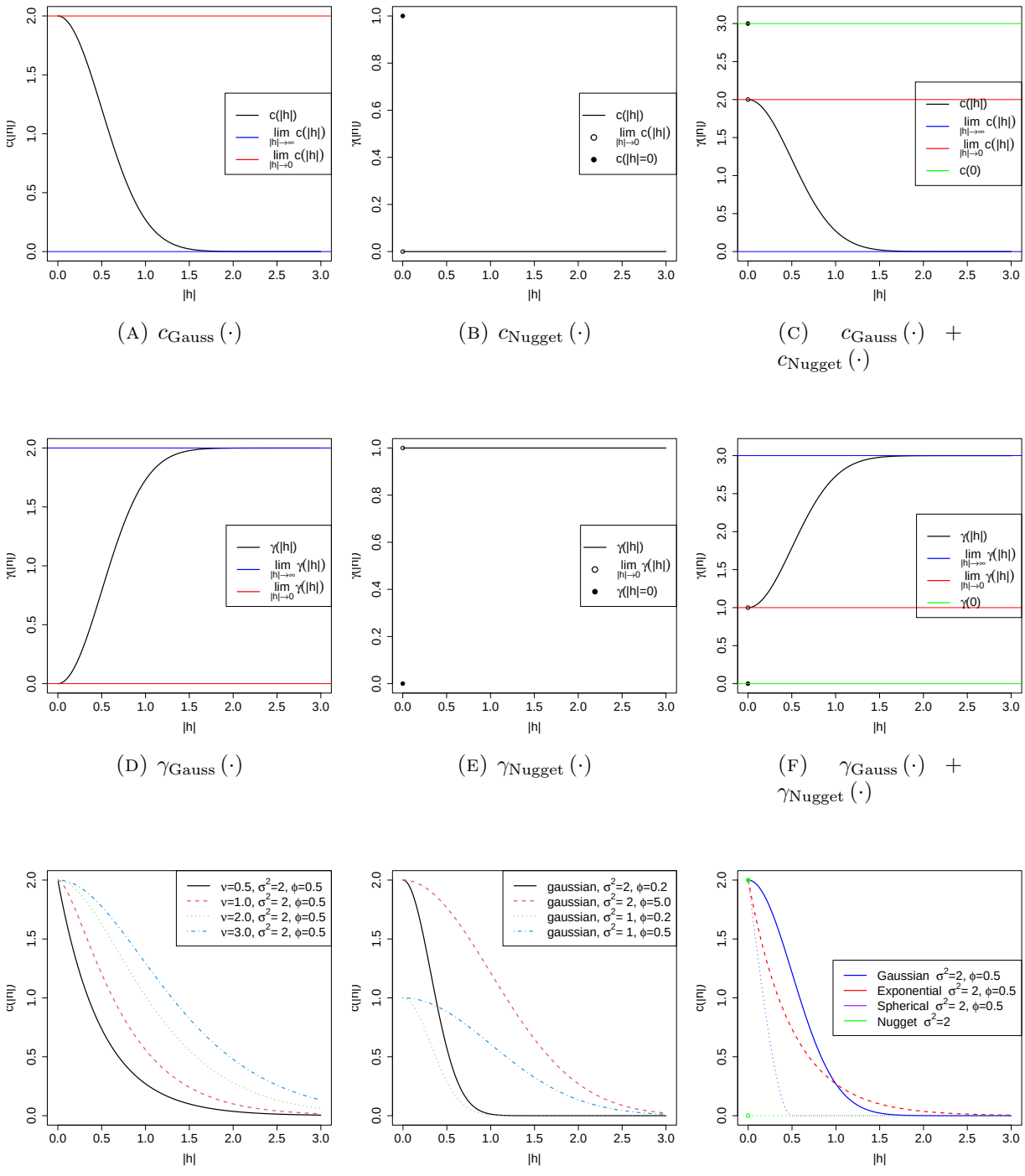
for $\sigma^2 > 0$, $\phi > 0$, and $\nu \geq 0$. Parameter ν controls the variogram's regularity at 0 which in turn controls the quadratic mean (q.m.) regularity of the associated process. For $\nu = 1/2$, we get the exponential c.f.,

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

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.



(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 11.1. Covariograms $c(\cdot)$ and semivariograms $\gamma(\cdot)$

11.1.3. *Spherical c.f.*

Note 77. [1]Spherical covariograme takes the form

No need to memorize (11.3)

$$c(h) = \begin{cases} \sigma^2 \left(1 - \frac{3}{2}\frac{\|h\|_1}{\phi} + \frac{1}{2}\left(\frac{\|h\|_1}{\phi}\right)^3\right) & \|h\|_1 \leq \phi \\ 0 & \|h\|_1 > \phi \end{cases}, \ h \in \mathbb{R}^3. \tag{11.3}$$

for $\sigma^2 > 0$ and $\phi > 0$. 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.

12. Anisotropy

Note 78. 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 79. The semivariogram $\gamma : \mathcal{H} \to \mathbb{R}$ 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 semivariograms $\gamma(h_1) \neq \gamma(h_2)$.

Definition 80. The intrinsically random field $(Z(s) : s \in \mathcal{S})$ is anisotropic if its semivariogram is anisotropic.

Definition 81. The covariogram $c : \mathcal{H} \to \mathbb{R}$ 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 82. The (weakly) stationary random field $(Z(s) : s \in \mathcal{S})$ is anisotropic if its covariogram is anisotropic.

Note 83. For brevity, below we discuss about intrinsic random fields and semivariograms, however the concepts/definitions apply to stationary random fields and covariograms when defined, as in Defs 79 & 81.

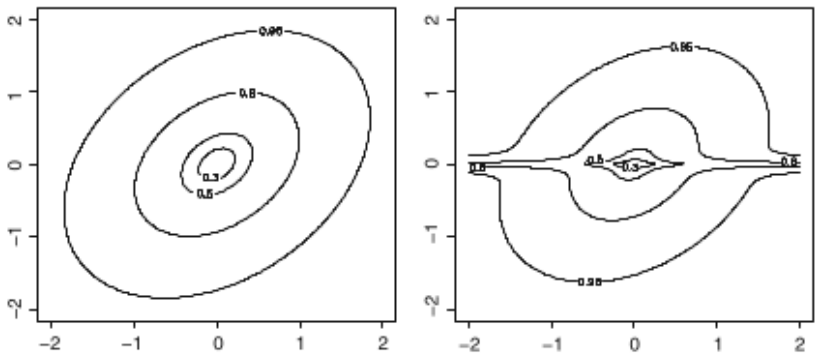
12.1. **Geometric anisotropy.**

Definition 84. The semivariogram $\gamma_{\text{g.a.}} : \mathcal{H} \to \mathbb{R}$ exhibits geometric anisotropy if it results from an A-linear deformation of an isotropic semivariogram with function $\gamma_{\text{iso}}(\cdot)$; i.e.

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

Note 85. Such semivariograms have the same sill in all directions but with ranges that vary depending on the direction. See Figure 12.1a.

[1]For it's derivation see Ch 8 in [4]



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 12.1. Isotropy vs Anisotropy

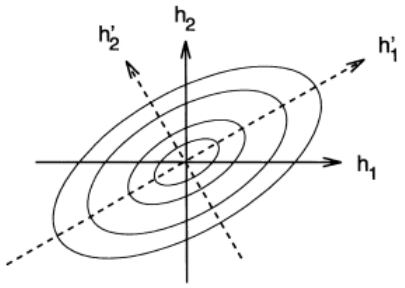


FIGURE 12.2. Rotation of the 2D coordinate system

Example 86. For instance, if $\gamma_{\text{g.a.}}(h) = \gamma_{\text{iso}}\left(\sqrt{h^\top Q h}\right)$, where $Q = A^\top A$.

Example 87. [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-semivariogram lines are spherical.

(1) [Rotate] Apply rotation matrix R to h such as $h' = Rh$. 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.

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

Definition 88. Support anisotropy is called the type of anysotropy that the semivariogram $\gamma(h)$ depends only on certain coordinates of h.

Example 89. If it is $\gamma(h = (h_1, h_2)) = \gamma(h_1)$, then I ve support anisotropy

Definition 90. Zonal anisotropy occurs when the semivariogram $\gamma(h)$ is the sum of several components each with a support anisotropy.

Example 91. Let γ' and γ'' be semivariogram with sills v' and v'' correspondingly. If it is $\gamma(h = (h_1, h_2)) = \gamma'(\|h_1\|) + \gamma''\left(\sqrt{\|h_1\| + \|h_2\|}\right)$, then I 've Zonal anisotropy because γ has a sill $v' + v''$ is direction $(0, 1)$ and a sill v' in direction $(1, 0)$.

Note 92. We have Zonal anisotropy then the semivariogram calculated in different directions suggest a different value for the sill (and possibly the range).

Note 93. If in 2D case, the sill in h_1 is larger than that in h_2, we can model zonal anysotropy of random field $(Z(s) : s \in \mathcal{S})$ by assuming $Z(s) = I(s) + A(s)$, where $I(s)$ is an isotropic random field with isotropic semivariogram γ_I along dimension of h_1 and $A(s)$ is an process with anisotropic semivariogram γ_I without effect on dimension h_1; i.e. $\gamma_Z(h) = \gamma_I(h) + \gamma_A(h)$.

12.3. **Non-linear deformations.**

Note 94. A (rather too general) non-stationary non-intrinsic random field model can be specified by considering semivariogram $2\text{Var}(Z(s) - Z(t)) = 2\gamma_o(\|G(s) - G(t)\|)$ such that a bijective non-linear (function) deformation $G(\cdot)$ of space $\mathcal{S}$ has been applied on the isotropic semivariogram 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 random field $(G(s) : s \in \mathcal{S})$ with semivariogram $2\text{Var}(G(s) - G(t)) = 2\gamma'_o(\|G'(s) - G'(t)\|)$ and so on...; then we will be talking about deep learning.

13. Geometrical properties of random fields

Note 95. We discuss basic geometric properties of random field we will use for modeling, as it can give us a deeper intuition on how to design appropriate spatial statistical models.

Definition 96. (Continuity in quadratic mean (q.m.)) Second-order random field $(Z(s) : s \in \mathcal{S})$ is q.m. continuous at $s \in \mathcal{S}$ if

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

Note 97. Consider random field $(Z(s) : s \in \mathcal{S})$ .Then

$$\mathrm{E}\left(Z(s+h) - Z(s)\right)^2 = \left(\mathrm{E}\left(Z(s+h)\right) - \mathrm{E}\left(Z(s)\right)\right)^2 + \mathrm{Var}\left(Z(s+h) - Z(s)\right)$$

- If Z is intrinsic r.f., then

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

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

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

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

Definition 98. Differentiable in quadratic mean (q.m.)) Second-order random field $(Z(s) : s \in \mathcal{S})$ is q.m. differentiable at $s \in \mathcal{S}$ there exist

$$\dot{Z}(s) = \lim_{h \to 0} \frac{Z(s+h) - Z(s)}{h}. \text{ in q.m.} \tag{13.1}$$

Proposition 99. *Let $c(s,t)$ be the covariance function of $Z = (Z(s) : s \in \mathcal{S})$. 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 (13.1).*

Example 100. The process with exponential c.f. $c(h) = \sigma^2 \exp\left(-|h|/\phi\right)$ is continuous because $\lim_{h \to 0} c(h) = \sigma^2 = c(0)$ but not differentiable because $\frac{\partial^2}{\partial h^2} c(h)$ does not exist at $h = 0$.

Part 2. Model building & related parametric inference

14. The Geostatistical model (the big picture)

14.1. Linear Model of Regionalization.

Note 101. A spatial phenomenon can be thought as being the sum of several independent subphenomena acting at different characteristic scales. A linear model can be set up to split the stochastic process representing the phenomenon into several uncorrelated stochastic processes, each with a different variogram or covariance function and characterizing different aspect of the overall phenomenon under investigation.

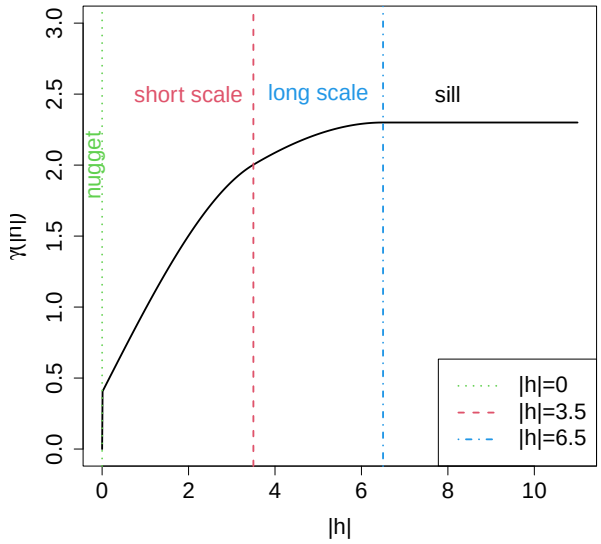


FIGURE 14.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|; \sigma^2 = 0.4\right)$.

14.1.1. *Decomposition of the random field.*

Note 102. The linear model of regionalization consider the decomposition of the random field of interest $Z(s)$ as a summation of m independent zero-mean random fields $\{Z_j(s); s \in \mathcal{S}\}_{j=0}^{m}$ each of them characterizing different spatial scales, as

$$Z(s) = \mu(s) + Z_1(s) + ... + Z_m(s) \tag{14.1}$$

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

Remark 103. In (14.1), let $Z_j(\cdot)$ be intrinsic random field with semivariogram $\gamma_j(\cdot)$ and mutually independent, then the semivariogram of $Z(\cdot)$ is $\gamma(\cdot) = \sum_{j=1}^{m} \gamma_j(\cdot)$.

Example 104. For instance consider (14.1) with $\mu(s) = 0$, $m = 3$, $Z_1(s)$ with a spherical semivariogram (11.3) with range $\phi_1 = 3.5$, $Z_2(s)$ with a spherical semi-variogram (11.3) with range $\phi_2 = 6.5$, and $Z_3(s)$ with a nugget semi-variogram. See the "sudden" changes of the line in Figure 14.1 representing change of spatial behavior.

14.1.2. *Scale of variation.*

Note 105. Cressi [1] considers the following intuitive decomposition

$$Z(s) = \mu(s) + W(s) + \eta(s) + \varepsilon(s), \ s \in S \tag{14.2}$$

where

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

$W(s)$**:** is a zero mean second order continuous intrinsic random field 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 intrinsic random field 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 106. 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 107. Clearly, smaller scale components, such as $\eta(s)$, $\varepsilon(s)$ could be identified if the sampling grid is sufficiently fine.

Note 108. Decomposition (14.2) is not unique and the components are not clearly identifiable from the data when modeled; e.g. one may find two pairs of $\mu(s), W(s)$ doing the same thing; yet, separating $\eta(s)$ and $\varepsilon(s)$ is difficult as they often describe changes with range smaller than that of the sites (!)

Note 109. The geostatistical model is often presented (with reference to (14.2)) is a form

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

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

Note 110. Alternatively, the hierarchical statistical model (Handout 1, 3.5) is used

$$Z(s) = Y(s) + \varepsilon(s), \ s \in S \tag{14.3}$$

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

Note 111. Another decomposition we will use

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

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

Note 112. In several problems, additional covariates may be considered. The available dataset is of the form $\{(x_i, s_i, Z_i)\}_{i \in \mathcal{S}}$ where $Z_i := Z(s_i, x_i)$ is the observed response at

location s_i, associated with the p-dimensional covariate $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$. Although not a necessity rule, the effect of the associated p-dimensional covariates is often expressed via the deterministic drift function $\mu(s,x) = \mathrm{E}(Z(s,x))$. E.g. in decomposition in (14.2)

$$Z(s,x) = \mu(s,x) + W(s) + \eta(s) + \varepsilon(s),\ s \in S, x \in \mathcal{X}. \tag{14.4}$$

Here, to simplify the presentation, we suppress dependence on possible covariates $x \in \mathcal{X}$.

15. Learning the semivariogram

Note 113. Consider a random field $(Z(s); s \in \mathcal{S})$, $\mathcal{S} \in \mathbb{R}^d$ observed at n sites $S = \{s_1, ..., s_n\}$, and hence a dataset $\{(s_i, Z(s_i))\}_{i=1}^n$.

Note 114. Consider a decomposition

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

where $\mu(\cdot)$ is an unknown deterministic drift and $\delta(\cdot)$ is a zero mean intrinsic random field.

Example 115. (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). See Figure 15.1a. This is the R dataset `meuse{sp}`.

Example 116. (Wolfcamp-aquifer dataset) We also consider the Wolfcamp-aquifer dataset in Exercise 2 in the Exercise sheet. See Figure 15.2a

15.1. The semvariogram cloud.

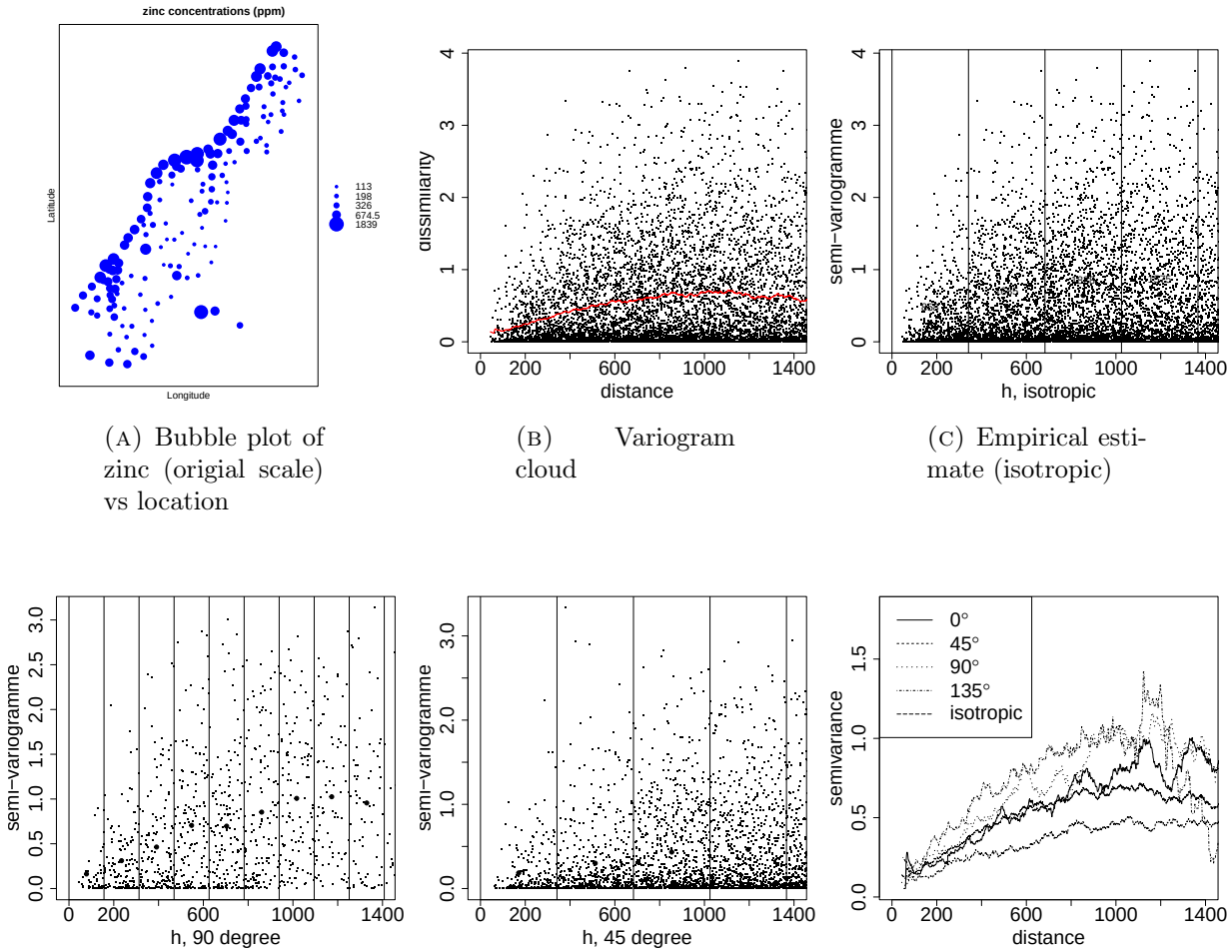
Assumption 117. *Assume that $(Z(s); s \in \mathcal{S})$ in an intrinsic random field with unknown constant mean; aka $Z(s) = \mu + \delta(s)$.*

Definition 118. Dissimilarity between pairs of data values $Z(s_a)$ and $Z(s_b)$ is called the measure

$$\gamma^\star(s_a, s_b) = \frac{1}{2}(Z(s_b) - Z(s_a))^2 \tag{15.1}$$

Definition 119. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (lag or orientation) then we get

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



(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)

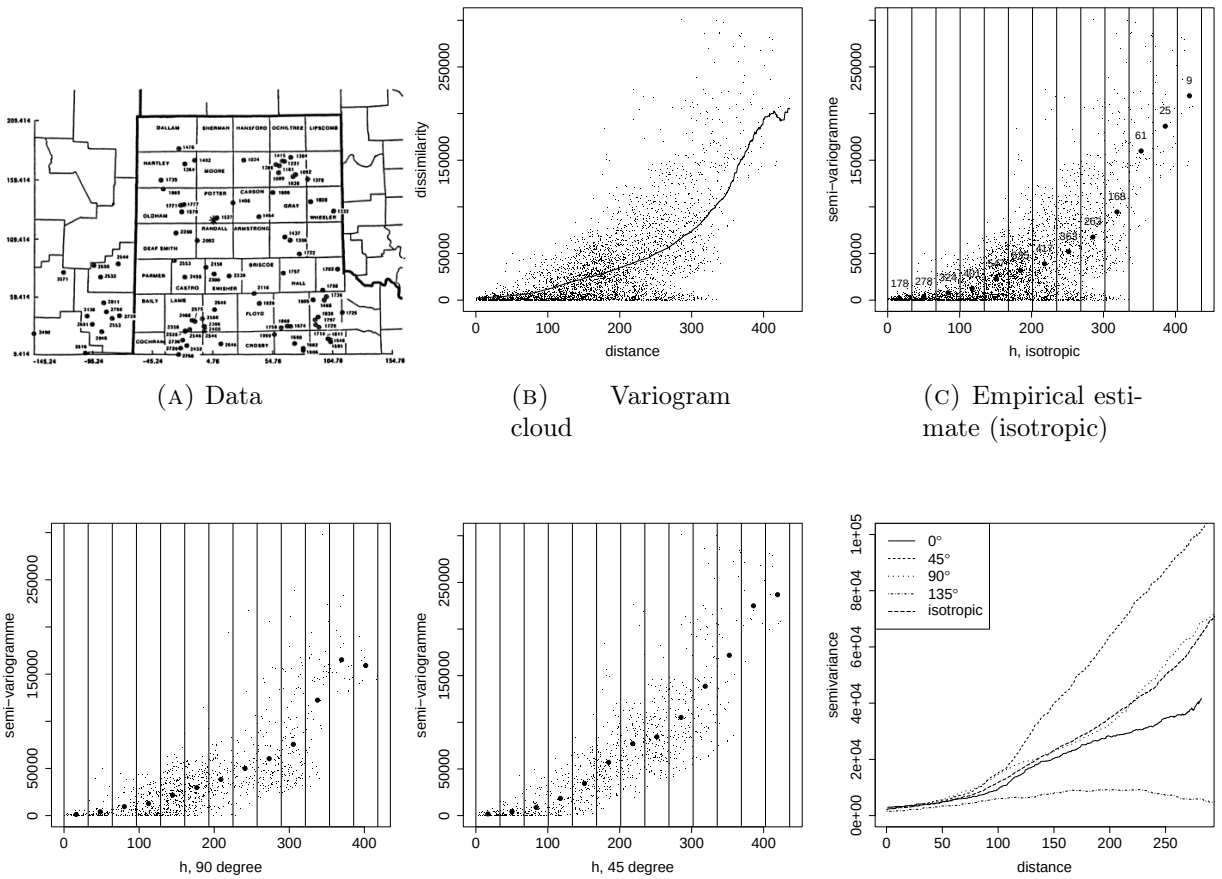
FIGURE 15.1. Meuse dataset variogram estimations (Zinc in log scale)

Definition 120. The semivariogram 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 121. Note that (15.1) is an unbiased estimator of the semivariogram and hence the semivariogram cloud is too.

Note 122. Often there is a smoothing of the cloud is superimposed onto the cloud itself to help us see semivariogram's characteristics (e.g., sill, nugget, range) which may be "hidden" due to potential outliers in the plot.



(A) Data

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 15.2. Wolfcamp-aquifer dataset variogram estimations

Example 123. Figure 15.1b and Figure 15.2b show the semivariogram 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.

15.2. Non-parametric semivariogram estimator of $\gamma(\cdot)$.

Assumption 124. *Assume that $(Z(s); s \in \mathcal{S})$ in an intrinsic random field with unknown constant mean; aka $\mu(\cdot)$ is an unknown constant.*

Proposition 125. *The Smoothed Matheron estimator $\hat{\gamma}(\cdot)$ of semivariogram $\gamma(\cdot)$ of an unknown constant mean intrinsic random field $Z(\cdot)$ is*

$$\hat{\gamma}_M(h) = \frac{1}{2\left|N_{r_1,r_2}(h)\right|} \sum_{\forall (s_i,s_j) \in N_{r_1,r_2}(h)} \left(Z(s_i) - Z(s_j)\right)^2 \tag{15.2}$$

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

$$B_{r_1,r_2}(h) = \left\{ x : |\|x\| - \|h\|| < r_1, \text{ and } \left\| \frac{x}{\|x\|_2} - \frac{h}{\|h\|_2} \right\|_2 < r_2 \right\} \tag{15.3}$$

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

Note 126. If we consider isotropic semivariogram $\gamma(\cdot)$ then the ball may just considerate only the length of the distance as

$$B_{r_1}(h) = \{x : |\|x\| - \|h\|| < r_1\} \tag{15.4}$$

because the direction does not have any effect.

Note 127. 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 128. 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 129. Figures 15.1c and 15.2c, show the nonparametric estimator ignoring the direction for the datasets Meuse and Wolfcamp-aquifer dataset. The estimator is calculated by using the ball in (15.4).

Example 130. Figures 15.2dand 15.1e show the nonparametric estimator considering directions 90° and 45° for the dataset Meuse. Figures 15.2d and 15.2e do the same for the Wolfcamp-aquifer dataset. The estimator is calculated by using the ball (15.3).

Note 131. In practice anisotropies are detected by inspecting experimental semivariograms in different directions and are induced into the model by tuning predefined anisotropy parameters.

Example 132. Figure 15.1f and 15.2a show the nonparametric semivariogram estimator for different directions for the two datasets. We observe possible anisotropy due to the differences in the lines.

15.3. Classic parametric estimator of $\gamma(\cdot)$.

Assumption 133. *Consider (for now) the assumption that $(Z(s) ; s \in \mathcal{S})$ in an intrinsic random field with unknown constant mean; aka $\mu(\cdot)$ is an unknown constant.*

Note 134. Smoothed Matheron estimator (15.2) does not necessarily satisfies semivariogram properties, such as negative definiteness. To address this we use a parametric family of appropriate semi-variogram functions and tune them against data.

Note 135. Popular parametrized isotopic semivariogrames are those Section 11.1. Anisotropic semi-variogrames/covariogrames can be specified by using isotropic ones and applying a rotation and dilation as in Example 86.

Note 136. Below are some properties that allow the specification of sophisticated semivariograms from simpler ones.

(1) $\tilde{\gamma}(h) = \gamma(Ah)$ where $\gamma(\cdot)$ is a semivariogram and A constant matrix.
(2) $\gamma(\cdot) = \sum_{i=1}^{n} a_i \gamma_i(\cdot)$, if $a_i \geq 0$, and $\{\gamma_i(\cdot)\}$ are semivariograms
(3) $\gamma(\cdot) = \prod_{i=1}^{n} \gamma_i(\cdot)$, if $\{\gamma_i(\cdot)\}$ are semivariograms
(4) $\gamma(\cdot) = \int \gamma_u(\cdot) \, \mathrm{d}F(u)$, if $\gamma_u(\cdot)$ is a semivariogram parametrized by $u \sim F$
(5) $\gamma(\cdot) = \lim_{n\to\infty} \gamma_n(\cdot)$ if $\gamma_n(\cdot)$ is semivariogram and the limit exists
(6) $\gamma_Z(h) = \gamma_Y(h) + \gamma_X(h)$ corresponds to random field $Z(s) = Y(s) + X(s)$ if $(Y(s) : s \in \mathcal{S})$ and $(X(s) : s \in \mathcal{S})$ are independent intrinsic random fields with semivariograms $\gamma_Y(\cdot)$ and $\gamma_X(\cdot)$.
(7) $\gamma(\cdot)$ is a semivariogram iff $\exp(-a\gamma(\cdot))$ is positive definite for any $a > 0$.

Example 137. $\gamma(h) = \|h\|^2$ is a semivariogram because $\exp\left(-a\|h\|_2^2\right)$ is a c.f. for any $a > 0$ and hence positive definite.

Note 138. For a q.m. continuous $(Z(s) : s \in \mathcal{S})$, it is $\lim_{\|h\|\to 0} \gamma(h) = 0$ 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 $\lim_{\|h\|\to 0} \gamma(h) = \sigma_\varepsilon^2 \neq 0$ aka a nugget. Nugget effect is often mathematically described by considering a decomposition ; [5; Ch 1.4.1]

$$Z(s) = Y(s) + \varepsilon(s)$$

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

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

Note 139. Let $\hat{\gamma}$ be the empirical semivariogram $\hat{\gamma}$ (e.g., Matheron (15.2)) computed at k classes, i.e. it is available $\{h_j, \hat{\gamma}(h_j)\}_{j=1}^{k}$. Let γ_θ be a parametrised semivariogram by the unknown θ. The Least Square Errors (LSE) estimator is $\hat{\gamma}_{\mathrm{LSE}}(h) = \gamma\left(h; \hat{\theta}_{\mathrm{LSE}}\right)$ where

$$\hat{\theta}_{\mathrm{LSE}} = \arg\min_{\theta} \left(\hat{\gamma} - \gamma(h;\theta)\right)^\top V(\theta) \left(\hat{\gamma} - \gamma(h;\theta)\right), \tag{15.5}$$

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

Note 140. An example is OLS

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\theta} \left(\sum_j (\hat{\gamma}(h_j) - \gamma(h;\theta))^2 \right) \tag{15.6}$$

Example 141. Figures 15.3a and 15.3b show the OLE and WLE estimates (15.6) and (18.1) of the exponential and spherical semivariogram for the Meuse dataset. Figure 15.3c shows the OLE and WLE estimates (15.6) and (18.1) of the exponential semi-variogram for the Wolfcamp dataset. The parametric semivariograms were tuned against the non-parametric estimator (15.2) presented in dots, as discussed in Proposition 139.

15.4. **Parametric learning of nonzero $\mu(\cdot)$ and $\gamma(\cdot)$.**

Note 142. Assume a random field model $(Z(s); s \in \mathcal{S})$ decomposed as

$$Z(s) = \mu(s) + \delta(s)$$

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

15.4.1. *Non-parametric inference.*

Note 143. Semi-parametric learning is as follows:

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

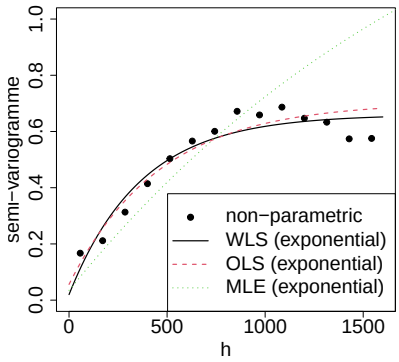
$$\hat{\beta}_{\text{LSE}} = \arg\min_{\theta} \left(\sum_i (Z(s_i) - \mu(s_i;\beta))^2 \right) \tag{15.7}$$

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

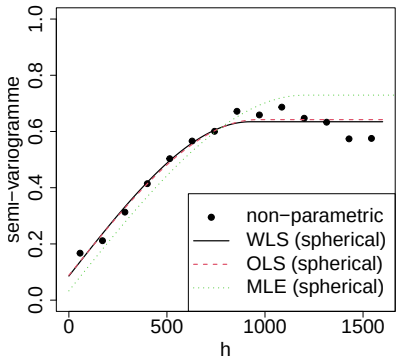
$$\hat{\delta}(s_i) = Z(s_i) - \mu\left(s_i; \hat{\beta}_{\text{LSE}}\right) \tag{15.8}$$

(3) Compute empirical variogram against $\hat{\delta}$ on $\mathscr{H}$ according to Proposition 125.
(4) Compute estimates $\hat{\theta}_{\text{LSE}}$ and $\hat{\gamma}_{\text{LSE}}(h)$ according to Proposition 139.

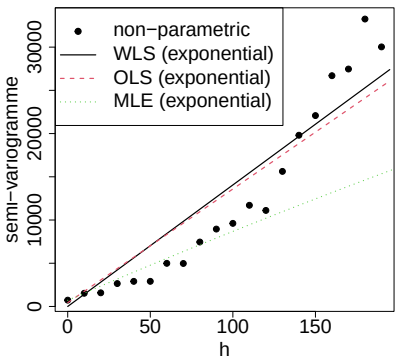
Example 144. Figure 15.3a and 15.3b fit an exponential c.f. and a spherical c.f. in the data of Meuse dataset (assuming constant mean); we cannot eyeball any big difference. Figure 15.3c fit an exponential c.f. in the data of Wolfcamp dataset (assuming constant mean); the fit looks really bad, possibly we should consider a non-constant mean and remove the trend.



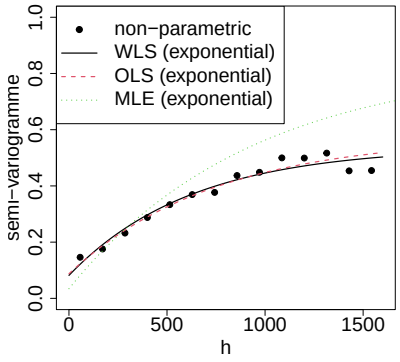
(A) Exponential c.f. with constant mean (Meuse dataset)



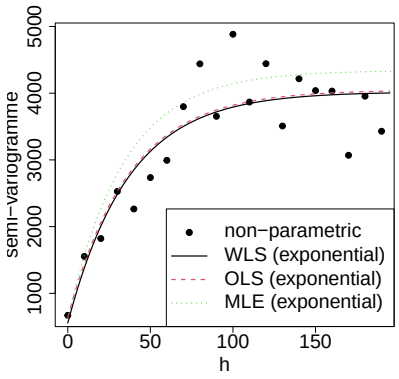
(B) Spherical c.f. with constant mean (Meuse dataset)



(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 15.3. Parametric training

Example 145. Figure 15.3d fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-42.8, -9.5 \cdot 10^{-4}, -6.6 \cdot 10^{-4})^\top$ in Meuse dataset. Possibly inference would suggest a constant mean function. Figure 15.3e fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-607, -1.12, -1.13)^\top$ in Wolfcamp dataset; we see an improvement in fit compared to Figure 15.3c.

15.4.2. *Parametric inference via MLE.*

Note 146. Assume that the probability distribution of the random field $(Z(s); s \in \mathcal{S})$ is known. The MLE estimates $\left(\hat{\beta}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}\right)$ of (β, θ) can be computed as

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

where $L(z_1, ..., z_n|\beta, \theta)$ is the associated likelihood function given observed data $\{(s_i, Z_i)\}_{i=1}^n$.

Example 147. If s $Z(\cdot) \sim \text{GP}(\mu(\cdot; \beta), c(\cdot, \cdot; \sigma^2, \phi))$, with $\mu(s; \beta) = \beta_0 + s_1\beta_1 + s_2\beta_2$ and $\gamma_{(\sigma^2,\phi)}(h) = \sigma^2(1 - \exp(-\phi h^2))$ then MLE of (β, σ^2, ϕ) is

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

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

16. (classical) Kriging (for prediction)

Note 148. "Kriging" 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 $\left\{s_j^*\right\}$ or a subregion v_0) of a spatial region $\mathcal{S}$ by properly averaging out data in the neighborhood around the location of interest.

16.1. **Universal Kriging.**

Note 149. Consider the statistical model specified as a stochastic process $(Z(s); s \in \mathcal{S})$ with

$$Z(s) = \mu(s) + \delta(s) \tag{16.1}$$

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 random field.

Note 150. Consider an 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. At these dataset points (16.1) is vectorized as

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

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

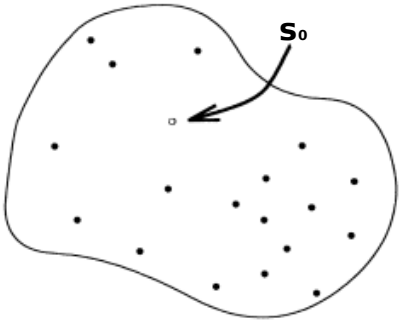


FIGURE 16.1. Kriging area

Note 151. Interested lies in learning/predicting $Z(s_0)$ at an unseen spatial location s_0 within the spatial domain $\mathcal{S}$ (Figure 16.1).

Note 152. "Universal Kriging" (UK) is the technique for producing a Best Linear Unbiased Estimator (BLUE) predictor for $Z_0 := Z(s_0)$ at spatial location $s_0 \in \mathcal{S}$ as a weighted average of the available data around in a neighborhood around that location .

Definition 153. The Universal Kriging (UK) predictor $Z_{\text{UK}}(s_0)$ of $Z(s_0)$ at location $s_0 \in \mathcal{S}$ is the Best Linear Unbiased Estimator (BLUE) of $Z(s_0)$ given the data $\{(s_i, Z_i)\}_{i=1}^n$.

Note 154. [**L**INEAR] The UK predictor $Z_{\text{UK}}(s_0)$ of $Z(s_0)$ at s_0 has the following linear form weighted by a set of tunable unknown weights $\{w_i\}$

$$Z_{\text{UK}}(s_0) = w_{n+1} + \sum_{i=1}^{n} w_i Z(s_i) \tag{16.2}$$

$$= w_{n+1} + w^\top Z \tag{16.3}$$

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

Note 155. [**U**nbiased] For (16.2), to satisfy unbiasness (that is zero systematic error"), we need

$$\mathrm{E}(Z_{\text{UK}}(s_0)) = \mathrm{E}(Z(s_0)) \Leftrightarrow w_{n+1} + \sum_{i=1}^{n} w_i \mathrm{E}(Z(s_i)) = \mu(s_0)$$

$$\Leftrightarrow w_{n+1} + \sum_{i=1}^{n} w_i \mu(s_i) = \mu(s_0) \Leftrightarrow w_{n+1} + \sum_{i=1}^{n} w_i (\psi(s_i))^\top \beta = (\psi(s_0))^\top \beta$$

$$\Leftrightarrow w_{n+1} + w^\top \Psi \beta = \Psi_0 \beta \tag{16.4}$$

where Ψ is matrix with $[\Psi]_{i,j} = \psi_j(s_i)$ and Ψ_0 is a (column) vector with $[\Psi_0]_{1,j} = \psi_j(s_0)$. Because in (16.4) both sides are polynomial w.r.t β all coefficients must be equal; hence

sufficient and necessary conditions for unbiaseness are

$$\text{(16.5)} \qquad \text{Assumption:} \qquad \left(\psi\left(s_0\right)\right)^{\top} = \sum_{i=1}^{n} w_i \left(\psi\left(s_i\right)\right)^{\top} \qquad \Leftrightarrow \Psi_0 = w^{\top}\Psi$$

$$\text{(16.6)} \qquad \text{Assumption:} \qquad w_{n+1} = 0$$

Note 156. We set $\psi_0\left(\cdot\right) = 1$; then (16.4) implies

$$\text{(16.7)} \qquad \text{Assumption:} \qquad \sum_{i=1}^{n} w_i = 1 \qquad \Leftrightarrow w^{\top}\underline{1} = 1$$

Note 157. The MSE of $Z_{\text{UK}}\left(s_0\right)$, given some of the above Assumptions, is

$$\begin{aligned}
\text{(16.8)} \qquad \text{MSE}\left(Z_{\text{UK}}\left(s_0\right)\right) &= \text{E}\left(Z_{\text{UK}}\left(s_0\right) - Z\left(s_0\right)\right)^2 \\
&= \text{E}\left(\Psi\beta + \delta\left(s_0\right) - w^{\top}\Psi\beta - w^{\top}\delta\right)^2; \qquad \left\{\text{let } \delta = \left(\delta\left(s_1\right), ..., \delta\left(s_n\right)\right)^{\top}\right\} \\
\text{(16.9)} \qquad &= \text{E}\left(\sum_{i=1}^{n} w_i \delta\left(s_i\right) - \delta\left(s_0\right)\right)^2 \quad \overset{\sum_{i=1}^{n} w_i = 1}{=} \quad \text{E}\left(\sum_{i=1}^{n} w_i \left(\delta\left(s_i\right) - \delta\left(s_0\right)\right)\right)^2 \\
\text{(16.10)} \qquad &= -\text{E}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left(\delta\left(s_i\right) - \delta\left(s_j\right)\right)^2 - 2\frac{1}{2}\sum_{i=1}^{n} w_i \left(\delta\left(s_i\right) - \delta\left(s_0\right)\right)^2\right) \\
\text{(16.11)} \qquad &= -\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j \frac{1}{2}\text{E}\left(\delta\left(s_i\right) - \delta\left(s_j\right)\right)^2 + 2\sum_{i=1}^{n} w_i \frac{1}{2}\text{E}\left(\delta\left(s_i\right) - \delta\left(s_0\right)\right)^2
\end{aligned}$$

Note 158. To impose tractability, we consider the following additional assumption

$$\text{(16.12)} \quad \text{Assumption:} \qquad \left(\delta\left(s\right); s \in \mathcal{S}\right) \text{ intrinsic random field} \quad \text{with semivariogram } \gamma\left(\cdot\right)$$

Note 159. Since $\left(\delta\left(s\right); s \in \mathcal{S}\right)$ is intrinsic stationary, and its semivariogram exists, $\text{MSE}\left(Z_{\text{UK}}\left(s_0\right)\right)$ can be expressed w.r.t. the semivariogram as

$$\begin{aligned}
\text{(16.13)} \qquad \text{MSE}\left(Z_{\text{UK}}\left(s_0\right)\right) &= -\gamma\left(s_0 - s_0\right) - \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j \gamma\left(s_i - s_j\right) + 2\sum_{i=1}^{n} w_i \gamma\left(s_i - s_0\right) \\
\text{(16.14)} \qquad &= -\gamma_{00} - w^{\top}\boldsymbol{\Gamma} w + 2w^{\top}\boldsymbol{\gamma}_0
\end{aligned}$$

where $w = \left(w_1, ..., w_n\right)^{\top}$, $\boldsymbol{\gamma}_0 = \left(\gamma\left(s_0 - s_i\right), ..., \gamma\left(s_0 - s_n\right)\right)^{\top}$, $\gamma_{00} = \gamma\left(s_0 - s_0\right)$, and $\left[\boldsymbol{\Gamma}\right]_{i,j} = \gamma\left(s_i - s_j\right)$.

Note 160. [**B**est] The Lagrange function for minimizing the MSE (16.13) under (16.4) is

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

Note 161. The UK system of equations is

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

$$\begin{cases} 0 = -2\sum_{j=1}^{n} w_{\text{UK},j}\gamma(s_i - s_j) + 2\gamma(s_0 - s_i) - \sum_{j=0}^{p} \lambda_{\text{UK},j}\psi_j(s_i), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1,...,n \\ \psi_j(s_0) = \sum_{i=1}^{n} w_{\text{UK},i}\psi_j(s_i), \quad j = 0,...,p \end{cases} \iff \tag{16.15}$$

$$\begin{cases} 0 = -2\boldsymbol{\Gamma} w + 2\boldsymbol{\gamma}_0 - \Psi\lambda_{\text{UK}} \\ \Psi_0 = w_{\text{UK}}^{\top}\Psi \end{cases} \tag{16.16}$$

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

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

$$\lambda_{\text{UK}} = 2\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right) \tag{16.17}$$

Then by substituting (16.17) in (16.16), I get the UK weights as

$$w_{\text{UK}} = \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\gamma}_0 - \Psi\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)\right) \tag{16.18}$$

Note 162. Then by substituting 16.18 in (19.2) , the UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

$$Z_{\text{UK}}(s_0) = \left(\boldsymbol{\gamma}_0 - \Psi\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)\right)^{\top}\boldsymbol{\Gamma}^{-1}Z \tag{16.19}$$

and by substituting 16.18 in (16.14) its standard error is

$$\sigma_{\text{UK}}(s_0) = \sqrt{-w_{\text{UK}}^{\top}\boldsymbol{\Gamma} w_{\text{UK}} + 2w_{\text{UK}}^{\top}\boldsymbol{\gamma}_0} \tag{16.20}$$

$$= \sqrt{\boldsymbol{\gamma}_0\Gamma^{-1}\boldsymbol{\gamma}_0 - \left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)^{\top}\left(\Psi^{\top}\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^{\top}\Gamma^{-1}\boldsymbol{\gamma}_0 - \Psi_0^{\top}\right)} \tag{16.21}$$

Note 163. $(1-\alpha)\,100\%$ Prediction interval of UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

$$\left(Z_{\text{UK}}(s_0) - q_{\alpha/2}\sqrt{\sigma^2_{\text{UK}}(s_0)}\ ,\ Z_{\text{UK}}(s_0) + q_{1-\alpha/2}\sqrt{\sigma^2_{\text{UK}}(s_0)}\right) \tag{16.22}$$

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

Note 164. Note that we have not assumed a particular distribution of $Z(\cdot)$ or $\delta(\cdot)$, but only stationarity assumptions.

Note 165. Assumption in (21.3) restricts the available models (16.2) under consideration. Essentially, it addresses the non-determination issue caused by the assumption (16.12) (introducing intrinsic stationarity δ) which can only partially characterize the covariance function of δ. Also it ensures conditional negative definiteness for 16.14.

Note 166. It was not necessary to consider the intrinsic stationarity in Note 158 to derive Universal Kriging predictor equations; formulas (16.19) & (16.20) could have been derived with respect to the covariance function $c(\cdot,\cdot)$ of $(\delta(\cdot))$ instead of its semivariogram $\gamma(\cdot)$. Here, intrinsic stationarity was assumed for practical reasons; it allowed us to express 16.19 and (16.20) as functions of the semivariogram whose estimation has been discussed in Section 15.

Note 167. Practical use of (16.19), (16.20), and (16.22) requires knowledge of the linear drift coefficients $\{\beta_j\}$ and the semivariogram $\gamma(\cdot)$. If $\{\beta_j\}$ and $\gamma(\cdot)$ are unknown, a way to learn them, is as follows. Consider a separate dataset training dataset $\{(s_i', Z_i')\}_{i=1}^{n'}$ (although in practice we use the same $\{(s_i, Z_i)\}_{i=1}^{n}$). Model the semivariogram $\gamma(\cdot)$ with a conditional negative semi-definite function $\gamma_\theta(\cdot)$ parameterized by an unknown parameter θ. Then train/estimate $\{\beta_j\}$ and $\gamma_\theta(\cdot)$ against dateset $\{(s_i', Z_i')\}_{i=1}^{n'}$ by using the semi-parametric procedure in Note 143 or Note 146, and plug the estimates in (16.19), (16.20), and (16.22) .

Note 168. The dataset $\{(s_i', Z_i')\}_{i=1}^{n'}$ used to train $\{\beta_j\}$ and $\gamma(\cdot)$ should be different than the dataset $\{(s_i, Z_i)\}_{i=1}^{n'}$ used to produce the Kriging predictive equations (16.19), (16.20), and (16.22). This is because in theory the same dataset should not be used twice (it leads to overconfidence). However often in practice the same dataset is use in violation of the theory because usually this has a small impact in the results.

Example 169. [2] Consider the example with the Meuse dataset. Fig 16.2b presents the UK prediction $Z_{\text{UK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (16.1) for which the deterministic drift mean has a linear form $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$. Following Notes 167 and 143, we computed the $\hat{\beta}_{\text{LSE}}$ of β by (15.7), we computed the residual process $\left\{\hat{\delta}_i\right\}$ by removing the linear trend by (15.8), we computed the non-parametric estimate of semivariogram $\hat{\gamma}$ (15.2) of δ as in Proposition 125. We considered a (parametric) isotropic exponential semivariogram $\gamma_{(\sigma^2,\phi)}$ 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 (15.6) (see Figure 15.3d).Then we plugged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)}$ in (16.19) to compute the UK weights w_{UK} for the UK predictor $Z_{\text{UK}}(s_0) = w_{\text{UK}} Z$ for any $s_0 \in \mathcal{S}$.

Example 170. (Cont. Examples 115, 132) Consider the example with the Meuse dataset. The dataset has another measurement (a potential regressor in the deterministic mean $\mu(s)$), the "distance to the Meuse river bed" $\{d_i\}$ at the associated locations $\{s_i\}$, let's denote it by d. Figure 16.2c shows a rather linear relationship between Z and $\sqrt{d}$. Consider deterministic drift mean has a linear form $\mu(s,d) = \beta_0 + \beta_1\sqrt{d(s)}$ while the rest specification is the same

[2] https://github.com/georgios-stats/Spatio-Temporal_Statistics_Michaelmas_2024/blob/main/Lecture_notes/R_scripts/03.Geostatistical_data_meuse_gstats.R

as in Example 169. We follow the same procedure as in Example 169 and we get the UK predictor in Figure 16.2d.

16.2. **Ordinary Kriging.**

Note 171. Ordinary Kriging (OK) addresses spatial prediction in cases that the specified statistical model on $(Z(s); s \in \mathcal{S})$ has the form

$$Z(s) = \beta_0 + \delta(s) \tag{16.23}$$

with unknown $\beta_0 \neq 0$ and intrinsically stationary process $(\delta(s); s \in \mathcal{S})$.

Note 172. 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 173. [The derivation is in (Exercise 17 Exercise sheet).] For demonstration, we mention some key equations of OK

$$\mathfrak{L}(w, \lambda) = \underbrace{-w^\top \mathbf{\Gamma} w + 2w^\top \boldsymbol{\gamma}_0}_{=\mathrm{MSE}(Z_{\mathrm{OK}}(s_0))} - \lambda \quad \underbrace{\left(w^\top \underline{1} - 1\right)}_{\text{Assumption } \sum_{i=1}^n w_i = 1} \tag{16.24}$$

The OK system of equations is $0 = \nabla_{(\{w_i\},\lambda)} L(w, \lambda)\big|_{(w,\lambda)}$ producing

$$\begin{cases} 0 = -2\mathbf{\Gamma} w_{\mathrm{OK}} + 2\boldsymbol{\gamma}_0 - 1\lambda \\ w_{\mathrm{OK}}^\top 1 = 1 \end{cases} \tag{16.25}$$

the weights are

$$w_{\mathrm{OK}} = \mathbf{\Gamma}^{-1}\left(\boldsymbol{\gamma}_0 + \frac{1 - 1^\top \mathbf{\Gamma}^{-1}\boldsymbol{\gamma}_0}{1^\top \mathbf{\Gamma}^{-1} 1} 1\right) \tag{16.26}$$

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

$$\sigma^2_{\mathrm{OK}}(s_0) = \boldsymbol{\gamma}_0^\top \mathbf{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \frac{\left(1 - 1^\top \mathbf{\Gamma}^{-1}\boldsymbol{\gamma}_0\right)^2}{1^\top \mathbf{\Gamma}^{-1} 1}. \tag{16.27}$$

Example 174. (Cont. Examples 115, 132, 170) Consider the example with the Meuse dataset. with $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$. We do not see much difference between OK in Figure 16.2a and UK in Figure 16.2b possibly because the slops in the linear trend (mean) of UK are rather small and insignificant (See Example 145).

16.3. **Simple Kriging.**

Note 175. Simple Kriging (SK) addresses spatial prediction in cases that the specified statistical model on $(Z(s); s \in \mathcal{S})$ has the form

$$Z(s) = \mu(s) + \delta(s) \tag{16.28}$$

where the deterministic mean $\mu(s)$ is known, and $(\delta(s); s \in \mathcal{S})$ is a weakly stationary process with covariogram $c(\cdot)$.

Example 176. [The derivation is in (Exercise 15 in the Exercise sheet).] It does not require any assumption in the weights such as (16.5) or (16.25).As a supplementary and for demonstration, we mention the SK predictor at s_0 and standard error:

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

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 177. Consider the example with the Meuse dataset. Fig 16.2a presents the OK prediction $Z_{\mathrm{OK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (16.23) that is the UK case (16.1) for when $\mu(s) = \beta_0$. First we computed the non-parametric semivariogram $\hat{\gamma}$ (15.2) as in Proposition 125; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ where we computed the OLS $\hat{\theta}_{\mathrm{OLS}} = \left(\hat{\sigma}^2_{\mathrm{OLS}}, \hat{\phi}_{\mathrm{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (15.6) (see Figure 15.3a); and then we plugged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\mathrm{OLS}}, \hat{\phi}_{\mathrm{OLS}}\right)}$ in (16.26) to compute the OK weights w_{OK} for the OK predictor $Z_{\mathrm{OK}}(s_0) = w_{\mathrm{OK}} Z$ for any $s_0 \in \mathcal{S}$.

17. The Bayesian Kriging paradigm (Hierarchical modeling)

17.1. General framework.

Note 178. The Bayesian framework provides an elegant solution for taking into account the uncertainty on variogram or covariance parameters.

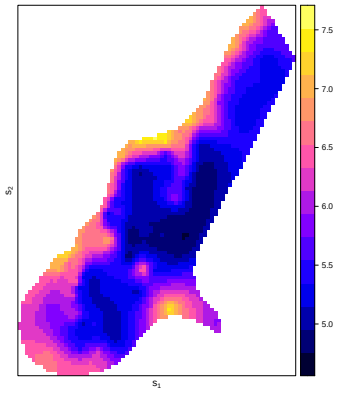
Note 179. Consider the geostatistical model for $(Z(s); s \in \mathcal{S})$ with a scale decomposition such as in (14.3)

$$Z(s) = \underbrace{\mu(s) + w(s)}_{=Y(s)} + \varepsilon(s), \quad s \in \mathcal{S} \tag{17.1}$$

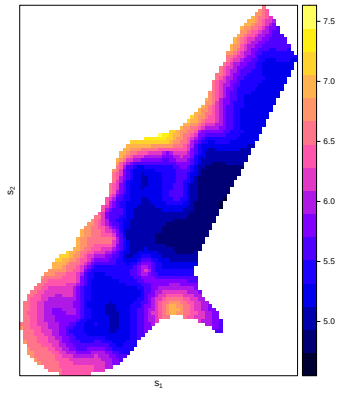
where, $\mu(s) = \mathrm{E}(Z(s))$ is an unknown drift function modelling large scale variations, $(w(s); s \in \mathcal{S})$ is a zero mean random field modelling lower scale variation, and $(\varepsilon(s); s \in \mathcal{S})$ is a nugget random field modeling measurement errors. Also $Y(s) = \mu(s) + w(s)$.

Note 180. Consider a dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $Z_i = Z(s_i)$ being a realization of (17.1) at site $s_i \in \mathcal{S}$.

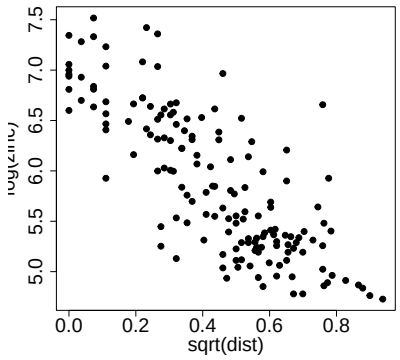
Note 181. Unlike in the traditional kriging framework, in Bayesian kriging, we have to specify a certain probabilistic model on the spatial random fields. Uncertainty can be decomposed



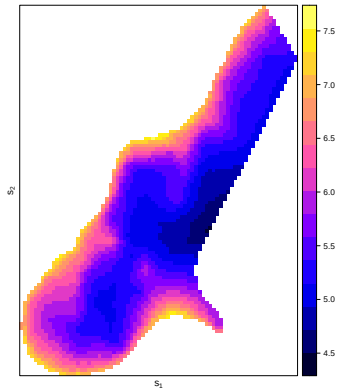
(A) Ordinary Kriging, constant mean, exponential covariogram



(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.

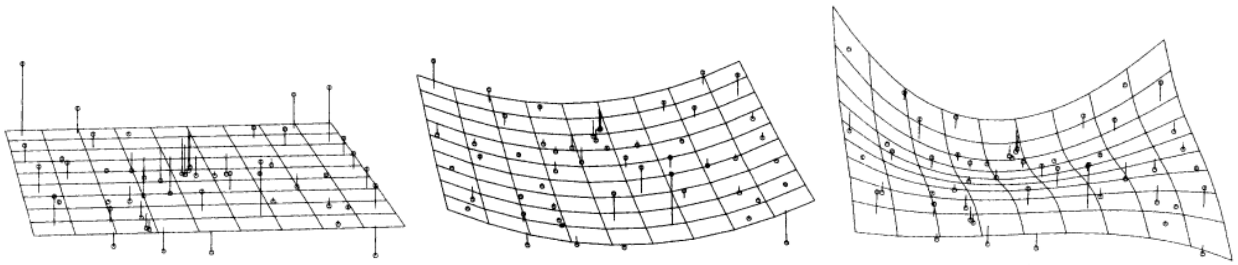


(D) Universial Kriging, mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d}$, exponential covariogram

FIGURE 16.2. Kriging Meuse dataset.

according to the Hierarchical spatial model

$$\begin{cases} Z|Y,\theta_1,\theta_2 & \text{data model} \\ Y|\theta_1,\theta_2 & \text{spatial process model} \\ \theta_1 & \text{hyper-priormodel (optional layer)} \end{cases} \tag{17.2}$$



(A) (B) (C)

FIGURE 17.1. Examples representing the hierarchical spatial model 17.2 for different values of ϑ

with

$$\text{pr}\left(Z, Y, \theta_1|\theta_2\right) = \text{pr}\left(Z|Y, \theta_1, \theta_2\right) \text{pr}\left(Y|\theta_1, \theta_2\right) \text{pr}\left(\theta_1|\theta_2\right)$$

where $Z = \left(Z_1, ..., Z_n\right)^\top$, $Y = \left(Y_1, ..., Y_n\right)^\top$, with $Y_i = Y\left(s_i\right)$. Here, $\theta_1 \in \Theta_1$ is an unknown random hyper parameter following prior $\theta_1|\theta_2 \sim \text{pr}\left(\theta_1|\theta_2\right)$, and $\theta_2 \in \Theta_2$ is an unknown fixed parameter without a specified prior.

Note 182. Spatial process model expresses the scientific uncertainty (e.g., that coming from $\left(Y\left(\cdot\right)\right)$ as quantified via the specified distribution $\text{pr}\left(Y|\theta\right)$ possibly labeled by some hyper-parameter $\theta = \left(\theta_1, \theta_2\right)^\top$. Data model expresses the measurement uncertainty (e.g., that coming from $\left(\varepsilon\left(\cdot\right)\right)$) as quantified via the distribution $\text{pr}\left(Z|Y, \theta\right)$ possibly labeled by some parameter θ.

Note 183. Figure 17.1 presents a visualization of the hierarchical model in Notes 181. The surfaces can be considered as a realization of the spatial process model $Y\left(\cdot\right)$, and the dotes can be considered as realizations of the data model $\varepsilon\left(\cdot\right)$ at specific sites given the spatial process.

Note 184. Under Bayesian model (17.2), unknown but fixed θ_2 can be learned pointwise by computing a point estimator $\hat{\theta}_2$ via marginal likelihood maximization

$$\hat{\theta}_2 = \arg\min_{\theta_2} \left(-2 \log\left(\text{pr}\left(Z|\theta_2\right)\right)\right),$$

$$\text{where} \quad \text{pr}\left(Z|\theta_2\right) = \int \text{pr}\left(Z, Y, \theta_1|\theta_2\right) \text{d}Y \, \text{d}\theta_1$$

Note 185. Under Bayesian model (22.8), uncertainty about unknown random θ_1 can be represented by the posterior distribution

$$\text{pr}\left(\theta_1|Z, \theta_2 = \hat{\theta}_2\right) = \frac{\text{pr}\left(Z|\theta_1, \theta_2 = \hat{\theta}_2\right) \text{pr}\left(\theta_1|\theta_2 = \hat{\theta}_2\right)}{\text{pr}\left(Z|\theta_2 = \hat{\theta}_2\right)}$$

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

Note 186. The posterior predictive distributions of the spatial process model $(Y(\cdot))$ given the data Z is

$$\text{pr}\left(Y(s_0)|Z, \theta_2 = \hat{\theta}_2\right) = \int \text{pr}\left(Y(s_0)|Z, \theta_1, \theta_2 = \hat{\theta}_2\right) \text{pr}\left(\theta_1|Z, \theta_2 = \hat{\theta}_2\right) \text{d}\theta_1 \tag{17.3}$$

and the marginal process $(Z(\cdot))$ given the data Z is

$$\text{pr}\left(Z(s_0)|Z, \theta_2 = \hat{\theta}_2\right) = \int \text{pr}\left(Z(s_0)|Z, \theta_1, \theta_2 = \hat{\theta}_2\right) \text{pr}\left(\theta_1|Z, \theta_2 = \hat{\theta}_2\right) \text{d}\theta_1 \tag{17.4}$$

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

Note 187. The Bayes Kriging predictor $\hat{Y}_{\text{BK}}(s_0)$ of $Y(s_0)$ at unseen location s_0 equations is the optimizer

$$\hat{Y}_{\text{BK}}(s_0) = \underset{Y_{\text{BK}}(s_0)}{\arg\min} \left(\text{E}\left(\ell\left(Y_{\text{BK}}(s_0), Y(s_0)\right)|Z\right)\right) \tag{17.5}$$

given a pre-specified loss function $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$. The expectation is under the probability (17.4). For Bayes Kriging predictor $\hat{Z}_{\text{BK}}(s_0)$ of $Z(s_0)$ at s_0 is

$$\hat{Z}_{\text{BK}}(s_0) = \underset{Z_{\text{BK}}(s_0)}{\arg\min} \left(\text{E}\left(\ell\left(Z_{\text{BK}}(s_0), Z(s_0)\right)|Z\right)\right) \tag{17.6}$$

17.2. **An example: Gaussian process regression.**

Note 188. We are going through a particular example of the Bayesian Kriging (or Bayesian Gaussian process regression) to demonstrate how the "Bayesian Kriging" works.

List of useful formulas.

Fact 189. *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*

$$AX + BY + c \sim N\left(A\mu_X + B\mu_Y + c, A\Sigma_X A^\top + B\Sigma_Y B^\top\right) \tag{17.7}$$

Fact 190. *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 191. *(Woodbury matrix identity) It is*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1} DA^{-1}$$

Fact 192. *[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

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

Note 193. Consider an available dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $\{Z_i \in \mathbb{R}\}$ and $\{s_i \in \mathcal{S}\}$. The i-th datum $Z_i = Z(s_i)$ may be an independent measurement of an unobserved quantity $Y_i = Y(s_i)$ at location s_i but contaminated by additive random error $\varepsilon_i = \varepsilon(s_i)$; i.e. $Z_i = Y_i + \epsilon_i$ for $i = 1, ..., n$. Interest lies in recovering functions $Z(\cdot)$ and/or $Y(\cdot)$ over the spatial domain $\mathcal{S}$.

Specifying the hierarchical model.

Note 194. The geostatistical model can be considered as

$$Z(s) = \underbrace{\mu(s) + w(s)}_{=Y(s)} + \varepsilon(s), \quad s \in \mathcal{S} \tag{17.8}$$

$\mu(\cdot)$**:** is a deterministic systematic drift $\mu(s) = \mathrm{E}(Z(s))$ modeling large scale variation;
$w(\cdot)$**:** models smaller scale variation, it is modeled as a Gaussian process

$$w(\cdot) \sim \mathrm{GP}(0, c(\cdot, \cdot))$$

with covariance function $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and mean function zero.

$\varepsilon(\cdot)$**:** models the additive random measurement error, it is modeled as a Gaussian process

$$\varepsilon(\cdot) \sim \mathrm{GP}\left(0, c_{\text{nugget}}\left(\cdot,\cdot|\sigma^2\right)\right)$$

with nugget covariance function $c_{\text{nugget}}(s,s'|\sigma^2) = \sigma^2 1_{\{0\}}(|s-s'|)$.

Assuming, $w(\cdot) \perp \varepsilon(\cdot)$, it is implied

$Y(\cdot)$**:** models the noisiness signal

$$Y(\cdot) \sim \mathrm{GP}(\mu(\cdot), c(\cdot,\cdot))$$

$Z(\cdot)$**:** is the marginal process

$$Z(\cdot) \sim \mathrm{GP}(\mu(\cdot), c_Z(\cdot,\cdot))$$

where $c_Z(s,s') = c(s,s') + \sigma^2 1_{\{0\}}(|s-s'|)$ assuming that $w(\cdot)$ and $\varepsilon(\cdot)$ are independent.

Note 195. The resulted hierarchical model is

$$\begin{cases} Z|Y \sim \mathrm{N}(Y, I\sigma^2) & \text{data model} \\ Y \sim \mathrm{N}(\mu(S), C(S,S)) & \text{spatial process model} \end{cases} \tag{17.9}$$

where $[Z]_i = Z(s_i)$, $[Y]_i = Y(s_i)$, $[\mu(S)]_i = \mu(s_i)$, and $[C(S,S)]_{i,j} = c(s_i,s_j)$.

The predictive distribution of $Y(\cdot)|Z$ under *(17.9).*

Note 196. Assume a vector of "unseen" sites $S_* = (s_{*,1},...,s_{*,q})^\top$ for any $q \in \mathbb{N}_0$. The join marginal distribution of $(Y_*, Z)^\top$ where $Z := Z(S)$ and $Y_* := Y(S_*)$ is

$$\begin{pmatrix} Y_* \\ Z \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(S_*) \\ \mu(S) \end{pmatrix}, \begin{pmatrix} C(S_*,S_*) & (C(S_*,S))^\top \\ C(S_*,S) & C(S,S)+I\sigma^2 \end{pmatrix}\right) \tag{17.10}$$

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

Note 197. The (posterior) predictive distribution of $Z_*|Z$ is the conditional distribution

$$Y_*|Z \sim \mathrm{N}(\mu_1(S_*), C_1(S_*,S_*)) \tag{17.11}$$

where

$$C_1(S_*,S_*) = C(S_*,S_*) - (C(S,S_*))^\top \left(C(S,S)+I\sigma^2\right)^{-1} C(S,S_*) \tag{17.12}$$

$$\mu_1(S_*) = \mu(S_*) - (C(S,S_*))^\top \left(C(S,S)+I\sigma^2\right)^{-1} (\mu(S)-Z) \tag{17.13}$$

[We used Fact 192].

Note 198. Since derivation of (17.11) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (17.11) can be extended to a Gaussian Process

$$Y(\cdot)|Z \sim \mathrm{GP}(\mu_1(\cdot), c_1(\cdot,\cdot)) \tag{17.14}$$

with

$$\begin{aligned} c_1(s,s') =& c(s,s') - (C(S,s))^\top \left(C(S,S) + I\sigma^2\right)^{-1} C(S,s') \\ \mu_1(s) =& \mu(s) - (C(S,s))^\top \left(C(S,S) + I\sigma^2\right)^{-1} (\mu(S) - Z) \end{aligned}$$

for any $s, s' \in \mathcal{S}$. This is the predictive process for noiseless signal $Y(s)$ at any $s \in \mathcal{S}$ given Z. *[Here we used the definition of GP (Definition 17) given Note* 196*].*

The predictive distribution of $Z(\cdot)|Z$ under *(17.9).*

Note 199. It is a Gaussian Process

$$Z(\cdot)|Z \sim \mathrm{GP}(\mu_1(\cdot), c_1^z(\cdot,\cdot)) \tag{17.15}$$

with $c_1^z(s,s') = c_1(s,s') + \sigma^2 1_{\{0\}}(|s-s'|)$. This is because $Z(\cdot) = Y(\cdot) + \varepsilon(\cdot)$ with $Y(\cdot) \perp \varepsilon(\cdot)$.

Learning $\mu(\cdot)$ *and* $c(\cdot,\cdot)$ *via parametrization.*

Note 200. Let $\mu(\cdot)$ be parametrised as $\mu(s|\beta) = \psi(s)^\top \beta$ with known basis functions $\psi(\cdot) = (\psi_1(\cdot), ..., \psi_p(\cdot))^\top$, and unknown random parameter $\beta \in \mathbb{R}^p$ following prior distribution $\beta \sim \mathrm{N}(b, B)$, $B > 0$. Let $c(\cdot,\cdot|\theta)$ be parametrised as $c(\cdot,\cdot) = c(\cdot,\cdot|\theta)$ with unknown fixed parameter $\theta \in \Theta$. Let σ^2 be an unknown fixed parameter.

Note 201. The Bayesian hierarchical model summaries to

$$\begin{cases} Z|Y,\sigma^2 \sim \mathrm{N}(Y, I\sigma^2) & \text{data model} \\ Y|\beta,\theta \sim \mathrm{N}(\mu(S), c(S,S)) & \text{spatial process model} \\ \beta \sim \mathrm{N}(b,B) & \text{hyper-prior model} \end{cases} \tag{17.16}$$

Note 202. The posterior of β given data Z and θ is computed via the Bayes theorem

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

and results as

$$\beta|Z,\theta,\sigma^2 \sim \mathrm{N}\left(b_n\left(\theta,\sigma^2\right), B_n\left(\theta,\sigma^2\right)\right) \tag{17.17}$$

with

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

[For the calculations of the derivation see Exercise 18 in *the Exercise sheet.]*

Note 203. Given a vector of "unseen" sites $S_* = \left(s_{*,1},...,s_{*,q}\right)^\top$ for any $q \in \mathbb{N}_0$, I integrate 17.11 with respect to (17.17) i.e.

$$\begin{aligned}\operatorname{pr}\left(Y_*|Z,\theta,\sigma^2\right) &= \int \operatorname{pr}\left(Y_*|Z,\beta,\theta,\sigma^2\right)\operatorname{pr}\left(\beta|Z,\theta,\sigma^2\right)\mathrm{d}\beta \\ &= \int \mathrm{N}\left(Y_*|\mu_1\left(S_*|\beta,\theta,\sigma^2\right),C_1\left(S_*,S_*|\theta,\sigma^2\right)\right)\mathrm{N}\left(\beta|b_n\left(\theta,\sigma^2\right),B_n\left(\theta,\sigma^2\right)\right)\mathrm{d}\beta\end{aligned}$$

where now $C\left(|\theta,\sigma^2\right)$, $C_1\left(|\theta,\sigma^2\right)$ and $\mu_1\left(|\theta,\sigma^2,\beta\right)$ are parameterized by β, θ, and σ^2. By using Fact 190

$$Y_*|Z,\theta,\sigma^2 \sim \mathrm{N}\left(\mu_2\left(S_*|\theta,\sigma^2\right),C_2\left(S_*,S_*|\theta,\sigma^2\right)\right) \tag{17.18}$$

where

(17.19)

$$\begin{aligned}C_2\left(S_*,S_*|\theta,\sigma^2\right) =& C_1\left(S_*,S_*|\theta,\sigma^2\right) \\ &+ \left[\Psi\left(S_*\right) - \left(C\left(S,S_*|\theta\right)\right)^\top \left(C\left(S,S|\theta\right) + I\sigma^2\right)^{-1}\Psi\left(S_*\right)\right] \\ &\times B_n\left(\theta,\sigma^2\right)\left[\Psi\left(S_*\right) - \left(C\left(S,S_*|\theta\right)\right)^\top \left(C\left(S,S|\theta\right) + I\sigma^2\right)^{-1}\Psi\left(S_*\right)\right]^\top\end{aligned}$$

(17.20)

$$\mu_2\left(S_*|\theta,\sigma^2\right) = \Psi\left(S_*\right)b_n\left(\theta,\sigma^2\right) - \left(C\left(S,S_*|\theta\right)\right)^\top \left(C\left(S,S|\theta\right) + I\sigma^2\right)^{-1}\left(\Psi\left(S\right)b_n\left(\theta,\sigma^2\right) - Z\right)$$

Note 204. Since derivation of (17.18) holds for all vectors $S_* \in \mathbb{R}^q$ and all $q > 0$, (17.18) can be extended to a Gaussian Process

$$Y\left(\cdot\right)|Z,\theta,\sigma^2 \sim \mathrm{GP}\left(\mu_2\left(\cdot|\theta,\sigma^2\right),c_2\left(\cdot,\cdot|\theta,\sigma^2\right)\right) \tag{17.21}$$

(17.22)

$$\mu_2\left(s|\theta,\sigma^2\right) = \psi\left(s\right)b_n\left(\theta,\sigma^2\right) - \left(C\left(s|\theta\right)\right)^\top \left(C\left(\theta\right) + I\sigma^2\right)^{-1}\left(\Psi b_n\left(\theta,\sigma^2\right) - Z\right)$$

(17.23)

$$\begin{aligned}c_2\left(s,s'|\theta,\sigma^2\right) =& c_1\left(s,s'|\theta,\sigma^2\right) \\ &+ \left[\psi\left(s\right) - \left(C\left(s|\theta\right)\right)^\top \left(C\left(\theta\right) + I\sigma^2\right)^{-1}\Psi\right]B_n\left(\theta,\sigma^2\right)\left[\psi\left(s\right) - \left(C\left(s|\theta\right)\right)^\top \left(C\left(\theta\right) + I\sigma^2\right)^{-1}\Psi\right]\end{aligned}$$

with column vector $C\left(s|\theta\right) = \left(c\left(s,s_1\right),...,c\left(s,s_n\right)\right)^\top$, and matrix $C\left(\theta\right) := C\left(S,S|\theta\right)$.

Note 205. Estimates $\hat{\theta}$ and $\hat{\sigma}^2$ of the unknown fixed hyper-parameters θ and σ^2 are computed by maximizing the marginal likelihood of Z given θ and σ^2

$$\text{pr}\left(Z|\theta,\sigma^2\right) = \int \text{pr}\left(Z|\beta,\theta,\sigma^2\right)\text{pr}\left(\beta\right)\mathrm{d}\beta \tag{17.24}$$

$$= \int \text{N}\left(Z|\Psi\beta, C\left(\theta\right)+I\sigma^2\right)\text{N}\left(\beta|b,B\right)\mathrm{d}\beta \tag{17.25}$$

$$= \text{N}\left(Z|\Psi b, C\left(\theta\right)+I\sigma^2+\Psi B\Psi^\top\right) \tag{17.26}$$

where $C\left(\theta\right) := C\left(S,S|\theta\right)$ [from Fact 190] by computing

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

The predictive distribution of $Z\left(\cdot\right)|Z$ after parameterization.

Note 206. It is a Gaussian Process

$$Z\left(\cdot\right)|Z,\theta,\sigma^2 \sim \text{GP}\left(\mu_2\left(\cdot|\theta,\sigma^2\right), c_2^z\left(\cdot,\cdot|\theta,\sigma^2\right)\right) \tag{17.27}$$

with $c_2^z\left(s,s'|\theta,\sigma^2\right) = c_2\left(s,s'|\theta,\sigma^2\right)+\sigma^2 1_{\{0\}}\left(|s-s'|\right)$. This is because $Z\left(\cdot\right) = Y\left(\cdot\right)+\varepsilon\left(\cdot\right)$ with $Y\left(\cdot\right)\perp\varepsilon\left(\cdot\right)$.

Note 207. The estimated "Kriging predictor" results by plugging $\hat{\theta}$ and $\hat{\sigma}^2$ in (17.21)

$$Y\left(\cdot\right)|Z,\hat{\theta},\hat{\sigma}^2 \sim \text{GP}\left(\mu_2\left(\cdot|\hat{\theta},\hat{\sigma}^2\right), c_2\left(\cdot,\cdot|\hat{\theta},\hat{\sigma}^2\right)\right) \tag{17.28}$$

Part 3. Spatial misalignment (special topic)

18. Regularization (an intro)

Note 208. Let $\left(Z\left(s\right): s\in\mathcal{S}\right)$ be a random field with mean $\mu\left(s\right)$ at $s\in\mathcal{S}$, covariance function $c\left(s,s'\right)$ at $s,s'\in\mathcal{S}$, and semivariogram (if exists) $\gamma\left(h\right)$.

Definition 209. Let $\varphi\left(t\right)$ be an integrable function of t such as $\int\left|\varphi\left(t\right)\right|\mathrm{d}t<\infty$. The regularized random field $\left(Z_\varphi\left(s\right): s\in\mathcal{S}\right)$ is defined as

$$Z_\varphi\left(s\right) = \int Z\left(s+t\right)\varphi\left(t\right)\mathrm{d}t \tag{18.1}$$

Example 210. The mean, covariance function, and semivariogram $\left(Z_\varphi\left(s\right): s\in\mathcal{S}\right)$ in (18.1) is

$$\mu_\varphi\left(s\right) = \text{E}\left(Z_\varphi\left(s\right)\right) = \int \mu\left(s+t\right)\varphi\left(t\right)\mathrm{d}t$$

$$c_\varphi\left(s,s'\right) = \text{Cov}\left(Z_\varphi\left(s\right), Z_\varphi\left(s'\right)\right) = \int\int c\left(s+t,s'+t'\right)\varphi\left(t\right)\varphi\left(t'\right)\mathrm{d}t\mathrm{d}t'$$

Proof. See Exercise 21 in the Exercise sheet. □

Note 211. Assume the random field $(Z(s) : s \in \mathcal{S})$ is stationary and has spectral measure $\mathrm{d}F(\omega)$. Then regularized random field $(Z_\varphi(s) : s \in \mathcal{S})$ defined as in (18.1) has spectral measure

$$\mathrm{d}F_\varphi(\omega) = |\tilde{\varphi}(\omega)|^2 \mathrm{d}F(\omega)$$

where $\tilde{\varphi}(\omega) = \int_t e^{it^\top \omega} \varphi(t) \mathrm{d}t$. This is because

$$\begin{aligned} c_\varphi(s, s-h) &= \int_t \int_{t'} c(s+t, s-h+t') \varphi(t) \varphi(t') \mathrm{d}t\mathrm{d}t' \\ &= \int_t \int_{t'} c(t+h-t') \varphi(t) \varphi(t') \mathrm{d}t\mathrm{d}t' \\ \text{(Bohner's theorem)} &= \int_t \int_{t'} \int_h e^{i(t+h-t')^\top \omega} \mathrm{d}F(\omega) \varphi(t) \varphi(t') \mathrm{d}t\mathrm{d}t' \\ &= \int_h e^{ih^\top \omega} \left(\int_t e^{it^\top \omega} \varphi(t) \mathrm{d}t \right) \left(\int_t e^{it'^\top \omega} \varphi(t') \mathrm{d}t' \right) \mathrm{d}F(\omega) \end{aligned}$$

Note 212. Regularization is often motivated by applications such as

Prediction: In mining, one wants to predict not a single value of the process at a new site, but instead the total ore content in a block of rock.

Data collection: One cannot observe a random field at an individual site t, but can observe only the average value of the process in a small region about t. E.g., in mining, the smallest practical measurement that can be made is the average mineral content over a section of a borehole.

Image: there is often some smearing between neighboring pixels of an image.

Note 213. Here the concept is presented on simple applications of spatial misalignment and change of support

$$\varphi(t) = \frac{1}{|B|} 1(t \in B), \quad B \subseteq \mathcal{S}.$$

19. Intro to spatial misalignment

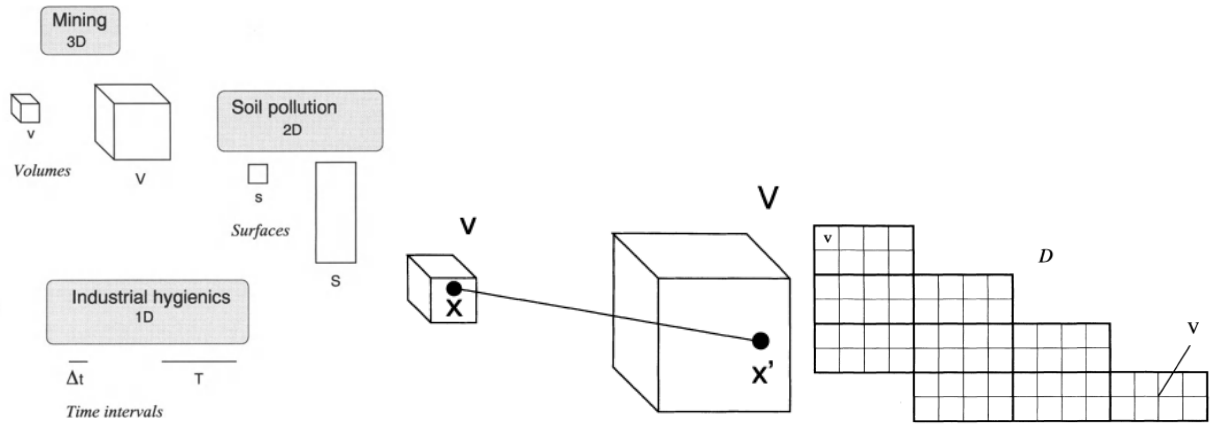
Note 214. Consider a stochastic process $(Z(s) : s \in \mathcal{S})$ where $\mathcal{S} \in \mathbb{R}^d$, $d \in \mathbb{N}_{>0}$, with $\mathrm{Var}(s) < \infty$ for all $s \in \mathcal{S}$.

Definition 215. We define the block average $Z(B)$ with $B \subseteq \mathcal{S}$ as

$$Z(B) = \begin{cases} \frac{1}{|B|} \int_B Z(s) \mathrm{d}x & |B| > 0 \\ \text{average} \{Z(s) : s \in B\} & |B| = 0 \end{cases} \tag{19.1}$$

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

Definition 216. The integral in (19.1) can be defined by Riemann sums. E.g. in 2D if $B = [a_1, a_2] \times [b_1, b_2]$, $a_1 < u_0 < ... < u_n < a_2$, $b_1 < v_0 < ... < v_n < b_2$, $u'_j \in [u_{j-1}, u_j]$, and



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

FIGURE 19.1. Change of support

$v_i' \in [v_{i-1}, v_i]$, then

$$\int_B Z(s)\,\mathrm{d}s = \lim_{n\to\infty, m\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{m} (v_i - v_{i-1})(u_j - u_{j-1})\, Z\left(v_i', u_j'\right) \tag{19.2}$$

Note 217. Notice that the integral in (19.1) is a linear operator, hence for $A, B \subseteq \mathcal{S}$ it is

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

Note 218. A common problem is to predict the block average $Z(B)$ of a process $(Z(s) : s \in \mathcal{S})$ over a block B whose location and geometry are known and whose d-dimensional volume is $|B|$. (See Figure 21.1)

Definition 219. The support of the block average $Z(B)$ in (19.1) is B and involves the geometry, size, and spatial orientation of the line, area, or volume of the input.

Change of support problem.

Note 220. Changing the support of a variable creates a new variable related to the original one but with different statistical characteristics: mean, co-variance, dependencies, etc...

Definition 221. Change of support problem refers to making inference on block of averages whose supports are different than those of the data. $Z = (Z(B_1), ..., Z(B_n))^\top$.

20. Extension and Dispersion Variance

Note 222. With spatial variables it is necessary to take account the spatial disposal of points, surfaces or volumes for which the variance of a quantity should be computed.

Definition. Extension variance $\sigma_E^2(v, V)$ of a small volume v to a larger volume V is defined by

$$\sigma_E^2(v, V) := \mathrm{Var}(Z(v) - Z(V))$$

Definition 223. The dispersion variance of of the small identical volumes v_j partitioning a larger volume V is

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

Notation 224. Let v be a small volume v and let V be a larger volume. Then we denote a semivariogram integral

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

Proposition 225. *Let $(Z(s) : s \in \mathcal{S})$ be an intrinsic random field with semivariogram $\gamma(\cdot)$.*

(1) The extension variance $\sigma_E^2(v, V)$ of a small volume v to a larger volume V is

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

(2) Suppose a large volume V is partitioned into n smaller units $\{v_j\}_{j=1}^n$ of equal size and geometry. The dispersion variance of each unit v_j is

$$\sigma^2(v|V) = \bar{\gamma}(V, V) - \bar{\gamma}(v, v) \tag{20.1}$$

Proof. For part 1, see Exercise 22 in the Exercise sheet. For part 2, see Exercise 24 in the Exercise sheet. □

Note 226. The following result (Krige's relationship) resembles to "ANOVA".

Proposition 227. *[Krige's relationship] Consider that the domain S is partitioned into volumes V which are partitioned into smaller volumes v; i.e. $v \subseteq V \subseteq S$. Then the relation between the three supports is*

$$\sigma^2(v|\mathcal{S}) = \sigma^2(v|V) + \sigma^2(V|\mathcal{S}) \tag{20.2}$$

Proof. (Sketch of the proof) (20.1) becomes

$$\sigma^2(v|\mathcal{S}) = \bar{\gamma}(\mathcal{S}, \mathcal{S}) - \bar{\gamma}(v, v)$$

similar

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

so (20.1) gives

$$\sigma^2(v|\mathcal{S}) - \sigma^2(v|V) = \sigma^2(V|\mathcal{S}) \tag{20.3}$$

□

Change of support effect.

Note 228. Consider the case that the domain S is partitioned into volumes V which are partitioned into smaller volumes v. Assume there are available samples at "point" locations s each of them lies to the center of one of the smaller volumes v. Making the assumption that the sampled value at each point location s is extended to each area of influence v implies that the distribution of average values of the blocks is the same as the distribution of the values at the sample points. However from (20.3), we see that this is not true; in fact the distribution of the values for a support v is narrower than the distribution of point values because the variance $\sigma^2(s|v)$ of the points in v generally is not negligible; i.e. $\sigma^2(s|V) - \sigma^2(v|V) = \sigma^2(s|v) > 0$.

Change of support: affine model.

Note 229. Consider a stationary process $Z(s)$ for $s \in \mathcal{S}$, and consider a block process $Z_v(s)$ on a block v. The affine model assumes that the standardized point variable $Z(s)$ follows the same distribution as the standardized block variable $Z(v)$.

Example 230. An example of the use of affine models is the Gaussian process case, where $Z(s) \sim \mathrm{N}(\mu, \sigma^2)$ and $Z(v) \sim \mathrm{N}(\mu, \sigma_v^2)$, –same mean but different variances– it is

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

which implies the relation

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

21. Block Kriging

Note 231. Block Kriging (BK) aims to predict a block value $Z(v_0)$ at block v_0 instead of at a point value s_0; see Figure 21.1. It can be used within the framework of Universal, Ordinary, Simple, and Bayesian Kriging cases we saw in Section 16.1.

Note 232. Assume we want the estimate a block value $Z(v_0)$ at block v_0 with some volume $|v_0|$ given that my data $\{(s_i, Z_i)\}_{i=1}^n$ are realizations $Z_i = Z(s_i)$ at point values s_i (Figure 21.1).

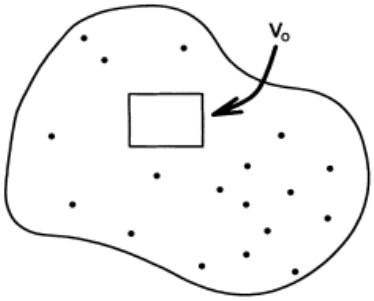


FIGURE 21.1. Block Kriging cartoon

21.1. **(Classical) Universal Block Kriging.**

Note 233. Here, we present the Block Kriging (BK) in the (Classical) Universal Kriging framework (Section 16.1). We will refer to the UK in Section 16.1 as point-to-point UK.

Note 234. Consider that the statistical model is the stochastic process $(Z(s) : s \in \mathcal{S})$ with

$$Z(s) = \mu(s) + \delta(s) \tag{21.1}$$

Assume

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

with vector of unknown coefficients $\beta = (\beta_0, ..., \beta_p)^\top$ and vector of known basis functions $\psi(s) = (\psi_0(s), ..., \psi_p(s))^\top$. Assume $\delta(s)$ is a zero mean process. Assume $\delta(s)$ is an intrinsic stationary process with a semi-variogram $\gamma(\cdot)$ –as in UK in Section 16.1; intrinsic stationarity is not a necessary assumption if one can estimate the covariance function directly.

Note 235. The Block UK predictor $Z_{\text{BK}}(v_0)$ of $Z(v_0)$ at block v_0 with support $|v_0| > 0$ has the following linear form weighted by a set of tunable unknown weights

$$Z^*_{\text{BK}}(v_0) = w_{n+1} + \sum_{i=1}^{n} w_i Z(s_i) = w_{n+1} + w^\top Z \tag{21.2}$$

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

Note 236. Following the steps in (point-to-point) UK (Note 155), consideration of $\psi_0(\cdot)=1$ unbiasness implies conditions

$$\text{ASSUMPTION:} \quad \Psi_0=\sum_{i=1}^{n} w_i\psi(s_i) \Leftrightarrow \Psi_0=w^\top\Psi \tag{21.3}$$

$$\text{ASSUMPTION:} \quad w_{n+1}=0 \tag{21.4}$$

$$\text{ASSUMPTION:} \quad \sum_{i=1}^{n} w_i=1 \Leftrightarrow w^\top\underline{1}=1 \tag{21.5}$$

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

Note 237. Following the steps in (point-to-point) UK (Note 159), I get

$$\begin{aligned}\mathrm{MSE}(Z_{\mathrm{BK}}(v_0))-\sum_{i=1}^{n} w_i\sum_{j=1}^{n} w_j\gamma(s_i-s_j)+2\sum_{i=1}^{n} w_i\bar{\gamma}(s_i,v_0) \\ =-w^\top\mathbf{\Gamma}w+2w^\top\bar{\boldsymbol{\gamma}}_0-\bar{\gamma}_{00}\end{aligned} \tag{21.6}$$

where $\bar{\gamma}_{00}=\bar{\gamma}(v_0,v_0)\neq 0$, $\bar{\boldsymbol{\gamma}}_0=(\bar{\gamma}(s_1,v_0),...,\bar{\gamma}(s_n,v_0))^\top$, and $\bar{\gamma}(s_i,v_0)$ be the average variogram of each sample point with the block of interest. This is the same as that of point-to-point UK in (16.8) where the point $\gamma(s_i,s_0)$ is substituted by the integral $\bar{\gamma}(s_i,v_0)$.

Note 238. The Block Universal Kriging equations then are

$$\begin{cases}0=-2\mathbf{\Gamma}w_{\mathrm{BK}}+2\bar{\boldsymbol{\gamma}}_0-\Psi\lambda_{\mathrm{UK}}\\ \Psi_0=w_{\mathrm{BK}}^\top\Psi\end{cases} \tag{21.7}$$

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

$$w_{\mathrm{BK}}=\mathbf{\Gamma}^{-1}\left(\bar{\boldsymbol{\gamma}}_0-\Psi\left(\Psi^\top\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^\top\Gamma^{-1}\bar{\boldsymbol{\gamma}}_0-\Psi_0^\top\right)\right) \tag{21.8}$$

$$\lambda_{\mathrm{BK}}=2\left(\Psi^\top\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^\top\Gamma^{-1}\boldsymbol{\gamma}_0-\Psi_0^\top\right) \tag{21.9}$$

Note 239. Hence the UK predictor $Z_{\mathrm{BK}}(s_0)$ at s_0 is

$$Z_{\mathrm{BK}}(s_0)=\left(\bar{\boldsymbol{\gamma}}_0-\Psi\left(\Psi^\top\Gamma^{-1}\Psi\right)^{-1}\left(\Psi^\top\Gamma^{-1}\bar{\boldsymbol{\gamma}}_0-\Psi_0^\top\right)\right)^\top\mathbf{\Gamma}^{-1}Z \tag{21.10}$$

with standard error (by substituting (21.7) in(21.6))

$$\sigma_{\mathrm{BK}}(v_0)=\sqrt{-\bar{\gamma}_{00}-w_{\mathrm{BK}}^\top\mathbf{\Gamma}w_{\mathrm{BK}}+2w_{\mathrm{BK}}^\top\bar{\boldsymbol{\gamma}}_0} \tag{21.11}$$

Note 240. Block Kriging as a concept can be implemented even when s_i are not points but have some volume $|s_i|>0$. Then we call the case as aggregation if $|s_i|<|v_0|$, or disaggregation if $|s_i|>|v_0|$.

21.2. **Bayesian Block Kriging.**

Note 241. If $Z(s)$ is a Gaussian process defined on points $s \in \mathcal{S}$, then the block average $Z(v)$ with $v \subseteq \mathcal{S}$ is a Gaussian process as well. This is because integration (or averaging) in (19.1) is a linear operation as seen in (19.2), and linear combinations of Gaussians is Gaussian as well.

Note 242. Bayesian Block Kriging predictive distribution and moments are derived in the same way as the "point to point" Bayesian Kriging in Section 17. Let $(Z(s) : s \in \mathcal{S})$ be a GRF, and$\{(Z(s_i), s_i); s_i \in \mathcal{S}\}$ be observables. I want to compute the Bayesian Block Kriging predictive distribution of $Z(v_0)$ at unseen volume/unit $v_0 \subseteq \mathcal{S}$. Main steps involve.

(1) compute the joint distribution in (17.10) i.e.

$$\begin{pmatrix} Z(v_0) \\ Z(\{s_i\}) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(v_0) \\ \mu(\{s_i\}) \end{pmatrix}, \begin{pmatrix} c(v_0, v_0) & c(v_0, \{s_i\}) \\ (c(v_0, \{s_i\}))^\top & c(\{s_i\}, \{s_i\}) \end{pmatrix}\right)$$

with

$$\begin{aligned} \mu(v_0) &= \frac{1}{|v_0|}\int_{x \in v_0} \mu(x)\,\mathrm{d}x \\ c(v_0, s_i) &= \frac{1}{|v_0|}\int_{x \in v_0} c(x, s_i)\,\mathrm{d}x \\ c(v_0, v_0') &= \frac{1}{|v_0|\,|v_0'|}\int_{x \in v_0}\int_{y \in v_0'} c(x, y)\,\mathrm{d}x\mathrm{d}y \end{aligned}$$

(2) compute the the predictive distribution as the conditional Normal distribution $\mathrm{pr}(Z(v_0)|Z)$ (Note 196), and

(3) recognize the corresponding Gaussian process as in Note 198.

The derivation is identical to that Section 17.2.

Part 4. Extensions to multivariate Geostatistics (special topic)

22. Extensions to Multivariate Geostatistics

Note 243. So far we have limited our attention to a single real-valued measurement $Z(s) \in \mathcal{Z} \subseteq \mathbb{R}$ at each site $s \in \mathcal{S} \subseteq \mathbb{R}^d$. A natural extension is to allow a vector of measurements $Z(s) \in \mathcal{Z} \subseteq \mathbb{R}^k$, $k \geq 1$, with elements $Z_i(s)$, $i = 1, ..., k$ at each site s.

22.1. **Cross-variance functions.**

Definition 244. Let $Z_1(s), ..., Z_k(s)$ be k random fields on $s \in \mathcal{S}$. The cross-covariance function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

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

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

Definition 245. Let $Z_1(s),...,Z_k(s)$ be k (weakly) stationary random fields on $s\in\mathcal{S}$. The under stationarity cross-covariance function (or cross-covariogram function) of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

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

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

Definition 246. Let $Z_1(s),...,Z_k(s)$ be k (weakly) stationary random fields on $s\in\mathcal{S}$. The matrix valued function

$$C(h)=\mathrm{Cov}(Z(s),Z(s+h))$$

with $[C(h)]_{i,j}=C_{i,j}(h)$ is the covariance function matrix (or under stationarity cross-covariance function) of $Z(s)=(Z_1(s),...,Z_k(s))^\top$.

Example 247. Cross-covariograms have the following properties

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

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

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

Definition 248. Let $Z_1(s),...,Z_k(s)$ be k intrinsic random fields on $s\in\mathcal{S}$. The cross-variogram function of $Z_i(\cdot)$ and $Z_j(\cdot)$ is defined as

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

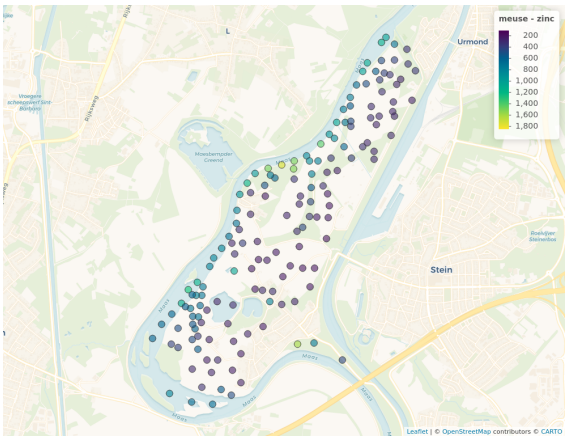
for $i,j=1,...,k$ and $s,s+h\in\mathcal{S}$. In matrix form

$$\Gamma(h)=\frac{1}{2}\mathrm{Cov}(Z(s+h)-Z(s),Z(s+h)-Z(s))$$

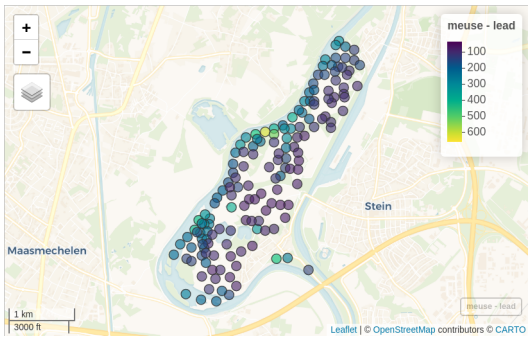
with $[\Gamma(h)]_{i,j}=\gamma_{i,j}(h)$.

22.2. **Co-Kriging.**

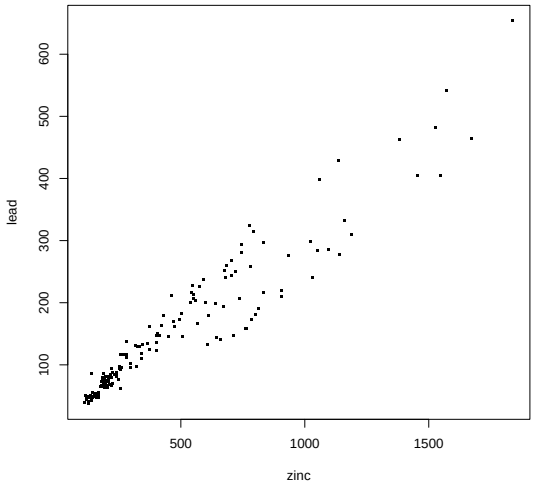
Note 249. CoKriging procedure is a natural extension of kriging when the cross-covariance function is available. A quantity of interest (QoI) (response variable) is coKriged at a specific location from data about itself and/or about auxiliary variables in the neighborhood. The purpose of coKriging is to "borrow strength" from the measurements on the auxiliary variables to improve the accuracy when predicting the QoI.



(A) zinc concentration (obs)



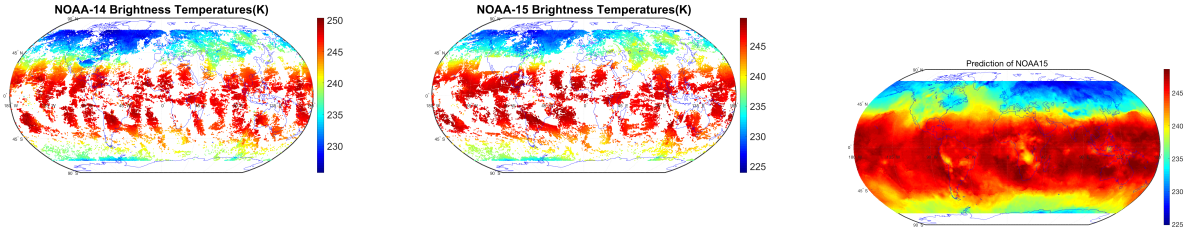
(B) lead concentration (obs)



(C) scatter plot (obs)

FIGURE 22.1. Map of the meuse dataset

Example 250. Different variables correspond to a different characteristics measures at same/different locations and associated with same/different covariates. In the meuse{sp} R dataset, the QoI $\left\{\left(Z_i^{\text{zinc}}, Z_i^{\text{lead}}, Z_i^{\text{cooper}}, Z_i^{\text{carmium}}\right)\right\}$ are the concentrations of zinc, lead, copper, and cadmium, at these locations. Interest lies on the prediction of QoI at unobserved locations (i.e. interpolation), computation of the joint distribution of the QoI (evaluate the distribution of a random function $Z(s) = \left(Z^{\text{zinc}}(s), Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s)\right)$, for all $s \in S$), and how each of QoI depends each other along the flood plain of the river Meuse is of interest. Alternatively, one may be interested in the prediction of of zinc $Z^{\text{zinc}}(s)$ at any point s given all others $\left\{Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s)\right\}$ in a combined way; i.e. $Z^{\text{zinc}}(s) | Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s)$. See Figure 22.1.



(A) N14 (less accurate)

(B) N15 (more accurate)

(C) N15 (co-kriging prediction)

FIGURE 22.2. Satellite temperature readings data

Example 251. Different variables correspond to a different accuracy level or support.

- The QoI could be the precipitation over a spatial domain, and the auxiliary variables could be: $\{Z_1(s_i)\}_{i=1}^{n_1}$ the precipitation measured by a weather station, and $\{Z_2(s_i)\}_{i=1}^{n_2}$ the precipitation measured by a satellite. Here the weather station measurements are much more accurate than those from the satellite however they are taken at a smaller number of locations $n_1 \ll n_2$.
- The QoI could be the temperature over a spatial domain; the available data may be the temperature readings by an old technology (less accurate) satellite $\{Z_2(s_i)\}_{i=1}^{n_1}$, and the temperature readings by a new technology (more accurate) satellite $\{Z_1(s_i)\}_{i=1}^{n_1}$.

  Interest lies in the predictive process $Z_1(\cdot) \mid \{Z_{1,i}\}, \{Z_{2,i}\}$ of $Z_1(s)$ at any point s given all the available data $\{Z_1(s_i)\}_{i=1}^{n_1}$ and $\{Z_2(s_i)\}_{i=1}^{n_1}$ (i.e. given the combined data). See Figure 22.2.

22.3. **Classical coKriking.**

Note 252. We present the concept in the ordinary Kriging framework.

Note 253. Consider k stochastic processes $Z_1(s), ..., Z_k(s)$, $s \in \mathcal{S}$. Consider data at n sites $\{s_i\}_{i=1}^{n}$. Let $\boldsymbol{Z}(\boldsymbol{s})$ be a $n \times k$ matrix $\boldsymbol{Z}(\boldsymbol{s}) = Z_j(s_i)$ for $i = 1, ..., n_j$, and $j = 1, ..., k$. It is desired to predict the j_0-th variable $Z_{j_0}(s_0)$ for some $j_0 \in \{1, ..., k\}$ at location $s_0 \in \mathcal{S}$.

Note 254. Assume

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

Note 255. Co-Kriging predictor $Z_{\mathrm{CK},j_0}(s_0)$ is the BLUE predictor $Z_{\mathrm{CK},j_0}(s_0)$ of $Z_{j_0}(\cdot)$ at s_0.

Note 256. The Co-Kriging predictor has the linear form

$$Z_{\mathrm{CK},j_0}(s_0) = w_{0,0} + \sum_{j=1}^{k}\sum_{i=1}^{n_j} w_{j,i} Z_j(s_i) = w_{0,0} + \sum_{j=1}^{k} w_j^\top Z_j \tag{22.1}$$

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

Note 257. Parametrization (22.1) requires that all $Z_j(\cdot)$ components are observed at each site s_i. However the concept of co-kriging can also be adjusted to consider more general cases such as those where different processes $Z_j(\cdot)$ are observed at different sets of sites from each other.

Note 258. To enforce unbiassness, we find sufficient conditions for $\{w_{j,i}\}$

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

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

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

Notation 259. Set convenient notation for the calculations below as

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

Note 260. The MSE (or Variance) is

$$\begin{aligned}
\operatorname{MSE}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)\right) =& \mathrm{E}\left(Z_{\mathrm{CK},j_0}\left(s_0\right)-Z_{j_0}\left(s_0\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)-Z_{j_0}\left(s_0\right)\right)^2 \\
=& \mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}Z_j\left(s_i\right)-Z_{j_0}\left(s_0\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=0}^{n_j}w_{j,i}Z_j\left(s_i\right)\right)^2 \\
=& \mathrm{E}\left(\sum_{j=1}^{k}\left(\sum_{i=0}^{n_j}w_{j,i}Z_j\left(s_i\right)-\sum_{i=1}^{n_j}w_{j,i}\mu_j\right)\right)^2=\mathrm{E}\left(\sum_{j=1}^{k}\sum_{i=0}^{n_j}\left(Z_j\left(s_i\right)-\mu_j\right)\right)^2 \\
=& \sum_{j=1}^{k}\sum_{i=0}^{n_j}\sum_{j'=1}^{k}\sum_{i'=0}^{n_{j'}}w_{j,i}w_{j',i'}\mathrm{E}\left(Z_j\left(s_i\right)-\mu_j\right)\left(Z_{j'}\left(s_{i'}\right)-\mu_{j'}\right) \\
=& \sum_{j=1}^{k}\sum_{i=0}^{n_j}\sum_{j'=1}^{k}\sum_{i'=0}^{n_{j'}}w_{j,i}w_{j',i'}C_{j,j'}\left(s_i,s_{i'}\right) \\
=& \sum_{j=1}^{k}\sum_{i=1}^{n_j}\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j,i}w_{j',i'}C_{j,j'}\left(s_i,s_{i'}\right) \\
& -2\sum_{j'=1}^{k}\sum_{i'=1}^{n_{j'}}w_{j',i'}C_{j_0,j'}\left(s_0,s_{i'}\right)+C_{j_0,j_0}\left(s_0,s_0\right) \qquad (22.2)
\end{aligned}$$

Note 261. The Lagrange function is

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

Note 262. The CK system of equations produced by $0=\nabla_{(w,\lambda)}\mathfrak{L}\left(w,\lambda\right)\big|_{(w_{\mathrm{CK}},\lambda_{\mathrm{CK}})}$ is

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n_j}w_{j,i}C_{j,j'}\left(s_i,s_{i'}\right)-\lambda_{j'}=C_{j_0,j'}\left(s_0,s_{i'}\right),\quad \forall j',i' \qquad (22.3)\\
\sum_{i=1}^{n_{j_0}}w_{j_0,i}=1,\quad \sum_{i=1}^{n_{j'}}w_{j',i}=0,\ \forall j'
\end{aligned}$$

Note 263. Plugin (22.3) in (22.2), I can get the co-Kriging variance

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

Note 264. The above derivation can be done wrt the cross-variogram as in UK, OK (by making extra assumptions). I choose to presented wrt the cross-covariance as more general.

22.4. **Bayesian coKriking.**

Note 265. Regarding the Bayesian framework. Consider the paradigm that $Z_j(\cdot)$ are GP, where $\mu_j(\cdot) = \mathrm{E}(Z_j(\cdot))$ and $c_{j,j'}(\cdot) = \mathrm{Cov}(Z_j(\cdot), Z_{j'}(\cdot))$. Let set of sites $S_j = \{s_{j,1}, ..., s_{j,n_j}\}$ and assume there is an available dataset $\{(Z_{j,i}, s_{j,i})\}_{i=1}^{n_j}$ for $j = 1, ..., k$. The procedure is the same as discussed in Section 17, with the only difference that the predictive Gaussian process will be $Z_{j_0}(\cdot) \,|\, \{Z_{1,i}\}, ..., \{Z_{k,i}\}$, for $j_0 \in \{1, ..., k\}$ and resulted after computing the joint distribution

(22.4)

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

and conditioning as

$$Z_{j_0}(S_*) \,|Z_1, ..., Z_k \sim \mathrm{N}\left(\mu_{j_0|1,...,k}(S_*), C_{j_0|1,...,k}(S_*, S_*)\right) \tag{22.5}$$

(22.6)

$$\mu_{j_0|1,...,k}(S_*) = \mu_{j_0}(S_*) - C_{j_0} C^{-1}[\mu - Z], \; C_{j_0|1,...,k}(S_*, S_*) = C_{j_0,j_0}(S_*, S_*) - C_{j_0} C^{-1} C_{j_0}^{\top}$$

Note 266. If k is large with moderate large n_j for each (or some) j's, the calculations in (22.3) and (22.4) can be too computationally challenging and have unrealistic computational requirements for a standard PC. E.g., we will have to solve a huge system of equations in (22.3), while we will have to do operations with a huge covariance matrix in (22.4). In 90's your computer (particularly its CPU and its RAM) would complain with a blue screen...

22.5. **Intrinsic correlation model.**

Note 267. Tricks to mitigate challenges with large k and n_j involve imposing restrictions on the cross-covariance functions $C_{i,j}(s,t)$ having special structure often by introducing conditional independences (hence restricting the model), as well as using suitable experimental designs.

Definition 268. Intrinsic Multivariate Correlation model is the model which describes relations between variables by the covariance matrix $B > 0$ and the relations between points in space by a spatial correlation $\varrho(h)$ which is the same for all variables; i.e.

$$C(h) = B\varrho(h)$$

Definition 269. Intrinsic Multivariate Correlation model is the model which describes relations between variables by the covariance matrix $B > 0$ and the relations between points in space by a spatial correlation $\varrho(h)$ which is the same for all variables; i.e.

$$C(h) = B\varrho(h)$$

Example 270. Consider a set of processes $\{Z_j(\cdot)\}$ where the cross-covariance is modeled as $C_{i,j}(s,t) = \sigma_{i,j}\varrho(|s-t|)$ where $\sigma_{i,j}$ is the (i,j) element of a semi-positive matrix Σ and $\varrho(h) = c(h)/c(0)$ is the correlogram of some isotropic covariogram function $c(\cdot)$. Show that this co-kriging model is Intrinsic Multivariate Correlation model.

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

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

Note 271. In Co-Kriging, using the Intrinsic Multivariate Correlation model for the cross-covariance, and using sites in a grid allows the use of Kronecker product operations in cases (22.3) and 22.4 for mitigating the computational requirements.

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

Note 272. A simple way to construct a multiple covariance function is through linear combinations of independent random fields.

Note 273. Let $\{(Z_j(s) : s \in \mathcal{S}); j = 1, ..., k\}$ be a set of random fields on $s \in \mathcal{S}$.

Note 274. Assume each $Z_j(s)$ can be decomposed into sets $\left\{\left(Z_j^{(u)}(s) : s \in \mathcal{S}\right); u = 0, ..., m\right\}$ of spatially uncorrelated components as

$$Z_j(s) = \mu_j(s) + \sum_{u=0}^{m} Z_j^{(u)}(s) \tag{22.7}$$

which are zero mean stationary random fields such as

$$\mathrm{E}\left(Z_j^{(u)}(s)\right) = 0$$

$$\mathrm{Cov}\left(Z_i^{(u)}(s), Z_j^{(v)}(s+h)\right) = \begin{cases} C_{i,j}^{(u)}(h), & u = v \\ 0 & u \neq v \end{cases} \tag{22.8}$$

for all $i, j = 1, ..., k$.

Note 275. For the cross-covariance functions (22.8) of the spatial components, assume intrinsic correlation model

$$C_{i,j}^{(u)}(h) = b_{i,j}^{(u)}\varrho^{(u)}(h) \tag{22.9}$$

imposing a covariance function matrix

$$C^{(u)}(h) = B^{(u)}\varrho^{(u)}(h) \tag{22.10}$$

where $\varrho^{(u)}(h)$ are correlation functions for all $u = 0, ..., k$, and $B^{(u)}$ are coregionalization matrices with order $n \times n$, $\left[B^{(u)}\right]_{i,j} = b_{i,j}^{(u)}$ and $B^{(u)} > 0$ which can be set-up.

Note 276. The intrinsic correlation model assumption (22.9) allows each spatial component $Z_j^{(u)}(s)$ to be represented as

$$Z_j^{(u)}(s) = \sum_{p=1}^{k} a_{j,p}^{(u)} w_p^{(u)}(s)$$

where $\left\{w_p^{(u)}(s)\right\}$ are random fields (independent, zero-mean, variance one) with

$$\mathrm{E}\left(w_p^{(u)}(s)\right) = 0$$

$$\mathrm{Cov}\left(w_p^{(u)}(s), w_q^{(v)}(s+h)\right) = \begin{cases} \varrho^{(u)}(h), & u = v \text{ and } p = q \\ 0 & u \neq v \text{ or } p \neq q \end{cases}$$

Note 277. In matrix form,

$$Z^{(u)}(s) = A^{(u)} w^{(u)}(s)$$

where $\left[A^{(u)}\right]_{j,p} = a_{j,p}^{(u)}$ and $w^{(u)}(s) = \left(w_1^{(u)}(s), ..., w_k^{(u)}(s)\right)^\top$ for $u = 0, ..., m$.

Note 278. From (22.10) , we observe that $B^{(u)} = A^{(u)}\left(A^{(u)}\right)^\top$ and hence $\left\{a_{j,p}^{(u)}\right\}$ satisfy $B^{(u)} > 0$, or equivalently

$$\left|b_{i,j}^{(u)}\right| \leq \sqrt{b_{i,i}^{(u)} b_{j,j}^{(u)}},\ \forall i, j, u$$

Note 279. To sum up (22.7) becomes

$$Z_j(s) = \mu_j(s) + \sum_{u=0}^{m}\sum_{p=1}^{k} a_{j,p}^{(u)} w_p^{(u)}(s)$$

for $j = 1, ..., k$. In matrix form, it is

$$Z(s) = \mu(s) + \sum_{u=0}^{m} A^{(u)} w^{(u)}(s)$$

Note 280. To check some properties of the LMC, consider $m = 0$, i.e. $Z(s) = \mu(s) + A^{(0)}W^{(0)}(s)$, and $A^{(0)}$ as a lower triangular matrix. It is

$$\begin{aligned}
Z_1(s) &= \mu(s) + a_{1\ ,1}^{(0)} w_1^{(0)}(s) \\
&\vdots \\
Z_{j-1}(s) &= \mu(s) + a_{j-1,1}^{(0)} w_1^{(0)}(s) + \ldots + a_{j-1,j-1}^{(0)} w_{j-1}^{(0)}(s) \\
Z_j(s) &= \mu(s) + a_{j\ ,1}^{(0)} w_1^{(0)}(s) + \ldots + a_{j\ ,j-1}^{(0)} w_{j-1}^{(0)}(s) + a_{j,j}^{(0)} w_j^{(0)}(s) \\
&\vdots \\
Z_k(s) &= \mu(s) + a_{k\ ,1}^{(0)} w_1^{(0)}(s) + \ldots + a_{k\ ,j-1}^{(0)} w_{j-1}^{(0)}(s) + a_{k,j}^{(0)} w_j^{(0)}(s) + \ldots + a_{k,k}^{(0)} w_k^{(0)}(s)
\end{aligned}$$

Then

$$\begin{aligned}
\operatorname{Cov}\left(Z_{j-1}(s), Z_j(t) \mid \{Z_i(t)\}_{i=1}^{j-1}\right) &= \operatorname{Cov}\left(\sum_{i=1}^{j-1} a_{i,j-1}^{(0)} w_i^{(0)}(s), \sum_{i'=1}^{j} a_{i',j}^{(0)} w_{i'}^{(0)}(t) \mid \left\{w_{i''}^{(0)}(t)\right\}_{i''=1}^{j-1}\right) \\
&= \sum_{i=1}^{j-1}\sum_{i'=1}^{j} a_{i,j-1}^{(0)} a_{i',j}^{(0)} \operatorname{Cov}\left(w_i^{(0)}(s), w_{i'}^{(0)}(t) \mid \left\{w_{i''}^{(0)}(t)\right\}_{i''=1}^{j-1}\right) \\
\text{(due to independency)} &= 0 + \sum_{i=1}^{j-1} a_{i,j-1}^{(0)} a_{i,j}^{(0)} \operatorname{Cov}\left(w_i^{(0)}(s), w_i^{(0)}(t) \,|w_i^{(0)}(t)\right) \\
\text{(due to conditioning)} &= 0
\end{aligned}$$

This Markovian property may facilitate the computations as implies that the inverse sample covariance matrix of $\{Z_j(s)\}$ has zeros and may be sparse.

Note 281. Training of the parameterized LMC, can be performed in the semiparametric framework by specifying parametric semi-variogrames $\left\{\gamma_\theta^{(u)}(h)\right\}$ for each $\left\{w^{(u)}(\cdot)\right\}$ which yields a cross-semivariogram for $Z(s)$

$$\Gamma_\theta(h) = \sum_{u=0}^{m} B^{(u)} \gamma_\theta^{(u)}(h)$$

and minimizing as

$$\hat{\theta} = \arg\min_\theta \left(\sum_{l=1}^{c} \varpi(h_l) \operatorname{trace}\left(\left(\Gamma_\theta(h_l) - \hat{\Gamma}(h_l)\right)^2\right)\right)$$

where $\hat{\Gamma}(h_l)$ is the sample cross-semivariogram at separation value h_l, and $\varpi(\cdot)$ specified weights.

Note 282. Training of the parametrized LMC, can be performed (1.) in the Frequentist framework via MLE by maximizing the associated likelihood ; or (2.) in the Bayesian framework by specifying hyper priors and computing the posterior expectations

Example 283. Consider the geostatistical model as a mixed effect model

$$Z(s) = \mu(s) + \underbrace{\sum_{u=0}^{m} A^{(u)} w^{(u)}(s)}_{=v(s)} + \varepsilon(s).$$

Further, assume (1.) $[\mu(s)]_j = \Psi_j^\top(s)\beta_j$ with known bases functions $\Psi_j^\top(s)$, (2.) $w^{(u)}(s)$ are zero-mean Gaussian random fields, and (3.) $\varepsilon(s) \overset{\text{iid}}{\sim} \mathrm{N}(0, V)$ iid for any $s \in \mathcal{S}$. Let $v(s) = \sum_{u=0}^{m} A^{(u)} w^{(u)}(s)$. Consider dataset $\{(s_i, Z_1(s_i), ..., Z_k(s_i))\}_{i=1}^n$. Then the hierarchical model is

$$Z(s_i) \mid \{v(s_i)\} \overset{\text{ind}}{\sim} \mathrm{N}(\mu(s_i) + v(s_i), V), \quad i = 1, ..., n$$

$$(v(s_1), ..., v(s_n))^\top \sim \mathrm{N}\left(0, \sum_{u=0}^{m} R_\theta^{(u)} \otimes B^{(u)}\right)$$

where $R_\theta^{(u)}$ such as $\left[R_\theta^{(u)}\right]_{i,j} = \varrho_\theta(|s_i - s_j|)$ where θ are some additional parameters. The use of knonecker product operator $\otimes$ facilitates the inversion matrix operations required for MLE and Bayesian training to learn θ, β, and V.

Appendix A. Kronecker product

Note 284. Let A be an $m \times n$ matrix with elements $[A]_{i,j} = a_{i,j}$ and let B be a $p \times q$ matrix. The Kronecker product $A \otimes B$ is the $pm \times qn$ block matrix

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1,n}B \\ \vdots & \ddots & \vdots & \vdots & \vdots & \iddots & \vdots \\ \vdots & \cdots & a_{i-1,j-1}B & a_{i-1,j}B & a_{i-1,j+1}B & \cdots & \vdots \\ \vdots & \cdots & a_{i,j-1}B & a_{i,j}B & a_{i,j+1}B & \cdots & \vdots \\ \vdots & \cdots & a_{i+1,j-1}B & a_{i+1,j}B & a_{i+1,j+1}B & \cdots & \vdots \\ \vdots & \iddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1}B & \cdots & \cdots & \cdots & \cdots & \cdots & a_{m,n}B \end{bmatrix}$$

Note 285. Some properties:

- If A is an $m \times n$ matrix and B is a $p \times q$ matrix

$$(A \otimes B) \otimes C = A \otimes (B \otimes C)$$
$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$
$$(A \otimes B)(C + D) = (A \otimes C) + (B \otimes D)$$
$$(A \otimes B)^{\top} = A^{\top} \otimes B^{\top}$$

- If A is an $n \times n$ matrix and B is a $m \times m$ matrix

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$
$$\operatorname{trace}(A \otimes B) = \operatorname{trace}(A) \operatorname{trace}(B)$$
$$\det(A \otimes B) = (\det(A))^{m} (\det(B))^{n}$$

Appendix B. Vec operator

Note 286. The vec-operator applied on a $m \times n$ matrix A with elements $[A]_{i,j} = a_{i,j}$ stacks the columns into a vector as

$$\operatorname{vec}(A) = (a_{i,j}; i = 1, ..., m; j = 1, ..., n)^{\top} \ ; \ (\text{index } i \text{ runs faster than index } j)$$
$$= \left(\underbrace{a_{1,1}, ..., a_{n,1}}_{a_{1:n,1}}, \underbrace{a_{1,2}, ..., a_{n,2}}_{a_{1:n,2}}, ..., \underbrace{a_{i,j}, ..., a_{i+n,j}}_{a_{(i+1):(i+n),j}}, ..., \underbrace{a_{1,m-1}, ..., a_{n,m-1}}_{a_{1:n,(m-1)}}, \underbrace{a_{1,m}, ..., a_{n,m}}_{a_{1:n,m}} \right)^{\top}$$

Note 287. Some properties:

- Let A, B, and X be matrices with suitable sizes and let a, and c be vectors with suitable lengths, then

$$\begin{aligned}
\operatorname{vec}(A+B) &= \operatorname{vec}(A) + \operatorname{vec}(B) \\
\operatorname{trace}\left(A^{\top}B\right) &= (\operatorname{vec}(A))^{\top} \operatorname{vec}(B) \\
\operatorname{vec}(AXB) &= \left(B^{\top} \otimes A\right) \operatorname{vec}(X)
\end{aligned}$$

Note 288. Further info:

- Petersen, K. B., & Pedersen, M. S. (2008). The matrix cookbook. Technical University of Denmark, 7(15), 510.

Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

Lecture notes part 3: Aerial unit data / spatial data on lattices

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

Aim. To introduce Aerial unit data modeling: the basic building models.

Reading list & references:

[1] Cressie, N. (2015; Part II). Statistics for spatial data. John Wiley & Sons.
[2] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
[3] Gaetan, C., & Guyon, X. (2010; Ch 3). Spatial statistics and modeling (Vol. 90). New York: Springer.

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

Note 1. Recall from Section 2.2 of "Lecture notes part 1: Types of spatial data" that modeling aerial unit / lattice data types involves the use of random field models with a discrete index set. Such data are collected over areal units such as pixels, census districts or tomographic bins. Often, there is a natural neighborhood relation or neighborhood structure.

Note 2. This means we need to introduce suitable basic building models able to represent the characteristics of the underline data generating mechanisms. These as the "Discrete Random Fields".

1. Discrete Random Fields

Note 3. We re-introduce the definition of the random field with regards to the aerial unit data framework.

Definition 4. A random field $Z = (Z_s; s \in \mathcal{S})$ on a set of indexes $\mathcal{S}$ taking values in $\mathcal{Z}^{\mathcal{S}}$ is a family of random variables $\{Z_s := Z_s(\omega); s \in \mathcal{S}, \omega \in \Omega\}$ where each $Z_s(\omega)$ is defined on the same probability space $(\Omega, \mathfrak{F}, \mathrm{pr})$ and taking values in $\mathcal{Z}$.

Note 5. In aerial unite data modeling, the (spatial) set of sites $\mathcal{S}$, at which the process is defined, is discrete, it can be finite or infinite (e.g. $\mathcal{S} \subseteq \mathbb{Z}^d$), regular (e.g. pixels of an image) or irregular (states of a country).

Note 6. The general state space $\mathcal{Z}$ of the random field can be quantitative, qualitative or mixed. E.g., $\mathcal{Z} = \mathbb{R}_+$ in a Gamma random field, $\mathcal{Z} = \mathbb{N}$ in a Poisson random field, $\mathcal{Z} = \{0, 1\}$ in a binary random field.

Note 7. If $\mathcal{Z}$ is finite or countably infinite, the (joint)distribution of Z has a PMF

$$\mathrm{pr}_Z(z) = \mathrm{pr}(Z = z) = \mathrm{pr}(\{Z_s = z_s;\ s \in \mathcal{S}\}),\ \forall z \in \mathcal{Z}^{\mathcal{S}}$$

otherwise if $\mathcal{Z} \subseteq \mathbb{R}^d$ and Z continuous we will use the joint PDF.

Definition 8. The discrete set of sites $\mathcal{S} = \{s_i; i = 1, ..., n\}$ is often called lattice of sites.

Notation 9. Often we will use the notation Z_s instead of $Z(s)$ or Z_i instead of $Z(s_i)$. Hence, since $\mathcal{S} = \{s_i; i = 1, ..., n\}$, we can consider a more convenient notation

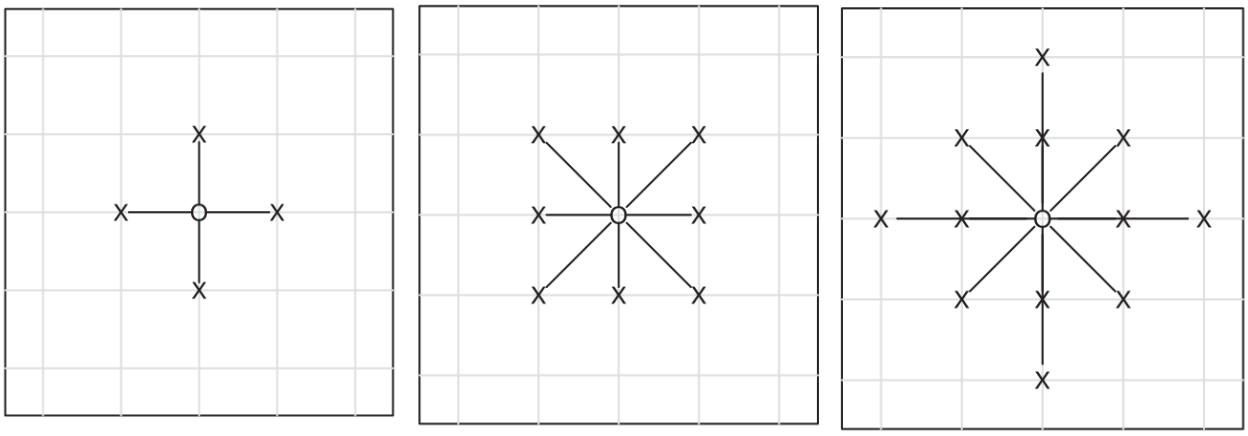
$$Z = (Z_s; s \in \mathcal{S})^\top = (Z_i = Z(s_i); i = 1, ...n)^\top.$$

Note 10. Modeling aerial unit data often requires the specification of a neighborhood relation or neighborhood structure.

Notation 11. The notation $i \sim j$ between two sites $i, j \in \mathcal{S}$ means that "sites i and j are neighboring" according to a "neighborhood relation" $\sim$.

Definition 12. Given a lattice of sites $\mathcal{S}$ and "neighborhood relation" $\sim$, we can define the neighborhood $\mathcal{N}_s$ of $s \in \mathcal{S}$ as

$$\mathcal{N}_s = \{s' \in \mathcal{S} : s \sim s'\}$$



Definition 13. Proximity matrix W is called a matrix W which aims at spatially connecting unites i and j in some fashion given some symmetric neighborhood relation $\sim$ on $\mathcal{S}$. Usually $[W]_{i,i} = 0$.

Note 14. Proximity matrix W may be such that it represents the neighborhood relation $\sim$ in a binary fashion e.g.

$$[W]_{i,j} = \begin{cases} 1 & \text{if } i \sim j \text{ and } i \neq j \\ 0 & \text{if } i \not\sim j \text{ or } i = j \end{cases}$$

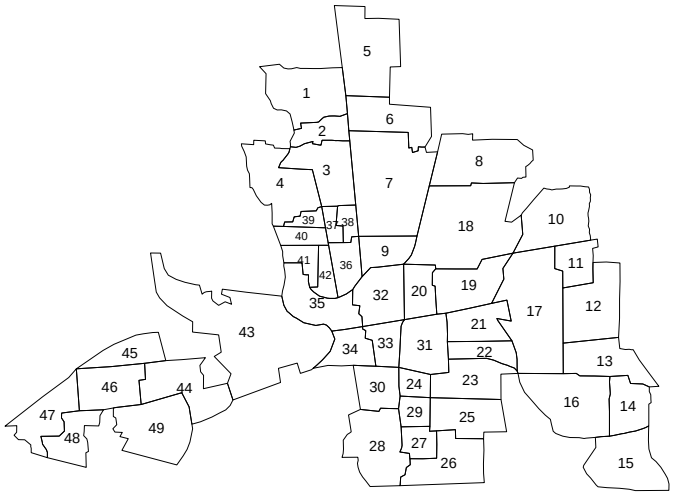


FIGURE 1.1. Lattice of spatial sites for Columbus dataset. Each neighborhood is a site. Each site is labeled. The collection of sites is the lattice of sites.

or how close site i is to site j based on some distance $d(i,j)$, e.g.

$$[W]_{i,j} = \begin{cases} 1/d(i,j) & \text{if } i \sim j \text{ and } i \neq j \\ 0 & \text{if } i \not\sim j \text{ or } i = j \end{cases}$$

Note 15. Proximity matrix W does not necessarily need to be symmetric, some times it is standardized as $[W]_{i,j} \leftarrow [W]_{i,j} / \sum_j [W]_{i,j}$.

Example 16. Consider the Columbus OH dataset which concerns spatially correlated count data arising from 49 districts/neighborhood in Columbus, OH in 1980. This is the R dataset `columbus{spdep}`. Figure 1.1 presents the sites and the lattice of sites. Each neighborhood is a site. Each site is label. The collection of sites is the lattice of sites coded with a unique labeled according to some order. One may define the "neighborhood relation $i \sim j$ considering counties that share common boarders (adjacent). Then for site $i = 43$, $i \sim j$ involves any $j \in \{44, 35, 34\}$ and for site $i = 20$, $i \sim j$ involves any $j \in \{32, 9, 18, 19, 31, 33\}$. Here $\mathcal{N}_{43} = \{44, 35, 34\}$ and $\mathcal{N}_{20} = \{32, 9, 18, 19, 31, 33\}$. The proximity matrix based on binary scheme will contain elements $W_{43,35} = 1$, $W_{43,43} = 0$, and $W_{43,33} = 0$.

Example 17. (Logistic/Ising model) Let variable Z_i denote the presence of a characteristic as $Z_i = 1$ or absence of it as $Z_i = 0$ on a site labeled by $i \in \mathcal{S}$. Then $\mathcal{Z} = \{0, 1\}$. The Ising model is defined by the (joint) PMF

$$\text{pr}_Z(z) \propto \exp\left(\alpha \sum_{i \in \mathcal{S}} z_i + \beta \sum_{\{i,j\}: i \sim j} z_i z_j\right), \ \forall z \in \mathcal{Z}^{\mathcal{S}} \tag{1.1}$$

E.g., it can model a black & white noisy image, where $\mathcal{S}$ denotes the labels of the image pixels, and Z_i denotes the presence of a black pixel ($Z_i = 1$) or its absence ($Z_i = 0$). Under Ising model (1.1), the characteristic is observed with probability $\mathrm{pr}_{Z_i}(z_i = 1) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ when $\beta = 0$. The characteristic's presence is encouraged in neighboring sites when $\beta > 0$, and discouraged when $\beta < 0$.

Notation 18. We use notation, for $\mathcal{A} \subset \mathcal{S}$

$$\mathrm{pr}_{\mathcal{A}}\left(z_{\mathcal{A}}|z_{\mathcal{S}\setminus\mathcal{A}}\right) = \mathrm{pr}\left(Z_{\mathcal{A}} = z_{\mathcal{A}}|Z_{\mathcal{S}\setminus\mathcal{A}} = z_{\mathcal{S}\setminus\mathcal{A}}\right)$$

Definition 19. Local characteristics of a random field Z on $\mathcal{S}$ with values in $\mathcal{Z}$ are the conditionals

$$\mathrm{pr}_i\left(z_i|z_{\mathcal{S}-i}\right) = \mathrm{pr}_{\{i\}}\left(z_{\{i\}}|z_{\mathcal{S}\setminus\{i\}}\right),\ i \in \mathcal{S},\ z \in \mathcal{Z}^{\mathcal{S}}$$

Example 20. (Cont. Example 17) The local characteristics of the Ising model in (1.1) are

$$\mathrm{pr}_i\left(z_i = 1|z_{\mathcal{S}-i}\right) = \frac{\exp\left(\alpha + \beta\sum_{\{i,j\}:i\sim j} z_j\right)}{1 + \exp\left(\alpha + \beta\sum_{\{i,j\}:i\sim j} z_j\right)}$$

2. Lattice Random Fields (background)

Note 21. (Recall basic properties Fourier transform) Let $\{\beta_s : s \in \mathcal{S}\}$, $\mathcal{S} \subseteq \mathbb{Z}^d$ be a set of real coefficients

- The Fourier transform
$$\tilde{\beta}(\omega) = \sum_{s\in\mathcal{S}} \beta_s e^{is^\top\omega},\ \omega \in (-\pi,\pi]^d.$$
- The inverse Fourier transform is
$$\beta_s = \int_{(-\pi,\pi]^d} e^{-is^\top\omega}\tilde{\beta}(\omega)\,\mathrm{d}\omega$$
- Regularity conditions for the Fourier transform to be a well-defined function are
  (1) If $\{\beta_s\}$ are summable, $\sum_{s\in\mathcal{S}}|\beta_s| < \infty$, then $\tilde{\beta}(\omega)$ is bounded and continuous function of ω.
  (2) If $\{\beta_s\}$ are square summable, $\sum_{s\in\mathcal{S}}|\beta_s|^2 < \infty$ then $\tilde{\beta}(\omega)$ is square-integrable over $(-\pi,\pi]^d$ and (visa versa).

Note 22. Let $(Z_s : s \in \mathcal{S})$ be a (weakly) stationary random field where $\mathcal{S} \subseteq \mathbb{Z}^d$. The following is a tool (similar to Bochner's theorem) for specifying covariance function of stationary lattice random field. It results from Herglotz's theorem.

Proposition 23. *(Herglotz's theorem) Let $c : \mathbb{Z}^d \to \mathbb{R}$ be a real-valued function on integers for $d \geq 1$. Then $c(\cdot)$ is positive semidefinite (stationary covariance function) if and only if*

it can be represented as

$$c(h) = \int_{(-\pi,\pi]^d} \exp\left(i\omega^\top h\right) dF(\omega)$$

where F is a *symmetric positive bounded finite measure on* $(-\pi,\pi]^d$ *and* $F(-\pi) = 0$. F *is called spectral measure of* $c(h)$. f *is called spectral density of* $c(h)$ *if*

$$dF(\omega) = f(\omega)\, d\omega$$

Proposition 24. *If* $c(\cdot)$ *is integrable, the spectral density* $f(\cdot)$ *can be computed by inverse Fourier transformation*

$$f(\omega) = \left(\frac{1}{2\pi}\right)^d \sum_h \exp\left(-i\omega^\top h\right) c(h)$$

Note 25. Let $\left(Y(s) : s \in \mathbb{R}^d\right)$ be a stationary random field with spectral measure $F_Y(\omega)$, $\omega \in \mathbb{R}^d$ and let $\left(Z_s : s \in \mathbb{Z}^d\right)$ with $Z_s = Y(s)$ for $s \in \mathbb{Z}^d$, then Z_s has spectral measure

$$F_Z(\omega) = \sum_{k\in\mathbb{Z}^d} F_Y(2\pi k + d\omega),\ \omega \in (-\pi,\pi]^d$$

where frequencies separated by a lag $2\pi k$, $k \in \mathbb{Z}^d$, are aliased together in the construction of$F_Z(\omega)$. Hence, there are infinitely many ways to interpolate a stationary process on $\mathbb{Z}^d$ to give a stationary random field $\mathbb{R}^d$.

Note 26. Let $(U_s : s \in \mathcal{S})$, $\mathcal{S} \subseteq \mathbb{Z}^d$, be a stationary random field covariance function

$$c_U(h) = \int_{(-\pi,\pi]^d} e^{i\omega^\top h} f(\omega)\, \mathrm{d}\omega$$

and spectral density $f(\omega)$ over $(-\pi,\pi]^d$. Let $(V_s : s \in \mathcal{S})$, $\mathcal{S} \subseteq \mathbb{Z}^d$, be a random field such as

$$V_s = \sum_{h\in\mathbb{Z}^d} U_{s+h}\beta_h,\ s \in \mathcal{S} \subseteq \mathbb{Z}^d$$

where $\left\{\beta_h : h \in \mathbb{Z}^d\right\}$ are summable functions, i.e. $\sum_h |\beta_h| < \infty$. The covariance function of $(V_s : s \in \mathcal{S})$ is

$$c_V(h) = \mathrm{Cov}(V_s, V_{s+h}) = \sum_{t\in\mathbb{Z}^d}\sum_{t'\in\mathbb{Z}^d} \beta_t\beta_{t'} c_U(h + t - t')$$

with spectral measure

$$\mathrm{d}F_V(\omega) = \left|\tilde{\beta}(\omega)\right|^2 \mathrm{d}F_U(\omega),\ \omega \in (-\pi,\pi]^d \tag{2.1}$$

where

$$\tilde{\beta}(\omega) = \sum_{h\in\mathbb{Z}^d} \beta_h e^{ih^\top\omega},\ \omega \in (-\pi,\pi]^d \tag{2.2}$$

is the Fourier transform of β_h.

Proof. This is straightforward from Proposition 23 (Herglotz's theorem) □

3. Compatibility of conditional distributions

Note 27. Here, we discuss how to represent a joint probability distribution via its full conditionals. We need this for model building purposes.

Definition 28. Let random vector $Z = (Z_1, ..., Z_n)$ with joint distribution $\pi(Z_1, ..., Z_n)$. The set of distributions $\{\pi_i(\cdot|Z_{-i}); i = 1, ...n\}$ is called compatible to the joint distribution $\pi(Z_1, ..., Z_n)$ if the joint distribution $\pi(Z_1, ..., Z_n)$ has conditionals $\{\pi_i(Z_i|Z_{-i}); i = 1, .., n\}$.

Note 29. To specify suitable building models representing spatial dependency of a random field $(Z_i)_{i\in\mathcal{S}}$, it is often easier to visualize the joint distribution pr_z in terms of conditional distributions $\{\pi_i(Z_i|Z_{\mathcal{S}-i}); i \in \mathcal{S}\}$ rather than directly.

Note 30. Thus, instead of specifying a joint model for $(Z_i)_{i\in\mathcal{S}}$, a researcher may propose putative families of conditional distributions $\{\pi_i(Z_i|Z_{\mathcal{S}-i}); i \in \mathcal{S}\}$. However, an arbitrary chosen set of conditional distributions $\{\pi_i(\cdot|\cdot); i \in \mathcal{S}\}$ is not generally compatible, in the sense that there exists a joint distribution for $(Z_i)_{i\in\mathcal{S}}$, and hence we need to impose conditions.

Note 31. In what follows, we discuss necessary and sufficient conditions regarding compatibility.

Proposition 32. *(Compatibility condition) Let F be a joint distribution with $dF(x, y) = f(x, y)\, d(x, y)$ on $\mathcal{S}_x \times \mathcal{S}_y$. Let candidate condition distributions*

$$G \text{ with } dG(x|y) = g(x|y)\, dx, \text{ on } x \in \mathcal{S}_x$$
$$Q \text{ with } dQ(y|x) = q(y|x)\, dy, \text{ on } y \in \mathcal{S}_y$$

and let $N_g = \{(x, y) : g(x|y) > 0\}$ and $N_q = \{(x, y) : q(y|x) > 0\}$. A distribution F with conditionals exists iff

(1) $N_g = N_q = N$
(2) there exist functions u and $v$where $g(x|y)/q(y|x) = u(x)\, v(y)$ for all $(x, y) \in N$ and $\int u(x)\, dx < \infty$

Proof. Omitted[1]. □

[1]See Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. Journal of the American Statistical Association, 84(405), 152-156.

Note 33. Essentially the above conditions guarantee that

$$k(y)\, g(x|y) = f(x,y) = h(x)\, q(y|x)$$

where k, g, h, q are densities.

Example 34. The conditionals $x|y \sim \mathrm{N}(a+by, \sigma^2+\tau^2 y^2)$ and $y|x \sim \mathrm{N}(c+dx, \tilde{\sigma}^2+\tilde{\tau}^2 x^2)$ are compatible if $\tau^2 = \tilde{\tau}^2 = 0$, $d/\tilde{\sigma}^2 = b/\sigma^2$, and $|db| < 1$.

Solution. See Exercise 29 in the Exercise sheet.

Note 35. Proposition 32 can be extended to more dimensions. For more info see (Arnold, B. C., & Press, S. J. (1989). in footnote 1)

Note 36. The following theorem shows that local characteristics can determine the entire distribution in certain cases.

Theorem 37. *(Besag's factorization theorem; Brook's Lemma) Let Z be a $\mathcal{Z}$ valued random field taking values in $\mathcal{Z}^{\mathcal{S}}$ where $\mathcal{S} = \{1, ..., n\}$ with $n \in \mathbb{N}$, and such as $pr_Z(z) > 0$, $\forall z \in \mathcal{Z}^{\mathcal{S}}$. Then for all*

$$\frac{pr_Z(z)}{pr_Z(z^*)} = \prod_{i=1}^{n} \frac{pr_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{pr_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}, \quad \forall z, z^* \in \mathcal{Z}^{\mathcal{S}} \tag{3.1}$$

Proof. I will show that

$$\mathrm{pr}_Z(z) = \prod_{i=1}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)} \mathrm{pr}_Z(z^*)$$

It is

$$\mathrm{pr}_Z(z_1, ..., z_n) = \frac{\mathrm{pr}_n(z_n|z_1, ..., z_{n-2}, z_{n-1})}{\mathrm{pr}_n(z_n^*|z_1, ..., z_{n-2}, z_{n-1})} \mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*)$$

Let proposition P_j be

$$\mathrm{pr}_Z(z) = \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)} \mathrm{pr}_Z\left(z_1, ..., z_{n-j-1}, z_{n-j}^*, ..., z_n^*\right)$$

Proposition P_0 is true

$$\mathrm{pr}_Z(z) = \frac{\mathrm{pr}_n(z_n|z_1, ..., z_{n-1})}{\mathrm{pr}_n(z_n^*|z_1, ..., z_{n-1})} \mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*) \tag{3.2}$$

Proposition P_1 is true

$$\mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*) = \frac{\mathrm{pr}_{n-1}(z_{n-1}|z_1, ..., z_{n-2}, z_n^*)}{\mathrm{pr}_{n-1}(z_{n-1}^*|z_1, ..., z_{n-2}, z_n^*)} \mathrm{pr}_Z\left(z_1, ..., z_{n-2}, z_{n-1}^*, z_n^*\right)$$

Assume that P_j is true. Then proposition $\mathrm{P}_{j+1}$is true as well, because

$$\begin{aligned}
\mathrm{pr}_Z(z) &= \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)}{\mathrm{pr}_i\left(z^*_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-j-1},z^*_{n-j},...,z^*_n\right) \\
&= \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)}{\mathrm{pr}_i\left(z^*_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)} \\
&\quad \times \frac{\mathrm{pr}_{n-j-1}\left(z_{n-j-1}|z_1,...,z_{n-j-2},z^*_{n-j},...,z^*_n\right)}{\mathrm{pr}_{n-j-1}\left(z^*_{n-j-1}|z_1,...,z_{n-j-2},z^*_{n-j},...,z^*_n\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-j-2},z^*_{n-j-1},...,z^*_n\right) \\
&= \prod_{i=n-(j+1)}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)}{\mathrm{pr}_i\left(z^*_i|z_1,...,z_{i-1},z^*_{i+1},...,z^*_n\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-(j+1)-1},z^*_{n-(j+1)},...,z^*_n\right)
\end{aligned}$$

Then (3.1) is correct according to the induction principle. □

Note 38. Theorem 37 shows that the joint $\mathrm{pr}_Z(\cdot)$ can be constructed from its conditionals $\{\mathrm{pr}_i(\cdot|\cdot)\}$ if distributions $\{\mathrm{pr}_i(\cdot|\cdot)\}$ are compatible for $\mathrm{pr}_Z(\cdot)$, under the requirement that this construction is invariant wrt the coordinate permutation $\{1,...,n\}$ and the reference state z^*– these invariances correspond to the conditions in Proposition 32.

4. Spatial autoregressive models

We present two basic spatial Autoregressive models, the SAR and CAR, able to represent spatial dependency.

4.1. **Simultaneous AutoRegressive (SAR) models.**

Definition 39. "Autoregressive" representation for lattice random field $(Z_s; s \in \mathcal{S})$, $\mathcal{S} \subseteq \mathbb{Z}^d$ is called

$$\sum_{h \in \mathcal{H} \cup \{0\}} a_h \left(Z_{s+h} - \mu_{s+h}\right) = \varepsilon_s, \ s \in \mathcal{S} \tag{4.1}$$

where $\{a_h; h \in \mathcal{H}\}$, $\mathcal{H} = \left\{h \in \mathbb{Z}^d - \{0\} : s + h \in \mathcal{S}\right\}$, are coefficients and $\{\varepsilon_s; s \in \mathcal{S}\}$ is a discrete white noise random field with variance $\lambda_s = \mathrm{Var}(\varepsilon_s)$. We will denote it as SAR($\mathcal{H}$).

4.1.1. *Assuming stationarity.*

Note 40. We will focus our study on stationary CAR random field $(Z_s; s \in \mathcal{S})$, i.e. $\mu_s = \mu$ and $\lambda_s = \sigma^2_\varepsilon$ for $s \in \mathcal{S} \subseteq \mathbb{Z}^d$.

Note 41. $\{\varepsilon_s; s \in \mathcal{S}\}$ has c.f. $c_\varepsilon(h) = \sigma_\varepsilon^2 \delta_{\{0\}}(h)$ and hence spectral density $f_\varepsilon(\omega) = \sigma_\varepsilon^2/(2\pi)^d$. The spectral density f for $(Z_s; s \in \mathcal{S})$ is

$$f(\omega) = \frac{1}{|\tilde{a}(\omega)|^2} f_\varepsilon(\omega) = \frac{1}{|\tilde{a}(\omega)|^2} \left(\frac{1}{2\pi}\right)^d \sum_h e^{-i\omega^\top h} c_\varepsilon(h)\, \mathrm{d}h \tag{4.2}$$
$$= \frac{1}{|\tilde{a}(\omega)|^2} \left(\frac{1}{2\pi}\right)^d \sum_{h\in\mathbb{Z}^d} e^{-i\omega^\top h} \sigma_\varepsilon^2 \delta_{\{0\}}(h)\, \mathrm{d}h = \frac{1}{|\tilde{a}(\omega)|^2} \frac{\sigma_\varepsilon^2}{(2\pi)^d}$$

where $\tilde{a}(\omega) = \sum_h a_h e^{ih^\top \omega}$ since (26).

Note 42. For random field $(Z_s; s \in \mathcal{S})$ to be stationary, $f(\omega)$ in (5.1) must be integrable function and bounded function on $\omega \in (-\pi, \pi]^d$. Hence, we can set restrictions on coefficients

$$a_0 > 0, \text{ and } \sum_h |a_h| < a_0 \tag{4.3}$$

satisfying regularity conditions in (21).

Note 43. To make model (6.3) identifiable and use it for inference, we can introduce further restrictions $a_h = a_{-h}$.

Definition 44. Lattice random field $(Z_s; s \in \mathcal{S})$ is called Simultaneous AutoRegressive (SAR) if it is given in an "Autoregression" representation (6.3)

$$\sum_{h\in\mathcal{N}} a_h Z_{s+h} = \varepsilon_s,\ s \in \mathcal{S}$$

whose coefficients $\{a_h; h \in \mathcal{H}\}$ satisfy the symmetry condition

$$a_h = a_{-h},\ \forall h$$

and $f_Z(\omega)$ in (5.1) is an integrable function over $\omega \in (-\pi, \pi]^d$.

4.1.2. *Assuming Gaussian distribution.*

Note 45. Following we provide the matrix form of the definition used in software implementations.

Definition 46. Consider discrete set of sites $\mathcal{S} = \{s_i; i = 1, ..., n\}$ and a lattice random field $(Z_s; s \in \mathcal{S})$. Vectorize $Z = (Z_1, ..., Z_n)^\top$ with $Z_i = Z(s_i)$ and set

$$Z = \mu + A(Z - \mu) + E \Longleftrightarrow E = (I - A)(Z - \mu)$$

Assume that matrix A is such that $[A]_{i,i} = 0$ and $(I - A)^{-1}$ exists. Assume n-dimensional Gaussian random vector $E \sim \mathrm{N}_n(0, \Lambda)$ with $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. We say that Z follows a "Gaussian" Simultaneous Autoregressive model.

Note 47. The joint distribution of Z following the SAR model in Definition 46 is

$$Z \sim \mathrm{N}\left(\mu, (I-A)^{-1}\Lambda\left(I-A^{\top}\right)^{-1}\right) \tag{4.4}$$

Proof. Z is a linear combination of Gaussian random vectors, hence if follows a Gaussian distribution. It's mean and variance are

$$\begin{aligned}\mathrm{E}(Z) &= \mathrm{E}\left((I-A)^{-1}E+\mu\right)=\mu,\\ \mathrm{Var}(Z) &= \mathrm{Var}\left((I-A)^{-1}E+\mu\right)=(I-A)^{-1}\mathrm{Var}(E)\left(I-A^{\top}\right)^{-1}=(I-A)^{-1}\Lambda\left(I-A^{\top}\right)^{-1}\end{aligned}$$

□

4.2. **Conditional autoregressive models (CAR).**

Definition 48. A lattice random field $(Z_s; s\in\mathcal{S})$, $\mathcal{S}\subseteq\mathbb{Z}^d$ is called Conditional AutoRegressive (CAR) model if

$$\begin{aligned}\mathrm{E}(Z_s|Z_{S-s}) &= \mu_s+\sum_{h\in\mathcal{H}} b_h\left(Z_{s+h}-\mu_{s+h}\right)\\ \mathrm{Var}(Z_s|Z_{S-s}) &= \kappa_s\end{aligned}$$

where $b_h=-b_h$ and $\mathcal{H}=\left\{h\in\mathbb{Z}^d-\{0\}: s+h\in\mathcal{S}\right\}$. We will denote it as CAR($\mathcal{H}$).

Note 49. Alternatively CAR lattice random field $(Z_s; s\in\mathcal{S})$, $\mathcal{S}\subseteq\mathbb{Z}^d$can be written as

$$Z_s=\mu_s+\sum_{h\in\mathcal{H}} b_h\left(Z_{s+h}-\mu_{s+h}\right)+\epsilon_s \tag{4.5}$$

where $(\epsilon_s : s\in\mathcal{S})$ is the residual random field with mean $\mathrm{E}(\epsilon_s)=0$ and variance $\mathrm{Var}(\epsilon_s)=\kappa_s$. Also $b_h=-b_h$ and $\mathcal{H}=\left\{h\in\mathbb{Z}^d-\{0\}: s+h\in\mathcal{S}\right\}$.

4.2.1. *Assuming stationary.*

Note 50. We will focus our study on stationary CAR random field $(Z_s; s\in\mathcal{S})$, and hence we set $\mu_s=\mu$ and $\kappa_s=\kappa$ for $s\in\mathcal{S}\subseteq\mathbb{Z}^d$.

Note 51. It is

$$\begin{aligned}\mathrm{E}(Z_s\epsilon_s) &= \mathrm{E}\left(\left(\epsilon_s-\mu-\sum_{h\in\mathcal{H}} b_h\left(Z_{s+h}-\mu\right)\right)\epsilon_s\right)\\ &= \mathrm{E}\left(\epsilon_s^2\right)-\mu\mathrm{E}(\epsilon_s)-\sum_{h\in\mathcal{H}} b_h\mathrm{E}\left(\left(Z_{s+h}-\mu\right)\epsilon_s\right)\\ &= \mathrm{E}\left(\epsilon_s^2\right)+0+0=\kappa\end{aligned}$$

Note 52. The covariance function $c(\cdot)$ of stationary CAR random field $(Z_s; s\in\mathcal{S})$ is

$$c(h)=\sum_{h'\in\mathcal{H}} b_{h'}c\left(h-h'\right)+\kappa\delta_{\{0\}}(h)$$

as, for $h \neq 0$ it is

$$\begin{aligned} c(h) &= \mathrm{E}\left(\left(Z_s-\mu\right)\left(Z_{s+h}-\mu\right)\right) = \mathrm{E}\left(\left(\sum_{h'\in\mathcal{H}} b_{h'}\left(Z_{s+h'}-\mu\right)+\epsilon_s\right)\left(Z_{s+h}-\mu\right)\right) \\ &= \mathrm{E}\left(\left(\sum_{h'\in\mathcal{H}} b_{h'}\left(Z_{s+h'}-\mu\right)+\epsilon_s\right)\left(Z_{s+h}-\mu\right)\right) \\ &= \sum_{h'\in\mathcal{H}} b_{h'}\mathrm{E}\left(Z_{s+h'}Z_{s+h}\right) - \mu^2\sum_{h'\in\mathcal{H}} b_{h'} + \mathrm{E}\left(\epsilon_s Z_{s+h}\right) \\ &= \sum_{h'\in\mathcal{H}} b_{h'}b_{h'}c\left(h-h'\right) - 0 + 0 \end{aligned}$$

and for $h = 0$

$$\begin{aligned} c(0) &= \mathrm{E}\left(\left(Z_s-\mu\right)\left(Z_s-\mu\right)\right) = \\ &= \mathrm{E}\left(\left(\sum_{h\in\mathcal{H}} b_h\left(Z_{s+h}-\mu\right)+\epsilon_s\right)\left(\sum_{h'\in\mathcal{H}} b_{h'}\left(Z_{s+h'}-\mu\right)+\epsilon_s\right)\right) \\ &= \sum_{h\in\mathcal{H}}\sum_{h'\in\mathcal{H}} b_{h'}b_{h'}c\left(h-h'\right)+\kappa \end{aligned}$$

Note 53. The spectral density of the covariance function $c(h)$ of stationary CAR random field $(Z_s; s \in \mathcal{S})$ is computed by inverse Fourier transform as

$$\begin{aligned} f(\omega) &= \left(\frac{1}{2\pi}\right)^d \sum_{\mathrm{h}} e^{-i\omega^\top h} c(h) \\ &= \left(\frac{1}{2\pi}\right)^d \sum_{\mathrm{h}} e^{-i\omega^\top h}\left(\sum_{h'\in\mathcal{H}} b_{h'}c\left(h-h'\right)+\kappa\delta_{\{0\}}(h)\right) \\ &= \tilde{b}(\omega) f(\omega) + \kappa\frac{1}{(2\pi)^d} \implies \\ f(\omega) &= \frac{\kappa}{(2\pi)^d}\frac{1}{1-\tilde{b}(\omega)},\ \tilde{b}(\omega) = \sum_{h\in\mathcal{H}} b_h e^{ih^\top\omega} = \sum_{h\in\mathcal{H}} b_h\cos\left(\omega^\top h\right) \end{aligned} \tag{4.6}$$

Hence sufficient conditions for CAR random field $(Z_s; s \in \mathcal{S})$ in (5.3) to be stationary is the spectral density (5.4) to be bounded. This is true if $\tilde{b}(\omega) < 1$ which is implied by

$$\sum_{h\in\mathcal{H}} |b_h| < \infty$$

satisfying regularity conditions in (21).

4.2.2. *Assuming Gaussian distribution.*

Note 54. Following we provide the matrix form of the definition used in software implementations.

Definition 55. "Gaussian" Conditional autoregressive model, CAR, assumes that the local characteristics $\{\mathrm{pr}_i(z_i|z_{\mathcal{S}-i})\}$ are Gaussian distributions

$$Z_i|z_{\mathcal{S}-i} \sim \mathrm{N}\left(\underbrace{\mu_i + \sum_{j\neq i} b_{i,j}(Z_j - \mu_j)}_{=\mathrm{E}(Z_i|Z_{\mathcal{S}-i})}, \underbrace{\kappa_i}_{=\mathrm{Var}(Z_i|Z_{\mathcal{S}-i})}\right), \quad \forall i \in \mathcal{S} \tag{4.7}$$

Proposition 56. *Let $K = diag(\{\kappa_i\})$ with $\kappa_i > 0$, matrix B with $B_{i,i} := [B]_{i,i} = 0$, and real vector μ with suitable dimensions. If Z follows a Gaussian CAR (Definition 55), $I - B$ is non-singular, and $(I-B)^{-1}K > 0$, then the joint distribution of Z is*

$$Z \sim N\left(\mu, (I-B)^{-1}K\right). \tag{4.8}$$

Proof. Without lose of generality, consider zero mean $\mu = 0$ (or equivalently set $Z := Z - \mu$). The full conditionals $Z_i|z_{\mathcal{S}-i}$ in (4.7) are compatible with the joint distribution $\mathrm{pr}_Z(z)$. By using Besag's factorization theorem (Theorem 37) with reference state/configuration $z^* = 0$ we get

$$\begin{aligned}
\mathrm{pr}_Z(z) &= \prod_{i=1}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z^*_{i+1}=0,...,z^*_n=0\right)}{\mathrm{pr}_i\left(z^*_i=0|z_1,...,z_{i-1},z^*_{i+1}=0,...,z^*_n=0\right)}\mathrm{pr}_Z(z^*=0) \\
&= \prod_{i=1}^{n} \frac{\mathrm{N}\left(z_i|\sum_{j<i} b_{i,j}z_j + 0, \kappa_i\right)}{\mathrm{N}\left(0|\sum_{j<i} b_{i,j}z_j + 0, \kappa_i\right)}\mathrm{pr}_Z(z^*=0) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i - \sum_{j<i} b_{i,j}z_j\right)^2 + \frac{1}{2\kappa_i}\left(0 - \sum_{j<i} b_{i,j}z_j\right)^2\right) \\
&= \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i^2 - 2z_i\sum_{j<i} b_{i,j}z_j\right)\right)\mathrm{pr}_Z(z^*=0) \\
&= \exp\left(-\sum_i \frac{z_i^2}{2\kappa_i} + \frac{1}{2}2\sum_i\sum_{j<i}\frac{b_{i,j}}{\kappa_i}z_iz_j\right)\mathrm{pr}_Z(z^*=0) \\
&= \exp\left(-\frac{1}{2}z^\top K^{-1}z + \frac{1}{2}z^\top K^{-1}Bz\right)\mathrm{pr}_Z(z^*=0) = \exp\left(-\frac{1}{2}z^\top\left[K^{-1}(I-B)\right]z\right)\mathrm{pr}_Z(z^*=0) \\
&= \mathrm{N}\left(z|0,(I-B)^{-1}K\right)
\end{aligned} \tag{4.9}$$

Recovering the mean from (4.9), it is

$$\mathrm{pr}_Z(z) = \mathrm{N}\left(z-\mu|0,(I-B)^{-1}K\right) = \mathrm{N}\left(z|\mu,(I-B)^{-1}K\right)$$

□

Note 57. When CAR is used for modeling, B is often specified to be sparse either due to some natural problem specific property, or for our computational convenience as is may allow the use of sparse solvers. To achieve this, one way is to specify $B = \phi W$ where $\phi > 0$ and W is an adjacency/proximity matrix; that is $[B]_{i,j} = \phi 1(i \sim j)\, 1(i \neq j)$ will be non-zero only for adjacent pairs i and j.

Note 58. The system in (4.8) can be rewritten as

$$Z = \mu + B(Z-\mu) + E \Longleftrightarrow E = (I-B)(Z-\mu) \tag{4.10}$$

by setting $E = (I-B)(Z-\mu)$. The distribution of Z in (4.8) induces a distribution on E as $E \sim \mathrm{N}\left(0, K(I-B)^\top\right)$ because

$$\mathrm{E}(E) = \mathrm{E}((I-B)(Z-\mu)) = (I-B)\,\mathrm{E}(Z-\mu) = 0$$
$$\mathrm{Var}(E) = \mathrm{Var}((I-B)Z) = (I-B)\,\mathrm{Var}(Z)\,(I-B)^\top = (I-B)(I-B)^{-1}K(I-B)^\top$$

4.3. **A comparison between CAR and SAR.**

Note 59. In most of the cases a SAR model can be written as a CAR model implying that CAR family of models can be more general than that of SAR models.

Note 60. We compare the use and flexibility of the two models, in particular the stationary CAR($\mathcal{H}_{\mathrm{CAR}}$) and the stationary SAR($\mathcal{H}_{\mathrm{SAR}}$) on $\mathbb{Z}^d$.

Example 61. Consider the stationary SAR and CAR models on $\mathbb{Z}^d$.

(1) Every stationary SAR model on $\mathbb{Z}^d$ is a stationary CAR model on $\mathbb{Z}^d$.
(2) When $d \geq 2$, the family of CAR models is larger than that of the SAR models.

Proof. Consider the stationary models SAR($\mathcal{H}_{\mathrm{SAR}}$) and CAR($\mathcal{H}_{\mathrm{CAR}}$) :

$$\mathrm{CAR}(\mathcal{H}_{\mathrm{CAR}}): Z_s = \mu + \sum_{h\in\mathcal{H}_{\mathrm{CAR}}} b_h(Z_{s+h}-\mu) + \epsilon_s,\ s\in\mathcal{S} \tag{4.11}$$

$$\mathrm{Var}(\epsilon_s) = \kappa,\ \mathrm{E}(\epsilon_s) = 0 \tag{4.12}$$

$$\mathrm{SAR}(\mathcal{H}_{\mathrm{SAR}}): \sum_{h\in\mathcal{H}_{\mathrm{SAR}}} a_h(Z_{s+h}-\mu) = \varepsilon_s,\ s\in\mathcal{S} \tag{4.13}$$

$$\mathrm{Var}(\varepsilon_s) = \lambda,\ \mathrm{E}(\epsilon_s) = 0 \tag{4.14}$$

(1) The spectral density of CAR($\mathcal{H}_{\text{CAR}}$) is

$$f_{\text{CAR}(\mathcal{H}_{\text{CAR}})}(\omega) = \frac{\kappa}{(2\pi)^d} \frac{1}{1 - \sum_{h \in \mathcal{H}_{\text{CAR}}} b_h e^{ih^\top \omega}} = \frac{\lambda}{(2\pi)^d} \frac{1}{\frac{\lambda}{\kappa} - \sum_{h \in \mathcal{H}_{\text{CAR}}} \frac{\lambda}{\kappa} b_h e^{ih^\top \omega}}$$

I can choose to set $\mathcal{H}_{\text{CAR}} = \{i - j : i \in \mathcal{H}_{\text{SAR}} \cup \{0\} \text{ and } j \in \mathcal{H}_{\text{SAR}} \cup \{0\}\}$,

$$b_h = \begin{cases} -\frac{\kappa}{\lambda} \sum_{v \in \mathcal{H}_{\text{SAR}}} a_v a_{v+h} & , h \neq 0 \\ 1 & , h = 0 \end{cases}, \text{ for } h \in \mathcal{H}_{\text{CAR}}$$

then

$$\begin{aligned} f_{\text{CAR}(\mathcal{H}_{\text{CAR}})}(\omega) &= \frac{\lambda}{(2\pi)^d} \frac{1}{\frac{\lambda}{\kappa} - \sum_{h \in \mathcal{H}_{\text{CAR}}} \sum_{v \in \mathcal{H}_{\text{SAR}}} a_v a_{v+h} e^{i(h \pm v)^\top \omega}} \\ &= \frac{\lambda}{(2\pi)^d} \frac{1}{\left| \sum_{h \in \mathcal{H}_{\text{SAR}}} a_h e^{ih^\top \omega} \right|^2} \end{aligned}$$

which is the spectral density of SAR($\mathcal{H}_{\text{SAR}}$).

(2) We consider

$$Z_s = c \sum_{h \in \mathcal{H}_{\text{CAR}}} Z_{s+h} + \epsilon_s$$

where $\mathcal{H}_{\text{CAR}} = \{h \in \mathbb{Z}^2 : |h|_1 = 1\}$ and $c \neq 0$. Then

$$f_{\text{CAR}(\mathcal{H}_{\text{CAR}})}(\omega) = \frac{\kappa}{(2\pi)^d} \frac{1}{1 - \sum_{h \in \mathcal{H}_{\text{CAR}}} b_h e^{ih^\top \omega}} = \frac{\lambda}{(2\pi)^d} \frac{1}{c\left(1 - 2\cos(\omega_1 h_1) - 2\cos(\omega_2 h_2)\right)}$$

If some SAR($\mathcal{H}_{\text{SAR}}$) had this spectral density, then $\mathcal{H}_{\text{SAR}} \subset \mathcal{H}_{\text{CAR}}$. Noting that it must be either $a_{(1,0)} \neq 0$ or $a_{(0,-1)} \neq 0$ and either either $a_{(0,1)} \neq 0$ or $a_{(-1,0)} \neq 0$. Assume $a_{(1,0)} \neq 0$ and $a_{(0,1)} \neq 0$. In this case the spectral density should contain a term $\cos(\omega_1 h_1 - \omega_2 h_2)$ which is not the case. So stationary CAR model has no stationary SAR representation.

□

5. Related Random Fields with particular properties

Note 62. We introduce more general modeling structures for basic spatial models which are computationally convenient yet quite descriptive for spatial statistical modeling. Convenient because they aim to break a high-dimensional problem into smaller ones using conditional independence, and reasonable because they allow representation of spatial dependence as well. We introduce the Gibbs Random Fields and the Markov Random Fields. The aforesaid Ising, CAR, and SAR models are just special cases of modeling structures.

5.1. **Gibbs Random Fields.**

Notation 63. Recall notation $z_{\mathcal{A}} = (z_i : i \in \mathcal{A})$ and $\mathcal{Z}^{\mathcal{A}} = \left\{z_{\mathcal{A}} : z \in \mathcal{Z}^{\mathcal{S}}\right\}$ for $\mathcal{A} \subseteq \mathcal{S}$.

Definition 64. Let $\mathcal{S} \neq \emptyset$ be a finite collection of sites. Let $\mathcal{Z} \subset \mathbb{R}$. Interaction potential is a family $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ of potential functions $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$ such that $V_{\emptyset}(\cdot) := 0$ and for every set $\mathcal{A} \subseteq \mathcal{S}$ the sum

$$U_{\mathcal{A}}^{\mathcal{V}}(z) = \sum_{\{\forall \mathcal{B} \subseteq \mathcal{S} : \mathcal{A} \cap \mathcal{B} \neq \emptyset\}} V_{\mathcal{B}}(z_{\mathcal{B}}) \tag{5.1}$$

exists.

Definition 65. In Definition 64, the function $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$ is called potential on $\mathcal{A}$.

Definition 66. In Definition 64, the function $U_{\mathcal{A}}^{\mathcal{V}}(z)$ in (5.1) is called energy function of interaction potential $\mathcal{V}$ on $\mathcal{A}$ is called.

Definition 67. The interaction potential $\mathcal{V}$ is said to be admissible if for all $\mathcal{B} \subseteq \mathcal{S}$ and $z_{\mathcal{S}\setminus\mathcal{B}} \in \mathcal{Z}^{\mathcal{S}\setminus\mathcal{B}}$

$$C_{\mathcal{A}}^{\mathcal{V}}\left(z_{\mathcal{S}\setminus\mathcal{A}}\right) = \int \exp\left(U_{\mathcal{A}}^{\mathcal{V}}\left(\left(z_{\mathcal{A}}, z_{\mathcal{S}\setminus\mathcal{A}}\right)\right)\right) \mathrm{d}z_{\mathcal{A}} < \infty$$

Note 68. This allow as to define a distribution corresponding to the energy.

Definition 69. Let Z be $\mathcal{Z}$ valued Random Field on a finite collection of sites $\mathcal{S}$ with $\mathcal{S} \neq \emptyset$, and let $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ be an interaction potential of functions $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$. Assume that $\mathcal{V}$ is admissible. Then Z is a Gibbs Random Field with interaction potentials $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ if

$$\mathrm{pr}_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\setminus\mathcal{A}}\right) = \frac{1}{C_{\mathcal{A}}^{\mathcal{V}}\left(z_{\mathcal{S}\setminus\mathcal{A}}\right)} \exp\left(\underbrace{\sum_{\{\mathcal{B} \subseteq \mathcal{S} \,:\, \mathcal{A} \cap \mathcal{B} \neq \emptyset\}} V_{\mathcal{B}}(z_{\mathcal{B}})}_{= U_{\mathcal{A}}^{\mathcal{V}}(z)}\right), \quad z \in \mathcal{Z}^{\mathcal{S}} \tag{5.2}$$

Definition 70. The normalizing integral $C_{\mathcal{A}}^{\mathcal{V}}$ in (5.2) is called partition function.

Notation 71. For the marginal $\mathrm{pr}_Z(z_{\mathcal{S}})$ we will denote

$$\mathrm{pr}_Z(z_{\mathcal{S}}) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(U_{\mathcal{S}}^{\mathcal{V}}(z)\right) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(\sum_{\mathcal{B} \subseteq \mathcal{S}} V_{\mathcal{B}}(z_{\mathcal{B}})\right), \; z \in \mathcal{Z}^{\mathcal{S}}$$

where $C_{\mathcal{S}}^{\mathcal{V}} < \infty$ is the constant. In this case (and when it is clear), to easy the notation, we can omit $\cdot_{\mathcal{S}}^{\mathcal{V}}$ and just write

$$\mathrm{pr}_Z(z_{\mathcal{S}}) = \frac{1}{C} \exp\left(\sum_{\mathcal{B} \subseteq \mathcal{S}} V_{\mathcal{B}}(z_{\mathcal{B}})\right), \; z \in \mathcal{Z}^{\mathcal{S}}$$

Example 72. (Ising model) In Example 17, the Ising model (1.1) has potentials

$$\begin{aligned} V_{\emptyset}(z) &= 0 \\ V_{\{i\}}(z), &= \alpha z_i \, \forall i \in \mathcal{S} \\ V_{\{i,j\}}(z) &= \begin{cases} \beta z_i z_j & \text{if } i \sim j \\ 0 & \text{if } i \not\sim j \end{cases} \\ V_{\mathcal{A}}(z) &= 0, \text{ if } \operatorname{card}(\mathcal{A}) > 2 \end{aligned}$$

it has energy function

$$U(z) := U_{\mathcal{S}}^{\mathcal{V}}(z_{\mathcal{S}}) = \alpha \sum_{i \in \mathcal{S}} z_i + \beta \sum_{\{i \in \mathcal{S}, j \in \mathcal{S}: i \sim j\}} z_i z_j$$

and it has energy function conditional on $\mathcal{S} \backslash \mathcal{B}$

$$U_{\mathcal{B}}^{\mathcal{V}}\left(z_{\mathcal{B}} | z_{\mathcal{S} \backslash \mathcal{B}}\right) = \alpha \sum_{i \in \mathcal{B}} z_i + \beta \sum_{\{i \in \mathcal{B}, j \in \mathcal{S}: i \sim j\}} z_i z_j$$

Note 73. In what follows we discuss identifiability matters related to the potential.

Definition 74. The interaction potential $\mathcal{V}$ is said to be normalized with respect to a normalizing reference point $\zeta \in \mathcal{Z}$ if there is $i \in \mathcal{S}$ which for any $z \in \mathcal{Z}^{\mathcal{S}}$ with $z_i = \zeta$ implies that $V_{\mathcal{B}}(z) = 0$ for every $\mathcal{B} \neq \emptyset$.

Note 75. In (5.2), the mapping $\mathcal{V} \to \mathrm{pr}_Z$ is in general non-identifiable because (5.2) can be constructed from a different interaction potential $\tilde{\mathcal{V}} = \{V_{\mathcal{B}} + c : \mathcal{B} \subseteq \mathcal{S}\}$ for any constant c. I.e. $U_{\mathcal{S}}^{\mathcal{V}}(z) = U_{\mathcal{S}}^{\tilde{\mathcal{V}}}(z)$.

Note 76. One way to make $\mathcal{V}$ identifiable is to impose restriction

$$\forall \mathcal{A} \neq \emptyset, \; V_{\mathcal{A}}(z) = 0, \text{ if for some } \; i \in \mathcal{A}, \; z_i = \zeta \tag{5.3}$$

Notation 77. For convenience, consider notation related to $z^{[\mathcal{B},\zeta]}$ such as

$$\left[z^{[\mathcal{B},\zeta]}\right]_i = \begin{cases} \zeta, & \text{if } i \notin \mathcal{B} \\ z_i, & \text{if } i \in \mathcal{B} \end{cases}$$

and $z_{\mathcal{A}}^{[\mathcal{B},\zeta]} = \left(z_s^{[\mathcal{B},\zeta]}; s \in \mathcal{A}\right)$, and $z_s^{[\mathcal{B},\zeta]} = z_{\{s\}}^{[\mathcal{B},\zeta]}$ for some fixed ζ.

Example 78. For instance if $z \in \mathcal{Z}^{\mathcal{S}}$ where $\mathcal{S} = \{1, ..., n\}$ then

$$z^{[\emptyset,\zeta]} = \left(\overbrace{\zeta, ..., \zeta}^{n \text{ times}} \right)^{\top}; \qquad z^{[\{i\},\zeta]} = \left(\zeta, ..., \zeta, \overbrace{z_i}^{i\text{th location}}, \zeta, ..., \zeta \right)^{\top};$$

$$z^{[\{i,j\},\zeta]} = \left(\zeta, ...\zeta, \overbrace{z_i}^{i\text{th location}}, \zeta, ..., \zeta, \overbrace{z_j}^{j\text{th location}}, ..., \zeta \right)^{\top}; \quad z^{[\mathcal{S},\zeta]} = (z_1, ..., z_n)^{\top};$$

Note 79. The following theorem uniquely associates potentials satisfying (5.3) with (5.2) with regards a normalizing point.

Theorem 80. *Let Z be an $\mathcal{Z}$-valued random field on a finite collection $\mathcal{S} \neq \emptyset$ of sites such that $pr_Z(z) > 0$ for all $z \in \mathcal{Z}^{\mathcal{S}}$. Then Z is a Gibbs Random Field with respect to the canonical potential*

$$\begin{aligned} V_{\mathcal{A}}(z_{\mathcal{A}}) &= \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} U_{\mathcal{B}}^{\mathcal{V}}\left(z^{[\mathcal{B},\zeta]}\right), \ z \in \mathcal{Z}^{\mathcal{S}} \\ &= \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} \log\left(pr_Z\left(z^{[\mathcal{B},\zeta]}\right)\right), \ z \in \mathcal{Z}^{\mathcal{S}} \end{aligned} \tag{5.4}$$

where $\zeta \in \mathcal{Z}$ is a fixed value and notation $z^{[\mathcal{B},\zeta]}$ denotes the vector based on $z \in \mathcal{Z}^{\mathcal{S}}$ but modified such that its i-th element is $\left[z^{[\mathcal{B},\zeta]}\right]_i = z_i$ if $i \in \mathcal{B}$ and $\left[z^{[\mathcal{B},\zeta]}\right]_i = \zeta$ if $i \notin \mathcal{B}$. This is the unique normalized potential w.r.t $\zeta \in \mathcal{Z}$.

Proof. The proof is is based on Möbius inversion formula, and hence out of scope. □

Corollary 81. *From Theorem 80, for all $i \in \mathcal{A}$ it is*

$$V_{\mathcal{A}}(z_{\mathcal{A}}) = \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} \log\left(pr_i\left(z_i^{[\mathcal{B},\zeta]} | z_{\mathcal{S} \setminus \{i\}}^{[\mathcal{B},\zeta]}\right)\right), \ z \in \mathcal{Z}^{\mathcal{S}} \tag{5.5}$$

Note 82. The following example explains the use of Theorem 80 in terms of the Definition 64.

Example 83. Consider $\mathcal{S} = \{1, 2\}$. Let $z = (z_1, z_2)^{\top}$. Consider a fixed $\zeta \in \mathcal{Z}$. Then $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\} = \left\{V_{\{1\}}, V_{\{2\}}, V_{\{1,2\}}\right\}$. The decomposition of the energy $U\left(z = (z_1, z_2)^{\top}\right) := U_{\mathcal{S}}^{\mathcal{V}}(z)$ is written as

$$U(z_1, z_2) - U(\zeta, \zeta) = V_{\{1\}}(z_1) + V_{\{2\}}(z_2) + V_{\{1,2\}}(z_1, z_2)$$

by using (5.1) with

$$V_{\{1\}}(z_1) = U(z_1, \zeta) - U(\zeta, \zeta)$$
$$V_{\{2\}}(z_2) = U(\zeta, z_2) - U(\zeta, \zeta)$$
$$V_{\{1,2\}}(z_1, z_2) = U(z_1, z_2) - U(z_1, \zeta) - U(\zeta, z_2) + U(\zeta, \zeta)$$

by (5.4).

Example 84. (Ising model) We revisit Example 17 where

$$\mathrm{pr}_Z(z) \propto \exp\left(\alpha \sum_{i \in \mathcal{S}} z_i + \beta \sum_{\{i,j\}: i \sim j} z_i z_j\right), \ \forall z \in \mathcal{Z}^{\mathcal{S}}$$

Consider Notation 77, for instance,

$$z^{[\emptyset,\zeta]} = \left(\overbrace{\zeta, ..., \zeta}^{n \text{ times}}\right)^\top; \qquad z^{[\{i\},\zeta]} = \left(\zeta, ..., \zeta, \overbrace{z_i}^{i\text{th location}}, \zeta, ..., \zeta\right)^\top;$$

$$z^{[\{i,j\},\zeta]} = \left(\zeta, ...\zeta, \overbrace{z_i}^{i\text{th location}}, \zeta, ..., \zeta, \overbrace{z_j}^{j\text{th location}}, ..., \zeta\right)^\top; \quad z^{[\mathcal{S},\zeta]} = (z_1, ..., z_n)^\top;$$

It is $V_\emptyset = 0$ by definition. By using Theorem 80 and considering a reference point $\zeta = 0$, we get

$$V_{\{i\}}(z) = (-1)^{1-1} U\left(z^{[\{i\},\zeta]}\right) + (-1)^{1-0} U\left(z^{[\emptyset,\zeta]}\right) = a z_i, \tag{5.6}$$

for any $i \in \mathcal{S}$ and

$$\begin{aligned} V_{\{i,j\}}(z) &= \left[(-1)^{2-2} U\left(z^{[\{i,j\},\zeta]}\right)\right] + \left[(-1)^{2-1} U\left(z^{[\{i\},\zeta]}\right)\right] \\ &\quad + \left[(-1)^{2-1} U\left(z^{[\{j\},\zeta]}\right)\right] + \left[(-1)^{2-0} U\left(z^{[\emptyset,\zeta]}\right)\right] \\ &= [\alpha z_i + \alpha z_j + \beta z_i z_j] + [-\alpha z_i] + [-\alpha z_j] + [0] = \beta z_i z_j \end{aligned} \tag{5.7}$$

for any $i, j \in \mathcal{S}$, with $i \sim j$. Obviously, it is $V_{\{i,j\}}(z) = 0$ for any $i, j \in \mathcal{S}$, with $i \not\sim j$; and it is $V_{\mathcal{A}}(z) = 0$ for $\mathrm{card}(\mathcal{A}) > 2$.

6. Markov Random Fields

Note 85. Regarding spatial modeling, $\sim$ can describe adjacent sites which is in accordance to the spatial statistics "dogma" that *near things are more related than distant things.* Also it may be computationally convenient for big data problems (large number of sites) as it introduces sparsity and allows specialized numerical algorithms to be implemented.

Note 86. Markov Random Fields constrain the problem such that the conditional distribution of the label at some site i given those at all other sites $j \in \mathcal{S} - \{i\}$ depends only on the labels at neighbors of site i.

Example 87. Recall the Ising model in Example 84 whose sites are equipped with a symmetric relation "$\sim$". It's potentials $V_{\mathcal{A}}$ are non-zero only when $\mathcal{A}$ is a pair of sites $\{i, j\}$ satisfying the relation $\sim$ (5.7) or when $\mathcal{A}$ a singleton (5.6). Consequently, its local characteristics $\mathrm{pr}_i\left(z_i | z_{\mathcal{S}\setminus\{i\}}\right)$ depend only on the values of the sites $j \in \mathcal{S}\setminus\{i\}$ that satisfy $\sim$.

Definition 88. We define as the boundary of $\mathcal{A}$, $\mathcal{A} \subseteq \mathcal{S}$, for a given relation $\sim$ the set

$$\partial\mathcal{A} = \{s \in \mathcal{S}\setminus\mathcal{A} : \exists t \in \mathcal{A} \,\text{s.t.}\, s \sim t\}$$

Definition 89. Let $\partial\mathcal{A}$ be the boundary of $\mathcal{A} \subseteq \mathcal{S}$ for a symmetric relation $\sim$ the finite set $\mathcal{S} \neq \emptyset$. $Z = (Z_s; s \in \mathcal{S})$ is a Markov random field on $\mathcal{S}$ taking values in $\mathcal{Z}$ with respect to the symmetric relation $\sim$ if for each $\mathcal{A} \subset \mathcal{S}$ and $Z_{\mathcal{A}\setminus\mathcal{S}} \in \mathcal{Z}_{\mathcal{A}\setminus\mathcal{S}}$ the distribution of Z on $\mathcal{A}$ conditional on $Z_{\mathcal{A}\setminus\mathcal{S}}$ only depends on $Z_{\partial\mathcal{A}}$ (i.e. the configuration of Z on the neighborhood boundary of $\mathcal{A}$) i.e.

$$\mathrm{pr}_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\setminus\mathcal{A}}\right) = \mathrm{pr}_Z\left(z_{\mathcal{A}} | z_{\partial\mathcal{A}}\right) \tag{6.1}$$

when $\mathrm{pr}_Z\left(z_{\mathcal{S}\setminus\mathcal{A}}\right) > 0$.

Note 90. Definition 89 implies that (6.1) becomes

$$\mathrm{pr}_Z\left(z_i | z_{-i}\right) = \mathrm{pr}_Z\left(z_i | z_{\partial\{i\}}\right), \quad \forall i \in \mathcal{S} \tag{6.2}$$

when $\mathrm{pr}_Z\left(z_{\mathcal{S}\setminus\{i\}}\right) > 0$

Definition 91. A non-empty subset $\mathcal{C}$, $\mathcal{C} \subset \mathcal{S}$, is a clique in $\mathcal{S}$ with respect to $\sim$ if for all $s, t \in \mathcal{C}$ with $s \neq t$ it is $s \sim t$ or if $\mathcal{C}$ is a singleton set.

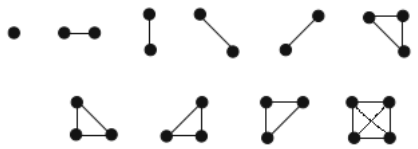


FIGURE 6.1. Examples of cliques

Notation 92. The set containing all the cliques in a lattice of sites in $\mathcal{S}$ equipped with a relation $\sim$ will be usually denoted as bold $\boldsymbol{\mathcal{C}}$.

Note 93. The following theorem shows that the distribution of any Markov random field such that $\mathrm{pr}_Z(z) > 0$ can be expressed in terms of interactions between neighbors.

Theorem 94. *(Hammersley–Clifford) Let $Z = (Z_s; s \in \mathcal{S})$ be an $\mathcal{Z}$-valued random field on a finite collection $\mathcal{S} \neq \emptyset$ of sites such that $pr_Z\left(z_{\mathcal{A}}|z_{\mathcal{C}\setminus\mathcal{A}}\right) > 0$ for all $\mathcal{A} \subset \mathcal{S}$ and $z \in \mathcal{Z}^{\mathcal{S}}$. Let $\sim$ be a symmetric relation on $\mathcal{S}$. Then Z is a Markov Random Field with respect to $\sim$ if and only if*

$$pr_Z(z) \propto \prod_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} \varphi_{\mathcal{C}}(z_{\mathcal{C}}) \tag{6.3}$$

for some interaction functions $\varphi_{\mathcal{C}} : \mathcal{Z}^{\mathcal{C}} \to \mathbb{R}^+$ defined on cliques $\mathcal{C} \in \boldsymbol{\mathcal{C}}$.

Proof. □

For convenience, let $\left[z^{\mathcal{B},\delta}\right]_i = \begin{cases} \delta, & \text{if } i \notin \mathcal{B} \\ z_i, & \text{if } i \in \mathcal{B} \end{cases}$, *and* $z_{\mathcal{A}}^{\mathcal{B},\delta} = \left(z_s^{\mathcal{B},\delta}; s \in \mathcal{A}\right)$, *and* $z_s^{\mathcal{B},\delta} = z_{\{s\}}^{\mathcal{B},\delta}$.

for $\Longrightarrow$: *By Theorem 80, Z is Gibbs with a canonical potential (5.4)*

$$V_{\mathcal{A}}(z_{\mathcal{A}}) = \sum_{\mathcal{A}\subseteq\mathcal{B}} (-1)^{Card(\mathcal{A}\setminus\mathcal{B})} \log\left(pr_Z\left(z^{[\mathcal{B},\zeta]}\right)\right),$$

for $z \in \mathcal{Z}^{\mathcal{S}}$. We need to show that for all $\mathcal{A}$ which are not a cliques, $\mathcal{A} \notin \boldsymbol{\mathcal{C}}$.

Assume a set $\mathcal{A}$ with $\mathcal{A} \subseteq \mathcal{S}$ which is not a clique, $\mathcal{A} \notin \boldsymbol{\mathcal{C}}$, there are two distinct sites $s, t \in \mathcal{A}$ with $s \not\sim t$. Then,

$$\begin{aligned}
V_{\mathcal{A}}(z) &= \sum_{\mathcal{B}\subseteq\mathcal{A}} (-1)^{Card(\mathcal{A}\setminus\mathcal{B})} \log\left(pr_Z\left(z_s^{\mathcal{B},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B},\delta}\right)\right) \\
&= \sum_{\mathcal{B}\subseteq\mathcal{A}\setminus\{s,t\}} (-1)^{Card(\mathcal{A}\setminus\mathcal{B})} \log\left(pr_Z\left(z_s^{\mathcal{B},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B},\delta}\right)\right) \\
&\quad + \sum_{\mathcal{B}\subseteq\mathcal{A}\setminus\{s,t\}} (-1)^{Card(\mathcal{A}\setminus(\mathcal{B}\cup\{s\}))} \log\left(pr_Z\left(z_s^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s\},\delta}\right)\right) \\
&\quad + \sum_{\mathcal{B}\subseteq\mathcal{A}\setminus\{s,t\}} (-1)^{Card(\mathcal{A}\setminus(\mathcal{B}\cup\{t\}))} \log\left(pr_Z\left(z_s^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{t\},\delta}\right)\right) \\
&\quad + \sum_{\mathcal{B}\subseteq\mathcal{A}\setminus\{s,t\}} (-1)^{Card(\mathcal{A}\setminus(\mathcal{B}\cup\{s,t\}))} \log\left(pr_Z\left(z_s^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s,t\},\delta}\right)\right)
\end{aligned}$$

Rearranging I get simplifies

$$V_{\mathcal{A}}(z) = \sum_{\mathcal{B}\subseteq\mathcal{A}} (-1)^{Card(\mathcal{A}\setminus\mathcal{B})} \log\left(\frac{pr_Z\left(z_s^{\mathcal{B},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B},\delta}\right)}{pr_Z\left(z_s^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{t\},\delta}\right)} \frac{pr_Z\left(z_s^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s,t\},\delta}\right)}{pr_Z\left(z_s^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s\},\delta}\right)}\right)$$

Because $s \not\sim t$, it is $pr_Z\left(z_s^{\mathcal{B},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B},\delta}\right) = pr_Z\left(z_s^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{t\},\delta}\right)$ and $pr_Z\left(z_s^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s,t\},\delta}\right) = pr_Z\left(z_s^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\setminus s}^{\mathcal{B}\cup\{s\},\delta}\right)$. This implies $V_{\mathcal{A}}(z) = 0$ for any subset $\mathcal{A}$ with $\mathcal{A} \subseteq \mathcal{S}$ which is not a clique. Hence (6.3) holds.

for $\Longleftarrow$: *By using (5.2), I can write*

$$pr_Z\left(z_{\mathcal{A}}|z_{\mathcal{S}\setminus\mathcal{A}}\right)=\frac{1}{C_{\mathcal{A}}\left(z_{\mathcal{S}\setminus\mathcal{A}}\right)}\exp\left(U_{\mathcal{A}}\left(z\right)\right)$$

where

$$U_{\mathcal{A}}\left(z\right)=\sum_{\{\mathcal{C}\subseteq\mathcal{S}\,:\,\mathcal{A}\cap\mathcal{C}\neq\emptyset\}}V_{\mathcal{C}}\left(z_{\mathcal{C}}\right)$$

depends only on $\{z_i : i\in\mathcal{A}\cup\partial\mathcal{A}\}$ *as* $pr_Z\left(\cdot\right)$ *is a Markov Random Field.*

Note 95. Because $\mathrm{pr}_Z\left(z\right)>0$, the Markov Random Field in (6.3) is a Gibbs Random Field as

$$\mathrm{pr}_Z\left(z\right)\propto\exp\left(\sum_{\mathcal{C}\in\boldsymbol{\mathcal{C}}}\log\left(\varphi_{\mathcal{C}}\left(z_{\mathcal{C}}\right)\right)\right)$$

with non-zero interaction potentials restricted to cliques $\mathcal{C}\in\boldsymbol{\mathcal{C}}$.

Note 96. Essentially Theorem 94 gives guidelines on using Markov RF and Gibbs RF that:

for $\Longrightarrow$: we need to show that there exists an interaction potential $\boldsymbol{\varphi}=\{\varphi_{\mathcal{C}}:\mathcal{C}\in\boldsymbol{\mathcal{C}}\}$ defined on the cliques $\boldsymbol{\mathcal{C}}$ such that $\mathrm{pr}_Z\left(\cdot\right)$ is a Gibbs Random Field with iteration potential $\boldsymbol{\varphi}$.

for $\Longleftarrow$: a Gibbs Random Field with potentials $\{\varphi_{\mathcal{C}}:\mathcal{C}\in\boldsymbol{\mathcal{C}}\}$ defined on the cliques $\boldsymbol{\mathcal{C}}$ is a Markov Random Field.

Example 97. (Ising model; Cont. Example 17). The joint PMF of the Ising model in Example 17 is

$$\begin{aligned}\mathrm{pr}\left(z\right)=&\frac{\exp\left(\alpha\sum_{i\in\mathcal{S}}z_i+\beta\sum_{\{i,j\}:i\sim j}z_iz_j\right)}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}}z_i+\beta\sum_{\{i,j\}:i\sim j}z_iz_j\right)}\\=&\frac{1}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}}z_i+\beta\sum_{\{i,j\}:i\sim j}z_iz_j\right)}\prod_{i\in\mathcal{S}}\exp\left(\alpha z_i\right)\prod_{i\in\mathcal{S}}\prod_{j:j\sim i}\exp\left(\beta z_iz_j\right)\end{aligned}$$

I can find that

$$\varphi_{\emptyset}=1/\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}}z_i+\beta\sum_{\{i,j\}:i\sim j}z_iz_j\right)$$

$$\varphi_{\{i\}}\left(z_{\{i\}}\right)=\exp\left(\alpha z_i\right),\qquad\forall i\in\mathcal{S} \tag{6.4}$$

$$\begin{aligned}\varphi_{\{i,j\}}\left(z_{\{i,j\}}\right)&=\exp\left(\beta z_iz_j\right), && \forall i,j\in\mathcal{S}\quad\text{s.t.}\quad i\sim j\\ \varphi_{\{i,j\}}\left(z_{\{i,j\}}\right)&=1, && \forall i,j\in\mathcal{S}\quad\text{s.t.}\quad i\not\sim j\\ \varphi_{\mathcal{A}}\left(z_{\mathcal{A}}\right)&=1, && \forall\mathcal{A}\subset\mathcal{S}\quad\text{s.t.}\quad\mathrm{card}\left(\mathcal{A}\right)>2\end{aligned} \tag{6.5}$$

where $\{i\}$ and $\{i,j\}$ satisfying $i \sim j$ are cliques. Alternatively, as $\emptyset$ is not a clique if that $\varphi_\emptyset$ is just the constant term which can be absorbed by (6.4) and (6.5) and correspond to cliques.

Part 2. Model building for aerial data & related inference

7. Automodels

Note 98. We introduce a general class of models, the automodels and their special case Besag's automodels, which are associated to the exponential family of distributions and able to represent spatial dependence.

Definition 99. A random variable X taking values in $\mathcal{X}$ follows an exponential family labeled by parameter $\theta \in \Theta$ if the associated PMF/PDF $\mathrm{pr}_X(x|\theta)$ can be expressed in the form

$$\mathrm{pr}_X(x|\theta) = \exp\left(A(\theta)^\top B(x) + C(x) + D(\theta)\right),\ \forall x \in \mathcal{X}$$

where $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, and $D(\cdot)$ are known functions.

7.1. Multi-parameter automodels.

Theorem 100. *Consider Markov random field $Z = (Z_s; s \in \mathcal{S})$ that takes values in $\mathcal{Z}$ on a finite set of points $\mathcal{S}$ and has energy function $U(\cdot)$, Assume that the following assumptions are satisfied with some fixed normalization configuration $\zeta = (\zeta, ..., \zeta)^\top \in \mathcal{Z}^{\mathcal{S}}$:*

(1) In the energy function $U(\cdot)$ the dependence between the sites is pairwise only, i.e.

$$U(z) = \sum_{i \in \mathcal{S}} V_i(z_i) + \sum_{\{\{i,j\} \in \mathcal{S}^2 : i \sim j\}} V_{i,j}(z_i, z_j),\ z \in \mathcal{Z}^{\mathcal{S}}$$

with $V_i(\zeta) = V_{i,j}(z_i, \zeta) = V_{i,j}(\zeta, z_j) = 0$ for all $i, j \in \mathcal{S}$.

(2) For all $i \in \mathcal{S}$, the conditional distributions (characteristics) are such that

$$\log\left(pr_i(z_i|z_{-i})\right) = \left(A_i(z_{-i})\right)^\top B_i(z_i) + C_i(z_i) + D_i(z_{-i}), \tag{7.1}$$

where $A_i(z_{-i}) \in \mathbb{R}^\ell$, $B_i(z_i) \in \mathbb{R}^\ell$, for $\ell \geq 1$ and $C_i(z_i) \in \mathbb{R}$, and $D_i(z_{-i}) \in \mathbb{R}$ with $C_i(\zeta) = 0$ and $B_i(\zeta) = 0$.

(3) For all $i \in \mathcal{S}$, span$\{B_i(z_i); z_i \in \mathcal{Z}\} = \mathbb{R}^\ell$, for $\ell \geq 1$.

Then,

(1) the functions $A_i(z_{-i}) \in \mathbb{R}^\ell$ take the form

$$A_i(z_{-i}) = \alpha_i + \sum_{i \neq j} \beta_{i,j} B_j(z_j),\ i \in \mathcal{S}$$

where $\{\alpha_i; i \in \mathcal{S}\}$ *is a family of* ℓ*-dimensional vectors, and* $\{\beta_{i,j}; i,j \in \mathcal{S}, i \neq j\}$ *is a family of* $\ell \times \ell$ *symmetric matrices, and*

(2) the potentials are given by

$$V_i(z_i) = (\alpha_i)^\top B_i(z_i) + C_i(z_i) \tag{7.2}$$

$$V_{i,j}(z_i, z_j) = (B_i(z_i))^\top \beta_{i,j} B_j(z_j) \tag{7.3}$$

Proof. Omitted, but can be found in

(1) Hardouin, C., & Yao, J. F. (2008). Multi-parameter automodels and their applications. Biometrika, 95(2), 335-349.

(2) Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 192-225.

□

Theorem 101. *Consider Markov random field* $Z = (Z_i; i \in \mathcal{S})$ *that takes values in* $\mathcal{Z}$ *on a finite set of points* $\mathcal{S}$ *and has energy function* $U(\cdot)$*, Assume that the following assumptions are satisfied with some fixed normalization configuration* $\zeta = (\zeta, ..., \zeta)^\top \in \mathcal{Z}^{\mathcal{S}}$ *:*

(1) energy function $U(\cdot)$ *involves only pairwise dependence between the sites, i.e.*

$$U(z) = \sum_{i \in \mathcal{S}} V_i(z_i) + \sum_{\{\{i,j\} \in \mathcal{S}^2 : i \sim j\}} V_{i,j}(z_i, z_j), \ z \in \mathcal{Z}^{\mathcal{S}}$$

with potentials

$$V_i(z_i) = (\alpha_i)^\top B_i(z_i) + C_i(z_i) \tag{7.4}$$

$$V_{i,j}(z_i, z_j) = (B_i(z_i))^\top \beta_{i,j} B_j(z_j) \tag{7.5}$$

and $V_i(\zeta) = V_{i,j}(z_i, \zeta) = V_{i,j}(\zeta, z_j) = 0$ *for all* $i, j \in \mathcal{S}$*.*

(2) energy function $U(\cdot)$ *is admissible; i.e.*

$$\int \exp(U(z))\, dz < \infty$$

Then,

(1) the family of conditional distributions $pr_i(z_i | z_{-i})$ *belongs to a multiparameter exponential family distributions such as*

$$\log(pr_i(z_i | z_{-i})) = (A_i(z_{-i}))^\top B_i(z_i) + C_i(z_i) + D_i(z_{-i}), \tag{7.6}$$

whose natural parameters $A_i(z_{-i}) \in \mathbb{R}^\ell$ *take the form*

$$A_i(z_{-i}) = \alpha_i + \sum_{i \neq j} \beta_{i,j} B_j(z_j), \ i \in \mathcal{S}$$

where $\{\alpha_i; i \in \mathcal{S}\}$ *is a family of* ℓ*-dimensional vectors, and* $\{\beta_{i,j}; i,j \in \mathcal{S}, i \neq j\}$ *is a family of* $\ell \times \ell$ *symmetric matrices.*

Proof. Omitted, but can be found in

(1) Hardouin, C., & Yao, J. F. (2008). Multi-parameter automodels and their applications. Biometrika, 95(2), 335-349.

(2) Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 192-225.

□

Definition 102. Automodel is called the model satisfying the assumptions of Theorem 100.

Definition 103. Univariate automodel is the automodel with $\ell = 1$ in Theorem 100.

Definition 104. Multi-parameter is the automodel with $\ell > 1$ in Theorem 100.

Remark 105. In the univariate automodel, $\ell = 1$, assumption 3 in Theorem 100 is not needed; it is automatically satisfied as B_i's are not identically zero. Yet, for $\ell = 1$, (7.2) and (7.3) become

$$V_i(z_i) = \alpha_i B_i(z_i) + C_i(z_i) \tag{7.7}$$

$$V_{i,j}(z_i, z_j) = \beta_{i,j} B_i(z_i) B_j(z_j) \tag{7.8}$$

7.2. **Besag auto-models.**

Definition. Random field $Z = (Z_s; s \in \mathcal{S})$ follows a Besag's auto-model if Z is real-valued and its joint distribution $\mathrm{pr}_Z(z)$ is given by

$$\mathrm{pr}_Z(z) = \frac{1}{C} \exp\left(\sum_{i \in \mathcal{S}} V_i(z_i) + \sum_{\{i,j\} \in \mathcal{S}^2 : i \sim j} \beta_{i,j} z_i z_j \right),\ z \in \mathcal{Z}^{\mathcal{S}} \tag{7.9}$$

with $\beta_{i,j} = \beta_{j,i}$ for all $i, j \in \mathcal{S}$.

Note 106. The following allows us to define a Markov Random Field model from a set of conditional distributions (characteristics) whose compatibility is automatically satisfied.

Proposition 107. *If each of the*

$$pr_i(z_i | z_{-i}), \quad for \quad i \in \mathcal{S}$$

is a family of real-valued $z_i \in \mathbb{R}$ *conditional distributions which are members of the exponential family of distributions (7.1) with* $B_i(z_i) = z_i$ *for* $i \in \mathcal{S}$*, then they are compatible a Besag's auto-model with distribution (7.9) if* $\beta_{i,j} = \beta_{j,i}$ *for all* $i, j \in \mathcal{S}$*.*

Proof. For

$$\text{pr}_i\left(z_i|z_{-i}\right)=\exp\left(A_i\left(z_{-i}\right)z_i+C_i\left(z_i\right)+D_i\left(z_{-i}\right)\right)$$

it is

$$\begin{aligned}V_i\left(z_i\right)=&\alpha_i B_i\left(z_i\right)+C_i\left(z_i\right)=\alpha_i z_i+C_i\left(z_i\right)\\V_{i,j}\left(z_i,z_j\right)=&\beta_{i,j}B_i\left(z_i\right)B_j\left(z_j\right)=\beta_{i,j}z_i z_j\end{aligned}$$

so

$$\text{pr}_Z\left(z\right)\propto\exp\left(\sum_i\left[\alpha_i z_i+C_i\left(z_i\right)\right]+\sum_{i\sim j}\beta_{i,j}z_i z_j\right),\ z\in\mathcal{Z}^{\mathcal{S}}$$

□

Example 108. (Logistic automodel / Ising model) Consider that $Z\left(s\right)$ represents presence or absence of a characteristic at location $s\in\mathcal{S}$. Mathematically, assume random field Z taking values on a set of indices $\mathcal{S}$ in $\mathcal{Z}=\{0,1\}$ on $\mathcal{S}=\{1,...,n\}$, $n\in\mathbb{N}-\{0\}$.

Consider that for a given z_{-i} it is

$$z_i|z_{-i}\sim\text{Logit}\left(\theta_i\left(z_{-i}\right)\right),\quad i\in\mathcal{S}.$$

Hint:: The PMF of distribution $x|\theta\sim\text{Logit}\left(\theta\right)$ can be written as $\text{pr}\left(x|\theta\right)=\frac{\exp(x\theta)}{1+\exp(\theta)}1\left(x\in\{0,1\}\right)$. Then the characteristics are

$$\text{pr}_i\left(z_i|z_{-i}\right)=\frac{\exp\left(z_i\theta_i\left(z_{-i}\right)\right)}{1+\exp\left(\theta_i\left(z_{-i}\right)\right)}1\left(z_i\in\{0,1\}\right)\tag{7.10}$$

Now, let's parameterize $\{\theta_i\left(\cdot\right)\}$ as

$$\theta_i\left(z_{-i}\right)=\alpha_i+\sum_{j:j\sim i}\beta_{i,j}z_j\tag{7.11}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j}=\beta_{j,i}$. Then (7.10) becomes

(7.12)

$$\log\left(\text{pr}_i\left(z_i|z_{-i}\right)\right)=\underbrace{z_i}_{B_i(z_i)}\underbrace{\left(\alpha_i+\sum_{j\sim i}\beta_{i,j}\overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})}+\underbrace{0}_{C_i(z_i)}+\underbrace{\left(-\log\left(1+\exp\left(\alpha_i+\sum_{j:j\sim i}\beta_{i,j}z_j\right)\right)\right)}_{D_i(z_{-i})}$$

Notice that all the conditionals $z_i|z_{-i}$ follow an Exponential family with

$$\begin{aligned} A_i\left(z_{-i}\right) =& \alpha_i + \sum_{j:j\sim i} \beta_{i,j} B_i\left(z_j\right) \\ B_i\left(z_i\right) =& z_i \\ C_i\left(z_i\right) =& 0 \\ D_i\left(z_{-i}\right) =& -\log\left(1+\exp\left(\alpha_i + \sum_{j:j\sim i}\beta_{i,j} z_j\right)\right) \end{aligned}$$

Also, I can get $C_i(\zeta) = 0$ and $B_i(\zeta) = 0$ by considering a reference point $\zeta = 0$. From Theorem 107, (7.10) with (7.11), the conditionals $z_i|z_{-i}$ are compatible as a Besag automodel with marginal distribution

$$\text{pr}_Z(z) \propto \exp\left(\overbrace{\sum_i \alpha_i \underbrace{z_i}_{B_i(z_i)} + \sum_i \underbrace{\sum_{j:j\sim i}}_{\sum_{\{i,j\}:j\sim i}} \beta_{i,j} z_i z_j}^{U(z)=}\right) \tag{7.13}$$

I observe that:

- Here the Ising model has spatially dependent coefficients $\{\alpha_i\}$ and $\{\beta_{i,j}\}$, unlike the Ising model in Example 17 where we considered $\{\alpha_i = \alpha\}$ and $\{\beta_{i,j} = \beta\}$.
- When $\beta_{i,j} = 0$, for all j such as $j \sim i$, it is $\text{pr}_i(z_i|z_{-i}) = \frac{\exp(z_i\alpha_i)}{1+\exp(\alpha_i)}$.
- Characteristic's present at site i is encouraged in neighboring site j when $\beta_{i,j} > 0$, and discouraged when $\beta_{i,j} < 0$.

The resulting spatial model is called Logistic automodel or Ising model (the latter name is from physics).

Example 109. (Poisson automodel) Consider that $Z(s)$ represents counts at location $s \in \mathcal{S}$. Mathematically we can consider Z taking values in $\mathcal{Z} = \mathbb{N}$ on a set of sites $\mathcal{S} = \{1, ..., n\}$, where $n \in \mathbb{N} - \{0\}$.

Consider that for a given z_{-i} it is

$$z_i|z_{-i} \sim \text{Poisson}\left(\lambda_i\left(z_{-i}\right)\right)$$

Hint:: The PMF of Poisson distribution $x|\lambda \sim \text{Poisson}(\lambda)$ can be written as

$$\text{pr}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)\, 1(x \in \mathbb{N})$$

with mean $\text{E}(x|\lambda) = \lambda$.

Then the full conditionals (characteristics) are

$$\text{pr}_i\left(z_i|z_{-i}\right)=\frac{1}{z_i!}\left(\lambda_i\left(z_{-i}\right)\right)^{z_i}\exp\left(-\lambda\right)1\left(z_i\in\mathbb{N}\right) \tag{7.14}$$

Now, let's parameterize $\{\lambda_i(\cdot)\}$ as

$$\log\left(\lambda_i\left(z_{-i}\right)\right)=\alpha_i+\sum_{j:j\sim i}\beta_{i,j}z_j \tag{7.15}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j}=\beta_{j,i}$. So (7.14) becomes

$$\log\left(\text{pr}_i\left(z_i|z_{-i}\right)\right)=\underbrace{z_i}_{B_i(z_i)}\underbrace{\left(\alpha_i+\sum_{j\sim i}\beta_{i,j}\overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})}+\underbrace{\log\left(z_i!\right)}_{C_i(z_i)}+\underbrace{0}_{D_i(z_{-i})}$$

with

$$\begin{aligned}A_i\left(z_i\right)=&\alpha_i+\sum_{j\sim i}\beta_{i,j}B_i\left(z_j\right)\\B_i\left(z_{-i}\right)=&z_i\\C_i\left(z_i\right)=&\log\left(z_i!\right)\\D_i\left(z_{-i}\right)=&0\end{aligned}$$

I can notice that all the conditionals $z_i|z_{-i}$ follow exponential of exponential. Also, I can get $C_i(\zeta)=0$ and $B_i(\zeta)=0$ by considering a reference point $\zeta=0$. From Theorem 107, (7.14) with (7.15), the the conditionals $z_i|z_{-i}$ are compatible as a Besag auto-model with marginal distribution

$$\text{pr}_Z\left(z\right)\propto\exp\left(\overbrace{\sum_i\left(\underbrace{\alpha_i\overbrace{z_i}^{B_i(z_j)}+\overbrace{\log\left(z_i!\right)}^{C_i(z_i)}}_{V_i(z_i)}\right)+\sum_i\sum_{j:j\sim i}\beta_{i,j}z_iz_j}^{U(z)=}\right)$$

or otherwise the energy function is

$$U\left(z\right)=\sum_i\left(\alpha_iz_i+\log\left(z_i!\right)\right)+\sum_{j\sim i}\beta_{i,j}z_iz_j$$

Furthermore, to ensure that $U(z)$ is admissible, we need to consider additional conditions. I observe that

$$\sum_{z\in\mathbb{N}^S}\exp\left(U\left(z\right)\right)=\sum_{z\in\mathbb{N}^S}\prod_i\left(\exp\left(\alpha_iz_i+\log\left(z_i!\right)\right)+\sum_{j\sim i}\beta_{i,j}z_iz_j\right)$$

- If we use additional condition $\beta_{i,j} \leq 0$ then

$$\sum_{z \in \mathbb{N}^{\mathcal{S}}} \exp\left(U\left(z\right)\right) \leq \sum_{z \in \mathbb{N}^{\mathcal{S}}} \prod_{i} \left(\exp\left(\alpha_i z_i + \log\left(z_i!\right)\right)\right) = \sum_{z \in \mathbb{N}^{\mathcal{S}}} \prod_{i} \frac{1}{z_i!} \exp\left(\alpha_i z_i\right) < \infty$$

  which converges. Modeling-wise, $\beta_{i,j} < 0$ introduces competition among the neighbors similar to the Ising model. So by introducing a competition such as $\beta_{i,j} \leq 0$ in the model I prevent the count z_i at i to explode.
- If $\beta_{i,j} > 0$, I discourage competition among neighboring sites. Admissibility can be satisfied if we truncate the state space as $z_i < M$ for some fixed upper bound M. For instance, the characteristics $z_i | z_{-i}$ can follow a Poisson distribution truncated at M.

$$\mathrm{pr}_i\left(z_i | z_{-i}\right) = \frac{1}{z_i!} \left(\lambda_i\left(z_{-i}\right)\right)^{z_i} \exp\left(-\lambda\right) 1\left(z_i \in \{0, 1, ..., M\}\right)$$

  So I can prevent z_i at i to explode by forcefully bounding it $z_i < M$ with a big enough value $M > 0$.

The resulting spatial model is called Poisson automodel.

Note 110. A CAR model is an automodel. Recall that CAR model is defined such as its local characteristics (full conditional distributions) are Gaussian distributions; however Gaussian distribution is an exponential distribution family. Hence the joint distribution of CAR model in Proposition 56 could have been derived from Theorem 100 as well.

7.3. **Parameterization matters in automodels.**

Remark 111. The unknown parameter vector $\theta = \left(\left(\alpha_i; i \in \mathcal{S}\right), \left(\beta_{i,j}; i, j \in \mathcal{S}\right)\right)$ in automodels (e.g., Besag's automodel (7.9)) can be further parameterised to have a particular structure without the need to consider any additional constrains in Theorems 100 & 101.

Remark 112. The dimensionality of automodel parameters $\theta = \left(\left(\alpha_i; i \in \mathcal{S}\right), \left(\beta_{i,j}; i, j \in \mathcal{S}\right)\right)$ may be too large leading to an over-parameterized model or prohibitively large computational cost when the size of the set of sites $\mathcal{S}$ is large (a usual case). To mitigate this issue, a way is to set a structure on $\left\{\alpha_i, \beta_{i,j}\right\}_{i,j \in \mathcal{S},}$ reducing its dimentionality.

E.g.: by setting

$$\alpha_i = a w_i, \quad \text{and} \quad \beta_{i,j} = b_i c_j; \text{ for } i, j \in \mathcal{S},$$

with some known weights $\{w_i;\ i \in \mathcal{S},\}$ and unknown $\{a, b_i, c_j;\ i, j \in \mathcal{S}\}$. Then learning $\mathrm{Card}\left(\mathcal{S}\right)\left(1 + \mathrm{Card}\left(\mathcal{S}\right)\right)$ unknown parameter $\{\alpha_i, \beta_{i,j};\ i, j \in \mathcal{S},\}$ reduces to learning just $1 + 2\mathrm{Card}\left(\mathcal{S}\right)$ unknown parameters $\{a, b_i, c_j;\ i, j \in \mathcal{S}\}$. Note, that $\beta_{i,j} = b_i c_j$ restricts the interaction between i, j.

Remark 113. When covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$ are available (such as other characteristics or time), one could "link" them to the model via the parameters $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S}}$, For instance

E.g.: by setting

$$\alpha_i = a_i + \sum_{k=1}^{p} d_k x_{i,k}, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j};\ \text{for } i,j \in \mathcal{S}, \tag{7.16}$$

where $\{a_i;\ i \in \mathcal{S}\}$, $\{d_k; k = 1, ..., k\}$ and $\{\beta_{i,j};\ i,j \in \mathcal{S}\}$ are unknown parameters. d_k represents the influence of k-th covariate $x_{i,k}$, for all $i \in \mathcal{S}$. $\beta_{i,j}$ represents the influence of the $z_{\partial i}$ at the neighboring sites of Z_i. Examination of the sign of $\beta_{i,j}$, and d_k or whether $\beta_{i,j} \neq 0$, $d_k \neq 0$ facilitates the discovery of pasterns and conditional dependencies.

E.g.: it t denotes time, we can make the automodel "dynamic" (aka spatio-temporal) by setting $x_i = (t_i, t_i^2)^\top$ for $i \in \mathcal{S}$ and

$$\alpha_i = a_i + d_1 t_i + d_2 (t_i)^2, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j};\ \text{for } i,j \in \mathcal{S}.$$

Example 114. In Example 109, given observable covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$, one may set (7.15) as

$$\log(\lambda_i(z_{-i})) = \left[a_i + \sum_{k=1}^{p} d_k x_{i,k}\right] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j\right] \tag{7.17}$$

Then d_k represent the influence of k-th covariate $x_{i,k}$, for all $i \in \mathcal{S}$, and $\beta_{i,j}$ represents the influence of the $z_{\partial i}$ at the neighboring sites of Z_i. For admissibility, a condition such as $\beta_{i,j} \leq 0$ should be specified (see Example 109). Further restrictions on the unknown parameters, or dimension reduction techniques, should be used because the number of unknowns is greater than the number of observations in (7.17).

Example 115. In Example 109, if the dataset is $\{(t_i, s_i, Z_i); i \in \mathcal{S}\}$ where Z_i is the measurement (e.g. counts of a characteristic), at time t_i, at location $s_i \in \mathbb{R}^2$ of the i-th observation, a researcher may consider a parametrization

$$\log(\lambda_i(z_{-i}, t_i)) = [a_i + d_1 t_i] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j\right] \tag{7.18}$$

and be interested in learning the unknown parameters $\{a_i\}$, d_1, and $\{\beta_{i,j}\}$. Obviously, the resulted model is space-time.

8. Frequentist modeling and likelihood based inference

Note 116. Consider a dataset $\{(s_i, Z_i = Z(s_i)); i = 1, ..., n\}$ where Z_i is the observation at site s_i for $i = 1, ..., n$. Assume that the the sampling distribution of $(Z_i)_{i=1}^n$ is specified by the researcher to be

$$Z \sim \text{pr}_Z(Z|\theta) \tag{8.1}$$

labeled by unknown parameter vecrtor θ. Parametric and predictive inference can be performed based based on the associated likelihood or its approximation PsuedoLikelihood.

Note 117. For easy of presentation we assume that the observables Z are a realization of an automodel and hence their sampling distribution (8.1) is that of an automodel with potentials 7.2 and 7.3, and unknown parameter $\theta = (\{\alpha_i\}, \{\beta_{i,j}\})^\top$.

8.1. **MLE: Maximum likelihood estimation.**

Note 118. We describe the maximum likelihood estimation in the automodel framework.

Remark 119. In the MLE framework, given a dataset $\{(s_i, Z_i = Z(s_i)); i = 1, ..., n\}$, estimation of the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ of an automodel can be performed by maximizing the likelihood, as

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\text{pr}_Z(Z|\{\alpha_i\}, \{\beta_{i,j}\})\right) \tag{8.2}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i, j \in \mathcal{S}$$

$$\text{...and any other problem specific restrictions}$$

where $\text{pr}_Z(Z|\{\alpha_i\}, \{\beta_{i,j}\})$ is the joint distribution (7.9) given the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$.

Example 120. (Logistic automodel / Ising model) Assume that observables Z follow the Logistic automodel (7.13) in Example 108. Computing MLE $\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}$ of $\{\alpha_i\}, \{\beta_{i,j}\}$ requires

$$\begin{aligned}\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\log\left(\text{pr}_Z(Z|\{\alpha_i\}, \{\beta_{i,j}\})\right)\right) \\ &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j - \log\left(C(\{\alpha_i\}, \{\beta_{i,j}\})\right)\right) \end{aligned} \tag{8.3}$$

where

$$C(\{\alpha_i\}, \{\beta_{i,j}\}) = \sum_{\forall z \in \mathcal{Z}^{\mathcal{S}}} \exp\left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j\right) \tag{8.4}$$

is the normalizing constant. Optimization in (8.3) can be done numerically by using a recursive optimization algorithm such as Newton-Raphson.

Note 121. The optimization problem (8.2) can be too computationally expensive. For instance, in Example 120, a recursive optimization algorithm, like Newton-Raphson, requires several iterations. At each iteration the evaluation of the (parameter dependent) constant (8.4) has to be evaluated. A computation of that constant can be too expensive when the set of sites $i \in \mathcal{S}$ is large because the sum $\sum_{\forall z \in \mathcal{Z}^{\mathcal{S}}}$ in (8.4) implies scanning all the possible configurations of $z \in \mathcal{Z}^{\mathcal{S}}$. A way to mitigate this is to use instead an "approximation" of the likelihood, such as the Pseudo-likelihood.

8.2. **MPLE: Maximum pseudo likelihood estimation.**

Note 122. We describe the maximum pseudo likelihood estimation in the automodel framework.

Definition. The pseudo likelihood pseudo$L(Z;\theta)$ of observables $Z = (Z_1, ..., Z_n)^\top$ given parameters θ is an approximation of the (exact) likelihood $L(Z;\theta)$ of observables $Z = (Z_1, ..., Z_n)^\top$ given parameters θ which is equal to

$$\text{pseudo}L(Z;\theta) = \prod_i \text{pr}(Z_i|Z_{-i},\theta)$$

where $\text{pr}(Z_i|Z_{-i},\theta)$ are the conditionals of the joint pdf/pmf of the sampling distribution $\text{pr}(Z|\theta)$ of Z given parameter θ.

Definition. (Maximum PseudoLikelihood Estimator) The Maximum Pseudo-Likelihood Estimator (MPLE) $\tilde{\theta}$ of θ is the maximizer of the pseudo likelihood function pseudo$L(Z;\theta)$ where the parameter θ is the argument and the observables $Z = (Z_1, ..., Z_n)^\top$ are fixed values.

$$\tilde{\theta} = \arg\max_\theta \left(\text{pseudo}L(Z;\theta)\right)$$

Remark 123. Then (8.2) becomes: In the MPLE framework, given a dataset $\{(s_i, Z_i = Z(s_i)); i = 1, ..., n\}$, e
of the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ of an automodel can be performed by maximizing the pseudo-likelihood, as

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}} \left(\prod_{i\in\mathcal{S}} \text{pr}_Z(Z_i|Z_{-i},\theta)\right) \tag{8.5}$$

$$= \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}} \left(\sum_{i\in\mathcal{S}} \log\left(\text{pr}_Z(Z_i|Z_{-i},\theta)\right)\right) \tag{8.6}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i,j \in \mathcal{S}$$

...and any other problem specific restrictions

Example 124. (Logistic automodel / Ising model) (Cont. Example 120) Assume that observables Z follow the Logistic automodel (7.13) in Example 108. From (7.12), the conditionals (local characteristics) are computed to be such as

$$\log\left(\mathrm{pr}_i\left(z_i|z_{-i}\right)\right) = z_i\left(\alpha_i + \sum_{j\sim i}\beta_{i,j}z_j\right) - \log\left(1+\exp\left(\alpha_i+\beta_{i,j}\sum_{j:j\sim i}\beta_{i,j}z_j\right)\right)$$

and hence

$$\begin{aligned}\left(\left\{\hat{\alpha}_i\right\},\left\{\hat{\beta}_{i,j}\right\}\right) &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max}\left(\sum_{i\in\mathcal{S}}\log\left(\mathrm{pr}_i\left(z_i|z_{-i}\right)\right)\right)\\ &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max}\left(\sum_{i\in\mathcal{S}}z_i\left(\alpha_i+\sum_{j\sim i}\beta_{i,j}z_j\right) - \sum_{i\in\mathcal{S}}\log\left(1+\exp\left(\alpha_i+\beta_{i,j}\sum_{j:j\sim i}\beta_{i,j}z_j\right)\right)\right)\end{aligned}$$

which does not depend on the normalizing constant (8.4) and hence its computation is less computationally demanding.

9. Hierarchical modeling (Bayesian modeling)

9.1. **A general framework for the hierarchical modeling.**

Note 125. Uncertainty in spatial statistics can be decomposed as a Bayesian hierarchical spatial model

$$\begin{cases} Z|Y,\vartheta_1,\vartheta_2 & \text{data model}\\ Y|\vartheta_1,\vartheta_2 & \text{spatial process model}\\ \vartheta_1|\vartheta_2 & \text{hyper-parameter prior model}\end{cases} \tag{9.1}$$

where uncertainty is described by

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

Let $\vartheta = \left(\vartheta_1,\vartheta_2\right)^\top$ be unknown hyper-parameters. Let ϑ_1 and ϑ_2 be unknown random and fixed hyper-parameters.

Data model: expresses the measurement uncertainty as it is quantified via the distribution $\mathrm{pr}\left(Z|Y,\vartheta\right)$ possibly labeled by some parameter ϑ. It is often specified/modeled so that it can measure the goodness of fit between Z and Y.

Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\mathrm{pr}\left(Y|\vartheta\right)$ possibly labeled by some parameter ϑ. It is often specified/modeled with purpose (among others) to encourage spatial coherence and represente spatial dependence.

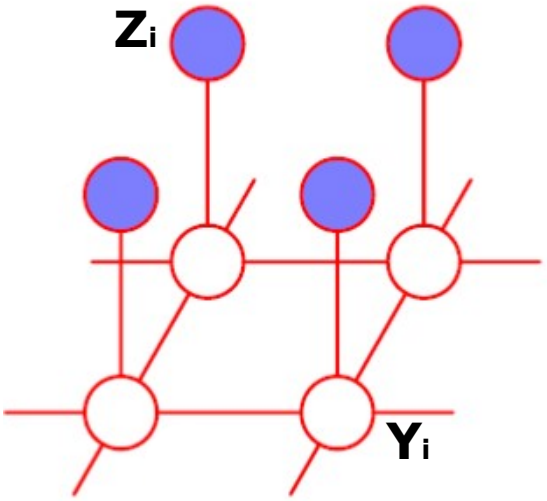


FIGURE 9.1. Hierarchical spatial model structure. $\{Y_i\}$ is the spatial process model which is hidden. $\{Z_i\}$ is the data model. The cartoon depicts a hierarchical spatial model with the special conditional independence structure $Z_i|\{Y_i\},\vartheta \sim \prod_i \text{pr}(Z_i|Y_i,\vartheta)$ and $Y|\vartheta \sim \text{pr}(Y|\vartheta)$

Hyper-parameter prior model: expresses uncertainty about specific unknown model hyper-parameters

See for example Figure 9.1

Note 126. Fixed ϑ_2 can be learned pointwise by computing the ML-II point estimator

$$\hat{\vartheta}_2 = \arg\min_{\vartheta_2} \left(-2\log\left(\text{pr}\left(Z|\vartheta_2\right)\right)\right) \tag{9.2}$$

as the maximizer of the marginal likelihood

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

or by computing the pseudo ML-II point estimator

$$\tilde{\vartheta}_2 = \arg\min_{\vartheta_2} \left(-2\log\left(\prod_i \text{pr}\left(Z_i|Z_{-i},\vartheta_2\right)\right)\right) \tag{9.3}$$

as the maximizer of the pseudo marginal likelihood

$$\text{pseudo}L\left(Z|\vartheta_2\right) = \prod_i \text{pr}\left(Z_i|Z_{-i},\vartheta_2\right)$$

$\tilde{\vartheta}_2$ in (9.3) is a computationally cheaper approximation of the MLE $\hat{\vartheta}_2$ in 9.2.

Note 127. Random ϑ_1 can be learned by computing the posterior pdf/pmf of ϑ_1 given Y and $\vartheta_2 = \hat{\vartheta}_2$

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

where the value $\hat{\vartheta}_2$ (or $\tilde{\vartheta}_2$) is plugged in.

Note 128. General interest lies in computing the posterior distributions of the spatial process model $(Y_i; i \in \mathcal{S})$, (or latent process, or noiseless process) given the data Z

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

Note 129. Below we give two examples in aerial data.

9.2. **Examples.**

9.2.1. *A simplified spatial model for binary data (e.g. Image denoising).*

Example 130. (Image denoising)