Spatio-temporal statistics (MATH4341)

Michaelmas term

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

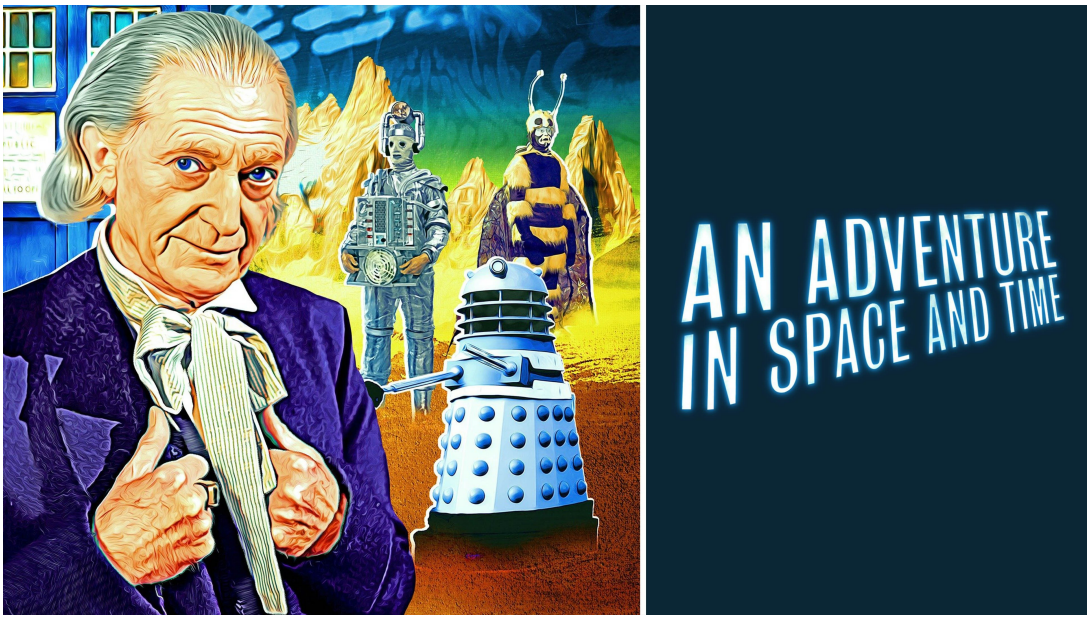
Department of Mathematical Sciences (Office MCS3088)
Durham University
Stockton Road Durham DH1 3LE UK

2023/12/12 at 14:42:02

Concepts

An introduction to spatial statistics:

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



Handouts

1. Handout 1: Types of spatial data
2. Handout 2: Computational methods
3. Handout 3: Point referenced data modeling / Geostatistics
4. Handout 4: Aerial unit data / spatial data on lattices

Reading list

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

Main texts:

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 Classic book in spatial statistics, but a bit outdated.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

Supplementary textbooks for various types of spatial data:

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 Classic book in spatial statistics, and perhaps one of the first, if not the first, textbook in the area, so outdated.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

Supplementary textbooks for Point reference data / Geostatistics:

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

Supplementary textbooks for Areal data:

- TBD

Supplementary textbooks for Point pattern data:

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

Supplementary textbooks for Software:

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

Supplementary textbooks for Theory:

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

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 1: Types of spatial data

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

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

Reading list & references:

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

1. Motivations

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

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

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

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

Shall I ignore spatial dependence? –No!

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

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

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

Spatial data and spatial process.

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

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

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

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

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

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

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

2. Principal spatial statistics areas

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

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

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

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

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

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

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

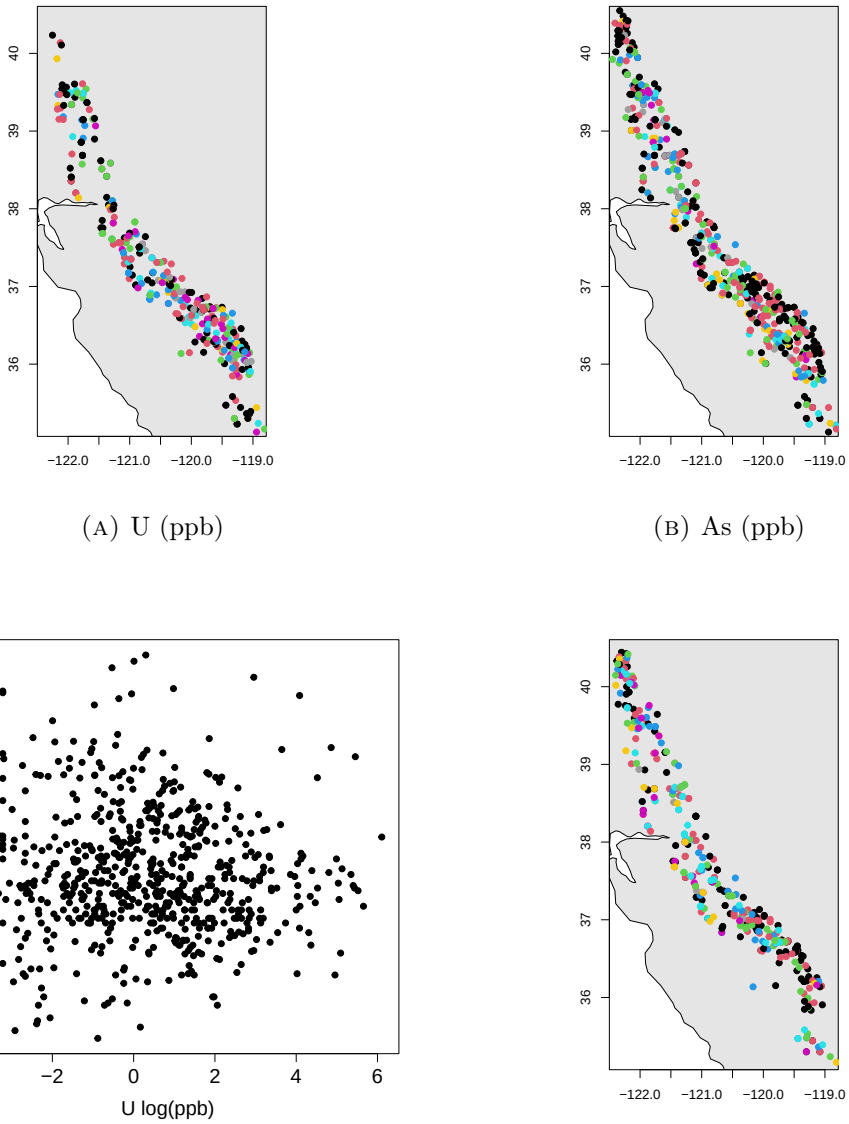
Example 15. (Ground water pollution in the Central Valley of California[1]) California's Central Valley is one of the most productive agricultural regions in the world. With an increase in population, groundwater consumption is expected to increase. Agricultural irrigation heavily draws on the groundwater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

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

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

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

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



(A) U (ppb)

(B) As (ppb)

(C) U vs As (log ppb)

(D) Cr (ppb)

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

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

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

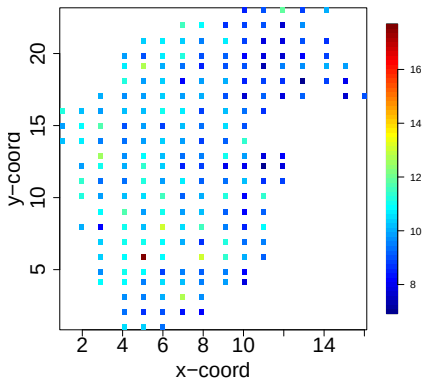


FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

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

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

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

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

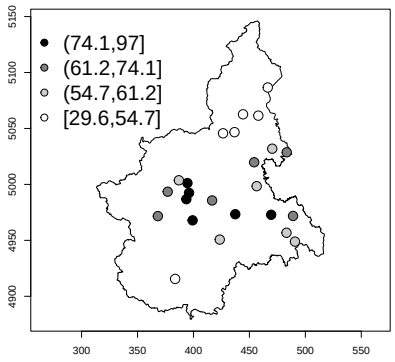


FIGURE 2.3. (Air pollution data) Average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte

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

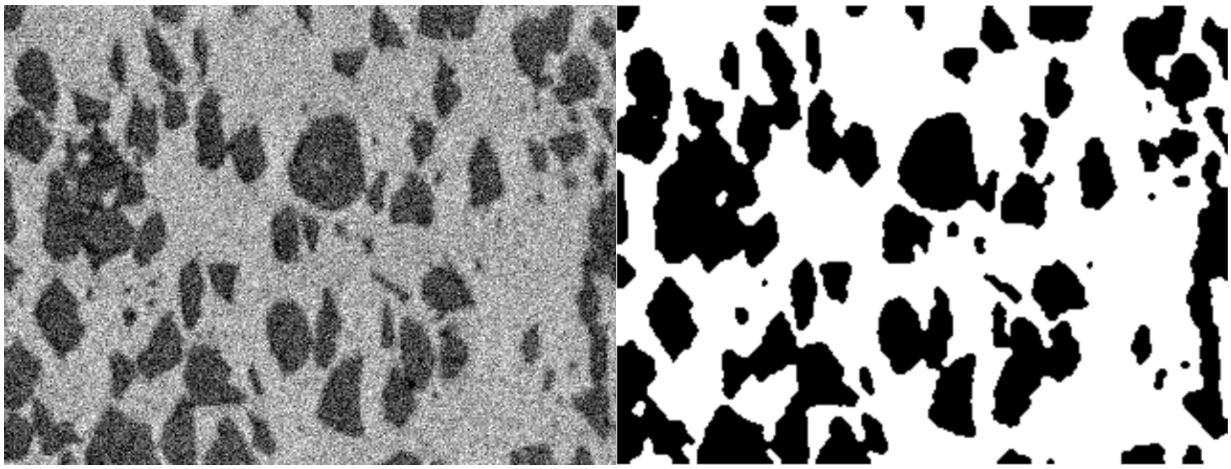
Example 20. In image processing, S may be a grid of pixels (locations are fixed and regular).

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

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

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

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



(A) Noised image (B) Reconstructed image

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

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

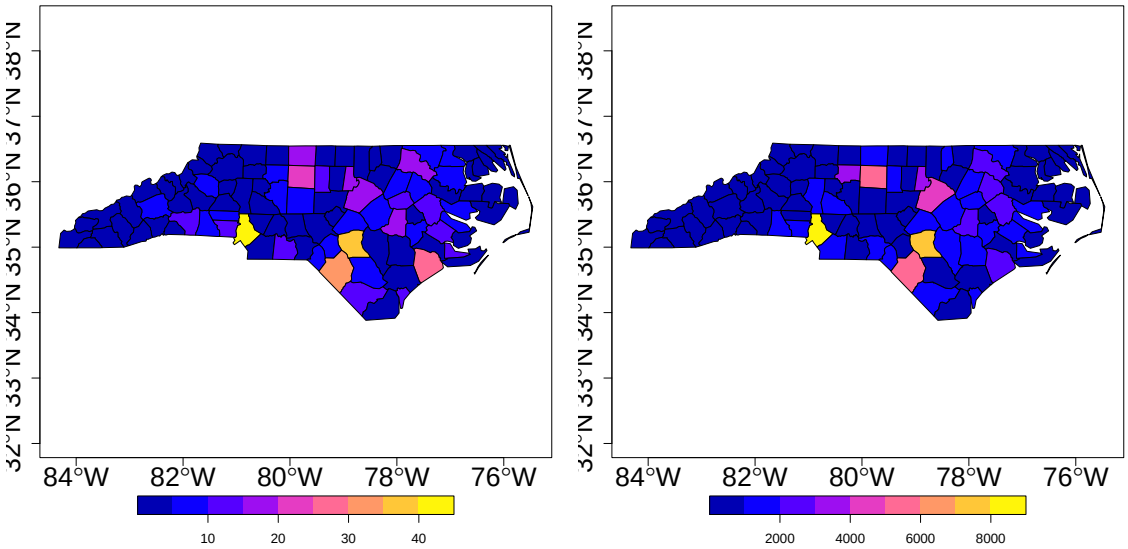
2.3. **Spatial point pattern data.**

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

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

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

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



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

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

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

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

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

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

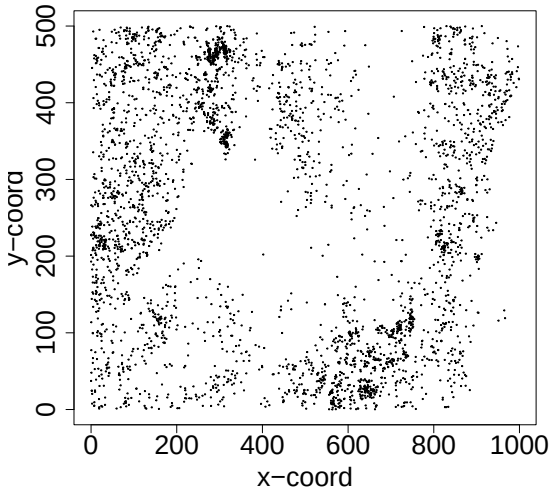


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

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

3. Uncertainty quantification and modeling

To be defined rigorously later

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

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

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

Spatial process model.

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

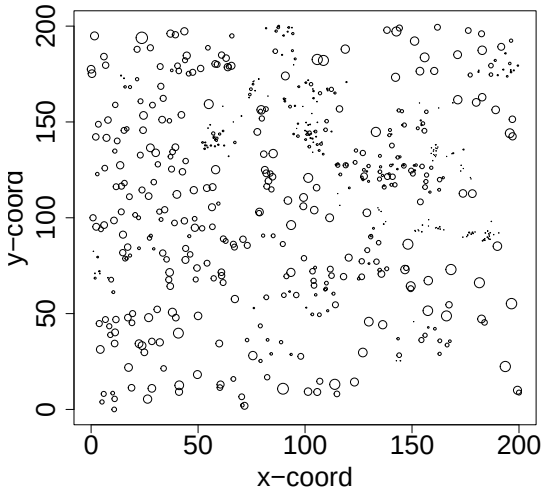


FIGURE 2.7. Longleaf Pines Point Pattern (Spatial point data)

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

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

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

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

Data model.

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

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

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

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

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

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

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

The hierarchical statistical model.

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

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

with

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

The Empirical (Bayes) hierarchical model.

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

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

with

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

...more details in the next lecture.

The Bayesian hierarchical model.

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

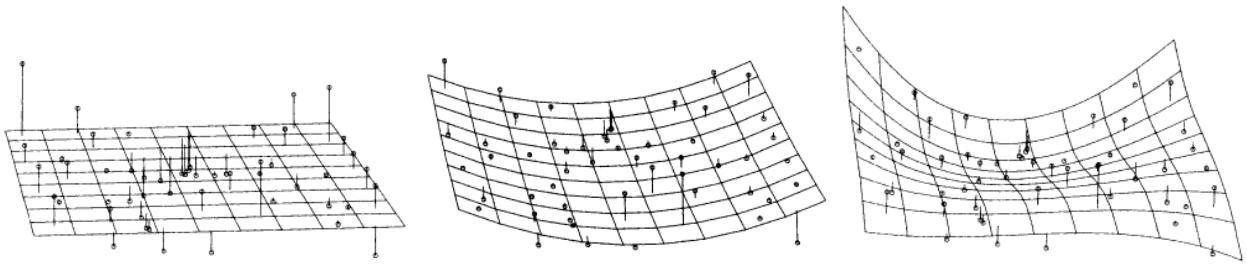
$$\begin{cases} Z|Y,S,\theta & \text{data model} \\ Y,S|\theta & \text{spatial process model} \\ \theta & \text{hyper-parameter prior model} \end{cases} \tag{3.7}$$

with

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

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

That's a snapsot for what follows.



(A) (B) (C)

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

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

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

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

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

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 2: Introduction to INLA & R-INLA

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

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

Reading list & references:

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

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

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

1. Laplace Approximation (LA)

Proposition 1. *Consider integral*

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

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

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

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

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

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

then

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

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

Example 2. Consider posterior expectation

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

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

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

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

Solution. (Sketch of the solution) It is

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

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

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

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

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

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

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

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

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

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

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

Solution. (Sketch of the solution) It is

$$\begin{aligned}\mathrm{pr}\left(\theta_1|z\right) &= \frac{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta_2}{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta} = \frac{\int \exp\left(nL^*_{\theta_1}\left(\theta_2\right)\right)\mathrm{d}\theta_2}{\int \exp\left(nL\ \ \left(\theta\right)\right)\mathrm{d}\theta} \\ &\overset{(\star)}{\approx} \left(\frac{\det\left(\Sigma^*\left(\theta_1\right)\right)}{2\pi n \det\left(\Sigma\right)}\right)^{\frac{1}{2}} \frac{\mathrm{pr}\left(\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\exp\left(\log\left(\mathrm{pr}\left(z|\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\right)\right)}{\mathrm{pr}\left(\hat{\theta}\right)\exp\left(\log\left(\mathrm{pr}\left(z|\hat{\theta}\right)\right)\right)}\end{aligned}$$

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

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

2. Integrated Nested Laplace Approximation (INLA)

2.1. **Motivations.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note 9. The joint posterior probability model is

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

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

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

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

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

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

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

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

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

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

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

2.3. **The general idea.**

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

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

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

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

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

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

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

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

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

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

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

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

Laplace approximation: Similar to Note 15, compute

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.5. **Byproducts.**

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

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

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

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

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

the model complexity here is measured via effective number of parameters

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

and hence DIC is defined as

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

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

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

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

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

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

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

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

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

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

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

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

Note 24. There are two essential steps:

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

The fitted model is returned as an inla object.

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

- Bayesian model

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

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

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

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

Data loading.

- Load R-INLA

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

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

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

Training via R-INLA.

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

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

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

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

Parametric inference.

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

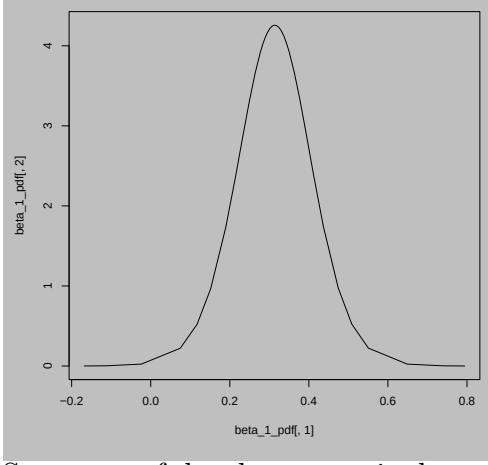
```
summary(result)
```

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

Random effects:
  Name    Model
    u IID model

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

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

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

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

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



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

result$marginals.fixed

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

result$summary.hyperpar

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

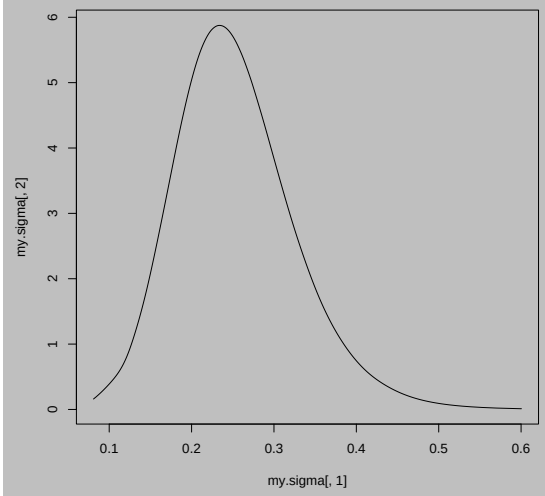
- To get the posterior summary of a function of the parameters, e.g. the posterior mean and standard deviation of $\sigma^2 = \exp(\tau)$

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

```
      mean         sd
0.25353753 0.07325247
```

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

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



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

```
?inla.marginal
```

Predictive inference.

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

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

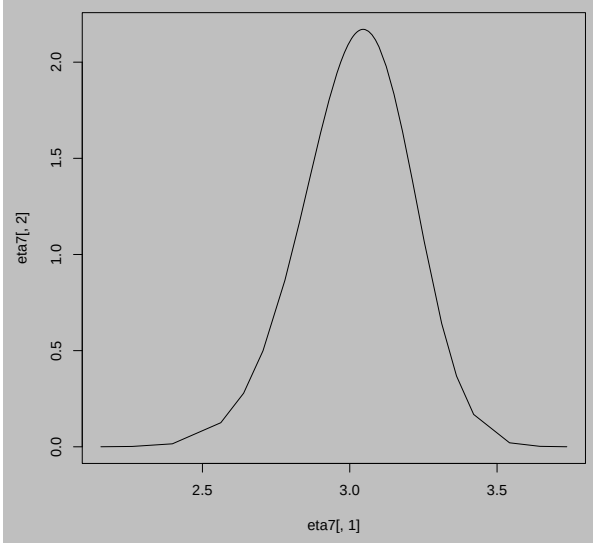
- Using the same settings as before, train the model by function `inla(INLA)`.

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

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

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

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

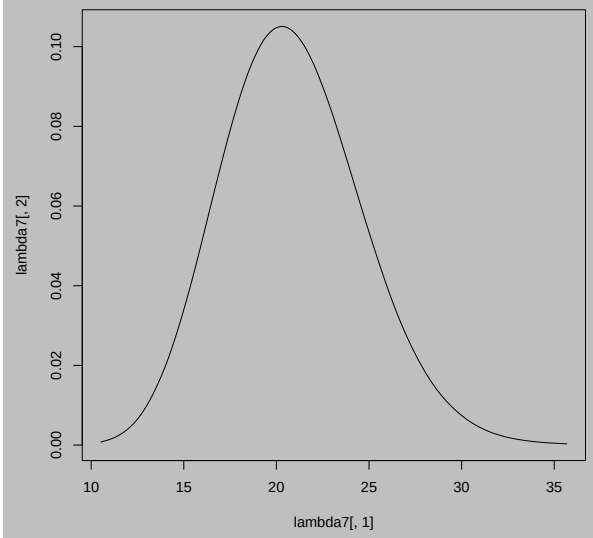


- Summary about is $\mathrm{pr}(\eta_7|z_{-7})$ taken by

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

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

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

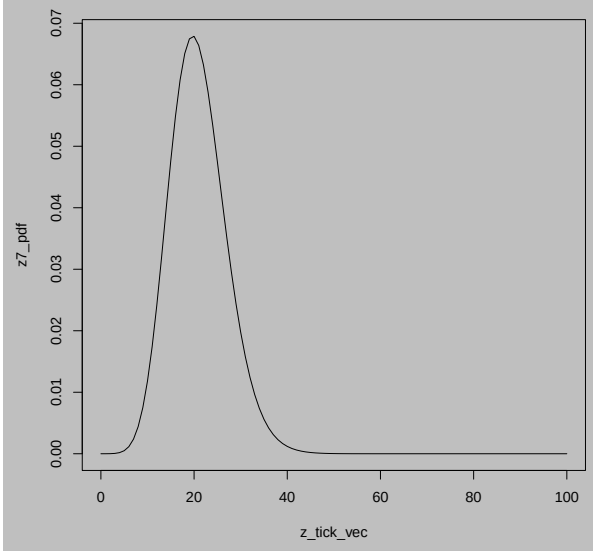


- To compute $\text{pr}\,(z_7|z_{-7})$ i.e. the predictive distribution (in this case) or the posterior distribution of the missing value (in principle), we can consider the following integration

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

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

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

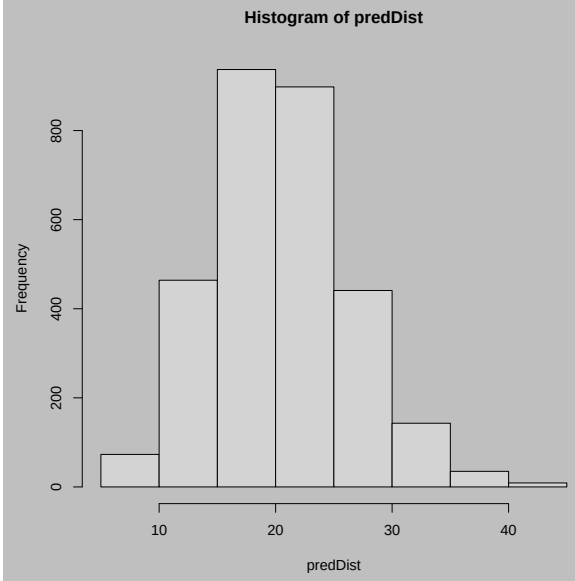


- alternatively one approximate (3.1) by Monte Carlo integration

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

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

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



Introduced in ASM3, MLNN3

Note 26. Assume we wish to address the minimization problem

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

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

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

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

Algorithm 29. *Newton algorithm consist of the recursion*

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

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

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

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

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

Example 31. Consider the marginal likelihood

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

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

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

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

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

Proof. Gradient descent's recursion is

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

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

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

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

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

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

Newton algorithm's recursion is

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

where additionally

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

□

Appendix B. Gaussian approximation of a (posterior) distribution

Introduced in SI2

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

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

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

Solution. The log posterior density is

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

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

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

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

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

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 3: Point referenced data modeling / Geostatistics

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

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

Reading list & references:

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

Specialized reading.

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

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

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

1. Stochastic processes (or Random fields)

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

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

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

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

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

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

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

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

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

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

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

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

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

1.1. **Mean and covariance functions.**

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

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

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

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

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

1.1.1. *Construction of covariance functions.*

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

Proposition 10. *The function* $c : S \times S \to \mathbb{R}$, $S \subseteq \mathbb{R}^d$ is a covariance function iff $c(\cdot,\cdot)$ is semi-positive definite; i.e. the Gram matrix $(c(s_i, s_j))_{i,j=1}^n$ is non-negative definite for any $\{s_i\}_{i=1}^n$, $n \in \mathbb{N}$.

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

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

Note 12. Proposition 13 uses the experience from basis functions, while Theorem 37 uses experience from characteristic functions to be incorporated int he process for modeling reasons.

Remark 13. One way to construct a c.f c is to set $c(s, s') = \psi(s)^\top \psi(s')$, for a given vector of basis functions $\psi(\cdot) = (\psi_1(\cdot), ..., \psi_n(\cdot))$.

Proof. From Proposition 10, as

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

□

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

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

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

Example 16. In second order processes $(Z_s)_{s \in \mathcal{S}}$, then the associated mean function $\mu(\cdot)$ and covariance function $c(\cdot,\cdot)$ exist, because $c(s,t) = \mathrm{E}(Z_s Z_t) - \mathrm{E}(Z_s)\mathrm{E}(Z_t)$ for $s, t \in \mathcal{S}$.

3. Gaussian process

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

Definition 18. $Z = (Z_s; s \in S)$ indexed by $S \subseteq \mathbb{R}^d$ is a Gaussian process (GP) or random field (GRF) if for any $n \in \mathbb{N}$ and for any finite set $\{s_1, ..., s_n; s_i \in \mathcal{S}\}$, the random vector $(Z_{s_1}, ..., Z_{s_n})^\top$ follows a multivariate normal distribution.

Also Example of Proposition

Proposition 19. *A GP* $Z = (Z_s; s \in S)$ *is fully characterized by its mean function* $\mu : S \to \mathbb{R}$ *with* $\mu(s) = E(Z_s)$*, and its covariance function with* $c(s, s') = Cov(Z_s, Z_{s'})$*.*

Notation 20. Hence, we denote the GP as $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$.

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

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

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

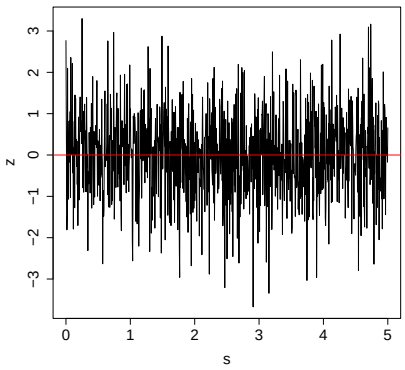
Example 23. Recall your linear regression lessons where you specified the sampling distribution to be $y_x|\beta,\sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(x^\top\beta,\sigma^2\right)$, $\forall x \in \mathbb{R}^d$. Well that can be considered as a GP $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ with $\mu(x) = x^\top\beta$ and $c(x,x') = \sigma^2 1(x=x')$ in (3).

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

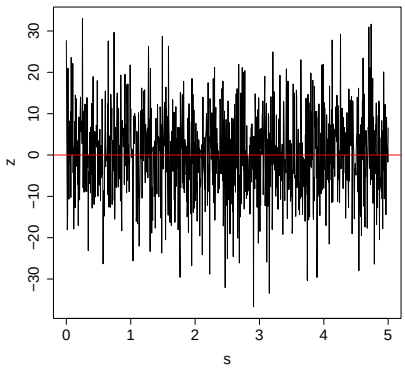
Algorithm 1 R script for simulating from a GP $(Z_s; s \in \mathbb{R}^1)$ with $\mu(s) = 0$ and $c(s,t) = \sigma^2 \exp\left(-\beta \|s-t\|_2^2\right)$

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

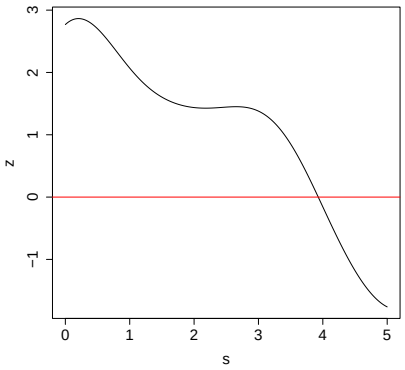
Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by σ^2 (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by σ^2 (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency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



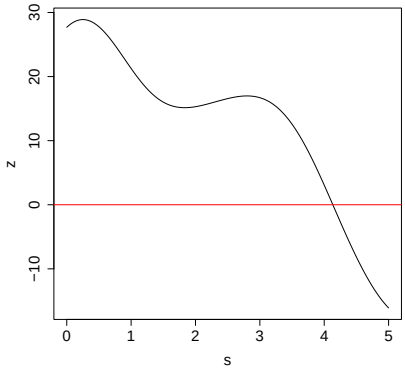
(A) Nugget c.f
($\sigma^2 = 1$)



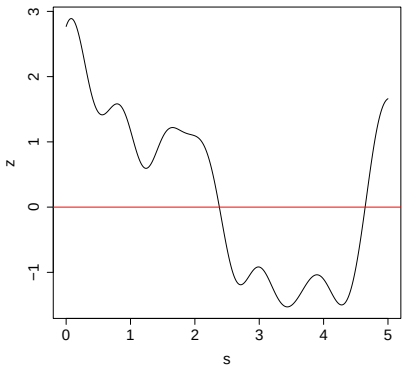
(B) Nugget c.f
($\sigma^2 = 100$)



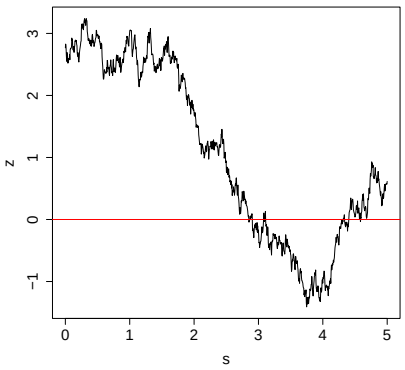
(C) Gauss c.f
($\sigma^2 = 1$,$\beta = 0.5$)



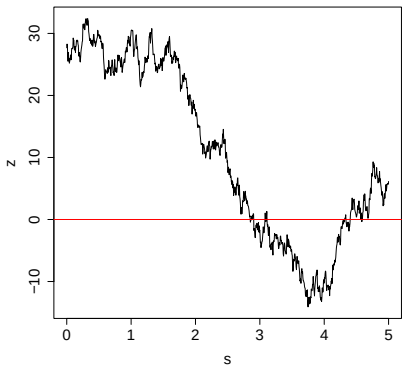
(D) Gauss c.f
($\sigma^2 = 100$,$\beta = 0.5$)



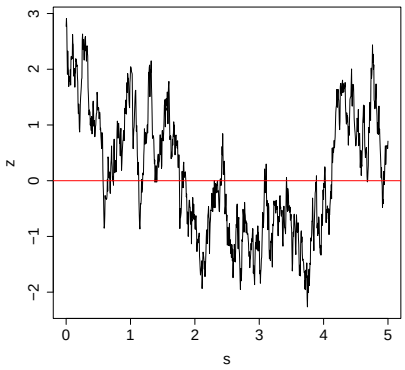
(E) Gauss c.f
($\sigma^2 = 1$,$\beta = 5$)



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

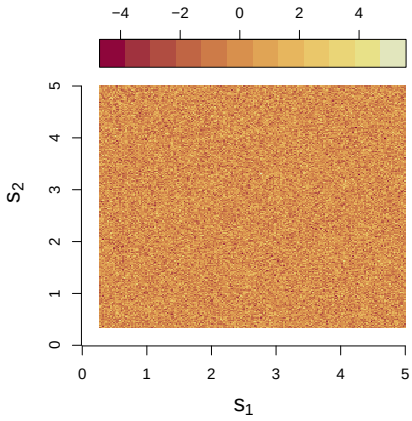


(G) Exp c.f
($\sigma^2 = 100$,$\beta = 0.5$)

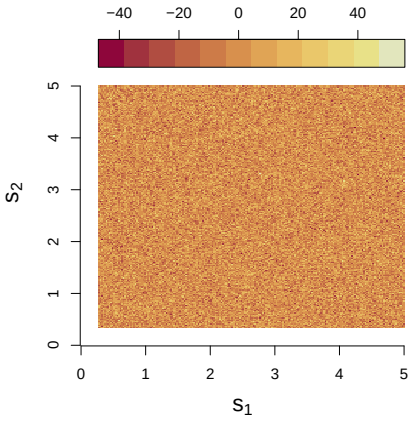


(H) Exp c.f
($\sigma^2 = 1$,$\beta = 5$)

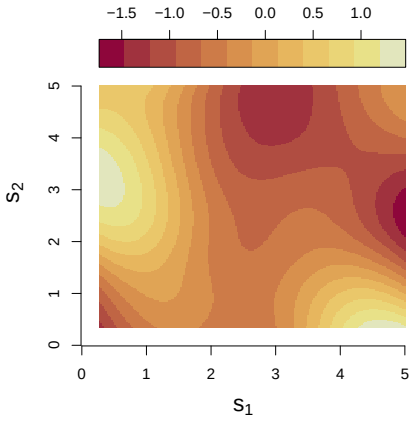
FIGURE 3.1. Realizations of GRF $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot,\cdot))$ when $s \in [0,5]$ (using same seed)



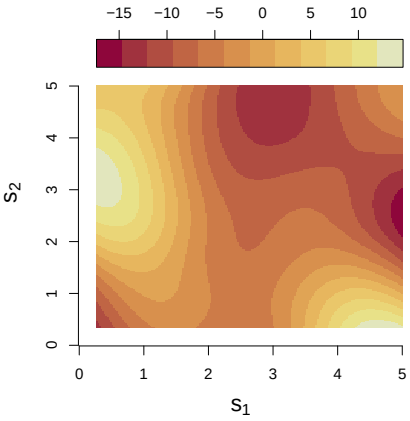
(A) Nugget c.f
($\sigma^2 = 1$)



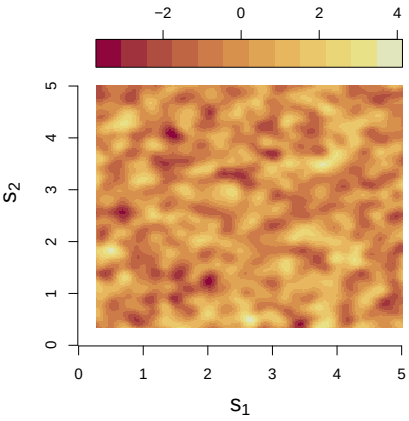
(B) Nugget c.f
($\sigma^2 = 100$)



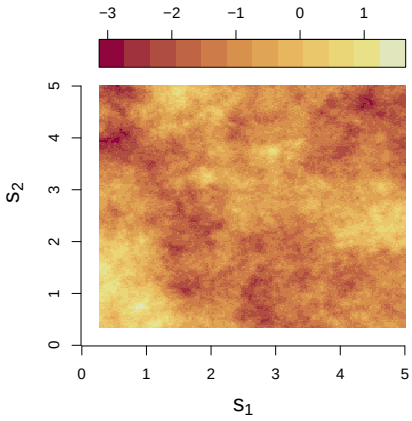
(C) Gauss c.f
($\sigma^2 = 1, \beta = 0.5$)



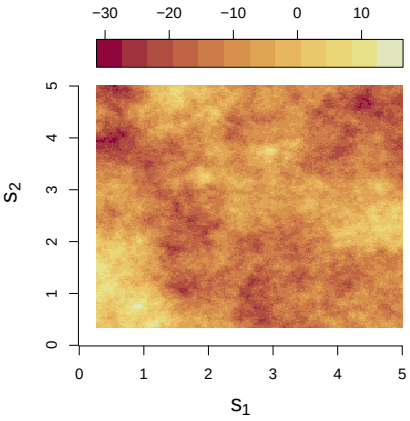
(D) Gauss c.f
($\sigma^2 = 100, \beta = 0.5$)



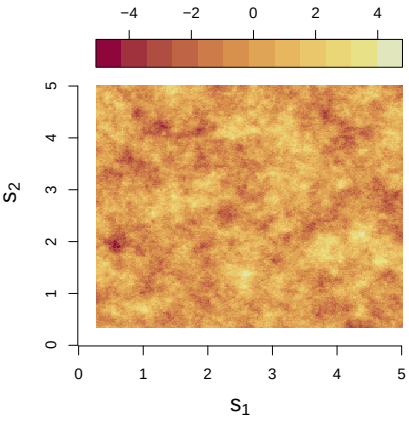
(E) Gauss c.f
($\sigma^2 = 1, \beta = 5$)



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



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



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

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

4. Strong stationarity

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

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

Definition 27. A random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ is strongly stationary if for all finite sets consisting of $s_1, ..., s_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, for all $k_1, ..., k_n \in \mathbb{R}$, and for all $h \in \mathbb{R}^d$

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

5. Weak stationarity (or second order stationarity)

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

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

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

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

Definition 31. Weakly (or second order) stationary covariance function is called the c.f. of a weakly stationary stochastic process.

6. Covariogram

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

Definition 33. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be a weakly stationary random field. The covariogram function of $Z = (Z_s)_{s\in\mathbb{R}^d}$ is defined by $c : \mathbb{R}^d \to \mathbb{R}$ with

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

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

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

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

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

Proposition 35. *If $c : \mathbb{R}^d \to \mathbb{R}$ is the covariogram of a weakly stationary random field $Z = (Z_s)_{s\in\mathbb{R}^d}$ then:*

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

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

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

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

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

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

Theorem 37. *(Bochner's theorem) A continuous even real-valued function $c : \mathbb{R}^d \to \mathbb{R}$ is a covariance function of a weakly stationary random process if and only if it can be represented as*

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

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

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

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

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

Theorem 38. *If $c(\cdot)$ is integrable, $F(\cdot)$ is absolutely continuous with spectral density $f(\cdot)$ of $Z = (Z_s; s \in \mathcal{S})$ then by Fast Fourier transformation*

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

Example 39. Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Then, by using Theorem 37, the spectral density is

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

i.e. it has a Gaussian form.

7. Intrinsic stationarity

Note 40. Getting greedier, we introduce an even weaker stationarity than the weak stationarity by considering lag dependent variance in the increments of the process with purpose to be able to use more inclusive models; Definition 30 implies that $\mathrm{Var}(Z_{s+h} - Z_s) = \mathrm{Var}(Z_{s+h}) + \mathrm{Var}(Z_s) - 2\mathrm{Cov}(Z_{s+h}, Z_s) = 2c(0) - 2c(h)$.

Definition 41. A random field $Z = (Z_s)_{s \in \mathbb{R}^d}$ is intrinsically stationary if, for all $h \in \mathbb{R}^d$, $(Z_{s+h} - Z_s)_{s \in \mathbb{R}^d}$ is weakly stationary; i.e.

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

Definition 42. Intrinsically stationary covariance function is called the c.f. of an intrinsically stationary stochastic process.

Example 43. The following covariance function is not weakly but intrinsically stationary

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

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

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

and

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

8. Semi variogram and Variogram

Note 44. The definition of the semi-variogram function requires the random field to be intrinsic stationarity; which is weaker assumption than weak stationary required by covariogram.

Definition 45. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be intrinsically stationary. The semi-variogram of Z is defined by $\gamma : \mathbb{R}^d \to \mathbb{R}$ with

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

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

Remark 47. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be weakly stationary with covariogram $c(\cdot)$. Then Z is intrinsic stationary with semi-variogram

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

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

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

Proposition 49. *Properties of semi-variogrames. Let $Z = (Z_s)_{s\in\mathbb{R}^d}$ be an intrinsically stationary process.*

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

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

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

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

(4) The following functions are semi-variograms

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

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

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

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

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

8.1. Behavior of variogram (Nugget effect, Sill, Range).

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

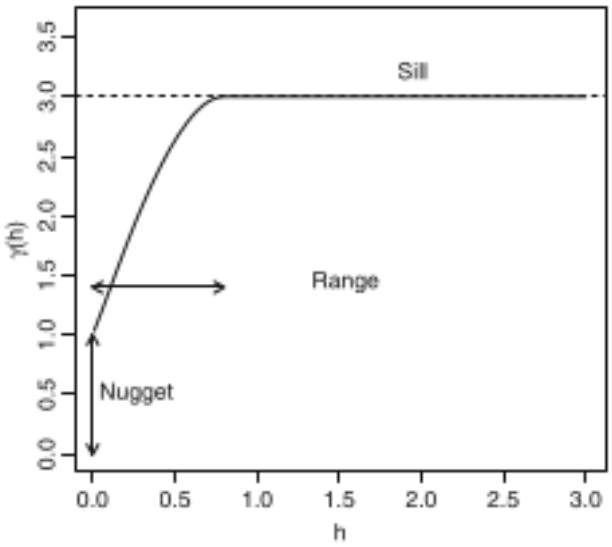


FIGURE 8.1. Semi Variogram's characteristics

Note 51. A semivariogram tends to be an increasing function of the lag $\|h\|$. Recall in weakly stationary processes, $\gamma(h) = c(0) - c(h)$ where common logic suggests that $c(h)$ is decreases with $\|h\|$.

Note 52. If $\gamma(h)$ is a positive constant for all lags $h \neq 0$, then $Z(s_1)$ and $Z(s_2)$ are uncorrelated regardless of how close $s_1$and s_2 are; and $Z = (Z_s)_{s \in \mathbb{R}^d}$ is often called white noise.

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

Nugget Effect.

Note 54. Nugget effect is the semivariagram's limiting value

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

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

Note 55. Nugget effect $\sigma_\varepsilon^2 \neq 0$ may expected or assumed to appear due to (1) measurement errors (e.g., if we collect repeated measurements at the same location s) or (2) due to some microscale variation causing discontinuity in the origin that cannot be detected from the data i.e. the spatial gaps because we collect a finite set of measurements at spatial locations. Hence theoretically, we could consider a more detailed decomposition $\sigma_\varepsilon^2 = \sigma_{\text{MS}}^2 + \sigma_{\text{MS}}^2$ where σ_{MS}^2 refers to the microscale and σ_{MS}^2 refers to the measurement error; however (my experience) this is non-identifiable.

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

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

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

Note 57. Nugget effect is often mathematically described by considering a decomposition ;

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

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

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

Sill.

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

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

Partial sill.

Definition 60. Partial sill is $\lim_{\|h\|\to\infty}\gamma(h) - \lim_{\|h\|\to 0}\gamma(h)$ which takes into account the nugget.

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

Note 61. An abrupt change in slope indicates the passage to a different structuration of the values in space. This is often modeled via decompoditions of processes with different semivariograms as in (8.2).

9. Isotropy

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

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

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

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

Definition 65. Isotropic semi-variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$ is the semi-variogram of the isotropic stochastic process. (sometimes for simplicity of notation we use $\gamma : \mathbb{R}^+ \to \mathbb{R}$ with $\gamma(\|h\|) = \frac{1}{2}\mathrm{Var}(Z_s - Z_{s-h})$.

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

Definition 67. Isotropic covariogram $c : \mathbb{R}^d \to \mathbb{R}$ of a weakly stationary process is the covariogram associated to an isotropic semi-variogram (sometimes for simplicity of notation we use $c : \mathbb{R}^+ \to \mathbb{R}$ with $c(\|h\|)$ from (9.1)).

9.1. **Popular isotropic covariance functions.**

Note 68. Given the covariogram $c(\cdot)$, and the semi-variogram can be computed from $\gamma(h) = c(0) - c(h)$ for any h.

9.1.1. *Nugget-effect.*

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

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

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

9.1.2. *Matern c.f.*

No need to memorize (9.2)

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

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

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

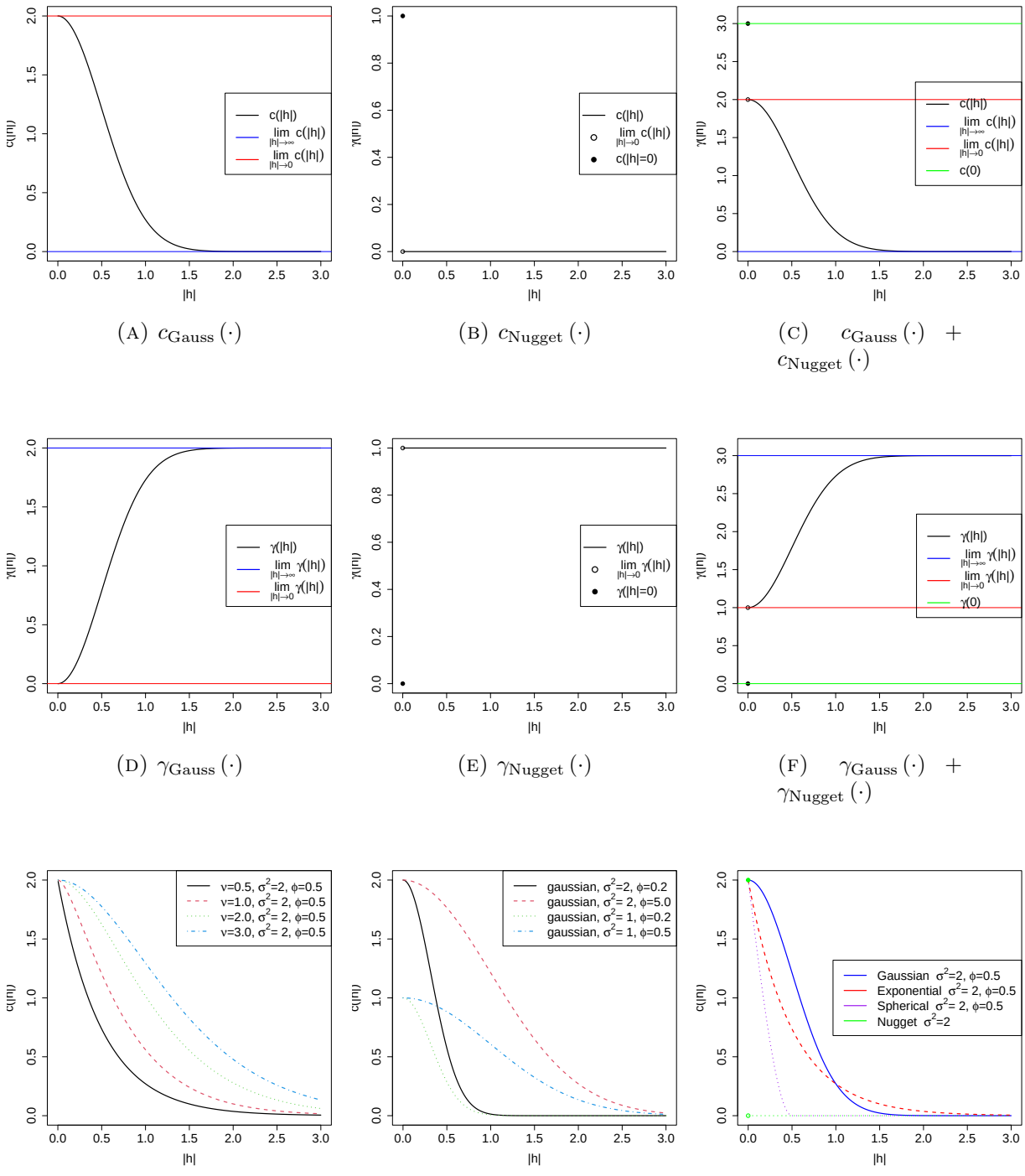
$$c(h) = \sigma^2 \exp\left(-\frac{1}{\phi} \|h\|_1 \right)$$

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

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

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

9.1.3. *Spherical c.f.*



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

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

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

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

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

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

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

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

10. Anisotropy

Note 72. Dependence between $Z(s)$ and $Z(s+h)$ is a function of both the magnitude and the direction of separation h. This can be caused by the underlying physical process evolving differently in space (e.g., vertical and horizontal axes).

Definition 73. The variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$ is anisotropic if there are h_1 and h_2 with same length $\|h_1\| = \|h_2\|$ but different direction $h_1/\|h_1\| \neq h_2/\|h_2\|$ that produce different variograms $\gamma(h_1) \neq \gamma(h_2)$.

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

Definition 75. The covariogram $c : \mathbb{R}^d \to \mathbb{R}$ is anisotropic if there are h_1 and h_2 with same length $\|h_1\| = \|h_2\|$ but different direction $h_1/\|h_1\| \neq h_2/\|h_2\|$ that produce different covariogram $c(h_1) \neq c(h_2)$.

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

Note 77. For brevity, below we discuss about intrinsically stationary process and variograms, however the concepts/definitions apply to weakly stationary process and covariograms when defined, as in Defs 73 & 75.

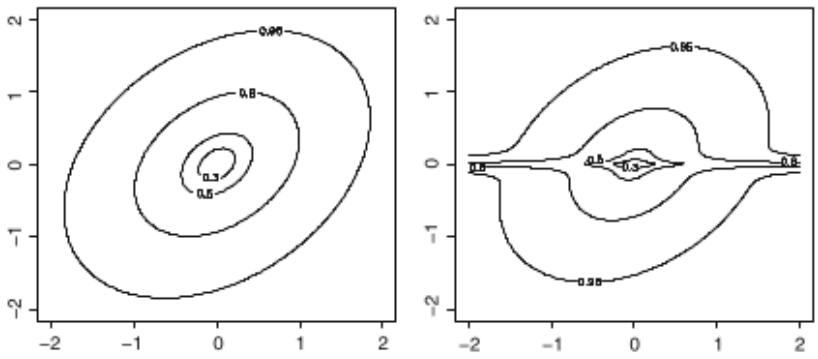
10.1. **Geometric anisotropy.**

Definition 78. The semi-variogram $\gamma_{\text{g.a.}} : \mathbb{R}^d \to \mathbb{R}$ exhibits geometric anisotropy if it results from an A-linear deformation of an isotropic semi-variogram with function $\gamma_{\text{iso}}(\cdot)$; i.e.

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

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

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



(A) Geometric anisotropy (B) Zonal anisotropy

FIGURE 10.1. Isotropy vs Anisotropy

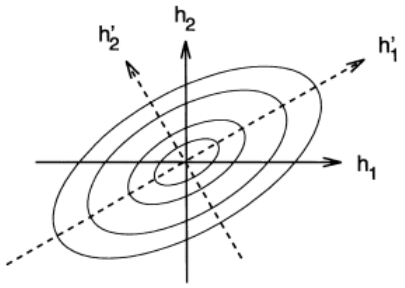


FIGURE 10.2. Rotation of the 2D coordinate system

Example 81. [Rotating and dilating an ellipsoid in 2D] Consider a coordinate system for $h = (h_1, ..., h_n)^\top$. We wish to find a new coordinate system for h in which the iso-variogram lines are spherical.

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

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

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

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

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

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

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

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

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

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

Definition 82. Support anisotropy is called the type of anysotropy when the semi-variogram $\gamma(h)$ of the process depends only on certain coordinates of h.

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

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

Example 85. Let γ' and γ'' be semi-variograms. If it is $\gamma(h = (h_1, h_2)) = \gamma'(\|h_1\|) + \gamma''\left(\sqrt{\|h_1\| + \|h_2\|}\right)$, then I 've Zonal anisotrotropy.

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

Note 87. If in 2D case, the sill in h_1 is larger than that in h_2, we can model zonal anysotropy of stochastic process (Z_s) by assuming $Z(s) = I(s) + A(s)$, where $I(s)$ is an isotropic process with isotropic semi-variogram γ_I along dimension of h_1 and $A(s)$ is an process with anisotropic semi-variogram γ_I without effect on dimension h_1; i.e. $\gamma_Z(h) = \gamma_I(h) + \gamma_A(h)$.

10.3. **Non-linear deformations.**

Note 88. A (rather too general) non-stationary model can be specified by considering semi-variogram $2\text{Var}(Z(s) - Z(t)) = 2\gamma_o(\|G(s) - G(t)\|)$ where we have performed a bijective non-linear (function) deformation $G(\cdot)$ of space $\mathcal{S}$ and applied on the isotropic semi-variogram γ_o. For instance, $\gamma_o(h) = \sigma^2 \exp(-\|h\|/\phi)$ and $G(s) = s^2$ as a deterministic function. Now, if function $G(\cdot)$ is considered as unknown, one can model it as a stochastic process $(G_s)_{s\in\mathcal{S}}$, and then we will be talking about deep learning modeling stuff.

11. Geometrical properties

Note 89. We discuss basic geometric properties of the basic models we will use for modeling, as it can give us a deeper intuition on how to design appropriate spatial statistical models.

Definition 90. (Continuity in quadratic mean (q.m.)) Second-order process $Z = (Z_s)_{s\in S}$ is q.m. continuous at $s \in \mathcal{S}$ if

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

Proposition 91. *For $Z=(Z_s)_{s\in S}$ it is*

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

- *If Z is intrinsically stationary, then*

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

and hence q.m. continuous iff $\lim_{h\to 0}\gamma(h)=\gamma(0)$.

- *If Z is weakly stationary, then*

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

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

Note 92. It has been shown that if a random field $Z=(Z_s)_{s\in S}$ has a variogram which is everywhere continuous apart from the origin i.e. $\lim_{s\to 0}\gamma(s)\neq\gamma(0)$ then Z it can be represented as $Z_s=Y_s+\varepsilon_s$ where (Y_s) has everywhere a continuous variogram and (ε_s) has a nugget effect, and Y_s, ε_s are independent. [2; Ch 1.4.1]

Definition 93. Differentiable in quadratic mean (q.m.)) Second-order process $Z=(Z_s)_{s\in\mathbb{R}}$ is q.m. differentiable at $s\in\mathbb{R}$ there exist

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

Proposition 94. *Let $c(s,t)$ be the covariance function of $Z=(Z_s)_{s\in S}$. Then Z is everywhere differentiable if $\frac{\partial^2}{\partial s\partial t}c(s,t)$ exists and it is finite. Also, $\frac{\partial^2}{\partial s\partial t}c(s,t)$ is the covariance function of (11.1).*

Example 95. The process with Gaussian c.f. $c(h)=\sigma^2\exp(-|h|/\phi)$ is continuous because $\lim_{h\to 0}c(h)=\sigma^2=c(0)$ but not differentiable because $\frac{\partial^2}{\partial h^2}c(h)$ does not exist at $h=0$.

Part 2. Model building & related parametric inference

12. The Geostatistical model

12.1. Linear Model of Regionalization.

Note 96. A spatial phenomenon can be thought as being the sum of several independent subphenomena acting at different characteristic scales. A linear model can be set up to split the stochastic process representing the phenomenon into several uncorrelated stochastic processes, each with a different variogram or covariance function and characterizing different aspect of the overall phenomenon under investigation.

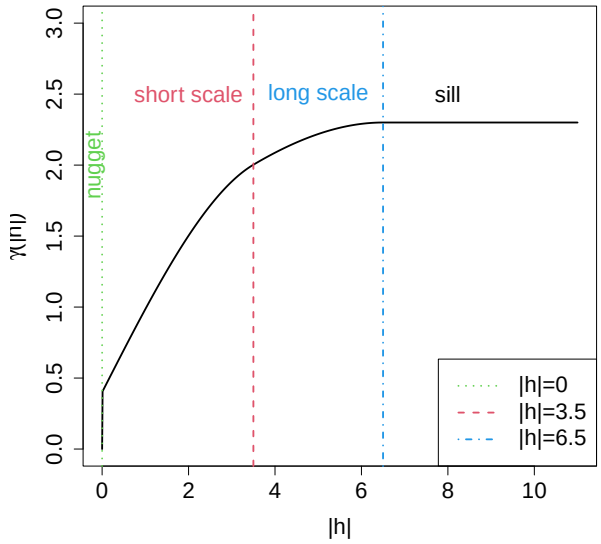


FIGURE 12.1. Variogram $\gamma(\cdot)$ of $Z(s) = Z_1(s) + Z_2(s) + Z_3(s)$ with spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 0.8, \phi = 3.5\right)$, spherical s.v. $\gamma_1\left(|h|; \sigma^2 = 1.1, \phi = 6.5\right)$, and nugget $\gamma_3\left(|h|; \sigma^2 = 0.4\right)$.

12.1.1. *Decomposition of the stochastic process.*

Note 97. The linear model of regionalization consider the decomposition of the stochastic process of interest $Z(s)$ as a summation of m independent zero-mean stochastic processes $\{Z_j(s)\}_{j=0}^{m}$ each of them characterizing different spatial scales, as

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

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

Remark 98. In (12.1), let $Z_j(\cdot)$ be intrinsically stationary with semi-variogram $\gamma_j(\cdot)$, then the semi-variogram of $Z(\cdot)$ is $\gamma(\cdot) = \sum_{j=1}^{m} \gamma_j(\cdot)$.

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

12.1.2. *Scale of variation.*

Note 100. Cressi [1] Considers the following intuitive decomposition

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

where

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

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

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

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

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

Note 101. Reasonably, larger scale components, such as $\mu(s)$, $W(s)$ can represented in the variogram if the diameter of the sampling domain is large $\mathcal{S}$ is large enough.

Note 102. Clearly, smaller scale components, such as $\eta(s)$, $\varepsilon(s)$ could be identified if the sampling grid is sufficiently fine.

Note 103. Decomposition (12.2) is not unique and the components are not clearly identifiable from the data when modeled; e.g. one may find two pairs of $\mu(s), W(s)$ doing the same thing; yet, separating $\eta(s)$ and $\varepsilon(s)$ is difficult as they often describe changes with range smaller than that of the sites (!)

Note 104. The geostatistical model is often presented (with reference to (12.2)) is a form

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

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

Note 105. Alternatively, the hierarchical statistical model (Handout 1, 3.5) is used

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

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

Note 106. A simpler decomposition is

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

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

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

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

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

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

13. Training & inference

Note 108. Suppose that the intrinsic stationary random field $(Z_s)_{s\in\mathcal{S}}$, $\mathcal{S} \in \mathbb{R}^d$ is observed at n sites $S = \{s_1, ..., s_n\}$, and we get n observed dataset $\{(s_i, Z(s_i))\}_{i=1}^n$.

Example 109. (Meuse river data set) The Meuse river dataset set, used as a running example gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Here, we use the topsoil zinc concentration, mg kg-1 soil ("ppm") as quantity of interest (Z). Heavy metal concentrations are from composite samples of an area of approximately 15m × 15m. See Figure 13.1a. This is the R dataset `meuse{sp}`.

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

13.1. The variogram cloud.

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

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

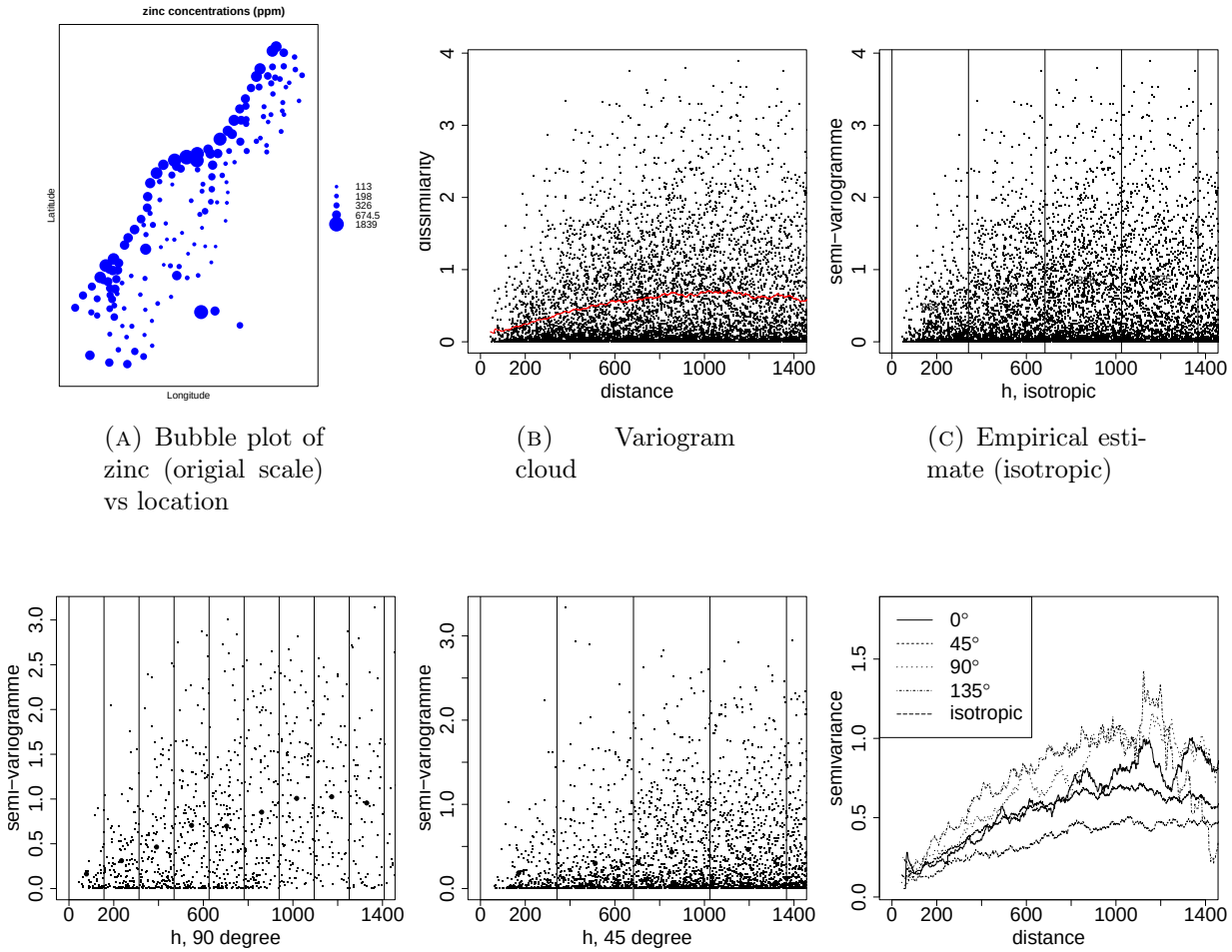
Definition 112. If we let dissimilarity between pairs of data values $Z(s)$ and $Z(s_b)$ depend on the separation $h = s_b - s$ (distance and orientation) then we get

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

Definition 113. The variogram cloud is the set of $n(n-1)/2$ points

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

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



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

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

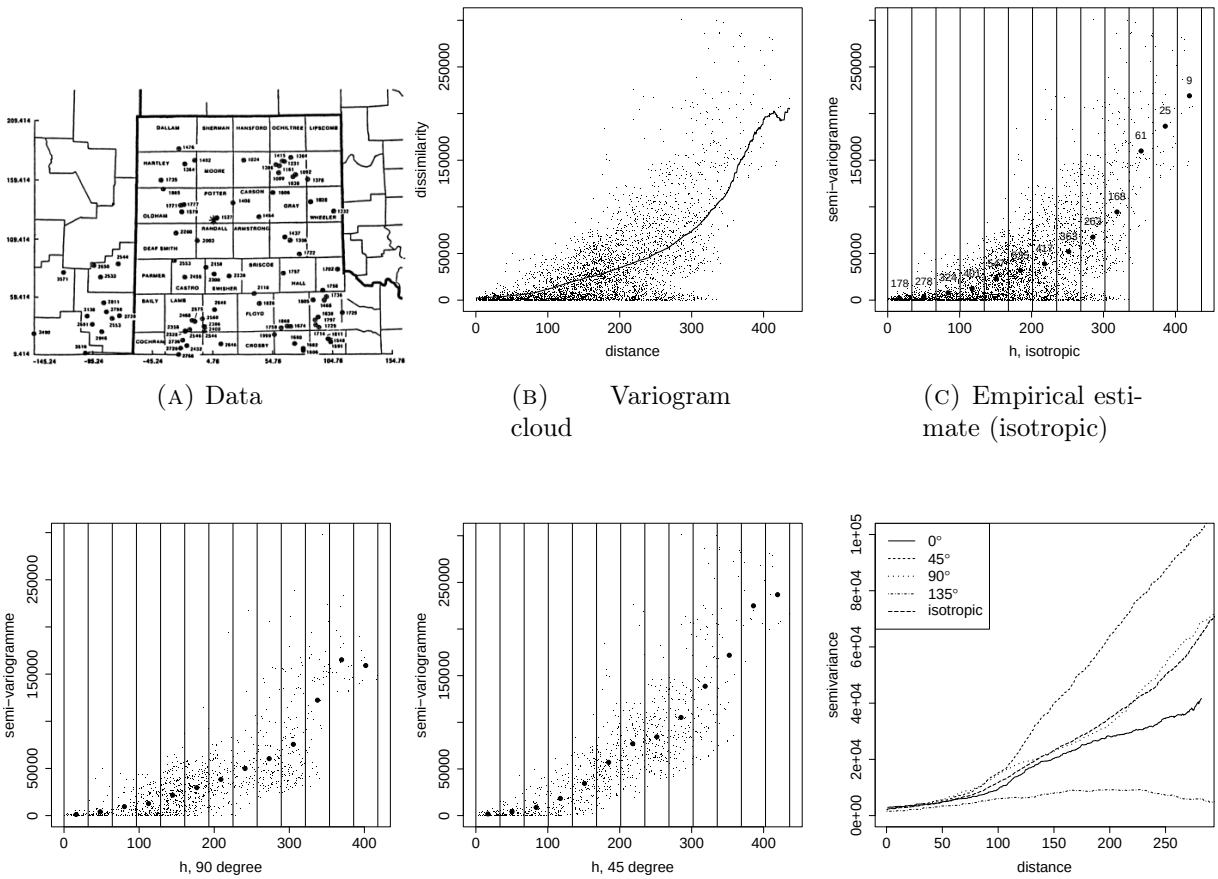
(F) Empirical estimate (different directions)

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

Note 115. Often there is a smoothing of the cloud is superimposed onto the cloud itself to help us see variogram's characteristics (e.g., sill, nugget, range) which may be "hidden" due to potential outliers in the plot.

Example 116. Figure 13.1b and Figure 13.2b show the variogram cloud plots (that is a point plot of the dissimilarities vs the distances) for the datasets Meuse and Wolfcamp-aquifer dataset. The red line is a smoother line of the cloud.

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



(A) Data

(B) Variogram cloud

(C) Empirical estimate (isotropic)

(D) Empirical estimate (90°)

(E) Empirical estimate (45°)

(F) Empirical estimate (different directions)

FIGURE 13.2. Wolfcamp-aquifer dataset variogram estimations

Proposition 117. *Smoothed Matheron estimator* $\hat{\gamma}(\cdot)$ *of semi-variogram* $\gamma(\cdot)$ *is*

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

where

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

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

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

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

Note 118. Estimator 13.2 can be written in matrix form as $\hat{\gamma}_{\mathrm{M}}(h) = Z^{\top} A(h) Z$, where $[A(h)]_{i,j} = 1(i \neq j) - 1/|N_{r_1,r_2}(h)|$ is a positive definite matrix.

Note 119. If we consider isotropic semi-variogram $\gamma(\cdot)$ then the ball may just considerate only the length of the distance as

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

because the direction does not have any effect.

Note 120. The choice of r_1 , r_2 is an art, and a trade-off between variance and bias, similar to the bin length in histograms.

Note 121. In practice, we consider a finite number of k separations $\mathscr{H} = \{h_1, ..., h_k\}$, we estimate in such a way that each class contains at least 30 pairs of points. Then compute $\{\hat{\gamma}_{\mathrm{M}}(h); h \in \mathscr{H}\}$, and plot $\{(h_j, \hat{\gamma}_{\mathrm{M}}(h_j)); j = 1, ..., k\}$.

Example 122. Figures 13.1c and 13.2c, show the nonparametric estimator ignoring the direction for the datasets Meuse and Wolfcamp-aquifer dataset. The estimator is calculated by using the ball in (13.4).

Example 123. Figures 13.2dand 13.1e show the nonparametric estimator considering directions 90° and 45° for the dataset Meuse. Figures 13.2d and 13.2e do the same for the Wolfcamp-aquifer dataset. The estimator is calculated by using the ball (13.3).

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

Example 125. Figure 13.1f and 13.2a show the nonparametric variogram estimator for different directions for the two datasets. We observe possible anisotropy due to the differences in the lines.

Proposition 126. *Assume a stationary Gaussian process* $(Z_s \sim GP(0, c(\cdot,\cdot)))_{s \in \mathcal{S}}$ *with semi-variogram* $\gamma(\cdot) = c(0) - c(\cdot)$. *The empirical semi-variogram* $\hat{\gamma}_{\mathrm{M}}$ in *(13.2) is*

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

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

Note 127. Estimation of the covariogram is done by

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

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

13.3. **Classic parametric estimation.**

Note 128. Smoothed Matheron estimator (13.2) does not necessarily satisfies semi-variogram properties, such as negative definiteness. To address this we use a parametric family of appropriate semi-variogram functions and tune them against data.

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

Proposition 130. *(Criteria checking variogram's validity.) A continuous function* $2\gamma(\cdot)$ *with* $\gamma(0) = 0$ *is a valid variogram iff: any of the following is satisfied:*

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

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

Example 131. Gaussian semi-variogram in Ex 48, it is

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

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

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

Proposition 132. *(Least Square Errors) Consider that the empirical semivariogram* $\hat{\gamma}$ *(e.g., Matheron (13.2)) of* γ *have been computed at* k *classes, i.e. it is available* $\{h_j, \hat{\gamma}(h_j)\}_{j=1}^{k}$. *The Least Square Errors (LSE) estimator of* $\gamma_\theta(h)$ *parametrised by the unknown* θ *for all* h *is* $\hat{\gamma}_{LSE}(h) = \gamma\left(h; \hat{\theta}_{LSE}\right)$, *where*

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

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

Proposition 133. *(Ordinary least squares) If use $V(\theta) = I$ in (13.6), we get the OLS $\hat{\gamma}_{OLS}(h) = \gamma\left(h; \hat{\theta}_{OLS}\right)$*

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

Proposition 134. *(Weighted least squares) If use $V(\theta) = diag(\varpi_1(\theta), ..., \varpi_k(\theta))$ for some weight function $\{\varpi_j(\theta)\}$, we get the WLE $\hat{\gamma}_{WLE}(h) = \gamma\left(h; \hat{\theta}_{WLE}\right)$*

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

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

Example 135. Figures 13.3a and 13.3b show the OLE and WLE estimates (13.7) and (13.8) of the exponential and spherical semi-variogram for the Meuse dataset. Figure 13.3c shows the OLE and WLE estimates (13.7) and (13.8) of the exponential semi-variogram for the Wolfcamp dataset. The parametric semi-variograms were tuned against the non-parametric estimator (13.2) presented in dots, as discussed in Proposition 132.

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

Note 136. Assume a stochastic process model (Z_s) decomposed as

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

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

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

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

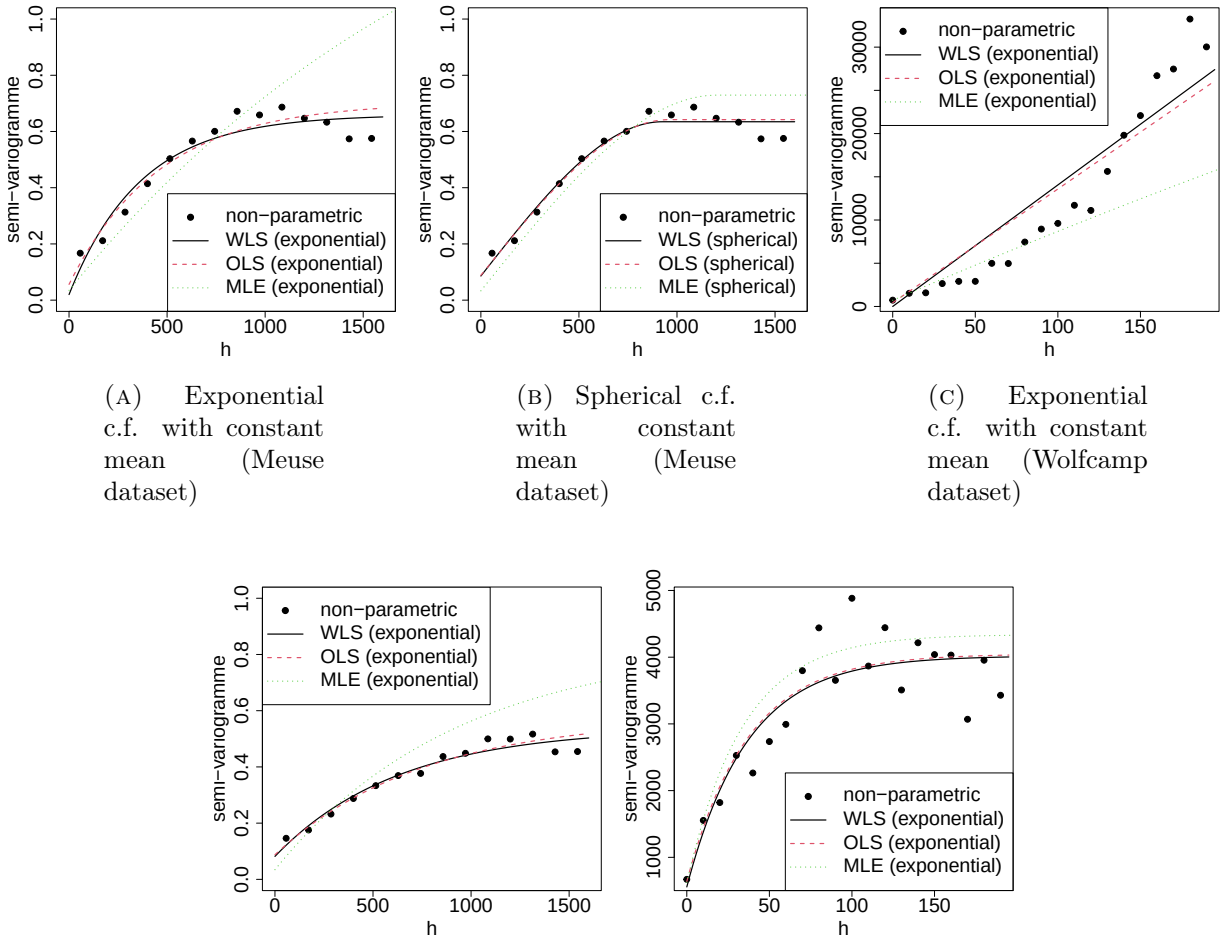
$$\hat{\beta}_{LSE} = \arg\min_{\theta} \left(\sum_i \left(Z(s_i) - \mu(s_i;\beta) \right)^2 \right)$$

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

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

(3) Estimate the empirical variogram for $\hat{\delta}$ on $\mathscr{H}$ according to Proposition 117, and estimate θ according to Proposition 132.

Example 138. Figure 13.3a and 13.3b fit an exponential c.f. and a spherical c.f. in the data of Meuse dataset (assuming constant mean); we cannot eyeball any big difference. Figure



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

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

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

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

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

FIGURE 13.3. Parametric training

13.3c fit an exponential c.f. in the data of Wolfcamp dataset (assuming constant mean); the fit looks really bad, possibly we should consider a non-constant mean and remove the trend.

Example 139. Figure 13.3d fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-42.8, -9.5 \cdot 10^{-4}, -6.6 \cdot 10^{-4})^{\top}$ in Meuse dataset. Possibly inference would suggest a constant mean function. Figure 13.3e fits an exponential c.f. in the residuals $\delta(s) = Z(s) - \mu(s)$ where $\mu(s) = \mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$ and $\hat{\beta}_{\text{OLS}} = (-607, -1.12, -1.13)^{\top}$ in Wolfcamp dataset; we see an improvement in fit compared to Figure 13.3c.

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

Note 140. Given that a probability distribution has been specified for the stochastic process $(Z_s)_{s\in\mathcal{S}}$, the MLE involves (1) the derivation of the associated pdf $\text{pr}(Z_1,...,Z_n|\beta,\theta)$ of the n-dimensional sampling distribution, (2) the computation of the associated likelihood function $L(z_1,...,z_n|\beta,\theta)$ given the data $\{(s_i,Z_i)\}_{i=1}^n$, and finally (3) the computation of the MLE estimates $\left(\hat{\beta}_{\text{MLE}},\hat{\theta}_{\text{MLE}}\right)$ of $\left(\hat{\beta},\hat{\theta}\right)$ as

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

Example 141. If $(Z_s)_{s\in\mathcal{S}}$ is specified as $Z(\cdot)\sim\text{GP}(\mu(\cdot;\beta),c(\cdot,\cdot;\theta))$, with $\mu(s;\beta)=\beta_0+s_1\beta_1+s_2\beta_2$ then MLE of (β,θ) is

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

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

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

Note 142. Given that a probability distribution has been specified for the stochastic process $(Z_s)_{s\in\mathcal{S}}$, the Bayesian training involves (1) the derivation of the pdf $\text{pr}(Z_1,...,Z_n|\beta,\theta)$ of the n-dimensional sampling distribution, (2) the computation of the associated likelihood function $L(z_1,...,z_n|\beta,\theta)$ given the data $\{(s_i,Z_i)\}_{i=1}^n$; and (3) the specification of the prior model $(\beta,\theta)\sim\text{pr}(\beta,\theta)$, leading to the Bayesian hierarchical model

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

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

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

(See Handout 1, Section 3)

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

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

and the posterior is given by the Bayes theorem as

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

The parametric variogram can be estimated via

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

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

Part 3. Prediction in geostatistics

14. The (traditional) Kriging paradigm

Note 144. "Kriging" is a general technique for deriving an estimator / predictor of $Z\left(\cdot\right)$ (or a function of it) at a location (such as a spatial point s_0, or a block of points $\left\{s_j^*\right\}$ or a subregion v_0) of a spatial region $\mathcal{S}$ by properly averaging out data in the neighborhood around the location of interest.

14.1. Universal Kriging.

Note 145. Consider we have specified the statistical model as a stochastic process $\left(Z_s\right)_{s\in\mathcal{S}}$ with

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

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

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

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

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

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

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

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

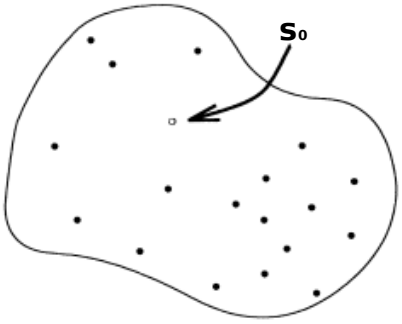


FIGURE 14.1. Kriging area

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

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

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

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

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

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

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

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

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

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

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

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

(14.5)

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

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

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

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

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

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

Note 155. The UK system of equations is

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

(14.10)

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

(14.11)

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

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

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

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

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

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

Note 156. Hence the UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

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

with standard error

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

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

Note 157. $(1-\alpha)\,100\%$ Prediction interval of UK predictor $Z_{\text{UK}}(s_0)$ at s_0 is

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

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

Note 158. Note that we have not assumed a particular distribution of Z_s or δ_s, but only stationarity assumptions.

Note 159. It was not necessary to consider the intrinsic stationarity assumption in Note 145 in order to derive the Universal Kriging predictor; we could have derived its formulas (14.14) & (14.15) with respect to the covariance function $c(\cdot,\cdot)$ of (Z_s) instead of its semivariogram $\gamma(\cdot)$. Here, intrinsic stationarity was assumed for practical reasons: it allowed us to express

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

Note 160. To use (14.14), (14.15), and (14.17), we need to learn the unknown coefficients $\{\beta_j\}$ and the semi-variogram $\gamma(\cdot)$, or "equivalently" the unknown hyper-parameter θ of the parametric semivariogram $\gamma_\theta(\cdot)$ used to cast $\gamma(\cdot)$. In practice, we use the same dataset used to compute (14.13), however in principle a fresh training dataset $\{(s_i', Z_i')\}_{i=1}^n$ is required (never use the same training data 2 times). A training procedure can be the following.

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

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

(2) Compute the residuals

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

(3) Compute the empirical variogram $\hat{\gamma}$ for $\hat{\delta}$ on $\mathscr{H}$ according to Proposition 117,

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

Example 161. [4] Consider the example with the Meuse dataset. Fig 14.2b presents the UK prediction $Z_{\text{UK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (14.1) for when the spatial mean has a linear form $\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$. Following Note 160, we computed the $\hat{\beta}_{\text{LSE}}$ of β by (14.18), then we removed the linear trend by (14.19) and computed the residual process $\left\{\hat{\delta}_i\right\}$, then we computed the semi-variogram $\hat{\gamma}$ (13.2) of δ as in Proposition 117; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ of δ where we computed the OLS $\hat{\theta}_{\text{OLS}} = \left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (13.7) (see Figure 13.3d); and then we pluged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)}$ in (14.14) to compute the UK weights w_{UK} for the UK predictor $Z_{\text{UK}}(s_0) = w_{\text{UK}} Z$ for any $s_0 \in \mathcal{S}$. The reason that we do not see much difference between OK in Figure 14.2a and UK in Figure 14.2b is possibly because the slops in the linear trend (mean) of UK are rather small and insignificant (See Example 139).

Example 162. (Cont. Examples 109, 125) Consider the example with the Meuse dataset. The dataset has another measurement (a potential regressor in the deterministic mean $\mu(s)$), the "distance to the Meuse river bed" $\{d_i\}$ at the associated locations $\{s_i\}$, let's denote it by d. Figure 14.2c shows a rather linear relationship between Z and $\sqrt{d}$, hence we can consider

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

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

14.2. **Ordinary Kriging.**

Note 163. Ordinary Kriging (OK) addresses spatial prediction in cases that the specified statistical model on $(Z_s)_{s\in\mathcal{S}}$ has the form

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

with unknown $\beta_0 \neq 0$ and intrinsically stationary process (δ_s).

Note 164. OK can be derived as a special case of the Universal Kriging by setting $p = 0$ and constant spatial mean $\mu(s) = \beta_0$.

Example 165. [The derivation is in (Exercise 19 Exercise sheet).] As a supplementary and for demonstration, we mention that the OK assumption is $\sum_{i=1}^{n} w_i = 1$; the OK system of equations is $0 = \nabla_{(\{w_i\},\lambda)} L(w,\lambda)\big|_{(w,\lambda)}$ producing

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

the weights are

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

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

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

14.3. **Simple Kriging.**

Note 166. Simple Kriging (SK) addresses spatial prediction in cases that the specified statistical model on $(Z_s)_{s\in\mathcal{S}}$ has the form

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

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

Example 167. [The derivation is in (Exercise 17 in the Exercise sheet).] It does not require any assumption in the weights such as (14.4) or (14.21).As a supplementary and for

demonstration, we mention the SK predictor at s_0 and standard error:

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

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

Example 168. Consider the example with the Meuse dataset. Fig 14.2a presents the OK prediction $Z_{\text{OK}}(s_0)$ at any point $s_0 \in \mathcal{S}$ under model (14.20) that is the UK case (14.1) for when $\mu(s) = \beta_0$. First we computed the non-parametric semivariogram $\hat{\gamma}$ (13.2) as in Proposition 117; then we considered a (parametric) isotropic exponential semi-variogram $\gamma_{(\sigma^2,\phi)}$ where we computed the OLS $\hat{\theta}_{\text{OLS}} = \left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)$ of the hyperparameters (σ^2, ϕ) as in (13.7) (see Figure 13.3a); and then we plugged in the estimated $\gamma_{\left(\hat{\sigma}^2_{\text{OLS}}, \hat{\phi}_{\text{OLS}}\right)}$ in (14.22) to compute the OK weights w_{OK} for the OK predictor $Z_{\text{OK}}(s_0) = w_{\text{OK}} Z$ for any $s_0 \in \mathcal{S}$.

15. The Bayesian Kriging paradigm (Hierarchical modeling)

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

Note 169. Consider the geostatistical model of (Z_s) with a scale decomposition such as in (12.3)

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

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

Note 170. Consider a dataset $\{(s_i, Z_i)\}_{i=1}^n$ with $Z_i = Z(s_i)$ being a realization of (15.1) at site $s_i \in \mathcal{S}$. Let $Z = (Z_1, ..., Z_n)^\top$, and $Y = (Y_1, ..., Y_n)^\top$.

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

Recall

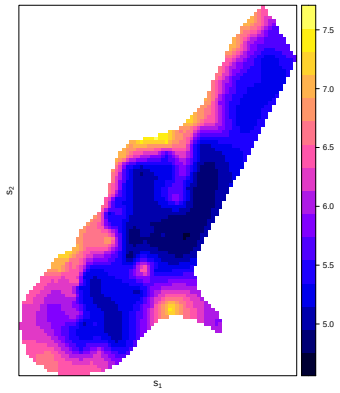
Note 172. Uncertainty can be decomposed according to the Hierarchical spatial model

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

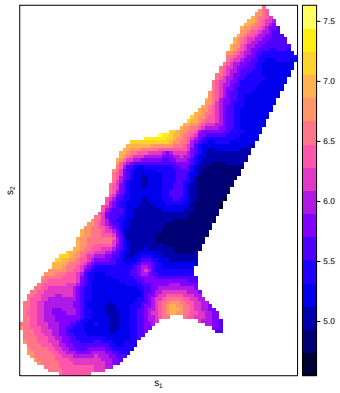
with

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

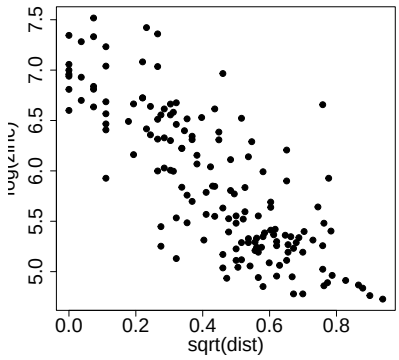
Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\text{pr}(Y|\vartheta)$ possibly labeled by some parameter ϑ.



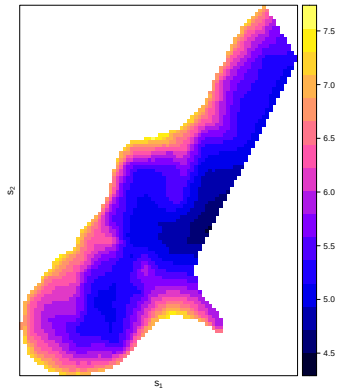
(A) Ordinary Kriging, constant mean, exponential covariogram



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



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

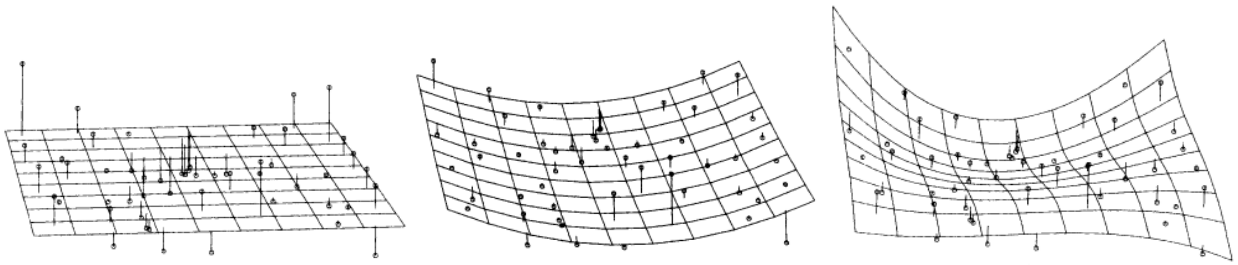


(D) Universial Kriging, mean $\mu(s,d) = \beta_0 + \beta_1\sqrt{d}$, exponential covariogram

FIGURE 14.2. Kriging Meuse dataset.

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

Note 173. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \text{pr}(\vartheta_1|\vartheta_2)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior; in certain problems, ϑ_2 can be considered as known and sometimes as unknown in what follows.



(A) (B) (C)

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

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

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

where uncertainty is described by

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

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

Note 176. Under Bayesian model (15.3), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

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

by maximizing the marginal likelihood

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

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

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

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

Note 178. General interest lies in computing the posterior predictive distributions of the spatial process model (Y_s), (or latent process, or noiseless process) given the data Z

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

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

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

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

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

Note 179. The above statistical problem is naturally addressed in the (either full or empirical) Bayesian statistical framework. It is often called Bayesian Kriging.

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

Inventory of useful formulas.

Fact 180. *Let $X\sim N(\mu_X,\Sigma_X)$ $Y\sim N(\mu_Y,\Sigma_Y)$ and Y,X independent. Let fixed matrices A and B and vector c of appropriate sizes. Then*

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

Fact 181. *Let $N(\beta|b,B)$ be the Gaussian pdf with mean b and covariance B at β. It is*

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

Fact 182. *[Marginalization & conditioning] Let $x_1\in\mathbb{R}^{d_1}$,* and $x_2\in\mathbb{R}^{d_2}$. If

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

then it is

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

where

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

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

A possible narrative - a story.

Note 184. Consider there is available a dataset $\{(s_i, Z_i)\}_{i=1}^n$ where $Z_i = Z(s_i)$ is a realization of a stochastic process (Z_s) with $\{Z_i \in \mathbb{R}\}$.

Note 185. In particular, assume that data are instances of an unknown function $Y(\cdot)$ at s_i but contaminated by additive random noise $\{\varepsilon_i \sim \mathrm{N}(0, \tau^2); i = 1, ..., n\}$ with scale $\tau > 0$; i.e. $Z_i = Y(s_i) + \varepsilon_i$.

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

Specifying the hierarchical model.

Note 187. A natural model to setup for this problem is the geostatistical model

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

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

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

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

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

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

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

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

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

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

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

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

where $c(s,s'|\theta) = c_Y(s,s'|\sigma^2,\phi) + c_\varepsilon(s,s'|\tau)$, and covariance function parameters $\theta = (\sigma^2,\phi,\tau)^\top$. *[We used the additive property of Gaussian random variables in Fact* 180 *].*

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

Note 190. Assume a vector of "unseen" sites $S_* = (s_{*,1},...,s_{*,q})^\top$ for any $q \in \mathbb{N}_0$. Let convenient notation $Z := Z(S)$, and $Z_* := Z(S_*)$. The join marginal distribution of $(Z_*,Z)^\top$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is

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

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

Note 191. Given that vector Z is observed/known, the (posterior) predictive distribution of $Z_*|Z$ given β, $\theta = (\sigma^2,\phi,\tau)^\top$ is the conditional distribution

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

where

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

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

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

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

with

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

for any $s,s' \in \mathcal{S}$. This is the predictive process of $Z(s)$ at any $s \in \mathcal{S}$ given Z, β, θ. *[Here we used the definition of GP (Definition 18) given Note* 191*].*

Note 193. Assume that the parameters (β,θ) are unknown but fixed (i.e. no prior is specified). Training can be performed by maximizing the marginal likelihood of Z given β,θ

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

derived from (15.9) by solving

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

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

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

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

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

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

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

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

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

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

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

Note 199. The marginal Bayesian model is now extended to

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

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

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

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

and results as

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

with

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

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

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

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

and it is again a GP

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

with

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

(15.18)

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

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

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

Note 203. Assume that $\theta = (\sigma^2, \phi, \tau)^\top$ is an unknown fixed hyper-parameter without a prior distribution being specified. Training can be performed by maximizing the marginal likelihood of Z given θ

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

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

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

[from Fact 181] by computing

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

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

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

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

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

Part 4. Spatial misalignment (special topic)

16. Intro to spatial misalignment

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

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

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

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

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

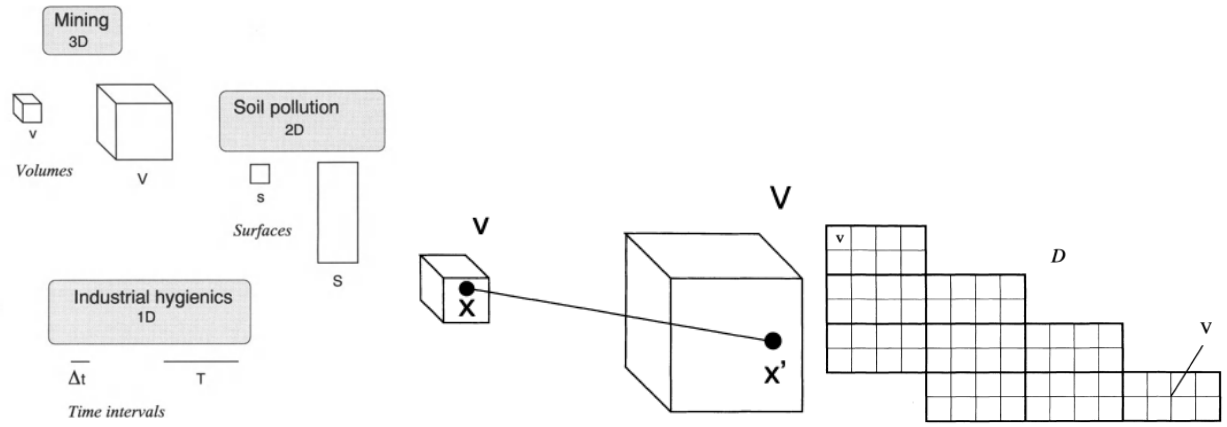
$$\int_B Z\left(x\right)\mathrm{d}x = \lim_{n\to\infty,m\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(v_j - v_{j-1}\right)\left(u_j - u_{j-1}\right)Z\left(u'_j,v'_j\right) \tag{16.2}$$

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

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

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

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



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

FIGURE 16.1. Change of support

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

Change of support problem.

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

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

16.1. **Extension and Dispersion Variance.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

similar

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

so (16.3) gives

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

In greater scale, the above can be extended to

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

□

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

Change of support effect.

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

Change of support: affine model.

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

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

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

which implies the relation

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

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

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

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

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

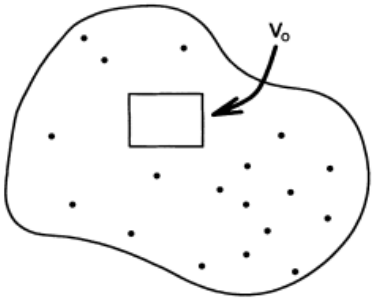


FIGURE 16.2. Block Kriging cartoon

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

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

Assume

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

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

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

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

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

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

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

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

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

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

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

Note 232. The Block Universal Kriging equations then are

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

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

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

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

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

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

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

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

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

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

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

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

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

with

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

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

Part 5. Extensions to multivariate Geostatistics (special topic)

17. Extensions to Multivariate Geostatistics

17.1. Cross-variance functions.

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

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

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

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

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

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

Example 239. Cross-covariograms have the following properties

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

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

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

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

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

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

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

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

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

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

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

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

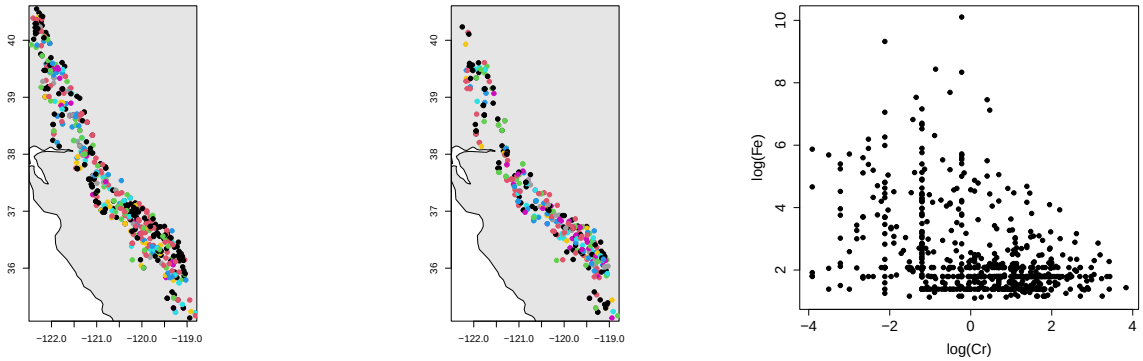
for $i,j=1,...,k$ and $s,s+h\in\mathcal{S}$.

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

17.2. **Co-Kriging.**

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

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



(A) Cr concentration

(B) Fe concentration

(C) Scatterplot Cr vs Fe

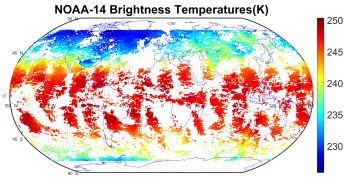
FIGURE 17.1. Central valley Groundware example data

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

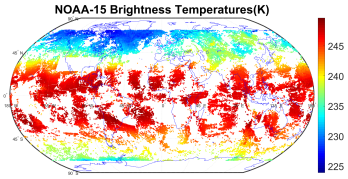
(3) Different variables correspond to a different accuracy level or support. The variable of interest could be the temperature reading in a location, and the auxiliary variables could be $\{Z_1(s_i)\}_{i=1}^{n_1}$ the temperature readings by an old technology (less accurate) satellite, and $\{Z_2(s_i)\}_{i=1}^{n_1}$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

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

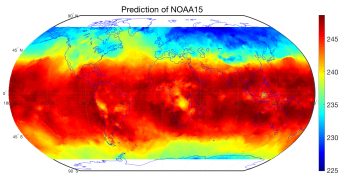
Note 247. Consider k stochastic processes $Z_1(s), ..., Z_k(s)$, $s \in \mathcal{S}$. Consider data at n sites $\{s_i\}_{i=1}^{n}$. Let $\boldsymbol{Z}(\boldsymbol{s})$ be a $n \times k$ matrix $\boldsymbol{Z}(\boldsymbol{s}) = Z_j(s_i)$ for $i = 1, ..., n_j$, and $j = 1, ..., k$. It is desired to predict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



(A) N14 (less accurate)



(B) N15 (more accurate)



(C) N15 (co-kriging prediction)

FIGURE 17.2. Satellite temperature readings data

Note 248. Assume

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

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

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

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

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

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

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

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

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

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

Notation 253. Set convenient notation for the calculations below as

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

Note 254. The MSE (or Variance) is

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

Note 255. The Lagrange function is

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

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

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

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

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

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

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

(17.4)

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

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

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

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

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

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

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

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

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

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

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

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

17.3.1. *Intrinsic multivariate correlation.*

17.3.2. *The Linear model of coregionalisation.*

Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 4: 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] Gaetan, C., & Guyon, X. (2010; Ch 3). Spatial statistics and modeling (Vol. 90). New York: Springer.

Specialized reading.

[3] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons. (on Spatial analysis)

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

Note 1. Recall from Section 2.2 of "Handout 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.

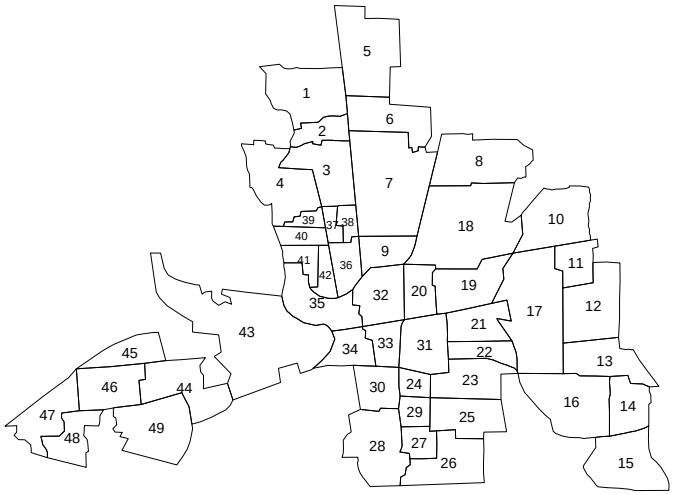


FIGURE 1.1. Lattice of spatial sites for Columbus dataset. Each neighborhood is a site. Each site is labeled. The collection of sites is the lattice of sites.

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. More 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

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

Notation 10. 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" $\sim$.

Example 11. 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 $j \in \{32, 9, 18, 19, 31, 33\}$.

Example 12. (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 $\text{pr}_{Z_i}(z_i = 1) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ when $\beta = 0$. The characteristic's presence is encouraged in neighboring sites when $\beta > 0$, and discouraged when $\beta < 0$.

Notation 13. We use notation, for $\mathcal{A} \subset \mathcal{S}$

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

Definition 14. Local characteristics of a random field Z on $\mathcal{S}$ with values in $\mathcal{Z}$ are the conditionals

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

Example 15. (Cont. Example 12) The local characteristics of the Ising model in (1.1) are

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

2. Compatibility of conditional distributions

Note 16. Here, we discuss how to represent a joint probability distribution via its full conditionals. We need this for model building purposes.

Definition 17. 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 18. 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 19. 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 20. In what follows, we discuss necessary and sufficient conditions regarding compatibility.

Proposition 21. *(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 $\int u(x)\,dG(x|y) < \infty$

Proof. Omitted[1]. □

Note 22. 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 23. The conditionals $x|y \sim \mathrm{N}(a+by, \sigma^2+\tau^2y^2)$ and $y|x \sim \mathrm{N}(c+dx, \tilde{\sigma}^2+\tilde{\tau}^2x^2)$ are compatible if $\tau^2 = \tilde{\tau}^2 = 0$ and $d/\tilde{\sigma}^2 = b/\sigma^2$.

Solution. See Exercise 24 in the Exercise sheet.

Note 24. Proposition 21 can be extended to more dimensions. For more info see (Arnold, B. C., & Press, S. J. (1989).)

Note 25. The following theorem shows that local characteristics can determine the entire distribution in certain cases.

Theorem 26. *(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*

$$\frac{pr_Z(z)}{pr_Z(z^*)} = \prod_{i=1}^{n} \frac{pr_i\left(z_i|z_1, ..., z_{i-1}, z^*_{i+1}, ..., z^*_n\right)}{pr_i\left(z^*_i|z_1, ..., z_{i-1}, z^*_{i+1}, ..., z^*_n\right)}, \quad \forall z, z^* \in \mathcal{Z}^{\mathcal{S}} \tag{2.1}$$

Proof. I will show that

$$\mathrm{pr}_Z(z) = \prod_{i=1}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1, ..., z_{i-1}, z^*_{i+1}, ..., z^*_n\right)}{\mathrm{pr}_i\left(z^*_i|z_1, ..., z_{i-1}, z^*_{i+1}, ..., z^*_n\right)} \mathrm{pr}_Z(z^*)$$

[1] See Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. Journal of the American Statistical Association, 84(405), 152-156.

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

$$\mathrm{pr}_Z(z) = \frac{\mathrm{pr}_n(z_n|z_1,...,z_{n-1})}{\mathrm{pr}_n(z_n^*|z_1,...,z_{n-1})}\mathrm{pr}_Z(z_1,...,z_{n-1},z_n^*) \tag{2.2}$$

Proposition P_1 is true

$$\mathrm{pr}_Z(z_1,...,z_{n-1},z_n^*) = \frac{\mathrm{pr}_{n-1}(z_{n-1}|z_1,...,z_{n-2},z_n^*)}{\mathrm{pr}_{n-1}\left(z_{n-1}^*|z_1,...,z_{n-2},z_n^*\right)}\mathrm{pr}_Z\left(z_1,...,z_{n-2},z_{n-1}^*,z_n^*\right)$$

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 (2.1) is correct according to the induction principle. □

Note 27. Theorem 26 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 21.

3. Gaussian Autoregressive models

We present two basic spatial models, the CAR and SAR, able to represent spatial dependency.

Definition 28. Adjacency matrix N is called a matrix N with $[N]_{i,j} = 1(i \sim j)$ (it is implied that $[N]_{i,i} = 0$) for some symmetric neighborhood relation $\sim$ on $\mathcal{S}$. It aims at spatially connecting unites i and j.

Definition 29. Proximity matrix W is called a matrix W which aims at spatially connecting unites i and j in some fashion for some symmetric neighbourhood relation $\sim$ on $\mathcal{S}$. Usually $[W]_{i,i} = 0$

3.1. **Conditional autoregressive models (CAR).**

Definition 30. "Gaussian" Conditional autoregressive model, CAR, assumes that the local characteristics $\{\mathrm{pr}_i(z_i|z_{\mathcal{S}-i})\}$ are Gaussian distributions with mean $\mathrm{E}(Z_i|Z_{\mathcal{S}-i}) = \mu_i + \sum_{j\neq i} b_{i,j}(Z_j - \mu_j)$ and variance $\mathrm{Var}(Z_i|Z_{\mathcal{S}-i}) = \kappa_i$ for $i \in \mathcal{S}$;

$$Z_i|z_{\mathcal{S}-i} \sim \mathrm{N}\left(\mu_i + \sum_{j\neq i} b_{i,j}(Z_j - \mu_j), \kappa_i\right), \quad \forall i \in \mathcal{S} \tag{3.1}$$

Proposition 31. *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 30), $I - B$ is non-singular, and $(I-B)^{-1}K > 0$, then the joint distribution of Z is*

$$Z \sim N\left(\mu, (I-B)^{-1}K\right). \tag{3.2}$$

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 (3.1) are compatible with the joint distribution $\mathrm{pr}_Z(z)$. By using Besag's factorization theorem (Theorem 26) with reference state/configuration $z^* = 0$ we get

$$\begin{aligned}
\mathrm{pr}_Z(z) &= \prod_{i=1}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z^*_{i+1}=0,...,z^*_n=0\right)}{\mathrm{pr}_i\left(z^*_i=0|z_1,...,z_{i-1},z^*_{i+1}=0,...,z^*_n=0\right)}\mathrm{pr}_Z(z^*=0) \\
&= \prod_{i=1}^{n} \frac{\mathrm{N}\left(z_i|\sum_{j<i} b_{i,j}z_j + 0, \kappa_i\right)}{\mathrm{N}\left(0|\sum_{j<i} b_{i,j}z_j + 0, \kappa_i\right)}\mathrm{pr}_Z(z^*=0) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i - \sum_{j<i} b_{i,j}z_j\right)^2 + \frac{1}{2\kappa_i}\left(0 - \sum_{j<i} b_{i,j}z_j\right)^2\right) \\
&= \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i^2 - 2z_i\sum_{j<i} b_{i,j}z_j\right)\right)\mathrm{pr}_Z(z^*=0) \\
&= \exp\left(-\sum_i \frac{z_i^2}{2\kappa_i} + \frac{1}{2}2\sum_i\sum_{j<i}\frac{b_{i,j}}{\kappa_i}z_iz_j\right)\mathrm{pr}_Z(z^*=0) \\
&= \exp\left(-\frac{1}{2}z^\top K^{-1}z + \frac{1}{2}z^\top K^{-1}Bz\right)\mathrm{pr}_Z(z^*=0) = \exp\left(-\frac{1}{2}z^\top\left[K^{-1}(I-B)\right]z\right)\mathrm{pr}_Z(z^*=0) \\
&= \mathrm{N}\left(z|0,(I-B)^{-1}K\right)
\end{aligned} \tag{3.3}$$

Recovering the mean from (3.3), 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 32. 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 $B = \phi N$ where $\phi > 0$ and N is an adjacency 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 33. The system in (3.2) can be rewritten as

$$Z = \mu + B(Z-\mu) + E \Longleftrightarrow E = (I-B)(Z-\mu) \tag{3.4}$$

by setting $E = (I-B)(Z-\mu)$. The distribution of Z in (3.2) induces a distribution on E as $E \sim \mathrm{N}\left(0, K(I-B)^{\top}\right)$ because

$$\begin{aligned}
\mathrm{E}(E) &= \mathrm{E}((I-B)(Z-\mu)) = (I-B)\,\mathrm{E}(Z-\mu) = 0 \\
\mathrm{Var}(E) &= \mathrm{Var}((I-B)Z) = (I-B)\,\mathrm{Var}(Z)\,(I-B)^{\top} = (I-B)(I-B)^{-1}K(I-B)^{\top}
\end{aligned}$$

3.2. **Simultaneus Autoregressive (SAR) models.**

Note 34. CAR sets the AR relation, and specifies the distribution on Z which induces the distribution on E; see (3.4). SAR does does the reverse; sets the same AR relation but it specifies the distribution on E which induces the distribution on Z –this is more might be more intuitive (?).

Definition 35. Consider discrete set of sites $\mathcal{S} = \{s_i; i = 1, ..., n\}$. Consider a random field $Z = (Z_s; s \in \mathcal{S})^{\top} = (Z_i = Z(s_i); i = 1, ...n)^{\top}$ on the discrete set of indexes $\mathcal{S}$ with values in $\mathcal{Z}$. Define

$$Z = \mu + \tilde{B}(Z-\mu) + E \Longleftrightarrow E = \left(I-\tilde{B}\right)(Z-\mu)$$

Assume that matrix $\tilde{B}$ is such that $\left(I-\tilde{B}\right)^{-1}$ exists, and $\left[\tilde{B}\right]_{i,i} = 0$. Assume that $E = (E_i; i = 1, ...n)$ is an n-dimensional Gaussian random vector $E \sim \mathrm{N}_n(0, \Lambda)$ with $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ whose elements are indexed by $\mathcal{S}$. Then we say that Z follows a "Gaussian" Simultaneous Autoregressive, SAR, model.

Proposition 36. *The joint distribution of Z following the SAR model in Definition 35 is*

$$Z \sim N\left(\mu, \left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}\right) \tag{3.5}$$

Proof. Z is a linear combination of Gaussian random vecors, hence if follows a Gaussian distribution. It's mean and variance are

$$\mathrm{E}(Z) = \mathrm{E}\left(\left(I-\tilde{B}\right)^{-1}E+\mu\right) = \mu,$$

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\left(I-\tilde{B}\right)^{-1}E+\mu\right) = \left(I-\tilde{B}\right)^{-1}\mathrm{Var}(E)\left(I-\tilde{B}^{\top}\right)^{-1} = \left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}$$

□

3.3. **A comparison between CAR and SAR.**

Note 37. We compare the use and flexibility of the two models.

Remark 38. From (3.2) and (3.5), CAR and SAR are equivalent iff

$$\underbrace{(I-B)^{-1}K}_{\text{CAR}} = \underbrace{\left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}}_{\text{SAR}}$$

Note 39. Following, we show that any SAR model can be written as a CAR model, however the converse is not always true.

Proposition 40. *Any positive-definite covariance matrix Σ can be expressed as the covariance matrix of a CAR model $\Sigma_{CAR} = (I-B)^{-1}K$, for a unique pair of matrices B and K where $(I-B)$ is non-singular and K is diagonal.*

Proof. (This proof can be considered as an exercise for understanding CAR) Express

$$\Sigma^{-1} = D - R$$

for

$$[D]_{i,j} = \begin{cases} [\Sigma^{-1}]_{i,i} & i=j \\ 0 & i\neq j \end{cases}, \text{ and } [R]_{i,j} = \begin{cases} 0 & i=j \\ -[\Sigma^{-1}]_{i,j} & i\neq j \end{cases}$$

then

$$\Sigma = (D-R)^{-1} = \left(D\left(I-D^{-1}R\right)\right)^{-1} = \left(I-D^{-1}R\right)^{-1}D^{-1}$$

Now define $B = D^{-1}R$ and $K = D^{-1}$, and you get $\Sigma = \Sigma_{\mathrm{CAR}}$. Now regarding the uniqueness, assume there is another pair of $\mathring{B}$, and $\mathring{K}$ such that $\Sigma_{\mathrm{CAR}} = \left(I-\mathring{B}\right)^{-1}\mathring{K}$. Then

$$\mathrm{diag}\left(\Sigma^{-1}\right) = \mathrm{diag}\left(\Sigma_{\mathrm{CAR}}^{-1}\right) = \mathrm{diag}\left(\mathring{K}^{-1}\left(I-\mathring{B}\right)\right) = \mathrm{diag}\left(\mathring{K}^{-1}\right)$$

and similarly $\mathrm{diag}(\Sigma^{-1}) = \mathrm{diag}(K^{-1})$. Hence it has to be $\mathring{K} = K$ because both are diagonal matrices. Then it is

$$\left(I-\mathring{B}\right)^{-1}\mathring{K} = (I-B)^{-1}K \overset{\mathring{K}=K}{\Longleftrightarrow} \mathring{B} = B.$$

So the representation is unique. □

Proposition 41. *Any positive-definite covariance matrix Σ can be expressed as the covariance matrix of a SAR model $\Sigma_{SAR} = \left(I - \tilde{B}\right)^{-1} \Lambda \left(I - \tilde{B}^\top\right)^{-1}$ for a (non-unique) pair of matrices $\tilde{B}$ and Λ where $\left(I - \tilde{B}\right)$ is non-singular, $\left[\tilde{B}\right]_{i,i} = 0$, and Λ is diagonal.*

Proof. (This proof can be considered as an exercise for understanding SAR) Express

$$\Sigma^{-1} = LL^\top$$

where L is a lower triangular matrix with $[L]_{i,i} > 0$. Such matrix decomposition can be done by Cholesky decomposition, square-matrix decomposition, etc... and hence it is not always unique. Then

$$\Sigma = \left(LL^\top\right)^{-1} = L^{-\top} L^{-1}$$

Now express, $L = D - C$ for

$$[D]_{i,j} = \begin{cases} [L]_{i,i} & i = j \\ 0 & i \neq j \end{cases}, \text{ and } [C]_{i,j} = \begin{cases} 0 & i = j \\ -[L]_{i,j} & i \neq j \end{cases}$$

then

$$\begin{aligned} \Sigma &= (D - C)^{-\top} (D - C)^{-1} = \left(I - D^{-1}C\right)^{-\top} D^{-\top} D^{-1} \left(I - D^{-1}C\right)^{-1} \\ &= \left(I - C^\top D^{-\top}\right)^{-1} D^{-\top} D^{-1} \left(I - \left(C^\top D^{-\top}\right)^\top\right)^{-1} \end{aligned}$$

Set $\tilde{B} = C^\top D^{-\top}$ and $\Lambda = D^{-\top} D^{-1}$ and you get $\Sigma_{\text{SAR}} = \Sigma$ for non-unique pairs of $\tilde{B}$ and Λ. □

Proposition 42. *Any SAR model can be written as a unique CAR model.*

Proof. (This proof can be considered as an exercise for understanding CAR/sar) SAR and CAR are both Gaussian's with the same mean. SAR's variance matrix is positive definite, and hence it can be written in a unique manner as a CAR's variance matrix by Proposition 40. □

Proposition 43. *Any CAR model can be written as a non-unique SAR model.*

Proof. SAR and CAR are both Gaussian's with the same mean. CAR's variance matrix is positive definite, and hence it can be written in a non-unique manner as a SAR's variance matrix by Proposition 41. □

Example 44. Show that

(1) Z_i and E_j are independent for $i \neq j$ in Gaussian CAR
(2) Z_i and E_j are not necessarily independent for $i \neq j$ in Gaussian SAR

Solution.

(1) For Gaussian CAR,

$$\operatorname{Cov}(E,Z)=\operatorname{Cov}((I-B)Z,Z)=(I-B)\operatorname{Var}(Z)=(I-B)(I-B)^{-1}K=K$$

which is a diagonal; hence Z_i and E_j are independent for $i\neq j$.

(2) For Gaussian SAR,

$$\operatorname{Cov}(Z,E)=\operatorname{Cov}\left(\left(I-\tilde{B}\right)^{-1}E,E\right)=\left(I-\tilde{B}\right)^{-1}\operatorname{Var}(E)=\left(I-\tilde{B}\right)^{-1}\Lambda$$

which is not a diagonal matrix in general; hence Z_i and E_j may be dependent for $i\neq j$.

4. Related Random Fields with particular properties

Note 45. 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.

4.1. **Gibbs Random Fields.**

Notation 46. 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 47. 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

$$U_{\mathcal{A}}^{\mathcal{V}}(z)=\sum_{\{\forall\mathcal{B}\subseteq\mathcal{S}:\mathcal{A}\cap\mathcal{B}\neq\emptyset\}}V_{\mathcal{B}}(z_{\mathcal{B}}) \tag{4.1}$$

exists.

Definition 48. In Definition 47, the function $V_{\mathcal{A}}:\mathcal{Z}^{\mathcal{A}}\to\mathbb{R}$ is called potential on $\mathcal{A}$.

Definition 49. In Definition 47, the function $U_{\mathcal{A}}^{\mathcal{V}}(z)$ in (4.1) is called energy function of interaction potential $\mathcal{V}$ on $\mathcal{A}$ is called.

Definition 50. 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 51. This allow as to define a distribution corresponding to the energy.

Definition 52. 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

$$\text{pr}_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\setminus\mathcal{A}}\right) = \frac{1}{C_{\mathcal{A}}^{\mathcal{V}}\left(z_{\mathcal{S}\setminus\mathcal{A}}\right)} \exp\left(\underbrace{\sum_{\{\mathcal{B}\subseteq\mathcal{S}\,:\,\mathcal{A}\cap\mathcal{B}\neq\emptyset\}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)}_{=U_{\mathcal{A}}^{\mathcal{V}}(z)}\right), \quad z \in \mathcal{Z}^{\mathcal{S}} \tag{4.2}$$

Definition 53. The normalizing integral $C_{\mathcal{A}}^{\mathcal{V}}$ in (4.2) is called partition function.

Notation 54. For the marginal $\text{pr}_Z\left(z_{\mathcal{S}}\right)$ we will denote

$$\text{pr}_Z\left(z_{\mathcal{S}}\right) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(U_{\mathcal{S}}^{\mathcal{V}}\left(z\right)\right) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(\sum_{\mathcal{B}\subseteq\mathcal{S}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)\right),\ z \in \mathcal{Z}^{\mathcal{S}}$$

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

$$\text{pr}_Z\left(z_{\mathcal{S}}\right) = \frac{1}{C} \exp\left(\sum_{\mathcal{B}\subseteq\mathcal{S}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)\right),\ z \in \mathcal{Z}^{\mathcal{S}}$$

Example 55. (Ising model) In Example 12, the Ising model (1.1) has potentials

$$\begin{aligned}
V_{\emptyset}\left(z\right) &= 0 \\
V_{\{i\}}\left(z\right), &= \alpha z_i \,\forall i \in \mathcal{S} \\
V_{\{i,j\}}\left(z\right) &= \begin{cases} \beta z_i z_j & \text{if } i \sim j \\ 0 & \text{if } i \not\sim j \end{cases} \\
V_{\mathcal{A}}\left(z\right) &= 0, \text{ if } \text{ card}\left(\mathcal{A}\right) > 2
\end{aligned}$$

it has energy function

$$U\left(z\right) := U_{\mathcal{S}}^{\mathcal{V}}\left(z_{\mathcal{S}}\right) = \alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i\in\mathcal{S}, j\in\mathcal{S}: i\sim j\}} z_i z_j$$

and it has energy function conditional on $\mathcal{S}\setminus\mathcal{B}$

$$U_{\mathcal{B}}^{\mathcal{V}}\left(z_{\mathcal{B}} | z_{\mathcal{S}\setminus\mathcal{B}}\right) = \alpha \sum_{i\in\mathcal{B}} z_i + \beta \sum_{\{i\in\mathcal{B}, j\in\mathcal{S}: i\sim j\}} z_i z_j$$

Note 56. In what follows we discuss identifiability matters related to the potential.

Definition 57. 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 $V_{\mathcal{B}}(z) = 0$.

Note 58. In (4.2), the mapping $\mathcal{V} \to \mathrm{pr}_Z$ is in general non-identifiable because (4.2) 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 59. One way to make $\mathcal{V}$ identifiable is to impose restriction

$$\forall \mathcal{A} \neq \emptyset,\ V_{\mathcal{A}}(z) = 0,\ \text{if for some}\ \ i \in \mathcal{A},\ z_i = \zeta \tag{4.3}$$

Notation 60. 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 61. 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 62. The following theorem uniquely associates potentials satisfying (4.3) with (4.2) with regards a normalizing point.

Theorem 63. *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*

$$\begin{aligned} V_{\mathcal{A}}(z_{\mathcal{A}}) &= \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} U_{\mathcal{B}}^{\mathcal{V}}\left(z^{[\mathcal{B},\zeta]}\right),\ z \in \mathcal{Z}^{\mathcal{S}} \\ &= \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} \log\left(pr_Z\left(z^{[\mathcal{B},\zeta]}\right)\right),\ z \in \mathcal{Z}^{\mathcal{S}} \end{aligned} \tag{4.4}$$

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 64. *From Theorem 63, for all $i \in \mathcal{A}$ it is*

$$V_{\mathcal{A}}(z_{\mathcal{A}}) = \sum_{\mathcal{B} \subseteq \mathcal{A}} (-1)^{Card(\mathcal{A} \setminus \mathcal{B})} \log\left(pr_i\left(z_i^{[\mathcal{B},\zeta]} | z_{\mathcal{S}\setminus\{i\}}^{[\mathcal{B},\zeta]}\right)\right),\ z \in \mathcal{Z}^{\mathcal{S}} \tag{4.5}$$

Note 65. The following example explains the use of Theorem 63 in terms of the Definition 47.

Example 66. 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 (4.1) with

$$\begin{aligned} V_{\{1\}}(z_1) &= U(z_1, \zeta) - U(\zeta, \zeta) \\ V_{\{2\}}(z_2) &= U(\zeta, z_2) - U(\zeta, \zeta) \\ V_{\{1,2\}}(z_1, z_2) &= U(z_1, z_2) - U(z_1, \zeta) - U(\zeta, z_2) + U(\zeta, \zeta) \end{aligned}$$

by (4.4).

Example 67. (Ising model) We revisit Example 12 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 60, for instance,

$$z^{[\emptyset,\zeta]} = \left(\overbrace{\zeta, ..., \zeta}^{n \text{ times}}\right)^\top; \qquad z^{[\{i\},\zeta]} = \left(\zeta, ..., \zeta, \overset{i\text{th location}}{\overbrace{z_i}}, \zeta, ..., \zeta\right)^\top;$$

$$z^{[\{i,j\},\zeta]} = \left(\zeta, ...\zeta, \overset{i\text{th location}}{\overbrace{z_i}}, \zeta, ..., \zeta, \overset{j\text{th location}}{\overbrace{z_j}}, ..., \zeta\right)^\top; \quad z^{[\mathcal{S},\zeta]} = (z_1, ..., z_n)^\top;$$

It is $V_\emptyset = 0$ by definition. By using Theorem 63 and considering a reference point $\zeta = 0$, we get

$$V_{\{i\}}(z) = (-1)^{1-1} U\left(z^{[\{i\},\zeta]}\right) + (-1)^{1-0} U\left(z^{[\emptyset,\zeta]}\right) = a z_i\,, \tag{4.6}$$

for any $i \in \mathcal{S}$ and

$$V_{\{i,j\}}(z) = \left[(-1)^{2-2} U\left(z^{[\{i,j\},\zeta]}\right)\right] + \left[(-1)^{2-1} U\left(z^{[\{i\},\zeta]}\right)\right] + \left[(-1)^{2-1} U\left(z^{[\{j\},\zeta]}\right)\right] + \left[(-1)^{2-0} U\left(z^{[\emptyset,\zeta]}\right)\right] = [\alpha z_i + \alpha z_j + \beta z_i z_j] + [-\alpha z_i] + [-\alpha z_j] + [0] = \beta z_i z_j \tag{4.7}$$

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 $\text{card}(\mathcal{A}) > 2$.

4.2. **Markov Random Fields.**

Note 68. 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 69. 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 70. Recall the Ising model in Example 67 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 $\sim$ (4.7) or when $\mathcal{A}$ a singleton (4.6). Consequently, its local characteristics $\text{pr}_i\left(z_i | z_{\mathcal{S}\setminus\{i\}}\right)$ depend only on the values of the sites $j \in \mathcal{S}\setminus\{i\}$ that satisfy $\sim$.

Definition 71. 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 72. 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 is a 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.

$$\text{pr}_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\setminus\mathcal{A}}\right) = \text{pr}_Z\left(z_{\mathcal{A}} | z_{\partial\mathcal{A}}\right) \tag{4.8}$$

when $\text{pr}_Z\left(z_{\mathcal{S}\setminus\mathcal{A}}\right) > 0$

Note 73. Definition 72 implies that (4.8) becomes

$$\text{pr}_Z\left(z_i | z_{-i}\right) = \text{pr}_Z\left(z_i | z_{\partial\{i\}}\right), \quad \forall i \in \mathcal{S} \tag{4.9}$$

when $\text{pr}_Z\left(z_{\mathcal{S}\setminus\{i\}}\right) > 0$

Definition 74. 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.

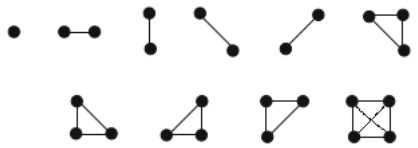


FIGURE 4.1. Examples of cliques

Notation 75. 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 76. 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 77. *(Hammersley–Clifford) Let Z 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*

$$pr_Z(z) \propto \prod_{\mathcal{C} \in \boldsymbol{\mathcal{C}}} \varphi_{\mathcal{C}}(z_{\mathcal{C}}) \tag{4.10}$$

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 63, Z is Gibbs with a canonical potential (4.4)*

$$V_{\mathcal{A}}(z_{\mathcal{A}}) = \sum_{\mathcal{A} \subseteq \mathcal{B}} (-1)^{Card(\mathcal{A}\setminus\mathcal{B})} \log\left(pr_Z\left(z^{[\mathcal{B},\zeta]}\right)\right),$$

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 (4.10) holds.

for $\Longleftarrow$: *By using (4.2), I can write*

$$pr_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\setminus\mathcal{A}} \right) = \frac{1}{C_{\mathcal{A}}\left(z_{\mathcal{S}\setminus\mathcal{A}} \right)} \exp\left(U_{\mathcal{A}}(z) \right)$$

where

$$U_{\mathcal{A}}(z) = \sum_{\{\mathcal{C} \subseteq \mathcal{S} \,:\, \mathcal{A} \cap \mathcal{C} \neq \emptyset\}} V_{\mathcal{C}}(z_{\mathcal{C}})$$

depends only on $\{z_i : i \in \mathcal{A} \cup \partial\mathcal{A}\}$ as $pr_Z(\cdot)$ is a Markov Random Field.

Note 78. Because $\mathrm{pr}_Z(z) > 0$, the Markov Random Field in (4.10) is a Gibbs Random Field as

$$\mathrm{pr}_Z(z) \propto \exp\left(\sum_{\mathcal{C} \in \boldsymbol{\mathcal{C}}} \log\left(\varphi_{\mathcal{C}}(z_{\mathcal{C}}) \right) \right)$$

with non-zero interaction potentials restricted to cliques $\mathcal{C} \in \boldsymbol{\mathcal{C}}$.

Note 79. Essentially Theorem 77 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 $\mathrm{pr}_Z(\cdot)$ 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 80. (Ising model; Cont. Example 12). The joint PMF of the Ising model in Example 12 is

$$\begin{aligned}\mathrm{pr}(z) =& \frac{\exp\left(\alpha\sum_{i\in\mathcal{S}} z_i + \beta\sum_{\{i,j\}:i\sim j} z_i z_j\right)}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}} z_i + \beta\sum_{\{i,j\}:i\sim j} z_i z_j\right)} \\ =& \frac{1}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}} z_i + \beta\sum_{\{i,j\}:i\sim j} z_i z_j\right)} \prod_{i\in\mathcal{S}}\exp(\alpha z_i)\prod_{i\in\mathcal{S}}\prod_{j:j\sim i}\exp(\beta z_i z_j)\end{aligned}$$

I can find that

$$\varphi_{\emptyset} = 1/\sum_{z\in\mathcal{Z}^{\mathcal{S}}}\exp\left(\alpha\sum_{i\in\mathcal{S}} z_i + \beta\sum_{\{i,j\}:i\sim j} z_i z_j\right)$$

$$\varphi_{\{i\}}\left(z_{\{i\}}\right) = \exp(\alpha z_i), \qquad \forall i \in \mathcal{S} \tag{4.11}$$

$$\begin{aligned}\varphi_{\{i,j\}}\left(z_{\{i,j\}}\right) &= \exp(\beta z_i z_j), && \forall i,j\in\mathcal{S} \quad \text{s.t.} \quad i\sim j \qquad (4.12)\\ \varphi_{\{i,j\}}\left(z_{\{i,j\}}\right) &= 1, && \forall i,j\in\mathcal{S} \quad \text{s.t.} \quad i\not\sim j \\ \varphi_{\mathcal{A}}(z_{\mathcal{A}}) &= 1, && \forall \mathcal{A}\subset\mathcal{S} \quad \text{s.t.} \quad \mathrm{card}(\mathcal{A}) > 2\end{aligned}$$

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 (4.11) and (4.12) and correspond to cliques.

Part 2. Model building for aerial data & related inference

5. Automodels

Note 81. 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 82. 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.

5.1. **Multi-parameter automodels.**

Theorem 83. *Consider Markov random field Z that takes values in $\mathcal{Z}$ on a finite set of points $\mathcal{S}$ and has marginal probability*

$$pr_Z(z) = \frac{\exp(U(z))}{\int \exp(U(z))\, dz}, \quad z \in \mathcal{Z}^{\mathcal{S}}. \tag{5.1}$$

Consider some fixed normalization configuration $\zeta = (\zeta, ..., \zeta)^\top \in \mathcal{Z}^{\mathcal{S}}$. Assume that the following assumptions are satisfied:

(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 exponential family distributions

$$\log(pr_i(z_i | z_{-i})) = (A_i(z_{-i}))^\top B_i(z_i) + C_i(z_i) + D_i(z_{-i}), \tag{5.2}$$

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 $D_i(\zeta, ..., \zeta) = 0$.

(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, necessarily,

(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

$$V_i(z_i) = (\alpha_i)^\top B_i(z_i) + C_i(z_i) \tag{5.3}$$

$$V_{i,j}(z_i, z_j) = (B_i(z_i))^\top \beta_{i,j} B_j(z_j) \tag{5.4}$$

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 84. Automodel is called the model satisfying the assumptions of Theorem 83.

Definition 85. Univariate automodel is the automodel with $\ell = 1$ in Theorem 83.

Definition 86. Multi-parameter automodel with $\ell \geq 1$ in Theorem 83.

Remark 87. In the univariate automodel, $\ell = 1$, assumption 3 in Theorem 83 is not needed; it is automatically satisfied as B_i's are not identically zero. Yet, for $\ell = 1$, (5.3) and (5.4) become

$$V_i(z_i) = \alpha_i B_i(z_i) + C_i(z_i) \tag{5.5}$$

$$V_{i,j}(z_i, z_j) = \beta_{i,j} B_i(z_i) B_j(z_j) \tag{5.6}$$

5.2. **Besag auto-models.**

Definition. Z follows a Besag's auto-model if Z is real-valued and its joint distribution $\mathrm{pr}_Z(z)$ is given by

$$\mathrm{pr}_Z(z) = \frac{1}{C} \exp\left(\sum_{i \in \mathcal{S}} V_i(z_i) + \sum_{\{i,j\} \in \mathcal{S}^2 : i \sim j} \beta_{i,j} z_i z_j \right), \ z \in \mathcal{Z}_{\mathcal{S}} \tag{5.7}$$

with $\beta_{i,j} = \beta_{j,i}$ for all $i, j \in \mathcal{S}$.

Note 88. The following allows us to define a Markov Random Field model from a set of conditional distributions (characteristics) whose compatibility is automatically satisfied.

Proposition 89. *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 (5.2) with $B_i(z_i) = z_i$ for $i \in \mathcal{S}$, then they are compatible a Besag's auto-model with distribution (5.7) if $\beta_{i,j} = \beta_{j,i}$ for all $i, j \in \mathcal{S}$.

Proof. For

$$\mathrm{pr}_i(z_i | z_{-i}) = \exp\left(A_i(z_{-i}) z_i + C_i(z_i) + D_i(z_{-i}) \right)$$

it is

$$V_i(z_i) = \alpha_i B_i(z_i) + C_i(z_i) = \alpha_i z_i + C_i(z_i)$$

$$V_{i,j}(z_i, z_j) = \beta_{i,j} B_i(z_i) B_j(z_j) = \beta_{i,j} z_i z_j$$

so

$$\mathrm{pr}_Z(z) \propto \exp\left(\sum_i \left[\alpha_i z_i + C_i(z_i) \right] + \sum_{i \sim j} \beta_{i,j} z_i z_j \right), \ z \in \mathcal{Z}_{\mathcal{S}}$$

□

Example 90. (Logistic automodel / Ising model) Consider that $Z(s)$ 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

$$z_i | z_{-i} \sim \text{Logistic}\left(\theta_i\left(z_{-i}\right)\right), \quad i \in \mathcal{S}.$$

Hint:: The PMF of Logistic distribution $x|\theta \sim \text{Logistic}(\theta)$ can be written as $\text{pr}(x|\theta) = \left(1 - \exp(x\theta)\right)^{-1} 1\left(x \in \{0, 1\}\right)$.

Then the characteristics are

$$\text{pr}_i\left(z_i | z_{-i}\right) = \frac{\exp\left(z_i \theta_i\left(z_{-i}\right)\right)}{1 + \exp\left(\theta_i\left(z_{-i}\right)\right)} 1\left(z_i \in \{0, 1\}\right) \tag{5.8}$$

Now, let's parameterize $\{\theta_i(\cdot)\}$ as

$$\theta_i\left(z_{-i}\right) = \alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j \tag{5.9}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j} = \beta_{j,i}$. Then (5.8) becomes

(5.10)

$$\log\left(\text{pr}_i\left(z_i | z_{-i}\right)\right) = \underbrace{z_i}_{B_i(z_i)} \underbrace{\left(\alpha_i + \sum_{j\sim i} \beta_{i,j} \overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})} + \underbrace{0}_{C_i(z_i)} + \underbrace{\left(-\log\left(1 + \exp\left(\alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j\right)\right)\right)}_{D_i(z_{-i})}$$

Notice that all the conditionals $z_i|z_{-i}$ follow an Exponential family with

$$\begin{aligned}
A_i\left(z_{-i}\right) =& z_i \\
B_i\left(z_i\right) =& \alpha_i + \sum_{j\sim i} \beta_{i,j} B_i\left(z_j\right) \\
C_i\left(z_i\right) =& 0 \\
D_i\left(z_{-i}\right) =& -\log\left(1 + \exp\left(\alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j\right)\right)
\end{aligned}$$

Also, I can get $C_i(\zeta) = 0$ and $D_i(\zeta, ..., \zeta) = 0$ by considering a reference point $\zeta = 0$. From Theorem 89, (5.8) with (5.9), the conditionals $z_i|z_{-i}$ are compatible as a Besag auto-model

with marginal distribution

$$\text{pr}_Z(z) \propto \exp\left(\overbrace{\sum_i \alpha_i \underbrace{z_i}_{B_i(z_i)} + \sum_i \underbrace{\sum_{j:j\sim i} \beta_{i,j} z_i z_j}_{\sum_{\{i,j\}:j\sim i}}}^{U(z)=}\right) \tag{5.11}$$

I observe that:

- Here the Ising model has spatially dependent coefficients $\{\alpha_i\}$ and $\{\beta_{i,j}\}$, unlike the Ising model in Example 12 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 $\text{pr}_i(z_i|z_{-i}) = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}$.
- 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 91. (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

$$z_i|z_{-i} \sim \text{Poisson}(\lambda_i(z_{-i}))$$

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

$$\text{pr}_i(z_i|z_{-i}) = \frac{1}{z_i!}(\lambda_i(z_{-i}))^{z_i} \exp(-\lambda)\, 1(z_i \in \mathbb{N}) \tag{5.12}$$

Now, let's parameterize $\{\lambda_i(\cdot)\}$ as

$$\log(\lambda_i(z_{-i})) = \alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j \tag{5.13}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j} = \beta_{j,i}$. So (5.12) becomes

$$\log(\text{pr}_i(z_i|z_{-i})) = \underbrace{z_i}_{B_i(z_i)} \underbrace{\left(\alpha_i + \sum_{j\sim i} \beta_{i,j} \overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})} + \underbrace{\log(z_i!)}_{C_i(z_i)} + \underbrace{0}_{D_i(z_{-i})}$$

with

$$\begin{aligned} A_i\left(z_{-i}\right) =& z_i \\ B_i\left(z_i\right) =& \alpha_i + \sum_{j\sim i} \beta_{i,j} B_i\left(z_j\right) \\ C_i\left(z_i\right) =& \log\left(z_i!\right) \\ D_i\left(z_{-i}\right) =& 0 \end{aligned}$$

I can notice that all the conditionals $z_i|z_{-i}$ follow exponential of exponential. Also, I can get $C_i(\zeta) = 0$ and $D_i(\zeta, ..., \zeta) = 0$ by considering a reference point $\zeta = 0$. From Theorem 89, (5.12) with (5.13), the the conditionals $z_i|z_{-i}$ are compatible as a Besag auto-model with marginal distribution

$$\mathrm{pr}_Z(z) \propto \exp\left(\overbrace{\sum_i \left(\underbrace{\alpha_i \overbrace{z_i}^{B_i(z_j)} + \overbrace{\log\left(z_i!\right)}^{C_i(z_i)}}_{V_i(z_i)}\right) + \sum_i \sum_{j:j\sim i} \beta_{i,j} z_i z_j}^{U(z)=}\right)$$

or otherwise the energy function is

$$U(z) = \sum_i \left(\alpha_i z_i + \log\left(z_i!\right)\right) + \sum_{j\sim i} \beta_{i,j} z_i z_j$$

Furthermore, to ensure that $U(z)$ is admissible, we need to consider additional conditions. I observe that

$$\sum_{z\in\mathbb{N}^{\mathcal{S}}} \exp\left(U(z)\right) = \sum_{z\in\mathbb{N}^{\mathcal{S}}} \prod_i \left(\exp\left(\alpha_i z_i + \log\left(z_i!\right)\right) + \sum_{j\sim i} \beta_{i,j} z_i z_j\right)$$

- If we use additional condition $\beta_{i,j} \leq 0$ then

  $$\sum_{z\in\mathbb{N}^{\mathcal{S}}} \exp\left(U(z)\right) \leq \sum_{z\in\mathbb{N}^{\mathcal{S}}} \prod_i \left(\exp\left(\alpha_i z_i + \log\left(z_i!\right)\right)\right) = \sum_{z\in\mathbb{N}^{\mathcal{S}}} \prod_i \frac{1}{z_i!} \exp\left(\alpha_i z_i\right) < \infty$$

  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 92. 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 31 could have been derived from Theorem 83 as well.

5.3. **Parameterization matters in automodels.**

Remark 93. The unknown parameter vector $\theta = ((\alpha_i; i \in \mathcal{S}), (\beta_{i,j}; i, j \in \mathcal{S}))$ in the automodel (5.1) (eg Besag's automodel (5.7)) can be further parameterised to have a particular structure without the need to consider any additional constrains in Theorem 83.

Remark 94. The dimensionality of θ 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 $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$ reducing its dimentionality.

- For instance, one may specify

$$\alpha_i = aw_i, \quad \text{and} \quad \beta_{i,j} = b_i c_j; \text{ for } i, j \in \mathcal{S},$$

with some known weights $\{w_i;\ i \in \mathcal{S},\}$ and unknown $\{a, b_i, c_j;\ i, j \in \mathcal{S}\}$. Then learning $\text{Card}(\mathcal{S})(1 + \text{Card}(\mathcal{S}))$ unknown parameter $\{\alpha_i, \beta_{i,j};\ i, j \in \mathcal{S},\}$ reduces to learning just $1 + 2\text{Card}(\mathcal{S})$ 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 95. When observable covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$ are available, one could "link" time to the model via the parameters $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$ and make if time dependent.

- For instance by setting

$$\alpha_i = a_i + \sum_{k=1}^{p} d_k x_{i,k}, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j}; \text{ for } i, j \in \mathcal{S}, \tag{5.14}$$

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.

Remark 96. Perhaps, in Remark 95, one (or many) of the observable covariates in vector x_i for $i \in \mathcal{S}$ can be "time" t dependent. One could "link" them to the model via the parameters $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$, and make the automodel dynamical (aka spatio-temporal).

- For instance, if one consider $x_i = (t_i, t_i^2)^\top$ for $i \in \mathcal{S}$ and set

$$\alpha_i = a_i + d_1 t_i + d_2 (t_i)^2, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j}; \text{ for } i, j \in \mathcal{S},$$

essentially he/she makes the model dynamic (or space-time, or spatio-temporal)

Of course, how to parameterize the covariates in (5.14) is problem dependent, (similar to the linear regression) and a model assessment/comparison is often required.

Example 97. In Example 91, given observable covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$, one may set (5.13) as

$$\log (\lambda_i (z_{-i})) = \left[a_i + \sum_{k=1}^{p} d_k x_{i,k} \right] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j \right] \tag{5.15}$$

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 91). 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 (5.15).

Example 98. In Example 91, 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

$$\log (\lambda_i (z_{-i}, t_i)) = [a_i + d_1 t_i] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j \right] \tag{5.16}$$

and be interested in learning the unknown parameters $\{a_i\}$, d_1, and $\{\beta_{i,j}\}$. Obviously, the resulted model is space-time.

6. Frequentist modeling and likelihood based inference

Note 99. 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

$$Z \sim \text{pr}_Z (Z|\theta) \tag{6.1}$$

labeled by unknown parameter vecrtor θ. Parametric and predictive inference can be performed based based on the associated likelihood or its approximation PsuedoLikelihood.

Note 100. For easy of presentation we assume that the observables Z are a realization of an automodel and hence their sampling distribution (6.1) is that of the automodel (5.1) with potentials 5.3 and 5.4, and unknown parameter $\theta = (\{\alpha_i\}, \{\beta_{i,j}\})^\top$.

6.1. **MLE: Maximum likelihood estimation.**

Note 101. We describe the maximum likelihood estimation in the automodel framework.

Remark 102. 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

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\mathrm{pr}_Z\left(Z \middle| \{\alpha_i\}, \{\beta_{i,j}\}\right)\right) \tag{6.2}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i, j \in \mathcal{S}$$

$$\text{...and any other problem specific restrictions}$$

where $\mathrm{pr}_Z\left(Z \middle| \{\alpha_i\}, \{\beta_{i,j}\}\right)$ is the joint distribution (5.7) given the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$.

Example 103. (Logistic automodel / Ising model) Assume that observables Z follow the Logistic automodel (5.11) in Example 90. Computing MLE $\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}$ of $\{\alpha_i\}, \{\beta_{i,j}\}$ requires

$$\begin{aligned}\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\log\left(\mathrm{pr}_Z\left(Z \middle| \{\alpha_i\}, \{\beta_{i,j}\}\right)\right)\right) \\ &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j - \log\left(C\left(\{\alpha_i\}, \{\beta_{i,j}\}\right)\right)\right) \end{aligned} \tag{6.3}$$

where

$$C\left(\{\alpha_i\}, \{\beta_{i,j}\}\right) = \sum_{\forall z \in \mathcal{Z}^{\mathcal{S}}} \exp\left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j\right) \tag{6.4}$$

is the normalizing constant. Optimization in (6.3) can be done numerically by using a recursive optimization algorithm such as Newton-Raphson.

Note 104. The optimization problem (6.2) can be too computationally expensive. For instance, in Example 103, a recursive optimization algorithm, like Newton-Raphson, requires several iterations. At each iteration the evaluation of the (parameter dependent) constant (6.4) 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 (6.4) 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.

6.2. **MPLE: Maximum pseudo likelihood estimation.**

Note 105. 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 106. Then (6.2) 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

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}} \left(\prod_{i\in\mathcal{S}} \text{pr}_Z(Z_i|Z_{-i},\theta)\right) \tag{6.5}$$

$$= \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}} \left(\sum_{i\in\mathcal{S}} \log\left(\text{pr}_Z(Z_i|Z_{-i},\theta)\right)\right) \tag{6.6}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i,j \in \mathcal{S}$$

...and any other problem specific restrictions

Example 107. (Logistic automodel / Ising model) (Cont. Example 103) Assume that observables Z follow the Logistic automodel (5.11) in Example 90. From (5.10), the conditionals (local characteristics) are computed to be such as

$$\log\left(\text{pr}_i(z_i|z_{-i})\right) = z_i\left(\alpha_i + \sum_{j\sim i}\beta_{i,j}z_j\right) - \log\left(1 + \exp\left(\alpha_i + \beta_{i,j}\sum_{j:j\sim i}\beta_{i,j}z_j\right)\right)$$

and hence

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\sum_{i\in\mathcal{S}} \log\left(\mathrm{pr}_i\left(z_i|z_{-i}\right)\right)\right)$$
$$= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\sum_{i\in\mathcal{S}} z_i\left(\alpha_i + \sum_{j\sim i}\beta_{i,j}z_j\right) - \sum_{i\in\mathcal{S}}\log\left(1+\exp\left(\alpha_i + \beta_{i,j}\sum_{j:j\sim i}\beta_{i,j}z_j\right)\right)\right)$$

which does not depend on the normalizing constant (6.4) and hence its computation is less computationally demanding.

7. Hierarchical modeling (Bayesian modeling)

7.1. **A general framework for the hierarchical modeling (A revision).**

Note 108. Uncertainty in spatial statistics can be decomposed according to the Hierarchical spatial model

$$\begin{cases} Z|Y,\vartheta & \text{data model} \\ Y|\vartheta & \text{spatial process model} \end{cases} \tag{7.1}$$

with

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

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.

See for example Figure 7.1

Note 109. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior; it can be considered sometimes as known and sometimes as unknown in what follows. (!)

Note 110. Then the Bayesian spatial hierarchical model becomes

$$\begin{cases} Z|Y,\vartheta_1,\vartheta_2 & \text{data model} \\ Y|\vartheta_1,\vartheta_2 & \text{spatial process model} \\ \vartheta_1|\vartheta_2 & \text{hyper-parameter prior model} \end{cases} \tag{7.2}$$

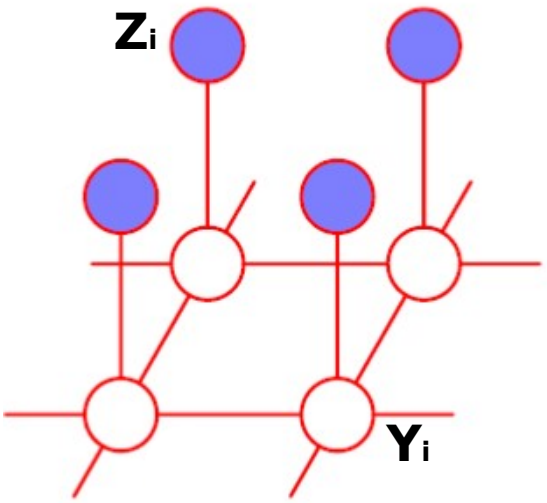


FIGURE 7.1. Hierarchical spatial model structure. $\{Y_i\}$ is the spatial process model which is hidden. $\{Z_i\}$ is the data model. The cartoon depicts a hierarchical spatial model with the special conditional independence structure $Z_i|\{Y_i\}, \vartheta \sim \prod_i \text{pr}(Z_i|Y_i, \vartheta)$ and $Y|\vartheta \sim \text{pr}(Y|\vartheta)$

where uncertainty is described by

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

Note 111. Under Bayesian model (7.2), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

$$\hat{\vartheta}_2 = \arg\min_{\vartheta_2} \left(-2\log\left(\text{pr}(Z|\vartheta_2)\right)\right) \tag{7.3}$$

by maximizing the marginal likelihood

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

or as a MPLE

$$\tilde{\vartheta}_2 = \arg\min_{\vartheta_2} \left(-2\log\left(\prod_i \text{pr}(Z_i|Z_{-i}, \vartheta_2)\right)\right)$$

by maximizing the pseudo marginal Likelihood

$$\text{pseudo}L(Z|\vartheta_2) = \prod_i \text{pr}(Z_i|Z_{-i}, \vartheta_2)$$

as a computational cheap approximation of the MLE $\hat{\vartheta}_2$ in 7.3.

Note 112. Under Bayesian model (7.2), when ϑ_1 is considered as unknown (but random), namely, the a prior $\vartheta_1 \sim \text{pr}(\vartheta_1|\vartheta_2)$ has been specified, uncertainty about unknown ϑ_1 given

Y and ϑ_2 can be represented by the posterior distribution

$$\text{pr}\left(\vartheta_1|Z,\vartheta_2=\hat{\vartheta}_2\right)=\frac{\text{pr}\left(Z|\vartheta_1,\vartheta_2\right)\text{pr}\left(\vartheta_1|\vartheta_2=\hat{\vartheta}_2\right)}{\text{pr}\left(Z|\vartheta_2=\hat{\vartheta}_2\right)}$$

where the value $\hat{\vartheta}_2$ is plugged in. Alternatively, we can plug-in $\tilde{\vartheta}_2$.

Note 113. 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 114. The above statistical problem is naturally addressed in the (either full or empirical) Bayesian statistical framework.

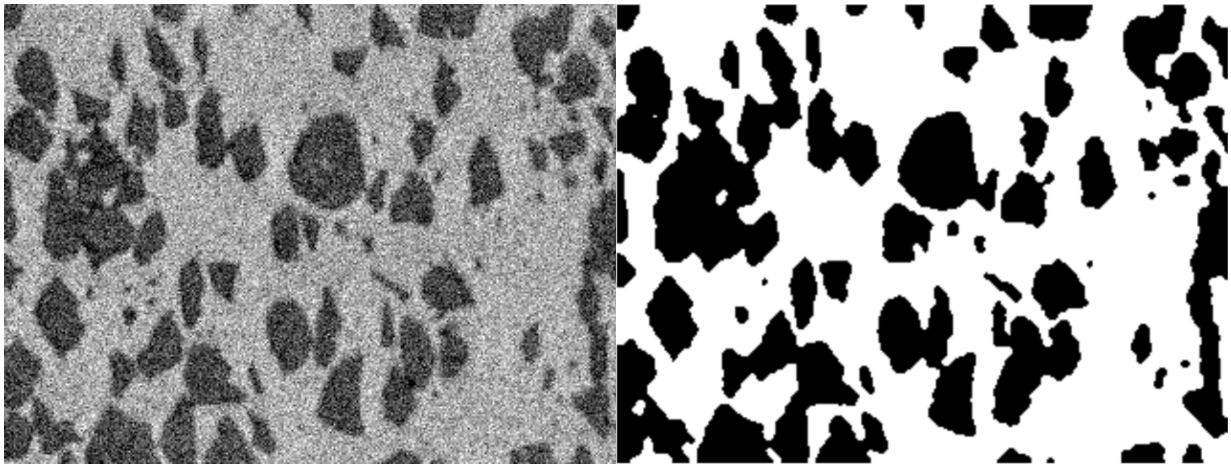
Note 115. Below we give two examples in aerial data.

7.2. **Examples.**

7.2.1. *A simplified spatial model for binary data (e.g. Image denoising).*

Example 116. (Image denoising) A central aim in image processing is to reconstruct an object (e.g. image) $Y=(Y_i;i\in\mathcal{S})$ based on a measurement (observation) $Z=(Z;i\in\mathcal{S})$ which is contaminated by errors $\varepsilon=(\varepsilon_i;i\in\mathcal{S})$. The framework of hierarchical modeling for aerial spatial data is suitable to address such cases.

Consider the image restoration dataset in Example 23 in Handout 1: Types of spatial data. (Figure 7.2a) We have a black and white noisy image with size 240×320 pixels.



(A) Noised image (B) Reconstructed image

FIGURE 7.2. Ferrite-Pearlite steel image (Image restoration)

Mathematically, denote $(Z_i)_{i\in\mathcal{S}}$ as the error contaminated (observed) image. The observables are coded as $Z_i = 1$ for black and $Z_i = 0$ for white at site $i \in \mathcal{S} = \{1, ..., 240 \times 320\}$. Let $n = \text{Card}(\mathcal{S})$. Hence $(Z_i)_{i\in\mathcal{S}}$ is a realization from the data model. The aim is to recover/learn the unknown real (error free) image $(Y_i)_{i\in\mathcal{S}}$ given the measurement/observation $(Z_i)_{i\in\mathcal{S}}$.

The data model can be specified (for instance) by "assuming" that the observation Z_i has been contaminated by iid noise with some "probability" p for all pixels $i \in \mathcal{S}$; i.e. $p = \text{pr}(\{Z_i \neq Y_i\}|p) = 1 - \text{pr}(\{Z_i = Y_i\})$ for all $i \in \mathcal{S}$. Hence

$$\text{pr}(Z_i|Y_i, p) = p^{1-1(\{Z_i=Y_i\})}(1-p)^{1(\{Z_i=Y_i\})},\ i \in \mathcal{S}$$

Consequently, the data model is

$$\begin{aligned}\text{pr}(Z|Y, p) &= \prod_{i=1}^{n} p^{1-1(\{Z_i=Y_i\})}(1-p)^{1(\{Z_i=Y_i\})} = p^{n_{(Z,Y)}}(1-p)^{n-n_{(Z,Y)}} \\ &= \exp\left(n_{(Z,Y)}\log\left(\frac{p}{1-p}\right) + (1-p)^n\right)\end{aligned}$$

where $n_{(Z,Y)} = \sum_{i\in\mathcal{S}} 1(\{Z_i = Y_i\})$.

The spatial process $(Y_i)_{i\in\mathcal{S}}$ is unknown (unobserved), and, according the Bayesian paradigm, we need to specify a prior process on $(Y_i)_{i\in\mathcal{S}}$ account for the uncertainty. To introduce spatial dependence, the researcher may judge to specify (for example) an Logistic automodel (Ising model) process prior such as

$$\text{pr}(Y|\alpha, \beta) \propto \exp\left(\alpha\sum_{i\in\mathcal{S}} Y_i + \beta\sum_{\{i,j\}:i\sim j} Y_iY_j\right),\ \{0,1\}^{\mathcal{S}}$$

with symmetric relation $i \sim j$ considering only the adjacent pixels.

The researcher may be uncertain about the "real" value of p and hence he/she may want to specify a conjugate Beta prior[2] $p \sim \text{Be}(g, h)$ with known g and h to account for the uncertainty. The researcher may set certain fixed values on α and β; hence consider that g, h α, and β are known values.

The Hierarchical Bayesian model is summarized as

$$\begin{cases} Z|Y, p \sim \text{pr}(Z|Y, p) & \text{data model} \\ Y \sim \text{pr}(Y|\alpha, \beta) & \text{spatial process model} \\ p \sim \text{Be}(g, h) & \text{hyper-parameter prior model} \end{cases} \tag{7.4}$$

[2] $\text{Be}(p|g, h) = p^{g-1}(1-p)^{h-1}1_{(0,1)}(p)/\text{B}(g, h)$

To learn $Y|Z$, one can compute the Bayesian MAP estimator of Y, i.e.

$$\begin{aligned}\hat{Y} &= \arg\max_{Y} \left(\log\left(\text{pr}\left(Z|Y\right)\text{pr}\left(Z\right)/\text{pr}\left(Z\right)\right)\right)\\ &= \arg\min_{Y} \left(-\log\left(\text{pr}\left(Z|Y\right)\right)-\log\left(\text{pr}\left(Y\right)\right)\right)\end{aligned}$$

where

$$\begin{aligned}\text{pr}\left(Z|Y\right) &= \int p^{n(Z,Y)}\left(1-p\right)^{n-n(Z,Y)}\text{Be}\left(p|g,h\right)\text{d}p\\ &= \int p^{n(Z,Y)}\left(1-p\right)^{n-n(Z,Y)}\frac{p^{g-1}\left(1-p\right)^{h-1}}{\text{B}\left(g,h\right)}\text{d}p\\ &= \frac{1}{\text{B}\left(g,h\right)}\int p^{n(Z,Y)+g-1}\left(1-p\right)^{n-n(Z,Y)+h-1}\text{d}p\\ &= \frac{1}{\text{B}\left(g,h\right)}\text{B}\left(n\left(Z,Y\right)+g,n-n\left(Z,Y\right)+h\right)\end{aligned}$$

via an optimization numerical algorithm, or perhaps the posterior expectation, i.e.

$$\hat{Y} = \text{E}\left(Y|Z\right) = \int Z\text{pr}\left(Y|Z\right)\text{d}Y$$

via MCMC, INLA, etc... Here the marginal posterior can be computed analytically as

$$\begin{aligned}\text{pr}\left(Y|Z\right) &= \int \text{pr}\left(Y,p|Z\right)\text{d}p = \int \frac{\text{pr}\left(Z|Y,p\right)\text{pr}\left(Y\right)\text{pr}\left(p\right)}{\int \text{pr}\left(Z|Y,p\right)\text{pr}\left(Y\right)\text{pr}\left(p\right)\text{d}p\text{d}Z}\text{d}p\\ &\propto \underbrace{\int \text{pr}\left(Z|Y,p\right)\text{pr}\left(p\right)\text{d}p}_{=\text{pr}(Z|Y)}\text{pr}\left(Y\right) = \text{pr}\left(Z|Y\right)\text{pr}\left(Y\right)\\ &\propto \underbrace{\frac{1}{\text{B}\left(g,h\right)}\text{B}\left(n_{(Z,Y)}+g,n-n\left(Z,Y\right)+h\right)}_{=\text{pr}(Z|Y)}\\ &\quad\times \underbrace{\frac{\exp\left(\alpha\sum_i Y_i+\beta\sum_{j\sim i}Y_iY_j\right)}{\sum_{Y\in\{0,1\}^n}\exp\left(\alpha\sum_i Y_i+\beta\sum_{j\sim i}Y_iY_j\right)}}_{=\text{pr}(Y)}\end{aligned}$$

$$\propto \text{B}\left(n\left(Z,Y\right)+g,n-n\left(Z,Y\right)+h\right)\exp\left(\alpha\sum_i Y_i+\beta\sum_{j\sim i}Y_iY_j\right) \tag{7.5}$$

Note that the only reason that we ignored the constant from the Ising process prior in (7.5) was because, in this particular example, the researcher considered α and β as known constants. Of course, that constant should not have be ignored if α and β had been considered as unknown, and hence we had to learn them.

Figure 7.2b shows the restored image as the Bayesian MAP estimator $\hat{Y} = \arg\max_{Y} (\log (\text{pr} (Y|Z)))$ of $Y|Z$ by using an R optimization function against (7.5).

7.2.2. *A simplified spatial model for count data (e.g. Counts analysis).*

Example 117. Consider the statistical problem scenario where there is available a dataset $\{(X_i, s_i, Z_i) ; i = 1, ..., n\}$, where $Z_i \in \mathbb{N}$ is the count of the occurrence of an event in a particular time interval, at a location s_i, and associated with a vector of covariates (other measurements) $X_i = (X_{i,1}, ..., X_{i,k})^\top$, for $i \in \mathcal{S}$, with $\mathcal{S} = \{1, ..., n\}$, and $n \in \mathbb{N} - \{0\}$ fixed. So, denote $(Z_i)_{i \in \mathcal{S}}$ as the observed vector. Assume that $Z_i \in \mathcal{Z}^{\mathcal{S}}$, with $\mathcal{Z} = \mathbb{N}$ and $\mathcal{S} = \{1, ..., n\}$.

- Such a scenario is suitable for the Columbus OH data set which concerns spatially correlated count data arising from small area sampling of some underlying process. This is the R dataset `columbus{spdep}`. Briefly, the Columbus data frame has 49 rows and 22 columns. Unit of analysis is 49 neighborhoods in Columbus, OH, 1980 data. The date frame has among others variables

  NEIG: neighborhood id value (1-49); conforms to id value used in Spatial Econometrics book.

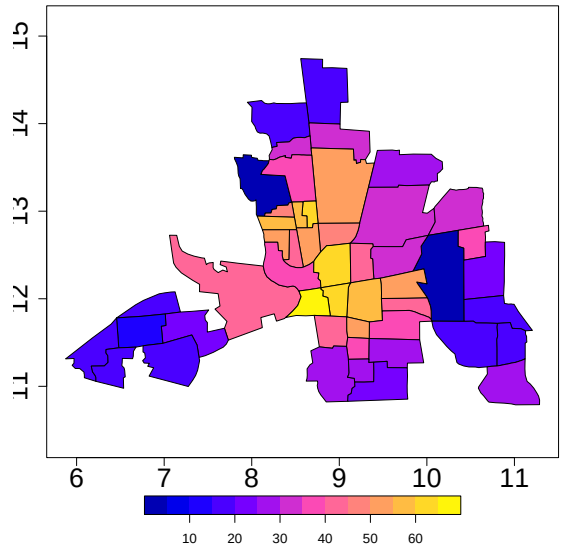
  CRIME: residential burglaries and vehicle thefts per thousand households in the neighborhood

  HOVAL: housing value (in 1,000 USD)

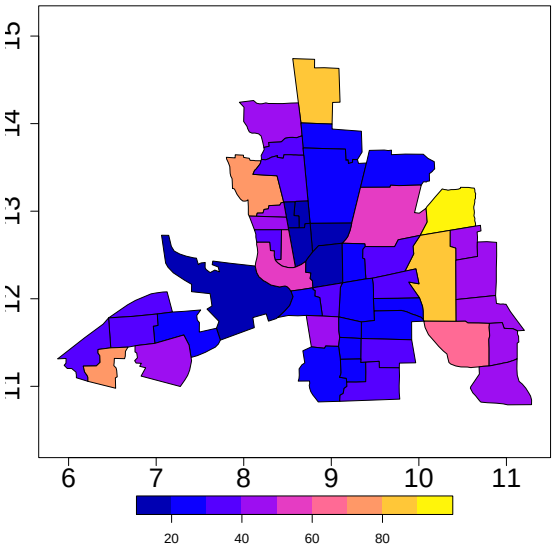
  INC: household income (in 1,000 USD)

  $$\left\{ \left(\underbrace{\text{HOVAL}_i, \text{INC}_i}_{X_i}, \underbrace{\text{NEIG}}_{s_i}, \underbrace{\text{CRIME}_i}_{Z_i} \right)^\top ; i = 1, ..., \underbrace{49}_{n} \right\}$$

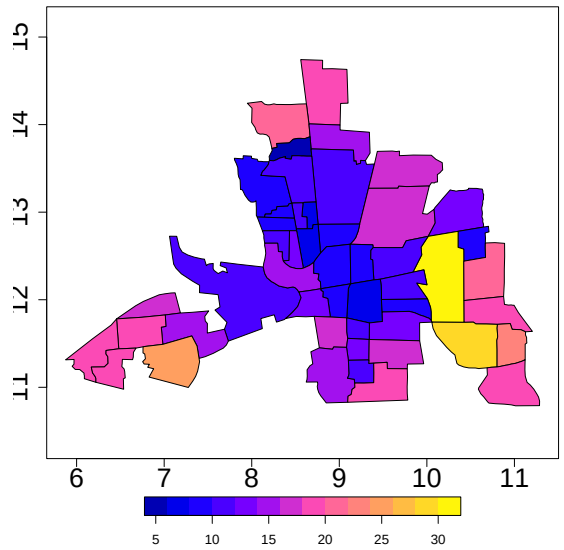
- Figure 2a shows the Property crime (number per thousand households) in 49 districts/neighborhood in Columbus in 1980, as well as the average value of the house in USD. Figure 2b presents the corresponding average house value. For privacy reasons, these are typically aggregated over areas that are large enough to ensure that the counts cannot be traced back to individuals.
- Interest may lie to find whether high rates of crime are clustered in a particular areas, and, perhaps what is the association of it with the value of the houses in the area.



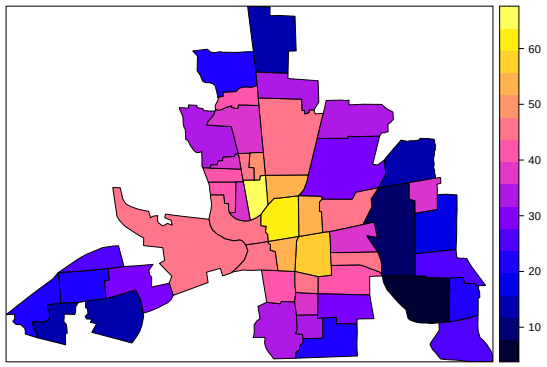
(A) CRIME: residential burglaries and vehicle thefts per thousand households in the neighborhood



(B) HOVAL: housing value (in 1,000 USD)



(C) INC: household income (in 1,000 USD)



(D) Fitted CRIME ~ HOVAL+INC in a Poisson model with rate modeled as CAR with mean CRIME ~ HOVAL+INC

FIGURE 7.3. Columbus Columbus OH spatial analysis dataset

For the data model, it is natural to assume that for site s_i the observable count Z_i is sampled from a Poisson distribution with some given rate $\lambda_i := \mathrm{E}(Z_i|Y_i) = \log(Y_i)$, different for different sites s_i and depending on an unknown/unobserved and underpinning spatial

process $(Y_i)_{i\in\mathcal{S}}$. The researcher may specify the data model as

$$Z_i|Y_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\lambda_i\left(Y_i\right)\right), \text{ for } i=1,...,n \tag{7.6}$$
$$\text{where} \quad \log\left(\lambda_i\right) = Y_i$$

This imposes the (rather strong) assumption that Z_i and Z_j are conditionally independent given the spatial process $(Y_i)_{i\in\mathcal{S}}$.

- For the Columbus dataset, data model (7.6) is reasonable because the observation Z_i represents count namely number of event in a specific time and space and with fixed rate λ_i for each individual neighborhood $i \in \mathcal{S}$.

The spatial process Y is unknown. To specify the uncertainty on Y, the researcher may judge to assign a CAR model prior on $(Y_i)_{i\in\mathcal{S}}$, for instance

$$Y_i|Y_{-i} \sim \text{N}\left(\mu_i + \beta_{i,j}\left(Y_j - \mu_j\right), \kappa_i\right), \text{ for } i=1,...,n \tag{7.7}$$
$$\mu_i = X_i^\top \alpha.$$

To reduce parametric dimensionality, we impose a more restrictive structure such that $\kappa_i = 1/\tau$ for all $i = 1, ..., n$, and

$$\beta_{i,j} = \begin{cases} \phi & \text{if } i \sim j \\ 0 & \text{if } i \not\sim j \text{ or } i = j \end{cases}$$

- In the Columbus example we Spatial process (7.7) is suitable with regressors $X_i = (1, \text{HOVAL}_i, \text{INC}_i)^\top$; that is CRIME ~ HOVAL + INC.

$$\mu_i = \alpha_0 + \alpha_1 \text{HOVAL}_i + \alpha_2 \text{INC}_i, \quad i = 1, ..., n$$

  This is because we are interested in investigating "whether high rates of crime (CRIME) are clustered in a particular areas ($i \in \mathcal{S}$), eg areas with expensive houses (HOVAL), and if yes, perhaps what is the association of it with the value of the houses in the area (INC)". Hence we can use our model in order to see the association of CRIME with SPACE (i.e. $i \in \mathcal{S}$), HOVAL (i.e. house value), and INC (i.e. income).

- Note that unlike the usual linear model, here I have managed to introduce spatial dependence in the model as well by

$$\begin{aligned}\mathrm{E}\left(Y_i|Y_{-i}\right) =& \underbrace{\alpha_0+\alpha_1\mathrm{HOVAL}_i+\alpha_2\mathrm{INC}_i}_{\text{cocariate dependence only}}+\underbrace{\sum_{i\sim j}\phi\left(Y_j-\mu_j\right)}_{\text{spatial dependence}}\\ =& \underbrace{\alpha_0+\alpha_1\mathrm{HOVAL}_i+\alpha_2\mathrm{INC}_i}_{\text{cocariate dependence only}}+\underbrace{\sum_{i\sim j}\phi Y_j}_{\text{spatial dependence only}}\\ &\underbrace{-\sum_{i\sim j}\phi\left[\alpha_0+\alpha_1\mathrm{HOVAL}_j+\alpha_2\mathrm{INC}_j\right]}_{\text{interaction of space and covariates}}\end{aligned}$$

This essentially produces a joint model (see Section 3.1 or just use Theorem 83)

$$Y\sim\mathrm{N}\left(X\alpha,\left(I-\phi N\right)^{-1}\tau^{-1}\right)$$

where N is an $n\times n$ matrix with $[N]_{i,j}=1\left(\{i\sim j\},i\neq j\right)$, where $\sim$ is defined to denote adjacent sites (or otherwise spatial locations sharing same boarders).

For the unknown hyper-parameters α, ϕ and κ, the researcher may consider hyper-priors $\alpha\sim\mathrm{N}\left(0,\Sigma_\alpha\right)$, $\phi\sim\mathrm{U}\left(0,\phi_{\max}\right)$, and $\tau\sim\chi^2\left(\nu\right)$; the prior distributions here are chosen for demonstration. The rest hyper-parameters $\Sigma_\alpha>0$, $\phi_{\max}\in\{\phi>0:I-\phi N\text{ is non singular}\}$, $\nu>0$ are considered as unknown fixed constants set by the researcher based on his/her subjective believes.

The Bayesian spatial hierarchical model becomes

$$\begin{cases} Z_i|Y_i\overset{\text{ind}}{\sim}\mathrm{Poisson}\left(\exp\left(Y_i\right)\right),\forall i & \text{data model}\\ Y|\alpha,\phi,\tau\sim\mathrm{N}\left(X\alpha,\left(I-\phi N\right)^{-1}\tau^{-1}\right) & \text{spatial process model}\\ \alpha|\tau\sim\mathrm{N}\left(\mu_\alpha,\tau^{-1}\Sigma_\alpha\right) & \text{hyper-prior model}\\ \tau\sim\chi^2\left(\nu\right) & \text{hyper-prior model}\\ \phi\sim\mathrm{U}\left(0,\phi_{\max}\right) & \text{hyper-prior model}\end{cases}\tag{7.8}$$

The joint probability model becomes

$$\begin{aligned}\mathrm{pr}\left(Z,Y,\alpha,\beta,\tau\right)&=\prod_{i\in\mathcal{S}}\mathrm{pr}\left(Z_i|Y_i\right)\mathrm{pr}\left(Y|\alpha,\phi,\tau\right)\mathrm{pr}\left(\alpha|\tau\right)\mathrm{pr}\left(\tau\right)\mathrm{pr}\left(\phi\right)\\ &=\prod_{i\in\mathcal{S}}\mathrm{Poisson}\left(Z_i|\exp\left(Y_i\right)\right)\mathrm{N}\left(Y|X\alpha,\left(I-\phi N\right)^{-1}\tau^{-1}\right)\\ &\quad\times\mathrm{N}\left(\alpha|\mu_\alpha,\tau^{-1}\Sigma_\alpha\right)\mathrm{ChiSq}\left(\tau|\nu\right)\mathrm{U}\left(\phi|0,\phi_{\max}\right)\end{aligned}$$

Interest lies in learning $Y|Z$ which can be addressed for instance by the Bayesian MAP estimator

$$\hat{\lambda} = \arg\max_{Y} \left(\text{pr} \left(\lambda | Z \right) \right)$$

or the posterior expectation estimator

$$\hat{\lambda} = \text{E}_{\text{pr}} \left(\lambda | Z \right) = \text{E}_{\text{pr}} \left(\exp \left(Y \right) | Z \right)$$

$\text{pr} \left(\lambda | Z \right)$ can be computed via random variable transformation from $\text{pr} \left(Y | Z \right)$ which is given by the Bayesian theorem as

$$\text{pr} \left(Y | Z \right) = \int \text{pr} \left(Y, \alpha, \tau, \phi | Z \right) \text{d}\alpha \text{d}\tau \text{d}\phi$$

$$\text{pr} \left(Y, \alpha, \tau, \phi | Z \right) \propto \text{pr} \left(Z, Y, \alpha, \tau, \phi \right) = \text{pr} \left(Z | Y \right) \text{pr} \left(Y | \alpha, \tau, \phi \right) \text{pr} \left(\alpha | \tau \right) \text{pr} \left(\phi \right)$$

The above integration is analytically intractable, and hence its numerical computation can be performed by methods such as MCMC, INLA, etc...

Some marginal pdf/pmf

$$\begin{aligned} \text{pr} \left(Y, \alpha, \tau, \phi | Z \right) &\propto \text{pr} \left(Z, Y, \alpha, \tau, \phi \right) \\ &= \prod_{i \in \mathcal{S}} \text{Poisson} \left(Z_i | \exp \left(Y_i \right) \right) \text{N} \left(Y | X\alpha, \left(I - \phi N \right)^{-1} \tau^{-1} \right) \text{N} \left(\alpha | \mu_\alpha, \Sigma_\alpha \tau^{-1} \right) \\ &\quad \times \text{U} \left(\phi | 0, \phi_{\max} \right) \text{ChiSq} \left(\tau | \nu \right) \end{aligned}$$

and

$$\begin{aligned} \text{pr} \left(Y, \tau, \phi | Z \right) &\propto \int \text{pr} \left(Z, Y, \alpha, \tau, \phi \right) \text{d}\alpha \\ &= \int \prod_{i \in \mathcal{S}} \text{pr} \left(Z_i | Y_i \right) \text{pr} \left(Y | \alpha, \tau, \phi \right) \text{pr} \left(\alpha | \tau \right) \text{pr} \left(\tau \right) \text{pr} \left(\phi \right) \text{d}\alpha \\ &= \prod_{i \in \mathcal{S}} \text{pr} \left(Z_i | Y_i \right) \underbrace{\int \text{pr} \left(Y | \alpha, \tau, \phi \right) \text{pr} \left(\alpha | \tau \right) \text{d}\alpha}_{= \text{pr}(Y | \phi, \tau)} \text{pr} \left(\tau \right) \text{pr} \left(\phi \right) \\ &= \prod_{i \in \mathcal{S}} \text{Poisson} \left(Z_i | \exp \left(Y_i \right) \right) \text{N} \left(Y | X\mu_\alpha, \Sigma \left(\phi \right) \tau^{-1} \right) \text{ChiSq} \left(\tau | \nu \right) \text{U} \left(\phi | 0, \phi_{\max} \right) \end{aligned}$$

where

$$\begin{aligned} \text{pr} \left(Y | \phi \right) &= \int \text{pr} \left(Y | \alpha, \tau, \phi \right) \text{pr} \left(\alpha | \tau \right) \text{d}\alpha \\ &= \text{N} \left(Y | X\mu_\alpha, \left(\left(I - \phi \right)^{-1} + X \Sigma_\alpha X^\top \right) \tau^{-1} \right) \\ &= \text{N} \left(Y | X\mu_\alpha, \left(\left(I - \phi N \right) + \left(I - \phi N \right) X \left(\Sigma_\mu^{-1} + X^\top \left(I - \phi N \right) X \right)^{-1} X^\top \left(I - \phi N \right) \right) \tau^{-1} \right) \\ &= \text{N} \left(Y | X\mu_\alpha, \Sigma \left(\phi \right) \tau^{-1} \right) \end{aligned}$$

where

$$\Sigma(\phi) = \left((I - \phi N) + (I - \phi N) X \left(\Sigma_{\mu}^{-1} + X^{\top} (I - \phi N) X \right)^{-1} X^{\top} (I - \phi N) \right)$$

and

$$\begin{aligned}
\operatorname{pr}(Y, \phi | Z) &\propto \int \operatorname{pr}(Z, Y, \tau, \phi) \, \mathrm{d}\tau \\
&= \int \prod_{i \in \mathcal{S}} \operatorname{pr}(Z_i | Y_i) \operatorname{pr}(Y | \tau, \phi) \operatorname{pr}(\tau) \operatorname{pr}(\phi) \, \mathrm{d}\alpha \\
&= \prod_{i \in \mathcal{S}} \operatorname{pr}(Z_i | Y_i) \underbrace{\int \operatorname{pr}(Y | \tau, \phi) \operatorname{pr}(\tau) \, \mathrm{d}\tau}_{= \operatorname{pr}(Y | \phi)} \operatorname{pr}(\phi) \\
&= \prod_{i \in \mathcal{S}} \operatorname{Poisson}(Z_i | \exp(Y_i)) \operatorname{T}\left(Y | X \mu_{\alpha}, \frac{1}{\nu} \Sigma(\phi), \nu \right) \operatorname{U}(\phi | 0, \phi_{\max})
\end{aligned}$$

because

Hint:: The following definition is given:
A d dimensional random vector y follows a Student t distribution with degrees of freedom ν, mean parameter μ, and scale parameter Σ iff it can be represented as $y = \mu + \sqrt{\nu/\xi} x$ results as $\xi \sim \chi^2_{\nu}$, and $x \sim \mathrm{N}(0, \Sigma)$. It is denoted as $y \sim \mathrm{T}(\mu, \Sigma, \nu)$
If $\Sigma > 0$, then x has pdf

$$\operatorname{T}(y | \mu, \Sigma, \nu) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2) \nu^{\frac{d}{2}} \pi^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu} (y - \mu)^{\top} \Sigma^{-1} (y - \mu) \right)^{-\frac{\nu + d}{2}}$$

$$\begin{aligned}
\operatorname{pr}(Y | \phi) &= \int \operatorname{pr}(Y | \phi, \tau) \operatorname{pr}(\tau) \, \mathrm{d}\tau \\
&= \int \mathrm{N}\left(Y | X \mu_{\alpha}, \Sigma(\phi) \tau^{-1} \right) \operatorname{ChiSq}(\tau | \nu) \, \mathrm{d}\tau \\
&= \int \mathrm{N}\left(Y | X \mu_{\alpha}, \Sigma(\phi) \tau^{-1} \frac{1}{\nu} \nu \right) \operatorname{ChiSq}(\tau | \nu) \, \mathrm{d}\tau \\
&= \int \mathrm{N}\left(Y | X \mu_{\alpha}, \frac{\nu}{\tau} \left(\frac{1}{\nu} \Sigma(\phi) \right) \right) \operatorname{ChiSq}(\tau | \nu) \, \mathrm{d}\tau \\
&= \int \mathrm{N}\left(Y | X \mu_{\alpha}, \frac{\nu}{\tau} \left(\frac{1}{\nu} \Sigma(\phi) \right) \right) \operatorname{ChiSq}(\tau | \nu) \, \mathrm{d}\tau \\
&= \operatorname{T}\left(y | X \mu_{\alpha}, \frac{1}{\nu} \Sigma(\phi), \nu \right)
\end{aligned}$$

Notice that

$$\operatorname{pr}(Y | Z) = \int \operatorname{pr}(Y, \phi | Z) \, \mathrm{d}\phi$$

involves an analytically intractable one-dimensional integral which can be easily approximated by using a standard integration algorithm (e.g. parallelogram rule).

- In Columbus example, the resulted Bayesian hierarchical model is as in (7.8). Estimation was facilitated via INLA. The estimates are computed as the posterior expected values given the data Z as

$$\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_{\text{const}} \\ \hat{\alpha}_{\text{HOVAL}} \\ \hat{\alpha}_{\text{INC}} \end{pmatrix} = \begin{pmatrix} 54.3139189 \\ -0.2821969 \\ -0.9882862 \end{pmatrix}$$

  $\hat{\phi} = 0.1589004$, and $\hat{\kappa} = 87.65$. The fitted counts $\hat{Y}$ are presented in Figure 7.3d.

By comparing Figure 7.3a and Figure 7.3d, we see that there are certain locations where the fitted counts $\hat{Y}$ and Z are substantially different. Perhaps, we could improve our parameterization in (7.7) by considering a less restrictive $\beta_{i,j}$ or by including more covariates in the mean μ_i.

Remark 118. Example 117 demonstrates that if INLA is used to facilitate inference in this particular model there is no reason to approximate $\text{pr}\left(Y|Z,\phi\right)$ as $\widetilde{\text{pr}}_{\text{G}}\left(Y|Z,\phi\right)$ (e.g. by Laplace approx.) because it is

$$\text{pr}\left(Y|Z,\phi\right) \propto \text{pr}\left(Y,\phi|Z\right) \propto \prod_{i\in\mathcal{S}} \text{Poisson}\left(Z_i|\exp\left(Y_i\right)\right) \text{T}\left(Y|X\mu_\alpha, \frac{1}{\nu}\Sigma\left(\phi\right), \nu\right)$$

Hence, as discussed in Note 18 of the "Handout 2: Introduction to INLA & R-INLA", the approximation step 3 in Algorithm 17 is omitted and the associated approximation error does not exist!

Proposition 119. *Following, we show that any SAR model can be written as a CAR model, however the inverse is not always true.*

Proof. Let Λ be $n \times n$ positive diagonal matrix. Let $\tilde{B}$ be $n \times n$ positive matrix where $I - \tilde{B}$ is non-singular and $\tilde{B}_{i,i} := \left[\tilde{B}\right]_{i,i} = 0$. Then $\left(I - \tilde{B}\right)^{-1} \Lambda \left(I - \tilde{B}^\top\right)^{-1}$ is well defined and I need to solve wrt B and $K = \text{diag}(\kappa_1, ..., \kappa_n)$

$$\begin{aligned} (I - B)^{-1} K &= \left(I - \tilde{B}\right)^{-1} \Lambda \left(I - \tilde{B}^\top\right)^{-1} \Leftrightarrow \\ K^{-1} (I - B) &= \left(I - \tilde{B}^\top\right) \Lambda^{-1} \left(I - \tilde{B}\right) \Leftrightarrow \\ K^{-1} - K^{-1} B &= \Lambda^{-1} - \tilde{B}^\top \Lambda^{-1} - \Lambda^{-1} \tilde{B} + \tilde{B}^\top \Lambda^{-1} \tilde{B} \end{aligned}$$

If I focus of the diagonal part and set $B_{i,i} := [B]_{i,i} = 0$

$$\left[K^{-1}\right]_{i,i} - \cancelto{= 0}{\left[K^{-1} B\right]_{i,i}} = \left[\Lambda^{-1}\right]_{i,i} - \cancelto{= 0}{\left[\tilde{B}^\top \Lambda^{-1}\right]_{i,i}} - \cancelto{= 0}{\left[\Lambda^{-1} \tilde{B}\right]_{i,i}} + \left[\tilde{B}^\top \Lambda^{-1} \tilde{B}\right]_{i,i}$$

so

$$\kappa_i = \left(\frac{1}{\lambda_i} + \sum_{j=1}^{n} \frac{\tilde{B}_{j,i}^2}{\lambda_j}\right)^{-1} > 0, \ \forall i = 1, ..., n$$

and hence I can solve with respect to K and B in a manner that they satisfy the assumptions of CAR. □

Remark 120. The converse of Proposition 119 is not true.