

# Spatio-temporal statistics (MATH4341)

## Michaelmas term

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Department of Mathematical Sciences (Office MCS3088)

Durham University

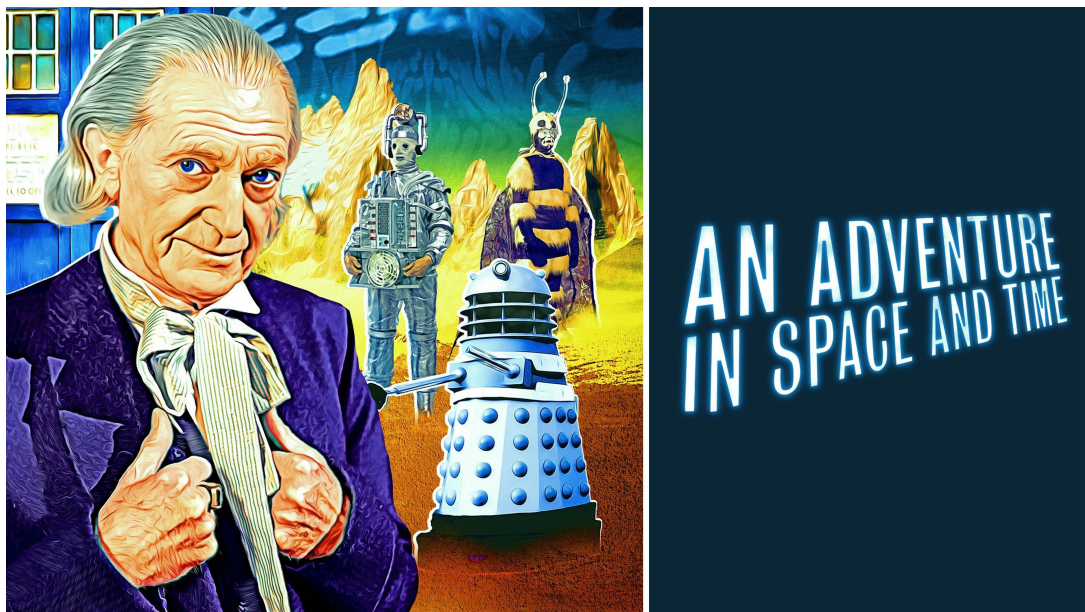
Stockton Road Durham DH1 3LE UK

2023/10/03 at 19:45:03

### Concepts

An introduction to spatial statistics:

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



# 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. Covers all the three the spatial stats concepts we will introduce. Classic book in spatial statistics, but a bit outdated.
- 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. A bit more theoretical than Cressie, N. (2015).
- 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. The authors support it is written for students.
- 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:

- 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 multivariate 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 Classic book in spatial statistics, and perhaps one of the first, if not the first, textbook in the area, so outdated.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.
- 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).
- 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
- 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).
- 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.

## Handout 1: Types of spatial data

Lecturer &amp; 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/](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  $\{Z_{s_1}, \dots, Z_{s_n}\}$  observed at locations spatial locations  $\{s_1, \dots, s_n\}$ . 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)

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

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

## **2. PRINCIPAL SPATIAL STATISTICS AREAS**

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

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

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

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

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

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

*Note 14.* Geostatistics tries 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  $S$ .

#### 2.1.1. Examples.

**Example 15.** (Ground water pollution in the Central Valley of California<sup>1</sup>) 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

---

<sup>1</sup>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](http://www.edf.org/sites/default/files/documents/groundwater-contaminants-report.pdf)

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 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 fixed/known and hence part of the training observations. The locations are irregularly scattered/spaced and hence not on a regular grid of points. 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’.

**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

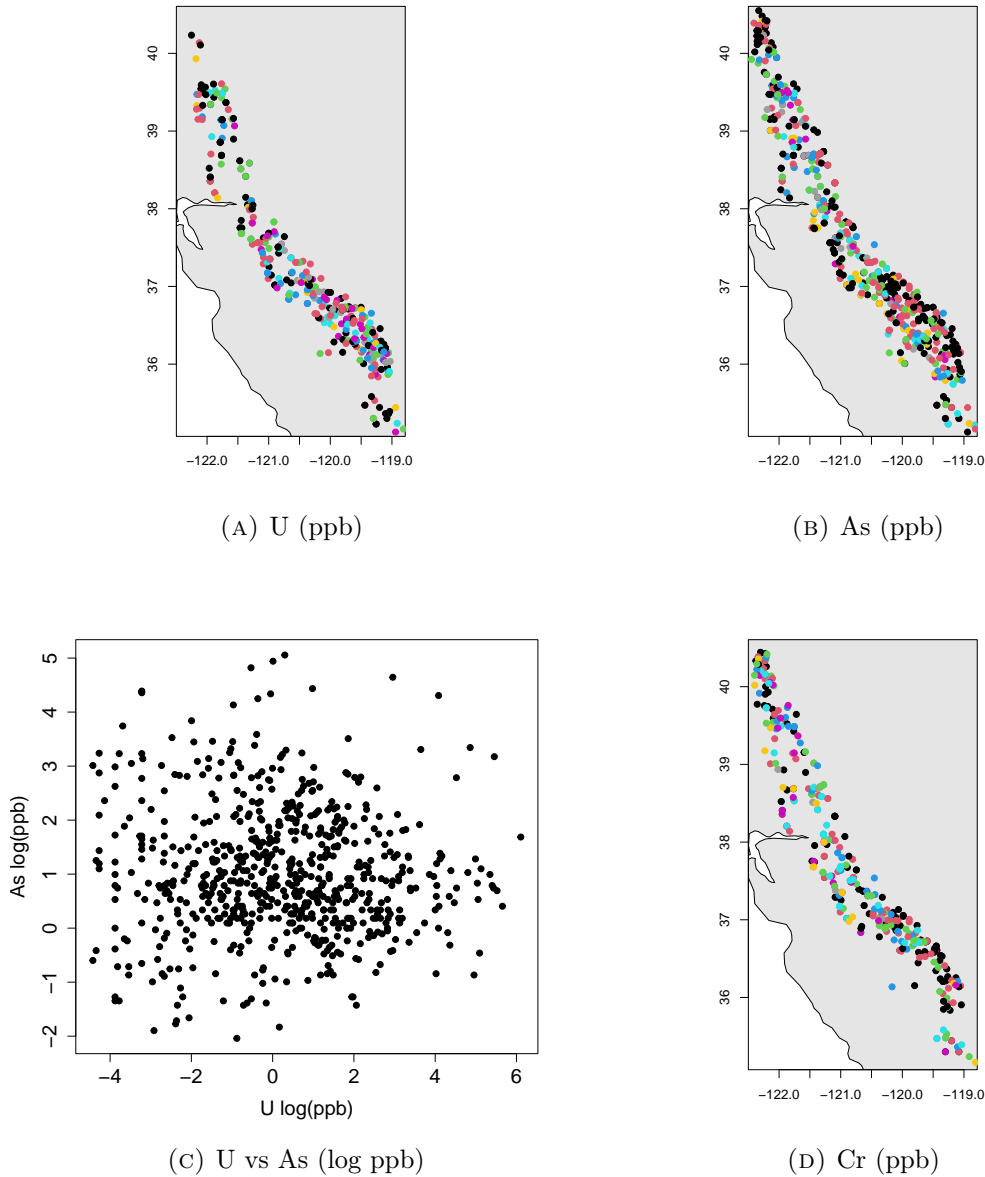


FIGURE 2.1. Ground water pollution in the Central Valley of California

County, Pennsylvania. The percentage of coal ash at the sampled locations is denoted by the colorbar. The sampled locations are fixed, and regularly spaced in a grid. 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. 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.

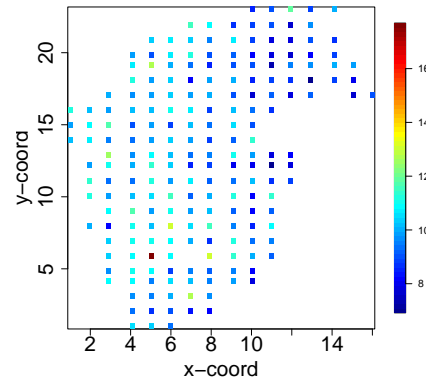


FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

**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.

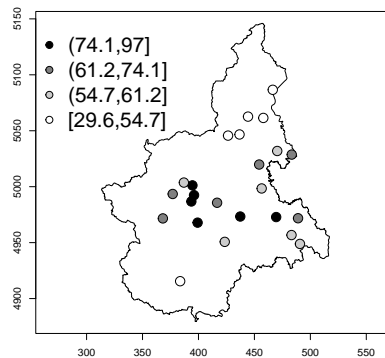


FIGURE 2.3. (Air pollution data) Average PM10 ( $\mu g/m^3$ ) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte

## 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 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 epidemiological 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  $240 \times 320$  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 continuous). The each observation  $Z(s)$  is the color of a pixel  $s$ .

**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. 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

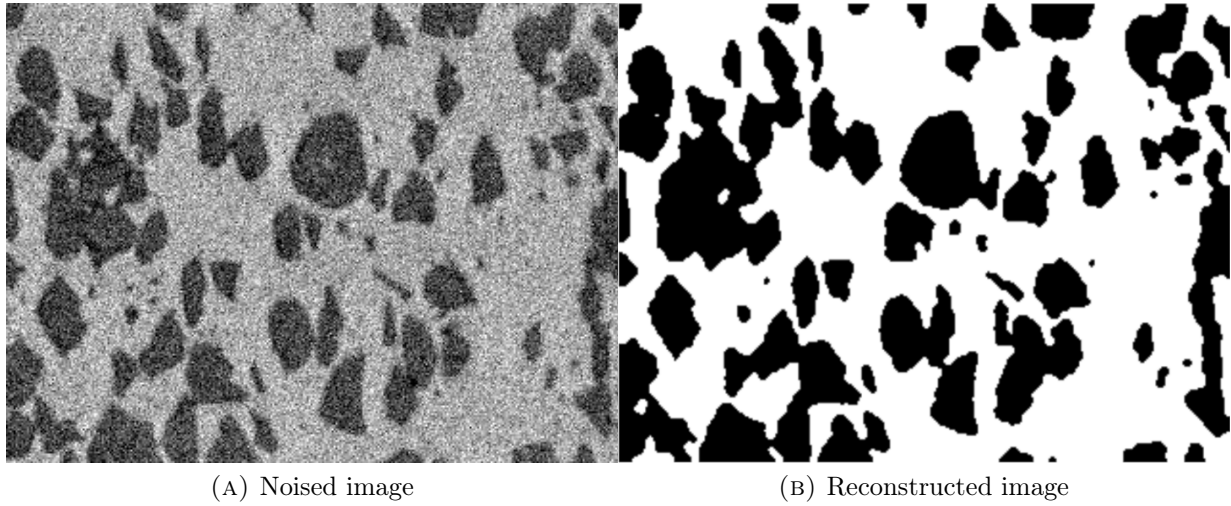


FIGURE 2.4. Ferrite-Pearlite steel image (Image restoration)

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.

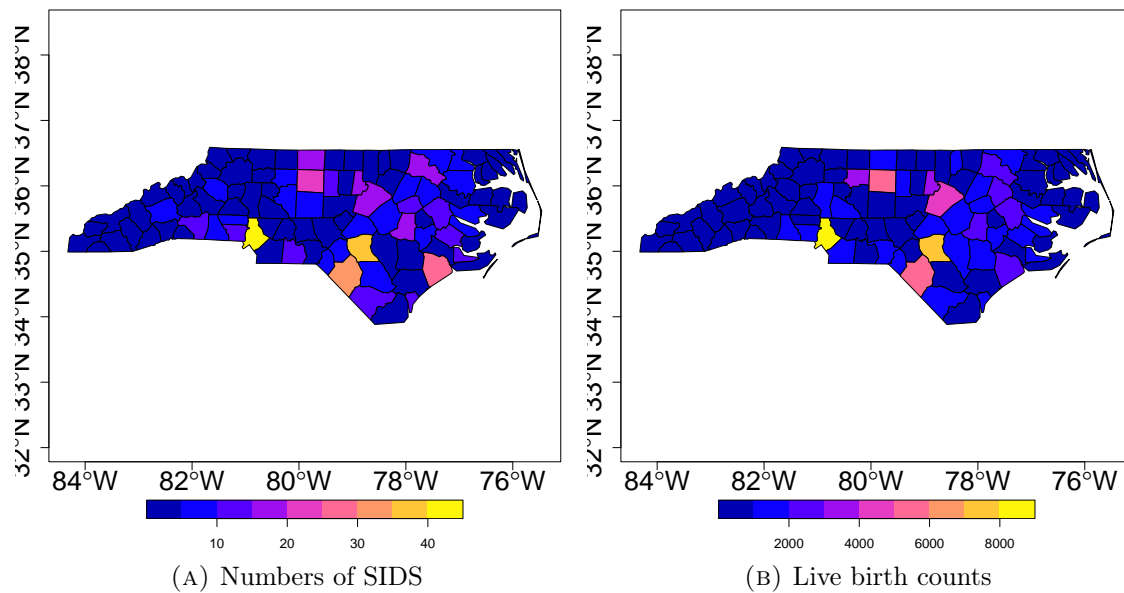


FIGURE 2.5. North Carolina SIDS data (Aerial unit data)

### 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  $Z(s) = 1$  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 patterns 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  $1000 \times 5000$  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 arranged 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 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

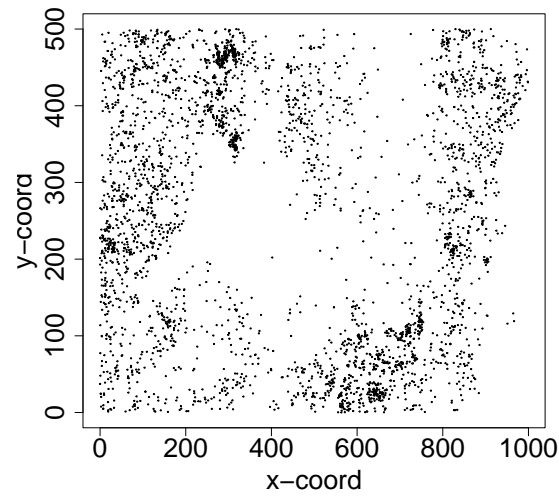


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

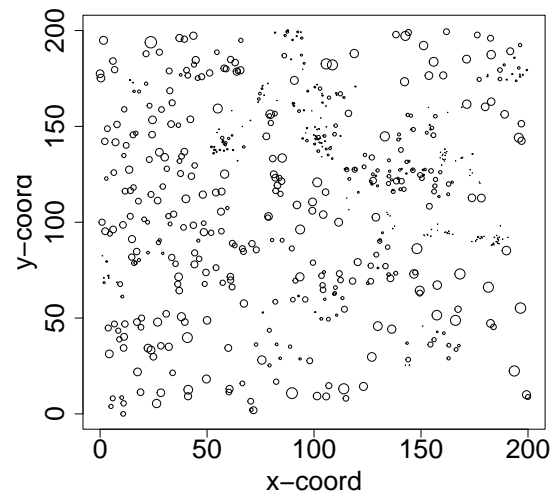


FIGURE 2.7. Longleaf Pines Point Pattern (Spatial point data)

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

*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

$$(3.1) \quad \{Y(s); s \in \mathcal{S}\},$$

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

*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

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

*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 where 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

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

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

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

To be  
defined  
rigorously  
later

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

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

### 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

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

with

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

### 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.

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

with

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

...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  $\theta \sim \text{pr}(\cdot)$  adding a third layer leading to the Bayesian hierarchical model

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

with

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

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

That's a snapshot for what follows.

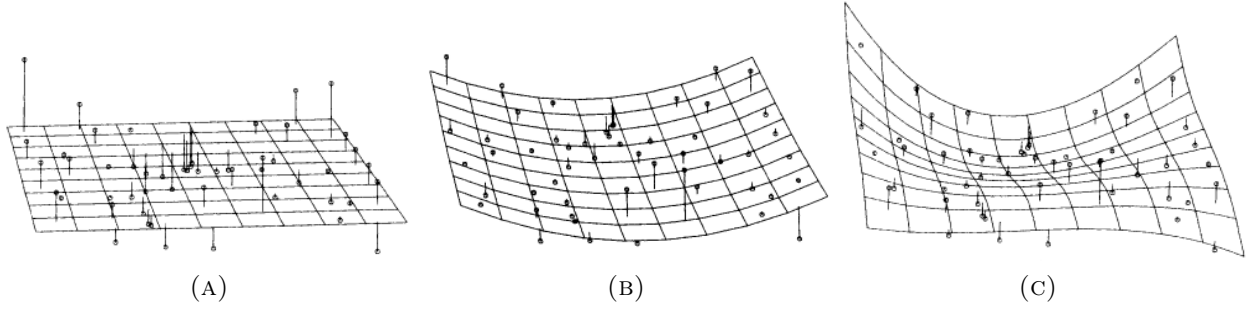


FIGURE 3.1. Examples representing the hierarchical spatial model (3.8) for different values of  $\theta = (\sigma^2, \beta, \phi)$

parametrized as  $[\mu]_i = s_{(0),i}\beta_0 + s_{(1),i}\beta_1 + s_{(2),i}\beta_2 + s_{(1),i}s_{(2),i}\beta_{12} + \dots$  at a location  $s$  (to consider spatial dependence) with some unknown parameter  $\beta$ , and covariance matrix parametrized as  $[C]_{i,j} = c(s_i, s_j)$  with  $c(s, s') = \sigma^2 \exp(-\phi \|s - s'\|)$ ; here  $\beta$ ,  $\phi$ , and  $\sigma^2$  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  $Z_i = Y_i + \epsilon_i$ ,  $\epsilon_i \sim N(0, \sigma^2)$ . To sum up, we have build the hierarchical model

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

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

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



## Handout 2: Computational methods Introduction to INLA

Lecturer &amp; 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

### 1. LAPLACE APPROXIMATION (LA)

**Proposition 1.** Consider integral

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

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

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

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

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

$$(1.1) \quad L(\theta) \approx L(\hat{\theta}) + (\theta - \hat{\theta})^\top \nabla L(\hat{\theta}) - \frac{1}{2} (\theta - \hat{\theta})^\top H(\hat{\theta}) (\theta - \hat{\theta})$$

then

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

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

**Example 2.** Consider posterior expectation

$$(1.2) \quad E(g(\theta) | z) = \int g(\theta) \text{pr}(\theta | z) d\theta$$

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

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

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

**Solution.** (Sketch of the solution) It is

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

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  $\text{pr}(\theta_1 | z) = \widehat{\text{pr}(\theta_1 | z)} (1 + O_{\theta_1}(n^{-1}))$  where the lower index indicates the dependence of the constant on  $\theta_1$ .

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

$$(1.4) \quad \text{pr}(\theta_1 | z) = \int \text{pr}(\theta_1, \theta_2 | z) d\theta_2$$

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

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

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

$\Sigma$  is the minus Hessian of  $n^{-1}(\log(\text{pr}(\theta_1, \theta_2)) + \log(\text{pr}(z|\theta_1, \theta_2)))$ ,

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

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

**Solution.** (Sketch of the solution) It is

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

where  $L_{\theta_1}^*(\theta_2) = n^{-1}(\log(\text{pr}(\theta_1, \theta_2)) + \log(\text{pr}(z|\theta_1, \theta_2)))$  and  $L(\theta) = n^{-1}(\log(\text{pr}(\theta)) + \log(\text{pr}(z|\theta)))$ . 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  $\text{pr}(\theta_1|z) = \widehat{\text{pr}(\theta_1|z)}(1 + O_{\theta_1}(n^{-1}))$  where the lower index indicates the dependence of the constant on  $\theta_1$ .

## 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  $z = (z_1, \dots, z_n)^\top$  can be assumed to be conditionally independent, given a latent random field  $y = (y_1, \dots, y_n)^\top$  and hyper-parameter  $\theta_1$ , i.e.

$$(2.1) \quad z|y, \theta_1 \sim \text{pr}(z|y, \theta_1) = \prod_{i=1}^n \text{pr}(z_i|y_i, \theta_1).$$

The second level assumes that  $y$  follows a multivariate Gaussian distribution (Essentially a Gaussian random field) given hyper-parameter  $\theta_2$ , i.e.

$$(2.2) \quad y|\theta \sim N(\mu(\theta_2), (Q(\theta_2))^{-1})$$

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  $\mathcal{G} = (\mathcal{V} = \{1, \dots, n\}, \mathcal{E})$  such that*

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

*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

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

*Note 9.* The joint posterior probability model is

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

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  $\theta$  is small, typically 2 to 5, but not exceeding 20.*
- (2)  *$\text{pr}(y|\theta)$  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  $\theta$ , 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  $\eta_i$  via a known link function  $g(\cdot)$ , as  $g(\mu_i) = \eta_i$  and

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

where  $\alpha$  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  $\epsilon_i$  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  $\eta_i$  instead of  $\epsilon$ ), and the rest hyperparameters (to be learned) constitute  $\theta$ .

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

### 2.3. The general idea.

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

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

$$(2.8) \quad \text{pr}(y_i|z) = \int \int \text{pr}(y, \theta|z) dy_{-i} d\theta = \int \text{pr}(y_i|z, \theta) \text{pr}(\theta|z) d\theta$$

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

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

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

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

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

$$(2.10) \quad \tilde{\text{pr}}(\theta|z) \propto \left. \frac{\text{pr}(z|y, \theta) \text{pr}(y|\theta) \text{pr}(\theta)}{\tilde{\text{pr}}_G(y|z, \theta)} \right|_{y=y^*(\theta)}$$

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  $\tilde{\text{pr}}_G(y|z, \theta)$  of  $\text{pr}(y|z, \theta)$  in Note 15. This is fast but not generally accurate.

**Laplace approximation:** Similar to Note 15, compute

$$(2.11) \quad \tilde{\text{pr}}(y_i|z, \theta) \propto \frac{\text{pr}(z|y, \theta) \text{pr}(y|\theta) \text{pr}(\theta)}{\tilde{\text{pr}}_G(y_{-i}|y_i, z, \theta)} \Big|_{y=y^*(\theta)}$$

where  $\tilde{\text{pr}}_G(y_{-i}|y_i, z, \theta)$  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 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 denominator 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  $\theta$ .
  - (a) Locate a collection of points  $\{\theta^{(k)}; k = 1, \dots, K\}$  in the area of high density of  $\tilde{\text{pr}}(\theta|z)$ .
  - (b) Find the mode of  $\tilde{\text{pr}}(\theta|z)$ .
- (2) Compute approximation  $\tilde{\text{pr}}(\theta|z)$  at points  $\{\theta^{(k)}; k = 1, \dots, K\}$  by using (2.10).
- (3) Compute approximation  $\tilde{\text{pr}}(y_i|z, \theta)$  at points  $\{\theta^{(k)}; k = 1, \dots, K\}$  of  $\theta$  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  $\tilde{\text{pr}}(y_i|z)$  of (2.8) via standard numerical approximation as

$$(2.12) \quad \tilde{\text{pr}}(y_i|z) = \sum_{k=1}^K \tilde{\text{pr}}(y_i|z, \theta^{(k)}) \tilde{\text{pr}}(\theta^{(k)}|z) \Delta^{(k)}$$

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

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

$$(2.13) \quad \tilde{\text{pr}}(\theta_j|z) = \sum_{k=1}^K \tilde{\text{pr}}(\theta_{-j}, \theta_{-j}^{(k)}|z) \Delta^{(k)}$$

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

*Note 18.* The error in (2.12) comes from the Laplace approximations in  $\tilde{\text{pr}}(\theta^{(k)}|z)$  and  $\tilde{\text{pr}}(y_i|z, \theta^{(k)})$ , as well as the numerical integration and the choice of locations  $\{\theta^{(k)}\}$ . When the likelihood  $\text{pr}(y|z, \theta^{(k)})$  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

$$\tilde{\text{pr}}(z) = \int \frac{\text{pr}(z|y, \theta) \text{pr}(y|\theta) \text{pr}(\theta)}{\tilde{\text{pr}}_G(y|z, \theta)} \Big|_{y=y^*(\theta)} d\theta$$

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

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

and hence DIC is defined as

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

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  $z^{\text{new}}$  (includes missing data) given the observables  $z$  and model (2.4) is

$$(2.14) \quad \text{pr}(z^{\text{new}}|z) = \int \text{pr}(z^{\text{new}}|y^{\text{new}}) \text{pr}(y^{\text{new}}|z) dy^{\text{new}}$$

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

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 in the 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("fmesh")
```

#### 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

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

where  $\{z_{i,j}\}$  (the observables) are number of colonies found on plate  $j$  for dose  $i$  and  $x_i$  indicate the  $i$ th dose. Let  $u_{ij} | \tau \sim N(0, \sigma^2)$  be the so-called random effects, while  $\{\beta_i\}$  are unknown parameters called fixed effects.

- In terms of model (2.4), the GMRF is  $y = (\{\lambda_{i,j}\}, \{\beta_i\}, \{u_{i,j}\})$ .
- We consider prior on  $\sigma^2$  such that

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

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

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  $\sigma^2$ ).
- 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  $\beta_1$  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(tau) and 1/sqrt(tau)^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))
```

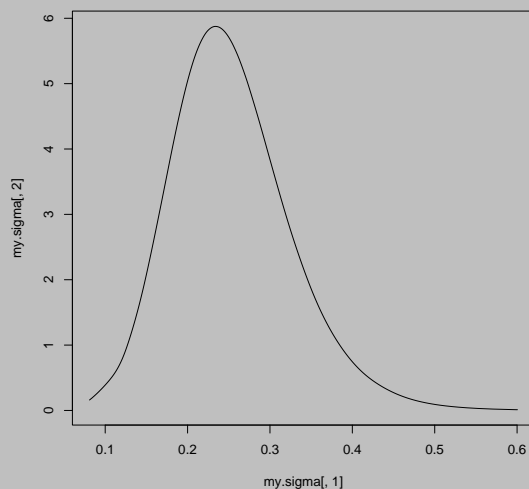
```

      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  $\lambda_i = E(z_i)$ , using the `control.family` object in order for `inla()` to compute the linear predictor  $\eta_i$ . 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 by default are not due to computational cost.

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

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



- Summary about  $\text{pr}(\eta_7|z_{-7})$  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

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

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

$$\begin{aligned}
 (3.2) \quad \text{pr}(z_7|z_{-7}) &\approx E_{\tilde{\text{pr}}(\lambda_7|z_{-7})}(\tilde{\text{pr}}(z_7|\lambda_7)) \\
 &\approx \frac{1}{T} \sum_{t=1}^T \tilde{\text{pr}}(z_7|\lambda_7^{(t)})
 \end{aligned}$$

where  $\left\{ \lambda_7^{(t)} \right\}_{t=1}^T$  is a sample drawn from  $\tilde{\text{pr}}(\lambda_7|z_{-7})$  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)
```





*Note 26.* Assume we wish to address the minimization problem

$$(A.1) \quad \hat{\theta} = \arg \min_{\theta} (C(\theta))$$

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

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

*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  $\{\theta^{(t)}\}$  to  $\hat{\theta}$  as  $\theta^{(t)} \rightarrow \hat{\theta}$  under regularity conditions (omitted here).

**Algorithm 29.** *Newton algorithm consist of the recursion*

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

where  $\nabla_{\theta} C(\theta)|_{\theta=\theta^{(t)}}$  is the gradient of  $C(\theta)$  at  $\theta = \theta^{(t)}$ ,  $\nabla_{\theta}^2 C(\theta)|_{\theta=\theta^{(t)}}$  is the Hessian matrix of  $C(\theta)$  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*

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

where  $\nabla_{\theta} C(\theta)|_{\theta=\theta^{(t)}}$  is the gradient of  $C(\theta)$  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  $\theta^* = \arg \min_{\theta} (-\ell_n(\theta))$  where  $\ell_n(\theta) = \log f(x|\theta)$  and  $\theta = (a, b)$ .

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

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

**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

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

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

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

Newton algorithm's recursion is

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

where additionally

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

□

## APPENDIX B. GAUSSIAN APPROXIMATION OF A (POSTERIOR) DISTRIBUTION

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

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

**Solution.** The log posterior density is

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

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

$$\begin{aligned} 0 = \frac{d}{d\theta} \log (\text{pr} (\theta|x)) \Big|_{\theta=\mu_n} &= \frac{a_n - 1}{\theta} - \frac{b_n - 1}{1 - \theta} \Big|_{\theta=\mu_n} \implies \mu_n = \frac{a_n - 1}{a_n + b_n - 2} \\ \Sigma_n = \frac{d^2}{d\theta^2} \log (\text{pr} (\theta|x)) \Big|_{\theta=\mu_n} &= \frac{a_n - 1}{\theta^2} - \frac{b_n - 1}{(1 - \theta)^2} \Big|_{\theta=\mu_n} \implies \Sigma_n = \frac{(a_n - 1)(b_n - 1)}{(a_n + b_n - 2)^3} \end{aligned}$$

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