Contents

1 Introduction

> *You can't do better design with a computer, but you can speed up your work enormously.*
>
> — **Wim Crouwel**
> (Graphic designer and typographer)

1

Corona Virus

2

> *Users do not care about what is inside the box, as long as the box does what they need done.*
>
> — **Jef Raskin**
> about Human Computer Interfaces

3 Introduction to Bayesian Inference

> *A picture is worth a thousand words. An interface is worth a thousand pictures.*
>
> **— Ben Shneiderman**
> (Professor for Computer Science)

Bayesian Inference is a method of statistical inference that uses Bayes' theorem to update the probability of a hypothesis as more data are observed or more information becomes available. It is an essential technique in mathematical statistics and the polar opposite of the frequentist approach, which makes predictions based solely on data from an experiment. In the Bayesian approach a *prior* distribution $p\left(\boldsymbol{\theta}, \sigma^2\right)$ is introduced as part of the model. This distribution is intended to express a state of knowledge or ignorance about $\boldsymbol{\theta}$ and σ^2 prior to obtaining the data. Using the prior distribution, the likelihood function $p\left(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2\right)$, and the observed data x, it is possible to calculate the probability distribution $p\left(\boldsymbol{\theta}, \sigma^2 \middle| \boldsymbol{x}\right)$ of $\boldsymbol{\theta}$ and σ^2 given the data $\boldsymbol{x}$. This distribution is called the *posterior* distribution of $\boldsymbol{\theta}$ and σ^2 an is used to make inferences about the parameters.[1]

3.1 Preliminaries

This work follows strict notation rules to easily represent different elements such as matrices or graphs and contains frequently used abbreviations. These and some other basic concepts used in this work are introduced below. The notation follows the one used by Rue and Held.[2]

[1] Cf. Box and Tiao 2011.
[2] Cf. Havard Rue and Held 2005.

Matrices and Vectors

Vectors and matrices are indicated by bold notation, such as $\boldsymbol{x}$ and $\boldsymbol{A}$. The transpose of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$. The element in the ith row and jth column of $\boldsymbol{A}$ is referenced by A_{ij}. This notation is also used for vectors and x_i denotes the ith element of a vector. The vector $(x_1, x_{i+1}, ..., x_j)^T$ is abbreviated to $\boldsymbol{x}_{i:j}$. If the columns $\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_m$ of a $n \times m$ matrix $\boldsymbol{A}$ are stacked on top of each other, this is denoted by $\text{vec}(\boldsymbol{A}) = \left(\boldsymbol{A}_1^T, \boldsymbol{A}_2^2, ..., \boldsymbol{A}_m^T\right)$. Deleting rows and/or columns from $\boldsymbol{A}$ creates a *submatrix*. If a submatrix of a $n \times n$ matrix $\boldsymbol{A}$ can be obtained by removing rows and columns of the same index, it is called a *principal submatrix*. If this matrix can be obtained by deleting the last $n - r$ rows and columns, it is called a *leading principal submatrix* of **A**.
A diagonal $n \times n$ matrix $\boldsymbol{A}$ is denoted by $\text{diag}(\boldsymbol{A})$ and has the following structure:

$$\text{diag}(\boldsymbol{A}) = \begin{pmatrix} A_{11} & & \\ & \ddots & \\ & & A_{nn} \end{pmatrix}.$$

The identity matrix is denoted by $\boldsymbol{I}$.
If $A_{ij} = 0$ for $i > j$ or $A_{ij} = 0$ where $i > j$, then $\boldsymbol{A}$ is called *upper triangular* and *lower triangular* respectively. The *bandwidth* of a matrix $\boldsymbol{A}$ is defined as $\max\{|i - j| : A_{ij} \neq 0\}$. The *lower bandwidth* is given by $\max\{|i - j| : A_{ij} \neq 0 \text{ and } i > j\}$. $|\boldsymbol{A}|$ denotes the *determinant* of a $n \times n$ matrix $\boldsymbol{A}$ and is equal to the product of the eigenvalues of $\boldsymbol{A}$. The *rank* of $\boldsymbol{A}$, referenced by $\text{rank}(\boldsymbol{A})$, is the number of linearly independent rows or columns of $\boldsymbol{A}$. The sum of the diagonal elements is called *trace* of $\boldsymbol{A}$, $\text{trace}(\boldsymbol{A}) = \sum_i A_{ii}$.
Finally, '$\odot$' denotes the element-wise multiplication of two matrices of size $n \times m$, '$\oslash$' denotes the element-wise division and raising each element of a matrix $\boldsymbol{A}$ to a scalar power uses the symbol '$\circledcirc$'.

Lattice and Torus

$\mathcal{I}_{\boldsymbol{n}}$ denotes a (regular) **lattice** (or grid) of size $\boldsymbol{n} = (n_1, n_2)$ (in the two-dimensional case). $\boldsymbol{x}$ can take values on $\mathcal{I}_{\boldsymbol{n}}$ and $x_{i,j}$ denotes the value of $\boldsymbol{x}$ at location ij, for $i = 1, ..., n_1$ and $j = 1, ..., n_2$. For easier reading this will be shortened to x_{ij}. On an *infinite lattice* $\mathcal{I}_{\boldsymbol{\infty}}$, ij are numbered as $i = 0, \pm 1, \pm 2, ...$, and $j = 0, \pm 1, \pm 2, ...$
A lattice with cyclic or toroidal boundary conditions is referred to as *torus* and is denoted by $\mathcal{I}_{\boldsymbol{\infty}}$. The dimension is $\boldsymbol{n} = (n_1, n_2)$ (in the two-dimensional case) and all indices are modulus $\boldsymbol{n}$ and run from 0 to $n_1 - 1$ or $n_2 - 1$. If a GMRF $\boldsymbol{x}$ is defined on

$\mathcal{I}_{\boldsymbol{n}}$, the toroidal boundary conditions imply that x_{-2,n_2} is equal to $x_{n_1-2,0}$ since $-2 \mod n_1$ is equal to $n_1 - 2$ and $n_2 \mod n_2$ is equal to 0.
An *irregular lattice* refers to a spatial configuration of regions $i = 1, ..., n$ where the regions (mostly) have common boundaries, for instance the states of a nation.

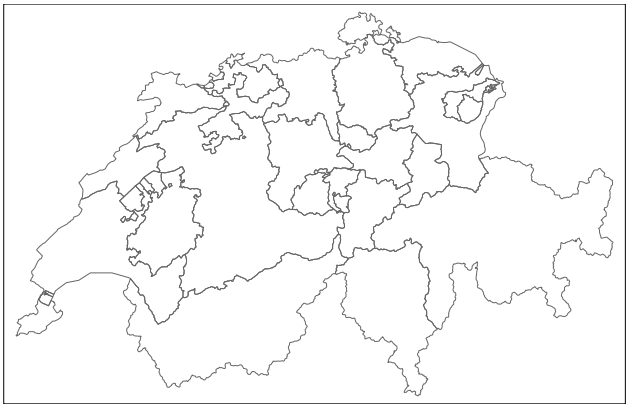

Fig. 3.1: The cantons of Switzerland, an example of an irregular lattice.

General Notation and Abbreviations

For $C \in \mathcal{I} = \{1, ..., n\}$ let $\boldsymbol{x}_C = \{x_i : i \in C\}$. $-C$ denotes the set $\mathcal{I} - \mathcal{C}$ such that $\boldsymbol{x}_{-C} = \{x_i : i \in C\}$. For two sets A and B, $A \setminus B = \{i : i \in A \text{ and } i \notin B\}$.
$\pi(\cdot)$ denotes the density of its arguments, for example $\pi(\boldsymbol{x})$ for the density of $\boldsymbol{x}$ and $\pi(\boldsymbol{x}_A|\boldsymbol{x}_{-A})$ for the conditional density of $\boldsymbol{x}_A$, given $\boldsymbol{x}_{-A}$. '$\sim$' is used when a variable is 'distributed' according to the law $\mathcal{L}$.
The expected value is denoted by $\mathbb{E}[\cdot]$, the variance by $\text{Var}(\cdot)$, the covariance by $\text{Cov}(\cdot)$, the precision by $\text{Prec}(\cdot) = \text{Cov}(\cdot)^{-1}$, the correlation by $\text{Corr}(\cdot, \cdot)$ and a probability by $\mathbb{P}(\cdot)$.

Symmetric Positive Definite Matrices

An $n \times n$ matrix $\boldsymbol{A}$ is *positive definite* exactly if

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0, \qquad \forall \boldsymbol{x} \neq \boldsymbol{0}.$$

If $\boldsymbol{A}$ is also symmetric, it is called a symmetric positive definite (SPD) matrix. Only SPD matrices are considered and sometimes the notation '$\boldsymbol{A} > 0$' is used for an SPD matrix $\boldsymbol{A}$.
An SPD matrix $\boldsymbol{A}$ has some of the following properties.

1. $\text{rank}\,(\boldsymbol{A}) = n$.
2. $|\boldsymbol{A}| > 0$.
3. $A_{ii} > 0$.
4. $A_{ii}A_{jj} - A_{ij}^2 > 0$, for $i \neq j$.
5. $A_{ii} + A_{jj} - 2|A_{ij}| > 0$ for $i \neq j$.
6. $\max A_{ii} > \max_{i \neq j} |A_{ij}|$.
7. $\boldsymbol{A}^{-1}$ is SPD.
8. All principal submatrices of $\boldsymbol{A}$ are SPD.

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are SPD, $\boldsymbol{A} + \boldsymbol{B}$ is also SPD, but the reverse is generally not true. Additionally, if $\boldsymbol{AB} = \boldsymbol{BA}$, $\boldsymbol{AB}$ is SPD.
The following conditions are all sufficient and necessary for a symmetric matrix $\boldsymbol{A}$ to be SPD:

1. All eigenvalues $\lambda_1, ..., \lambda_n$ of $\boldsymbol{A}$ are strictly positive.
2. There exists such a matrix $\boldsymbol{C}$ that $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^T$. If $\boldsymbol{C}$ is lower triangle, it is called the *Cholesky triangle* of $\boldsymbol{A}$.
3. All leading principal submatrices have strictly positive determinants.

A sufficient, but not necessary condition for a (symmetrical) matrix to be SPD is the criterion of *diagonal dominance*:

$$A_{ii} - \sum_{j:j \neq i} |A_{ij}| > 0, \qquad \forall i.$$

A $n \times n$ matrix $\boldsymbol{A}$ is called a *symmetric positive semidefinite* (SPSD) matrix. An SPSD matrix $\boldsymbol{A}$ is sometimes denoted '$\boldsymbol{A} \geq 0$'.[3]

[3] Cf. Havard Rue and Held 2005.

3.2 Basic Concepts of Bayesian Theory

3.2.1 Bayes' Theorem

At the heart of Bayesian inference is *Bayes' theorem*, which describes the probability of an event given prior knowledge of factors that might influence the event.
Let $\boldsymbol{x}^T = (x_1, ..., x_n)$ be a vector of n observations whose probability distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ depends on the values of k parameters $\boldsymbol{\theta}^T = (\theta_1, ..., \theta_k)$. Let $p(\boldsymbol{\theta})$ be the probability distribution of $\boldsymbol{\theta}$. Then

$$p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{x}). \tag{3.1}$$

Given the observed data $\boldsymbol{x}$, the conditional distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{x})}. \tag{3.2}$$

This last statement is known as Bayes' theorem. The *prior* distribution $p(\boldsymbol{\theta})$ contains knowledge about $\boldsymbol{\theta}$ without knowledge of the data. $p(\boldsymbol{\theta}|\boldsymbol{x})$ contains what is known about $\boldsymbol{\theta}$ given knowledge of the data and is the *posterior* distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$.
If $p(\boldsymbol{x}|\boldsymbol{\theta})$ is considered as a function of $\boldsymbol{\theta}$ instead of $\boldsymbol{x}$, it is called the *likelihood function* of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and can be written as $l(\boldsymbol{\theta}|\boldsymbol{x})$. Thus Bayes' theorem can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{\theta}). \tag{3.3}$$

It is evident that the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$ is proportional to the product of the distribution of $\boldsymbol{\theta}$ prior to observing the data and the likelihood function of $\boldsymbol{\theta}$ given $\boldsymbol{x}$. Therefore,

$$\text{posterior distribution} \propto \text{likelihood} \times \text{prior distribution}.$$

The data $\boldsymbol{x}$ modifies the prior knowledge of $\boldsymbol{\theta}$ through the likelihood function, and thus can be be regarded as a representation of the information about $\boldsymbol{\theta}$ derived from the data.[4]

3.2.2 Conditional Independence

In probability theory, two random variables x and y are *independent* given a third variable z if and only if the occurrence of x and y in their conditional probability

[4] Cf. Box and Tiao 2011.

distribution given z are independent events. To calculate the conditional density of $\boldsymbol{x}_A$, given $\boldsymbol{x}_{-A}$, the following statement will repeatedly be used,

$$\pi\left(\boldsymbol{x}_A|\boldsymbol{x}_{-A}\right) = \frac{\pi\left(\boldsymbol{x}_A, \boldsymbol{x}_{-A}\right)}{\pi\left(\boldsymbol{x}_A\right)} \propto \pi\left(\boldsymbol{x}\right). \tag{3.4}$$

It follows that x and y are independent precisely when $\pi\left(x, y\right) = \pi\left(x\right)\pi\left(y\right)$, which is expressed by $x \perp y$. x and y are conditionally independent for a given z if and only if $\pi\left(x, y|z\right) = \pi\left(x|z\right)\pi\left(y|z\right)$. The conditional independence can be easily validated with the help of the following *factorisation criterion*,

$$x \perp y|z \Longleftrightarrow \pi\left(x, y, z\right) = f\left(x, z\right) g\left(y, z\right), \tag{3.5}$$

for some functions f and g, and for all z with $\pi\left(z\right) > 0$.[5]

3.2.3 Undirected Graphs

Undirected graphs are used to represent the conditional independence structure in a Gaussian Markov random field. An *undirected graph* $\mathcal{G}$ is defined as a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains all nodes in the graph and $\mathcal{E}$ is the set of edges $\{i, j\}$, with $i, j \in \mathcal{V}$ and $i \neq j$. For $\{i, j\} \in \mathcal{E}$ there exists an undirected edge from node i to node j in the other case such an edge does not exist. If $\{i, j\} \in \mathcal{E} \forall i, j \in \mathcal{V}$ with $i \neq j$ a graph is *fully connected*. Most often $\mathcal{V} = \{1, 2, ..., n\}$ will be assumed, which is referred to as *labelled*. A simple example of an undirected graph is shown in Figure 3.2.

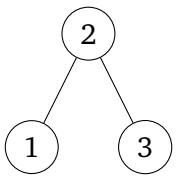

Fig. 3.2: An undirected labelled graph with 3 nodes, $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\}\{2, 3\}\}$.

The *neighbours* of node i are defined as all nodes in $\mathcal{G}$ with an edge to node i,

$$\text{ne}\left(i\right) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}.$$

[5] Cf. Havard Rue and Held 2005.

This definition can be extended to a set $A \subset \mathcal{V}$, where the neighbours of A are defined as

$$\text{ne}(A) = \bigcup_{i \in A} \text{ne}(i) \setminus A.$$

A *path* from i_1 to i_m is defined as a sequence of certain nodes in $\mathcal{V}, i_1, i_2, ..., i_m$, for which $(i_j, i_{j+1}) \in \mathcal{E}$ for $j = 1, ..., m-1$. Two nodes $i \notin C$ and $j \notin C$ are *separated* by a subset $C \subset \mathcal{V}$, if every path from i to j contains at least one node from C. Two disjoint sets $A \subset \mathcal{V} \notin C$ and $B \subset \mathcal{V} \notin C$ are separated by C, if all $i \in A$ and $j \in B$ are separated by C, that is, it is not possible to "wander" on the graph from somewhere in A and end somewhere in B without crossing C.
If i and j are neighbours in $\mathcal{G}$, this can be expressed by $i \overset{\mathcal{G}}{\sim} j$ or $i \sim j$ for the case where the graph is implicit. It follows that $i \sim j \iff j \sim i$.
Let A be a subset of $\mathcal{V}$. A *subgraph* $\mathcal{G}^A$ is a graph restricted to A, i.e., the graph obtained after removing all nodes that do not belong to A and all edges where at least one node does not belong to A. $\mathcal{G}^A = \left\{\mathcal{V}^A, \mathcal{E}^A\right\}$, where $\mathcal{V}^A = A$ and

$$\mathcal{E}^A = \{\{i, j\} \in \mathcal{A} \text{ and } \{i, j\} \in A \times A\}.$$

Let $\mathcal{G}$ be the graph in Figure 3.2 and $\mathcal{A} = \{2, 3\}$, then $\mathcal{V}^A = \{2, 3\}$ and $\mathcal{E}^A = \{\{1, 2\}\}$.[6]

3.2.4 The Exponential Family

In statistics and probability theory, the *exponential family* is a parametric set of probability distributions of a specific form. The distribution of a random variable $\boldsymbol{y}$ belongs to the exponential family if the discrete or continuous density with respect to a σ-finite measure of $\boldsymbol{y}$ has the form

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \lambda) = \exp\left(\frac{\boldsymbol{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\lambda} + c(\boldsymbol{y}, \lambda)\right), \tag{3.6}$$

with $c(\boldsymbol{y}, \lambda) \geq 0$ and measurable. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is the *natural* or *canonical* parameter of the exponential family, while $\lambda > 0$ is a *dispersion* or *nuisance* parameter. The natural parameter space Θ is the set of all $\boldsymbol{\theta}$ satisfying $0 < \int \exp\left(\left(\boldsymbol{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta})\right)/\lambda + c(\boldsymbol{y}, \lambda)\right) d\boldsymbol{y} <$

[6] Cf. Havard Rue and Held 2005.

∞. Moreover, $b(\boldsymbol{\theta})$ is a twice differentiable function and all moments of $\boldsymbol{y}$ exist. Specifically,

$$\mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \mu(\boldsymbol{\theta}) = \quad \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{3.7}$$

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \lambda \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{3.8}$$

The covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite in Θ^0, therefore $\mu : \Theta^0 \to M = \mu(\Theta^0)$ is injective. By substituting the inverse function $\theta(\mu)$ into $\frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, the variance function

$$v(\mu) = \frac{\partial^2 b(\boldsymbol{\theta}(\mu))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.9}$$

is given and the covariance can be written as

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \lambda v(\mu). \tag{3.10}$$

Important members of the exponential family are the normal, binomial, Poisson, gamma and multivariate normal distribution.[7]

3.2.5 The Multivariate Normal Distribution

The density of a normally distributed random variable $\boldsymbol{x} = (x_1, ..., x_n)^T, n < \infty$ with mean vector $\boldsymbol{\mu}$ $(n \times 1)$ and SPD covariance matrix Σ $(n \times n)$ is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^n \tag{3.11}$$

Here, $\mu_i = \mathbb{E}[x_i]$, $\Sigma_{ij} = \text{Cov}(x_i, x_j)$, $\Sigma_{ii} = \text{Var}(x_i) > 0$ and $\text{Corr}(x_i, x_j) = \Sigma_{ij} / (\Sigma_{ii} \Sigma_{jj})^{1/2}$. This is written as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For $n = 1$, $\mu = 0$ and $\Sigma_{11} = 1$ the standard normal distribution is obtained.
$\boldsymbol{x}$ is now split up into $\boldsymbol{x} = \left(\boldsymbol{x}_{\boldsymbol{A}}^T, \boldsymbol{x}_{\boldsymbol{B}}^T\right)$ and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are divided accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{A}} \\ \boldsymbol{\mu}_{\boldsymbol{B}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{AA}} & \boldsymbol{\Sigma}_{\boldsymbol{AB}} \\ \boldsymbol{\Sigma}_{\boldsymbol{BA}} & \boldsymbol{\Sigma}_{\boldsymbol{BB}} \end{pmatrix}.$$

Some basic properties of the multivariate normal distribution are the following.

1. $\boldsymbol{x}_{\boldsymbol{A}} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{A}}, \boldsymbol{\Sigma}_{\boldsymbol{AA}})$.
2. $\boldsymbol{\Sigma}_{\boldsymbol{AB}} = \boldsymbol{0}$ precisely when $\boldsymbol{x}_{\boldsymbol{A}}$ and $\boldsymbol{x}_{\boldsymbol{B}}$ are independent.

[7] Cf. Fahrmeir and Tutz 2013.

3. The conditional distribution $\pi(\boldsymbol{x_A}|\boldsymbol{x_B})$ is $\mathcal{N}\left(\boldsymbol{\mu_{A|B}}, \boldsymbol{\Sigma_{A|B}}\right)$, where

$$\boldsymbol{\mu_{A|B}} = \boldsymbol{\mu_A} + \boldsymbol{\Sigma_{AB}} + \boldsymbol{\Sigma_{BB}^{-1}}(\boldsymbol{x_B} - \boldsymbol{\mu_B}) \text{ and}$$
$$\boldsymbol{\Sigma_{A|B}} = \boldsymbol{\Sigma_{AA}} - \boldsymbol{\Sigma_{AB}}\boldsymbol{\Sigma_{BB}^{-1}}\boldsymbol{\Sigma_{BA}}.$$

4. If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{x}' \sim \mathcal{N}(\boldsymbol{\mu'}, \boldsymbol{\Sigma'})$ are independent, then $\boldsymbol{x}+\boldsymbol{x}' \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\mu'}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma'})$.[8]

3.3 Conjugate Priors

One property of exponential families is that they have conjugate priors, which is an important property in Bayesian statistics. If the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ and the prior distribution $p(\boldsymbol{\theta})$ belong to the same probability distribution family, the prior and posterior distributions are called *conjugate* distributions. Furthermore, the prior for the likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$ is called the *conjugate prior*. The term was introduced by Raiffa and Schlaifer,[9] and the property that all members of the exponential family have conjugate priorities was shown by Diaconis and Ylvisaker.[10]

3.3.1 Penalised Complexity Priors

One issue when selecting the prior distribution of a particular parameter is that it is not always intuitive when it comes to understanding and interpreting this distribution, something that is essential to ensure that it behaves as intended by the user. This problem can be addressed by using *penalised complexity priors*, which is a methodology that penalises the complexity of model components in relation to deviation from simple base model formulations.
PC priors provide a systematic and unified approach to calculating priority distributions for parameters of model components by using an inherited nested structure. This structure contains two models, the base model and a flexible version of the model. The first of the two is generally characterised by a fixed value of the relevant parameter, while the second version is considered a function of the random parameter. By penalising the deviation from the flexible model to the fixed base model, the PC prior is calculated.

[8] Cf. Havard Rue and Held 2005.
[9] Cf. Raiffa and Schlaifer n.d.
[10] Cf. Diaconis and Ylvisaker 1979.

The Principles Behind PC Priors

Four main principles should be followed to calculate priorities in a consistent way and to understand their properties.

Support to Occam's Razor

Let $\pi(x|\xi)$ denote the density of a model component x and ξ the parameter to which a prior distribution is to be assigned. The base model is characterised by a density $\pi(x|\xi = \xi_0)$, where ξ_0 is a fixed value. The prior for ξ should be such that proper shrinkage is given to ξ_0. The simplicity of the model is therefore prioritised over the complexity of the model, preventing overfitting.

Penalisation of Model Complexity

Let $f_1 = \pi(x|\xi)$ and $f_0(x|\xi = \xi_0)$ denote the flexible model and the base model respectively. The complexity of f_1 compared to f_0 is characterised using the Kullback-Leibler divergence to calculate a measure of complexity between the two models,

$$\text{KLD}(f_1||f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_0(x)}\right) dx. \tag{3.12}$$

This can be used to measure the information that is lost when f_1 is approximated by the simpler model f_0. For multinormal densities with zero mean, the calculation simplifies to

$$\text{KLD}(f_1||f_0) = \frac{1}{2}\left(\text{trace}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\right) - n - \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right), \tag{3.13}$$

where $f_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_i), i = 0, 1$, while n represents the dimension. For easier interpretation, the Kullback-Leibler divergence is transformed into a unidirectional distance measure

$$d(\xi) = d(f_1||f_0) = \sqrt{2\text{KLD}(f_1||f_0)} \tag{3.14}$$

which can be interpreted as a measure of distance from f_1 to f_0.

Constant Rate Penalisation

The derivation of the PC prior is based on a system of constant rate penalisation, given by

$$\frac{\pi_d\left(d\left(\xi\right)+\delta\right)}{\pi_d\left(d\left(\xi\right)\right)}=r^{\delta}, \qquad d\left(\xi\right),\delta\geq 0. \tag{3.15}$$

$r\in(0,1)$ represents the constant decay rate and thus implies that the relative change in the priority distribution for $d\left(\xi\right)$ is independent of the actual distance. Therefore, $d\left(\xi\right)$ is exponentially distributed with density $\pi\left(d\left(\xi\right)\right)=\lambda\exp\left(-\lambda d\left(\xi\right)\right)$ and rate $\lambda=-\ln\left(r\right)$. By a standard variable change transformation, the corresponding PC prior for ξ is given.

User-Defined Scaling

Since λ characterises the shrinkage properties of the prior, it is important that the rate can be chosen in an intuitive and interpretable way. One possibility is to determine λ by including a probability statement of tail events, for example

$$\mathbb{P}\left(Q\left(\xi\right)>U\right)=\alpha, \tag{3.16}$$

where U represents an assumed upper bound for an interpretable transformation $Q\left(\xi\right)$ and α denotes a small probability.

PC Priors for AR(1)

The first-order AR process is given by

$$x_t=\phi x_{t-1}+\epsilon_t, \qquad \epsilon\sim\mathcal{N}\left(0,\kappa^{-1}\right),\ \ t=2,...,n, \tag{3.17}$$

where x_1 is assumed to follow a normal distribution with mean 0 and marginal precision $\tau=\kappa\left(1-\phi^2\right)$. The variables $\{\epsilon_t\}_{t=1}^{n}$ are independent and follow a $\mathcal{N}\left(0,\kappa\right)$ distribution. The AR(1) model represents an important special case of AR processes where the autocorrelation coefficient ϕ specifies the complete dependence structure.

Base Model: No Dependency in Time

The correlation matrix of an AR(1) is generally defined as $\boldsymbol{\Sigma}_1 = \left(\phi^{|i-j|}\right)$. In the case of no autocorrelation, white noise results and the correlation matrix is equal to the identity matrix, $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$. The distance function is defined as $d(\phi) = \sqrt{(1-n)\log(1-\phi^2)}$. According to the constant rate penalty principle, $d(\phi)$ is assigned an exponential prior with rate $\theta/\sqrt{n-1}$. This leads to a prior distribution that is invariant to n, and the PC for the one-lag autocorrelation is given by

$$\pi(\phi) = \frac{\theta}{2}\exp\left(-\theta\sqrt{-\ln(1-\phi^2)}\right)\frac{|\phi|}{(1-\phi^2)\sqrt{-\ln(1-\phi^2)}}, \qquad |\phi|<1, \theta>0. \tag{3.18}$$

The rate parameter θ influences at what rate the prior shrinks towards the white noise base model. To infer θ, a tail event is used. In the case of $\phi = 0$ a tail event can be defined by the fact that large absolute correlations are less likely, i.e..,

$$\mathbb{P}(|\phi| > U) = \alpha.$$

This implies that $\theta = -\ln(\alpha)/\sqrt{-\ln(1-U^2)}$

Base Model: No Change in Time

As an alternative to the base model for the AR(1) process, it can be assumed that the process remains constant in time ($\phi = 1$), thus representing a limiting random walk case, which is a non-stationary and singular process. To derive the PC prior for ϕ, let $\boldsymbol{\Sigma}_1 = \left(\phi^{|i-j|}\right)$ and $\boldsymbol{\Sigma}_0 = \left(\phi_0^{|i-j|}\right)$, where ϕ_0 is close to 1 and $\phi < \phi_0$. The Kullback-Leibler divergence is

$$\begin{aligned}&\text{KLD}(f_1(\phi)\,||f_0) = \\ &\frac{1}{2}\left(\frac{1}{1-\phi_0^2}\left(n - 2(n-1)\phi_0\phi + (n-2)\phi_0^2\right) - n - (n-1)\ln\left(\frac{1-\phi^2}{1-\phi_0^2}\right)\right).\end{aligned}$$

Considering the limit as $\phi_0 \to 1$, the distance is

$$\begin{aligned}d(\phi) &= \lim_{\phi_0\to 1}\sqrt{2\text{KLD}(f_1(\phi)\,||f_0)} \\ &= \lim_{\phi_0\to 1}\sqrt{\frac{2(n-1)(1-\phi)}{1-\phi_0^2}} = c\sqrt{1-\phi}, \qquad |\phi|<1,\end{aligned}$$

constant for c, independent of ϕ. Since $0 \leq d(\phi) \leq c\sqrt{2}$, $d(\phi)$ is assigned a truncated exponential distribution with rate θ/c, resulting in the following PC prior,

$$\pi(\phi) = \frac{\theta \exp\left(-\theta\sqrt{1-\phi}\right)}{\left(1 - \exp\left(-\sqrt{2}\theta\right)\right) 2\sqrt{1-\phi}}, \qquad |\phi| < 1. \tag{3.19}$$

To scale the prior in terms of θ, (U, α) is determined in terms of $\mathbb{P}(\phi > U) = \alpha$. This equation is solved by

$$\frac{1 - \exp\left(-\theta\sqrt{1-U}\right)}{1 - \exp\left(-\sqrt{2}\theta\right)} = \alpha,$$

provided that α is larger than the lower limit $\sqrt{(1-U)/2}$.[11]

[11] Cf. Sørbye and Håvard Rue 2017.

3.4 Markov-Chain-Monte-Carlo-Methods

Markov chain Monte Carlo methods, also referred to as MCMC methods, are a set of algorithms that enable sampling from probability distributions based on the construction of Markov chains. After a sufficient number of iterations, the stationary distribution of a Markov chain can be taken as the desired distribution, with the quality of this distribution improving as the number of iterations increases. Most of the time, the construction of such a chain is relatively simple; the real challenge is to determine how many steps are needed before convergence towards the stationary distribution is achieved. MCMC methods are mostly used to compute numerical approximations of multidimensional integrals, for instance in Bayesian statistics or computational biology. The two main concepts used in MCMC methods are Monte Carlo integration and the aforementioned Markov chains, hence the name Markov Chain Monte Carlo.

3.4.1 Monte Carlo Integration

Monte Carlo integration is a technique that uses the generation of random numbers for numerical computation of definite integrals and is especially useful for higher-dimensional integrals. The problem the method addresses is the computation of the integral

$$\mathbb{E}_f\left[h\left(X\right)\right] = \int_{\chi} h(x) f(x) dx. \tag{3.20}$$

The integral can be approximated by using a sample $(X_1, ..., X_m)$ generated from f and calculating the arithmetic mean

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h\left(x_j\right). \tag{3.21}$$

According to the Strong Law of Large Numbers, $\overline{h}_m$ is likely to converge to $\mathbb{E}_f\left[h\left(X\right)\right]$. When the expectation of h^2 under f is finite, the convergence speed of $\overline{h}_m$ can be assessed. The variance too can be estimated from the sample $(X_1, ..., X_N)$ through

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} \left[h\left(x_j\right) - \overline{h}_m\right]^2. \tag{3.22}$$

For m large,

$$\frac{\overline{h}_m - \mathbb{E}_f\left[h\left(X\right)\right]}{\sqrt{v_m}} \tag{3.23}$$

is approximately distributed as a $\mathcal{N}(0,1)$ variable. This can be used for constructing a convergence test and to calculate confidence bounds for the approximation of $\mathbb{E}_f[h(X)]$.[12]

3.4.2 Markov Chains

Markov chains are stochastic processes that aim to provide the probability of the occurrence of future events. A Markov chain is defined by the fact that even if only a limited history is known, predictions about future developments can be made just as reliably as if the entire history of a process were known. Thus, the probability of moving from the current state to any state depends only on the current state of the chain. These probabilities are defined by a *transition kernel*, which is a function K on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$, such that

i. $\forall x \in \mathcal{X}, K(x,\cdot)$ is a probability measure;

ii. $\forall A \in \mathcal{B}(\mathcal{X}), K(\cdot, A)$ is measureable.

In the discrete case, the transition kernel is a matrix $\boldsymbol{K}$ with elements

$$\mathbb{P}_{xy} = \mathbb{P}(X_n = y | X_{n-1} = x), \qquad x, y \in \mathcal{X}.$$

If $\mathcal{X}$ is continuous, the kernel denotes the conditional density $K\left(x, x^T\right)$ of the transition $K(x,\cdot)$,

$$\mathbb{P}(X \in A|x) = \int_A K\left(x, x^T\right) dx^T.$$

Given a transition kernel K, a sequence $X_0, X_1, ..., X_n$ of random variables is a *Markov chain* (X_n), if, for any t, the conditional distribution of X_t given the previous states is the same as the distribution of X_t given the last state, x_{t-1},

$$\begin{aligned}\mathbb{P}(X_{k+1} \in A | x_0, x_1, x_2, ..., x_k) &= \mathbb{P}(X_{k+1} \in A | x_k) \\ &= \int_A K(x_k, dx). \end{aligned} \tag{3.24}$$

Markov chains can have certain properties that affect their long-term behaviour and are of particular importance for MCMC algorithms. Next, some of them will be introduced.

[12] Cf. Robert and Casella 2013.

Irreducibility

Irreducibility is critical to the construction of Markov chain Monte Carlo algorithms, as it ensures the convergence of such an algorithm. A Markov chain is *irreducible* if all states communicate, that is, for all states i and j the probability of getting from i to j in finite time is true positive.
Formally speaking, given a measure φ, a Markov chain (X_n) with transition kernel $K(x, y)$ is *φ-irreducible*, if, for every $A \in B(\mathcal{X})$ with $\varphi(A) > 0$, there exists n such that $K^n(x, A)\, \forall x \in \mathcal{X}$. The chain is *strongly φ-irreducible* if $n = 1 \forall$ measurable A.

Periodicity

The behaviour of a Markov chain can sometimes be limited by deterministic constraints on the transitions from X_n to X_{n+1}. For discrete chains, the *period* of a state $w \in \mathcal{X}$ is defined as.

$$d(w) = \text{g.c.d. } \{m \geq 1; K^m(w, w) > 0\},$$

with g.c.d the greatest common denominator. If a Markov chain is irreducible, the transition matrix can be written as a block matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & \boldsymbol{D}_1 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{D}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ \boldsymbol{D}_d & 0 & 0 & & 0 \end{pmatrix}, \tag{3.25}$$

It is evident that at every d-th step there is a return to the initial group. There exists only one value for the period when a chain is irreducible. If this value is 1, the irreducible chain is *aperiodic*.

Transience and Recurrence

To guarantee an acceptable approximation of a simulated model, a Markov chain needs to have good stability properties. Irreducibility is not strong enough to ensure that the trajectory of (X_n) enters A often enough. This leads to the formalisation of *recurrence* and *transience*.
In a finite space $\mathcal{X}$, a state $w \in \mathcal{X}$ is *transient* if it is finitely often visited and *recurrent* if it is almost certainly infinitely often visited.

For irreducible chains, these two properties are properties of the chain, not of a particular state.

Ergodicity

When looking at a Markov Chain (X_n) from a temporal point of view, it is essential to establish to what the chain is converging. A natural candidate for the limiting distribution is the stationary distribution π which leads to the need to define sufficient conditions on (X_n) for X_n to be asymptotically distributed according to π. There are several conditions that can be imposed on the convergence of P^n, the distribution of X_n to π. The most fundamental ans important is that of *ergodicity*, that is, independence of initial conditions.
If a Markov chain (X_n) is both aperiodic and positive recurrent, it is called an *ergodic* Markov chain.

Stationary distribution

A chain (X_n) is more stable if the marginal distribution of X_n is independent of n. This is a requirement for the existence of a probability distribution π such that $X_{n+1} \sim \pi$ if $X_n \sim \pi$. Markov chain Monte Carlo methods rely on the fact that this condition can be satisfied.
A σ-finite measure π is *invariant* for the transition kernel $K(\cdot, \cdot)$ if

$$\pi(B) = \int_{\mathcal{X}} K(x, B)\, \pi(dx), \qquad \forall B \in \mathcal{B}(\mathcal{X}).$$

This distribution is referred to as *stationary* if π is a probability measure, as $X_0 \sim \pi$ implies that $X_n \sim \pi$ is $\forall n$. An irreducible Markov chain has a stationary distribution precisely if it is positively recurrent. The distribution is then given by

$$\pi_x = \left(\mathbb{E}_x[\tau_x]\right)^{-1}, \qquad x \in \mathcal{X}, \tag{3.26}$$

where $\mathbb{E}_x[\tau_x]$ can be interpreted as the average number of transitions between two passages in x.
In practice, the stationary distributions are often of special interest. If these distributions are defined as the starting distribution of X_0, then all following distributions of the states X_n for any n are equal to the starting distribution. The interesting

question here is when such distributions exist and when any distribution converges against a stationary distribution of this kind.[13]

3.4.3 The Metropolis-Hastings Algorithm

Having established the basics of MCMC methods, one of the best known MCMC algorithms, the Metropolis-Hastings algorithm, is introduced next. It is a procedure for drawing random samples from a probability distribution from which direct sampling is difficult if a function proportional to the *target density* f is known. This function $q(\boldsymbol{y}|\boldsymbol{x})$ is called the *proposal density* and must be easy to simulate in order for the Metropolis-Hastings algorithm to be implementable. Moreover, it must be either explicitly present or *symmetric*, meaning $q(\boldsymbol{x}|\boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})$.
The Metropolis-Hastings algorithm of a target density f and proposal density q produces a Markov chain $\left(X^{(t)}\right)$ by the following transition.

Algorithm 1 The Metropolis-Hastings Algorithm

Given $f(\boldsymbol{x})$ and $q(\boldsymbol{y}|\boldsymbol{x})$
1: Initialisation: Choose arbitrary x_t as the first sample
2: **for** each iteration t **do**
3: Generate $Y_t \sim q\left(\boldsymbol{y}|x^{(t)}\right)$
4: Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \mathbb{P}\left(x^{(t)}, Y_t\right) \\ x^{(t)} & \text{with probability } 1 - \mathbb{P}\left(x^{(t)}, Y_t\right) \end{cases}$$

where

$$\mathbb{P}(x, y) = \min\left\{\frac{f(\boldsymbol{y})}{f(\boldsymbol{x})}\frac{q(\boldsymbol{x}|\boldsymbol{y})}{q(\boldsymbol{y}|\boldsymbol{x})}, 1\right\}. \tag{3.27}$$

$\mathbb{P}(x, y)$ is the *Metropolis-Hastings acceptance probability*.
The algorithm always accepts values y_t that lead to an increase in the ratio $f(y_t)/q\left(y_t|x^{(t)}\right)$ compared to the previous value $f\left(x^{(t)}\right)/q\left(x^{(t)}|y_t\right)$. In the symmetric case, the acceptance probability simplifies to

$$\mathbb{P}(x, y) = \min\left\{\frac{f(\boldsymbol{y})}{f(\boldsymbol{x})}, 1\right\}.$$

[13] Cf. Robert and Casella 2013.

If the Markov chain starts with a value $x^{(0)} > 0$, then $f\left(x^{(t)}\right) > 0 \forall t \in \mathbb{N}$ since the values of y such that $f(y_t) = 0$ will all be rejected by the algorithm. As the number of iterations t increases, the distribution of saved states $x_0, ..., x_t$ will converge towards the target density $f(\boldsymbol{x})$.[14]

3.4.4 The Gibbs Sampler

Gibbs-Sampling is a special case of the Metropolis-Hastings Algorithm, that is used to generate a sequence of samples of the joint probability distribution of two or more random variables. The aim of the method is to approximate this unknown joint probability distribution. Gibbs sampling is especially suitable when the joint distribution of a random vector is unknown, but the conditional distribution of each random variable is known. The underlying principle is to repeatedly select a variable and generate a value according to its conditional distribution, depending on the values of the other variables. During this iteration step, the values of the other variables remain unchanged. A Markov chain can be derived from the resulting sequence of sample vectors, and it can be shown that the stationary distribution of this Markov chain is precisely the sought joint distribution of the random vector.

The Two-Stage Gibbs Sampler

A general introduction to Gibbs sampling is the two-stage Gibbs sampler, which is applicable to a wide range of statistical models that do not demand the generality of the multi-stage Gibbs sampler.
Implementing the algorithm is straightforward. If the random variables X and Y have a joint density $f(\boldsymbol{x}, \boldsymbol{y})$, the two-stage Gibbs sampler generates a Markov chain (X_t, Y_t) as shown below. $f_{Y|X}$ and $f_{X|Y}$ represent the conditional densities

Algorithm 2 The Two-Stage Gibbs Sampler

Take $X_0 = x_0$
1: **for** each iteration t **do**
2: Generate $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$
3: Generate $X_t \sim f_{X|Y}(\cdot|y_t)$

associated with f. It is worth noting that not only (X_t, Y_t) is a Markov chain, but also the subsequences (X_t) and (Y_t) are.

[14] Cf. Robert and Casella 2013.

Normal Bivariate Gibbs Sampler

In the case of the bivariate normal density

$$(X, Y) \sim \mathcal{N}_2 \left(0, \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}\right)$$

the Gibbs sampler reads as follows.

Algorithm 3 The Two-Stage Gibbs Sampler for a normal distribution

Given y_t
1: **for** each iteration t **do**
2: Generate $X_{t+1}|y_t \sim \mathcal{N}\left(py_t, 1-p^2\right)$
3: Generate $Y_{t+1}|x_{t+1} \sim \mathcal{N}\left(px_{t+1}, 1-p^2\right)$

The Multi-Stage Gibbs Sampler

Let $p > 1$, then the random variable $X \in \mathcal{X}$ can be written as $X = (X_1, ..., X_p)$, where the X_i's are either one-dimensional or multidimensional. Moreover, assume that a simulation is possible from the corresponding univariate conditional densities $f_1, ..., f_p$, i.e.,

$$X_i|\boldsymbol{x}_{-i} \sim f_i\left(x_i|\boldsymbol{x}_{-i}\right)$$

can be simulated for $i = 1, ..., p$. The Gibbs sampler is then specified by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

Algorithm 4 The Multi-Stage Gibbs Sampler

Given $x^{(t)} = \left(x_1^{(t)}, ..., x_p^{(t)}\right)$, generate
1: $X_1^{(t+1)} \sim f_1\left(x_1|x_2^{(t)}, ..., x_p^{(t)}\right)$;
2: $X_2^{(t+1)} \sim f_2\left(x_2|x_1^{(t)}, x_3^{(t)}, ..., x_p^{(t)}\right)$;
$\vdots$
$X_p^{(p+1)} \sim f_p\left(x_p|\boldsymbol{x}_{-p}\right)$

$f_1, ..., f_p$ are referred to as the *full conditionals* and these are the only densities used for simulation. Hence, all simulations can be univariate, even for a high-dimensional problems.

3.5 Latent Gaussian Models and INLA

In recent years, a growing amount of georeferenced data has become available, leading to an increased need for appropriate statistical modeling to handle large and complex datasets. Bayesian hierarchical models have proven to be effective in capturing complex stochastic structures in spatial processes. A large proportion of these models are based on latent Gaussian models, a subclass of structured additive regression models.

Notation and Basic Properties

For structured additive regression models, the distribution of the response variable y_i is assumed to be a member of the exponential family, with the mean μ_i linked to a structured additive predictor η_i by a link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$. The predictor η_i takes into account the effect of multiple covariates in an additive way,

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i. \tag{3.28}$$

The $\left\{f^{(j)}(\cdot)\right\}$s are unknown functions of the covariates u, while the $\{\beta_k\}$s represent the linear effect of the covariates z and the ϵ_is are unstructured terms. Latent Gaussian models assign a Gaussian prior to α, $\left\{f^{(j)}(\cdot)\right\}$ and $\{\epsilon_i\}$. In the following $\boldsymbol{x}$ shall denote the vector of all latent Gaussian variables ($\{\eta_i\}$, α, $\left\{f^{(j)}\right\}$ and $\{\beta_k\}$) and $\boldsymbol{\theta}$ the vector of hyperparameters.
The conditional density $\pi(\boldsymbol{x}|\theta_1)$ is Gaussian with an assumed zero mean and precision matrix $\boldsymbol{Q}(\theta_1)$. The Gaussian density $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ with mean μ and covariance $\boldsymbol{\Sigma}$ at configuration x is denoted by $\mathcal{N}(\boldsymbol{x}; \mu, \boldsymbol{\Sigma})$. For simplicity, $\{\eta_i\}$ has been included instead of $\{\epsilon_i\}$.
The distribution for the n_d observational variables $y = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\boldsymbol{y}|\boldsymbol{x}, \theta_2)$ and is assumed conditionally independent given $\boldsymbol{x}$ and θ_2. Let $\boldsymbol{\theta} = \left(\theta_1^T, \theta_2^T\right)^T$ with $\dim(\boldsymbol{\theta}) = m$. For non-singular $\boldsymbol{Q}(\boldsymbol{\theta})$ the posterior is given by

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\, \pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in I} \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})\, |\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\theta})\, \boldsymbol{x} + \sum_{i \in I} \log\left\{\pi(y_i|x_i, \boldsymbol{\theta})\right\}\right].\end{aligned} \tag{3.29}$$

Most latent Gaussian models satisfy two basic properties:

1. The latent field $\boldsymbol{x}$ is of large dimension, $n \approx 10^2 - 10^5$. Therefor, the latent field is a Gaussian Markov random field with sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$.
2. The number of hyperparameters, m, is small, $m \leq 6$.

In most cases, both properties are required to produce fast inference, and thus these will be assumed to be true for the remainder of this work.[15]

3.5.1 Applications for Latent Gaussian Models

Latent Gaussian models can be employed in a vast range of different domains, in fact most structured Bayesian models are of this particular form. Some of these domains are presented below.

Regression Models

Bayesian generalised linear models correspond to the linear relationship $\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki}$. Either the linear relationship of the covariates, random effects or both can be introduced using the $f(\cdot)$ terms. Smooth covariate effects are frequently modeled using penalised spline models or random walk models, continuous indexed spline models or Gaussian processes. The incorporation of random effects allows for the consideration of overdispersion caused by unobserved heterogeneity or correlation in longitudinal data and can be introduced by defining $f(u_i) = f_i$ and $\{f_1\}$ to be independent, zero mean and Gaussian.

Dynamic Models

Temporal dependence can be introduced by using i in (3.28) as temporal index t and defining $f(\cdot)$ and $\boldsymbol{u}$ such that $f(u_t) = f_t$. Both a discrete-time and a continuous-time autoregressive model can be modeled by $\{f_t\}$. Furthermore, a seasonal effect or the latent process of a structured time series model can be modeled. Alternatively, a smooth temporal function in the same sense as for regression models can be represented by $\{f_t\}$.

[15] Cf. Håvard Rue et al. 2009.

Spatial and Spatio-Temporal Models

Similar to the previous type of model, spatial dependence can be modeled by a spatial covariate $\boldsymbol{u}$ such that $f(u_s) = f_s$, where s denotes the spatial location or region s. The stochastic model for f_s is constructed to promote spacial smooth realisations of some sort. Popular models of this type include the Besag-York-Mollié[16] model with extensions for regional data, continuous indexed Gaussian models and texture models. The dependence between spatial and temporal covariates can be achieved either by using a spatio-temporal covariate (s, t) or a corresponding spatio-temporal Gaussian field.
Often the final model consists of a sum of several components, e.g. a spatial component, random effects and both linear and smooth effects of some covariates. In order to separate the effects of the different components in (3.28), sometimes linear or sum-to-zero constraints can be imposed.[17]

3.5.2 The MCMC Approach to Inference

The usual approach to inference for latent Gaussian models involves the previously introduced Markov chain Monte Carlo methods. Due to several factors, these methods may perform poorly when applied to such models. One factor is the interdependence of the components of the latent field $\boldsymbol{x}$ while another is that $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are highly dependent on each other, especially for large n. The first of these problems can potentially be overcome by constructing a joint proposal based on a Gaussian approximation of the full conditional of $\boldsymbol{x}$, while the second problem requires, at least in part, a joint update of $\boldsymbol{\theta}$ and $\boldsymbol{x}$. There are several proposals to solve these shortcomings, but MCMC sampling continues to show poor performance from the end user's point of view.[18]

3.5.3 Gaussian Random Fields

Let $\boldsymbol{s} = (s_1, ..., s_n)^T$ be a vector of locations. A *Gaussian random field* (GRF)

$$\left\{ Z(s) : s \in D \subset \mathbb{R}^2 \right\} \tag{3.30}$$

[16] Cf. Besag et al. 1991.
[17] Cf. Håvard Rue et al. 2009.
[18] Cf. Håvard Rue et al. 2009.

is a set of random variables where the observations occur in a continuous domain and where each finite set of random variables follows a multivariate normal distribution. A random process $Z(\cdot)$ is strictly stationary if it is invariant to shifts, i.e., if for each set of locations and each $h \in \mathbb{R}^2$ the distribution of $\boldsymbol{Z(s)} = (Z(s_1), ..., Z(s_n))$ is equal to that of $\boldsymbol{Z(s+h)} = (Z(s_1 + h), ..., Z(s_n + h))$. A less constraining requirement is given by second-order stationarity. Under this condition, the process has a constant mean value

$$\mathbb{E}[\boldsymbol{Z(s)}] = \mu, \qquad \forall s \in D, \tag{3.31}$$

and the covariances depend only on the differences between locations

$$\text{Cov}(\boldsymbol{Z(s)}, \boldsymbol{Z(s+h)}) = C(\boldsymbol{h}), \qquad \forall \boldsymbol{s} \in D, \forall \boldsymbol{h} \in \mathbb{R}^2. \tag{3.32}$$

Furthermore, if the covariances depend only on the distances between the locations and not on the directions, the process is called isotropic. Else, the process is anisotropic. An intrinsically stationary process has a constant mean value and satisfies

$$\text{Var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j), \qquad \forall s_i, s_j. \tag{3.33}$$

$2\gamma(\cdot)$ is the variogram and $\gamma(\cdot)$ is called the semivariogram. Under the assumption of intrinsic stationarity, the constant-mean assumption implies

$$2\gamma(\boldsymbol{h}) = \text{Var}(\boldsymbol{Z(s+h)} - \boldsymbol{Z(s)}) = \mathbb{E}\left[(\boldsymbol{Z(s+h)} - \boldsymbol{Z(s)})^2\right],$$

and the estimation of the semivariogram can be obtained using the empirical semivariogram as follows:

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (Z(s_i) - Z(s_j))^2, \tag{3.34}$$

where $N(\boldsymbol{h}) = \{(s_1, s_j) : s_i - s_j = \boldsymbol{h}, i, j = 1, ..., n\}$ denotes the number of pairs and $|N(\boldsymbol{h})|$ the number of distinct pairs. For isotropic processes, the semivariogram is a function of distance $h = ||\boldsymbol{h}||$.
Plotting the empirical semivariogram against the separation distance conveys essential information regarding the continuity and spatial variability of the process. Given relatively short distances, the semivariogram tends to be small but increases with distance, indicating the similarity of observations in close proximity. The semivariogram levels off to a nearly constant value, also called the sill, as the separation distance increases, indicating a decrease in spatial dependence with distance within the range and no spatial correlation outside the range, which is reflected in a nearly constant

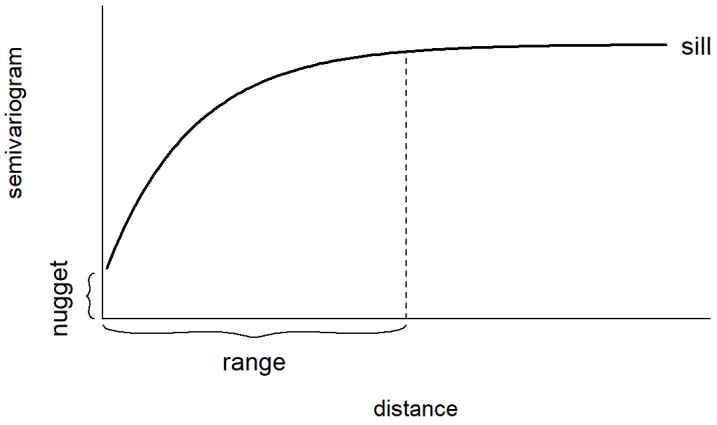

Fig. 3.3: A typical semivariogram

variance. If there is a discontinuity or a vertical jump at the origin, the process has a nugget effect, which is often due to a measurement error, but may also be indicative of a spatially discontinuous process. The empirical semivariogram is an exploratory tool useful for assessing whether data exhibit spatial correlation. Furthermore, it can be compared to a Monte Carlo envelope of empirical semivariograms calculated from random permutations of the data while keeping the locations fixed. If the empirical semivariogram lies outside the Monte Carlo envelope with increasing distance, this is an indication of spatial correlation.
The dependence structure of a GRF is given by the covariance matrix, which is constructed from a covariance function. Matérn models and exponential functions are conventionally used for this purpose. For the locations $s_i, s_j \in \mathbb{R}^2$ the exponential covariance function is given by

$$\text{Cov}\left(Z\left(s_i\right), Z\left(s_j\right)\right) = \sigma^2 \exp\left(-\kappa \left|\left|s_i - s_j\right|\right|\right), \tag{3.35}$$

where the distance between the locations s_i and s_j is denoted by $||s_i - s_j||$, the variance of the spatial field is given by σ^2, while $\kappa > 0$ controls the rate at which the correlation decays as the distance increases.
The Matérn family represents a flexible class of covariance functions that arises

naturally in a variety of scientific fields. The Matérn covariance function is written as

$$\text{Cov}\left(Z\left(s_i\right), Z\left(s_j\right)\right) = \frac{\sigma^2}{2^{\nu-1}\Gamma\left(\nu\right)}\left(\kappa\left|\left|s_i - s_j\right|\right|\right)^{\nu} K_{\nu}\left(\kappa\left|\left|s_i - s_j\right|\right|\right). \tag{3.36}$$

σ^2 denotes the marginal variance of the spatial field, $K_{\nu}\left(\cdot\right)$ represents the modified Bessel function of second kind and order $\nu > 0$, where ν is an integer. The mean square differentiability of the process is determined by ν and is usually fixed since it is difficult to identify in applications. For $\nu = 0.5$, this covariance function is the equivalent of the exponential covariance function. $\kappa > 0$ is related to the range ρ, which is defined as the distance at which there is approximately no correlation between two given points, $\rho = \sqrt{8\nu}/\kappa$ to be exact. Examples of these two covariance functions are shown below.[19]

3.5.4 Gaussian Markov Random Fields

Definition of GMRFs

Let $\boldsymbol{x} = (x_1, ..., x_n)^T$ be normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, ..., \}$ and $\mathcal{E}$ be such that there is no edge between nodes i and j exactly when $x_i \perp x_j | \boldsymbol{x}_{ij}$. Then $\boldsymbol{x}$ is a *Gaussian Markov random field* (GMRF) with respect to $\mathcal{G}$.
Since $\boldsymbol{\mu}$ does not affect the pairwise conditional independence properties of $\boldsymbol{x}$, this information is 'hidden' in $\boldsymbol{\Sigma}$. Hence,

$$x_i \perp x_j | x_{ij} \iff Q_{ij} = 0.$$

Therefore, the non-zero pattern of $\boldsymbol{Q}$ determines $\mathcal{G}$, i.e. whether x_i and x_j are conditionally independent, and can be derived from $\boldsymbol{Q}$. If $\boldsymbol{Q}$ is a fully dense matrix, then $\mathcal{G}$ is fully connected, implying that any normal distribution with SPD covariance matrix is a GMRF and vice versa.

[19] Cf. Moraga 2019.

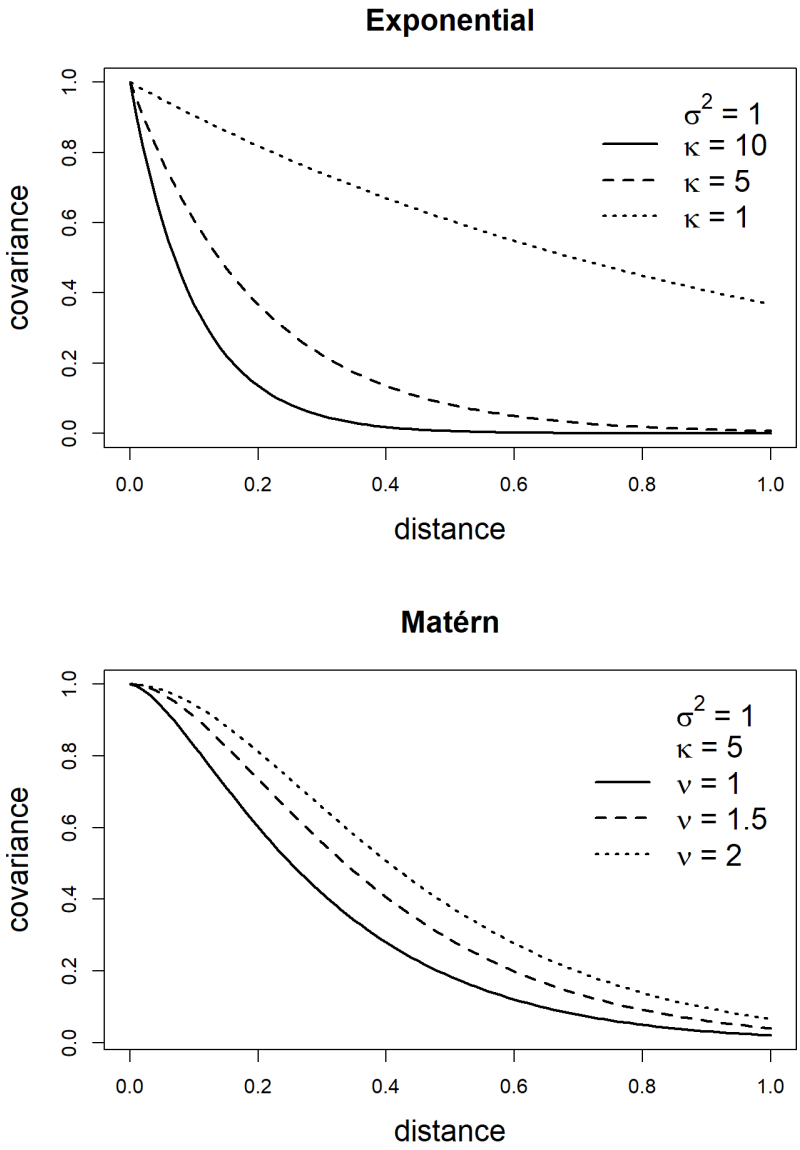

Fig. 3.4: Covariance functions corresponding to exponential and Matérn models.

The elements of $\boldsymbol{Q}$ are used for conditional interpretations. For any GMRF with respect to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$,

$$\mathbb{E}\left[x_i|\boldsymbol{x}_{-i}\right] = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}\left(x_j - \mu_j\right), \tag{3.37}$$

$$\text{Prec}\left(x_i|\boldsymbol{x}_{-i}\right) = Q_{ii} \quad \text{and} \tag{3.38}$$

$$\text{Corr}\left(x_i, x_j|\boldsymbol{x}_{ij}\right) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \qquad i \neq j. \tag{3.39}$$

On the main diagonal of $\boldsymbol{Q}$ are the conditional precisions of x_i given $\boldsymbol{x}_{-i}$ are placed, while the other elements, when scaled appropriately, provide information about the conditional correlation between x_i and x_j given $\boldsymbol{x}_{ij}$. Since $\text{Var}\left(x_i\right) = \Sigma_{ii}$ and $\text{Corr}\left(x_i, x_j\right) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, the information about the marginal variance of x_i and the marginal correlation between x_i and x_j is given by $\boldsymbol{\Sigma}$. The marginal interpretation provided by the correlation matrix is intuitive and informative, as the scope of the interpretation is reduced from an n-dimensional distribution to a one- or two-dimensional distribution. $\boldsymbol{Q}$ is difficult to interpret marginally because either $\boldsymbol{x}_{-i}$ or $\boldsymbol{x}_{ij}$ would have to be integrated out of the joint distribution parameterized with respect to $\boldsymbol{Q}$. $\boldsymbol{Q}^{-1} = \boldsymbol{\Sigma}$ by definition, and in general Σ_{ii} depends on each element in $\boldsymbol{Q}$ and vice versa.

Markov Properties of GMRFs

One property of GMRFs is that more information regarding conditional independence can be extracted from $\mathcal{G}$. The following three properties are equivalent.
The *pairwise Markov property*:

$$x_i \perp x_j|\boldsymbol{x}_{ij} \qquad \text{if } \{i,j\} \notin \mathcal{E} \text{ and } i \neq j.$$

The *local Markov property*:

$$x_i \perp \boldsymbol{x}_{-\{i,\text{ne}(i)\}}|\boldsymbol{x}_{\text{ne}(i)} \qquad \forall i \in \mathcal{V}.$$

The *global Markov property*:

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B|\boldsymbol{x}_C$$

for all disjoint sets A, B and C where A and B are non-empty and separated by C.

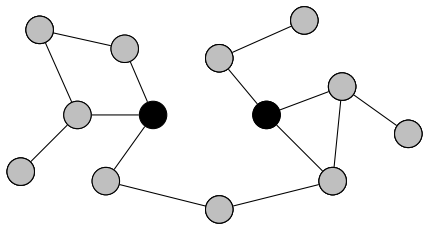

Fig. 3.5: The pairwise Markov property; the black nodes are conditionally independent given the light gray nodes.

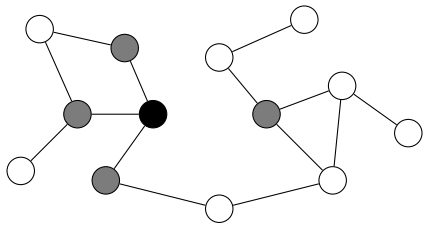

Fig. 3.6: The local Markov property; the black nodes and white nodes are conditionally independent given the dark gray nodes.

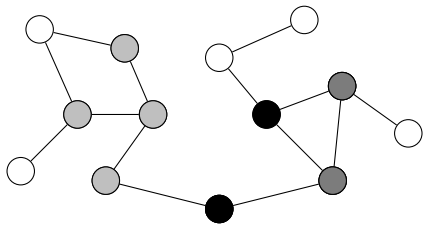

Fig. 3.7: The global Markov property; the dark gray and light gray nodes are globally independent given the black nodes.

Conditional Properties of GMRFs

An essential result of GMRFs is the conditional distribution for a subset $\boldsymbol{x}_a$ given $\boldsymbol{x}_{-A}$. Here the canonical parameterisation proves useful, since by definition it can be easily updated by successive conditioning.
By splitting the indices into the non-empty sets A and B, of which the latter is equal to -A,

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_A \\ \boldsymbol{x}_B \end{pmatrix}. \tag{3.40}$$

The mean and the precision are divided accordingly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix}. \tag{3.41}$$

The conditional distribution of $\boldsymbol{x}_A|\boldsymbol{x}_B$ is then a GMRF with respect to the subgraph $\mathcal{G}^A$ with mean $\boldsymbol{\mu}_{A|B}$ and precision matrix $\boldsymbol{Q}_{A|B} > 0$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}\left(\boldsymbol{x}_B - \boldsymbol{\mu}_B\right) \tag{3.42}$$

and

$$\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA}.$$

Thus, the explicit knowledge of $\boldsymbol{Q}_{A|B}$ is available through $\boldsymbol{Q}_{AA}$, i.e. no calculation is required to obtain the conditional precision matrix. Moreover, the conditional mean depends only on the values of $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ in $A \cup \text{ne}\left(A\right)$, since $Q_{ij} = 0 \, \forall j \not\in \text{ne}\left(i\right)$.
For successive conditioning, the canonical parameterisation for GMRF is useful.
A GMRF $\boldsymbol{x}$ with respect to $\mathcal{G}$ and canonical parameters $\boldsymbol{b}$ and $\boldsymbol{Q} > 0$ has the density

$$\pi\left(\boldsymbol{x}\right) \propto \exp\left(-1\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right).$$

The precision matrix is $\boldsymbol{Q}$ and the mean is $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$. The canonical parameterisation is written as

$$\boldsymbol{x} \sim \mathcal{N}_C\left(\boldsymbol{b}, \boldsymbol{Q}\right).$$

Furthermore,

$$\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right) \Longleftrightarrow \mathcal{N}_C\left(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q}\right).$$

If the indices are partitioned into two non-empty sets A and B and $\boldsymbol{x}$, $\boldsymbol{b}$ and $\boldsymbol{Q}$ are partitioned as in (3.40) and (3.41), then

$$\boldsymbol{x}_A|\boldsymbol{x}_B \sim \mathcal{N}_C\left(\boldsymbol{b}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B, \boldsymbol{Q}_{AA}\right). \tag{3.43}$$

Let $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x}, \boldsymbol{P}^{-1}\right)$ and $\boldsymbol{x} \sim \mathcal{N}_C\left(\boldsymbol{b}, \boldsymbol{Q}\right)$, then

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}_C\left(\boldsymbol{b} + \boldsymbol{P}\boldsymbol{y}, \boldsymbol{Q} + \boldsymbol{P}\right). \tag{3.44}$$

This allows the calculation of conditional densities with multiple sources of conditioning, e.g. conditioning on observed data and a subset of variables. Therefore, the canonical parameterisation can be repeatedly updated without explicitly calculating the mean until it is actually needed. The computation of the mean requires the

solution of $\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{b}$, but only matrix-vector products are needed for updating the canonical parameterisation.

Specification Through Full Conditionals

Alternatively, a GMRF can be specified by the full conditionals $\{\pi(x_i|\boldsymbol{x}_{-i})\}$ in place of $\boldsymbol{\mu}$ and $\boldsymbol{Q}$. Suppose the full conditionals are given as normals with

$$\mathbb{E}[x_i|\boldsymbol{x}_{-i}] = \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \quad \text{and} \tag{3.45}$$

$$\text{Prec}(x_i|\boldsymbol{x}_{-i}) = \kappa_i > 0 \tag{3.46}$$

for $i = 1, ..., n$, for $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$ and some $\{\eta_{ij}, i \neq j\}$. Evidently, $\sim$ is implicitly defined by the non-zero terms of $\{\beta_{ij}\}$. For there to exist a joint density $\pi(\boldsymbol{x})$ leading to these full conditional distributions, these full conditionals must be consistent. Since $\sim$ is symmetric, it follows that if $\beta_{ij} \neq 0$, then $\beta_{ji} \neq 0$. If the entries of the precision matrix are chosen such that

$$Q_{ii} = \kappa_i, \quad \text{and} \quad Q_{ij} = \kappa_i \beta_{ij}$$

and $\boldsymbol{Q}$ must be symmetrical, i.e.,

$$\kappa_i \beta_{ij} = \kappa_j \beta_{ji},$$

then $\boldsymbol{x}$ is a GMRF with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} = (Q_{ij})$.

Multivariate GMRFs

A *multivariate GMRF* (MGMRF) is a multivariate extension of a GMRF that has proven useful in applications. Let $\boldsymbol{x}$ be a GMRF with respect to $\mathcal{G}$, then the Markov property implies that

$$\pi(x_i|\boldsymbol{x}_{-i}) = \pi(x_i|\{x_j : j \sim i\}).$$

x_i is the value related to node i. Often the nodes have physical interpretations such as an administrative region of a country, which can be used to define the neighbours of node i. Let each of the n nodes have an associated vector $\boldsymbol{x}_i$ of dimension p,

resulting in a GMRF of size np. Such a GMRF is denoted by $\boldsymbol{x} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T\right)^T$. The Markov property with respect to the nodes is preserved, i.e.,

$$\pi\left(\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right) = \pi\left(\boldsymbol{x}_i|\left\{\boldsymbol{x}_j : j \sim i\right\}\right),$$

where $\sim$ is with respect to *the same graph* $\mathcal{G}$. Let $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^T, ..., \boldsymbol{\mu}_n^T\right)^T$ be the mean of $\boldsymbol{x}$, where $\mathbb{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{\mu}_i$, and $\widetilde{\boldsymbol{Q}} = \left(\widetilde{\boldsymbol{Q}}_{ij}\right)$ its precision matrix, where each element of the matrix is a $p \times p$ matrix.
It follows that

$$\boldsymbol{x}_i \perp \boldsymbol{x}_j|\boldsymbol{x}_{-ij} \Longleftrightarrow \widetilde{\boldsymbol{Q}}_{ij} = \boldsymbol{0}.$$

Formally, a random vector $\boldsymbol{x} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T\right)^T$ with $\dim\left(\boldsymbol{x}_i\right) = p$, is called a MGMRF$_p$ with respect to $\mathcal{G} = \left(\mathcal{V} = \{1, ..., n\}, \mathcal{E}\right)$ with mean $\boldsymbol{\mu}$ and precision matrix $\widetilde{\boldsymbol{Q}} > 0$, exactly when its density has the form

$$\begin{aligned}\pi\left(\boldsymbol{x}\right) &= \left(\frac{1}{2\pi}\right)^{np/2} \left|\widetilde{\boldsymbol{Q}}\right|^{1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \widetilde{\boldsymbol{Q}}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right) \\ &= \left(\frac{1}{2}\right)^{np/2} \left|\widetilde{\boldsymbol{Q}}\right|^{1/2} \exp\left(-\frac{1}{2}\sum_{ij}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_i\right)^T \widetilde{\boldsymbol{Q}}_{ij}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_j\right)\right)\end{aligned}$$

and

$$\widetilde{\boldsymbol{Q}}_{ij} \neq 0 \Longleftrightarrow \{i, j\} \in \mathcal{E}\, \forall\, i \neq j.$$

A MGMRF$_p$ is equivalent to a GMRF of dimension np with identical mean vector and precision matrix. Therefore, all results valid for a GMRF are also valid for a MGMRF$_p$, with modifications, since the graph for a MGMRF$_p$ has size n and is defined with respect to $\{\boldsymbol{x}_i\}$, while for a GMRF it has size np and is defined with respect to $\{x_i\}$.
The interpretation of $\widetilde{\boldsymbol{Q}}_{ii}$ and $\widetilde{\boldsymbol{Q}}_{ij}$ can be derived from the full conditional $\pi\left(\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right)$. The extensions of (3.37) and (3.38) are

$$\mathbb{E}\left[\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right] = \boldsymbol{\mu}_i - \widetilde{\boldsymbol{Q}}_{ii}^{-1}\sum_{j:j\sim i}\widetilde{\boldsymbol{Q}}_{ij}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_j\right) \tag{3.47}$$

$$\text{Prec}\left(\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right) = \widetilde{\boldsymbol{Q}}_{ii}. \tag{3.48}$$

In some applications, the full conditionals

$$\mathbb{E}\left[\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right] = \boldsymbol{\mu}_i - \sum_{j:j\sim i}\boldsymbol{\beta}_{ij}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_j\right) \tag{3.49}$$

$$\text{Prec}\left(\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right) = \boldsymbol{\kappa}_i > 0, \tag{3.50}$$

In some applications, the full conditionals are used to define the MGMRF$_p$, for given $p \times p$-matrices $\{\boldsymbol{\beta}_{ij}, i \neq j\}$, $\{\boldsymbol{\kappa}_i\}$, and vectors $\boldsymbol{\mu}_i$. Again, $\sim$ is implicitly defined by the non-zero matrices $\{\boldsymbol{\kappa}_i\}$. Similar requirements as for $p = 1$ apply to the existence of the joint density: $\boldsymbol{\kappa}_i\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{ij}^T\boldsymbol{\kappa}_j$ for $i \neq j$ and $\widetilde{\boldsymbol{Q}} > 0$. The $p \times p$ elements of $\widetilde{\boldsymbol{Q}}$ are

$$\widetilde{\boldsymbol{Q}}_{ij} = \begin{cases} \boldsymbol{\kappa}_i\boldsymbol{\beta}_{ij} & i \neq j \\ \boldsymbol{\kappa}_i & i = j \end{cases};$$

therefore $\widetilde{\boldsymbol{Q}} > 0 \Longleftrightarrow (\boldsymbol{I} + (\boldsymbol{\beta}_{ij})) > 0$.[20]

3.5.5 Integrated Nested Laplace Approximation

An alternative to MCMC methods that is both less computationally intensive and suitable for performing approximate Bayesian inference in latent Gaussian models is *Integrated nested Laplace Approximation* (INLA). The basis of INLA is the use of a combination of analytical approximations and numerical algorithms for sparse matrices to approximate the posterior distribution using closed-form expressions. This speeds up inference and circumvents problems of sample convergence and mixing, making it suitable for fitting large data sets or exploring other models.
INLA can be used for all models of the following form,

$$\begin{aligned} y_i|\boldsymbol{x}, \boldsymbol{\theta} &\sim \pi\left(y_i|x_i, \boldsymbol{\theta}\right), \qquad i = 1, ..., n, \\ \boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}\left(\mu\left(\boldsymbol{\theta}\right), \boldsymbol{Q}\left(\boldsymbol{\theta}\right)^{-1}\right), \\ \boldsymbol{\theta} &\sim \pi\left(\boldsymbol{\theta}\right). \end{aligned}$$

As introduced in section 3.5, $\boldsymbol{y}$ are the observed data, $\boldsymbol{x}$ is a Gaussian field, $\boldsymbol{\theta}$ represents the hyperparameters, while $\mu\left(\boldsymbol{\theta}\right)$ and $\boldsymbol{Q}\left(\boldsymbol{\theta}\right)$ denote the mean and precision matrix respectively.To ensure fast inference, the dimension of the hyperparameter vector $\boldsymbol{\theta}$ should be small, since the approximations are computed by numerical integration over the hyperparameter space.
In most cases, the observations y_i are assumed to belong to the exponential family with mean $\mu_i = g^{-1}\left(\eta_i\right)$. As shown in equation (3.28), η_i accounts for the effects of several covariates in an additive way, which makes it suitable for a wide range of models, including spatial and spatio-temporal models, since $\left\{f^{(j)}\right\}$ can take very different forms.
Let $\boldsymbol{x} = \left(\alpha, \{\beta_k\}\,|\theta \sim \mathcal{N}\left(\mu\left(\theta\right), \boldsymbol{Q}\left(\theta\right)^{-1}\right)\right)$ be the vector of latent Gaussian variables,

[20] Cf. Havard Rue and Held 2005.

and let $\boldsymbol{\theta}$ be the vector of hyperparameters, which are not required to be Gaussian. INLA calculates accurate and fast approximations for the posterior marginals of the components of the latent Gaussian variables

$$\pi\left(x_i|\boldsymbol{y}\right), \qquad i = 1, ..., n,$$

as well as the posterior marginals for the hyperparameters of the latent Gaussian model

$$\pi\left(\theta_j|\boldsymbol{y}\right), \qquad j = 1, ..., \dim\left(\boldsymbol{\theta}\right).$$

For each element x_i of $\boldsymbol{x}$ the posterior marginals are given by

$$\pi\left(x_i|\boldsymbol{y}\right) = \int \pi\left(x_i|\boldsymbol{\theta}, \boldsymbol{y}\right) \pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}, \tag{3.51}$$

and the posterior marginal for the hyperparameters can be expressed by

$$\pi\left(\theta_j|\boldsymbol{y}\right) = \int \pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}_{-j}. \tag{3.52}$$

$\pi\left(x_i|\boldsymbol{y}\right)$ is approximated by combining analytical approximations to the full conditionals $\pi\left(x_i|\boldsymbol{\theta}, \boldsymbol{y}\right)$ and $\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ and numerical integration routines to integrate out $\boldsymbol{\theta}$. Similarly, $\pi\left(\theta_j|\boldsymbol{y}\right)$ is approximated by approximating $\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ and integrating out $\boldsymbol{\theta}_{-j}$. In particular, the posterior density of $\boldsymbol{\theta}$ is obtained through Gaussian approximation for the posterior of the latent field, $\widetilde{\pi}_G\left(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}\right)$, evaluated at the posterior mode, $x^*\left(\boldsymbol{\theta}\right) = \arg\max_{\boldsymbol{x}} \pi_G\left(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}\right)$,

$$\widetilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto \left.\frac{\pi\left(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}\right)}{\widetilde{\pi}_G\left(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}\right)}\right|_{\boldsymbol{x}=x^*(\boldsymbol{\theta})}. \tag{3.53}$$

Next, the following nested approximations are constructed,

$$\widetilde{\pi}\left(x_i|\boldsymbol{y}\right) = \int \widetilde{\pi}\left(x_i|\boldsymbol{\theta}, \boldsymbol{y}\right) \widetilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}, \qquad \widetilde{\pi}\left(\theta_j|\boldsymbol{y}\right) = \int \widetilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}_{-j}. \tag{3.54}$$

Finally, these approximations are numerically integrated with respect to $\boldsymbol{\theta}$

$$\widetilde{\pi}\left(x_i|\boldsymbol{y}\right) = \sum_k \widetilde{\pi}\left(x_i|\theta_k, \boldsymbol{y}\right) \widetilde{\pi}\left(\theta_k|\boldsymbol{y}\right) \times \Delta_k, \tag{3.55}$$

$$\widetilde{\pi}\left(\theta_j|\boldsymbol{y}\right) = \sum_l \widetilde{\pi}\left(\theta_l^*|\boldsymbol{y}\right) \times \Delta_l^*, \tag{3.56}$$

with Δ_k and Δ_l^* representing the area weights corresponding to θ_k and θ_l^*.
To obtain the approximations for the posterior marginals for the x_i's conditioned on selected values of θ_k and $\widetilde{\pi}\left(x_i|\theta_k, \boldsymbol{y}\right)$, a Gaussian, Laplace or simplified Laplace approximation can be used. Using a Gaussian approximation derived from $\widetilde{\pi}_G\left(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}\right)$

is the simplest and fastest solution, but in some situations it produces errors in the location and is unable to capture skewness behaviour. Therefore, the Laplace approximation is favoured over the Gaussian approximation, although it is relatively expensive. The simplified Laplace approximation is associated with lower costs and addresses inaccuracies of the Gauss approximation in terms of location and skewness in a satisfactory manner.[21]

[21] Cf. Moraga 2019.

4 Analysis of Geospatial Health Data

> *Innovation distinguishes between a leader and a follower.*
>
> **— Steve Jobs**
> (CEO Apple Inc.)

4.1 Geographic Data

In spatial statistics, two fundamental types of geographic data exist, namely *vector data* and *raster data*. In the vector data model, the world is represented by points, lines and polygons with discrete, well-defined boundaries, which tends to result in high accuracy. Raster data, on the other hand, divides the surface into cells of uniform size, and raster datasets are used as the basis for background images in web mapping.
Determining which data type to use depends on the domain of the application. Vector data dominates in the social sciences because human settlements typically have discrete boundaries, while raster data are commonly used in many environmental sciences because they are based on remote sensing data. Naturally, there is also some overlap and both types can be used together or one form can be converted into the other.

4.1.1 Vector Data

The geographic vector data model is based on points located within a *coordinate reference system* (CRS), in which points either represent self-standing features or form more complex geometric shapes, i.e. lines and polygons. Using this system, Trondheim can be represented by the coordinates $(10.4, 63.4)$, meaning 10.4 degrees east of the prime meridian and 63.4 degrees north of the equator. It could also be written as $(1157722.70, 9199010.75)$, which is the position of Trondheim using the

Web Mercator projection, the de facto standard for web mapping applications. More will be said about CRS later, but for now it is sufficient to know that it is possible to display coordinates in various ways.

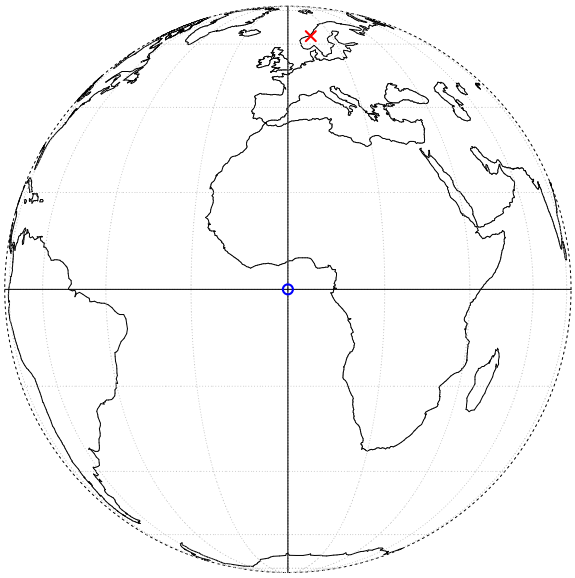

Fig. 4.1: A geographic CRS with an origin at 0° longitude an latitude. The red X denotes the location of Trondheim.

Different Types of Vector Data

As mentioned earlier, there are different types of vector data. There are 17 different geometry types in the standard *simple features*, but there are seven core types that can be used in most analysis software. These types are visualised in the following graphic.

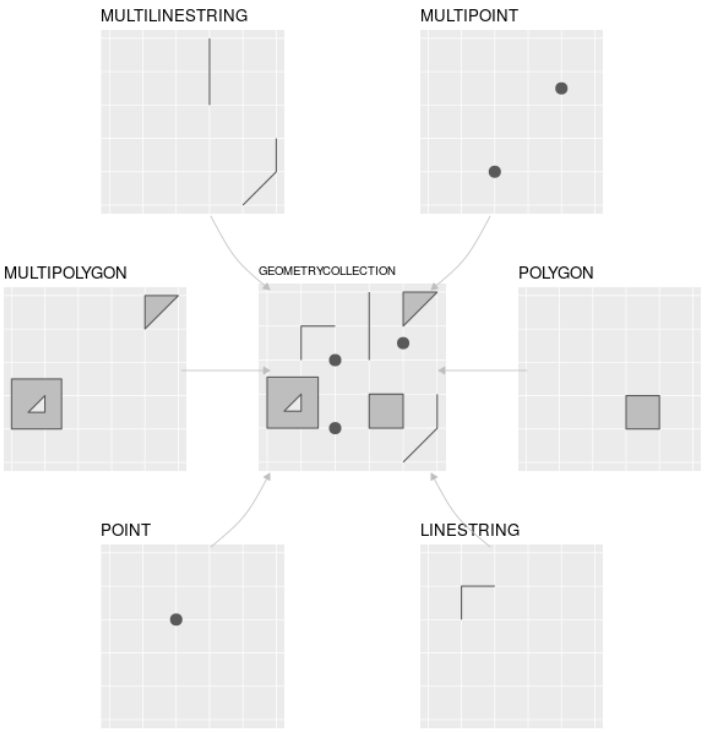

Fig. 4.2: The most commonly used simple feature types.

Simple Features was developed by the Open Geospatial Consortium and is an open, standardised, hierarchical data model that represents a wide range of geometry types. The use of this data model ensures that scientific work can be transferred to other institutions, e.g. when importing from and exporting to spatial databases.

4.1.2 Raster Data

The geographic raster data model consists in most cases of a raster header and a matrix representing uniformly distributed cells/pixels. The raster header defines the CRS, the origin (starting point) and the extent. Since the number of columns and rows and the resolution of the cell size are stored in the extent, starting from the origin, it is easy to access and change each cell by its ID or by specifying the row and column number. In this type of representation, the coordinates of the four vertices of each cell are not explicitly stored, instead only the origin is stored. This speeds up data processing and makes it more efficient, but each raster layer can only contain a single value, which can be either numeric or categorical. Typically, raster maps are

used to depict continuous features such as elevation or temperature, but categorical variables, for example soil or land cover.

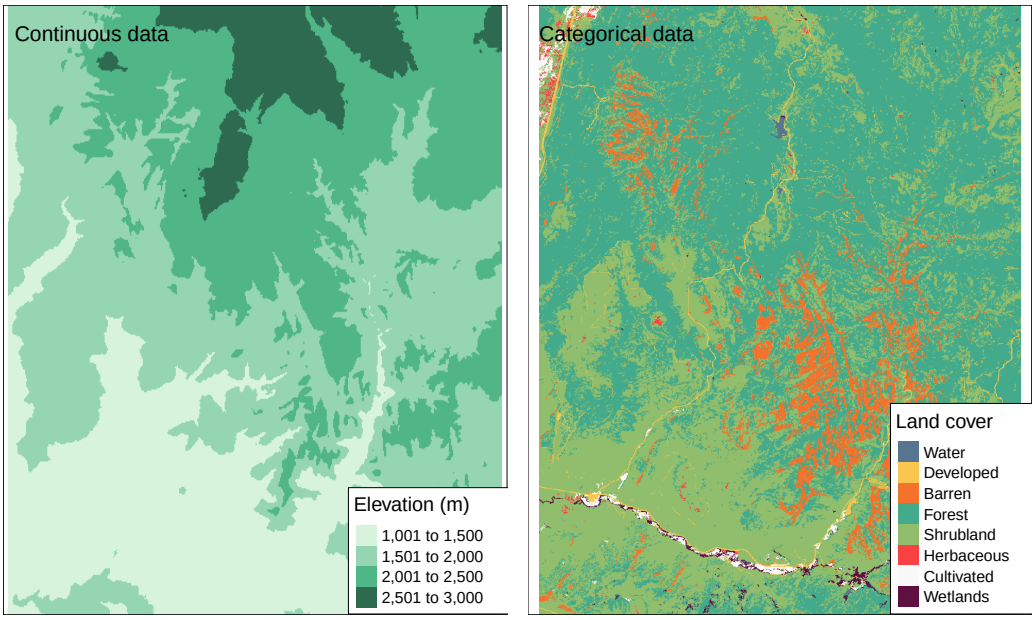

Fig. 4.3: An example of continuous and categorical raster data

Coordinate Reference Systems

A common denominator of vector and raster data are that both use the coordinate reference system (CRS), which defines how spatial elements relate to the surface of the Earth. The CRS can be either geographic or projected.

Geographic Coordinate Systems

Geographic coordinate systems use two values, *longitude* and *latitude*, to identify any location on Earth. Longitude is defined as the east-west location at an angular distance from the prime meridian plane, while latitude is the angular distance north or south of the equator. Consequently, distances in geographic CRS are not measured in metres.

The Earth's surface is typically represented in geographical coordinate systems by

a spherical or ellipsoidal surface. The former assumes that the Earth is a perfect sphere of a certain radius, which has the advantage of being a simplistic model, but is associated with inaccuracies owing to the fact that the Earth is not a sphere. Ellipsoidal models are defined by the equatorial radius and the polar radius, providing a better model since the equatorial radius is approximately 11.5 km longer than the polar radius.
The *datum* is a broader component of CRS that contains information about which ellipsoid to use and the exact relationship between Cartesian coordinates and the location on the Earth's surface. The notation *proj4string* is used to store these additional details. It allows for local variations of the Earth's surface, such as large mountain ranges, to be taken into account in local CRS. Datum can again be divided into two categories, *local* and *geocentric*, the difference being that in the local datum the ellipsoidal surface is shifted to match the surface at a particular location, whereas in the geocentric datum the centre of gravity of the Earth is the centre and the accuracy of the projections is not optimised for any particular location.

Projected Coordinate Systems

Projected CRS are based on Cartesian coordinates on an implicitly flat surface and have an origin, x and y axes, and a linear unit of measurement, metres for instance. They are based on geographic CRS and rely on map projections to convert between the three-dimensional surface of the Earth and the east/north values (x and y) in a projected CRS.
This transition always entails some distortion, skewing some of the properties of the earth's surface, such as area, direction, distance and shape. Generally, the name of a projection is based on a property it preserves, e.g. equal area projection preserves area, equidistant projection preserves distance and conformal projection preserves local shape.
Again, subgroups exist in projection coordinate systems, *conic*, *cylindrical* and *planar* projections. In a conic projection, the earth's surface is projected onto a cone along one or two tangent lines. Along these lines the distortions are minimised and increase with the distance to the lines. The projection is therefore best suited for maps of mid-latitude areas. Cylindrical projections map the surface onto a cylinder. These types of projections can be created by touching the surface of the earth along one or two tangent lines. They are often used to map the entire Earth. A planar projection projects data onto a flat surface that touches the globe at a point or along a tangent line, and is typically used in mapping polar projections.[1]

[1] Cf. Lovelace et al. 2019.

4.2 Spatial Point Processes

A stochastic process that describes the location of particular events/points that occur in a region is known as a point process. The number of points as well as the location of the points are random. An example of a point process would be the number of earthquakes and their locations.

4.2.1 Fundamentals of Point Processes

Let Z be a random, at most countable set of points in a space $\mathbb{X}$, for example $\mathbb{R}^d$. Ignoring measurability issues, Z can be thought of as a mapping $\omega \mapsto Z(\omega)$ from Ω into the set of countable subsets of $\mathbb{X}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ defines an underlying probability space. Z can then be identified with the family of mappings

$$\omega \mapsto \eta(\omega, B) := \text{card}\,(z(\omega) \cap B), \qquad B \subset \mathbb{X}, \tag{4.1}$$

which counts the number of points from Z in B. For any fixed $\omega \in \Omega$, $\eta(\omega, \cdot)$ is the counting measure supported by $Z(\omega)$.
For a general definition of a point process, let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let $N_{<\infty}(\mathbb{X}) \equiv N_{<\infty}$ be the space of all measures μ on $\mathbb{X}$ such that $\mu(B) \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}\, \forall B \in \mathcal{X}$. Let $N(\mathbb{X}) \equiv N$ be the space of all measures describable as a countable sum of measures from $N_{<\infty}$, for example the *zero measure* 0 which is equal to 0 on $\mathcal{X}$. In general, any sequence $(x_n)_{n=1}^k$ of elements of $\mathbb{X}$, where $k \in \overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ denotes the number of terms in the sequence, can be used to define a measure

$$\begin{aligned} \mu &= \sum_{n=1}^{k} \delta_{x_n}. \\ \Rightarrow \mu(B) &= \sum_{n=1}^{k} \mathbf{1}_B(x_n), \qquad B \in \mathcal{X}. \end{aligned} \tag{4.2}$$

More generally, for any measurable $f : \mathbb{X} \to [0, \infty]$,

$$\int f d\mu = \sum_{n=1}^{k} f(x_n) \tag{4.3}$$

For $k = 0$ in (4.2), μ is equal to the zero measure. The point set $\boldsymbol{x} = (x_1, ..., x_n)^T$ is said to be not pairwise different and if $x_i = x_j$ with $i \neq j$, μ is said to have multiplicities. The multiplicity of x_i is equal to the number

$$\text{card}\left\{j \leq k : x_j = x_i\right\}.$$

Any μ of the form (4.2) is interpreted as a counting measure with possible multiplicities, but in general it cannot be guaranteed that every $\mu\, inN$ can be written in this particular form.
A point process η on $\mathbb{X}$ is called *proper point process* if random elements $X_1, X_2, ...\, existin\mathbb{X}$ and a $\overline{\mathbb{N}}_0$-valued random variable κ such that almost surely

$$\eta = \sum_{n=1}^{\kappa} \delta_{X_n}. \tag{4.4}$$

For $\kappa = 0$ this is the zero measure on $\mathbb{X}$.
This terminology is motivated by the intuition that a point process is a (random) set of points, rather than an integer measure. A proper point process fits this intuition better, since it can be interpreted as a countable set of points in $\mathbb{X}$.

4.2.2 Poisson Processes

Poisson processes are defined by the fact that the number of points in a given set follows a Poisson distribution. Furthermore, the numbers of points in disjoint sets are stochastically independent.
In application, Poisson processes are used in a wide range of fields, including biology, economics and image processing.
Let λ be an s-finite measure on $\mathbb{X}$. Let an *Poisson process* with intensity measure λ be defined as a point process η on $\mathbb{X}$ with the following two properties:

1. $\forall B \in \mathcal{X} : \eta(B) \sim \text{Po}(\lambda(B); k)\, \forall k \in \mathbb{N}_0 \Longleftrightarrow \mathbb{P}(\eta(B) = k)$
2. $\forall m \in \mathbb{N}$ and all pairwise disjoint sets $B_1, ..., B_m \in \mathcal{X}$: the random variables $\eta(B_1), ..., \eta(B_m)$ are independent.

A point process satisfying the second of these conditions is called *completely independent*. If η is a Poisson process with intensity measure λ, then

$$\mathbb{E}[\eta(B)] = \lambda(B). \tag{4.5}$$

For the zero measure,

$$\mathbb{P}\left(\eta(\mathbb{X}) = 0\right) = 1,$$

with $\lambda = 0$.[2]

4.2.3 Random Measures and Cox Processes

A Poisson process with a random intensity measure, and thus the result of a *doubly stochastic* process, is called a *Cox process*. A random measure is a natural and important generalisation of a point process and it is the determining factor of the distribution of a Cox process.
Since a Cox process η can be interpreted as the result of a doubly stochastic process, a random measure ξ is generated first, followed by a Poisson process with intensity measure ξ.
These processes are often used to simulate spike trains or in financial mathematics for modeling the prices of financial instruments where credit risk is a major factor.

Random Measures

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let $M\left(\mathbb{X}\right) \equiv M$ denote the set of all s-finite measures μ on $\mathbb{X}$. Let $\mathcal{M}\left(\mathbb{X}\right) \equiv \mathcal{M}$ be the σ field generated by all sets of the form

$$\left\{\mu \in M : \mu(B) \leq t\right\}, \qquad B \in \mathcal{X}, t \in \mathbb{R}_+.$$

This is equal to the smallest σ-field of subsets of M such that $\mu \mapsto \mu(B)$ is a measurable mapping for all $B \in \mathbb{X}$.
A random measure on $\mathbb{X}$ is defined as a random element ξ of the space $(M, \mathcal{M})$, i.e. a measurable mapping $\xi : \Omega \mapsto M$.
If ξ is a random measure and $B \in \mathcal{X}$, then $\xi(B)$ denotes the random variable $\omega \mapsto \xi\left(\omega, B\right) := \xi\left(\omega\right)\left(B\right)$. This mapping represents a kernel from Ω to $\mathbb{X}$ with the additional property that the measure $\xi\left(\omega, \cdot\right)$ is s-finite for each $\omega \in \Omega$.
A random measure ξ on $\mathbb{X}$ follows the distribution of the probability measure $\mathbb{P}_\xi$ on $(M, \mathcal{M})$ given by $A \mapsto \mathbb{P}\left(\xi \in A\right)$. This distribution is again determined by the family of random vectors $\left(\xi\left(B_1\right), ..., \xi\left(B_m\right)\right)$ for pairwise disjoint $B_1, ..., B_m \in \mathcal{X}$ and $m \in \mathbb{N}$.

[2] Cf. Last and Penrose 2017.

Cox Processes

Let Π_λ denote the distribution of a Poisson process with intensity measure $\lambda \, in M\left(\mathbb{X}\right)$ and let ξ be a random measure on $\mathbb{X}$. A point process η on $\mathbb{X}$ is called a Cox process directed by ξ if

$$\mathbb{P}\left(\eta \in A | \xi\right) = \Pi_\xi(A), \qquad \mathbb{P}\text{-almost surely.}, A \in \mathcal{N}. \tag{4.6}$$

$\mathcal{N}\left(\mathbb{X}\right) \equiv \mathcal{N}$ denotes the σ field formed by the set of all subsets of N of the form

$$\left\{\mu \in N : \mu(B) = k\right\}, \qquad B \in \mathcal{X}, k \in \mathbb{N}_0.$$

Thus $\mathcal{N}$ denotes the smallest σ field on N such that $\mu \mapsto \mu(B)$ is measurable for all $B \in \mathcal{X}$.[3]

[3] Cf. Last and Penrose 2017.

4.3 Modeling and Visualising Health Data

4.3.1 Areal Data

Areal or lattice data are the result of segmenting a fixed domain into a finite number of sub-regions where results are aggregated, e.g. the number of infections with a specific disease in districts or the number of overweight people in provinces. Often the aim of disease risk models is to assess the risk within the same areas for which data are available. This can be done with a simple measure such as the *standardised incidence ratio* (SIR) or by using a Bayesian hierarchical model, which allows information to be drawn from neighbouring areas and incorporates covariates, thereby smoothing and reducing extreme values.
A widely used model is the *Besag-York-Mollié* (BYM), which takes spatial correlation and the potential for observations in neighbouring areas to be more similar than those in distant regions into account. It includes a spatial random effect that smoothes the data according to a neighbourhood structure, and an unstructured exchangeable component that models uncorrelated noise. In settings where disease numbers are monitored over time, spatio-temporal models account for temporal correlations in addition to spatial correlation, while also accounting for spatio-temporal interactions.

Spatial Neighbourhood Matrices

Spatial or proximity matrices are useful for exploratory analysis of area data. Let w_{ij} denote the (i, j) element of a *spatial neighbourhood matrix* $\boldsymbol{W}$. w_{ij} connects the two areas in some spatial way. The neighbourhood structure over the complete study region is defined by $\boldsymbol{W}$, and the elements of the matrix can be considered as weights. The closer j is to i, the more weight is associated with it. The simplest neighbourhood definition is given by the binary matrix

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{else} \end{cases} \tag{4.7}$$

Since a region cannot share a boundary with itself, $w_{ii} = 0$. Below, the number of shared borders of each canton in Switzerland are mapped.

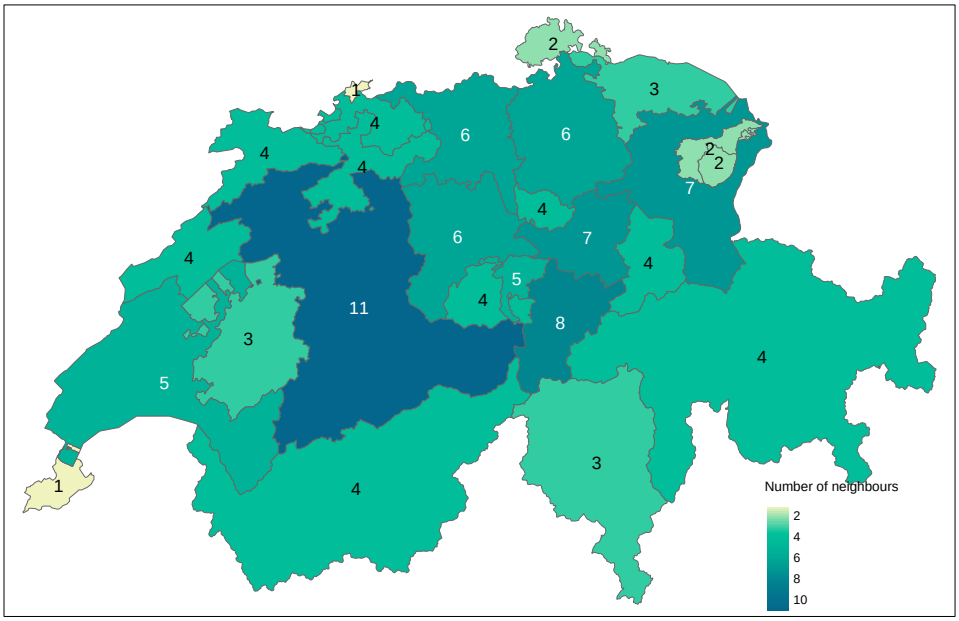

Fig. 4.4: The number of shared borders of cantons in Switzerland

Standardised Incidence Ratio

A basic measure of disease risk is the *standardised incidence ratio*, which yields an estimate in each of the areas that form a partition of the study region. It is defined as the ratio of observed counts to expected counts

$$\text{SIR}_i = \frac{Y_i}{E_i}. \tag{4.8}$$

E_i represents the sum of the expected number of cases of a given area i that behave according to the way the standard population behaves. It is calculated using indirect standardisation as

$$E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}, \tag{4.9}$$

with $r_j^{(s)}$ the rate in stratum j in the standard population and $n_j^{(i)}$ the population in stratum j of area i. If the stratum information is unavailable, the expected counts can be calculated as follows

$$E_i = r^{(s)} n^{(i)},$$

where $r^{(s)}$ denotes the rate in the standard population and $n^{(i)}$ is the population of area i. If the standardised incidence rate is greater than 1, area i has a higher risk than expected from the standard population, while for $\text{SIR}_i = 1$ the risk is the same

and for $\text{SIR}_i < 1$ it is lower than expected. The ratio is also called the standardised mortality ratio when applied to mortality data.

Spatial Small Area Disease Risk Estimation

While SIRs may prove useful in some situations, in areas with low population sizes or rare diseases, expected counts may be low, making SIRs insufficiently reliable for reporting. It is therefore preferable to assess disease risk using models that allow information to be borrowed from neighbouring areas and incorporate information from covariates, thus smoothing or shrinking extreme values due to small sample sizes.
The observed counts Y_i in area i are typically modeled with a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected counts and θ_i denotes the relative risk in area i. To account for extra Poisson reliability, the logarithm of the relative risk is expressed as the total of the intercept and the random effects. θ_i quantifies whether area i has a higher $(\theta_i > 1)$ or lower $(\theta_i < 1)$ risk than the average risk in the standard population. If the risk of an area i is half the average risk, then $\theta_i = 0.5$. The general model for spatial data is formulated as follows:

$$Y_i \sim \text{Po}\left(E_i\theta_i\right), \quad i = 1, ..., n, \tag{4.10}$$

$$\log\left(\theta_i\right) = \alpha + u_i + v_i. \tag{4.11}$$

The overall risk in the region of study is represented by α, u_i is a random effect specific to each area to model the spatial dependence between relative risks, and v_i is an unstructured exchangeable component that models uncorrelated noise, $v_i \sim \mathcal{N}\left(0, \sigma_v^2\right)$. Covariates are often included to measure risk factors and other random effects to deal with different sources of variability. For example,

$$\log\left(\theta_i\right) = \boldsymbol{d}_i\boldsymbol{\beta} + u_i + v_i,$$

with $\boldsymbol{d}_i = (1, d_{i1}, ..., d_{ip})$ a vector of the intercept and p covariates corresponding to the area i and $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$ the vector of coefficients. An increase in $\boldsymbol{d}_j$ $(j = 1, ..., p)$ by one unit, leads to an increase in the relative risk by a factor of $\exp\left(\beta_j\right)$, provided that all other covariates remain constant.
In the Besag-York-Mollié (BYM[4]) model, this spatial random effect u_i is assigned a conditional autoregressive (CAR) distribution that smoothes the data according to a

[4] Cf. Besag et al. 1991.

given neighbourhood structure that defines two areas as neighbours if they share a common boundary, specifically,

$$u_i|\boldsymbol{u}_{-i} \sim \mathcal{N}\left(\overline{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}}\right), \tag{4.12}$$

where $\overline{u}_{\delta_i}^{-1} = n_{\delta_i}^{-1} \sum_{j \in \delta_i} u_j$, while δ_i and n_{δ_i} represent the set and the amount of neighbours of the area i, respectively. The unstructured component v_i is modeled as an independent and identically distributed (i.i.d.) normal variable with zero mean and variance σ_v^2.
In 2017, Simpson et al. proposed BYM2, a new parameterisation of the BYM model that yields interpretable parameters and facilitates the assignment of meaningful penalised complexity priors. It uses a scaled, spatially structured component $\boldsymbol{u_*}$ and an unstructured component $\boldsymbol{v_*}$,

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\boldsymbol{v_*} + \sqrt{\phi}\boldsymbol{u_*}\right). \tag{4.13}$$

The precision parameter $\tau_b > 0$ controls the marginal variance contribution of the weighted sum of $\boldsymbol{u_*}$ and $\boldsymbol{v_*}$. The mixing parameter $0 \leq \phi \leq 1$ captures the proportion of the marginal variance explained by the structured effect $\boldsymbol{u}_*$. Therefore, the BYM2 model is equal to a pure spatial model for $\phi = 1$ and equal to unstructured spatial noise for $\phi = 0$. To define the prior for the marginal accuracy τ_b, the following probability statement is used:

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{\sqrt{\tau_b}} > U\right) = \alpha \\ &\Longleftrightarrow \mathbb{P}\left(\phi < U\right) = \alpha. \end{aligned} \tag{4.14}$$

Spatio-Temporal Small Area Disease Risk Estimation

When disease counts are monitored over time, spatio-temporal models are useful as they take into account not only the spatial structure but also temporal correlations and spatio-temporal interactions. Let Y_{ij} be the counts observed in area i and at time j, θ_{ij} be the relative risk, E_{ij} be the expected number of cases in area i and at time j, then

$$Y_{ij} \sim \text{Po}\left(E_{ij}, \theta_{ij}\right), \qquad i = 1, ..., I,\ j = 1, ..., J. \tag{4.15}$$

$\log\left(\theta_{ij}\right)$ is written as the sum of several components, including spatial and temporal structures, to consider that neighbouring areas and successive times may have similar risk. Spatio-temporal interactions can be included to account for the fact that

temporal trends may differ from area to area but may be more alike in neighbouring areas.
Bernardinelli et al.,[5] for example, propose a spatio-temporal model with parametric time trends that expresses the logarithm of relative risks as

$$\log(\theta_{ij}) = \alpha + u_i + v_i + (\beta + \delta_i) \times t_j. \tag{4.16}$$

The intercept is denoted by α, $u_i + v_i$ is a random area effect, β represents a global linear trend effect and δ_i is an interaction between space and time which is the difference between β and the area-specific trend. For modeling u_i and δ_i, a CAR distribution is used and v_i is i.i.d.. This specification allows each of the areas to have its individual time trend, where the spatial intercept is given by $\alpha + u_i + v_i$ and the slope by $\beta + \delta_i$. δ_i is referred to as the differential trend of the i-th area and represents the amount by which the time trend of area i deviates from the overall time trend β. If $\delta_i \neq is0$, then area i has a time trend with a slope that is either steeper or less steep than the overall time trend β.
For models that do not demand linearity of the time trend, non-parametric models such as the one proposed by Knorr-Held[6] can be used. This specific model incorporates spatial effects, temporal random effects and an interaction between space and time as follows:

$$\log(\theta_{ij}) = \alpha + u_i + v_i + \gamma_j + \phi_j + \delta_{ij}. \tag{4.17}$$

The intercept is again denoted by α, $u_i + v_i$ is a spatial random effect defined as before, i.e. u_i follows a CAR distribution and v_i is i.i.d.. $\gamma_j + \phi_j$ represents a temporal random effect and γ_j follows either a first order random walk in time (RW1)

$$\gamma_j | \gamma_{j-1} \sim \mathcal{N}\left(\gamma_{j-1}, \sigma_\gamma^2\right), \tag{4.18}$$

or second order random walk in time (RW2)

$$\gamma_j | \gamma_{j-1}, \gamma_{j-2} \sim \mathcal{N}\left(2\gamma_{j-1} - \gamma_{j-2}, \sigma_\gamma^2\right). \tag{4.19}$$

The unstructured temporal effect is given by $\phi_j, \phi_j \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\phi^2\right)$. The interaction between space and time, δ_{ij}, can be specified in a number of ways by combining the structure of the random effects that interact. The interactions proposed by Knorr-Held are those between the effects (u_i, γ_j), (u_i, ϕ_j), (v_i, γ_j) and (v_i, ϕ_j).
Using the last of these interactions leads to the assumption that there is no spatial or

[5] Cf. Bernardinelli et al. 1995.
[6] Cf. Knorr-Held 2000.

temporal structure on δ_{ij}. Thus, the interaction term can be modeled as $\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$.

Issues With Areal Data

The analysis of spatially aggregated data is subject to the "misaligned data problem" (MIDP), which arises when the data to be analysed is at a different scale from that at which it was collected. This may be solely due to the fact that the aim is to obtain the spatial distribution of a variable at a new spatial level of aggregation, e.g. if predictions are to be made at the county level with data that was originally collected at the postcode level. Another objective may be to try to find an association between variables available at different spatial scales, e.g. determining whether the risk of an unfavourable outcome provided at the country level correlates with exposure to an environmental pollutant measured at different stations, taking into account the population at risk and other demographic information available at the postcode level.

The Modifiable Area Unit Problem (MAUP) describes a problem where the inference may differ when the same underlying data are grouped at a new spatial level of aggregation. It consists of two interrelated effects, the first of which is the scale/aggregation effect. It relates to the different conclusions obtained when the same data are grouped into larger and larger areas. The other effect is the grouping/zoning effect, which accounts for the variability in results due to alternative formations of the areas, resulting in differences in area shape given the same or similar scales.

Ecological studies are defined by their reliance on aggregated data and the inherent potential for ecological fallacies. This phenomenon occurs when estimated associations obtained from the analysis of variables measured at the aggregate level lead to conclusions that differ from analyses based on the same variables measured at the individual level. This can be considered a special case of MAUP and the resulting so-called ecological bias is composed of two effects similar to the aggregation and zoning effects in MAUP. Namely, the aggregation bias caused by the aggregation of individuals and the specification bias due to the different distribution of confounding variables that results from the aggregation.[7]

[7] Cf. Moraga 2019.

4.3.2 Geostatistical Data

Geostatistical data are measurements of one or more spatially continuous features collected at specific locations. They can be a disease risk measured by a survey in different villages, the level of a pollutant recorded at several monitoring stations, or the density of mosquitoes responsible for disease transmission measured by traps set at different locations. Let $Z(s_1), ..., Z(s_n)$ be the observations of a spatial variable Z at locations $s_1, ..., s_n$. Geostatistical data are often assumed to be partial realisations of a random process

$$\left\{Z(s) : s \in D \subset \mathbb{R}^2\right\}, \tag{4.20}$$

where D denotes a fixed subset of $\mathbb{R}^2$ and the spatial index $\boldsymbol{s}$ varies continuously over D. For practical reasons, it is only possible to observe $Z(\cdot)$ at a finite set of locations. The inference of the characteristics, e.g. mean and variability of the process, of the spatial process is based on this partial realisation. Using these characteristics, it is possible to predict the process at unobserved locations and construct a spatially continuous surface of the variable of interest.

Stochastic Partial Differential Equation Approach

With geostatistical data, an underlying spatially continuous variable can often be assumed and modelled using a Gaussian random field. A spatial model can be fitted using the stochastic partial differential equation (SPDE) approach and the variable of interest can be predicted at new locations. A GRF with a Matérn covariance matrix can be written as a solution to the following continuous domain SPDE:

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau x(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}). \tag{4.21}$$

The GRF is represented by $x(\boldsymbol{s})$, where smoothness is controlled by α, while $\mathcal{W}(s)$ denotes a Gaussian spatial white noise process. $\kappa > 0$ is a scale parameter and Δ denotes the Laplacian given by $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$, where d is the dimension of the spatial domain D.
The smoothness parameter ν of the Matérn covariance function is linked to the SPDE by

$$\nu = \alpha - \frac{d}{2}$$

while the marginal variance σ^2 is related to the SPDE by

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}.$$

For $d = 2$ and $\nu = 0.5$ this corresponds to the exponential function.
The SPDE can be solved approximately using the *finite element* method, which partitions the spatial domain D into a set of non-intersecting triangles, resulting in a triangulated mesh with n vertices and n basis functions $\psi_k(\cdot)$. These functions are piecewise linear functions on each triangle, equal to 1 at vertex k and 0 otherwise. The continuously indexed Gaussian field x is thus represented as a discretely indexed Gaussian Markov random field by the finite basis functions defined on the triangulated mesh

$$x(\boldsymbol{s}) = \sum_{k=1}^{n} \psi_k(\boldsymbol{s}) x_k, \tag{4.22}$$

with n the number of vertices of the triangulate, $\psi_k(\cdot)$ the piecewise linear basis functions and $\{x_k\}$ zero-mean Gaussian distributed weights.
The joint distribution of the weight vector follows a Gaussian distribution, $\boldsymbol{x} = (x_1, ..., x_n) \sim \mathcal{N}(0, \boldsymbol{Q}^{-1}(\tau, \kappa))$, which approximates the solution $x(\boldsymbol{s})$ of the SPDE in the mesh nodes, and the basis functions transform $x(\boldsymbol{s})$ from the mesh nodes to the other spatial locations of interest.[8]

[8] Cf. Moraga 2019.

Bibliography

[Ber+95] Luisa Bernardinelli et al. “Bayesian analysis of space—time variation in disease risk”. In: *Statistics in medicine* 14.21-22 (1995), pp. 2433–2443.

[BT11] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*. Vol. 40. John Wiley & Sons, 2011.

[BYM91] Julian Besag, Jeremy York, and Annie Mollié. “Bayesian image restoration, with two applications in spatial statistics”. In: *Annals of the institute of statistical mathematics* 43.1 (1991), pp. 1–20.

[DY79] Persi Diaconis and Donald Ylvisaker. “Conjugate priors for exponential families”. In: *The Annals of statistics* (1979), pp. 269–281.

[FT13] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013.

[Kno00] Leonhard Knorr-Held. “Bayesian modelling of inseparable space-time variation in disease risk”. In: *Statistics in medicine* 19.17-18 (2000), pp. 2555–2567.

[LNM19] Robin Lovelace, Jakub Nowosad, and Jannes Muenchow. *Geocomputation with R*. CRC Press, 2019.

[LP17] Günter Last and Mathew Penrose. *Lectures on the Poisson process*. Vol. 7. Cambridge University Press, 2017.

[Mor19] Paula Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press, 2019.

[RC13] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[RMC09] Håvard Rue, Sara Martino, and Nicolas Chopin. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2 (2009), pp. 319–392.

[RS] Howard Raiffa and Robert Schlaifer. “Applied Statistical Decision Theory, Division of Research, GraduateSchool of Business Administration, Harvard University, Boston, 1961”. In: *RaiffaApplied Statistical Decision Theory1961* ().

[SR17] Sigrunn Holbek Sørbye and Håvard Rue. “Penalised complexity priors for stationary autoregressive processes”. In: *Journal of Time Series Analysis* 38.6 (2017), pp. 923–935.

List of Figures

List of Tables

57

List of Listings

58