Spatial Statistics

Autumn Semester 2013

Hansruedi Künsch
Seminar für Statistik
ETH Zürich

Version of February 3, 2014

Contents

1 Introduction **1**

2 Geostatistics **3**
2.1 Basic concepts . . . 3
2.1.1 Definitions . . . 3
2.1.2 Positive definite functions . . . 6
2.1.3 Smoothness properties of Gaussian processes . . . 8
2.1.4 Simulation of Gaussian random fields with R . . . 10
2.1.5 Convolution models . . . 13
2.2 Estimation of Gaussian process models . . . 16
2.2.1 Nonparametric estimation . . . 16
2.2.2 Parametric estimation . . . 17
2.2.3 Bayesian estimation . . . 19
2.2.4 Fitting models with R . . . 19
2.3 Kriging . . . 23
2.3.1 Simple Kriging . . . 24
2.3.2 Universal and ordinary Kriging . . . 26
2.3.3 Bayesian Kriging . . . 27

3 Models on a lattice **29**
3.1 Hierarchical models . . . 29
3.2 Spatial Markov property . . . 31
3.2.1 Gaussian Markov random fields . . . 31
3.2.2 Gibbs representation . . . 33
3.3 Inference for (latent) Markov fields . . . 37
3.3.1 The general case . . . 38
3.3.2 The Gaussian case . . . 40

4 Point patterns **45**
4.1 Basic concepts . . . 45
4.2 Moments and other characteristics . . . 47
4.3 Estimation of moments and other characteristics . . . 49
4.3.1 Estimation of the intensity . . . 50
4.3.2 Estimation of the second moment . . . 50
4.3.3 Estimation of F and G . . . 51
4.4 Models with dependence . . . 52

4.4.1 Cox point patterns 52
4.4.2 Neyman-Scott models (Cluster models) 53
4.4.3 Models with inhibition 53
4.4.4 Gibbs models 54
4.5 Estimation for parametric models 55

Chapter 1

Introduction

Spatial statistics deals with data where the location where the data were obtained is important. It is thus the analogue of time series analysis where the time is important. The most important distinction between time series and spatial statistics comes from the fact that time is linearly ordered, but space has no natural order. This makes the specification of models in spatial statistics more difficult.

Typically, in spatial statistics time is fixed to one instant or a given time interval. Space-time models combine spatial statistics with time series analysis.

The following examples show some typical research questions and data for which the methods of spatial statistics were developed.

Soil pollution At many places, the soil is polluted with metals, often due to industrial sources. An unpublished study by Andreas Papritz (Institute for Terrestrial Ecology, ETHZ) analyzes the copper content of the top 20 cm of soil at 775 points around a factory near Basel. The content varies spatially, generally decreasing with distance from the pollution source. The goal is to develop a model which describes the copper content as a function of explanatory variables and the spatial variability which remains unexplained, to identify unusual spots and to predict values at unobserved sites.

Estimating rainfall For hydrological models one needs to know as precisely as possible the rainfall intensity at all places in the catchment of a river. There are only a few stations where measurements are made, but less precise (and most likely biased) radar measurements are available in continuous space. Hence one would like to combine these two sources of data to obtain better prediction of rainfall intensity. See e.g. R. Erdin, Scientific Report MeteoSwiss No. 92, 2013.

Spatial distribution of diseases Data on the occurence of various types of cancer or other diseases are typically available by district, together with the age-standardized population of each district. The estimated cancer

rates vary considerably from district to district, and spatial patterns in the rates can give important information about possible other covariates that are missing. However, for the analysis, the random variability of estimated rates due to the Poisson nature of counts has to be taken into account. An example is in Chapter 4 of the book H. Rue and L. Held, Gaussian Markov Random Fields, Chapman & Hall, 2005, where lung cancer data from Germany are analyzed.

Agricultural field trials Crop yields depend not only on the variety and the treatement, but also on soil variability which is unobserved, but clearly has spatial structure. Examples are in J. Besag and D. Higdon, J. Royal Statist. Soc. B 61, (1999), 691–746.

Blurred images (low level vision) Here the goal is to get rid of the blur without smoothing out sharp edges and other true features of an image.

Object recognition (high level vision) In high level vision, the goal is to identify objects in an image with the goal to understand what the image shows. An example is Fleuret and Geman, Intern. J. of Computer Vision 41 (2001), 85-107.

Surfaces and textures The pore space in soil has a complicated structure which influences how water or other liquids are transported through the soil. It would be interesting to have simple stochastic models which produce realizations comparable to real soils. It would allow to study how both geometric features and transport properties change as we change parameters. See for instance, P. Lehmann et al., Adv. Water Resources 31 (2008). In two dimensions, textures are another example where geometrical structure is combined with random variations. Typical tasks are segmentation of images according to different textures or the reproduction of textures by stochastic models. See e.g. T. Hofmann et al., IEEE Transactions Pattern Analysis and Machine Intelligence, 1998, 20.

Position of trees The location of trees in a part of a forest generates a pattern of points. One expects that such a pattern differs in several aspects from a purely random distribution of points because trees compete for ressource (moisture, light, nutritions). There are statistical methods which describe how strongly and in which respect an observed point pattern differs from a pattern where points are distributed at random.

Epicenters of earthquakes Models for the space-time distribution of earthquake epicenters have been studied for instance by Ogata in a series of papers, see e.g. JASA 97 (2002), 369-380.

Chapter 2

Gaussian Processes and Geostatistics

2.1 Basic concepts

We discuss in this chapter statistical models, where observations are possible in principle at any point on $\mathbb{R}^d$ (or some open subset D of it). For this, we need the concept of a stochastic process. The simplest examples of stochastic processes are Gaussian processes which we define and discuss in some detail.

2.1.1 Definitions

Definition 2.1 *A* stochastic process *on a domain D is a collection of random variables $(Z(x); x \in D)$ indexed by D and defined on a common probability space $(\Omega, \mathcal{F}, P)$.*

Because random variables are mappings from Ω to $\mathbb{R}$, a stochastic process is therefore a mapping $Z : D \times \Omega \to \mathbb{R}$, $(x, \omega) \mapsto Z(x, \omega)$. For fixed $x \in D$, $Z(x, .)$ is a random variable, and for fixed ω, $Z(., \omega)$ is a function from D to $\mathbb{R}$. In most cases, $D \subseteq \mathbb{R}^d$, and for $d = 1$ x is usually interpreted as time whereas for $d > 1$, x has the interpretation of space or the combination of space and time. In the spatial interpretation, often the term *random field* is used instead of stochastic process.

In Bayesian nonparametric regression, stochastic processes arise if one defines a prior on the space of all regression functions: x then denotes the vector of explanatory variables, and D the domain of these explanatory variables.

Mathematically, the construction of a stochastic process is complicated and involves measure theory: One has to construct a probability space $(\Omega, \mathcal{F}, P)$ and the measurable mappings $Z(x, .)$. We skip here over these subtle points. The only thing which we will use (with a few exceptions) are the finite dimensional

distributions of the process, that is the joint distribution of $(Z(x_1), \ldots, Z(x_n))$ for any n and any values $x_1 \in D, \ldots, x_n \in D$.

The simplest class of stochastic processes are Gaussian processes

Definition 2.2 *A stochastic process is called* Gaussian*, if all finite-dimensional distributions are Gaussian.*

Because Gaussian distributions are determined by their first two moments, a Gaussian process is determined by its mean function $m : D \to \mathbb{R}, x \mapsto m(x) = \mathbb{E}(Z(x))$ and its covariance function $C : D \times D \to \mathbb{R}, (x, x') \mapsto C(x, x') = \mathrm{Cov}\,(Z(x), Z(x'))$. Whereas there are no restrictions on the mean function m, the covariance function must be symmetric $C(x, x') = C(x', x)$ and *positive definite* in the following sense:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} C(x_i, x_j)\beta_i\beta_j \geq 0 \quad (n \in \mathbb{N}, x_i \in D, \beta_i \in \mathbb{R}).$$

In time series and in spatial statistics, we usually have observations from only one realization of the underlying random function. (Situations where one has several independent realizations are typically discussed under functional data analysis). It is clear that from one realization it is not possible to find out something about the underlying finite dimensional distributions without additional assumptions. We need some redundancy for doing statistics. The simplest assumption which provides redundancy in the case $D = \mathbb{R}^d$ is *stationarity*.

Definition 2.3 *A stochastic process on $\mathbb{R}^d$ is called* stationary*, if the finite-dimensional distributions do not change when the points x_i are shifted by a common vector h, that is $(Z(x_1), \ldots, Z(x_n))$ and $(Z(x_1 + h), \ldots, Z(x_n + h))$ have the same distribution.*

It is easy to see that a Gaussian stochastic process on $\mathbb{R}^d$ is stationary iff the mean function is constant: $m(x) \equiv m$ and the covariance function depends only on the relative position $C(x, x') = C(x - x')$. Such invariance of the first two moments is called weak stationarity in the case of non-Gaussian processes.

The simplest form of non-stationarity has a non-constant mean whereas the covariance is still stationary

$$Z(x) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(x) + \tilde{Z}(x)$$

where $f_1, \ldots, f_p$ are some known functions (e.g. longitude or latitude or altitude), the β_j are parameters which can be used to fit the model to data and $\tilde{Z}$ is a stationary Gaussian process with mean zero.

The first step to relax the stationarity condition for the covariances is given by the concept of an intrinsic process where only increments are assumed to be stationary:

Definition 2.4 *A Gaussian random process on $\mathbb{R}^d$ is called* intrinsic *if the mean function is constant, that is $m(x) \equiv m$, and if $\mathrm{Var}\,(Z(x) - Z(x'))$ depends only on the difference $x - x'$. For an intrinsic process, we call $\gamma(h) = \frac{1}{2}\mathrm{Var}\,(Z(x+h) - Z(x))$ the semivariogram.*

A stationary Gaussian process is also intrinsic, and the semivariogram can be expressed with the covariance function

$$\gamma(h) = \frac{1}{2}(C(x+h, x+h) - 2C(x+h, x) + C(x, x)) = C(0) - C(h).$$

Processes which are intrinsic, but not stationary, are a generalization of ARIMA processes in time series analysis. The simplest example of a process on $\mathbb{R}$ which is intrinsic, but not stationary is Brownian motion. It has mean zero and the following covariance

$$C(x, x') = \begin{cases} \min(|x|, |x'|) & (xx' \geq 0) \\ 0 & (xx' \leq 0) \end{cases}$$

The semivariogram is then $\gamma(h) = \frac{1}{2}|h|$, and non-overlapping increments are independent.

The next simple lemma will be used repeatedly in the following

Lemma 2.1 *a) If $(Z(x))$ is a stationary process, then*

$$\mathrm{Cov}\left(\sum_{j=1}^{n} \beta_j Z(x_j), \sum_{j=1}^{n} \beta'_j Z(x_j)\right) = \sum_{j,k=1}^{n} \beta_j \beta'_k C(x_k - x_j). \tag{2.1}$$

for any $n \in \mathbb{N}$, any $x_i \in \mathbb{R}^d$ and any $\beta_i, \beta'_i \in \mathbb{R}$

b) If $(Z(x))$ is an intrinsic process, then

$$\mathrm{Cov}\left(\sum_{j=1}^{n} \lambda_j Z(x_j), \sum_{j=1}^{n} \lambda'_j Z(x_j)\right) = -\sum_{j \neq k} \lambda_j \lambda'_k \gamma(x_k - x_j) \tag{2.2}$$

for any $n \in \mathbb{N}$, any $x_i \in \mathbb{R}^d$ and any $\lambda_i, \lambda'_i \in \mathbb{R}$ such that $\sum_j \lambda_j = \sum_j \lambda'_j = 0$.

Proof:

Claim a) follows from basic properties of covariances. For claim b) we use

$$\sum_{j=1}^{n} \lambda_j Z(x_j) \;=\; \sum_{j=1}^{n} \lambda_j (Z(x_j) - Z(x_1))$$

and analogously for the second linear combination. Therefore

$$\mathrm{Cov}\left(\sum \lambda_j Z(x_j), \sum \lambda'_j Z(x_j)\right) \;=\; \sum_{j,k} \lambda_j \lambda'_k \,\mathrm{Cov}\,(Z(x_j) - Z(x_1), Z(x_k) - Z(x_1))\,.$$

In particular, if $\lambda_j = \lambda'_j = 1$, $\lambda_k = \lambda'_k = -1$ and all other values are zero, then

$$2\gamma(x_j - x_k) \quad = \quad 2\gamma(x_j - x_1) + 2\gamma(x_k - x_1) - 2\,\mathrm{Cov}\left(Z(x_j) - Z(x_1), Z(x_k) - Z(x_1)\right).$$

We thus can express the covariance of two increments with the help of the semivariogram. Plugging this into the equality above, we obtain

$$\begin{aligned}\mathrm{Cov}\left(\sum \lambda_j Z(x_j), \sum \lambda'_j Z(x_j)\right) &= \sum_{j,k} \lambda_j \lambda'_k (\gamma(x_j - x_1) + \gamma(x_k - x_1) - \gamma(x_j - x_k)) \\ &= 0 + 0 - \sum_{j \neq k} \lambda_j \lambda'_k \gamma(x_k - x_j).\end{aligned}$$

□

By the second part of this lemma, a semivariogram must be "conditionally negative definite" (conditionally because the inequality only holds for coefficients λ_j which sum to zero).

Besides invariance under translations, one can also consider invariance under rotations:

Definition 2.5 *A stationary or intrinsic Gaussian process on $\mathbb{R}^d$ is called* isotropic *if $C(h) = C(\|h\|)$ or $\gamma(h) = \gamma(\|h\|)$, respectively.*

For an isotropic process, the dependence is the same in all directions. If Z is an isotropic stationary process and B is any invertible $d \times d$ matrix which is not a multiple of the identity, then the process $\tilde{Z}(x) = Z(Bx)$ is again stationary but no longer isotropic:

$$\mathbb{E}(\tilde{Z}(x)) = m, \quad \mathrm{Cov}\left(\tilde{Z}(x+h), \tilde{Z}(x)\right) = C(\|B(x+h) - Bx\|) = C(\|Bh\|).$$

Therefore the shape of the covariance function is the same in all directions, but the scaling of the distance depends on the direction. An example is given in Section 2.1.4 below.

2.1.2 Positive definite functions

We start with a list of the most important parametric models for a stationary covariance functions or a semivariogram.

"White noise", "nugget-model" This model assumes that values at different positions are independent, no matter how close they are:

$$C(h) = \begin{cases} \sigma^2, & \text{if } h = 0 \\ 0, & \text{if } h \neq 0 \end{cases} \tag{2.3}$$

"General exponential model"

$$C(h) = \sigma^2 \cdot \exp\left(-\left(\frac{\|h\|}{\rho}\right)^{\nu}\right). \tag{2.4}$$

For positive definiteness we must have $0 < \nu \leq 2$. The two most popular models have $\nu = 1$ and $\nu = 2$, respectively, and are known as 'exponential model" and "Gaussian model".

"Spherical model"

$$C(h) = \sigma^2 \cdot \frac{|B(0,1) \cap B(h/\rho, 1)|}{|B(0,1)|} \tag{2.5}$$

where $B(x,r)$ is the sphere in $\mathbb{R}^d$ with center x and radius r, and the absolute value of a set in $\mathbb{R}^d$ denotes the volume (or the area in case $d = 2$) of this set.

"Matérn-Modell"

$$C(h) = \sigma^2 \cdot \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|h\|}{\rho}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{\|h\|}{\rho}\right), \tag{2.6}$$

where $K_\nu(\cdot)$ is the so-called modified Bessel function of order ν and $\Gamma(\cdot)$ the Gamma function. For $\nu = 1/2$, we obtain $C(h) = \sigma^2 \exp(-||h||/\rho)$, i.e. the exponential model, and for $\nu \to \infty$ we obtain $C(h) = \sigma^2 \exp(-||h||^2/(2\rho^2))$, i.e. the Gaussian model. There are different parametrizations of this model in the literature. The reason for not absorbing the factor $\sqrt{2\nu}$ in ρ is that models with different values of ν differ mainly with respect to their smoothness near zero whereas the distance where the covariance is close to zero is similar for all ν.

The modified Bessel function seems complicated, see Abramowitz and Stegun, Chapter 9.6-9.8, but it is coded in *R*. Therefore we need not to know much about it for applying the model. In Section 2.1.4 we show how to evaluate the covariance function and how to generate a plot for different shape parameters ν.

"Power model"

$$\gamma(h) = \sigma^2 \cdot \|h\|^\nu \tag{2.7}$$

where $0 < \nu < 2$. For $\nu = 1$ and $d = 1$, it is the semivariogram of Brownian motion.

The unknown parameters are σ^2, ρ and ν. The parameter σ scales the observations Z ($Z \to const \cdot Z$), ρ scales the distance between points ($x \to const \cdot x$) and ν is a shape parameter which decides how smooth C or γ are at the origin.

Proving that the functions in the list above are positive definite is complicated. A powerful tool for doing this is Fourier analysis. A function $C : \mathbb{R}^d \to \mathbb{R}$ which has the representation

$$C(h) = \int_{\mathbb{R}^d} \cos(\lambda^T h) s(\lambda) d\lambda$$

where $s : \mathbb{R}^d \to \mathbb{R}_+$ is integrable is always positive definite, and this condition is close to being necessary: If C is a stationary covariance function which is continuous at the origin and satisfies

$$\int_{\mathbb{R}^d} |C(h)|dh < \infty$$

(that is the covariance decays to zero fast enough as the distance increases), then

$$s(\lambda) = (2\pi)^{-d} \int C(h) \cos(\lambda^T h)dh \geq 0$$

for all $\lambda \in \mathbb{R}^d$ and

$$C(h) = \int_{\mathbb{R}^d} \cos(\lambda^T h)s(\lambda)d\lambda.$$

The function s is called the spectral density of the covariance function. For $d = 1$, these results are usually discussed in time series analysis.

In this course, we do not use spectral methods much, so I will not go into this topic in more detail. But I would like to mention that the spectral density of the Matérn model is

$$s(\lambda) = \sigma^2 \text{const.}(\nu, \rho) \left(2\nu/\rho^2 + \lambda^2\right)^{-\nu - d/2}.$$

Since this is much simpler than the covariance function, it adds plausibility to the claim that the Matérn model is a very fundamental class of models.

2.1.3 Smoothness properties of Gaussian processes

How smooth the realization of an intrinsic Gaussian process are, depends on the smoothness of the semivariogram at the origin. In order to make precise statements, we recall different convergence concepts of random variables first. A sequence of random variables $X_1, X_2, \ldots$ (on a common probability space) converges to X in probability or stochastically if for all $\epsilon > 0$

$$P(|X_n - X| \geq \epsilon) \to 0 \quad (n \to \infty).$$

By Chebyshev's inequality, convergence of probability holds if $E((X_n - X)^2) \to 0$ which is called mean-square convergence or L_2-convergence. There is a stronger convergence concept, almost sure convergence, where $X_n(\omega) \to X(\omega)$ for all values ω from a set with probability one.

A stochastic process on $\mathbb{R}^d$ is called *mean-square continuous* if $Z(x + h_n)$ converges in mean-square to $Z(x)$ for any x and any sequence $h_n \to 0$. An intrinsic process is obviously mean-square continuous iff γ is continuous at the origin. It can be shown that γ is then continous everywhere. Unfortunately, realizations $Z(., \omega)$ of a mean-square continuous process are not necessarily continuous functions with probability one (convergence in probability does not imply almost sure convergence, and even if realizations are almost surely continuous for any fixed x, they can still be discontinuous for some x because the exceptional set

may depend on x). A Gaussian intrinsic process has almost surely continuous realisations if for some $\delta > 0$

$$\gamma(h) = O(|\log(||h||)|^{1+\delta}) \quad (h \to 0),$$

see e.g. Theorem 1.4.4 in Adler and Taylor (2007). This covers all continuous semivariograms which are used in practice.

The semivariogram of the white noise model is not continuous at the origin. In fact, the realizations of the corresponding process are so irregular that the Riemann integral over any compact domain $D \subset \mathbb{R}^d$ has a constant value: If we partition D as $D = \cup V_i$ and approximate the integral $\int_D Z(x)dx$ by a Riemann sum $\sum Z(x_i)|V_i|$ where $x_i \in V_i$, then this approximation converges to $m = E(Z(x))$ in mean square as $\sup_i |V_i| \to \infty$. However for any fixed x, $Z(x) \neq m$. Hence, white noise is not a reasonable model if taken literally. Usually it is considered as an approximation of a process whose semivariogram is continuous, but is constant for $||h|| \geq h_0$ where h_0 is much smaller than the distance between any pair of points where the process is observed.

Mathematically, there is a rigorous theory for a slightly different white noise model. For this model, heuristically $\sigma^2 = \infty$ and thus the values at any point x are not defined. Only the integrals $\int Z(x)f(x)dx$ exist for any $f : \mathbb{R}^d \to \mathbb{R}$ with $\int f(x)^2 dx < \infty$, and the random variables $(\int Z(x)f_1(x)dx, \ldots, \int Z(x)f_n(x)dx)$ have a multivariate Gaussian distribution with means $m \int f_i(x)dx$ and covariances $\sigma^2 \int f_i(x)f_j(x)dx$. We will use this version of white noise at a few places later.

Higher order smoothness of γ is related to higher smoothness of the process. We say that a stochastic process is differentiable in quadratic mean if for any $x \in \mathbb{R}^d$ and any direction $h \in \mathbb{R}^d$

$$D_h Z(x) = \lim_{t \to 0} \frac{Z(x + th) - Z(x)}{t}$$

exists in the mean-square sense. An intrinsic process is mean-square differentiable if the semivariogram is twice differentiable at the origin. Moreover the semi variogram is then everywhere twice differentiable and we have

$$\mathrm{Cov}(D_h Z(x), D_h Z(x')) = h^T D^2 \gamma(x - x')h$$

where $D^2\gamma$ is the matrix of partial second derivatives of the semi variogram. For Gaussian processes, a slightly stronger condition implies continuous differentiability of almost all realizations $Z(., \omega)$. This can be extended to conditions for the existence of higher derivatives. There is also a theory of smoothness of fractional order of the process which is related to smoothness of the semivariogram of twice the same order. We do not go into the details.

The main conclusion of this section is that in the Matérn model, the realizations will be almost surely n times continuously differentiable if $\nu > n$. The Gaussian covariance function is obtained as the limit $\nu \to \infty$, and its realizations are almost surely infinitely often differentiable. The possibility to have

the whole range of differentiability propoerties for the realizations makes the Matérn model suitable for many applications. The next Section shows simulated realizations of the Matérn model for different values of the shape parameter.

2.1.4 Simulation of Gaussian random fields with R

In this section we begin to illustrate the use of *R* for spatial statistics. We will use the packages `RandomFields`, `geoR` and `spatstat` in this course.

The package `RandomFields` has the largest list of parametric models for the covariance function and the most advanced methods for simulation. The package `geoR` offers more choices for the analysis of spatial data.

We begin by showing how one can compute covariance functions with `RandomFields` for some specified models.

```
## Evaluation of covariance functions with the package RandomFields:
library(RandomFields)
C <- CovarianceFct(x= 1,model="matern",
                   param=c(mean=0, variance=1,
                           nugget=0, scale=1, nu=0.5))
```

In figure 2.1, the covariance function for the Matérn model with various smoothness parameter and two different intervals for the distance is plotted.

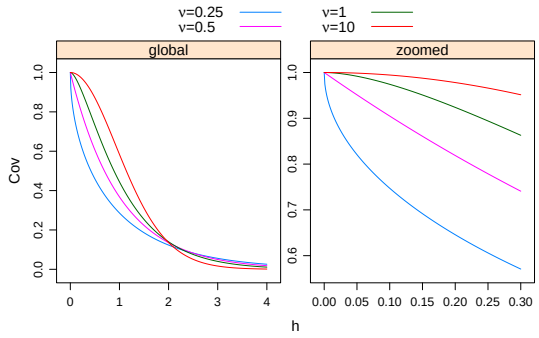

Figure 2.1: Covariance function of the Matérn model with various smoothness parameters.

Next we demonstrate how to simulate a Gaussian random field with the Matérn model and different shape parameters.

First we load the package `RandomFields`, `lattice` for plotting and `foreach` for convenient wrapping of simulation:

```
## Simulation of Gaussian random field with a given
## covariance function with the package RandomFields:
```

```
library(RandomFields)
library(lattice)
library(foreach) #not compulsory but useful to combine results

#PrintModelList() ## the complete list of implemented models
```

Then we create a grid on the 4x4 square with 10,000 points on which to simulate:

```
nside <- 100 ## nside * nside = number of cells in the grid
grid <- expand.grid(seq(0, 4, length.out=nside),
                    seq(0, 4, length.out=nside))
x <- grid[,1]
y <- grid[,2]
```

Set the parameters of the model:

```
cov.model <- "matern"
mean <- 0 # no trend
variance <- 1 # sigma^2
nugget <- 0 # no nugget added
scale <- 1
nu.vec <- c(0.25, 0.5, 1, 10) # control the smoothness
```

We then simulate a realization of the field for each of the smoothness parameters.

```
## Simulate:
set.seed(13) ## for reprocudibility
## foreach loops through values of i and combines each
## output with .combine, here rbind
sim <- foreach (i = 1:length(nu.vec), .combine="rbind") %do% {
   nu <- nu.vec[i]
   z <- GaussRF(x=x, y=y, model=cov.model,
                param=c(mean, variance, nugget, scale, nu))
   data.frame(z, x, y, nu)
 }
```

Finally we format the data and plot them nicely with lattice (Figure 2.2)

```
## for nice labels:
sim$nu <- factor(sim$nu)
facnames <- c(expression(paste(nu, "=0.25")),
              expression(paste(nu, "=0.5 (Exponential)")),
              expression(paste(nu, "=1")),
              expression(paste(nu, "=10 (->Gaussian)")))
# Conditional plot:
```

```
levelplot(z ~ x*y | nu, data=sim,
          col.regions = terrain.colors(100),
          asp=1, index.cond=list(c(3,4,1,2)), #to change order
          strip=strip.custom(factor.levels=facnames))
```

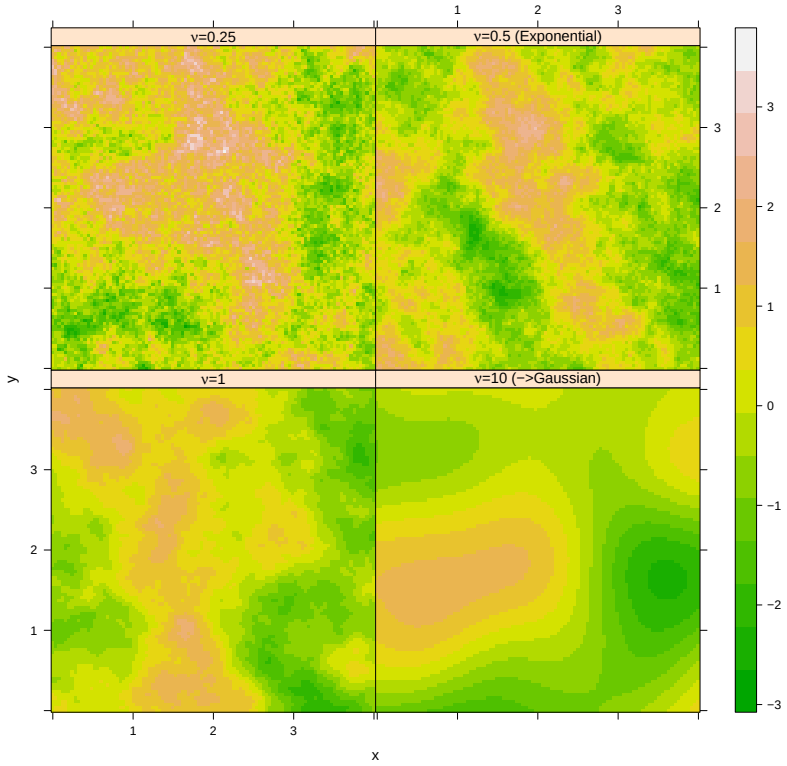

Figure 2.2: Realization of random fields with the Matérn model and various smoothness parameter.

Isotropy vs. anisotropy Often spatial patterns occur in an asymmetric way for various reason (e.g. because of the orientation of the topography in Switzerland). Here we demonstrate by simulation the qualitative difference between an isotropic and an anisotropic random field.

First we load the package `geoR` and set the parameters:

```
library(geoR)
cov.model <- "matern"
mean <- 0 # no trend
variance <- 1 # sigma^2
nugget <- 0 # no nugget added
nu <- .5
rho <- .2/sqrt(2*nu)
aniso <- c(pi/4, 3) # (angle, amplitude)
```

Then we proceed to the simulation on a regular grid of 10,000 points on the unit square:

```
set.seed(12)
## Simulation of isotropic field:
## grid is 100x100 on unit square
sim.iso <- grf(nx=100, ny=100, grid="reg",
               cov.model="matern",
               cov.pars = c(variance, scale),
               nugget=nugget,
               kappa=nu)
## Simulation of anisotropic field:
## grid is 100x100 on unit square
sim.anis <- grf(nx=100, ny=100, grid="reg",
                cov.model="matern",
                cov.pars = c(variance, scale),
                nugget=nugget,
                kappa=nu,
                aniso.pars= aniso) # anistropy
```

Finally we put the simulations together and plot them with lattice (Figure 2.3)

```
## put together for plotting:
output1 <- data.frame(z= sim.iso$data, x, y, iso="isotropic")
output2 <- data.frame(z= sim.anis$data, x, y, iso="anisotropic")
data.iso <- rbind(output1, output2)
## Conditional plot:
levelplot(z ~ x*y | iso, data=data.iso,
          col.regions = terrain.colors(100),
          index.cond=list(c(1,2)), ## to change order...
          xlab='', ylab='', useRaster=T, asp=1)
```

2.1.5 Convolution models

This is an alternative way to define a Gaussian process directly. It uses a kernel K and a Gaussian white noise process ξ (the version with infinite variance) and sets

$$Z(x) = \int_{\mathbb{R}^d} K(x - x')\xi(x')dx'.$$

This is a generalization of the moving average process to continuous time and higher dimensions.

By the formulae in the previous subsection,

$$E(Z(x)) = m_\xi \int K(x)dx, \quad \mathrm{Cov}(Z(x+h), Z(x)) = \sigma_\xi^2 \int K(x'+h)K(x')dx'.$$

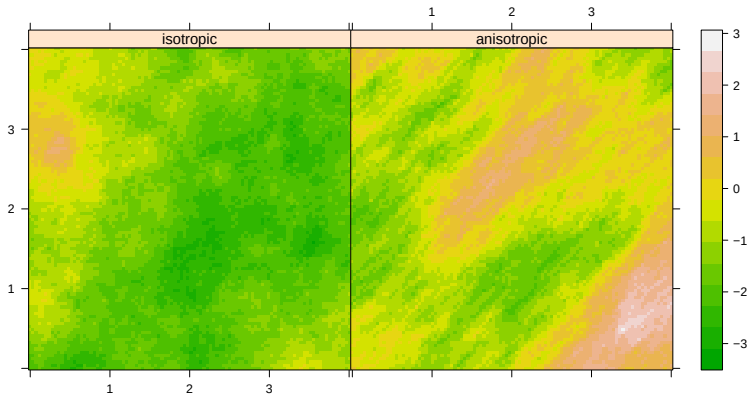

Figure 2.3: An isotropic and an anisotropic simulated random field.

Hence the model is stationary. In particular, if we take as K the indicator function of a sphere centered at the origin, we see that the spherical model belongs to the class of convolution models.

For simulation and for other purposes, one has to approximate the integral by a Riemann sum. Doing this, one obtains the model

$$Z(x) = \sum_{j=1}^{\infty} K(x - x_j)\xi_j.$$

where the ξ_j are independent $\mathcal{N}(\mu_j, \sigma_j^2)$ variables. This means that the process is a linear combination of a fixed function K shifted by deterministic vectors x_j and scaled by random amplitudes ξ_j. This discrete convolution model is usually no longer stationary. It will be approximately stationary if the centers x_i form a regular grid and m_j and σ_j^2 are constant.

One can also move further away from stationarity and replace $K(x - x_j)$ by arbitrary basis functions $K_j(x)$ (subject to some condition which guarantees that the sum is well defined). Such a representation holds for many Gaussian processes.

Theorem 2.1 *Let Z be a Gaussian mean zero process on a compact set $T \subset \mathbb{R}^d$ with continuous covariance function $C(x, x')$. Then there exists a sequence of orthnormal continuous eigenfunctions K_j, and eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, $\sum_j \lambda_j^2 < \infty$ with*

$$\begin{aligned} \int_T C(x, x')K_j(x')dx' &= \lambda_j K_j(x), \\ C(x, x') &= \sum_{j=1}^{\infty} \lambda_j K_j(x)K_j(x'). \end{aligned}$$

Moreover, if we set $\xi_j = \int_T Z(x)K_j(x)dx$*, then the* ξ_j*'s are independent standard normal random variables, and*

$$Z(x) = \sum_{j=1}^{\infty} K_j(x)\sqrt{\lambda_j}\xi_j$$

where the series converges in quadratic mean.

The first part is Mercer's theorem, and the second part is called the Karhunen-Loéve expansion. The results are generalizations of the diagonalisation of symmetric matrices with orthogonal matrices in linear algebra and principle component analysis in statistics. Explicit computation of the eigenfunctions and eigenvalues is however possible only in exceptional cases.

One such exception is Brownian motion restricted to $D = [0, 1]$. Then the eigenfunctions and eigenvalues are

$$K_j(x) = \sqrt{2}\sin((j - 0.5)\pi x), \quad \lambda_j = \frac{1}{\pi^2(j - 0.5)^2} \quad (j = 1, 2, \ldots).$$

We show the approximation for different truncations of the infinite sum in Figure 2.4.

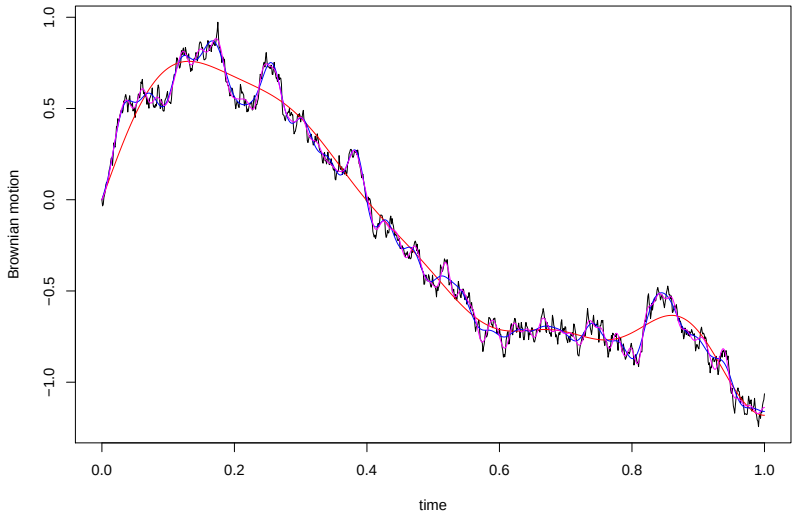

Figure 2.4: Karhunen-Loeve expansion of Brownian motion with 11 (red), 51 (blue), 101 (purple) and 1001 (black) terms

```
psi <- function(i,t) sqrt(2)*sin((i-0.5)*pi*t) #Eigenfunctions K_j
time <- seq(0,1,length=1001)
n <- 1001
set.seed(14)
xi <- rnorm(n) #random coefficients
lambda <- 1/(pi*((1:n)-0.5)) #square root of eigenvalues
```

```
bm <- (xi*lambda)%*%outer(1:n,time,FUN="psi") #Approximation with n terms
plot(time,bm,type="l",ylab="Brownian motion")
bm.11 <- (xi[1:11]*lambda[1:11])%*%outer(1:11,time,FUN="psi")
                  #Approximation with 11 terms
lines(time,bm.11,col=2)
bm.51 <- (xi[1:51]*lambda[1:51])%*%outer(1:51,time,FUN="psi")
                  #Approximation with 51 terms
lines(time,bm.51,col=4)
bm.101 <- (xi[1:101]*lambda[1:101])%*%outer(1:101,time,FUN="psi")
                  #Approximation with 101 terms
lines(time,bm.101,col=6)
```

2.2 Estimation of Gaussian process models

2.2.1 Nonparametric estimation

Here we consider estimation of the semivariogram γ of an intrinsic process, based on observations $(Z(x_1), \ldots, Z(x_n))$. The assumption of Gaussianity is not needed.

The idea is simple: If the semivariogram is isotropic, we average squared increments of the process for pairs of observation points which have approximately the same distance:

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{||x_i - x_j|| \approx h} (Z(x_i) - Z(x_j))^2$$

where $N(h)$ denotes the number of terms in the sum on the right-hand side. The simplest way to implement this, is to partition the pairwise distances $x_i - x_j$ into mutually disjoint classes like for a histogram and then to estimate the semivariogram at each class center by averaging over the class. A slighlty more sophisticated method uses a kernel estimator:

$$\widehat{\gamma}(h) \quad = \quad \frac{\sum_{i,j} (Z(x_i) - Z(x_j))^2 \cdot K(\|x_i - x_j\| - \|h\|)}{2 \sum_{i,j} K(\|x_i - x_j\| - h)}.$$

If one does not want to assume isotropy, then one has to consider pairs of points with approximately the same distance and orientation.

If $\widehat{\gamma}(h)$ does not reach a sill for large h, one concludes that the observations do not come from a stationary process. If a sill is reached, then the stationarity assumption is accepted, the sill is considered as an estimate of $C(0)$ and one estimates the autocovariance as $\widehat{C}(0) - \widehat{\gamma}(h)$. As an alternative one can also proceed like in time series and estimate first the mean m by the average of all observations and the autocovariance by

$$\widehat{C}(h) = \frac{1}{N(h)} \sum (Z(x_i) - \widehat{m})(Z(x_j) - \widehat{m}).$$

The method based on the semivariogram is usually prefered because the effect of subtracting $\widehat{m}$ instead of the true mean is in some cases not negligible.

These methods are simple and give some basic information about the dependence in the data, but the statistical errors can be large and difficult to quantify and the fitted semivariogram need not be a valid model (that is it is not guaranteed to be conditionally negative definite).

2.2.2 Parametric estimation

We consider here the following model

$$Z(x) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(x) + Y(x) + \varepsilon(x)$$

where $f_1, \ldots, f_p$ are known functions, $(Y(x))$ is a mean zero Gaussian stationary process with covariance function

$$C(h) = \sigma^2 r(\|h\| / \rho; \nu)$$

and $(\varepsilon(x))$ is a nugget model with variance τ^2. The unknown parameters are $\beta = (\beta_0, \ldots, \beta_p)$ and $\theta = (\tau^2, \sigma^2, \rho, \nu)$. We want to estimate these parameters based on observations $Z(x_i)$ for $i = 1, 2, \ldots, n$. Written in vector/matrix form the model for the observations is

$$\mathbf{Z} = F_n \beta + \mathbf{Y} + \varepsilon$$

where F_n is the $n \times (p+1)$ matrix of the fixed effects and the errors $\mathbf{Y} + \varepsilon$ have covariance matrix $K_n(\theta) = C_n(\theta) + \tau^2 I_n$. (The matrix C_n has elements $C(||x_i - x_j||)$).

The simplest method is based on the nonparametric estimate of the covariance or the semivariogram (If $p > 0$, we estimate β by ordinary least-squares and use residuals). The parameters of the semivariogram are then estimated by matching the nonparametric semivariogram, either "by eye" or by minimizing a weighted sum of squared deviations:

$$\sum_{j=1}^{k} w_j \cdot \left(\widehat{\gamma}(h_j) - \tau^2 - \sigma^2(1 - r(h_j/\rho; \nu))\right)^2$$

The weights w_j are for instance the number of pairs used to estimate $\widehat{\gamma}(h_j)$.

A more systematic method is Maximum-Likelihood (ML). By the Gaussian assumption for (Y) and (ε), the log likelihood is given by

$$2 \log L_n(\beta, \theta) = -\log \det(K_n(\theta)) - (\mathbf{z} - F_n\beta)^T K_n(\theta)^{-1} (\mathbf{z} - F_n\beta).$$

The parameters $\widehat{\beta}(\theta)$ which maximize the likelihood for fixed value of θ are the so-called generalized least squares (GLS) estimators

$$\widehat{\beta}(\theta) = (F_n^T K_n(\theta)^{-1} F_n)^{-1} F_n^T K_n(\theta)^{-1} \mathbf{z}.$$

If we plug this into the formula for the likelihood we obtain the so-called profile likelihood which is a function of θ alone.

Although we have an explicit expression for the likelihood, finding the MLE can be difficult for several reasons. First, the computational effort to compute the inverse or the determinant of K_n can be huge because it is of the order $O(n^3)$ (even if we solve linear equations instead of inverting K_n). Second, there is the possibility of multiple local maxima. Third, the presence of regression parameters β_j can bias the estimated θ substantially. This can be seen already in ordinary regression where the MLE for the error variance is biased, and with correlated errors this effect is typically more pronounced.

Several methods have been proposed to deal with the first problem. One is the use of reduced rank models where C_n is replaced by a matrix of rank $r \ll n$:

$$C_n = K_n D_r K_n^T, \quad K_n \in \mathbb{R}^{n\times r}, \quad D_r = \mathrm{diag}(d_1^2, \ldots, d_r^2).$$

This means that the original model is replaced by a convolution model: $Z(x_i) = \sum_{j=1}^r K_{ij} d_j \xi_j$ with ξ_j i.i.d. standard normal. There are formulae in linear algebra which allow one to compute the inverse or the determinant of a matrix which is the identity plus a low rank perturbation (Sherman-Morrison-Woodbury formula and matrix determinant lemma). If the model for C is already in convolution form, a reduced rank is achieved by truncating the sum. We will see later a different approach that can be used for any model and is based on kriging. Another method to avoid the computation of inverses and determinants of large matrices is to replace the likelihood by the so-called pseudo-likelihood. For this, one chooses any number of small size subsets $V_i \subset \{x_1, \ldots, x_n\}$, usually overlapping, and maximizes

$$\prod_i L_{V_i}(\beta, \theta)$$

where L_{V_i} is the likelihood based on $(z(x_j); x_j \in V_i)$. A third method is based on sparse matrix techniques and will be discussed in Chapter 3 when we discuss lattice models.

The second problem is not specific to spatial statistics. It emphasizes the importance of having good starting values. The third problem can be avoided by the use of so-called Restricted Maximum Likelihood Estimation (REML). The idea is to transform the observations $\mathbf{Z}$ linearly to a $n-p-1$-dimensional vector $\mathbf{Z}' = A\mathbf{Z}$ which has mean zero for any value of β and to use Maximum Likelihood for $\mathbf{Z}'$. One can show that the resulting estimator does not depend on the choice of A (as long as A has rank $n-p-1$) and is obtained by minimizing

$$\log\det(K_n(\theta)) + \log\det(F_n^T K_n(\theta)^{-1} F_n) + (\mathbf{z} - F_n\widehat{\beta}(\theta))^T K_n(\theta)^{-1}(\mathbf{z} - F_n\widehat{\beta}(\theta)).$$

Hence the difference to the profile likelihood is the additional term $\log\det(F_n^T K_n(\theta)^{-1} F_n)$.

2.2.3 Bayesian estimation

In Bayesian statistics, the unknown parameters β and θ are assigned prior distributions $p(\beta)p(\theta)$ and inference is based on the posterior distribution

$$p(\beta, \theta \mid \mathbf{z}) \propto L_n(\beta, \theta)p(\beta)p(\theta).$$

Exploration of this posterior distribution is done by Markov chain Monte Carlo (MCMC) sampling: A dependent sample $(\beta^{(j)}, \theta^{(j)}; j = 1, 2, \ldots, N)$ which is approximately distributed according to the posterior is generated. With this sample it is then straightforward to compute point estimators as means or medians of each component or credible intervals via a pair of symmetric quantiles.

The construction of good Markov Chain Monte Carlo estimators is however not trivial, and we do not discuss this problem here.

2.2.4 Fitting models with R

We show variogram fitting using the package `gstat`. First we simulate 200 points data at random locations with a Matérn model ($\nu = 0.25$).

```
npoints <- 200
## Set parameters of the model:
cov.model <- "matern"; mean <- 0 ;variance <- 1; nugget <- 0
scale <- 1; nu <- 0.25
set.seed(13) ## for reprocudibility
sim <- grf(n=npoints, xlims=c(0,10), ylim=c(0,10),
     cov.model="matern",
     cov.pars = c(variance, scale),
     nugget=nugget,
     kappa=nu)
```

Nonparametric estimation Then we can use the function `variog` to compute the empirical variogram of the simulated data. The option *cloud* calculate the raw value for each pair versus their distance. In a second time the distance is binned. The last way is to calculate only the mean value at binned distances. Another option is to give a smooth estimate of the variogram.

In figure 2.5, first the simulated data points are plotted, then the variogram cloud, then the boxplot version. In the last plot, the mean values at binned distance are plotted together with the smooth estimate in blue and the theoretical variogram (known because the data are simulated) in dashed black.

```
library(geoR)
# variogram cloud
vario.c <- variog(sim, max.dist=3, op="cloud")
#binned variogram with storing the cloud
```

```
vario.bc <- variog(sim, max.dist=3, bin.cloud=TRUE)
# binned variogram
vario.b <- variog(sim, max.dist=3)
# smoothed variogram
vario.s <- variog(sim, max.dist=3, op="sm", band=0.2)
### plotting:
par(mfrow=c(2,2))
# plot the data:
zz <- cut(sim$data, 20)
mycol <- terrain.colors(20)
plot(sim[[1]], col=mycol[zz], pch=19, main="simulated points")
# cloud
plot(vario.c, main="variogram cloud")
# cloud + bin
plot(vario.bc, bin.cloud=TRUE);title(main="clouds for binned variogram")
# bin + smooth + theoretical variogram
plot(vario.b, main="binned variogram")
points(vario.s$u, vario.s$v, main="smoothed variogram", type="l", col="blue")
curve(variance-CovarianceFct(x=x,model="matern",
                             param=c(mean, variance, nugget, scale, nu)),
      0, 3, add=TRUE, col="black", lty=2, lwd=2)
```

Isotropy vs. anisotropy It is possible to study anisotropy through directional variograms. Separate variograms are computed for a given angle and tolerance. If they appear to be the same in all directions, one can say that there is no obvious anisotropic pattern. In the following we show a simulation of an anisotropic pattern and the corresponding directional variogram. On figure 2.6 it can be seen that the variogram for direction 45° reaches the sill much later than for other directions, as should be expected when looking at the point pattern.

```
aniso <- c(pi/4, 4) # (angle, amplitude)
set.seed(136)
## simulate an anisotropic pattern:
sim.anis <- grf(n=1000, xlims=c(0,10), ylim=c(0,10),
                cov.model="matern",
                cov.pars = c(variance, scale=2), #sigma^2
                nugget=nugget,
                kappa=0.5,
                aniso.pars= aniso) # anistropy
## plot the points:
zz <- cut(sim.anis$data, 20) # binning (for colors)
mycol <- terrain.colors(20)  # color palette
par(mfrow=c(1,2))
plot(sim.anis[[1]], col=mycol[zz], pch=19, main="simulated points")
```

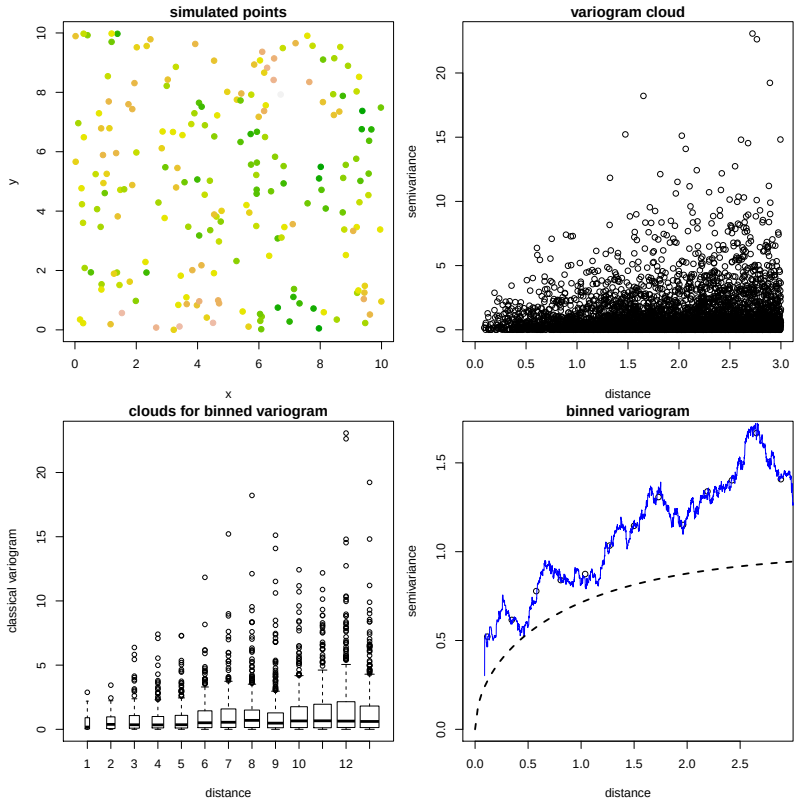

Figure 2.5: 200 simulated data points with the Matérn model and various variogram fitting displays.

```
## compute and plot variogram in 4 directions (0, 45, 90 and 135 degree):
vario.4 <- variog4(sim.anis, max.dist=6, tol=pi/8)
plot(vario.4)
```

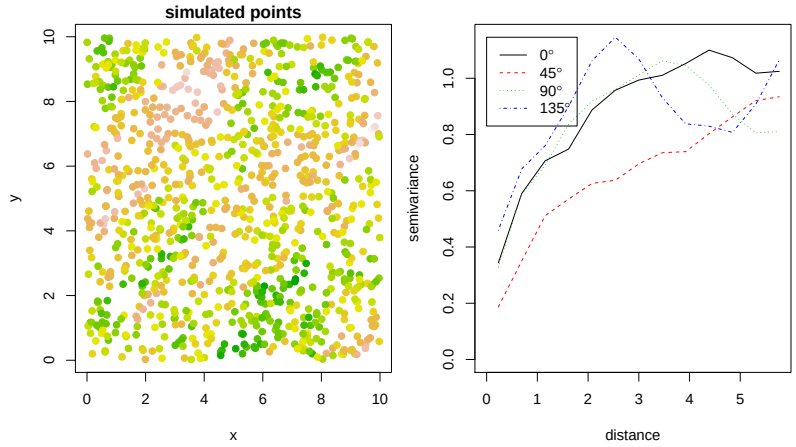

Figure 2.6: Simulated data for an anisotropic model and the respective directional variograms.

Parametric estimation Now we will show how to estimate parametrically the variogram on the same dataset simulated earlier and displayed in figure 2.5. The first method is weighted least square (WLS) and the second maximul likelihood (ML). For the WLS the weights are simply proportional to the number of pairs in the bin used to calculate the variogram. In both cases, we try to impose a nugget of 0 and see how it influences the solution. The results are displayed in figure 2.7.

```
# first we have to change the class of the simulated data:
simdat  <- as.geodata(sim)
# vario.b is the empirical variogram computed above
# ini.cov.pars are a set of initial values to try from
# kappa is the initial value bounded between 0.1 and 10
# WLS
wls <- variofit(vario.b, cov.model='matern',
                  ini.cov.pars=
                   expand.grid(seq(0,2,by=0.3), seq(0,2, by=0.3)),
                  fix.kappa=FALSE, kappa=1,
                 limits=pars.limits(kappa=c(lower=0.1, upper=10)))
# impose nugget =0
wls0 <- variofit(vario.b, cov.model='matern',
                  ini.cov.pars=
                    expand.grid(seq(0,2,by=0.3), seq(0,2, by=0.3)),
                  fix.kappa=FALSE, kappa=1,
                  limits=pars.limits(kappa=c(lower=0.1, upper=10)),
                  fix.nugget=TRUE, nugget=0)
## REML estimate:
## we use the WLS estimates as starting values:

ml <- likfit(simdat, cov.model='matern',
               ini.cov.pars=wls$cov.pars,
               fix.kappa=FALSE, kappa=1,
               limits=pars.limits(kappa=c(lower=0.1, upper=10)),
               lik.method="REML")
# impose nugget =0
ml0 <- likfit(simdat, cov.model='matern',
                ini.cov.pars=wls0$cov.pars,
                fix.kappa=FALSE, kappa=1,
                limits=pars.limits(kappa=c(lower=0.1, upper=10)),
                fix.nugget=TRUE, nugget=0, lik.method="REML")
## plot the results:
col.vec <- c('red', 'blue', 'green', 'orange', 'black')
lty.vec <- c(2,1,2,1, 2)
plot(vario.b)
lines(wls0, col=col.vec[1], lty=lty.vec[1], lwd=2)
lines(wls, col=col.vec[2], lty=lty.vec[2], lwd=2)
lines(ml0, col=col.vec[3], lty=lty.vec[3], lwd=2)
```

```
lines(ml, col=col.vec[4], lty=lty.vec[4], lwd=2)
# with the true variogram:
curve(nugget + variance -
         cov.spatial(x, cov.model="matern",
                     cov.pars= c(variance, scale), kappa=nu),
       0, 3, add=TRUE, col=col.vec[5], lty=lty.vec[5], lwd=2)
legend("bottomright",
        c("WLS (nugget=0)","WLS","REML (nugget=0)",  "REML", "truth"),
        col=col.vec, lty=lty.vec, lwd=2)
```

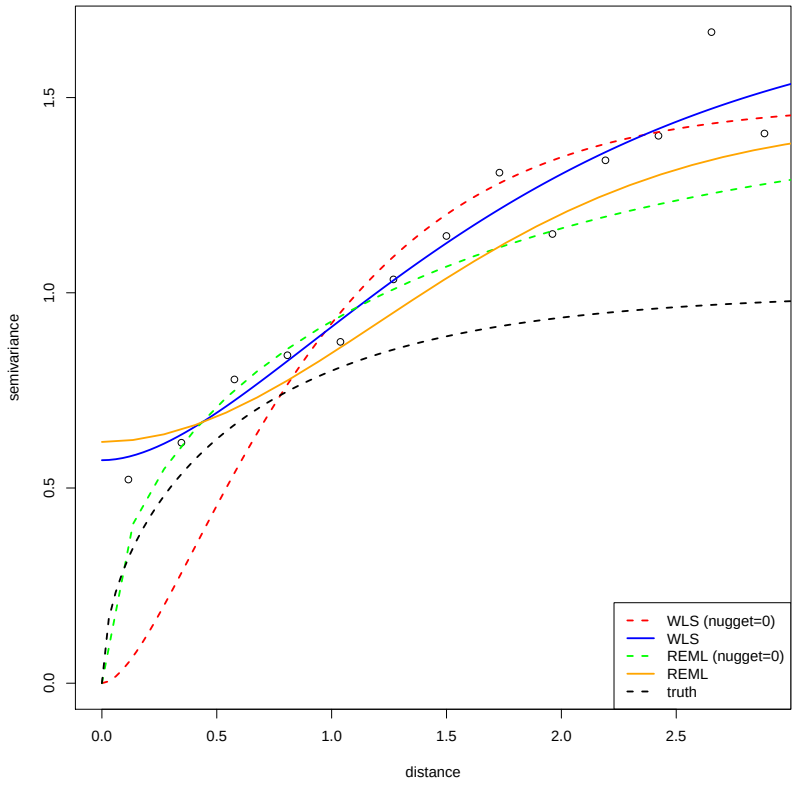

Figure 2.7: Parametric estimation of the variogram.

2.3 Kriging

We consider here the problem how to predict the value of a stochastic process $(Z(x))$ at new sites if we have observed $Z(x_i)$ for $i = 1, 2, \ldots, n$, and how to quantify the uncertainty of such a prediction. We will restrict ourselves to the case of Gaussian processes, or to the special class of predictors which are linear in the observed values $Z(x_i)$. We will see that both approaches lead to the same solution.

2.3.1 Simple Kriging

First we assume that the mean function m and the covariance function C of the process is known. If we also assume that the process is Gaussian, we can compute the conditional distribution $Z(x)$ given $Z(x_1), \ldots, Z(x_n)$ which quantifies (via expectation, standard deviation, quantiles etc.) the knowledge and the uncertainty about $Z(x)$. We can even compute the joint distribution at any number of new sites.

Theorem 2.2 *If $(Z(x))$ is a Gaussian process with arbitrary mean function m and covariance function C, then conditionally on $Z(x_1), \ldots, Z(x_n)$ $(Z(x))$ is again a Gaussian process with mean function*

$$m_n(x) = m(x) + c(x)^T K_n^{-1}(Z(x_1) - m(x_1), \ldots, Z(x_n) - m(x_n))^T$$

and covariance function

$$C_n(x, x') = C(x, x') - c(x)^T K_n^{-1} c(x')$$

where $c(x) = (C(x, x_1), \ldots, C(x, x_n))^T$ and K_n is the $n \times n$ matrix with elements $C(x_i, x_j)$. In particular the conditional mean is linear in the observed values and the conditional covariance is independent of the observed values.

Proof: We need to show that for any number of new sites $x_{n+1}, \ldots, x_{n+p}$ the conditional distribution of $\mathbf{Z}_{new} := (Z(x_{n+1}), \ldots, Z(x_{n+p}))$ given $\mathbf{Z}_{old} = (Z(x_1), \ldots, Z(x_n))$ is Gaussian with means $m_n(x_{n+i})$ and and covariances $C_n(x_{n+i}, x_{n+j})$. To simplify the notation, we assume that $m \equiv 0$ (for the general case, we simply replace $Z(x)$ by $Z(x) - m(x)$ in all formulae below). We use the result that the conditional density of $\mathbf{Z}_{new}$ given $\mathbf{Z}_{old}$ is obtained by multiplying the joint density by a function of $\mathbf{z}_{old}$ so that it becomes a density in $\mathbf{z}_{new}$ if we fix the value of $\mathbf{z}_{old}$. The joint density is equal to a constant times

$$\exp\left(-\frac{1}{2}\mathbf{z}_{old}^T Q_{old,old}\mathbf{z}_{old} - \frac{1}{2}\mathbf{z}_{new}^T Q_{new,new}\mathbf{z}_{new} - \mathbf{z}_{old}^T Q_{new,old}^T \mathbf{z}_{new}\right).$$

Here $Q_{old,old}$, $Q_{new,old}$ and $Q_{new,new}$ are the submatrices that we obtain if we partition the inverse of K_{n+p}, the joint covariance matrix of $(\mathbf{Z}_{old}, \mathbf{Z}_{new})$ in the obvious way

$$(K_{n+p})^{-1} = \begin{pmatrix} K_{old,old} & (K_{new,old})^T \\ K_{new,old} & K_{new,new} \end{pmatrix}^{-1} = \begin{pmatrix} Q_{old,old} & (Q_{new,old})^T \\ Q_{new,old} & Q_{new,new} \end{pmatrix}.$$

For later use, we state a result from linear algebra:

$$\begin{aligned} Q_{new,new} &= (K_{new,new} - K_{new,old}K_{old,old}^{-1}K_{new,old}^T)^{-1} \\ Q_{new,old} &= -Q_{new,new}K_{new,old}K_{old,old}^{-1} \\ Q_{old,old} &= K_{old,old}^{-1} - K_{old,old}^{-1}K_{new,old}^T Q_{new,old} \end{aligned}$$

which can be verified by blockwise matrix multiplication.

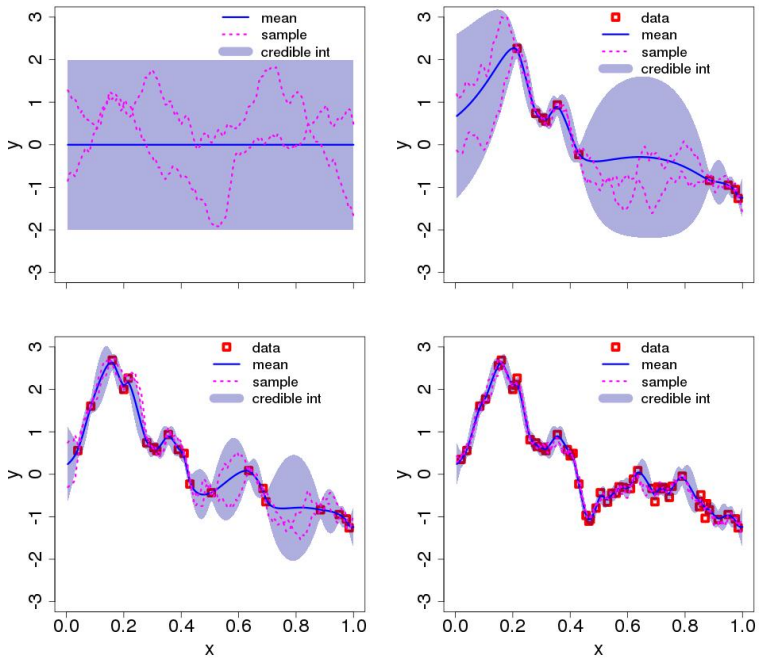

Figure 2.8: Reduction of uncertainty about a Gaussian random function by observing the values at an increasing number of points

We obtain the conditional density by collecting all terms in the exponent that contain $\mathbf{z}_{new}$ and normalizing. That is

$$f(\mathbf{z}_{new} \mid \mathbf{z}_{old}) \propto \exp\left(-\frac{1}{2}\mathbf{z}_{new}^T Q_{new,new}\mathbf{z}_{new} - \mathbf{z}_{old}^T Q_{new,old}^T \mathbf{z}_{new}\right)$$

where the proportionality sign $\propto$ means "up to a function of $\mathbf{z}_{old}$". The formula for completing a square,

$$\mathbf{z}^T A\mathbf{z} + 2\mathbf{b}^T\mathbf{z} = (\mathbf{z} + A^{-1}\mathbf{b})^T A(\mathbf{z} + A^{-1}\mathbf{b}) - \mathbf{b}^T A^{-1}\mathbf{b},$$

shows that $f(\mathbf{z}_{new} \mid \mathbf{z}_{old})$ is the Gaussian density with covariance $Q_{new,new}^{-1}$ and mean $-Q_{new,new}^{-1}Q_{new,old}\mathbf{z}_{old}$. The proof is now completed by using the formulae above for $Q_{new,new}$ and $Q_{new,old}$. □

Figure 2.8 illustrates how observations of the random function Z at points more and more points reduces the uncertainty about this function at other points. In each plot, the conditional mean given the observed values is shown together with two realizations from the conditional distribution of the process and a pointwise 95% prediction interval of the form mean $\pm$ 1.96 times the standard deviation.

Without assuming Gaussian distributions, we look for the optimal predictor in the class of linear predictors which have the form

$$\widehat{Z}(x) = \alpha_0(x) + \sum_{i=1}^{n} \alpha_i(x)(Z(x_i) - m(x_i)).$$

The functions α_i are determined such that the mean square error $\mathbb{E}((Z(x) - \widehat{Z}(x))^2)$ is minimal. By the rules for the variance we obtain

$$\begin{aligned}\mathbb{E}((Z(x) - \widehat{Z}(x))^2) &= (\mathbb{E}(Z(x)) - \mathbb{E}(\widehat{Z}(x)))^2 + \mathrm{Var}\left(Z(x) - \widehat{Z}(x)\right) \\ &= (m(x) - \alpha_0(x))^2 + C(x,x) - 2\sum_{i=1}^{n} \alpha_i(x)C(x,x_i) + \sum_{i,j=1}^{n} \alpha_i(x)\alpha_j(x)C(x_i,x_j).\end{aligned}$$

Thus clearly $\alpha_0(x) = m(x)$, and setting the partial derivatives with respect to the other $\alpha_i(x)$ equal to zero leads to the equation

$$K_n(\alpha_1(x), \ldots, \alpha_n(x))^T = c(x)$$

where $c(x)$ and K_n have the same meaning as in the Theorem above. The solution $\widehat{Z}(x)$ is also characterized by the following orthogonality conditions

$$\mathbb{E}(Z(x) - \widehat{Z}(x)) = 0, \quad \mathbb{E}((Z(x) - \widehat{Z}(x))Z(x_i)) = 0.$$

In particular

$$\mathrm{Cov}\left(Z(x) - \widehat{Z}(x), Z(x') - \widehat{Z}(x')\right) = \mathbb{E}((Z(x) - \widehat{Z}(x))Z(x')) = C(x,x') - c(x)^T K_n^{-1} c(x').$$

Therefore the covariance of the prediction errors at two sites is the same as the conditional covariance under Gaussian assumptions.

2.3.2 Universal and ordinary Kriging

The assumption of knowing the mean and the covariance functions is hardly ever satisfied. A more realistic approach is based on a parametric model for the mean function and a stationary or intrinsic covariance function. One can then estimate the mean and covariance function also from $\mathbf{Z}$ by one of the methods discussed before and use the estimated instead of the true quantities in the formulae for simple kriging. This plug-in approach will typically underestimate the uncertainty of the predictions because the uncertainty about the mean and the covariance is neglected. Moreover, it is not clear whether the resulting predictor still has some optimality property.

It turns out that one can find easily a procedure which deals correctly with the unknown mean function. For the unknown covariance the situation is more complicated, and the best solution seems to be a Bayesian approach.

We assume that the mean has the form

$$m(x) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(x) =: f(x)^T \beta$$

where f_j are known functions and the β_j are unknown parameters whereas the covariance function is assumed to be known. If we estimate the vector β by generalized least squares

$$\widehat{\beta} = (F_n^T K_n^{-1} F_n)^{-1} F_n^T K_n^{-1} \mathbf{z},$$

and use it instead of the true β in the formula for simple kriging, we have the predictor

$$\widehat{Z}(x) = f(x)^T \widehat{\beta} + c(x)^T K_n^{-1} (\mathbf{z} - F_n \widehat{\beta}).$$

This is called the universal kriging predictor, or – in the case $p = 0$ – the ordinary kriging predictor. It is a linear homogeneous function of the observations

$$\widehat{Z}(x) = \lambda(x)^T \mathbf{z}$$

where

$$\lambda(x)^T = (f(x)^T - c(x)^T K_n^{-1} F_n)(F_n^T K_n^{-1} F_n)^{-1} F_n^T K_n^{-1} + c(x)^T K_n^{-1}.$$

It follows therefore that for any β

$$\mathbb{E}(\widehat{Z}(x)) = \mathbb{E}(Z(x)) \Leftrightarrow \lambda(x)^T F_n = f(x)^T,$$

that is the universal kriging predictor is unbiased. Moreover, one obtains

$$\begin{aligned} \mathbb{E}((Z(x) - \widehat{Z}(x))^2) &= C(x,x) - c(x)^T K_n c(x) + \\ &\quad (f(x) - F_n K_n^{-1} c(x))^T (F_n^T K_n^{-1} F_n)^{-1} (f(x) - F_n K_n^{-1} c(x)). \end{aligned}$$

The first two terms give the mean square prediction error in case β is known. The third term thus represents the additional uncertainty due to estimating β. Under Gaussian assumptions $\widehat{Z}(x)$ is again Gaussian, and we can immediately compute for instance prediction quantiles.

If instead of the generalized least squares estimator one uses any other linear unbiased estimator of β, one obtains another homogeneous linear unbiased predictor, that is a predictor of the form $\lambda(x)^T \mathbf{z}$ with $\lambda(x)^T F_n = f(x)^T$. Using Lagrange multiplicators one can show that among all such predictors the one obtained from generalized least squares has the smallest mean square prediction error.

As said before, we still assume that the covariance function is known. However if the centered process $(Z(x) - m(x))$ is intrinsic, then the optimal coefficient vector $\lambda(x)$ depends only on the semivariogram. This follows from Lemma 2.1 because the first column of F_n is $(1, \ldots, 1)^T$ and therefore $1 - \sum_{j=1}^n \lambda_j(x) = 0$.

2.3.3 Bayesian Kriging

In Bayesian statistics the unknown parameters β and θ are conceptually the same as the unobserved value $Z(x)$ and we have

$$p(z(x), \beta, \theta \mid \mathbf{z}) = p(z(x) \mid \beta, \theta, \mathbf{z}) p(\beta, \theta \mid \mathbf{z}),$$

and therefore

$$p(z(x) \mid \mathbf{z}) = \int p(z(x) \mid \beta, \theta, \mathbf{z}) p(\beta, \theta | \mathbf{z}) d\beta d\theta.$$

The first term in the integrand on the right is Gaussian with mean and variance as given by Theorem 2.2. In the Bayesian approach one therefore averages the unknown parameters according to the posterior distribution instead of plugging in a point estimator. In particular, this takes uncertainty with respect to unknown parameters into account in a coherent way. The drawback of this approach is that the integral with respect to the posterior cannot be computed in closed form. If one can simulate from the posterior, the integral can be approximated by an average and the prediction density $p(z(x) \mid \mathbf{z})$ is approximated by a mixture of Gaussian densities.

If only β is unknown and if one chooses a Gaussian prior for β, then both the posterior $p(\beta \mid \mathbf{z})$ and the prediction density $p(z(x) \mid \mathbf{z})$ can be computed and they are again Gaussian. If one chooses an improper prior $p(\beta) \equiv 1$, then $p(z(x) \mid \mathbf{z})$ is Gaussian with mean equal to the universal kriging predictor and variance equal to its mean square prediction error.

Chapter 3

Models on a lattice

The specification of non-Gaussian processes is more difficult since mean and covariance are not enough. To simplify things one often discretizes space and considers only locations x which belong to a finite set $L \subset \mathbb{R}^d$. Often L is defined by the data one considers: When analyzing images, L is a regular lattice of pixels, or when analyzing epidemiological data, L is the lattice of districts which were used to aggregate the data. In other cases, one chooses a grid with a certain spatial resolution and then uses some deterministic interpolation scheme for values at points not belonging to this lattice.

Some notation: For models on a lattice, we denote the value at the site x by Z_x instead of $Z(x)$. We use bold-face $\mathbf{Z}$ for the random vector $(Z_x; x \in L)$ and for $A \subset L$ $\mathbf{Z}_A$ for the vector $(Z_x; x \in A)$. Finally $\mathbf{Z}_{-x}$ denotes the random vector $(Z_{x'}; x' \neq x)$. We use p as the generic symbol of any (conditional or unconditional) density or probability mass function. The arguments of p indicate in each case the density of which random vector we are considering. If we want to emphasize the spatial character of the model, we speak about random fields instead of random vectors.

3.1 Hierarchical models

Once we have decided on the grid L, one can define directly the joint distribution of the random vector $\mathbf{Z}$ which we assume generated the data we have. In many applications, this joint distribution is however specified indirectly by assuming that $\mathbf{Z}$ depends on some other unobserved random vector $\mathbf{Y}$ and we specify the marginal distribution of $\mathbf{Y}$ together with the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$. Assuming that densities exist, we then have

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y})p(\mathbf{z} \mid \mathbf{y}), \quad p(\mathbf{z}) = \int p(\mathbf{y})p(\mathbf{z} \mid \mathbf{y})d\mathbf{y}.$$

Examples

- Binary fields: One way to obtain a field with $Z_x \in \{0, 1\}$ is to assume a binary regression model with a spatially correlated random effect:

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathbf{y}) = \prod_{x;z_x=1} \Phi(\beta_0 + y_x) \prod_{x;z_x=0} (1 - \Phi(\beta_0 + y_x)),$$

  where $\mathbf{Y}$ is a Gaussian field with mean zero. The product formula means that given $\mathbf{Y}$ the variables Z_x are assumed to be independent. This can be written equivalently as

$$Z_x = \begin{cases} 1 & \beta_0 + Y_x - \varepsilon_x > 0 \\ 0 & \beta_0 + Y_x - \varepsilon_x \leq 0 \end{cases}$$

  where the ε_x are i.i.d. standard normal. We can include explanatory variables, or take other distributions for ε_x. For instance, the cumulative distribution function $G(u) = e^u/(1+e^u)$ leads to spatial logistic regression.

- Disease maps: Here Z_x denotes the number of cases with some disease in district x during a given period. One assumes that the Z_x are conditionally independent given $\mathbf{Y}$ with

$$Z_x \mid \mathbf{y} \sim \text{Poisson}(N_x \exp(\beta^T f_x + y_x)).$$

  Here N_x is the (age-adjusted) population in district x, $\exp(\beta^T f_x + y_x)$ is the excess risk, f_x is the vector of covariates at x and the spatial random effect $\mathbf{Y}$ is usually assumed to be a Gaussian field.

- Precipitation: In order to take into account the possibility of no rain and the skewness of the distribution of the amount of rain given that it rains, we can take

$$Z_x = \begin{cases} Y_x^{1/\lambda} & Y_x > 0 \\ 0 & Y_x \leq 0 \end{cases}$$

  where again $\mathbf{Y}$ is a Gaussian field which can be interpreted as the "potential rainfall".

- Blurred images: Here Z_x are the grey values in an observed image, and one assumes that they result from the "clean image" $\mathbf{Y}$ by taking a weighted average of values at nearby pixels and adding noise:

$$Z_x = \sum_{x'} H(x, x') Y_{x'} + \varepsilon_x$$

  Hence ε_x is Gaussian white noise and H describes how much the value at a pixel x' contributes to the observed value at pixel x. The probability distribution for Y is often chosen to reflect some basic properties of images like that sudden changes between neighboring pixels are rare. But since images can have (a few) sharp edges, non-Gaussian models for $\mathbf{Y}$ are preferable.

The basic tasks that we would like to solve are to infer what the observations $\mathbf{Z}$ tell us about the unobserved field $\mathbf{Y}$, and to estimate unknown parameters in the marginal distribution of $\mathbf{Y}$ or the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$. This is similar to what we have discussed in the chapter about kriging: Instead of an unobserved field $\mathbf{Y}$ we had the values of $\mathbf{Z}$ at sites without observations. Before we discuss methods for these basic tasks, we introduce the class of Markov random field models which are often used as a prior for $\mathbf{Y}$.

3.2 Spatial Markov property

In order to have some compromise between a realistic and a tractable model, one is looking for dependence which is essentially local. By this we mean that the value at one site is independent of the rest if we know the values at sites close-by:

$$p(z_x \mid \mathbf{z}_{-x}) = p(z_x \mid \mathbf{z}_{N(x)})$$

where $N(x)$ denotes the set of "neighbors" of x. The implicit assumption is that $N(x)$ consists only of a small number of sites. In order to check this property, one computes $p(z_x \mid \mathbf{z}_{-x})$ and then one sees whether it depends only on $\mathbf{z}_{N(x)}$.

If the above equation holds for all $x \in L$, then we call $\mathbf{Z}$ a *Markov random field* with respect to the underlying neighborhoods $N(x)$. In a hierarchical model as described in the previous section, a Markov random field is typically used for the unobserved field $\mathbf{Y}$ and not the observed field $\mathbf{Z}$, but in this section the distinction between $\mathbf{Y}$ and $\mathbf{Z}$ does not matter.

3.2.1 Gaussian Markov random fields

If $\mathbf{Z} \sim \mathcal{N}(m, K)$, then we can easily compute the conditional density $p(z_x \mid \mathbf{z}_{-x})$ (compare the derivation of Theorem 2.2)

$$\begin{aligned} p(z_x \mid \mathbf{z}_{-x}) &\propto \exp\left(-\frac{1}{2}(K^{-1})_{xx}(z_x - m_x)^2 - (z_x - m_x)\sum_{x' \neq x}(K^{-1})_{xx'}(z_{x'} - m_{x'})\right) \\ &\propto \exp\left(-\frac{(K^{-1})_{xx}}{2}\left(z_x - m_x + \sum_{x' \neq x}\frac{(K^{-1})_{xx'}}{(K^{-1})_{xx}}(z_{x'} - m_{x'})\right)^2\right), \end{aligned}$$

This is the density of the Gaussian distribution with mean value

$$\mathbb{E}(Z_x \mid \mathbf{Z}_{-x}) = m_x - \sum_{x' \neq x}\frac{(K^{-1})_{xx'}}{(K^{-1})_{xx}}(z_{x'} - m_{x'})$$

and variance $1/(K^{-1})_{xx}$. A Gaussian field is therefore Markovian iff the inverse of the covariance matrix is sparse:

$$x' \notin N(x) \Rightarrow (K^{-1})_{xx'} = 0.$$

In the following we denote the inverse of the covariance matrix K, the so-called precision matrix, by Q.

A Gaussian Markov random field can also be written in an autoregressive form

$$Z_x = m_x + \sum_{x' \neq x} w(x, x')(Z_{x'} - m_{x'}) + \varepsilon_x$$

where $w(x, x') = -Q_{xx'}/Q_{xx}$. Because the regression part is the conditional mean, a Gaussian Markov random field is also called a *conditional autoregression*. However, in contrast to the usual regression model, the errors ε_x are spatially correlated.

Lemma 3.1 *Assume that* $\mathbf{Z} \sim \mathcal{N}(0, K)$, $Q = K^{-1}$, $w(x, x') = -Q_{xx'}/Q_{xx}$ *and set* $\varepsilon_x = Z_x - \sum_{x' \neq x} w(x, x')Z_{x'}$. *Then*

$$\begin{aligned} \mathrm{Cov}(\varepsilon_x, Z_{x'}) &= \begin{cases} 0 & x \neq x' \\ 1/Q_{xx} & x = x' \end{cases} \\ \mathrm{Cov}(\varepsilon_x, \varepsilon_{x'}) &= \begin{cases} -\frac{w(x',x)}{Q_{xx}} & x \neq x' \\ 1/Q_{xx} & x = x' \end{cases} \end{aligned}$$

Proof: The first claim is nothing else than the orthogonality condition that we already found when we discussed simple kriging. The second claim follows from the first by a simple computation. □

If we set $w(x, x) = 0$ and $V = \mathrm{diag}(Q_{xx})$, then the above formulae can be summarized by

$$Q = V(I - W), \quad \varepsilon = (I - W)Z.$$

Thus instead of specifying the distribution by giving K or Q, we can also give the diagonal matrix V containing the conditional precisions and W containing the coefficients of the conditional autoregression. The meaning of these coefficients is easier to understand than that of some elements of K or Q. However, one then has the restriction that VW must be symmetric and $V - VW$ must be positive definite.

Example: Consider the regular lattice $L = \{1, 2, \ldots, n\}^2$ and assume that each site in the interior has 4 neighbors and that $Q_{xx} = \tau^2$, $W_{xx'} = \alpha$ for horizontal nearest neighbors and $W_{xx'} = \beta$ for vertical nearest neighbors. How to define Q and W for boundary sites is not so clear. Two simple solutions are periodic boundary conditions where we put the lattice on a torus, or free boundary conditions where the four corner sites have two and the other boundary sites three neighbors, but the values of Q and W are the same everywhere. $I - W$ is then positive definite if $|\alpha| + |\beta| < \frac{1}{2}$ because the matrix is diagonally dominant. We will see later the full admissible range of the parameters (α, β) when we give the eigenvalues of $I - W$. Clearly, we can add more neighbors as long as they are symmetric: With $2k$ neighbors, W has k parameters.

Because the deviations ε_x in a conditional autoregression are dependent, the joint density is not equal to the product of the conditional densities:

$$p(\mathbf{z}) \neq \prod_{x \in L} p(z_x \mid \mathbf{z}_{-x}).$$

One sees immediately that in the product on the right hand side the mixed terms $Q_{xx'}(z_x - m_x)(z_{x'} - m_{x'})$ appear twice as often than on the left hand side. Note the difference to the temporal Markov property used in time series where the product formula holds

$$p(z_1, z_2, \ldots, z_T) = p(z_1)p(z_2 \mid z_1)p(z_3 \mid z_2) \cdots p(z_T \mid z_{T-1}).$$

This is not a contradiction because in the temporal Markov property, one conditions only on the past and not on past and future.

In the literature, also so-called simultaneous autoregressions are discussed. They have the same form $Z = WZ + \varepsilon$, but with independent deviations ε_x. Therefore WZ is different from the vector of conditional means, and ε_x is correlated with some $Z_{x'}$, $x' \neq x$. If V is again the diagonal matrix containing the inverse variances of the ε_x, then we obtain

$$K = (I - W)^{-1}V^{-1}(I - W^T)^{-1} \Leftrightarrow Q = (I - W^T)V(I - W).$$

If W is sparse, then the same is true for Q (with a neighborhood of double size). Therefore, simultaneous autoregressions are a subclass of conditional autoregressions.

3.2.2 Gibbs representation

For a Gaussian Markov random field, we have seen that we can specify the joint distribution by specifying the conditional distributions of Z_x given its neighbors, but the connection is more complicated than a simple product formula. We show here that the same result holds true in general. This is useful because it is usually easier to choose a plausible form of the conditional distribution.

In order to formulate our main result, we need a concept from graph theory. We assume that the neighborhood relation is symmetric

$$x' \in N(x) \Leftrightarrow x \in N(x').$$

Then a set of sites $C \subset L$ is called a *clique* if $x \in N(x')$ for any two points $x, x' \in C$ with $x' \neq x$. Note that by definition, the empty set and every singleton $\{x\}$ is a clique. With nearest neighbors on the regular lattice, the only other cliques are the sets which consists of two adjacent sites. With more neighbors, there are also cliques which contain more than two sites.

Theorem 3.1 (Hammersley-Clifford ~ 1970**)** *If a random field* $\mathbf{Z}$ *has a strictly positive and continuous density* $p(\mathbf{z})$ *(or a strictly positive probability mass function), then it is a Markov random field iff* $p(\mathbf{z})$ *can be factored as a product of functions which depend only on the restriction of* $\mathbf{z}$ *to cliques:*

$$p(\mathbf{z}) = \prod_{Cliques\ C} g_C(\mathbf{z}_C).$$

The functions g_C *can be obtained from the conditional densities* $p(z_x \mid \mathbf{z}_{N(x)})$ *and they are unique under the additional standardization* $g_C(\mathbf{z}_C) = 1$ *if* $z_x = 0$ *for some* $x \in C$ *(or any other fixed value instead of* 0*).*

Proof: Sufficiency of the factorization for the Markov property is easy, because $p(z_x \mid \mathbf{z}_{-x})$ is equal to $p(\mathbf{z})$ times a function which depends only on $\mathbf{z}_{-x}$, and this function can be obtained from the normalization condition $\int p(z_x \mid \mathbf{z}_{-x})dz_x \equiv 1$. Therefore

$$p(z_x \mid \mathbf{z}_{-x}) \propto \prod_{C;\ x \in C} g_C(\mathbf{z}_C). \tag{3.1}$$

The converse is the difficult part. For $A \subseteq L$ and for any $\mathbf{z}$, we set $t_A(\mathbf{z})_x$ equal to z_x for $x \in A$ and equal to zero for $x \notin A$ (so t_A truncates the values outside A to zero). We then define

$$\Psi_A(\mathbf{z}_A) = -\log p(t_A(\mathbf{z})).$$

Note that for $x \notin A$ and $B = A \cup \{x\}$,

$$\Psi_B(\mathbf{z}_B) - \Psi_A(\mathbf{z}_A) = \log \frac{p(t_A(\mathbf{z}))}{p(t_B(\mathbf{z}))} = \log \frac{p(0_x \mid t_A(\mathbf{z})_{-x})}{p(z_x \mid t_A(\mathbf{z})_{-x})}$$

because $t_A(\mathbf{z})$ and $t_B(\mathbf{z})$ agree everywhere except at x. Therefore Ψ is determined by the conditional densities.

Next we introduce the *Moebius transform* of Ψ, denoted by Φ:

$$\Phi_B(\mathbf{z}_B) = \sum_{A \subset B} (-1)^{|B|-|A|} \Psi_A(\mathbf{z}_A).$$

If $z_x = 0$ for some $x \in B$, then $\Phi_B(\mathbf{z}_B) = 0$ because we can write

$$\Phi_B(\mathbf{z}_B) = \sum_{A \subset B \setminus \{x\}} (-1)^{|B|-|A|} (\Psi_A(\mathbf{z}_A) - \Psi_{A \cup \{x\}}(\mathbf{z}_{A \cup \{x\}}))$$

and each term in the sum is equal to zero.

The Moebius transform has an inverse which is given by

$$\Psi_A(\mathbf{z}_A) = \sum_{B \subset A} \Phi_B(\mathbf{z}_B),$$

see e.g. Lauritzen, Graphical Models (1996), Lemma A.2. Therefore

$$p(\mathbf{z}) = \exp\left(-\Psi_L(\mathbf{z})\right) = \exp\left(-\sum_{B \subset L} \Phi_B(\mathbf{z}_B)\right) = \prod_{B \subset L} g_B(\mathbf{z}_B)$$

where $g_B(\mathbf{z}_B) = \exp(-\Phi_B(\mathbf{z}_B))$. Moreover, the functions g_B are standardized as defined in the theorem.

So far we have used only the positivity of $p(\mathbf{z})$. It remains to show that the Markov property implies that $g_B \equiv 1$ or $\Phi_B \equiv 0$ unless B is a clique. If B is not a clique, there are two points x and x' in B which are not neighbors. For $A \subset B \setminus \{x, x'\}$ we set $A_1 = A \cup x$, $A_2 = A \cup x'$ and $A_3 = A \cup \{x, x'\}$. Then as we have seen above

$$\begin{aligned} \Psi_{A_1}(\mathbf{z}_{A_1}) - \Psi_A(\mathbf{z}_A) &= \log \frac{p(0_x \mid t_A(\mathbf{z})_{-x})}{p(z_x \mid t_A(\mathbf{z})_{-x})} \\ \Psi_{A_3}(\mathbf{z}_{A_3}) - \Psi_{A_2}(\mathbf{z}_{A_2}) &= \log \frac{p(0_x \mid t_{A_2}(\mathbf{z})_{-x})}{p(z_x \mid t_{A_2}(\mathbf{z})_{-x})}. \end{aligned}$$

Because $t_A(\mathbf{z})$ and $t_{A_2}(\mathbf{z})$ differ only at x' and x' is not a neighbor of x, it follows from the Markov property that

$$\Psi_{A_3}(\mathbf{z}_{A_3}) - \Psi_{A_2}(\mathbf{z}_{A_2}) - \Psi_{A_1}(\mathbf{z}_{A_1}) + \Psi_A(\mathbf{z}_A) = 0.$$

Therefore

$$\begin{aligned} \Phi_B(\mathbf{z}_B) &= \sum_{A \subset B} (-1)^{|B|-|A|} \Psi_A(\mathbf{z}_A) \\ &= \sum_{A \subset B \setminus \{x,x'\}} (-1)^{|B|-|A|} \left(\Psi_A(\mathbf{z}_A) - \Psi_{A_1}(\mathbf{z}_{A_1}) - \Psi_{A_2}(\mathbf{z}_{A_2}) + \Psi_{A_3}(\mathbf{z}_{A_3})\right) = 0. \end{aligned}$$

To prove uniqueness, we assume that we have a factorization with standardized g_C's. If we set $\Phi_C(\mathbf{z}_C) = -\log g_C(\mathbf{z}_c)$ if C is a clique and $\Phi_B(\mathbf{z}_B) = 0$ if B is not a clique, then

$$-\log p(\widetilde{\mathbf{z}_A}) = \sum_{B \subset A} \Phi_B(\mathbf{z}_B).$$

Applying the Moebius transform once again, we see that Φ_B is determined by $p(\mathbf{z})$. □

A few comments and corollaries of this theorem:

- If the factors are standardized, $\Phi_\emptyset = \Psi_\emptyset = -\log p(0)$ is a constant, and we can write
$$p(\mathbf{z}) \propto \prod_{C \text{ Clique}, C \neq \emptyset} g_C(\mathbf{z}_C).$$
However, there is in general no closed form expression of the normalization $\Phi_\emptyset$ if the other Φ_C's are given. (It is a sum over an exponential number of terms, or a high-dimensional integral).

- Models of the form given in the theorem arise naturally in statistical physics and are called *Gibbs-distributions*. There $\mathbf{Z}$ describes the microscopical state of a system of particles on the lattice L. By physical principles, the distribution of such a system is given by
$$p(\mathbf{z}) \propto \exp\left(-\frac{1}{T} \sum_C \Phi_C(\mathbf{z}_C)\right)$$

where T is the temperature and Φ_C is the interaction potential of the subsystem within C. Then $\sum_C \Phi_C(\mathbf{z}_c)$ is the total energy of the system, and the system prefers states with small energy. The lower the temperature the stronger this preference is.

- The Gaussian case is special because we have a so-called pair potential, Φ_C is identically zero for cliques with more than two elements.

- As in the Gaussian case, the joint density factorizes, but not as a product of the conditional densities of the value at one site given the rest:

$$p(\mathbf{z}) \neq \prod_{x \in L} p(z_x \mid \mathbf{z}_{N(x)}).$$

  On the right, each g_C appears $|C|$ times, and because of the normalization of the conditional densities there are factors g_A where A is not a clique.

- For any $A \subset L$ it holds that

$$p(\mathbf{z}_A \mid \mathbf{z}_{A^c}) \propto \prod_{C;\ C \cap A \neq \emptyset} g_C(\mathbf{z}_C),$$

  that is $p(\mathbf{z}_A \mid \mathbf{z}_{A^c})$ depends only on z_x for x in a clique C which intersects A. In other words, x must be a neighbor of some $x' \in A$. We call the set of such x the boundary of A and denote it by ∂A. Therefore we have the "global Markov property"

$$p(\mathbf{z}_A \mid \mathbf{z}_{A^c}) = p(\mathbf{z}_A \mid \mathbf{z}_{\partial A}).$$

Example: If Z_x can take only the values 0 or 1, then for a standardized potential $\Phi_C(\mathbf{z}_C) = 0$ unless $z_x = 1$ for all $x \in C$. Therefore we can write

$$\Phi_C(\mathbf{z}_C) = \theta_C \prod_{x \in C} z_x.$$

with one free parameter θ_C for each clique C. If $L \subset \mathbb{Z}^d$ and we also assume that the potentials do not change if we shift a clique, then the number of free parameters is equal to the number of cliques which contain a fixed site x. For instance, for $d = 2$ and 4 nearest neighbors, we have for an interior site

$$\mathbb{P}(Z_x = 1 \mid \mathbf{z}_{N(x)}) = \frac{\exp(-\theta_1 - \theta_2(z_{x+e_1} + z_{x-e_1}) - \theta_3(z_{x+e_2} + z_{x-e_2}))}{1 + \exp(-\theta_1 - \theta_2(z_{x+e_1} + z_{x-e_1}) - \theta_3(z_{x+e_2} + z_{x-e_2}))}.$$

where $e_1 = (1, 0)$ and $e_2 = (0, 1)$. Because this is like a logistic regression with the sum of horizontal and vertical neighbors as explanatory variables, the model is also called the *auto-logistic model*. In particular, the formula shows that we cannot specify $\mathbb{P}(Z_x = 1 \mid \mathbf{z}_{N(x)})$ arbitrarily for all 2^4 boundary conditions $\mathbf{z}_{N(x)}$.

For a pair potential (still in the binary case), we can obtain a more intuitive parametrization as follows. We write $x \sim x'$ if $x' \in N(x)$ and use that $2z_1z_2 = z_1 + z_2 - (z_1 - z_2)^2$ for binary z_1, z_2. Then

$$\sum_{x\in L} \theta_x z_x + \sum_{x\sim x'} \theta_{xx'} z_x z'_x = \sum_{x\in L} z_x(\theta_x + \sum_{x'\in N(x)} \theta_{xx'}) - \frac{1}{2}\sum_{x\sim x'} \theta_{xx'}(z_x - z_{x'})^2.$$

(Note that $\theta_{xx'} = \theta_{x'x}$, but in the sum $\sum_{x\sim x'}$ we take each pair twice). The second term is a weighted sum of conflicts between neighboring values and it is invariant if we exchange 0 and 1. The first term therefore regulates the preference of one value over the other.

This has motivated the use of models of the form

$$p(\mathbf{z}) \propto \exp\left(-\sum_x \Phi(z_x) - \sum_{x\sim x'} \Psi(z_x - z_{x'})^2\right)$$

with Ψ symmetric also for cases where Z_x takes arbitrary values in $\mathbb{R}$. If $\Psi(z)$ increases for positive z, this distribution favors realizations $\mathbf{z}$ which are locally constant. If both $\Phi(z)$ and $\Psi(z)$ are proportional to z^2, we are back in the Gaussian case, but if Ψ increases less steeply for $|z| \to \infty$, then there is a higher chance for occasional large jumps between neighboring values. This feature is important in image reconstruction which keeps sharp edges. If the range of Z_x is unbounded, then Φ must increase fast enough to make the right hand side integrable, otherwise there is no probability density proportional to the right hand side. This excludes in particular the choice $\Phi \equiv 0$. However, when we specify the prior distribution of a latent field $\mathbf{Y}$ often the posterior

$$p(\mathbf{y} \mid \mathbf{z}) \propto p(\mathbf{y})p(\mathbf{z} \mid \mathbf{y})$$

is a proper density even if $p(\mathbf{y})$ has infinite total mass. Hence in such cases we can take $\Phi \equiv 0$ and we obtain then what is called an *intrinsic prior*. Intrinsic Gaussian priors are discussed in detail e.g. in Rue and Held, Gaussian Markov Random Fields, 2005.

3.3 Inference for (latent) Markov fields

We study now more closely methods for the following problems:

1. Simulate realizations of a Markov random field.

2. Estimate unknown parameters θ in the potential of a Markov random field from one realization.

3. Estimate a latent Markov random field $\mathbf{Y}$ from observations $\mathbf{Z}$ if both the prior $p(\mathbf{y})$ and the conditional distribution $p(\mathbf{z} \mid \mathbf{y})$ are completely specified.

4. Estimate a latent Markov random field $\mathbf{Y}$ from observations $\mathbf{Z}$ together with unknown parameters in the prior $p(\mathbf{y})$ and/or the conditional distribution $p(\mathbf{z} \mid \mathbf{y})$.

As we will see, 2. and 4. are very difficult in general, but for Gaussian Markov random fields there are efficient methods.

3.3.1 The general case

Assume we are given a Markov random field

$$p(\mathbf{z}) \propto \exp(-\sum_C \Phi_C(\mathbf{z}_c))$$

and we would like to simulate realizations from this distribution. Direct simulation is usually not possible, and one uses iterative methods instead, so-called Markov chain Monte Carlo (MCMC). The *Gibbs sampler* is the following algorithm

1. Start with an arbitrary $\mathbf{z}^{(0)}$.
2. For $t = 1, 2, \ldots$, choose a site $x \in L$ either at random or according to a deterministic visiting schedule, set $\mathbf{z}_{-x}^{(t)} = \mathbf{z}_{-x}^{(t-1)}$ and draw $z_x^{(t)}$ from $p(z_x \mid \mathbf{z}_{N(x)}^{(t)})$.

By the definition of conditional densities, it is clear that if $\mathbf{z}^{(t-1)}$ is a realization from $p(\mathbf{z})$ then the same is true for $\mathbf{z}^{(t)}$. From the general theory of Markov chains it follows that asymptotically for large t $\mathbf{z}^{(t)}$ is a draw from $p(\mathbf{z})$ and for any function f we can estimate the mean

$$\int f(\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

by the average

$$\frac{1}{T-r} \sum_{t=r+1}^{T} f(\mathbf{z}^{(t)}).$$

Here r is a burn-in period which allows the simulations to reach the distribution $p(\mathbf{z})$. Sampling from $p(z_x \mid \mathbf{z}_{N(x)}^{(t)})$ is a one-dimensional problem for which there are good general methods, and if $N(x)$ is small the computation of the conditional densities is fast. Still the Gibbs sampler is a computationally intensive method because it typically needs large T and r, and their choice is not always easy.

For inference about $\mathbf{Y}$ based on $\mathbf{Z}$ when there are no unknown parameters, we use the posterior distribution

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{p(\mathbf{y}) p(\mathbf{z} \mid \mathbf{y})}{p(\mathbf{z})} \propto p(\mathbf{y}) p(\mathbf{z} \mid \mathbf{y}).$$

If we are interested in a point estimate, we can use the posterior mean or the posterior mode

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{z}) = \arg\max_{\mathbf{y}} p(\mathbf{y})p(\mathbf{z} \mid \mathbf{y}) = \arg\max_{\mathbf{y}} \left(\log p(\mathbf{y}) + \log p(\mathbf{z} \mid \mathbf{y})\right).$$

Posterior quantiles or posterior standard deviations can be used to quantify uncertainty. There are usually no closed form expressions of these estimates. If the prior $p(\mathbf{y})$ is Markov and if observations at a site x depend only on the latent field at the same site, i.e.

$$p(\mathbf{z} \mid \mathbf{y}) = \prod_{x \in L} p(z_x|y_x),$$

then $p(\mathbf{y} \mid \mathbf{z})$ is again Markov with the same neighborhood:

$$p(y_x \mid \mathbf{y}_{-x}, \mathbf{z}) \propto p(y_x \mid \mathbf{y}_{-x})p(z_x \mid y_x) = p(y_x \mid \mathbf{y}_{N(x)})p(z_x|y_x).$$

Thus only the single site potentials are affected by $\mathbf{z}$, and one can use the Gibbs sampler to approximate posterior means, standard deviations and quantiles.

For the posterior mode, one can use iterative componentwise maximization (called iterated conditional mode). One iterates

$$\widehat{y_x} = \arg\max_{y_x} \left(\log p(y_x \mid \widehat{\mathbf{y}}_{N(x)}) + \log p(z_x \mid y_x)\right)$$

until the changes become negligible. In general, one can however get stuck at local maxima, and the result depends on both the starting value and the visiting schedule of the sites. An algorithm which is in principle able to find the global maximum is *simulated annealing* which at iteration t sets $\mathbf{y}_{-x}^{(t)} = \mathbf{y}_{-x}^{(t-1)}$ and draws $y_x^{(t)}$ from the density proportional to $p(y_x \mid \mathbf{y}_{N(x)}^{(t)}, z_x)^{\beta_t}$. If $\beta_t \equiv 1$, this is just the Gibbs sampler, and for β_t large it is close to the iterated conditional mode algorithm. The idea of simulated annealing is to let β_t increase to ∞ sufficiently slowly so that the randomness allows to escape from local maxima.

If the potentials depend on unknown parameters θ, we have to estimate them. This is difficult even if the Markov random field is fully observed because the normalizing constant in $p_\theta(\mathbf{z})$ depends on θ:

$$p_\theta(\mathbf{z}) = \frac{\exp(-\sum_{C \neq \emptyset} \Phi_C(\mathbf{z}_c; \theta))}{\int \exp(-\sum_{C \neq \emptyset} \Phi_C(\mathbf{z}'_c; \theta)) d\mathbf{z}'},$$

but there is no closed form expression for it. This makes both maximum likelihood and Bayesian estimation difficult. A simple alternative is the pseudo-MLE

$$\widehat{\theta} = \arg\max_{\theta} \prod_{x \in L} p_\theta(z_x \mid \mathbf{z}_{N(x)}) = \arg\max_{\theta} \prod_{x \in L} \frac{\exp(-\sum_{C \ni x} \Phi_C(\mathbf{z}_C; \theta))}{\int \exp(-\sum_{C \ni x} \Phi_C(\mathbf{z}'_C; \theta)) dz'_x}.$$

This involves again normalizing constants, but this time we have one-dimensional integrals which are easier. In particular, if the Markov field is discrete, the

integral becomes a sum over a few terms. There are methods to estimate normalizing constants and to approximate the MLE, but these are computationally intensive, see Chapter 5 in Chen, Shao and Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer 2000.

If we have a latent field $\mathbf{y}$ and unknown parameters θ, then the joint estimation of $\mathbf{y}$ and θ is again difficult because most methods require to evaluate $p_\theta(\mathbf{y})$ for given θ and y. Simple alternatives are based on calibration or on cross validation. Calibration means to choose θ such that $\widehat{y}$ has the desired properties for a class of simple, but somehow typical $\mathbf{y}$'s. Cross validation means to choose θ such that $\log p(z_x \mid \mathbf{y} = \widehat{\mathbf{y}}(\mathbf{z}_{-x}))$ is maximal where $\widehat{\mathbf{y}}(\mathbf{z}_{-x})$ is the estimate of $\mathbf{y}$ based on z_{-x} (that is we take z_x as missing for estimating $\mathbf{y}$ and compare the observed value z_x with its distribution assuming that the estimate is correct). There is also a Bayesian version of cross validation, see Gelfand and Dey, JRSS B 56 (1994), 501-514.

A few examples for this chapter are the paper of Tjelmeland and Besag (Scand. J. Statistics 25, 1998, 415-433) for choosing the potential so that the realizations of the associated binary Markov random field have desired features, and the paper by Geman and Reynolds (IEEE Transactions Pattern Analysis and Machine Intelligence 14, 1992, 367-383) on restoration of blurred images where $p(\mathbf{y})$ is not log concave.

3.3.2 The Gaussian case

In the Gaussian case, the situation is much better, due to the work of Havard Rue and collaborators.

To begin, we discuss how to simulate a Gaussian Markov random field $\mathbf{Z}$ specified by a sparse precision matrix Q and how to estimate parameters in Q if $\mathbf{Z}$ is fully observed. In simple cases, the eigenvalues λ_i and eigenvectors u^i of the precision matrix Q are available in closed form. In that case, we can compute the normalizing constant

$$\det(Q)^{-1/2} = \prod_i \lambda_i^{-1/2}$$

and we can simulate $\mathbf{Z} \sim \mathcal{N}(0, Q^{-1})$ by simulating ξ_i i.i.d. standard normal and setting

$$\mathbf{Z} = \sum_i \frac{u^{(i)}}{||u^{(i)}||} \xi_i / \sqrt{\lambda_i}.$$

Example: Consider the regular lattice $L = \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ with nearest neighbors, free boundary conditions and shift invariant parameters. This means $Q_{xx} = \tau^2$ for all x, $Q_{xx'} = -\alpha\tau^2$ for horizontal nearest neighbors x, x' and $Q_{xx'} = -\beta\tau^2$ for vertical nearest neighbors x, x'. Then the eigenvalues are (without proof)

$$\lambda = \tau^2 \left(1 - 2\alpha \cos\left(\frac{\pi i_1}{n_1 + 1}\right) - 2\beta \cos\left(\frac{\pi i_2}{n_2 + 1}\right)\right) \quad (1 \le i_1 \le n_1, 1 \le i_2 \le n_2),$$

with corresponding eigenvectors

$$u_{x_1,x_2} = \sin\left(\frac{\pi i_1 x_1}{n_1+1}\right)\sin\left(\frac{\pi i_2 x_2}{n_2+1}\right).$$

In particular, we see that the eigenvalues are positive iff

$$|\alpha|\cos\left(\frac{\pi}{n_1+1}\right) + |\beta|\cos\left(\frac{\pi}{n_2+1}\right) < \frac{1}{2}.$$

For $n_1 = n_2 = 10$, this means for example $|\alpha| + |\beta| < 0.5211$. Unfortunately, it turns out that in order to have a nearest neighbor correlation of say greater than 0.9, one has to go extremely close to the boundary of the allowed parameter space.

If one cannot obtain eigenvalues and eigenvectors in closed form, then one uses the Cholesky decomposition $Q = FF^T$ in order to compute $\det(Q)^{-1/2}$ (i.e. the normalizing constant) and to simulate from $\mathcal{N}(0, Q^{-1})$. It turns out that the Cholesky decomposition depends on the ordering of the sites in L, and one can try to choose an ordering such that the elements of F have many zeroes at places known in advance. Then there is no need to compute these elements and the computation becomes efficient. Finding sparse Cholesky factorizations has been studied extensively in numerical linear algebra. Because the results can be understood in probabilistic terms, we go into some detail.

First we have to understand the probabilistic meaning of the Cholesky decomposition. We set $U = F^T = D(I - A)$ where D is diagonal and A is upper diagonal with zeroes in the diagonal. Then $Q = U^T U$ and therefore (with N equal the number of sites in L)

$$\mathbf{z}^T Q \mathbf{z} = ||U\mathbf{z}||^2 = ||D(I - A\mathbf{z})||^2 = \sum_{i=1}^{N} d_i^2 (z_{x_i} - \sum_{j=i+1}^{N} A_{ij} z_{x_j})^2.$$

Because also $\det(Q) = \prod_i d_i^2$, this implies

$$p(\mathbf{z}) = \prod_{i=N}^{1} \frac{d_i}{\sqrt{2\pi}} \exp\left(-\frac{d_i^2}{2}\left(z_{x_i} - \sum_{j=i+1}^{N} A_{ij} z_{x_j}\right)^2\right).$$

On the other hand, we can always write

$$p(\mathbf{z}) = \prod_{i=N}^{1} p(z_{x_i} \mid z_{x_{i+1}}, \ldots, z_{x_n}),$$

that is the Cholesky decomposition gives explicit expressions for the conditional mean and variance of Z_{x_i} given $Z_{x_{i+1}}, \ldots, Z_{x_N}$.

The global Markov property gives us now information about elements in the Cholesky factor F which have to be zero. Denote by $L_i^+ = \{x_j; j > i\}$ and

$L_i^- = \{x_j, j \le i\}$ the "future" and "present plus past" of x_i with respect to the chosen ordering in the Cholesky decomposition. By the global Markov property

$$p(\mathbf{z}_{L_i^-} \mid \mathbf{z}_{L_i^+}) = p(\mathbf{z}_{L_i^-} \mid \mathbf{z}_{\partial L_i^-}).$$

Integrating both sides over all z_{x_j} for $j < i$ shows that $p(z_{x_i} \mid \mathbf{z}_{L_i^+}) = p(z_{x_i} \mid \mathbf{z}_{\partial L_i^-})$ and therefore

$$\left(j > i \text{ and } x_j \notin \partial L_i^-\right) \;\Rightarrow\; A_{ij} = 0 \;\Leftrightarrow\; F_{ji} = 0.$$

Hence we obtain many zeroes in F if we take an ordering where the boundaries between past and future values are small. In the one-dimensional case $L = \{1, \ldots, n\}$ with k-nearest neighbors, the usual order gives a Cholesky factor which is as sparse as the precision Q. In higher dimensions, things are less fortunate. If $L = \{1, \ldots, n\}^2$ with nearest neighbors, the lexicographic order leads to a Cholesky factor with $F_{ij} = 0$ if $j < i - n$. A general algorithm to obtain a good ordering of any lattice is "nested dissection" where you successively split the set of sites into two approximate halves with a small boundary in between. For more details, see the book by Rue and Held.

Next, we turn to the case where we want to estimate both, a latent Gaussian Markov random field $\mathbf{y}$ and parameters θ, from observations $\mathbf{z}$:

$$p(\mathbf{y} \mid \theta) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T Q(\theta)\mathbf{y}\right), \quad p(\mathbf{z} \mid \mathbf{y}, \theta) = \prod_{x \in L} p(z_x \mid y_x, \theta).$$

Here $Q(\theta)$ is sparse, and the non-zero elements are the same for all θ.

We discuss two methods: One simulates from $p(\theta, \mathbf{y} \mid \mathbf{z})$ and is described in the book of H. Rue and L. Held, Section 4.4. The other uses analytical approximations of $p(\theta \mid \mathbf{z})$ and $p(\mathbf{y} \mid \mathbf{z})$, called Laplace approximations. It is due to H. Rue, S. Martino and N. Chopin, JRSS B 71 (2009), 319-392. Both methods use a Gaussian approximation $\tilde{p}(\mathbf{y} \mid \mathbf{z}, \theta)$ of

$$p(\mathbf{y} \mid \mathbf{z}, \theta) = \frac{p(\mathbf{z} \mid \mathbf{y}, \theta)p(\mathbf{y} \mid \theta)}{p(\mathbf{z} \mid \theta)} \propto p(\mathbf{z} \mid \mathbf{y}, \theta)p(\mathbf{y} \mid \theta).$$

We start with an approximation of the right hand side around a value $\mathbf{y}^* = \mathbf{y}^*(\theta, \mathbf{z})$ to be determined later:

$$p(\mathbf{z} \mid \mathbf{y}, \theta)p(\mathbf{y} \mid \theta) \approx p(\mathbf{z} \mid \mathbf{y}^*, \theta)p(\mathbf{y}^* \mid \theta) \exp\left((\mathbf{y} - \mathbf{y}^*)^T a^* - \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T Q^* (\mathbf{y} - \mathbf{y}^*)\right)$$

where both Q^* and a^* can depend on θ and $\mathbf{z}$. By completing the square, the exponential term on the right hand side is equal to

$$\exp\left(\frac{1}{2}a^{*T} Q^{*-1} a^*\right) \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{y}^* - Q^{*-1}a^*)^T Q^* (\mathbf{y} - \mathbf{y}^* - Q^{*-1}a^*)\right).$$

Therefore we can approximate $p(\mathbf{y} \mid \mathbf{z}, \theta)$ by the normal density with mean the $\mathbf{y}^* + Q^{*-1}a^*$ and precision Q^*. Moreover, because

$$p(\mathbf{z} \mid \theta) = \int p(\mathbf{z} \mid \mathbf{y}, \theta)p(\mathbf{y} \mid \theta)d\mathbf{y},$$

we also obtain

$$p(\mathbf{z} \mid \theta) \approx \tilde{p}(\mathbf{z} \mid \theta) = (2\pi)^{N/2} \frac{p(\mathbf{z} \mid \mathbf{y}^*, \theta) p(\mathbf{y}^* \mid \theta)}{\det(Q^*)^{1/2}} \exp\left(\frac{1}{2} a^{*T} Q^{*-1} a^*\right).$$

That is, we have found an explicit approximation of the intractable likelihood of θ given the observations $\mathbf{z}$. For later use we also note that for $\mathbf{y}$ close to $\mathbf{y}^*$

$$p(\mathbf{z} \mid \theta) = \frac{p(\mathbf{y} \mid \theta) p(\mathbf{z} \mid \mathbf{y}, \theta)}{p(\mathbf{y} \mid \mathbf{z}, \theta)} \approx \frac{p(\mathbf{y} \mid \theta) p(\mathbf{z} \mid \mathbf{y}, \theta)}{\tilde{p}(\mathbf{y} \mid \mathbf{z}, \theta)} \approx \tilde{p}(\mathbf{z} \mid \theta).$$

(For $\mathbf{y} = \mathbf{y}^*$ the last $\approx$ becomes an equality).

How do we choose $\mathbf{y}^*$, a^* and Q^* ? One possibility is to choose

$$\mathbf{y}^* = \mathbf{y}^*(\theta, \mathbf{z}) = \arg\max_{\mathbf{y}} \left(\log p(\mathbf{z} \mid \mathbf{y}, \theta) + \log p(\mathbf{y} \mid \theta)\right),$$

the posterior mode of $p(\mathbf{y} \mid \mathbf{z}, \theta)$, and to determine a^* and Q^* through a second order Taylor approximation at $\mathbf{y}^*$. This gives

$$a^* = 0, \quad Q^* = Q(\theta) + \operatorname{diag}\left(\frac{\partial^2}{\partial y_x^2} \log p(z_x \mid y_x, \theta)\Big|_{y_x = y_x^*}\right).$$

Because Q^* and Q have zero entries at the same positions, the same ordering of sites can be used for any θ to compute the determinant of Q^* (the normalizing constant) and to simulate from $\tilde{p}$. This choice of $\mathbf{y}^*$ is likely to give a good approximation of $p(\mathbf{z} \mid \theta)$ because the integrand is well approximated where it is large and where its contribution to the integral is therefore biggest. A disadvantage is that $\mathbf{y}^*$ usually has to be computed iteratively and this has to be done for many values of θ . Typically computing $\mathbf{y}^*$ is easier if we take the maximum likelihood estimator

$$y_x^* = \arg\max_{y_x} \log p(z_x \mid y_x, \theta).$$

For instance, in the case of disease mapping without explanatory variables, there are no unknown parameters in $p(z_x \mid y_x)$:

$$\log p(z_x \mid y_x) = -N_x \exp(y_x) + z_x y_x + z_x \log N_x - \log(z_x!)$$

and the arg max is equal to $y_x^* = \log(z_x/N_x)$. To determine a^* and Q^*, one can use again a second order Taylor approximation at $\mathbf{y}^*$. Then as before, Q^* and Q differ only on the diagonal, but $a* = -Q(\theta)\mathbf{y}^* \neq 0$.

The analytical method works directly with the above approximation for $p(\mathbf{z} \mid \theta)$, choosing $\mathbf{y}^*$ as the mode of the posterior $p(\mathbf{y} \mid \mathbf{z}, \theta)$. If θ is low-dimensional (say it has at most 6 components), one can then numerically find the posterior mode $\widehat{\theta} = \arg\max p(\theta)\tilde{p}(\mathbf{z} \mid \theta)$, a set of points $\theta^{(i)}$ around $\widehat{\theta}$ and weights $w^{(i)}$ such that

$$\int f(\theta)\tilde{p}(\theta \mid \mathbf{z}) d\theta \approx \sum_{i=1}^{M} f(\theta^{(i)}) w^{(i)}.$$

In order to estimate the latent field, one can use

$$p(\mathbf{y} \mid \mathbf{z}) = \int p(\mathbf{y} \mid \mathbf{z}, \theta) p(\theta \mid \mathbf{z}) d\theta \approx \sum_{i=1}^{M} \tilde{p}(\mathbf{y} \mid \mathbf{z}, \theta^{(i)}) w^{(i)},$$

or – because formulae for high-dimensional densities are not very useful – its marginalization

$$p(y_x \mid \mathbf{z}) \approx \sum_{i=1}^{M} \tilde{p}(y_x \mid \mathbf{z}, \theta^{(i)}) w^{(i)}.$$

It turns out that further improvements are possible by modifying the approximation $\tilde{p}$. Details are in the paper by Rue, Martino and Chopin.

The simulation method uses a Markov chain Monte Carlo method which updates simultaneously θ and $\mathbf{y}$ and is thus different from a Gibbs sampler. It belongs to the class of so-called Metropolis-Hastings algorithms which combine a proposal step with an acceptance step to obtain convergence to the target distribution (if you have never heard of this algorithm, please skip this paragraph!). In our case it proceeds iteratively from an arbitrary starting value $\theta^{(0)}, \mathbf{y}^{(0)}$ as follows

1. Propose new values $\theta^* \sim \mathcal{N}(\theta^{(t-1)}, \Sigma)$, and $\mathbf{y}^* \sim \tilde{p}(\mathbf{y} \mid \mathbf{z}, \theta^*)$.
2. Generate U uniform on $(0, 1)$ and set $\theta^{(t)} = \theta^*, \mathbf{y}^{(t)} = \mathbf{y}^*$ if

$$U \leq \frac{p(\theta^*) p(\mathbf{y}^* \mid \theta^*) p(\mathbf{z} \mid \mathbf{y}^*, \theta^*) \tilde{p}(\mathbf{y}^{(t-1)} \mid \mathbf{z}, \theta^{(t-1)})}{p(\theta^{(t-1)}) p(\mathbf{y}^{(t-1)} \mid \theta^{(t-1)}) p(\mathbf{z} \mid \mathbf{y}^{(t-1)}, \theta^{(t-1)}) \tilde{p}(\mathbf{y}^* \mid \mathbf{z}, \theta^*)}.$$

   Otherwise set $\theta^{(t)} = \theta^{(t-1)}$ and $\mathbf{y}^{(t)} = \mathbf{y}^{(t-1)}$.

Note that if the approximation $\tilde{p}$ has no error, then the acceptance condition becomes

$$U \leq \frac{p(\theta^*) p(\mathbf{z} \mid \theta^*)}{p(\theta^{(t-1)}) p(\mathbf{z} \mid \theta^{(t-1)})} = \frac{p(\theta^* \mid \mathbf{z})}{p(\theta^{(t-1)} \mid \mathbf{z})}.$$

This is the standard Metropolis algorithm for sampling from the target distribution $p(\theta \mid \mathbf{z})$ if the latter were available in closed form. If $\tilde{p}$ is a good approximation, the acceptance probabilities should not change much compared with such an idealized algorithm.

Which of the two methods should be prefered ? The simulation method converges in principle to the true joint posterior as the number of iterations goes to infinity, whereas the analytical method has a fixed error coming from $p(\mathbf{z} \mid \theta) \neq \tilde{p}(\mathbf{z} \mid \theta)$. In practical situations, typically this error is small and one needs a huge number of simulation iterations to detect it. However, the analytical method is much faster.

Chapter 4

Point patterns

4.1 Basic concepts

Definition 4.1 *A point pattern, also called a point process, on a set $B \subseteq \mathbb{R}^d$ is a random, locally finite subset $Z = Z(\omega) = \{x_1(\omega), x_2(\omega), \ldots\}$ of B. Here, locally finite means that every compact (i.e. closed and bounded) set $A \subset B$ contains only a finite number of points. Note that the enumeration of the points of Z is arbitrary, and points with different indices are assumed to be different (in accordance with the notion of a set).*

Basic examples are the location of plants of some species, or the epicenters of earthquakes. In some applications, the points of Z carry additional information, a so-called mark, which may be categorical (plants of different species in a region) or continuous (the magnitude of an earthquake). We will not discuss marked point processes here.

The number of points in A, $A \subset B$, is denoted by

$$N(A) = |A \cap Z| = \sum_{i=1}^{\infty} 1_A(x_i)$$

It takes values in $\{0, 1, 2, \ldots, \infty\}$. Clearly, N is a measure on B with the Borel σ-algebra $\mathcal{B}(B)$. In this whole chapter, $|A|$ denotes the number of elements of A if A is finite, and the area or volume of A if A has non-empty interior.

In order to define precisely what is meant by a random set (a generalization of a random variable), some measure theory would be needed. We omit these technical parts. The distribution of a point pattern Z is defined to be the collection of joint distributions of $(N(A_1), \ldots, N(A_k))$ for all k and all bounded $A_1, \ldots A_k$ in B. A model for a point pattern specifies all these joint distributions in a coherent way. If B itself is compact, this can be done by specifying a distribution for $N(B)$ and, for any $n \in \mathbb{N}$, a density f_n for putting n points $x_1, \ldots x_n$ in B. Because the enumeration of the points is arbitrary, f_n must be symmetric, that is its value is the same for any permutation of the arguments.

Example 4.1 (Poisson point pattern) *For a bounded set B, a Poisson point pattern has $N(B) \sim$ Poisson(λ) and for any n the points $x_1, \ldots, x_n$ are i.i.d. $\sim f(x)dx$ where f is a probability density on B. The distribution of a Poisson point pattern is given in the following lemma.*

Lemma 4.1 *If Z is a Poisson point pattern as defined above, then for any number of pairwise disjoint subsets $A_1, \ldots, A_k$ of B, the random variables $N(A_i)$ for $1 \leq i \leq k$ are independent and Poisson($\lambda \int_{A_i} f(x)dx$)-distributed.*

Proof: *We first prove the result for $k = 1$.*

$$\begin{aligned}
\mathbb{P}(N(A) = j) &= \sum_{n=j}^{\infty} \mathbb{P}(N(A) = j \mid N(B) = n) \cdot \mathbb{P}(N(B) = n) \\
&= \sum_{n=j}^{\infty} \binom{n}{j} \left(\int_A f(x)dx \right)^j \left(1 - \int_A f(x)dx \right)^{n-j} \cdot \exp(-\lambda) \frac{\lambda^n}{n!} \\
&= \exp(-\lambda) \frac{\left(\lambda \int_A f(x)dx\right)^j}{j!} \sum_{n=j}^{\infty} \frac{1}{(n-j)!} \lambda^{n-j} (1 - \int_A f(x)dx)^{n-j} \\
&= \exp\left(-\lambda \int_A f(x)dx \right) \frac{\left(\lambda \int_A f(x)dx\right)^j}{j!}.
\end{aligned}$$

The first equality is the law of total probability. The second equality holds by the definition of the Poisson point pattern because if there are n points in total, the number of points in A has a binomial distribution. The last equality follows from the Taylor series of the exponential function.

The proof for $k > 1$ is similar, instead of the binomial one has the multinomial distribution. □

Because for a Poisson point pattern the points are independent, it is not an interesting model by itself, but it is often used as a null model and one wants to describe in which sense the observed pattern differs from a Poisson point pattern.

Although observations are always restricted to some bounded domain W, it would be difficult to justify an assumption that there are no points outside W. One therefore needs a model with a B that is substantially larger than W. For the concept of stationarity, one even has to consider point patterns which are defined on the whole of $\mathbb{R}^d$.

Definition 4.2 *A point pattern Z on $\mathbb{R}^d$ is called* stationary *if the distribution is the same everywhere, that is $(N(A_1), \ldots, N(A_k))$ has the same distribution as $(N(A_1 + h), \ldots, N(A_k + h))$ for all $h \in \mathbb{R}^d$, all k and all $A_1, \ldots, A_k$. A stationary point pattern is called* isotropic *if the distribution is the same in all directions, that is $(N(A_i), 1 \leq i \leq k)$ and $(N(R(A_i)), 1 \leq i \leq k)$ has the same distribution for any rotation R about the origin.*

As in the case of stochastic processes, stationarity and isotropy justify spatial averaging for estimation purposes and make it possible to infer the distribution of a point pattern from one single realization.

Poisson point patterns exists also for unbounded domains B. They are specified by a function $\lambda : B \to \mathbb{R}_+$ such that for all bounded sets A $\int_A \lambda(x) < \infty$, and their distribution is such that for any number of bounded, pairwise disjoint subsets $A_1, \ldots, A_k$ of B, the random variables $(N(A_i), 1 \leq i \leq k)$ are independent and Poisson$(\int_{A_i} \lambda(x)dx)$-distributed. We omit the proof that such a point pattern exists for any λ. The function λ is called the intensity. In our previous notation, the intensity is $\lambda f(x)$. A Poisson point pattern is stationary iff the intensity is constant, and then it is automatically isotropic.

4.2 Moments and other characteristics

We start with the intuitive definition of the first and second moment.

The first moment is called the *intensity function*. It is defined

$$\lambda(x) = \frac{\mathbb{P}(N(dx) = 1)}{dx} = \lim_{U \downarrow \{x\}} \frac{\mathbb{P}(N(U) = 1)}{|U|},$$

assuming that the limit exists. The intensity tells us how likely it is that a point x belongs to the random set Z. In the case of the Poisson process, this definition of intensity coincides with the one given above. If a point process is stationary, then $\lambda(x) = \lambda(0)$ for any x, that is the intensity is constant.

The *second-order product density* is defined for $x \neq x'$

$$\lambda_2(x, x') = \frac{\mathbb{P}(N(dx) = 1, N(dx') = 1)}{dxdx'} = \lim_{U \downarrow \{x\}, U' \downarrow \{x'\}} \frac{\mathbb{P}(N(U) = 1, N(U') = 1)}{|U||U'|}.$$

It quantifies how likely it is to find points of Z at two locations. By definition, $\lambda_2(x, x') = \lambda_2(x', x)$. For a Poisson process, $\lambda_2(x, x') = \lambda(x)\lambda(x')$ because the number of points in two disjoint subsets are independent. If a point process is stationary, then $\lambda_2(x, x') = \lambda_2(x - x', 0)$, that is the second-order product density is a function of relative position $x - x'$. In the isotropic case, it is a function of the distance $||x - x'||$.

The second-order product density appears in the conditional intensity given that there is a point of Z at x:

$$\begin{aligned} \mathbb{P}(N(dx') = 1 \mid N(dx) = 1) &= \frac{\mathbb{P}(N(dx) = 1, N(dx') = 1)}{\mathbb{P}(N(dx) = 1)} = \frac{\lambda_2(x, x')}{\lambda(x)} dx' \\ &= \frac{\lambda_2(x, x')}{\lambda(x)\lambda(x')} \mathbb{P}(N(dx') = 1) \quad (x' \neq x), \end{aligned}$$

whereas for $x = x'$, obviously $\mathbb{P}(N(dx') = 1 \mid N(dx) = 1) = 1$. The ratio

$$g(x, x') = \frac{\lambda_2(x, x')}{\lambda(x)\lambda(x')}$$

is called the *pair correlation* function. If it is greater than one, the presence of a point at x increases the likelihood that another point is present at x' which is also called attraction. If the pair correlation is less than one, there is repulstion.

Mathematically, one avoids the problem of the existence of limits by defining integrated moments:

$$\Lambda(A) = \mathbb{E}(N(A)), \quad \Lambda_2(A_1 \times A_2) = \mathbb{E}(N(A_1)N(A_2)).$$

It is not difficult to show that Λ and Λ_2 are σ-additive measures on $(B, \mathcal{B}(B))$ and $(B^2, \mathcal{B}(B^2))$, respectively, and that for all positive functions $\psi : B \to \mathbb{R}$

$$\mathbb{E}(\sum_{i=1}^{\infty} \psi(x_i)) = \int_B \psi(x)\Lambda(dx)$$

and for all positive $\psi_2 : B^2 \to \mathbb{R}$

$$\mathbb{E}(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \psi_2(x_i, x_j)) = \int_{B\times B} \psi_2(x, x')\Lambda_2(dx, dx').$$

We then obtain the intuitive interpretation of $\lambda(x)$ if we assume that Λ has a density λ and that $\mathbb{P}(N(A) \geq 2) = o(|A|)$ as $|A| \to 0$. The second assumption is needed in order that $\mathbb{E}(N(A)) \sim \mathbb{P}(N(A) = 1)$ as $|A| \to 0$. Conditions for these two assumptions are complicated, and we ignore this difficulty.

Similarly λ_2 is obtained as the density of the measure $\Lambda_2(A_1 \times A_2) - \Lambda(A_1 \cap A_2)$ (we have to subtract the part of Λ_2 which is concentrated on the diagonal $x = x'$, otherwise the density does not exist).

The integrated conditional intensity is – by the formulae above for $\mathbb{P}(N(dx') = 1 \mid N(dx) = 1)$ – given by

$$\begin{aligned} \Lambda(A \mid x) &= \mathbb{E}(N(A) \mid x \in Z) = \int_A \mathbb{P}(N(dx') = 1 \mid N(dx) = 1) \\ &= 1_A(x) + \int_A \frac{\lambda_2(x', x)}{\lambda(x)} dx' = 1_A(x) + \int_A g(x', x)\lambda(x')dx'. \end{aligned}$$

(The first term comes from $x' = x$, the second from $x' \neq x$). Mathematically, it can be defined by a factorization

$$\Lambda_2(A_1 \times A_2) = \int_{A_2} \Lambda(A_1 \mid x)\Lambda(dx).$$

In the stationary case, $\Lambda(A \mid x) = \Lambda(A - x \mid 0)$. If the point pattern is also isotropic, it is sufficient to consider $\Lambda(A \mid 0)$ for $A = B(0, r) = \{x; ||x|| \leq r\}$, the ball with center at the origin and radius r. The so-called *K-function* is defined as

$$K(r) = \frac{\mathbb{E}(N(B(0, r)) \mid 0 \in Z) - 1}{\lambda} = \frac{1}{\lambda^2} \int_{B(0,r)} \lambda_2(||x||)dx = \int_{B(0,r)} g(||x||)dx.$$

For the Poisson process, we have $K(r) = r^d|B(0,1)|$. Knowing K, we can obtain the pair correlation function g by differentiating. For instance for $d = 2$ we have

$$g(r) = \frac{1}{2\pi r}K'(r).$$

The pair correlation function describes the character of dependence in a point pattern. It gives however not a full characterization, and there are examples of stationary point patterns with $g \equiv 1$, but the realizations are visually quite different from a Poisson pattern, see Baddeley und Silverman, Biometrics 40, 1984.

This led to the consideration of other characteristics of point patterns. The two most important are the "empty space function" F and the "nearest neighbor function" G. The empty space function is the cumulative distribution function of the distance from a fixed position $x \in B$ to the nearest point in Z. Because this distance is less than r if the ball $B(x,r) = \{x'; ||x' - x|| \leq r\}$ contains at least one point of Z, we have

$$F_x(r) = \mathbb{P}(N(B(x,r)) > 0).$$

For a stationary point pattern, this does not depend on x and we drop the subscript.

The nearest neighbor function is the cumulative distribution function of the distance from a point in Z to the nearest other point in Z. Assuming stationarity, this means that

$$G(r) = \mathbb{P}(N(B(0,r)) > 1 \mid 0 \in Z).$$

We look at the event of having more than one point in $B(0,r)$ because the origin is by conditioning always a point of Z.

For the stationary Poisson pattern $F = G$, and in dimension 2 we have

$$F(r) = G(r) = 1 - \exp(-\lambda\pi r^2).$$

Comparing F and G with the values for a Poisson pattern reveals which type of dependence is present: F indicates whether there are gaps in the pattern, whereas G describes the clustering of points at small distances.

4.3 Estimation of moments and other characteristics

We consider the situation where we have observed a realization of the point pattern in some observation window W, that is the set $Z \cap W$ consisting of all points $x_i \in Z$ which also belong to W. We want to account for the possibility that there are unobserved points of Z outside W, that is the domain B of the point pattern is (much) larger than W.

4.3.1 Estimation of the intensity

In the stationary case, the intensity is constant and has the simple estimator

$$\widehat{\lambda} = \frac{N(W)}{|W|}.$$

Its variance is

$$\mathrm{Var}\left(\widehat{\lambda}\right) = \frac{\Lambda_2(W \times W) - \Lambda(W)^2}{|W|^2}.$$

Nonparametric estimation of the intensity in the non-stationary case is possible if it varies slowly. We can choose a kernel k, i.e. a probability density k on $\mathbb{R}^d$, which has mean zero, support in $B(0,1)$ and its maximum at the origin. We then estimate $\lambda(x)$ by

$$\widehat{\lambda}(x) = \frac{1}{\int_W k((x-x')/h)dx'} \sum_{i; x_i \in W} k((x-x_i)/h)$$

where h is a bandwidth. For a point x with distance at least h from the boundary, it holds that

$$\mathbb{E}(\widehat{\lambda}(x)) = h^{-d} \int k((x-x')/h)\lambda(x')dx' = \int k(u)\lambda(x-hu)du.$$

If λ is twice differentiable, the bias is therefore of the order h^2. The variance of $\widehat{\lambda}$ can be expressed with the second moment product density λ_2. For points near the boundary, the bias of $\widehat{\lambda}$ is of the order h, and one has to use asymmetric kernels to reduce it.

In a Bayesian approach, one would put a prior distribution on the set of intensities which then acts as a regularizer. We will come back to this idea briefly when we discuss Cox point processes.

4.3.2 Estimation of the second moment

We consider here only the stationary and isotropic case where $\lambda_2(x, x') = \lambda_2(||x - x'||)$. We first discuss estimation of the K-function. The following estimator replaces the mean in the definition by an average.

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} N(W)} \sum_{i=1}^{N(W)} (N(B(x_i, r) \cap W) - 1) = \frac{1}{\widehat{\lambda} N(W)} \sum_{i=1}^{N(W)} \sum_{j \neq i} 1_{[\|x_i - x_j\| \leq r]}.$$

Because of the unobserved points outside W this estimator is severely biased. A simple alternative uses a reduced window and takes the average only over those x_i where $B(x_i, r) \subset W$. This makes however not full use of the available observations. A better way is to compensate for unobserved points by giving weights greater or equal to one to observed pairs (x_i, x_j):

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda} N(W)} \sum_{i=1}^{N(W)} \sum_{j \neq i} 1_{[\|x_i - x_j\| \leq r]} w_{ij}$$

where for $d = 2$

$$w_{ij} = \frac{2\pi \cdot \|x_j - x_i\|}{\text{Length of } (\{x|\, \|x - x_i\| = \|x_j - x_i\|\} \cap W)}$$

is the inverse of the probability that a random point with distance $||x_j - x_i||$ from x_i falls into W. One can show that this estimator is ratio unbiased.

Theorem 4.1 *If the point pattern is stationary and isotropic, then*

$$\mathbb{E}(\widehat{\lambda^2 K}(r)) = \lambda^2 K(r)$$

We omit the proof.

In principle, we can estimate λ_2 by taking the derivative of $\widehat{K}$. In contrast to K, the estimate $\widehat{K}$ is typically not smooth and thus one has to smooth it before taking the derivative. This is essentially the same as using the following kernel estimator:

$$\widehat{\lambda_2}(r) = \frac{1}{2\pi r} \sum_{x_i \in W} \sum_{x_j \in W; x_j \neq x_i} \frac{k(||x_i - x_j|| - r)}{|W \cap (W - (x_j - x_i))|}.$$

4.3.3 Estimation of F and G

We restrict ourselves to the stationary and isotropic case in $d = 2$ dimensions. The obvious estimators of F and G,

$$\begin{aligned} \widehat{G}(r) &= \frac{1}{N(W)} \sum_{i=1}^{N(W)} 1_{[N(B(x_i,r) \cap W) > 1]} \\ \widehat{F}(r) &= \frac{1}{|W|} |\{x \in W | N(B(x,r) \cap W) > 0\}| = \frac{|\cup_i\, B(x_i,r) \cap W|}{|W|} \end{aligned}$$

are again severly biased because of the unobserved points outside W.

A better method to estimate G uses an analogy to censored observations in survival analysis. We define for $x_i \in Z \cap W$

$$T_i = \min_j \|x_i - x_j\|, \quad C_i = \min_{x \notin W} (\|x - x_i\|).$$

Because points outside W are unknown, we cannot observe T_i, only $U_i = \min(T_i, C_i)$ together with the indicator $\Delta_i = 1_{[T_i \leq C_i]}$ whether T_i has been observed. The same situation arises in survival analysis where T_i is the survival time of patient i and C_i is the censoring time, the time when the study finishes or patient i drops out of the study. The *Kaplan-Meier* estimate estimates the distribution function of the survival time T_i based on censored observations $U_i = \min(T_i, C_i)$. Using it in the situation of point patterns, it gives the estimator

$$\widehat{G}(r) = 1 - \prod_{i; t_i \leq r} \left(1 - \frac{|\{j; u_j = t_i, \Delta_j = 1\}|}{|\{j; u_j \geq t_i\}|}\right).$$

Similar ideas can be used for the empty space function F. There, we have for any $x \in W$ the "censored" observation $U(x) = \min(T(x), C(x))$ together with the "censoring indicator" $\Delta(x) = 1_{[T(x) \le C(x)]}$ where $T(x) = \min_{x_j \in Z}(||x - x_j||)$ is the distance to the nearest point of the pattern and $C(x) = \min_{x' \notin W}(||x' - x||)$ is the distance to the boundary of W. Then the product in G becomes the exponential of an integral

$$\widehat{F}(r) = 1 - \exp\left(- \int_0^r \frac{|\{x \in W \mid u(x) = s, \Delta(x) = 1\}|}{|\{x \in W \mid u(x) \ge s\}|} ds \right).$$

4.4 Models with dependence

4.4.1 Cox point patterns

This is a hierarchical model where the point pattern is an inhomogeneous Poisson point pattern with an unobserved random intensity $Y(x)$. For instance, if Z is the point pattern of positions of plants of a certain species, $Y(x)$ can be interpreted as the environmental condition (light, humidity, nutrients) at x. In the Bayesian setting, we obtain a Cox model as the marginal distribution of Z if we put a prior on the intensity.

The distribution of Z for a Cox model is obtained by averaging the conditional distribution of Z given Y with respect to the distribution of Y. In particular

$$\lambda(x) = \frac{\mathbb{E}(\mathbb{P}(N(dx) = 1 \mid Y))}{dx} = \frac{\mathbb{E}(Y(x))dx}{dx} = \mathbb{E}(Y(x))$$

or

$$\lambda_2(x, x') = \frac{\mathbb{E}(\mathbb{P}(N(dx) = 1, N(dx') = 1 \mid Y))}{dx dx'} = \mathbb{E}(Y(x)Y(x')).$$

This implies that the pair correlation function is

$$g(x, x') = \frac{\mathbb{E}(Y(x)Y(x'))}{\mathbb{E}(Y(x))\mathbb{E}(Y(x'))} = 1 + \frac{\mathrm{Cov}\,(Y(x), Y(x'))}{\mathbb{E}(Y(x))\mathbb{E}(Y(x'))}.$$

One can also show that Z is stationary (and isotropic) iff Y is stationary (and isotropic). This is somewhat counter-intuitive because a single realization of Z looks exactly like a realization of a non-stationary Poisson pattern. However, if Y is stationary, then deviations of Y from the mean occur randomly and are statistically the same everywhere.

For a concrete model, we can choose a stationary Gaussian process model for $(\log Y(x))$ with one of the covariance functions discussed in Chapter 2, or $Y(x) = \sum_i k(x, x_i')$, where $\{x_1', x_2', \ldots\}$ is a stationary Poisson pattern and $k \ge 0$ satisfies $\int k(x, x')dx' < \infty$ for all x. In both cases, we can compute the mean and covariance function of Y and thus also of Z.

4.4.2 Neyman-Scott models (Cluster models)

This model explains clusters of close-by points in Z by the existence of some unobserved common parents or cluster centers from where the observed points originated. The parents form a stationary Poisson point process $Z' = \{x'_1, x'_2, \ldots\}$ with constant intensity λ_0. Each parent x'_i has M_i children which are located at positions $x_{ij} = x'_i + D_{ij}$ $(j = 1, 2, \ldots M_i)$ where both M_i and D_{ij} are i.i.d.. We then observe the positions of all children, but not those of the parents

$$Z = \cup_{i=1}^{\infty} \cup_{j=1}^{M_i} \{x_{ij}\}.$$

The information which points of Z have a common parent is also lost.

Although the intuitive idea is quite different from Cox models, the two classes have a non-empty intersection: If the number M_i of children is Poisson, then the Neyman-Scott model has the same distribution as a Poisson process with the random intensity

$$Y(x) = \lambda \sum_i f(x - x'_i)$$

where f is the density of the D_{ij}. This can be seen as follows: For each i the points x_{ij} form an inhomogeneous point pattern with intensity $\lambda_0 f(x - x'_i)$ (compare Lemma 4.1). Moreover the superposition of independent Poisson point patterns is again a Poisson point pattern (because the sum of independent Poisson variables is again Poisson).

The Neyman-Scott model is stationary, and one can compute first and second moments. One obtains

$$\lambda = \lambda_0 \mathbb{E}(M_i)$$

and

$$\lambda_2(x) = \lambda^2 + \lambda_0 \mathbb{E}(M_i(M_i - 1)) \int f(x - u) f(-u) du.$$

See for instance Section 5.3 in Stoyan et al.

4.4.3 Models with inhibition

Such models aim to have a minimal distance r_0 between any two points of Z. This can be done in different ways.

Sequential inhibition assumes a bounded domain B: The first point is distributed uniformly on B and the n-th point is distributed uniformly on $\{x; x \in B, ||x - x_1|| > r_0, \ldots, ||x - x_{n-1}|| > r_0\}$. The procedure stops when it is not possible to position more points.

Another model starts from a stationary Poisson point pattern $\{x'_1, x'_2, x'_3, \ldots\}$ with intensity λ_0 and eliminates one of the two points of any pair with $||x'_i - x'_j|| \leq r_0$. In order to have a unique definition, we attach to every x'_i an independent uniform random number U_i on $(0, 1)$ and we eliminate all points x'_i for which there is an x'_j with $||x'_i - x'_j|| \leq r_0$ and $U_j > U_i$. It is easy to

see that this procedure eliminates in some cases more points than necessary to keep the minimal distance.

The second model is stationary, and one can compute the first and second moments, see Stoyan et al. p. 164. The intensity λ is monotone increasing in the intensity λ_0 of the Poisson pattern that one starts with, and in the limit $\lambda_0 \to \infty$ one obtains $\lambda = 1/|B(0, r_0)|$. If we place at each point of Z a ball with radius $r_0/2$, we obtain a random arrangement of nonoverlapping balls in $\mathbb{R}^d$. The coverage is however not very dense. Because a large set A contains about $\lambda|A|$ balls, the fraction of A covered by the balls is about $\lambda|B(0, r_0/2)| = 2^{-d}$.

4.4.4 Gibbs models

We assume again that the domain B is bounded. As remarked earlier, a point pattern can then be specified by the probabilities $p_n = \mathbb{P}(N(B) = n)$ and by the densities f_n for putting n points $x_1, \ldots, x_n$ in B. A Gibbs model assumes a special form of p_n and f_n. For simplicity, we only discuss models with pairwise interactions. This means that for some potentials $\Phi_1 : B \to \mathbb{R}$, $\Phi_2 : B \times B \to \mathbb{R}$, Φ_2 symmetric, f_n and p_n take the following form

$$f_n(x_1, \ldots, x_n) = \frac{1}{M_n} \exp\left(- \sum_i \Phi_1(x_i) - \sum_{i<j} \Phi_2(x_i, x_j) \right),$$

where

$$M_n = \int_{B^n} \exp\left(- \sum_i \Phi_1(x_i) - \sum_{i<j} \Phi_2(x_i, x_j) \right) dx_1 \cdots dx_n$$

and

$$p_n = \text{const.} \frac{M_n}{n!} \ (n > 0), \quad p_0 = \text{const.}.$$

For a valid model we must therefore have

$$\frac{1}{\text{const.}} = 1 + \sum_{n=1}^{\infty} \frac{M_n}{n!} < \infty.$$

The simplest example of such a model is the Strauss process where

$$\Phi_1(x) = \theta_1, \quad \Phi_2(x, x') = \theta_2 1_{[||x-x'|| \le r_0]}.$$

For $\theta_2 > 0$, the more pairs of points with distance less than r_0, the smaller is f_n, that is we have a model with repulsion. For $\theta_2 = \infty$ we obtain another model with inhibition. For $\theta_2 < 0$ it seems that we have a model with attraction. However, it is not difficult to see that in this case the sum which defines const. is infinite: The attraction is so strong that the model would like to put infinitely many points in the domain B.

One possible solution is to introduce a repulsion at very small distances

$$\Phi_2(x,x') = \begin{cases} \theta_2 > 0 & \|x-y\| < \delta \\ \theta_3 < 0 & \delta \le \|x-y\| < r_0 \\ 0 & \|x-y\| \ge r_0 \end{cases}$$

Another possibility is to define p_n independently of f_n so that they sum to one.

The reason behind the above definition of p_n in Gibbs models is that it implies the following formula for the "probability that $Z = \{x_1, \ldots, x_n\}$":

$$p_n n! f_n(x_1, \ldots, x_n) = \text{const.} \exp\left(- \sum_i \Phi_1(x_i) - \sum_{i<j} \Phi_2(x_i, x_j) \right).$$

(Note that M_n cancels, and $n!$ arises on the left because the indexing of points in Z is arbitrary). This form is consistent with the physical argument that the probability of a configuration is proportional to the exponential of minus the energy of the configuration divided by the temperature (compare the same argument for lattice models). Mathematically, the "probability that $Z = \{x_1, \ldots, x_n\}$" has to be understood as the Radon-Nikodym density with respect to the Poisson pattern with constant intensity 1 where this "probability" is constant for all n and all $x_1, \ldots, x_n$.

If $\Phi_2(x,x') = 0$ for $||x - x'|| > r_0$, then the point pattern has a spatial Markov property. The so-called Papangelou conditional intensity is the conditional probability that a point x belongs to Z given that $Z = \{x_1, \ldots, x_n\}$ everywhere on $B \cap dx^c$:

$$\begin{aligned} \lambda(x \mid x_1, \ldots, x_n) &= \frac{\mathbb{P}(N(dx) = 1 \mid Z = \{x_1, \ldots, x_n\} \text{ outside } dx)}{dx} \\ &= \frac{p_{n+1}(n+1)! f_{n+1}(x, x_1, \ldots, x_n)}{p_n n! f_n(x_1, \ldots, x_n)}. \end{aligned}$$

In this conditional intensity, the whole pattern outside x is given whereas the conditional intensity that we defined through the second-order product density is based only on the information that another point of the pattern is at some point $x' \neq x$. For a pairwise interaction potential, we obtain

$$\lambda(x \mid x_1, \ldots, x_n) = \exp\left(-\Phi_1(x) - \sum_j \Phi_2(x, x_j) \right),$$

all other terms cancel. This intensity depends only on those points x_i of the pattern which are "r_0-close to x" which is the spatial Markov property.

4.5 Estimation for parametric models

For a Poisson point pattern with a parametric intensity $\lambda_\theta(x)$, the likelihood based on an observation $Z \cap W = \{x_1, \ldots, x_n\}$ in a window W is equal to

$$\exp\left(- \int_W \lambda_\theta(x) dx \right) \prod_i \lambda_\theta(x_i).$$

Heuristically, we multiply the probabilities $\mathbb{P}_\theta(N(dx) = 0) = 1 - \lambda_\theta(x)dx = \exp(-\lambda_\theta(x)dx)$ over all $x \neq x_i$ with the probabilities $\mathbb{P}_\theta(N(dx_i) = 1) = \lambda_\theta(x_i)dx_i$. We can therefore use maximum likelihood or Bayesian methods to estimate θ.

In practically all other cases, the likelihood is not available explicitely. This makes inference for point pattern by maximum likelihood or Bayesian methods difficult. The method of moments is simpler. If we have a closed form expression of the K-function in dependence of the parameter θ, then we can match this function to the nonparametric estimate $\widehat{K}$ from Section 4.3.2:

$$\widehat{\theta} = \arg\min_\theta \int_{r_1}^{r_2} \left| \widehat{K}(r)^\alpha - K_\theta(r)^\alpha \right| w(r)dr.$$

Here, the range $[r_1, r_2]$, the exponent α and the weight function w can be chosen freely. Often, one takes $\alpha = 1/2$ because this stabilizes the variance of $\widehat{K}(r)$.

Finally, for Gibbs models we can maximize the following analogue of the pseudo-likelihood function from Section 3.3.1

$$\exp\left(- \int_W \lambda_\theta(x \mid x_1, \ldots, x_n)dx \right) \prod_{i=1}^n \lambda_\theta(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

where $\lambda(x \mid x_1, \ldots, x_n)$ is the Papangelou conditional intensity defined in the previous Section. Because this formula assumes implicitly that there are no points outside W, the estimated will be biased. If $\Phi_2(x, x') = 0$ for $||x - x'|| > r_0$, one should presumably use the reduced window of points with distance at least r_0 from the boundary.