Thesis for the Degree of Licentiate of Engineering

Spatial Mixture Models with Applications in Medical Imaging and Spatial Point Processes

Anders Hildeman

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2017

Spatial Mixture Models with Applications in Medical Imaging and Spatial Point Processes
Anders Hildeman

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31 772 1000

Author e-mail: `hildeman@chalmers.se`

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Göteborg, Sweden 2017

Spatial Mixture Models with Applications in Medical Imaging and Spatial Point Processes

Anders Hildeman

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Finite mixture models have proven to be a great tool for both modeling non-standard probability distributions and for classification problems (using the latent variable interpretation). In this thesis we are building spatial models by incorporating spatially dependent categorical latent random fields in a hierarchical manner similar to that of finite mixture models. This allows for non-linear prediction, better interpretation of estimated model parameters, and the added possibility of addressing questions related to classification.

This thesis consists of two papers. The first paper concerns a problem in medical imaging where substitutes of computed tomography (CT) images are demanded due to the risks associated with X-radiation. This problem is addressed by modeling the dependency between CT images and magnetic resonance (MR) images. The model proposed incorporates multidimensional normal inverse Gaussian distributions and a spatially dependent Potts model for the latent classification. Parameter estimation is suggested using a maximum pseudo-likelihood approach implemented using the EM gradient method. The model is evaluated using cross-validation on three dimensional data of human brains.

The second paper concerns modeling of spatial point patterns. A novel hierarchical Bayesian model is constructed by using Gaussian random fields and level sets in a Cox process. The model is an extension to the popular log-Gaussian Cox process and incorporates a latent classification field in order to handle sudden jumps in the intensity surface and to address classification problems. For inference, a Markov chain Monte Carlo method based on the preconditioned Crank-Nicholson MALA method is suggested. Finally, the model is applied to a popular data set of tree locations in a rainforest and the results show the advantage of the proposed model compared to the log-Gaussian Cox process that has been applied to the very same data set in several earlier publications.

Keywords: Spatial statistics, Point processes, Finite mixture models, Bayesian level set inversion, Substitute CT, Gaussian fields, Non-Gaussian

List of appended papers

Paper I **A. Hildeman**, D. Bolin, J. Wallin, A. Johansson, T. Nyholm, T. Asklund, and J. Yu.
Whole-brain substitute CT generation using Markov random field mixture models.
Preprint

Paper II **A. Hildeman**, D. Bolin, J. Wallin, J. Illian.
Level set Cox processes.
Preprint

My contribution to the appended papers:

Paper I: I participated in the development of the model, conducted the analysis, drafted the manuscript, and after consultation produced the final manuscript. I implemented the code jointly together with J. Wallin and D. Bolin.

Paper II: I co-developed the model together with the other authors of the paper. I implemented the code, conducted the analysis and wrote drafted the manuscript by myself. The theoretical results of appendix A and B were achieved jointly together with J.Wallin and D.Bolin. I produced the final manuscript after consultation with the co-authors.

Acknowledgements

I would like to thank my supervisor David Bolin for introducing me to the interesting field of spatial statistics, for constantly taking the time when I have questions even though you have so many other responsibilities, and for your vast and broad knowledge. Jonas Wallin for your many ideas and quick comprehension of new problems. Jun Yu for great collaboration and for your hospitality during my visit at Umeå University. Also Fekadu, Jianfeng, Ottmar, and the MRI group for the hospitality during the same visit. Adam Johansson and Tufve Nyholm for answering my questions concerning medical imaging. Janine Illian for your great insight in point process theory, how to communicate it, and the importance of interdisciplinary work.

Thank you Efthymios Karatzas and Milo Viviani for our "secret" project and friendship. Olle Nerman and Marco Longfils for our collaboration with the projects at the consultancy group. Medet for wrestling sessions and John for long walks.

Thank you also to people I have had the fortune to teach together with: Sandra, Sebastian, Fanny, Andreas, Johan T, Johan J, and Reimond. I think that our effort has helped many students to a greater understanding, confidence, and interest in the uncertain and strange world of randomness. Thanks also to the members, and former members, of the spatial statistics group: Henrike, Claes, Tuomas, Kaspar, and Anna-Kaisa.

Last, but certainly not least, I would like to thank Karin for your constant support, my family (Bodil, Leif, Arild, Kristina, Ingela, and Bumbum) for raising me, and all my friends.

Anders Hildeman
Gothenburg, August 14, 2017

List of abbreviations

LGCP	log-Gaussian Cox process
GRF	Gaussian random field
MRF	Markov random field
GMM	Gaussian mixture model
MCMC	Markov chain Monte Carlo
MALA	Metropolis adjusted Langevin algorithm
MRI	Magnetic resonance imaging
CT	Computed tomography
CSR	Complete spatial randomness
pdf	probability density function
pmf	probability mass function
cdf	cumulative density function
ML	Maximum likelihood
EM	Expectation-Maximization

Contents

1 Introduction

The thesis you are currently holding in your hand (or reading in the soothing light of your screen) is a work made up of two articles in the field of spatial statistics. Although the focus of them are quite different they share the common notion of spatial modeling through mixture models with spatially dependent class probability distributions.

The first paper concerns a prediction problem in medical imaging where the usage of possibly dangerous X-radiation could be reduced by instead utilizing magnetic resonance imaging and statistical analysis. The paper proposes modeling images from magnetic resonance imaging (MRI) and computed tomography (CT) jointly utilizing a spatial model. By learning the parameters of the model from available medical data, prediction is possible based on conditional distributions given only MRI images.

The second paper introduces a spatial point process model that is able to model several classes of point patterns observed on separate and unknown partitions of the observational window. It is particularly useful in settings where some categorical and unknown covariate introduces bias in the standard point process models. It can also be used to handle classification problems for point pattern data. A well known data set of point locations of trees in a tropical rain forest is analyzed in order to show the strengths of the model. Markov Chain Monte Carlo methods are developed for the proposed model and theoretical questions regarding the consistency of finite dimensional model approximations, which are required for practical inference, are addressed.

In order to set the stage for the presentation of the articles we need to know the background and main concepts on which the effort was based. The remainder of this chapter is devoted to a brief introduction to the field of spatial statistics. Chapter 2 introduces the important concept of a random fields, Chapter 3 introduces the basics of spatial point processes, Chapter 4 finite mixture models, Chapter 5 statistical inference and Monte Carlo simulations, and finally Chapter 6 summarizes the content of the two papers.

1.1 History

Spatial statistics is a subfield of statistics that grew out of demands in the industrial sectors of the early 19:th century. The purpose: to draw conclusions or aid in decision making based on observed spatial data. The word *spatial* means that data can be compared using geometrical concepts such as distance and direction. The methodology originated, basically independently, from the fields of forestry, agriculture, and mining (Gelfand et al., 2010). In agriculture they studied the yield of cereal and recognized how spatial variations in yield on a field could be attributed to soil constituents or other covariates. Stochastic

models were needed that could explain these variations. In forestry they studied the location of trees in forests and how these were distributed. Effects such as repulsion between trees due to competition over sunlight and other resources, and spatial dependency due to pollination paths and seed dispersal demanded new modeling methods. Lastly, the mining engineers wanted to predict the prevalence of certain minerals in the ground based on samples at specific points.

Spatial data can be characterized in to three main categories:

- Data sampled on a continuous spatial domain.

  Between any two points in (continuous) space, $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, there is an infinite number of other points. The data consists of values at some of these points. The analysts interest is how these measurements relate to the values on the entire spatial domain. Examples of such data sets are surface air temperature and water salinity.

- Data sampled on a discrete spatial domain.

  The spatial domains only has a countable number of points, and in between two of them there are only a finite number of other points. The data consists of values at some of these points. Example of such data sets are digital images (that are made up of a discrete set of pixels) or experimental designs with "blocked" regions.

- Spatial point pattern data.

  For point patterns, the location of events are studied. Here the question of interest is not the values at points in space but at which points in space that some event occurred. That is, the data is a countable collection of points spread out over a continuous spatial region. Typical examples of point patterns are locations of trees in a forest or location of robberies in a city. See Chapter 3 for further explanation.

1.2 Purpose

Spatial statistical analysis of data is usually needed to answer one or more of the following questions:

- What are the values at unobserved points in space? (Spatial prediction / Kriging)
- What are the parameters values of our model? (Model estimation)
- Are our model assumptions reasonable? (Model validation)

Spatial prediction, here meaning prediction of values at unobserved points given the values at some observed ones, was of interest to the South African mining engineer Danie Gerhardus Krige that first formalized the problem. In spatial statistics, spatial conditional prediction is hence referred to as *kriging* as a homage to him. Often kriging predictions are not just point values, but rather conditional distributions given the observed data. This is more informative and point estimates such as the expected value, median, or mode can then be acquired from the probability distribution. Additionally, estimates of the uncertainty such as the standard deviation or interquartile range will give important information about the prediction error.

Model estimation is the act of fitting a model to the observed data. This is typically needed in order to draw conclusions about the underlying process that generated the spatial data. For instance, a parametric model representing tree growth in a forest might have a parameter representing the repulsive effect between trees, an effect that might exist due to the competition for sunlight among neighboring trees. Estimation of that particular parameter give information about to what extent that repulsive effect is present among that particular specie of trees. Estimating model parameters is also usually needed in order to perform kriging.

Model validation assesses a models ability to explain the observed data. Since any conclusions drawn are based on some model assumptions, it is important to assess whether these assumptions are reasonable given the observed data. If the model does not explain the data well, the kriging estimates and model estimation will not be useful.

In order to perform meaningful spatial analysis some model of spatial dependence is often assumed. The model will usually be simplistic in its nature in order to make model estimation reliable and computationally feasible. The true unknown model on the other hand is not necessarily simple and some degree of model misspecification will often be present. This is the constant balancing act between what is possible and what is the truth. Closing this gap is one of the main aims of research in spatial statistics. Hopefully the papers of this thesis has helped closing this gap at least somewhat.

2 Random fields

In statistics, conclusions are drawn based on incomplete information using concepts from probability theory. Probability theory concerns processes where the outcome of an action is not determinstic, i.e. the same action can result in different outcomes under exactly the same surrounding conditions. We will call such an action an *experiment* and the outcome of the experiment a *realization*. A random variable is a mapping between a realization and a real value, i.e.

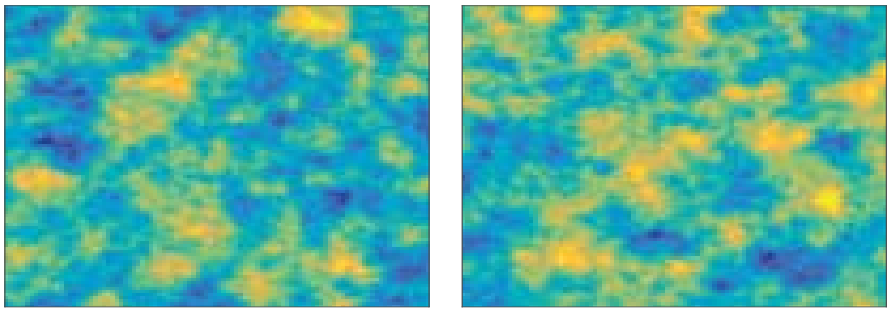


Fig. 1: Two realizations of the same stationary Gaussian random field on a bounded domain in $\mathbb{R}^2$.

$X : \Omega \to \mathbb{R}$, where X is the random variable, $X(\omega)$ a real value, $\omega \in \Omega$ a realization, and Ω is the set of all possible realizations. A random field is a mapping between a realization and a, possibly infinite, set of random variables, $\mathrm{X}(\boldsymbol{s}, \omega)$. Here $\boldsymbol{s}$ denotes a point in space and can be defined either on a bounded or unbounded spatial domain, $\mathcal{D}$.

One can think of a realization of a random field as a function mapping each point, $\boldsymbol{s} \in \mathcal{D}$ to a value, i.e. a random field is a random function with the domain $\mathcal{D}$. An example of two different realizations of the same random field on a bounded and continuous domain in $\mathbb{R}^2$ can be seen in Figure 1. Note how the two images show similar qualities even though they are completely different. As was mentioned in Chapter 1, spatial statistics concerns analysis of data observed on a spatial domain. In other words, the data can be seen as observations (or partial observations) of a realization of a random field.

A random field can have a discrete spatial domain, $\boldsymbol{s} \in \{\boldsymbol{s}_i\}_{i=1}^N$, or a continuous spatial domain, $\boldsymbol{s} \in \mathcal{D} \subseteq \mathbb{R}^d$. We will refer to a random field on a spatially discrete domain as a *spatially discrete random field* and the contrary as a *spatially continuous random field*. Likewise, the image of the random variables, $X(\boldsymbol{s})$, (all possible values attainable) at a point $\boldsymbol{s}$ can also be continuous or discrete. We will refer to a random field where $X(\boldsymbol{s})$ can only take on a discrete number of values for any $\boldsymbol{s}$ as a *discrete random field*. The dependence on a realization from the sample space will usually be omitted, i.e. $\mathrm{X}(\boldsymbol{s}, \omega) = \mathrm{X}(\boldsymbol{s})$.

2.1 Potts model

A particular type of random field that will be used in Paper I is the Potts model (Wu, 1982). It can be seen as a random field that is both discrete in space and in value. Hence it can be viewed as a finite collection of random variables,

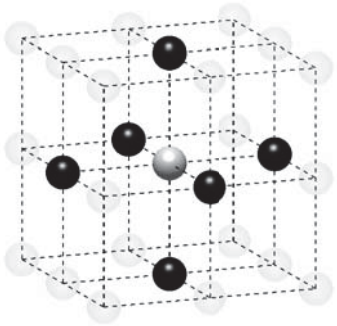


Fig. 2: A first order neighborhood structure on a regular lattice in three dimensions.

$\{X_i\}_{i=1}^N$, where $X_i = X(\boldsymbol{s}_i)$. It is defined through the conditional probability

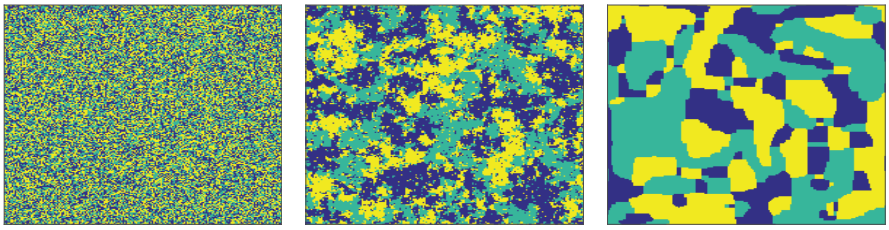
$$\mathbb{P}\left(X_i = k|\boldsymbol{X}_{-i}\right) = \frac{exp\left(-\alpha_k - \sum_{l=1}^K \beta_{kl} f_{il}(\boldsymbol{X}_{-i})\right)}{W(\alpha, \boldsymbol{\beta}, \boldsymbol{X}_{-i})}, \tag{1}$$

where $\boldsymbol{X}_{-i} = \{X_j\}_{j \neq i}$ denotes the set of all random variables except the i:th, and f_{il} denotes the number of points in the *neighborhood* of $\boldsymbol{s}_i$ that has the value l. The value of $X(\boldsymbol{s}_i)$ can be referred to as the class that $\boldsymbol{s}_i$ belongs to in a particular realization. The β-parameters controls the amount of attraction/repulsion between points of classes. The α-parameters control the marginal probabilities of the classes for any point $\boldsymbol{s}_i$.

The concept of a neighborhood defines which point in space that are connected to a specific point. This needs to be defined for every point in order to establish a Potts random field. An example of a first order neighborhood on a regular lattice in three dimensions can be seen in Figure 2 where the white ball denotes a point, $\boldsymbol{s}_i$, and the black balls corresponds to the neighborhood of $\boldsymbol{s}_i$, that is, the points that has the smallest euclidean distance to $\boldsymbol{s}_i$.

The Potts model in Paper I is used to model classification of points on a three dimensional lattice grid. There $X_i = 2$ would correspond to point $\boldsymbol{s}_i$ being a member of class 2. Figure 3 show three realizations of a Potts random field where the first having no spatial interaction, $\beta_k = 0$, the second having an attractive effect, $\beta_k = 1$, and the third figure having an even greater attractive effect, $\beta_k = 10$. As can be seen, the β_k parameters control the average size of the class regions.

The conditional probability of the random variable X_i in (1), $\mathbb{P}\left(X_i|\boldsymbol{X}_{-i}\right) = \mathbb{P}\left(X_i|\boldsymbol{X}_{\mathcal{N}_i}\right)$, does only depend on its neighbors, $\{\mathrm{X}_i\}_{i \in \mathcal{N}_i}$. Here, $\mathcal{N}_i$ denotes



(a) $\beta = 0$ (b) $\beta = -1$ (c) $\beta = -10$

Fig. 3: Example of realizations of a 3-class Potts field using three different values of the attraction parameter.

the neighborhood of $\boldsymbol{s}_i$. That is, given the values of all neighbors, the value at a specific point is completely independent of the remaining point in the random field. This is known as the local Markov property and the Potts model is hence a Markov random field (Winkler, 2003, chapter 3). The Markov property is used in Paper I in order to acquire a more computationally efficient prediction algorithm.

2.2 Gaussian random fields

A Gaussian random field (GRF) is a continuous random field such that any finite set of points on the spatial domain has a joint Gaussian distribution. A multivariate Gaussian distribution can be characterized by the mean value and covariance matrix. Likewise, a Gaussian random field can be characterized by the mean- and covariance-functions. The mean function, $\mu(\boldsymbol{s}) = \mathbb{E}[X(\boldsymbol{s})]$, is a first-order characteristic (only dependent on one point) describing the expected value at $\boldsymbol{s}$. The covariance function is a second-order characteristic (dependent on two points) describing the dependency between two points by their covariance, $\mathbb{C}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathbb{E}[(X(\boldsymbol{s}_1) - \mu(\boldsymbol{s}_1))(X(\boldsymbol{s}_2) - \mu(\boldsymbol{s}_2))]$. Often it is easier to work with a centered Gaussian random field, i.e. $\mu(\boldsymbol{s}) \equiv 0$. Such a field can easily be attained by subtracting the mean function from the original random field.

A common assumption used in Paper II is that of a stationary covariance function. This means that the covariance function is only dependent on the difference between two points, i.e. $\mathbb{C}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathbb{C}(0, \boldsymbol{s}_2 - \boldsymbol{s}_1)$. Since the dependency structure of a Gaussian random field is completely determined by the covariance function, a stationary covariance function will lead to a stationary Gaussian random field (if the GRF is centered).

If the covariance between points diminishes with distance such that spatial regions separated by a large enough distance behave as if they were from different realizations of the random field, the field is *ergodic*. If a random field satisfy

this ergodicity property it is possible to estimate the covariance function from only one realization of the random field as long as the *practical correlation range* is much smaller than the observed domain. Here, practical correlation range is loosely defined as the distance between two points at which the dependency is negligible.

2.2.1 Matérn covariance structure

In applications, the amount of data is limited and estimating an arbitrary covariance function is often not reliable. Therefore it is often necessary to assume that the covariance function is from a parametric class with only a small number of parameters. One such popular parametric class of stationary and isotropic covariance functions used frequently in spatial statistics are the Matérn class (Matérn, 1986; Handcock and Stein, 1993). This class can be parametrized by the marginal variance σ^2, the smoothness ν, and the practical correlation range, r. The smoothness parameter, ν, controls the differentiability of the covariance function at the origin. For a Gaussian random field this controls the smoothness of the realizations of the field itself in the sense that the field is almost surely Hölder continuous and ν is the corresponding Hölder constant. The practical correlation range, r, corresponds to the spatial distance between two points at which the correlation is 0.1. The marginal variance, σ^2, is the variance of the marginal distribution of $X(\boldsymbol{s})$ for any $\boldsymbol{s}$. The covariance function is defined as

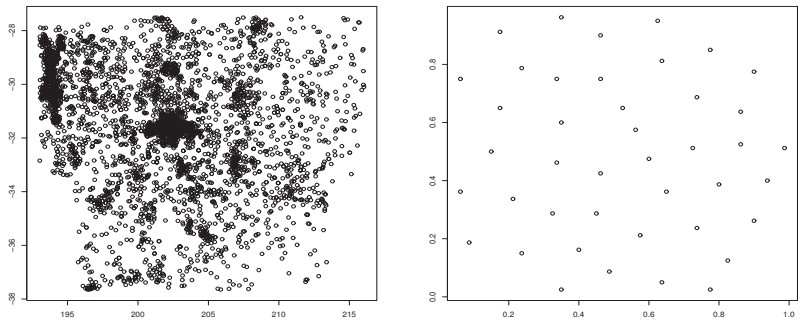
$$\mathbb{C}(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^\nu K_\nu(\kappa h),$$

where $\kappa = \frac{\sqrt{8\,\nu}}{\mathrm{r}}$, $h = \parallel \boldsymbol{s}_2 - \boldsymbol{s}_1 \parallel$, Γ is the gamma function, and K is the modified Bessel function of the second kind.

The popularity of the Matérn class of covariance functions is due to that it allows for a great flexibility in dependency structures while still keeping a small amount of easily interpretable parameters.

3 Spatial point processes

A spatial point pattern is a countable set of locations, $Y = \{x_1, x_2, ...\}$, $x_i \in \mathcal{D}$ for some continuous spatial domain, $\mathcal{D}$. Often the point pattern is observed in an observational window, W. That is, the point pattern exists in $\mathcal{D}$ but is only observed in $W \subset \mathcal{D}$. Here one can consider two types of point patterns, the finite and the infinite. The infinite point patterns consists of an infinite number of point and are typically defined on an open domain such as $\mathbb{R}^d$. Practically it is impossible to observe such a pattern on its full domain and the observational window will include only a strict subset of all the points. The finite point



(a) (b)

Fig. 4: (a) Observations of galaxies in the Shapley supercluster. (b) Location of centres of observed biological cells observed under optical microscopy.

patterns on the other hand will have a bounded spatial domain including all of the point locations. Hence, in applications, the observational window is more often the same as the spatial domain for finite point patterns.

Examples of spatial point patterns can be for instance the locations of galaxies in the Shapley super cluster as seen in Figure 4a (Drinkwater et al., 2004; Baddeley and Turner, 2005), locations of cell centres observed under optical microscopy as seen in Figure 4b (Baddeley and Turner, 2005; Ripley, 1977). Also the data set from paper II containing locations of the tree specie *Beilshmiedia Pendula* in a tropical rainforest is a point pattern, see Figure 11.

A point pattern can be defined as a counting measure, N, on the spatial domain $\mathcal{D}$, where $N(A)$ counts the number of points in the spatial region $A \subseteq \mathcal{D}$. A point process is a stochastic model for point patterns. Since a point pattern could be described as a counting measure, a point process can be described as a random counting measure. Statistical analysis of point patterns corresponds to analyzing the properties of the point process from which the point pattern was generated.

3.1 The Poisson process

One of the most important cases of a point process is the *homogeneous Poisson process*. This is the model of *complete spatial randomness* (CSR), i.e. unstructured point patterns. That is, locations of points are independent of each other and the number of point in a specific region is proportional to the measure of that region. In this work we will only consider spatial domains in euclidean

spaces with corresponding Lebesgue measure, $\mathcal{L}$. Hence, for a homogeneous Poisson process, $\mathbb{P}(N(A)) \propto \mathcal{L}(A)$, where $A \subseteq W$. Moreover, the counting measure of CSR is Poisson distributed with $N(A) \sim Pois(\lambda \cdot \mathcal{L}(A))$, where $\lambda \geq 0$.

Historically, most methods in point process statistics have been focused on differentiating between CSR and structured patterns. A structured pattern can either differ from CSR due to interaction between points and/or by spatial dependencies due to some available or unknown covariates. The differences between the two effects lies in the generative process more than the actual observed pattern. For example, assume that a seed is planted on a spatial region. The seed grows in to a tree and then a new seed is planted. If the second seed is planted too close to the, now fully grown, first tree it will be shaded and the possibility of growing in to a large tree is inhibited. This is an example of a repulsive interaction between points. On the other hand, the possibility to grow in to a large tree might also depend on the topography and soil constituents of the spatial region. Planting a seed close to a stream or in a dry desert will affect its chances of growing in to a large tree as well. This is an example of a dependency on covariates.

Definition 3.1 (Intensity measure)**.** *The intensity measure,* Λ*, of a point process is a deterministic measure defined as the expected value of the random counting measure, i.e.*

$$\Lambda(A) = \mathbb{E}\left[N(A)\right], A \in \mathcal{D}\,.$$

If Λ is absolutely continuous with respect to the Lebesgue measure it can be described by the *intensity function* λ as $\Lambda(A) = \int_A \lambda(\boldsymbol{s}) d\,\boldsymbol{s}$. For the homogeneous Poisson process, $\lambda(\boldsymbol{s}) = \lambda, \forall\, \boldsymbol{s} \in \mathcal{D}$, i.e. a constant intensity function. The *inhomogeneous Poisson process* is a point process which behaves as a homogeneous Poisson process on infinitesimal subregions of $\mathcal{D}$. Due to the additivity of Poisson distributed random variables, the counting measure of an inhomogeneous Poisson distribution is Poisson distributed as $N(A) \sim Pois(\Lambda(A))$. Hence, a Poisson process (homogeneous or not) is characterized solely by the intensity measure. The Poisson process assumes no interaction between points, a feature inherited from the CSR model due to the additivity of Poisson random variables. If an intensity function exists, covariates can be included in the model by letting λ be a function of the covariate values. In Paper II a log linear relation is consider where $\log \lambda(\boldsymbol{s}) = \sum_j B_j(\boldsymbol{s})\beta_j$ for covariates B_j and coefficients β_j.

3.2 Cox processes

A further extension of the Poisson process is that to a Cox process. A Cox process is defined using a positive random field, $\lambda(\boldsymbol{s})$. Conditioned on a given realization of $\lambda(\boldsymbol{s}$, the point process is an inhomogeneous Poisson process with λ as its intensity function. Hence, the model is doubly stochastic in the sense that it defines a generative process based on two steps of random objects. Cox processes can also be considered as Bayesian models of a Poisson processes where the latent intensity field, λ is given a prior probability distribution. A popular Cox process model is the log-Gaussian Cox process (LGCP) for which $\lambda(\boldsymbol{s}) = e^{\mathrm{X}(\boldsymbol{s})}$, where X is a Gaussian random field. The popularity of the LGCP model is due to its marriage of point processes with the well-studied Gaussian random fields.

The Cox process can model not only the effects of covariates but also clustering effects (attractive interaction effects). Regions with higher intensity in λ would correspond to cluster regions. A Cox process is however not enough to characterize all point processes. For instance, repulsive interaction effects, such as trees competing over sunlight, cannot be explained by such a model.

3.3 Characterizations of point processes

Just as moments, pdf's, and cdf's characterizes a random variable, point processes can be characterized by some statisticas. Generally, point processes are usually characterized by different kinds of measures and function valued statistics.

The moment measures characterizes the k-th order moments of $N(A)$ analogously to how the intensity measure was defined.

Definition 3.2 (k-th moment measure)**.** *The k-th moment measure of a spatial point process is defined as*

$$\mu^{(k)}(A_1 \times ... \times A_k) = \mathbb{E}\left[N(A_1)...N(A_k)\right].$$

Here, $A_1, ..., A_k$ are arbitrary and possibly equal spatial regions in $\mathcal{D}$.

Note that $\Lambda(A) = \mu^{(1)}(A)$ and higher order moment measures can characterize interaction behavior. Just as with random fields, the concepts of stationarity and isotropy are defined for point processeses as well.

Definition 3.3 (Stationarity)**.** *A point process with counting measure $N(A)$ is said to be stationary if,*

$$\mathbb{P}\left(N(A_1) = n_1, ..., N(A_k) = n_k\right) = \mathbb{P}\left(N(B_1) = n_1, ..., N(B_k) = n_k\right),$$

for any finite set $\{A_l\}_{l=1}^k$ where $B_l = A_l + t = \{\boldsymbol{s} : \boldsymbol{s} - t \in A_l\}$, i.e. a translation of A_l.

Definition 3.4 (Isotropy)**.** *A point process is isotropic if,*

$$\mathbb{P}\left(N(A_1) = n_1, ..., N(A_k) = n_k\right) = \mathbb{P}\left(N(B_1) = n_1, ..., N(B_k) = n_k\right),$$

for any finite set $\{A_l\}_{l=1}^k$ *where* $B_l = \{\boldsymbol{s} : R_\theta\, \boldsymbol{s} \in A_l\}$*, i.e. a rotation with angle* θ *of the points of* A_l *around the origin.*

The concept of ergodicity is also an important one. For an ergodic point process, the dependency between $N(A)$ and $N(B)$ will be neglible if the closest points in the two regions are sufficiently far away. This property means that if an ergodic point pattern is observed on a sufficiently large observational window, W, subregions far away from each other will have point patterns distributed as if from different realizations of the underlying point process. The implications being that, as long as the observational window is large enough, one point pattern can be used for statistical analysis of the underlying process parameters since it acts as having observed several independent realizations of point patterns.

For point processes it is sometimes useful to work with a probability distribution conditioned on that one of the points are located at some specific location. It shifts the viewpoint from "an absolute frame of reference outside the process under study, to a frame of reference inside the process" (Daley and Vere-Jones, 2003). Such conditional probability distributions for point processes are known as Palm distributions.

Definition 3.5 (Palm distribution)**.** *The Palm distribution is a probability distribution of a point process conditioned on that one of the points of a realization is located at a location* o*. For a stationary point process with intensity* λ*, the probability measure* $\mathbb{P}_o$ *of the Palm distribution is*

$$\mathbb{P}_o(Y \in \mathcal{A}) = \frac{1}{\lambda \mathcal{L}(W)} \mathbb{E}\left[\sum_{x \in Y \cap W} \mathbb{I}\left(Y - x \in \mathcal{A}\right)\right],$$

for some event $\mathcal{A}$*, some arbitrary spatial region* W *such that* $\mathcal{L}(W) > 0$*. Here,* $Y - x$ *denotes a translation of the points in the point pattern* Y *by* x*.*

We will denote the expectation with respect to the Palm distribution as $\mathbb{E}_o$ in contrast to the regular expectation with regards to the absolute frame of reference, $\mathbb{E}$.

In point process literature, some functional characteristics have been given particular attention. Here, functional characteristic refer to a function characterising aspects of the point process. Originally they were mostly used to test if point patterns behaved as the CSR model but it is nowadays common to use them to evaluate the goodness-of-fit where some observed point pattern

are compared to some null hypothesis model based on the functional characteristic of the two. In Paper II, a point pattern is compared to simulations from several assumed models. Evaluation of the models performance are based on the similarity of the functional characteristics between the real pattern and the simulated ones.

In the case of a stationary point process, Ripley's K-function (Ripley, 1977) (or estimates thereof) has been used extensively in order to investigate departures from complete spatial randomness.

Definition 3.6 (Ripley's K-function)**.** *For a point process with counting measure $N(A)$, the K-function is defined as,*

$$K(r) = \frac{1}{\lambda}\mathbb{E}_o\left[N(b(o,r) \setminus \{o\})\right],$$

where $b(o,r)$ is the ball with center in point o and radius r, and $\mathbb{E}_o$ is the expectation with respect to the Palm distribution with a point in o.

In words, $K(r)$ is the expected number of other points found inside a ball of radius r conditioned on that there is a point in the center of the ball. For the CSR model, $K(r) = b_d r^d$, where b_d is the value of the unit ball in $\mathbb{R}^d$ and d is the spatial dimension of the point pattern. Hence, by estimating the K-function from the point pattern it is possible to study the deviations from the theoretical K-function of the CSR model. For a point process with attractive spatial interaction (clustering), $K(r) > b_d r^d$. Likewise, a point process with repulsive spatial interaction (regularisation), $K(r) < b_d r^d$.

A variant of the K-function that represents the same information but is easier to interpret is Besag's L-function (Ripley, 1977, Besags comments),

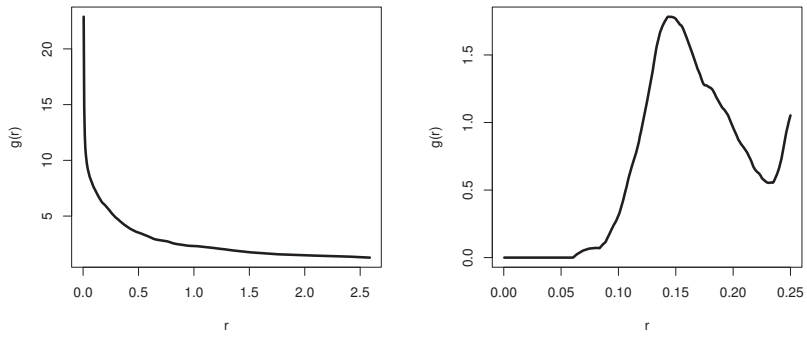
$$L(r) = \left(\frac{K(r)}{b_d}\right)^{1/d}.$$

The L-function is a modification of K such that for the CSR model $L(r) = r$ and estimations tend to be homoscedastic with respect to r. A further modification as $L^*(r) = L(r) - r$ transforms the L-function in to the centered L-function for which the CSR model would have $L^*(r) \equiv 0$. It is hence easier to interpret deviations from CSR.

The pair correlation function, $g(r)$, is another functional characteristic which relate to the K- function,

$$g(r) = \frac{K'(r)}{db_d r^{d-1} r},$$

where K' denotes the derivative of K. For the CSR model, $g(r) \equiv 1$. Values of $g(r)$ larger than 1 means that there are clustering effect at those distance



(a) (b)

Fig. 5: Estimated pair correlation functions. (a) Estimated g for the Shapley galaxy supercluster. (b) Estimated g for the cell data.

while $g(r) < 1$ mean that there are repelling effect. Typically, a point process might have attractive effects on some intervals and repulsive effects on others. Taking the example with tree locations, a repulsive effect exists for points very close to each other due to the competition for sun and the fact that the stem of the trees actually has a radius, however at medium distances there should be an attractive effect since the seed dispersal has a limited range. A realistic pair correlation function of such a point process would hence have a value lower than 1 for small r and then a value higher than 1 and finally a value approximately 1 for long distances. Figure 5 shows estimates of the pair correlation function for the two point patterns shown in Figure 4. The galaxy data set shows a clustering effect on short distances seen by $g(r) > 1$, while the cell data seem to be regularly spaced, seen by the peak above 1 at the range of $0.11 - 0.20$. This fits intuitively with the visual perception of the two point patterns seen in Figure 4.

Both the pair correlation function, K-function and L-function are second order characteristics since they are characterizing the behavior between two different points. The empty space function used in Paper II on the other hand characterizes first-order properties.

Definition 3.7 (Empty space function)**.** *The empty space function F is for a stationary point process with counting measure $N(A)$ defined as*

$$F(r) = \mathbb{P}\left(N(b(o,r)) > 0\right).$$

In words, $F(r)$ is the probability of observing a point in a ball of radius r centered at the origin, o. Of course, since it is assuming an isotropic and stationary point process, the origin can be interchanged with any point in the observational window. Once again, under CSR, $N(b(o,r)) \sim Pois(\lambda b_d r^d)$ and hence $F(r) = 1 - e^{-\lambda b_d r^d}$.

In the setting of Paper II we have used estimates of the L-, g-, and F-functions in order to compare our point pattern with simulations from the fitted models. In that setting we did neither assume isotropy nor stationary of the point process. However, we still expect the fitted model to yield estimated functions similar to the ones estimated from the actual point pattern. Hence, even though the interpretation of the functional characteristics is not clear in the non-stationary case, they can still be used for comparison. For details about estimating the functional characteristics mentioned above, see Illian et al. (2008).

4 Finite mixture models

A finite mixture model (Everitt and Hand, 1981) can be defined in two different but equivalent ways. Let us start by defining K classes; each associated with a random variable X_k with corresponding probability distributions D_k. Assume a random variable, Z, with probability distribution D_0, on a discrete sample space, $\{1, 2, ..., K\}$. The K different values that Z can assume corresponds to the K classes. The random variable Y will be distributed according to a finite mixture model if it is generated by first acquiring a realization z from Z, then assigning Y the value from a realization of X_z. Hence

$$Y = \sum_{k=1}^{K} \mathbb{I}(Z = k)\, X_k.$$

The finite mixture model can be viewed as a doubly stochastic model since it requires evaluation of random variables in two steps. If a probability density function (or probability mass function) exists, the mixture distribution can equivalently be defined by

$$f_Y(x) = \sum_{k=1}^{K} \pi_k f_k(x),$$

where f_Y is the pdf (or pmf) of Y, $\pi_k = \mathbb{P}(Z = k)$, and f_k is the pdf (or pmf) of X_k.

Typically the first definition is used when the properties of the latent variable Z is of interest, which is the case for classification problems. The second definition is more often used when a probability distribution should be approximated

by a set of simple ones. For instance explaining a multimodal distribution as a superposition of unimodal ones as in Figure 6.

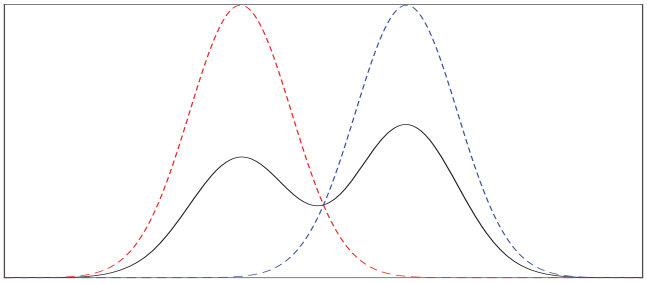


Fig. 6: Example of a finite mixture distribution (black) defined as the superposition of two Gaussian distributions (blue and red).

From here on out, finite mixture models will simply be referred to as mixture models.

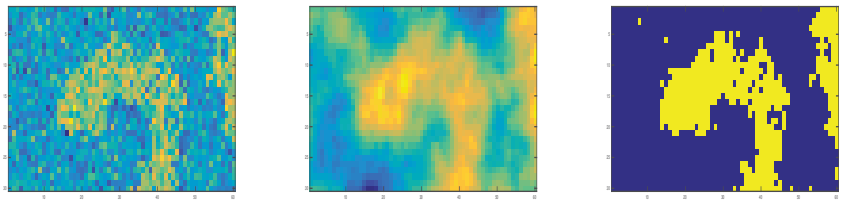
4.1 Spatial mixture models

In the papers of this thesis, mixture models are used in spatial models where Z is no longer a random variable but instead a random field, $Z(\boldsymbol{s})$. Likewise, X_k are no longer random variables but random fields as well, $X_k(\boldsymbol{s})$. In Paper I, a spatial mixture model was used to model the distribution of voxel values in medical images. A Potts model was used to model the latent classification for each voxel. Given this classification, each voxel was assigned a value from the distribution of the corresponding class.

In Paper II, a spatial finite mixture model was used to model the distribution of the intensity function of a Cox process. The latent classification field, $Z(\boldsymbol{s})$, was acquired from level sets of a Gaussian random field using the approach of Iglesias et al. (2016) and Dunlop et al. (2016). Compared to the Potts model, this model has the advantage that it defines a classification field in a continuous spatial domain.

Iglesias et al. (2016) and Dunlop et al. (2016) stated a probabilistic model for solutions to continuous geometric level set inversion problem observed with additive Gaussian noise. In a geometric level set inversion problem, some function known as the *level set function*, defines a partition of the spatial domain through its level sets. That is, $A_k = \{\boldsymbol{s} : c_{k-1} < \mathrm{X}(\boldsymbol{s}) \leq c_k\}$, where $\{A_k\}_k$ is the partition and $\{c_k\}_k$ are threshold values.

The aim of the inversion problem is to estimate the partitioning of the



(a) (b) (c)

Fig. 7: (a) Observed data corrupted by noise, Y. (b) Corresponding level set function, X. (c) Classification field.

domain given observations,

$$Y(\boldsymbol{s}_i) = \sum_{k=1}^{K} a_k \mathbb{I}\left(X(\boldsymbol{s}_i) \in]c_{k-1}, c_k]\right) + \epsilon_i,$$

where ϵ_i are Gaussian i.i.d random noise and a_k are parameters.

Figure 7 show the observed field, Y, the underlying (latent) field, X, and the classification field, Z, acquired from thresholding X in a realization of the level set model.

In paper II the Gaussian likelihood of Dunlop et al. (2016) is replaced by the more complicated scenario of point process data.

5 Inference

Statistical inference is the art of drawing conclusions based on the available data with the aid of some probabilistic model. In spatial statistics this is usually associated with estimating parameter values of a model or acquiring some prediction based on such parameters. Inference philosophies are divided in to either Bayesian or frequentistic.

From a frequentist's perspective there exists some true parameter values of the assumed model. The aim is to find the best estimate of these parameters given the observed data. Once the parameters have been estimated, prediction can be made using the model.

5.1 Maximum likelihood estimation using the EMG algorithm

Maximum likelihood (ML) estimation is a common frequentist approach to parameter estimation. The ML estimates are obtained as, $\hat{\Theta}_{ML} = \arg\max_\theta L(\Theta; \boldsymbol{x})$,

where L is the likelihood function, Θ are the parameters, and $\boldsymbol{x}$ the observed data. The ML estimators are consistent, asymptotically unbiased and asymptotically most efficient among all estimators (Olofsson and Andersson, 2012). Sometimes it is possible to find explicit analytical solutions to ML estimators but often numerical methods are required. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is an iterative method to find a local maximum of the likelihood function. This method is commonly used for finding $\hat{\Theta}_{ML}$ when no analytic solution is achievable due to missing information such as latent variables. Mixture models could be viewed as latent models where the classification values are the missing information, hence the EM algorithm is very often utilized to find ML estimates of mixture models.

The EM algorithm starts with some initial parameters values, $\Theta^{(0)}$. Then, in each iteration, an E-step is performed followed by an M-step. The E-step corresponds to computing the expected value of the latent variables given the current parameter values. The ensuing M-step maximizes the likelihood conditioned on the latent variables being equal to their expectation found in the E-step. The method is shown to converge for a very general class of problems (Wu, 1983).

In paper I the EM algorithm was not applicable since the M-step was not computationally feasible to compute, or even approximate, even with the latent variables known. Instead the EM gradient (EMG) algorithm of Lange (1995) was utilized. This method is based on the same concept as EM but the M-step is replaced with one step of the Newton-Raphson method,

$$\Theta^{(i+1)} = \Theta^{(i)} + H^{-1}(\Theta^{(i)})\mathbb{E}\left[\nabla \log L(\Theta^{(i)})\right].$$

That is, the M-step is replaced by one step in an iterative optimization procedure. Any strict local maximum point of the likelihood locally attracts the EM and EMG algorithm at the same rate of convergence.

5.2 Bayesian inference

The maximum likelihood approach to parameter estimation yields a point estimate of the "true" value of the assumed model. An alternative is the Bayesian perspective where the parameters of the model are considered random themselves. A distribution of the parameters is chosen prior to the analysis of the data, which should include all known information about the parameters. This distribution is known as the *prior* distribution. Inference can later be drawn from the, so called, *posterior* probability distribution which is the probability distribution conditioned on the observed data. This is acquired using Bayes

theorem and thereof the name, *Bayesian inference*,

$$f(\Theta|\boldsymbol{X}=\boldsymbol{x}) = \frac{f(\boldsymbol{X}=\boldsymbol{x}|\Theta)f(\Theta)}{f(\boldsymbol{X}=\boldsymbol{x})} \propto f(\boldsymbol{X}=\boldsymbol{x}|\Theta)f(\Theta),$$

where f denotes pdf:s. Bayes theorem can be generalized to handle more abstract probability spaces where no pdf:s exist if instead stated in terms of probability measures (see e.g. Stuart (2010)), this is used in Paper II.

The prior probability is hence weighted by the likelihood function to acquire the posterior probability. The posterior probability distribution is not only a point estimate but a whole probability distribution. Hence, more information is given since questions about uncertainties in the parameter estimation can be answered as well. How to choose the prior distribution depends on what is known about the problem. If nothing can be assumed there are two philosophies, either to choose an uninformative prior or chose a prior that penalizes the complexity of the model (Simpson et al., 2017). The first philosophy will let the data explain more of the behavior in lack of known information. The second philosophy assumes that a simpler model is better since it is more easily understood and is less prone to overfitting. Hence, in lack of information indicating the opposite, the simpler model should be preferred.

5.3 Monte Carlo simulation

The distribution can be known up to a normalizing constant through Bayes theorem but acquiring the normalizing constant is not always possible. Obtaining it through explicit integration of the posterior pdf is usually impossible, and Θ does often include a large set of parameters and are therefore high dimensional. In particular, in spatial statistics, Θ often includes latent random fields which are very high dimensional. The high dimensionality makes numeric integration computationally infeasible but Monte Carlo (MC) integration is often a viable alternative.

Monte Carlo integration is a method of approximating expected values by simulating samples from the correct probability distribution and computing the sample mean as a proxy for the true expectation. For instance, the probability of finding $\Theta \in A$ can be written as an expectation and estimated using MC simulation as

$$\mathbb{P}\left(\Theta \in A|\boldsymbol{X}=\boldsymbol{x}\right) = \mathbb{E}\left[\mathbb{I}\left(\Theta \in A\right)|\boldsymbol{X}=\boldsymbol{x}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(x^{(i)} \in A\right),$$

where $x^{(i)}$ are sampled from the distribution of $\Theta|\boldsymbol{X}=\boldsymbol{x}$. Hence, the posterior distribution can be approximated arbitrarily well if it is possible to sample from it.

Of course, MC simulation is not only applicable for Bayesian inference. For instance in Paper I, MC simulation was used to approximate some expectations that were infeasible to compute explicitly.

5.3.1 The Metropolis-Hastings algorithm

A Markov chain Monte Carlo (MCMC) simulation is a MC simulation where the samples are distributed as a Markov chain with stationary probability distribution equal to the *target distribution*. The target distribution meaning the probability distribution of interest, typically a posterior probability distribution.

Hence, there exists dependence in between consecutive samples but if the dependency declines fast enough compared to the number of iterations of the simulation, the MCMC integration will produce consistent estimates of the true expectation.

The main archetype of MCMC algorithms is the Metropolis-Hasting (MH) algorithm. It is based on suggesting a new sampled value, y, distributed according to some given probability distribution conditioned on the most recently sampled value, $x^{(i-1)}$. This probability distribution is known as the *proposal probability distribution* and we denote its pdf as $q(y|x^{(i-1)})$. Then given y, y is chosen as the new sample, $x^{(i)}$, with a probability α. This trial is known as the *accept/reject step* and α is known as the *acceptance probability* and is given by

$$\alpha = \min\left\{\frac{f(y)}{f(x^{(i-1)})}\frac{q(x^{(i-1)}|y)}{q(y|x^{(i-1)})}, 1\right\}.$$

Here, y is the proposed new sample, $x^{(i-1)}$ is the sample from the prior iteration, and f denotes the pdf of the target distribution. Of course, f is often only known up to a normalizing constant but that is no problem since in the ratio the normalizing constants are being canceled out. The first ratio weights the probability of accepting y with how probable y is compared to $x^{(i-1)}$ with respect to the target distribution. However, since there might be a higher probability of realizing y conditioned on $x^{(i-1)}$ than the opposite in the proposal distribution, the second ratio is needed to balance this. If y is accepted in the accept/reject step, then the former sample value is used instead, $x^{(i)} = x^{(i-1)}$.

Due to the Markovian structure, the initial value of the samples, $x^{(0)}$, will affect the distribution at later iterations. The first couple of iterations can be highly dependent on $x^{(0)}$ and the iterations until the dependency on the initial value has become insignificant is known as the burnin phase. How many iterations that the Markov chain spends in the burnin phase varies upon the choice of initial value, the target distribution and the proposal distribution. It is important that the chain is run for sufficiently many iterations as to leave the burnin phase and generated enough samples outside of the burnin phase to

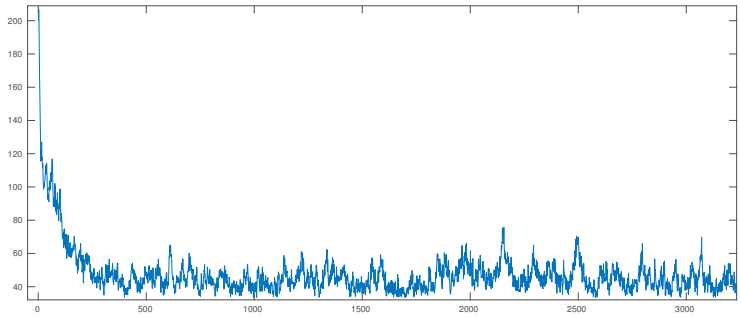


Fig. 8: Example of a parameter path during MCMC simulation.

be able to estimate the true expectation well. This is a common dilemma in Bayesian inference since it is not always obvious when the Markov chain leaves the burnin phase. Typically the burn in phase is identified visually using plots of the parameter paths for some of the parameters. When a parameter path no longer shows a clear trend, as it does in the beginning, it is considered that it has passed the burnin phase. Figure 8 shows an example of a parameter path during a MCMC simulation. By visual inspection we would conclude that the Markov chain passed the burnin phase after about 400 iterations. The samples from the burnin phase is then removed since it will bias the estimation.

An efficient MCMC chain should have as low dependence between consecutive samples as possible in order to make efficient use of the number of iterations available. This is known as quick mixing as compared to slow mixing where there are significant dependencies between samples in the Markov chain even when separated by a large number of iterations. Quick mixing requires small dependence on the prior sample in the proposal distribution while still allowing for a high acceptance probability. This is usually competing requirements that are hard to satisfy simultaneously.

A common proposal distribution for the MH algorithm is the Gaussian proposal centered at x, i.e. $q(y|x) \propto \exp\left(-\frac{(y-x)^T\Sigma^{-1}(y-x)}{2\delta}\right)$ for some chosen covariance Σ. Note that this is a symmetric proposal since $q(y|x) = q(x|y)$ and δ controls the stochastic step length. This is known as the random walk Metropolis-Hastings algorithm since the proposals would have behaved like a random walk if the accept/reject step had not been present.

5.3.2 Gibbs sampler

An important special case of the MH algorithm that even predates MH is the *Gibbs sampler*. Suppose that the samples are two dimensional, $x = [x_1, x_2]$. If the conditional probabilities are known, i.e. the distribution of $x_1|x_2$ and $x_2|x_1$, it is possible to use these conditional distributions as proposals. Hence the acceptance probability of the MH algorithm becomes $\alpha = \min\left\{\frac{f(x_1)}{f(x_2)}\frac{f(x_2|x_1)}{f(x_1|x_2)}, 1\right\} = \min\left\{\frac{f(x_1,x_2)}{f(x_1,x_2)}, 1\right\} = 1$. Since the acceptance probability is always 1, the accept/reject step is not necessary. Hence, if one samples first from $x_1^{(i)}|x_2^{(i-1)}$ and then from $x_2^{(i)}|x_1^{(i)}$ for each iteration, the corresponding sample path will be a realization of a Markov chain with stationary distribution equivalent to the target distribution. This works for x of arbitrary many dimensions.

It is also possible to mix the Gibbs sampler and MH algorithm such that two disjoint subsets of parameters for the target distribution are updated separately using the conditional distributions and the Gibbs sampler. Within each subset, a more general MH algorithm with arbitrary proposals could be used. This is known as Metropolis-within-Gibbs MCMC and is utilized in paper II.

5.3.3 MALA

The Metropolis adjusted Langevin algorithm (Roberts and Tweedie, 1996) is a method that, compared to the regular MH algorithm with symmetric Gaussian proposals, make use of the target distribution in designing the proposal distribution. This is achieved by making use of the gradient of the target pdf. Hence the extra cost of evaluating the gradient pays off in a more efficient MCMC sampler. It is based on the stochastic differential equation (SDE) of the Langevin diffusion process,

$$dX(t) = \Sigma\nabla \log f(X(t))dt + \sqrt{2}\Sigma^{\frac{1}{2}}dW(t), \tag{2}$$

where ∇ is the gradient operator with respect to the dimensions of $X(t)$, $W(t)$ is a Brownian motion, and Σ is the covariance operator of the proposal distribution. The solution to equation (2) has the target distribution as its stationary distribution. Hence, if the sample path of a Langevin diffusion would be available, taking samples at distances sufficiently far apart would correspond to sampling from the target distribution. The MALA algorithm uses the Euler-Maryuama method (Platen and Bruti-Liberati, 2010) to acquire a discretization of a sample path from the Langevin diffusion equation. However, the discretization introduces errors and an accept/reject step is hence necessary to enforce sampling from the target distribution.

The proposals generated from the regular random walk MH algorithm can be considered as being Euler-Maryuama time discretizations of a Brownian motion.

The Brownian motion does not have the target distribution as its stationary distribution and therefore the MALA algorithm will acquire fewer rejections for comparable steps lengths.

5.4 Crank-Nicholson MCMC

Cotter et al. remarked that the Euler-Maryuama scheme used for the MALA is not stable with respect to the step size and the number of dimensions of the random variable. With increased dimensionality, the step length needs to be decreased in order to keep a constant acceptance probability. That corresponds to a mixing of the MCMC chain that becomes slower with increased dimensionality. This can be a problem when approximating an infinite dimensional model by a finite dimensional approximation. Cotter et al. noticed that if the target probability measure, μ^Y, is absolutely continuous with respect to a Gaussian probability measure, μ_0, the SDE,

$$dX(t) = -\mathcal{K}\mathcal{Q}X(t)dt + \gamma\mathcal{K}\nabla \log f(X(t))dt + \sqrt{2\mathcal{K}}dW(t),$$

has the stationary probability measure μ_0 if $\gamma = 0$ and μ^Y when $\gamma = 1$. Here $\mathcal{K}$ can either be chosen as the covariance operator of μ_0 or the identity operator. $\mathcal{Q}$ is the precision operator of μ_0 and f is the Radon-Nikodym derivative $\frac{d\mu}{d\mu_0}$.

By discretizing this SDE using a Crank-Nicholson approximation on the linear part of the drift, stability is achieved and the discretization errors for a chosen step length are no longer dependent on the number of dimensions.

$$\left(I + \frac{1}{2}\mathcal{K}\mathcal{Q}\right) X(t_i) = \left(I - \frac{1}{2}\mathcal{K}\mathcal{Q}\right) X(t_{i-1}) + \gamma\mathcal{K}\nabla \log f(X(t_{i-1}))\delta + \sqrt{2\mathcal{K}\delta}\epsilon.$$

The choice of $\mathcal{K}$ should depends on if $(I + \frac{1}{2}\mathcal{L})$ can be efficiently inverted or if sampling from the prior distribution is straightforward. Setting $\mathcal{K}$ as the covariance operator corresponds to the case when sampling from the prior is possible and was utilized in Paper II.

Just as with MALA compared to random walk MH, choosing $\gamma = 1$ requires evaluation of the gradient but will lead to a higher acceptance probability.

The Crank-Nicholson MCMC scheme is particularly well-suited to spatial modeling including continuous Gaussian random fields since the infinite dimensional spatially continuous model has to be approximated by some finite dimensional ditto. The step length's invariance to the number of dimensions in the Crank-Nicholson MCMC algorithms is therefore important. Moreover, Gaussian fields clearly fulfill the requirement of a Gaussian prior needed for the method to be applicable. This was utilized in paper II in order to acquire an efficient posterior sampler.

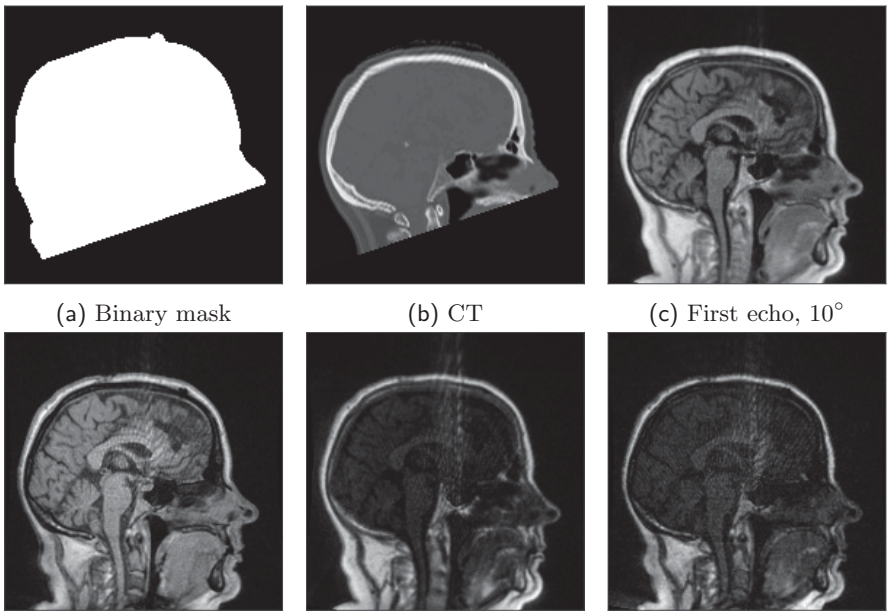
6 Summary of papers

6.1 Paper I: Whole-brain substitute CT generation using Markov random field mixture models

Computed tomography (CT) imaging is a technique for acquiring three dimensional internal images of electron density within living organisms as well as inanimate objects. The method relies on exposing the subject to X-radiation and measuring the attenuation as it passes through. Radiation therapy is another usage of X-radiation. Here, the ionizing property of the high-energy X-ray photons are used to damage cancerous tumors. In the dose planning of radiation therapy it is important to know how the subjects body will absorb X-radiation. That information is used to avoid damaging healthy tissue as much as possible while still exposing the tumor sufficiently. Such absorption properties can be acquired from a CT images and hence CT imaging is a vital tool in radiotherapy treatment. Another important application of CT imaging is for attenuation correction of PET images. In PET imaging, radiation is emitted by a tracer fluid inside the body of the subject and a PET scanner can sense this radiation and create images from it. In this way it is possible to follow the path of the tracer fluid as it passes through the body. Some of the radiation will however be absorbed by the body and the PET images need to be corrected for this attenuation in order to acquire reliable images. CT images are able to explain this attenuation and are hence an important tool in PET imaging as well.

A problem with CT imaging is that X-ray radiation is ionizing and hence there exists risks of damaging living tissue. This has triggered research in acquiring CT-equivalent information in other ways. Johansson et al. (2011) showed that it is possible to acquire a substitute CT (s-CT) image from magnetic resonance imaging (MRI) using statistical methods. More specifically, Johansson et al. (2011) modeled the voxels of a CT images and several MRI images with different flip angles and echo time jointly as independent realizations from a multidimensional Gaussian mixture model. The parameters of the model was learned from training data using a maximum likelihood estimator through the EM-algorithm. The s-CT images can then be acquired as the conditional mean given available MRI images. The classes of the mixture model could be thought of as different kinds of tissue. Hence, a separate distribution will be used to explain the joint behavior of the four MRI modalities and the CT image for each of the different tissue types. In this way it is possible acquire a non linear mapping from the four dimensional MRI space to the CT space. Figure 10 visualizes the different classifications of tissue types in a slice seen in profile of a human head.

Paper I extends on the Gaussian mixture model by adding two new concepts. First, the Gaussian distribution might be too restrictive and a more



(a) Binary mask (b) CT (c) First echo, 10°

(d) Second echo, 10° (e) First echo, 30° (f) Second echo, 30°

Fig. 9: A two dimensional profile slice of the three dimensional image of one of the subjects in the CT/MRI data. Binary data mask (panel a), CT image (panel b), four MRI UTE sequences (panels c-f).

flexible probability distribution could potentially explain the tissue behaviors better. For instance skewed probability distributions or distributions with a higher kurtosis could perform better for certain cases. For this reason, the Gaussian distribution is replaced by a normal inverse Gaussian (NIG) distribution, yielding a five dimensional NIG mixture model. Moreover, it is unrealistic to believe that the intensity distribution of the five dimensional voxels are independent of each other. Some spatial structure should exist in the images and hence voxels nearby should be dependent on each other. This dependence is captured by assuming a Potts model for the class memberships of the mixture model. Hence, the model of Paper I is a spatial model due to the distribution of the classification as a Potts spatial random field. However, conditioned on the class memberships there is no spatial dependency in the joint distribution of the five dimensional voxels.

The Potts model is given as a conditional probability distribution. The Hammersley-Clifford theorem states that the distribution of the class member-

ships can be stated as a neighbor Gibbs field (Winkler, 2003). Unfortunately, for a data set of realistic size, the normalizing constant is to demanding to compute and hence the probability measure of the joint distribution is only available up to a normalizing constant. This makes the likelihood function infeasible to evaluate and maximum likelihood-based parameter estimation can hence not be used. As an alternative we consider a pseudolikelihood, $\tilde{L}(\Theta; \boldsymbol{x})$, where the joint likelihood is approximated as a product of all conditional probabilities. Furthermore, even using the pseudolikelihood, the M-step of the EM-algorithm is infeasible to compute. Instead Paper I estimates the parameters using an EMG algorithm.

The proposed method is evaluated using cross-validation on a data set of brain scans from 14 different patients, see Figure 9 for an example. All four permutations of choosing Gaussian or NIG distributions on the random variables conditioned on the classes and choosing the Potts classification model or the original non spatial one are investigated. The model with Gaussian distribution and regular mixture model is considered to be the benchmark model since it corresponds to the model of (Johansson et al., 2011). The conclusions drawn from this cross-validation study is:

- The spatial classification model makes a significant difference in predictive ability.
- The NIG distribution did only show a consistent improvement when combined with the spatial classification model.

An example of classification for a slice of a head is shown in Figure 10. Classes correspnding to soft tissue, bone, air, and a class for bones and soft tissue mixed on a small scale are present in the figure.

6.2 Paper II: Level set Cox processes

A popular point process for modeling non-interacting point observations with varying spatial intensity is the log-Gaussian Cox process (LGCP). The latent Gaussian random field of the LGCP model is in practice assumed to have a simple parametrization that is both possible to estimate from available data and interpretable. A common assumption is that the covariance operator of the latent Gaussian field is a member of a parametric family of stationary operators and the mean field includes a finite number of fixed linear effects. Such a model is viable in many cases but the regularizing assumptions can also be too strong. An example of this can be seen in Figure 11 which shows a point pattern of observed locations of the tree *Beilschmiedia pendula* in a region of Barro Colorado island, Panama. The figure show a pattern that seem to be made up

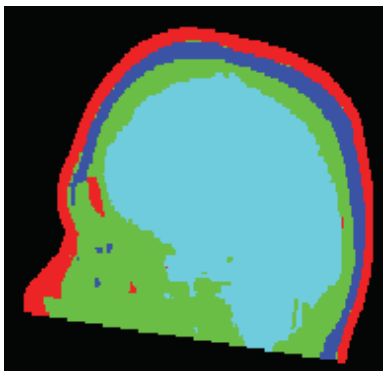


Fig. 10: Example classification using the proposed model. Here showing a two dimensional slice of a three dimensional classification field on one of the subjects in the CT/MRI data set. Each color denotes a class in the mixture model.

of two partitions of the observational window. One region of low intensity, and one region of high intensity with some spatial dependency structure.

Conditioned on knowing the spatial classification for the two distinct classes it would seem reasonable to model each region, separately, with LGCP models. This is exactly the idea of paper II where the LGCP model is extended with a latent classification field. The model assumes that there are an unknown partitioning of the spatial domain such that for each partition, the point observations are distributed according to a LGCP model with simple latent Gaussian random field structures, i.e. a Matérn covariance function. The proposed model is a Cox process where the logarithm of the intensity surface is distributed as a spatial mixture model between several classes of Gaussian random fields, i.e.

$$\log \lambda(\boldsymbol{s}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{s})\, \mathrm{X}_k(\boldsymbol{s}).$$

The spatially dependent classification probabilities, $\pi_k(\boldsymbol{s})$, are defined as a classification field generated by the level set approach described in Section 4.1. The model is named the *level set Cox process* (LSCP) due to this latent level set classification field.

The LSCP model is a latent Gaussian model since the intensity surface is completely defined by the realizations of latent Gaussian fields, one field for each mixture class as well as one field for the classification. Compared to the Potts model of paper I, the level set approach is defined on a continuous spatial domain which makes the LSCP model continuous in space.

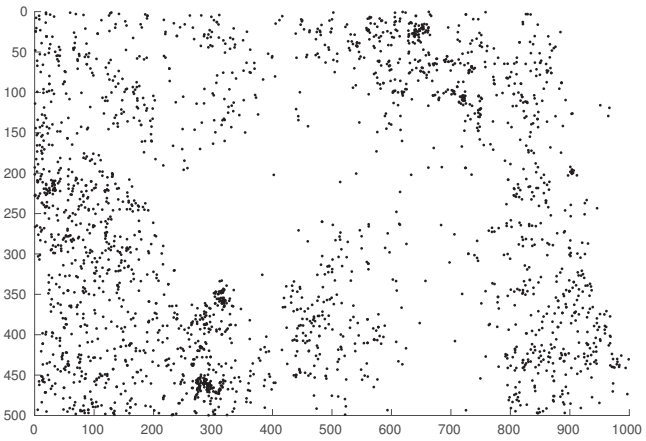


Fig. 11: Observations of the tree *Beilschmiedia pendula* on a 1000×500 square metres area of the Barro Colorado island of Panama.

Having a continuous point process model is important since point observations most often are observed in continuous space. However, to be able to do inference based on the data, some finite dimensional approximation is required. In paper II the observational domain, $\mathcal{D}$, is discretized in to a finite number of subregions on an equidistant lattice grid forming a partition of $\mathcal{D}$. It is shown that the posterior probability measures of the latent Gaussian fields of the finite dimensional model converges to the posterior measure of the continuous model under refinement of lattice grid.

Paper II proposes a Bayesian approach to statistical inference. The posterior marginal distributions of both parameters, latent Gaussian fields, and the intensity surface can be acquired by Monte Carlo simulations. A spectral approach using fast Fourier transforms together with the preconditioned Crank-Nicholson MCMC methods of Cotter et al. is proposed for efficient Monte Carlo based inference.

As an example, posterior inference using four different types of the LSCP model are compared on the tree locations of the Barro Colorado dataset with available covariate. These examples highlight the flexibility and potential of the model.

Bibliography

Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. URL `http://www.jstatsoft.org/v12/i06/`.

S.L. Cotter, G.O. Roberts, A.M. Stuart, and D. White. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster.

D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: Volume I: Elementary theory and methods*, volume 2. Springer, 2003. ISBN 0-387-95541-0.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

M.J. Drinkwater, Q.A. Parker, D. Proust, E. Slezak, and H. Quintana. The large scale distribution of galaxies in the Shapley Supercluster. *Publications of the Astronomical Society of Australia*, (21):89–96, 2004.

M.M. Dunlop, M.A. Iglesias, and A.M. Stuart. Hierarchical Bayesian level set inversion. *Statistics and Computing*, pages 1–30, 2016.

B.S. Everitt and D.J. Hand. *Finite mixture distributions*. Chapman and Hall, 1981. ISBN 0 412 22420 8.

A.E. Gelfand, P.J. Diggle, M. Fuentes, and P. Guttorp. *Handbook of spatial statistics*, volume 2. Taylor and Francis, 2010. ISBN 9781420072884.

M.S. Handcock and M.L. Stein. A Bayesian Analysis of Kriging. *Technometrics*, 35(4):403–410, 1993.

M.A. Iglesias, Y. Lu, and A.M. Stuart. A Bayesian level set method for geometric inverse problems. *Interfaces and free boundaries*, 18(2):181–217, 2016.

J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical analysis and modelling of spatial point patterns*. Wiley, 2008.

A. Johansson, M. Karlsson, and T Nyholm. CT substitute derived from MRI sequences with ultrashort echo time. *Medical Physics*, 38:2708–2714, 2011.

K. Lange. A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2): 425–437, 1995.

B. Matérn. *Spatial Variations*, volume 36. Springer-Verlag, 1986. ISBN 9780387963655.

P. Olofsson and M. Andersson. *Probability, statistics, and stochastic processes*, volume Second edition. Wiley, 2012. ISBN 9780470889749.

E. Platen and N. Bruti-Liberati. *Numerical solutions of stochastic differential equations with jumps in finance*, volume 64. Springer, 2010. ISBN 3642120571.

B.D. Ripley. Modelling spatial patterns. *Journal of the royal statistical society. Series B*, 39(2):172–212, 1977.

G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

D. Simpson, H. Rue, A. Riebler, T.G. Martins, and Sørbye S.H. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017.

A.M. Stuart. Inverse problems: A Bayesian perspective. *Acta numerica*, 19: 451–559, 2010.

Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer, 2003.

C.F.J. Wu. On the convergence properties of the EM algorithm. *The annals of statistics*, 11(1):95–103, 1983.

F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54(1):235–268, 1982.