Contributions to the statistical modelling of image data and spatial point patterns

Rasmus Waagepetersen
Department of Theoretical Statistics
University of Aarhus

Contents

Preface

This survey paper and the accompanying papers Møller & Waagepetersen (1996), Waagepetersen (1997*b*), Waagepetersen (1997*a*), and Møller, Syversveen & Waagepetersen (1996) constitute my Ph.D.-thesis in mathematical statistics.

The first part of the survey paper is mainly concerned with prior modelling in Bayesian image analysis. The subject of the second part is modelling and inference for spatial point patterns. Markov chain Monte Carlo (MCMC) has been an indispensable tool in my work, and an account of some basic notions and methods of MCMC is given in the final section of the survey paper. The papers Møller & Waagepetersen (1996) and Møller et al. (1996) are extensive, and I have therefore found it useful to let the survey paper contain sections with brief presentations of the main results of these papers.

I owe many thanks to my supervisor Jesper Møller for his careful, inspiring and enthusiastic guidance. I am also indebted to my wife Katrine for her patience and encouragement, and to my colleagues at the Department of Theoretical Statistics for stimulating discussions and helpful comments. Finally, I wish to thank Professor Hans R. Künsch and Professor Adrian Baddeley, and their colleagues at the Seminar für Statistik, ETH, and the Department of Mathematics, University of Western Australia, for the hospitality which I enjoyed during my stays abroad.

The work which led to this thesis was funded by the Faculty of Science, University of Aarhus.

Summary

In statistical image analysis much attention is devoted to construction of penalized likelihoods for point estimation of images. Parsimonious image parametrizations are rarely applicable, and likelihood estimation is consequently often encumbered with problems like overfitting and nonuniqueness of estimates. A penalizing term is therefore multiplied to the likelihood in order to move the maximum of the penalized likelihood away from the highly variable and "rough" likelihood estimates, towards more precise smoothed estimates.

For many commonly applied penalizing terms there are not both objective and practically applicable methods for choosing the smoothing parameter in the penalizing term. It is therefore relevant to analyze the residuals to check whether a suitable degree of smoothing has been applied. Summary statistics and tests to analyze residuals from penalized likelihood image segmentation are proposed and tried out on synthetic data in Waagepetersen (1997*a*).

Apart from the problem of choosing the smoothing parameter, it is typically difficult to assess uncertainty of penalized likelihood image estimates. These problems can in principle be solved within the Bayesian framework if a realistic prior modelling is available. The uncertainty is then given by the posterior, and prior parameters may be estimated either from the actual recorded image data by missing data methods, or from training data.

The construction of realistic prior image models is not straightforward. One contribution to the solution of this problem is Møller & Waagepetersen (1996) which introduce a class of image models denoted Markov connected component fields (MCCF's). These models are discrete-valued random fields on a lattice, and the probability of an image is proportional to a product of "interaction" functions evaluated on the maximal connected components in the image. Markov properties, extensions to infinite lattices, and other theoretical aspects are studied, and data examples demonstrate that a wide range of images can be modelled by MCCF's. The penalized Ising model is an example of a MCCF which is useful for modelling of vague prior information. Phase transition and a new simulation algorithm for this model is studied in Waagepetersen (1997*b*).

Bayesian inference concerning the intensity surface of an inhomogeneous Poisson process resembles Bayesian image analysis since the intensity surface may be regarded as a 2D image, and the observed point pattern as the noisy or transformed image. The aims and means of statistics for point patterns are, however, in general quite different from those of statistical image analysis. Interest is often focused on modelling of possible repulsion between points or clustering, and inference is mainly based on nonparametric second order summary statistics or parsimonious parametric models.

In Møller et al. (1996) a new class of parametric models for clustered point patterns is introduced. These models are Cox processes, where the random intensity surface is given by a log Gaussian random field, and the models are hence

denoted log Gaussian Cox processes (LGCP's). LGCP's provide a flexible class of models for clustered point patterns, where the clustering is due to an environmental heterogeneity, and LGCP's are furthermore appealing from a theoretical point of view. The product densities are e.g. given by simple expressions related to the mean and covariance of the Gaussian process, and this enables the construction of a third order summary statistic, and simple estimation procedures.

Inference concerning the unobserved intensity surface of a LGCP is possible by application of Markov chain Monte Carlo (MCMC). A Metropolis adjusted Langevin algorithm (MALA) is used to generate conditional simulations for Monte Carlo estimation of conditional means and variances of the Gaussian field and the intensity surface, given an observed point pattern. Prediction of unobserved parts of the point process is also possible, and the MALA can easily be modified so that a geometrically ergodic Markov chain is obtained.

The literature on analysis of nonstationary point patterns with interaction between points is not extensive. Ogata & Tanemura (1986) consider approximate maximum likelihood estimation for inhomogeneous Gibbs processes, and use the AIC criterion for comparison of fitted models. Baddeley, Møller & Waagepetersen (1997) study another class of models, where the observed point pattern is regarded as a nonstationary thinning of a stationary Markov point process. It is discussed how the K-function may be useful for analysis of such point patterns, and how MCMC may be applied to perform semiparametric inference.

1 Statistics for image data

1.1 Introduction

Image data appear in a wide range of contexts, and on a large variety of scales. Important sources of images are digitized photographs, electron microscopy, remote sensing, and medical imaging. Images of human organs are in the last case obtained by indirect methods such as ultrasound scanning, positron emission tomography (PET), and magnetic resonance (MR) scanning.

An example of a simple model for image data is the linear regression

$$Y = h(\theta) + \epsilon, \tag{1.1}$$

where $Y \in \mathbb{R}^I$ is the observed data, $\theta \in \mathbb{R}^I$ is a parameter representing the unobserved image, I is the index set of the digital image pixels, h is a linear mapping representing blur, and ϵ is *iid* Gaussian noise. Another example is PET imaging where θ represents the intensity surface of an inhomogeneous Poisson process, Y is a random field of Poisson variables, and $E(Y) = h(\theta)$, where h represents a complicated physical relation.

Image analysis tasks differ with respect to the amount of information which is built into the image parametrization. Noise filtering is a low level task, where the image parametrization is often based directly on the digital image representation, as in the model (1.1). Image segmentation is an intermediate level task, where it is desired to partition the image into a number of homogeneous regions belonging to a finite set V of region types. For this purpose one may e.g. in the model (1.1) take $\theta \in V^I$ instead of $\theta \in \mathbb{R}^I$. In a high level task like object recognition, the parametrization may represent type, outline, and position of the objects.

Images are usually very complex quantities for which simple and lowdimensinal parametrizations are seldom applicable. The problem of solving $z = h(\theta)$ with respect to θ for a given z, is furthermore often ill-posed (i.e. the mapping h is either not injective, or a small change in z may lead to a very different solution). Maximum likelihood estimation is therefore usually encumbered with problems like overfitting, multimodality, and high variability of the estimates, and pure likelihood methods are rarely encountered in statistical image analysis. One of few exceptions is Rudemo & Stryhn (1994), where image segmentation is considered as a two-dimensional change point problem, and asymptotic results are established for the distribution of the maximum likelihood estimator. Strong assumptions on the nature of the image are, however, required in order to obtain a sufficiently simple parametrization. The usual approach in statistical image analysis is to base inference on a penalized likelihood where a penalizing term is added to the log-likelihood to stabilize the estimation procedure. This approach is also known in physics and astronomy as regularized solutions to ill-posed inverse problems. Penalized likelihood is in the statistical image analysis literature best known as maximum a posteriori (MAP) estimation since the penalizing term is often interpreted as a prior model for the image θ.

If for example a quadratic penalizing term is used for estimation in the model (1.1), the penalized log-likelihood

$$-\frac{1}{2\sigma^2}\|Y - h(\theta)\|^2 + \lambda\theta C\theta^t \tag{1.2}$$

is obtained, where σ^2 is the noise variance, $\lambda > 0$ is the smoothing parameter, and C is a positive definite matrix. With an appropriate value of λ, the penalizing term serves to bias the maximum of the penalized likelihood away from the "rough" likelihood estimate, towards estimates which possess a suitable degree of smoothness according either to prior expectations, or some more objective criteria. A number of data driven methods like cross-validation and minimization of mean square error for choosing the smoothing parameter are reviewed in e.g. Thompson, Brown, Kay & Titterington (1991) and Künsch (1994). These methods are for computational reasons not practically applicable if e.g. θ is discrete, or a nonquadratic smoothing term is used, and for many commonly applied penalizing terms there do not exist both objective and practically applicable methods for choosing the smoothing parameter, see e.g. the discussion in Dinten, Guyon & Yao (1991). It therefore seems worthwhile to study the residuals from image analyses in order to check whether a suitable degree of smoothing has been applied. In Waagepetersen (1997*a*) (see also section 1.3) a rather simple image segmentation problem is considered, and various methods for analysis of residuals are proposed and tested on synthetic image data. Ad hoc criteria for choosing the smoothing parameter are discussed in Ripley (1988), Künsch (1994), and Waagepetersen (1997*a*).

Another problem related to penalized likelihood estimation of images is, apart from the choice of smoothing parameter, assessment of uncertainty of the estimates. Except for special cases like (1.2), the addition of the penalizing term to the log-likelihood makes derivation of exact or asymptotic results very complicated, and bootstrap methods may also be computationally prohibitive if extensive computations are required to obtain the penalized likelihood estimate.

Penalized likehood utilize prior information in the sense that one may discard values of the smoothing parameter which do not yield suitable smooth estimates, according to prior belief. The smoothing parameter can, however, strictly speaking, not be determined a priori since the appropriate value of the smoothing parameter depends on the noise level of the data. If e.g. σ^2 is large in (1.2), then smaller values of λ are needed to obtain the required smoothness, than if σ^2 is small. The penalizing term is, regarded as a prior model, furthermore usually a poor representation of prior knowledge in the sense that typical realizations are far from realistic image scenes. In Tjelmeland & Besag (1996) and Møller & Waagepetersen (1996) it is exemplified that even though a satisfactory MAP-estimate is obtained, it is not advisable to assess uncertainty by interpreting the penalized likelihood as a posterior distribution.

An alternative to penalized likelihood is to consider the image a realization of a stochastic process, and then assess uncertainty within a genuine Bayesian framework. In order that the posterior provides a reliable representation of the uncertainty, it is necessary that the prior model is realistic in the sense that typical realizations of the prior may be considered as likely realizations of the unobserved image. A realistic prior modelling also makes it meaningful to estimate prior parameters either from training data, or from the actual image data by using the EM algorithm (Quian & Titterington, 1991), or MCMC maximum likelihood for missing data situations, see Gelfand & Carlin (1991), Geyer (1994), and section 3.3.

In the last few years there has been a growing interest in creating realistic prior models. For high level tasks, such as object recognition, a current trend initiated by the group of Ulf Grenander is to use deformable template modelling, see e.g. Grenander & Miller (1994) and Grenander (1993). The prior is then a model for deformations and translations of the template which may be a simple polygon representing the outline of a typical object. Baddeley & Van Lieshout (1993) propose to use marked Markov point processes for prior modelling of the mutual placement of objects. Hurn & Rue (1997) define a template for each of a number of types of objects, and combine this with a Markov point process to obtain a prior for identification and classification of objects in multitype object scenes. Realistic Markov random field (MRF) models for images of objects against a background are considered in Tjelmeland & Besag (1996), and Møller & Waagepetersen (1996) introduce a new class of models named Markov connected component fields (MCCF's), which facilitate prior modelling of homogeneous connected regions in the image. It is demonstrated that a reasonable fit to real image data can be obtained with simple MCCF models. A more detailed discussion of MRF's and MCCF's is given in section 1.2. Künsch, Geman & Kehagias (1995) establish that any stationary discrete random field can be approximated arbitrarily close by hidden Markov random fields (HMRF's) and furthermore establish consistency of maximum likelihood estimation for HMRF's. They apply HMRF's as models for binary images, but the results are, as the authors point out, not very convincing.

It should be noted that the task of creating realistic prior models is much more difficult than that of constructing satisfactory penalizing terms. For a penalized likelihood estimate the large scale structures are determined by the data, and the penalizing term usually just serves to smooth the images at a local scale.

The Bayesian approach also has advantages when uncertainty of a functional of the image is required. One example is related to exploration of oil reservoirs which may be considered as 3D images consisting of different geological formations as e.g. chalk and shales. In this case direct well measurements are sparse and indirect seismic observations encumbered with great uncertainty, so that point estimation is often not meaningful. The real object of interest is moreover complicated functionals of the reservoir like e.g. production profiles. Bayesian modelling, combined with MCMC computation of posterior distributions, seems a very appropriate approach to assessment of uncertainty in such situations, see e.g. Omre & Tjelmeland (1996) and Syversveen & Omre (1996).

1.2 Markov connected component fields

1.2.1 Definitions

In practice images are represented digitally, i.e. as a matrix of pixel values indexed by a set I. An image parametrization commonly applied for image segmentation is based directly on the digital representation, so that the unobserved image belongs to $S = V^I$, where V is a finite set of colours or image labels for each region type in the image. For specificity we let $V = \{0, \dots, k-1\}$ where $k > 1$.

Suppose that a symmetric and reflexive relation $\sim$ on I is given. A $\sim$-connected component K is a nonempty subset of I, so that for all $i, j \in K$, there exists $i_1, \dots, i_n \in K$ with $i = i_1 \sim \cdots \sim i_n = j$. We let $\mathcal{K}$ denote the set of connected components. For each $x = (x_i)_{i \in I} \in S$, and any $A \subseteq I$, the connected component relation $\underset{x_A}{\sim}$ is defined by

$$\forall\, i, j \in A : i \underset{x_A}{\sim} j \Leftrightarrow \exists\, i_1, \dots i_n \in A : \; i = i_1 \sim \cdots \sim i_n = j \text{ and } x_{i_1} = \cdots = x_{i_n}.$$

The set of nonempty maximal cliques with respect to $\underset{x_A}{\sim}$ is denoted $\mathcal{K}(x_A)$, and $\mathcal{K}(x_A)$ hence constitutes the set of maximal homogeneous connected components of pixels in the image $x_A = (x_i)_{i \in A}$. Let further $\mathcal{K}^{(l)}(x_A) = \{K \in \mathcal{K}(x_A) \mid \forall j \in K : x_j = l\}$ be the set of connected components of colour $l \in V$ in the image $x_A \in V^A$. A random field $X = (X_i)_{i \in I}$ is now a MCCF if the density of X is of the form

$$p(x) \propto \prod_{K \in \mathcal{K}(x)} \Psi_K\Big(l(x_K)\Big), \; x \in S, \tag{1.3}$$

where $l(x_K)$ is the common value of x_i for $i \in K$, and Ψ is a nonnegative function defined on $\mathcal{K} \times V$. If $\Psi_K(0) = 1$ for all $K \in \mathcal{K}(x)$, then X is a *MCCF with background colour 0* which is a special case of the nearest-neighbour Markov point processes introduced in Baddeley & Møller (1989).

Prior knowledge of image components concerning e.g. size, shape, boundary complexity, or the Euler-Poincaré characteristic, can be incorporated in the model by choosing an appropriate function Ψ. The product form of p further implies that a MCCF possesses various interesting Markov properties.

Let for future use $A^c = I \setminus A$, $\partial A = \{j \in A^c \,|\, \exists i \in A : i \sim j\}$, and $\bar{A} = \partial A \cup A$. Further, for $K \subseteq I$ and $A \subseteq I$, we write $K \uparrow A$ if $K \cap A \neq \emptyset$.

1.2.2 Comparison with Markov random fields

A random field $X = (X_i)_{i \in I}$ is a MRF with respect to $\sim$, if the conditional distribution of X_i given $X_{-i} = (X_j)_{j \in I \setminus \{i\}}$ only depends on those X_j, where $j \sim i$. That is,

$$P(X_i = x_i | X_{-i} = x_{-i}) = p_i(x_i | x_{-i}) = p_i(x_i | x_{\partial \{i\}}), \; x \in S, \; i \in I, \tag{1.4}$$

where $\partial\{i\}$ is the $\sim$-neighbourhood of i. Suppose that p is hereditary, i.e.

$$p(x) > 0 \Rightarrow p(x^i) > 0 \text{ for all } i \in I \text{ and } x \in S, \tag{1.5}$$

where x^i is defined by $x^i_i = 0$ and $x^i_j = x_j$, $j \in I \setminus \{i\}$. It is then implied by the Hammersley-Clifford theorem (see the historical account in Clifford, 1991, and the references therein) that the density of X is of the form

$$p(x) \propto \prod_{C \in \mathcal{C}} \phi(x_C), \tag{1.6}$$

where $\mathcal{C}$ is the set of $\sim$-cliques excluding the empty clique, and ϕ is a nonnegative clique interaction function defined on $\cup_{C \in \mathcal{C}} V^C$.

Examples in appendix A in Møller & Waagepetersen (1996) show that neither is the class of $\sim$-MRF's contained in the class of $\sim$-MCCF's or vice versa. The intersection of the two model classes is characterized by Theorem 1 in Møller & Waagepetersen (1996). In this theorem it is under a positivity condition stated that the density of a random field which is both a MRF and a MCCF is of the form

$$p(x) \propto \prod_{K \in \mathcal{K}(x)} \prod_{C \in \mathcal{C} : C \subseteq K} \psi_C\Big(l(x_K)\Big), \tag{1.7}$$

where ψ is a positive function defined on $\mathcal{C} \times V$.

Consider the case $I \subseteq \mathbb{Z}^2$, and $\sim$ the second order neighbourhood relation $\sim_2$, where $i \sim_2 j$ if and only if $\|i - j\| < 2$. If p is both a MCCF and a MRF where the clique interaction function ϕ is motion invariant, then by the corollary to Theorem 1,

$$p(x) \propto \prod_{K \in \mathcal{K}(x)} \exp\Big(\alpha_{l(x_K)} a(K) + \beta_{l(x_K)} u(K) + \gamma_{l(x_K)} k_+(K) + \delta_{l(x_K)} \chi(K) + \epsilon_{l(x_K)} d(K)\Big), \tag{1.8}$$

where for $K \in \mathcal{K}$, $a(K) = |K|$ is the area measured by the cardinality of K, $u(K) = |\{(i,j) \in K \times \mathbb{Z}^2 \setminus K \mid i \sim_1 j\}|$ is the $\sim_1$-perimeter, $k_+(K)$ is the number of concave corners, $d(K)$ is the number of $\sim_1$ discontinuities, and $\chi(K)$ is the Euler-Poincáre characteristic (for further details see section 2.2 in Møller & Waagepetersen, 1996).

The Ising model is the special case of (1.8), where $V = \{0,1\}$, $\beta_l = \beta$, $l \in V$, and $\alpha_l = \gamma_l = \delta_l = \epsilon_l = 0$, $l \in V$.

1.2.3 Markov properties

Global Markov property

For a $\sim$-MRF X, the conditional distribution of X_A given X_{A^c} depends on X_{A^c} only through $X_{\partial A}$. For the relation $\underset{x}{\sim}$, the neighbourhood $\{j \in A^c | \exists i \in A : i \underset{x}{\sim} j\}$ of a set $A \subseteq I$ depends on x, and for MCCF's it is therefore required to condition on more than for MRF's, in order to obtain conditional independence results.

The spatial Markov property for MCCF's is given in terms of certain random partitions of I denoted random Markov partitions. A mapping $M(\cdot) = \big(A(\cdot), B(\cdot), C(\cdot)\big)$ defined on S and taking values in

$$\Delta = \{(A,B,C) \mid A, B, C \text{ are disjoint sets with } I = A \cup B \cup C\}$$

is said to be a *Markov partition* if the conditions (1.9)-(1.11) below are satisfied. The two first conditions are

$$\forall\, x \in S,\ i \in \overline{A(x)},\ j \in C(x) : i \underset{x}{\not\sim} j \tag{1.9}$$

and

$$\forall\, x, y \in S : B(x) = B(y) \Leftrightarrow M(x) = M(y). \tag{1.10}$$

Assuming (1.10), and defining for $B \subseteq I$ and $x_B \in V^B$,

$$\begin{aligned}
\mathcal{A}(B, x_B) &= \{y_A \in V^A \mid \exists y_C \in V^C : B(y_A, x_B, y_C) = B\},\\
\mathcal{C}(B, x_B) &= \{y_C \in V^C \mid \exists y_A \in V^A : B(y_A, x_B, y_C) = B\},\\
\mathcal{D}(B, x_B) &= \{(y_A, y_C) \in V^A \times V^C \mid B(y_A, x_B, y_C) = B\},
\end{aligned}$$

the last condition is

$$\forall x \in S : \mathcal{D}(B, x_B) = \mathcal{A}(B, x_B) \times \mathcal{C}(B, x_B) \text{ where } B = B(x). \tag{1.11}$$

Here and elsewhere in the following we identify (y_A, x_B, y_C) with the $z = (z_i)_{i \in I}$, where $z_i = x_i$, $i \in B$, and $z_i = y_i$, $i \in A \cup C$. The condition (1.9) implies that a component $K \in \mathcal{K}(x)$ can not intersect both $\overline{A(x)}$ and $C(x)$, and $B(x)$ is hence denoted the splitting set. The Markov partition is by (1.10) determined by $B(\cdot)$, and (1.11) implies that y_A and y_C can be checked separately to see if $(y_A, y_C) \in \mathcal{D}(B, x_B)$.

A random field X on S, and a Markov partition $M(\cdot)$ induce a random Markov partition $\mathbf{M} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) = M(X) = \big(A(X), B(X), C(X)\big)$. The spatial Markov property for a MCCF is given in Theorem 2 in Møller & Waagepetersen (1996) as follows:

Theorem 2 For a Markov partition $M(\cdot) = \big(A(\cdot), B(\cdot), C(\cdot)\big)$ and a MCCF X, we have that $X_{\mathbf{A}}$ and $X_{\mathbf{C}}$ are conditionally independent given $\mathbf{B}$ and $X_{\mathbf{B}}$.

Suppose for $B \subseteq I$ and $x_B \in V^B$ that $p(\mathbf{B} = B, X_{\mathbf{B}} = x_B) > 0$ or equivalently,

$$\exists (y_A, y_C) \in \mathcal{D}(B, x_B) : \prod_{K \in \mathcal{K}((y_A, x_B, y_C))} \Psi_K\big(l(y_K)\big) > 0.$$

Then the conditional distribution of $X_{\mathbf{A}}$ given that $\mathbf{B} = B$ and $X_{\mathbf{B}} = x_B$ is concentrated on the set $\mathcal{A}(B, x_B)$, and its conditional density is

$$p(x_A \mid B, x_B) = c(B, x_B) \prod_{K \in \mathcal{K}(x_{A \cup B}) : K \uparrow \bar{A}} \Psi_K\big(l(x_K)\big), \; x_A \in \mathcal{A}(B, x_B), \tag{1.12}$$

where $c(B, x_B)$ is a normalizing constant. Similarly,

$$p(x_C \mid B, x_B) \propto \prod_{K \in \mathcal{K}(x_{B \cup C}) : K \uparrow \bar{C}, K \not\uparrow \partial A} \Psi_K\big(l(x_K)\big), \; x_C \in \mathcal{C}(B, x_B). \tag{1.13}$$

□

Remark: The conditional densities (1.12) and (1.13) are slightly different due to the asymmetry in (1.9).

For $A \subseteq I$ we can e.g. consider the splitting set $B(x) = \cup_{K \in \mathcal{K}(x_{A^c}) : K \uparrow \bar{A}} K$ which depends only on x_{A^c}. A Markov partition is then given by $M(x) = \big(A, B(x_{A^c}), A^c \setminus B(x_{A^c})\big), x \in S$, and $\mathcal{A}(B, x_B)$ is in this case V^A if there exists a $y_{A^c \setminus B} \in V^{A^c \setminus B}$ such that $B(x_B, y_{A^c \setminus B}) = B$, and $\mathcal{A}(B, x_B)$ is empty otherwise. Other Markov partitions, which may be useful for handling of edge effects, are given in Møller & Waagepetersen (1996).

Local Markov property

For a hereditary density p, the ratio of local characteristics given by

$$\lambda_i(x) = p_i(x_i \mid x_{-i}) / p_i(0 \mid x_{-i}) \tag{1.14}$$

is well-defined for all $x \in S$ with $p(x) > 0$. If p is the density of a MCCF with background colour 0, then

$$\lambda_i(x) = \Psi_{K_i(x)}(x_i) / \prod_{K \in \mathcal{K}(x^i_{K_i(x)})} \Psi_K(x_i), \tag{1.15}$$

which depends on x only through $x_{K_i(x)}$ where $K_i(x) \in \mathcal{K}(x)$ is determined by $i \in K_i(x)$. In Theorem 3 in Møller & Waagepetersen (1996) it is established using Lemma 1 in Baddeley, Van Lieshout & Møller (1995) that the class of hereditary MCCF's with background colour 0 is characterized by the local Markov property (1.15).

If a MRF is specified in terms of local characteristics, the Hammersley-Clifford theorem yields that the joint distribution of the MRF is of the form (1.6). One can then apply the Brook expansion (see e.g. Besag, 1974) to calculate the clique interaction functions and check whether the local specification is consistent. The characterization result given by Theorem 3 in Møller & Waagepetersen (1996) can in principle be applied in a similar way to calculate Ψ from a local specification, but things are more complicated than for MRF's since $\mathcal{K}$ is typically of a very high cardinality.

1.2.4 Existence of MCCF's on infinite lattices

The question of existence of stationary MCCF's is important for establishment of almost sure consistency of estimators, and it is also of interest to study when extensions of finite lattice MCCF's to infinite lattices are nonunique, i.e. when phase transition occurs. Phase transition may have a dramatic effect also on the properties of the finite lattice random fields, and is further discussed in section 1.2.6.

For a symmetric and reflexive relation $\sim$ defined on I, and V a finite set of colours/labels, the sets $\mathcal{K}$ and $\mathcal{K}(x_A)$ for $x \in S = V^I$ and $A \subseteq I$, are defined as in section 1.2.1. Let further $\mathcal{G} = \{\Lambda \subseteq I \mid |\Lambda| < \infty\}$ be the set of finite subsets of I, and let $\{\Lambda_n\}_{n \geq 1}$ be an increasing sequence of subsets in $\mathcal{G} \cap \mathcal{K}$ such that for all $\Lambda \in \mathcal{G}$, there exists a Λ_k, $k \geq 1$, with $\Lambda \subseteq \Lambda_k \setminus \partial(\Lambda_k^c)$. It is assumed that $\sim$ is of finite range, so that $\bar{\Lambda} \subseteq \mathcal{G}$ for all $\Lambda \in \mathcal{G}$, and we shall without loss of generality assume, that $I \in \mathcal{K}$.

For a given function Ψ defined on $(\mathcal{K} \cap \mathcal{G}) \times V$ and $x \in S$, it is in section 4 of Møller & Waagepetersen (1996) discussed how to define a family $\{p_\Lambda\}_{\Lambda \in \mathcal{G}}$ of functions $p_\Lambda : V^\Lambda \times V^{\Lambda^c} \to [0,1]$, where $p_\Lambda(\cdot|x_{\Lambda^c})$, $x \in S$ is a probability density related to the conditional probabilities of a MCCF defined on a finite index set. If $\{p_\Lambda\}_{\Lambda \in \mathcal{G}}$ satisfies the consistency condition

$$p_{\tilde{\Lambda}}(x_{\tilde{\Lambda}}|x_{\tilde{\Lambda}^c}) = p_\Lambda(x_\Lambda|x_{\Lambda^c}) \sum_{y_\Lambda} p_{\tilde{\Lambda}}(y_\Lambda, x_{\tilde{\Lambda}\setminus\Lambda}|x_{\tilde{\Lambda}^c}),\ x \in S,$$

when $\Lambda \subseteq \tilde{\Lambda}$, it is a specification as defined in Preston (1976, p. 16-17). If further X is a random field with values in S, where the conditional density of X_Λ given $X_{\Lambda^c} = x_{\Lambda^c}$ is given by $p_\Lambda(\cdot|x_{\Lambda^c})$ for each $\Lambda \in \mathcal{G}$ and $x_{\Lambda^c} \in V^{\Lambda^c}$ with $P(X_{\Lambda^c} = x_{\Lambda^c}) > 0$, then X is specified by $\{p_\Lambda\}_{\Lambda\in\mathcal{G}}$, and will be considered as a MCCF extended to the infinite lattice.

In Møller & Waagepetersen (1996) the probability density $p_\Lambda(\cdot \mid x_{\Lambda^c})$ is obtained in a natural way as the limit of conditional densities of MCCF's defined on bounded observation windows $\Lambda_n, n \geq 1$. More specifically (c.f. (1.12)), let

$$p_{\Lambda,n}(x_\Lambda \mid x_{\Lambda_n\setminus\Lambda}) \propto \prod_{K\in\mathcal{K}(x_{\Lambda_n}):K\uparrow\bar{\Lambda}} \Psi_K\Big(l(x_K)\Big) \tag{1.16}$$

for all n such that $\Lambda \subseteq \Lambda_n$ and $x \in R_{\Lambda,n}$, where $R_{\Lambda,n}$ is the subset of S for which the right hand side of (1.16) can be normalized, i.e.

$$R_{\Lambda,n} = \{x \in S \,|\, \exists\, y \in S : x_{\Lambda^c} = y_{\Lambda^c} \text{ and } \prod_{\substack{K\in\mathcal{K}(y_{\Lambda_n}):\\ K\uparrow\bar{\Lambda}}} \Psi_K\Big(l(y_K)\Big) > 0\}.$$

Let

$$R_\Lambda = \cap_{n\geq N(\Lambda)} R_{\Lambda,n}$$

where $N(\Lambda)$ is defined in the proof of Lemma 1 in Møller & Waagepetersen (1996). Then p_Λ is defined as the limit

$$p_\Lambda(x_\Lambda|x_{\Lambda^c}) = \lim_{n\to\infty} p_{\Lambda_n}(x_{\Lambda_n}|x_{\Lambda_n^c}), \tag{1.17}$$

for all $x \in R_\Lambda$. The condition (1.18) below ensures that this limit exists. Let ψ be a strictly positive function defined on $\mathcal{C} \times V$. It is, informally, required that $\Psi_K(l)$ can be approximated by the product $\prod_{C\in\mathcal{C}:C\subseteq K} \psi_C(l)$ for large but finite $K \in \mathcal{K}$ and $l \in V$.

More precisely,

$$\begin{aligned} &\forall \Lambda \in \mathcal{G},\ \epsilon > 0\ \exists\, N \geq 1 : \\ &\qquad \forall n \geq N,\ l \in V, \text{ and } K \in \mathcal{K} \cap \mathcal{G} \text{ with } K \uparrow \bar{\Lambda} \text{ and } K \uparrow \overline{\Lambda_n^c} : \\ &\qquad\qquad \left| \frac{\Psi_K(l)}{\prod_{C\in\mathcal{C}:C\subseteq K} \psi_C(l)} - 1 \right| < \epsilon. \end{aligned} \tag{1.18}$$

This condition also implies that p_Λ is quasilocal, i.e. a uniform limit of local functions (a function defined on S is local if there is an $A \in \mathcal{G}$, such that the function depends on x only through x_A for $x \in S$). The condition is therefore sufficient to use results in Preston (1976) to establish the existence of a random field X specified by $\{p_\Lambda\}_{\Lambda\in\mathcal{G}}$, where $P(X \in \cap_{\Lambda\in\mathcal{G}} R_\Lambda) = 1$.

The condition (1.18) e.g. holds for the penalized Ising model considered in section 1.2.5. Other examples include cases where $\Psi_K(\cdot) \to 1$ as $|K| \to \infty$.

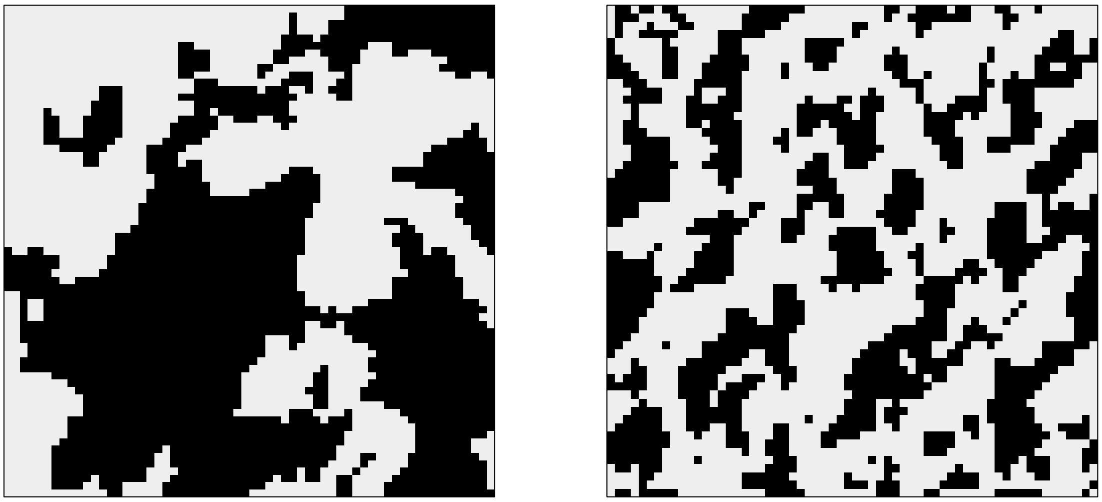

Figure 1.1: Data. Left: Rock sample image. Right: Handmade paper texture.

Assuming (1.18), it is in Lemma 1 in Møller & Waagepetersen (1996) shown that the limit in (1.17) exists, and in Lemma 2 that this limit is quasilocal. Theorem 4 is the existence theorem in which it is established that if Ψ satisfies (1.18), then there exists a random field whose conditional probabilities are given by the specification $\{p_\Lambda\}_{\Lambda\in\mathcal{G}}$ corresponding to Ψ.

Conditions for absence of phase transition may be established by using the uniqueness criterion of Dobrushin (see e.g. Georgii, 1988, p. 142), but the question of uniqueness is not addressed in Møller & Waagepetersen (1996).

1.2.5 Examples and simulations

In this section we shall consider some specific examples of MCCF's which will be fitted to the binary image data shown in Figure 1.1.

(a) *The penalized Ising model.* This MCCF model appears as a modification of the Ising model where realizations with presence of small connected regions of constant colour are penalized. The model is given by

$$p(x) \propto \exp\Big(-\Big(\sum_{K\in\mathcal{K}(x)} \beta u(K) + \frac{\gamma}{a(K)}\Big)\Big),\ x \in S, \tag{1.19}$$

so that small components are penalized when γ is positive. This model was fitted to the rock sample image using MCMC maximum likelihood estimation (see Geyer & Thompson, 1992 and section 3.3), whereby estimates $\hat{\beta} = 0.483$ and $\hat{\gamma} = 17$ were obtained. Figure 1.2 shows simulations of the fitted model. The large scale variability of the rock sample image is well reproduced by the penalized Ising model, but the boundaries in the data image appear somewhat smoother than the boundaries in the simulations. We shall return to the penalized Ising model in section 1.2.6.

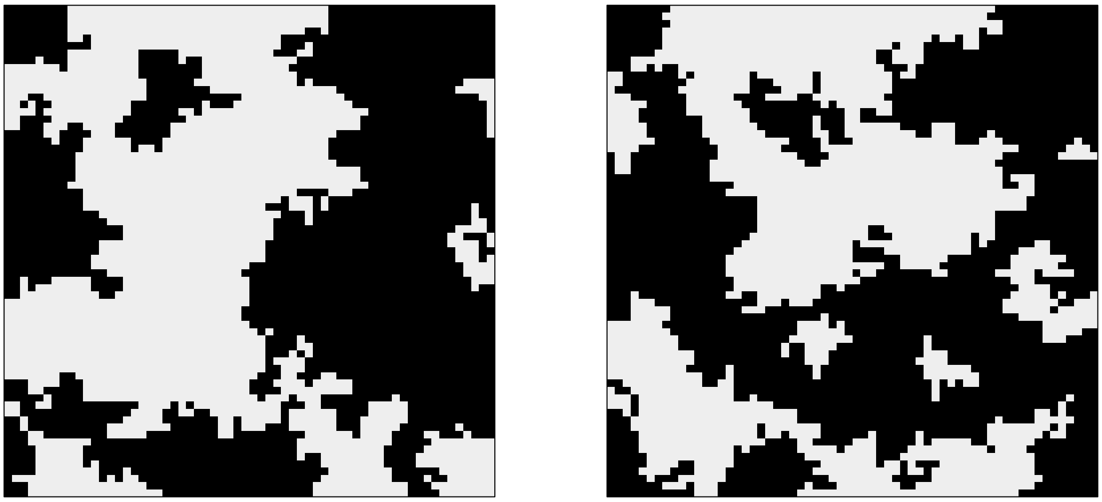

Figure 1.2: Realizations of the fitted model (1.19)

(b) A *MCCF model for black components on a white background.* The number and size of black components are important features of the handmade paper image in Figure 1.1. In order to model this image we first tried a binary second order MRF given in terms of the five geometric characteristics a, u, d, k_+, and χ. From the corollary to Theorem 1 in Møller & Waagepetersen (1996) it follows that this is equivalent to using the general parametrization (1.6) of a binary second order MRF with a motion invariant clique interaction function. This MRF model was not able to capture the variability displayed by the data.

Instead we considered the MCCF with background colour 0 given by

$$p(x) \propto \exp\Big(\sum_{K \in \mathcal{K}^{(1)}(x)} \alpha a(K) + \beta u(K) + \epsilon(\chi(K) - 1) + \phi + \gamma a(K)^2\Big), \ x \in S. \tag{1.20}$$

The parameter γ controls the size of the components since the squared area of the union of two components is greater than the sum of the squared areas of each component. The size of the components is also influenced by the parameter α together with the parameter ϕ which controls the number of components. The parameter β controls the perimeter of the black part of the image, and thereby also the shape (compactness) of the components. The statistic $1 - \chi(K)$ counts the number of holes in a component K.

The MCMC maximum likelihood estimates obtained from the data are $\hat{\alpha} = 0.087$, $\hat{\beta} = -0.834$, $\hat{\epsilon} = 1.550$, $\hat{\phi} = -1.045$, and $\hat{\gamma} = 1.923 \times 10^{-4}$. Simulations (see Figure 1.3) show that the variability of the number and size of the components is quite well reproduced by the model, and the characteristic $\chi - 1$ turned out to be very useful for modelling the number of holes in the components. The model was also judged by comparison of the empirical distributions of the black component characteristics $a(K)$, $u(K)$, and $u(K)/\sqrt{a(K)}$, $K \in \mathcal{K}^{(1)}(\text{data})$, with

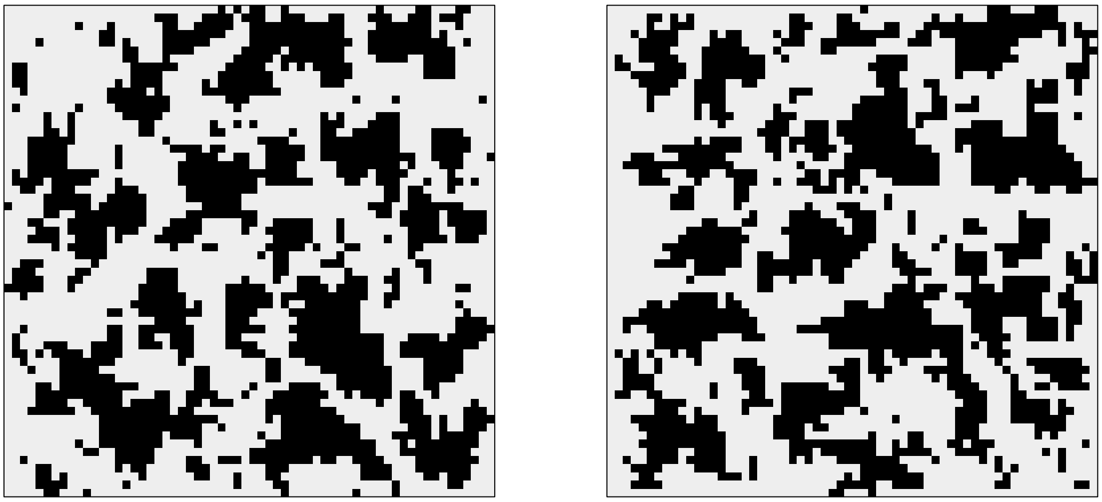

Figure 1.3: Realizations of the fitted model (1.20)

the distributions estimated from simulations of the fitted model, see Figure 8 in Møller & Waagepetersen (1996). Especially with respect to $a(K)$ and $u(K)$, the fit of the estimated model seems quite good.

(c) *An example of Bayesian image analysis.* This example from Møller & Waagepetersen (1996) shows the importance of using a realistic prior model when the full posterior distribution is required for the inference. Synthetic data obtained by adding a Gaussian noise with variance 1.25 to the rock sample image was considered. With this high noise level it is not possible to obtain precise point estimates, and an assessment of uncertainty is therefore relevant. The Ising model and the penalized Ising model, respectively, were used as priors, and the prior parameters were estimated from the rock sample image. This corresponds to estimation of prior parameters from training data. MCMC was used to estimate the posterior distributions of the number of black pixels, the number of components, and the marginal posterior probabilities of observing a black pixel. The results obtained with the two priors are shown in Figure 1.4, and the true numbers of components and black pixels together with posterior means and standard deviations are given in Table 1.1.

The posterior distribution obtained with the Ising prior is clearly a misleading representation of the knowledge concerning the number of components in the true image. This is not surprising since a large number of small components are present in the realizations of this posterior, see Figure 9 in Møller & Waagepetersen (1996). The posterior distribution of the number of components is more correctly centered around the true value when the penalized Ising prior is used, and this is also the case for the posterior distribution of the number of black pixels. The marginal posterior probabilities of observing a black pixel are furthermore in much better accordance with the pixel values of the true image, when the penalized

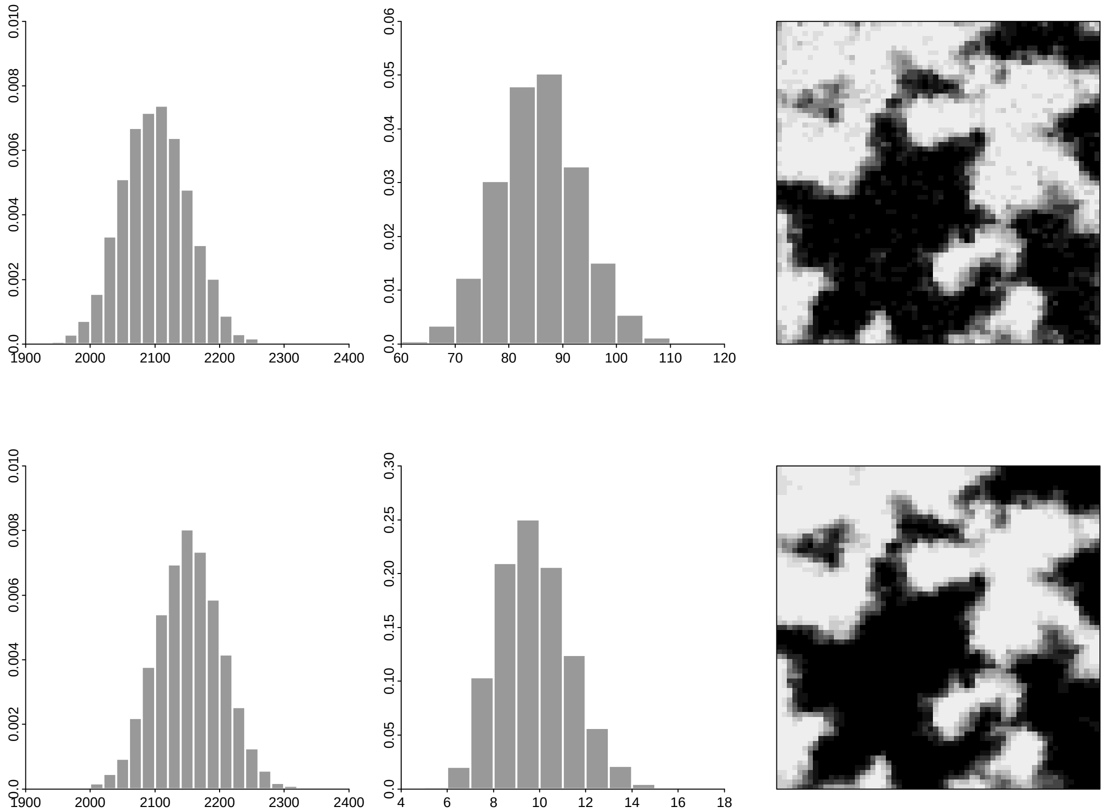

Figure 1.4: Upper row: posterior distribution with Ising prior of number of black pixels (left), number of components (middle), and the marginal posterior probabilities of observing a black pixel (where white in the gray level image corresponds to zero probability). Lower row: same as upper row, but with the penalized Ising prior.

Ising prior is used.

One may object that better results might be obtained with the Ising prior if another prior parameter was used instead of the maximum likelihood estimate. It is on the other hand not at all clear how to choose such an alternative parameter value.

1.2.6 Phase transition

The Ising model was originally introduced as a model for ferromagnets, and it is very interesting that this simple model is able to reproduce the phenomenon of spontaneous magnetization which is one example of a phase transition, see Kindermann & Snell (1980) or Georgii (1988). For Ising models defined on large finite lattices with a periodic or free boundary condition, the phase transition behaviour can be observed as an abrupt change from unimodality to bimodality of

True values	2157	11
Ising prior	2101.4 (51.8)	86 (7.7)
Penalized Ising prior	2153.1 (50.7)	10.3 (1.6)

Table 1.1: True values, posterior means, and standard deviations (in parenthesis) of number of black pixels and number of components.

the distribution of the number of black pixels. This change occurs when β exceeds a certain critical value, and corresponds to a transition from short range correlations to infinite range correlations for the infinite lattice Ising model. Mathematically, phase transition is defined as nonuniqueness of extensions to infinite lattices of finite lattice random fields. The set of infinite lattice Ising distributions is given by convex combinations of two extremal states which may be considered as perturbations of the point mass measures on the two modes (i.e. the constant images), see Georgii (1988) p. 101.

The phase transition behaviour is not an important issue when the Ising model is used as a penalizing term for MAP-estimation. In this case the Ising model, roughly speaking, serves as a device to clean images by removing components of pixels according to whether β is larger than a certain ratio between the log-likelihood ratio for removing the component, and the length of the boundary of the component, see the discussion in section 2 in Waagepetersen (1997*a*). Phase transition on the other hand seems undesirable if a realistic prior modelling is requested, since realizations of the subcritical Ising model has a rather chaotic and "noisy" appearance, while the supercritical realizations are dominated by one colour and contain no interesting structures.

It was therefore in Møller & Waagepetersen (1996) interesting to note that phase transition did not seem to occur for the fitted penalized Ising model described in section 1.2.5, despite that $\hat{\beta}$ is supercritical for the Ising model. This is investigated further in Waagepetersen (1997*b*), where it is demonstrated using the Peierls (1936) argument that phase transition actually does occur for the penalized Ising model for any positive γ and sufficiently large values of β. The apparent absence of phase transition behaviour for the penalized Ising model could be due to nonconvergence of the single-site updating Metropolis algorithm used for the simulations. When phase transition occurs one may expect that the model is concentrated on and around the modes, which are the constant images. It is in practice not possible for a single-site updating algorithm to move from typical nonconstant realizations of the fitted penalized Ising model to the constant images, since the chain then needs to move through images of very low probability. From the Monte Carlo estimate in Waagepetersen (1997*b*) it turned out, however, that the probability of the modes is very small ($\approx 10^{-13}$), so it seems safe to conclude that phase transition does not occur for the fitted penalized Ising model.

A natural approach to improve the single-site Metropolis algorithm is to introduce transitions, where all pixels in a component are updated simultaneously (whereby the component is removed from the image). A proposal mechanism for such a transition is easy to construct, but in order to maintain reversibility, it is also necessary to be able to make the reverse transition, i.e. to insert a component which is a subset of another existing component in the image. It is not straightforward to construct a proposal mechanism for such a move since the mechanism must generate components for which an insertion is likely to be accepted. For the penalized Ising model this especially means that the generated components must not contain small holes, since small holes occur with a very small probability under the penalized Ising model. The mechanism for generating a component must further be sufficiently simple, so that the probability of generating a given component can be easily calculated. This problem is in principle solved in Waagepetersen (1997*b*), but the constructed algorithm turned out to be useful only for a rather restricted set of parameter values for the penalized Ising model. For the Ising model the algorithm works for all values of β, but is not as efficient as the Swendsen & Wang (1987) algorithm.

The Ising model is an example of a pairwise interaction MRF. One approach to obtain parametrizations of higher order MRF's is to use morphological operations as suggested in Chen & Kelly (1992), Carstensen (1992), and Sivakumar & Goutsias (1997). Let $A(x) = \{i \in I | x_i = 1\}$ be the set of pixels of value one in a binary image x, and let $\bullet$ and $\circ$ denote the morphological operations of closure and opening (Serra, 1982), respectively. The result $A(x) \bullet B$ of the closure with a suitable structure element B, is a smoothed version of $A(x)$, where narrow "bays" and "channels" are filled out, and the opening operation smoothes $A(x)$ by removing sharp "capes" and thin "isthmuses". The inclusions $A(x) \circ B \subseteq A(x) \subseteq A(x) \bullet B$ hold, and in Sivakumar & Goutsias (1997) it is showed that the random field given by

$$p(x) \propto \exp(-\beta |A(x) \bullet B \setminus A(x) \circ B|) \qquad (1.21)$$

is a MRF with a local neighbourhood system. A class of MRF models may, as suggested in Sivakumar & Goutsias (1997), be obtained by introducing structure elements of different size and shape.

When $\beta > 0$ the model (1.21) favour images which are morphologically smooth, and the modes are given by those images $x \in S$ for which $A(x) \bullet B = A(x) \circ B$. Such images may contain many interesting structures, and it is therefore, from an applied point of view, probably not so important whether phase transition occurs for the model (1.21) (if one assumes that the extremal distributions are concentrated around the modes, when phase transition occurs). It is on the other hand by no means necessary that the modes of an image model contain interesting global structures. This is demonstrated by the penalized Ising model (1.19), and the MRF-models in Tjelmeland & Besag (1996), for which the modes are given by the constant images.

1.2.7 A discussion concerning MCCF's, MRF's and intermediate level priors

In the literature on MRF's there has traditionally been much focus on pairwise interaction MRF's where $\phi(x_C) = 1$ if $|C| > 3$ in the product (1.6). Models with a simple parametrization are thereby obtained, and pairwise interaction priors usually work well for smoothing purposes in MAP-estimation. It is, however, widely recognized that pairwise interaction models are not able to produce realizations which contain global structures of the kind that appear in real images scenes. The results in Tjelmeland & Besag (1996) show that interesting spatial scenes can be modelled by means of MRF's when higher order interactions are included. The number of parameters for the general parametrization of a MRF grows rapidly as the neighbourhood size increases, and the choice of a parsimonious parametrization therefore becomes an important and difficult problem. This highlights one of the attractive properties of MCCF's: large scale structures given by the connected components are modelled explicitly, and from Corollary 1 in Møller & Waagepetersen (1996) a guideline can be derived which suggests that one should use the five geometric component characteristics as a starting point for the modelling. The simulations in section 1.2.5 show that models for very different patterns can be obtained with this approach. Another problem with MRF's is related to change of resolution. If a MRF model has been chosen for a given digitized image, it is not clear how this model should be modified if a coarser or finer digitization is introduced. MCCF modelling based on the five geometric characteristics may be less sensitive to this problem, since at least the interpretation of the geometric characteristics is independent of scale.

MRF's and MCCF's are especially useful as intermediate level priors in situations where only vague prior information concerning spatial homogeneity, size of connected components, and boundary smoothness is available. If the images exhibit a high degree of regularity, so that one e.g. knows that ellipse shaped objects appear on a background, then continuum high level image priors based on deformable templates and marked Markov point processes are advantageous. With such priors relatively low dimensional image parametrizations are used, and rotations and translations can easily be applied to the image parameters. A high level of image interpretation can furthermore be obtained from the posterior distribution, and the image parametrization is independent of the choice of discretization for the digitized image.

Coloured triangulation models

The pixel based representation is flexible in the sense that all kinds of scenes can be represented, when the discretization is sufficiently fine. The drawback is the high dimensionality. An interesting approach to reduce dimensionality and still maintain flexibility is the coloured triangulation models in Nicholls (1996). In this paper $D \subseteq \mathbb{R}^2$ is an open bounded continuous index set whose boundary D_{bd} is given by a simple polygon, and the space of images is V^D, where V is as usual a finite set of colours. Let v_c denote the set of corner vertices of D_{bd} and let v_d and v_b be finite subsets of D and D_{bd}, respectively. The set of all possible triangulations of D with vertices given by $v_c \cup v_b \cup v_d$ is denoted $\Gamma_D(v_c, v_b, v_d)$. The space of coloured triangulations is

$$T = \cup_{\substack{v_b \subset D_{bd} \\ |v_b|<\infty}} \cup_{\substack{v_d \subset D: \\ |v_b|<\infty}} \cup_{\tau \in \Gamma_D(v_c,v_b,v_d)} \cup_{c \in C(\tau)} (\tau, c),$$

where $C(\tau)$ is the set of all possible colourings of the triangulation τ, where to each face in τ, one of the colours in V is assigned. An image $x = (x_s)_{s \in D}$ corresponding to a triangulation (τ, c) is given by $x_s = l$, $s \in D$, if $l \in V$ is the colour of the face to which s belongs. A finite measure χ_λ on T is defined by

$$\chi_\lambda(A) = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{\lambda^{2n+k}}{n!k!} \int_{D^n} \int_{(D_{bd})^k} \sum_{\tau \in \Gamma_D(v_c, s(v), s(u))} \sum_{c \in C(\tau)} 1\Big((\tau, c) \in A\Big) du dv, \ A \in \mathcal{T}, \tag{1.22}$$

where $\lambda > 0$ is an intensity parameter, $\mathcal{T}$ is an appropriate σ-algebra on T, and $s\Big((u_1, \ldots, u_n)\Big) = \{u_1, \ldots, u_n\}$, $n \geq 1$. The prior model for coloured triangulations is specified by a density with respect to χ_λ. The density may e.g. be similar to an Ising model ($V = \{0, 1\}$), so that

$$p(\tau, c) \propto \exp\Big(-\beta E(\tau, c)\Big),$$

where $\beta > 0$, and $E(\tau, c)$ is the number of pairs of different coloured faces of τ which have an edge in common. The advantage of the model in Nicholls (1996) is that one need not settle for a fixed lattice. The posterior distribution is instead concentrated on coloured triangulations, where the lattice corresponding to the triangulation is "adapted" to the actual observed image data. The prior parameter λ controls the number of vertices in the triangulations and thereby the coarseness of the lattice. A potential problem is parameter estimation. This is a missing data problem since there to a given image correspond infinitely many coloured triangulations. The approach in Nicholls (1996) is further discussed in section 2.4.

Noninformative image priors

The notion of a noninformative image prior is not well understood. For a univariate parameter in a finite parameter space, it is natural to choose the uniform distribution as a noninformative prior since all parameter values may a priori be equally plausible. For images the situation is different. Suppose that an image belongs to the finite state space $S = V^I$. Then the uniform distribution is noninformative in the sense that the MAP-estimate only depends on the likelihood.

But all typical realizations of this prior clearly differ strongly from prior expectations of typical images since no spatial homogeneity is present. A model like e.g. the penalized Ising model may seem a more reasonable choice as a noninformative or at least vague prior. Typical realizations of this model contain a suitable degree of spatial homogeneity, and the model assigns probabilities of the same order of magnitude to a very large variety of global structures.

1.3 Analysis of residuals from segmentation of noisy images

For many penalizing terms used for MAP-estimation of images there are not both objective and practically applicable methods for choosing the smoothing parameter. The smoothing parameter is then chosen in an ad hoc manner, and it is relevant to analyze residuals in order to check whether a suitable amount of smoothing has been applied. If a precise estimate is obtained, then the residual process is similar to the noise process. One may therefore investigate the residuals by comparing summary statistics of the residual process with the distribution of the summary statistics, when they are evaluated on the noise process. In Waagepetersen (1997*a*) three different methods for analyzing residuals are proposed and applied to a simple case of image segmentation, where a binary image is corrupted by an additive *iid* noise. Attention is especially focused on the case of a binary white noise. MAP-estimates are obtained using an Ising penalizing term, and simulated annealing is used to maximize the penalized likelihood.

If an oversmoothed estimate is obtained, one may expect clustering of pixels in the residual image where components in the true image are smoothed away. In the first approach considered in Waagepetersen (1997*a*) the distribution of the residual process is approximated by an Ising model with an external field. This model is fitted to the residual process, and a log-likelihood ratio test for absence of interactions in the residuals is applied.

A computationally simpler method is obtained by considering statistics U_k which for $k = 0, \ldots, 8$, are given by the fraction of 1-pixels in the residual image which have k neighbours of colour 0. The asymptotic distribution of these statistics is Gaussian, under the assumption that the residual process is a version of the noise process.

The third method is based on the empirical distribution of the boundary to area ratios $bd(K)/a(K)$, $K \in \mathcal{K}^{(1)}(r)$, for the 1-components in the residual image r. This statistic is compared with envelopes calculated from simulations of the noise process.

The experiments showed that there is scope for application of the second and third method, while the first method is too sensitive to the fact that the residual process is a perturbed version of the noise process, when very precise estimates can not be obtained.

For a noise process defined on an infinite lattice and for values of the noise rate where site-percolation (see e.g. Kesten, 1982) does not occur, it is possible to define the distribution of the boundary to area ratio of the "typical" 1-connected component. This distribution can be estimated unbiasedly using Monte Carlo, and it is discussed in section 2 of Waagepetersen (1997*a*) how this may be helpful to provide guidelines for the choice of smoothing parameter.

In Rue (1996) the residuals are analyzed from a somewhat different perspective. A statistic which measures deviations of the sample correlation structure of the residuals from the noise correlation structure is constructed, and this statistic is included in the penalizing term in order to enhance the image restoration.

2 Modelling of point patterns

2.1 Introduction

Statistical models for point pattern data are given in terms of point processes, where the points are usually locations in a subset S of $\mathbb{R}^d$. Let $\mathcal{B}(S)$ and $\mathcal{B}_0(S)$ denote the Borel sets and the bounded Borel sets of S, respectively, and let $N(S)$ denote the set of integer-valued counting measures on $\big(S, \mathcal{B}(S)\big)$. Let further $\mathcal{N}(S)$ denote the σ-algebra on $N(S)$ generated by the projections p_A, $A \in \mathcal{B}(S)$, given by $p_A(c) = c(A)$, $c \in N(S)$. A measure $c \in N(S)$ is locally finite if $c(B) < \infty$ for all $B \in \mathcal{B}_0(S)$, and a point process Φ on S is a locally finite integer-valued random measure, i.e. a measurable mapping from a probability space into $\big(N(S), \mathcal{N}(S)\big)$. A point process may equivalently be considered an integer-valued stochastic process $\{\Phi_B\}_{B \in \mathcal{B}(S)}$, where the stochastic variables Φ_B satisfy certain consistency conditions, see Proposition 7.1.X in Daley & Vere-Jones (1988). A point process is simple if $N(\{s\}) \in \{0, 1\}$ for all $s \in S$, and a simple point process may therefore be identified with a random countable subset X of points in S, where $s \in X$ if and only if $N(\{s\}) = 1$. We shall restrict attention to simple point processes. In the sequel ν_d denotes the Lebesgue measure on $\mathbb{R}^d$, and $|\cdot|$ as usual denotes the cardinality of a set.

The Poisson process corresponding to a locally finite mean measure m is the fundamental point process characterized by

P1 $|X \cap B|$ is Poisson distributed with mean $m(B)$, $B \in \mathcal{B}_0(S)$.

P2 $|X \cap B_i|$, $n > 1$, $i = 1, \ldots, n$, are independent whenever $B_1, \ldots, B_n \in \mathcal{B}_0(S)$ are disjoint.

The distribution of the unit rate homogeneous Poisson process for which $m = \nu_d$, will be denoted μ_S.

The class of Cox processes is obtained by substituting m by a random mean measure M which is often given in terms of a stochastic intensity process $\Lambda = \{\Lambda_s\}_{s\in\mathbb{R}^d}$ as

$$M(B) = \int_B \Lambda_s ds, \ B \in \mathcal{B}_0(S). \tag{2.1}$$

If e.g. the Cox process is a model for stands of trees, then Λ may represent an environmental heterogeneity related to for example soil fertility. For the log Gaussian Cox processes studied in Møller et al. (1996), see section 2.2, the intensity process is a log Gaussian process.

The Neyman-Scott processes constitute another important class of point processes. A Neyman-Scott process is generated by a three-step procedure: First, a Poisson process Z of parents is generated. Secondly, conditional on $Z = z$, *iid* random integers $J(z_i)$ are generated for each parent point $z_i \in z$. Thirdly, for each parent z_i, $J(z_i)$ *iid* offspring are generated according to a density g for the position of an offspring relative to the parent. The realization of the Neyman-Scott process is finally given by the superposition of the offspring. A Neyman-Scott process where the number of offspring is Poisson distributed is equivalent to a Cox process with intensity surface given by

$$\Lambda_s = \sum_{z_i\in Z} g(s - z_i), \ s \in S,$$

see Bartlett (1964). There may for Neyman-Scott processes as for Cox processes often be a natural interpretation as it is indicated by the terminology of “parents” and “offspring”. Log Gaussian Cox processes and certain Neyman-Scott processes are compared in section 4 of Møller et al. (1996).

Clustering of points is for Cox processes due to peaks in the underlying intensity process, while clusters of Neyman-Scott processes appear around the unobserved parent points. The points are in both cases conditionally independent given the underlying intensity or parent process. The class of *Markov point processes* (Ripley & Kelly, 1977; Baddeley & Møller, 1989) allows for modelling of clustering or repulsion which arise from direct attractive or repulsive interactions between the points. The definition of Markov point processes resembles that of MRF's. Suppose that a symmetric and reflexive relation $\sim$ on $\mathbb{R}^d$ is given, and let $\mathcal{C}$ denote the set of cliques given by $\sim$ (excluding the empty clique). For any point configuration $x \subseteq S$, we let $\mathcal{C}(x) = \{C \in \mathcal{C} | \, C \subseteq x\}$, i.e. $\mathcal{C}(x)$ is the set of $\sim$-cliques contained in x. The density with respect to μ_S of a finite $\sim$-Markov point process on a bounded subset $S \subseteq \mathbb{R}^d$, is of the form

$$f(x) \propto \prod_{C\in\mathcal{C}(x)} \phi(C), \ x \in N(S),$$

where ϕ is a nonnegative clique interaction function. A Markov point process X on $S = \mathbb{R}^d$ is specified by a set of conditional densities $\{f_B(\cdot|\cdot)\}_{B \in \mathcal{B}_0(S)}$ (see Preston, 1976) in a similar way that infinite lattice MRF's and MCCF's are specified, see section 1.2.4. For $A \in \mathcal{B}_0(S)$, f_A is the conditional density with respect to μ_A of $X \cap A$ given $X \setminus A$, and f_A is given in terms of the clique interaction function ϕ as

$$f_A(x \cap A | x \setminus A) \propto \prod_{C \in \mathcal{C}(x): C \cap A \neq \emptyset} \phi(C). \tag{2.2}$$

A standard example of a repulsive finite Markov point process is the Strauss process for which $s_1 \sim s_2 \Leftrightarrow \|s_1 - s_2\| \leq r$, $s_1, s_2 \in \mathbb{R}^d$, where $r \geq 0$ is the interaction radius. For parameters $\theta_1 \in \mathbb{R}$ and $\theta_2 \leq 0$, the clique interaction function is given by $\phi(s_1) = \exp(\theta_1)$, $\phi(\{s_1, s_2\}) = \exp(\theta_2), \{s_1, s_2\} \in \mathcal{C}$, and $\phi(C) = 1$ if $|C| > 2$. The density is thus

$$f(x) \propto \exp\Big(\theta_1 |x| + \frac{\theta_2}{2} \sum_{x_i \in x} \sum_{x_j \in x \setminus x_i} 1(\|x_i - x_j\| \leq r)\Big). \tag{2.3}$$

In section 3.4 we discuss simulation of the hard core process which is another example of a repulsive Markov point process.

For Markov point processes there has, as for MRF's, traditionally been much focus on pairwise interaction models like the Strauss process. Such models are useful for modelling of repulsive point patterns, but it is now generally believed that pairwise interaction models are not adequate models for clustered point patterns. More flexible models may be obtained by including higher-order interactions. The area-interaction process (Baddeley & Lieshout, 1995) has interactions of any order, and is in principle a model for both repulsive and clustered point patterns. My experience is that it is difficult from this model to generate point patterns which are clearly repulsive or clustered when judged visually, or in terms of the K-function (the K-function is defined in section 2.2.3). Geyer (1996) introduces two higher-order processes of which the first is the triplets process which appears by extending the sufficient statistic of the Strauss process (2.3) with the number of cliques of cardinality three. The second process, named the saturation process, is a modification of the Strauss process where the sum $\sum_{x_j \in x \setminus x_i} 1(\|x_i - x_j\| \leq r)$ in the sufficient statistic is replaced by $\min\Big(c, \sum_{x_j \in x \setminus x_i} 1(\|x_i - x_j\| \leq r)\Big)$ for a positive parameter $c \geq 0$. The triplets process and the saturation process are models both for clustering and repulsion, and from these models it is easy to generate patterns which are clearly repulsive or clustered. The continuum random cluster model (Møller, 1996) is also a flexible model for clustered and repulsive point patterns. This model belongs to the class of nearest neighbour Markov point processes (Baddeley & Møller, 1989) for which MCCF's with a background colour (see section 1.2.1 and section 1.2.3) are the discrete analogues.

Markov point processes may e.g. be applied to investigate the hypothesis of complete spatial randomness, i.e. whether data can be described by a homogeneous Poisson process. This hypothesis is traditionally investigated by considering various second order summary statistics like the K, F and G functions, see section 2.2.3. If the data is well described by a parametric model for a Markov point process, the likelihood ratio test probably provides a stronger alternative. A potential weakness of this approach is that parameters in models for Markov point processes do not always have a natural interpretation. It is e.g. not clear what to conclude from the triplets model if complete spatial randomness is rejected. This is, of course, of minor importance if it is just required to obtain a flexible prior for some Bayesian prediction procedure, as in Bayesian image analysis.

Maximum likelihood estimation for Markov point processes was once considered a difficult problem due to the unknown normalizing constant, but is now in general quite straightforward since the normalizing constant can be estimated accurately by MCMC. The likelihood for a Neyman-Scott process or a Cox process is given in terms of a mean value with respect to the distribution of the underlying parent or intensity process, and is in general analytically intractable. Also in this case MCMC may be used to estimate the likelihood as explained in Gelfand & Carlin (1991) and Geyer (1994), see also section 3.3. MCMC methods are in this case required to obtain conditional simulations of the intensity/parent process, given the data. It is not clear when likelihoods for Cox processes and Neyman-Scott processes are unimodal and free of local maxima, and maximization of the likelihood may therefore be difficult in practice. Minimum contrast methods based on the K-function have been the usual approach to fitting of Neyman-Scott processes, see Diggle (1983). In Møller et al. (1996) (see section 2.2.2) a minimum contrast method based on the covariance function of the Gaussian field yields a useful and computationally simple method to fit parametric models for log Gaussian Cox processes.

Nonparametric statistical methods based on summary statistics for point processes in general require stationarity, and nonstationary point patterns have mainly been analyzed as realizations of inhomogeneous Poisson processes, see e.g. Cressie (1991), and the references therein (Bayesian inference for the intensity surface of an inhomogeneous Poisson process is also discussed in section 2.2.5). It may also be of interest to investigate possible interactions between the points. Ogata & Tanemura (1986) consider approximate maximum likelihood estimation for inhomogeneous Gibbs processes, and use the AIC criterion for comparison of fitted models. Baddeley et al. (1997) study another class of models where the observed point pattern is assumed to be a nonstationary thinning of a stationary Markov point process. It is discussed how the K-function may be useful for analysis of such point patterns, and how MCMC may be used to perform semiparametric inference, see also section 2.3.

2.2 Log Gaussian Cox processes

Log Gaussian Cox processes (Møller et al., 1996) provide a flexible and tractable class of models for clustered point patterns, where the clustering of points is due to underlying spatial heterogeneity. The intensity surface Λ of a log Gaussian Cox process (LGCP) is given by $\Lambda_s = \exp(Y_s)$, $s \in \mathbb{R}^d$, where $Y = \{Y_s\}_{s\in\mathbb{R}^d}$ is a Gaussian random field. To ensure that the random mean measure M given by (2.1) is well-defined, it is assumed that Λ is given in terms of a continuous modification of Y. A condition on the covariance of Y which ensures the existence of a continuous modification is given in Lemma 1 in Møller et al. (1996). Multivariate LGCP's are straightforwardly defined when Y is a multivariate Gaussian field, see section 5 in Møller et al. (1996).

We shall restrict attention to stationary LGCP's. By Theorem 1 in Møller et al. (1996), stationarity of a LGCP implies stationarity of the corresponding Gaussian field Y, and the distribution of a stationary LGCP is thus uniquely determined by the mean $\mu \in \mathbb{R}$, the variance $\sigma^2 > 0$, and the correlation function $r : \mathbb{R}^d \to \mathbb{R}$ of Y.

In applications, a parametric correlation model is used. Table 2.1 contains examples of isotropic correlation models for which the condition in Lemma 1 in Møller et al. (1996) holds. The simulations in Figure 2 in Møller et al. (1996) show that a wide range of clustered point patterns can be generated from LGCP's when the correlation models in Table 2.1 and varying values of the parameters μ, σ^2, and $\beta > 0$, are used.

Gaussian:	$\exp(-(a/\beta)^2)$	Exponential:	$\exp(-a/\beta)$
Cardinal sine:	$\sin(a/\beta)/(a/\beta)$	Stable:	$\exp(-\sqrt{a/\beta})$

Table 2.1: Examples of correlation models.

2.2.1 Product densities and ergodicity

The n'th order product density $\rho^{(n)} : (\mathbb{R}^d)^n \to \mathbb{R}_+$, $n = 1, 2, \ldots$ of a point process X determines the n'th order moments of the random variables $|X \cap B|$, $B \in \mathcal{B}_0(S)$, and intuitively, $\rho^{(n)}(s_1, \ldots, s_n)ds_1 \cdots ds_n$ is the probability that the point process has precisely one point in each of n disjoint and infinitesimally small regions of volume $ds_1, \ldots, ds_n$. The first order product density $\rho^{(1)}(s) = \rho(s)$, $s \in \mathbb{R}^d$, is the intensity, and the pair correlation function g is given by $g(s_1, s_2) = \rho^{(2)}(s_1, s_2)/(\rho^2(s_1)\rho^2(s_2))$, $s_1, s_2 \in \mathbb{R}^d$.

For a Cox process, the product densities are given by

$$\rho^{(n)}(s_1, \ldots, s_n) = E \prod_{i=1}^{n} \Lambda_{s_i}$$

for distinct $s_1, \ldots, s_n \in \mathbb{R}^d$. For a stationary LGCP the following results (Theorem 1 in Møller et al., 1996) are easily obtained:

$$\rho = \exp(\mu + \sigma^2/2), \tag{2.4}$$

$$g(s_1 - s_2) = \exp\Big(\sigma^2 r(s_1 - s_2)\Big), \tag{2.5}$$

and

$$\begin{aligned} \rho^{(n)}(s_1, \ldots, s_n) &= \exp\Big(n\mu + \sigma^2\Big(\frac{n}{2} + \sum_{1 \le i < j \le n} r(s_i - s_j)\Big)\Big) \\ &= \rho^n \prod_{1 \le i < j \le n} g(s_i - s_j),\ n > 2. \end{aligned} \tag{2.6}$$

From (2.4) and (2.5) it follows that the distribution of a stationary log Gaussian Cox process is uniquely determined by the intensity and the pair correlation function. This becomes useful for construction of methods for parameter estimation in section 2.2.2. The expression (2.6) is used to construct a third order summary statistic in section 2.2.3.

Theorem 3 in Møller et al. (1996) relates ergodicity of a Cox process with ergodicity of the intensity process. A sufficient condition for ergodicity of a LGCP is

$$g(s) \to 1 \text{ as } \|s\| \to \infty. \tag{2.7}$$

2.2.2 Parameter estimation

Suppose that a realization $x = \{x_1, \ldots, x_n\}$ of $X \cap W$ is observed, where X is a stationary and isotropic LGCP, and $W \subset \mathbb{R}^d$ is a bounded observation window. Suppose further that a parametric model $r(\cdot;\beta)$ is chosen for the correlation function of the Gaussian field. The likelihood of $\theta = (\mu, \sigma^2, \beta)$ is then

$$L(\theta) = E_\theta\Big(\exp\Big(\int_W \Big(1 - \exp(Y_s)\Big) ds\Big) \prod_{i=1}^n \exp\Big(Y_{x_i}\Big)\Big). \tag{2.8}$$

This likelihood is in general analytically intractable. MCMC may be used to estimate the likelihood, see section 3.3, but this approach is computationally demanding since MCMC samples are required for a range of values of θ. This problem is especially prominent if a grid search is required to find the estimate.

Since the distribution of a LGCP is completely determined by the intensity and the pair correlation function, we instead propose to base inference on these summary statistics. Let $\hat{\rho} = |x|/\nu_d(W)$ be the natural estimate of the intensity, and let $\hat{g}$ be a nonparametric kernel estimate of g, see e.g. Stoyan & Stoyan

(1994). A nonparametric estimate of the covariance function of Y is given by $\log \hat{g}(a)$, $a \geq 0$, and estimates of σ^2 and β are obtained by minimization of

$$\int_{\epsilon}^{a_0} \Big(\big(\log \hat{g}(a)\big)^{\alpha} - \big(\sigma^2 r_\beta(a)\big)^{\alpha} \Big)^2 da, \tag{2.9}$$

where $\epsilon \geq 0$, $\alpha > 0$, and $a_0 > 0$ are user specified parameters. The parameters α and a_0 are determined by the shape of $\log \hat{g}$ and r_β, while ϵ is usually taken to be $\min_{i \neq j} \|x_i - x_j\|$. For fixed β, (2.9) is minimized at

$$\hat{\sigma}^2_\beta = \Big(\frac{B(\beta)}{A(\beta)}\Big)^{1/\alpha} \text{ with } B(\beta) = \int_{\epsilon}^{a_0} \big(\hat{c}(a) r_\beta(a)\big)^{\alpha} da,\ A(\beta) = \int_{\epsilon}^{a_0} \big(r_\beta(a)\big)^{2\alpha} da,$$

provided $B(\beta) > 0$. By inserting these expressions into (2.9) and using (2.4), the estimates

$$\hat{\beta} = \arg\max \frac{B(\beta)^2}{A(\beta)},\ \hat{\sigma}^2 = \hat{\sigma}^2_{\hat{\beta}},\ \hat{\mu} = \log(\hat{\rho}) - \hat{\sigma}^2/2,$$

are obtained. In our examples, the function $B(\beta)^2/A(\beta)$ was unimodal and easy to maximize, so the minimum contrast method appears to be a computationally feasible alternative to maximum likelihood estimation.

2.2.3 Model checking

The usual second order summary statistics for point processes is the empty space statistic F, the nearest neighbour distribution function G, and the reduced second moment measure K, given by

$$F(a) = P(X \cap b(0,a) \neq \emptyset),\ G(a) = P_0^!(X \cap b(0,a) \neq \emptyset),$$
$$\text{and } K(a) = \frac{1}{\rho} E_0^!\big(\#X \cap b(0,a)\big) = 2\pi \int_0^a b g(b) db,\ a > 0.$$

($P_0^!$ denotes the reduced Palm distribution at 0, and $E_0^!$ is expectation with respect to $P_0^!$, see Stoyan, Kendall & Mecke, 1995).

In Møller et al. (1996) a third order summary statistic z is introduced. This summary statistic is given by

$$z(a) = \frac{1}{\pi^2 a^4 \rho^2} E_0^! \sum_{\substack{x_1, x_2 \in X: \\ \|x_1\| \leq a, \|x_2\| \leq a}}^{\neq} \big(g(x_1) g(x_2) g(x_1 - x_2)\big)^{-1} =$$
$$\frac{1}{\pi^2 a^4 \rho^3} \int_{\|\xi\| \leq a} \int_{\|\eta\| \leq a} \frac{\rho^{(3)}(0, \xi, \eta)}{g(\xi) g(\eta) g(\xi - \eta)} d\xi d\eta,\ a > 0. \tag{2.10}$$

From (2.6) it follows that

$$z(a) = 1,\ a > 0,\ \text{for a log Gaussian Cox process.}$$

For a realization x of a stationary point process X and some observation window W, the sum

$$\sum_{x_1,x_2,x_3\in X: x_1\in W,\,\|x_2-x_1\|\le a,\,\|x_3-x_1\|\le a}^{\neq} \Big(g(x_2-x_1)g(x_3-x_1)g(x_2-x_3)\Big)^{-1} \tag{2.11}$$

is an unbiased estimate of $\nu_d(W)\pi^2a^4\rho^3z(a)$. This estimate is, however, not useful in practice since it includes unobserved points in $x\setminus W$. One solution is minus sampling where W in (2.11) is replaced by $W_{\ominus a}=\{s\in W\,|\, s+u\in W\;\forall\, u\in\mathbb{R}^2:\|u\|\le a\}$, so that the sum only includes x_1's for which all neighbouring points within the distance a are observed. The unbiased estimate of $\nu_d(W_{\ominus a})\pi^2a^4\rho^3z(a)$ thus obtained is especially for large a quite inefficient since not all triplets of points in $x\cap W$ contribute to the estimate. A more efficient estimate is proposed in Møller et al. (1996) in the case where X is a stationary and isotropic point process on $\mathbb{R}^2$. This estimate makes use of all observed triplets of points, but introduces an edge correction factor to account for possibly unobserved triplets which intersect W. For given $x_1\in W$, $a>0$, $b>0$, and $0\le\psi<2\pi$, let

$$U_{x_1,a,b,\psi}=\{\phi\in[0,2\pi[\,|\,x_1+a(\cos\phi,\sin\phi)\in W,\; x_1+b\Big(\cos(\phi+\psi),\sin(\phi+\psi)\Big)\in W\}.$$

The edge correction factor is

$$w_{x_1,a,b,\psi}=\frac{2\pi}{\nu_1(U_{x_1,a,b,\psi})},$$

where $\nu_1(U_{x_1,a,b,\psi})$ is the length of $U_{x_1,a,b,\psi}$, and Theorem 2 in Møller et al. (1996) states that

$$2\sum_{x_1\in X\cap W}\;\sum_{\substack{\{x_2,x_3\}\subseteq X\cap W\setminus\{x_1\}:\\ \|x_2-x_1\|\le a,\|x_3-x_1\|\le a}}^{\neq}\frac{w_{x_1,\|x_2-x_1\|,\|x_3-x_1\|,\psi(x_1,x_2,x_3)}}{g(x_2-x_1)g(x_3-x_1)g(x_2-x_3)} \tag{2.12}$$

is an unbiased estimate of $\nu_2(W)\pi^2a^4\rho^3z(a)$. Here $\psi(x_1,x_2,x_3)$ denotes the angle (anticlockwise) between x_2-x_1 and x_3-x_1. In practice, ρ and g are substituted with nonparametric estimates.

In Example 1 in Møller et al. (1996) a LGCP is fitted to the locations of 126 scots pine saplings, and the summary statistics F, G, z, and L (given by $L(a)=\sqrt{K(a)/\pi}$) were used to check the model assumption. The left plot in Figure 2.1 shows the estimated z for the data, and two sets of envelopes based on 20 unconditional simulations of the fitted log Gaussian Cox process and 20 simulations conditional on the observed number of points. The plot gives no reason to doubt the model assumption no matter whether the 'unconditional' or 'conditional' envelopes are considered.

In Penttinen, Stoyan & Henttonen (1992) and Stoyan & Stoyan (1994), a Matérn cluster process is applied as a model for the scots pines data. To check

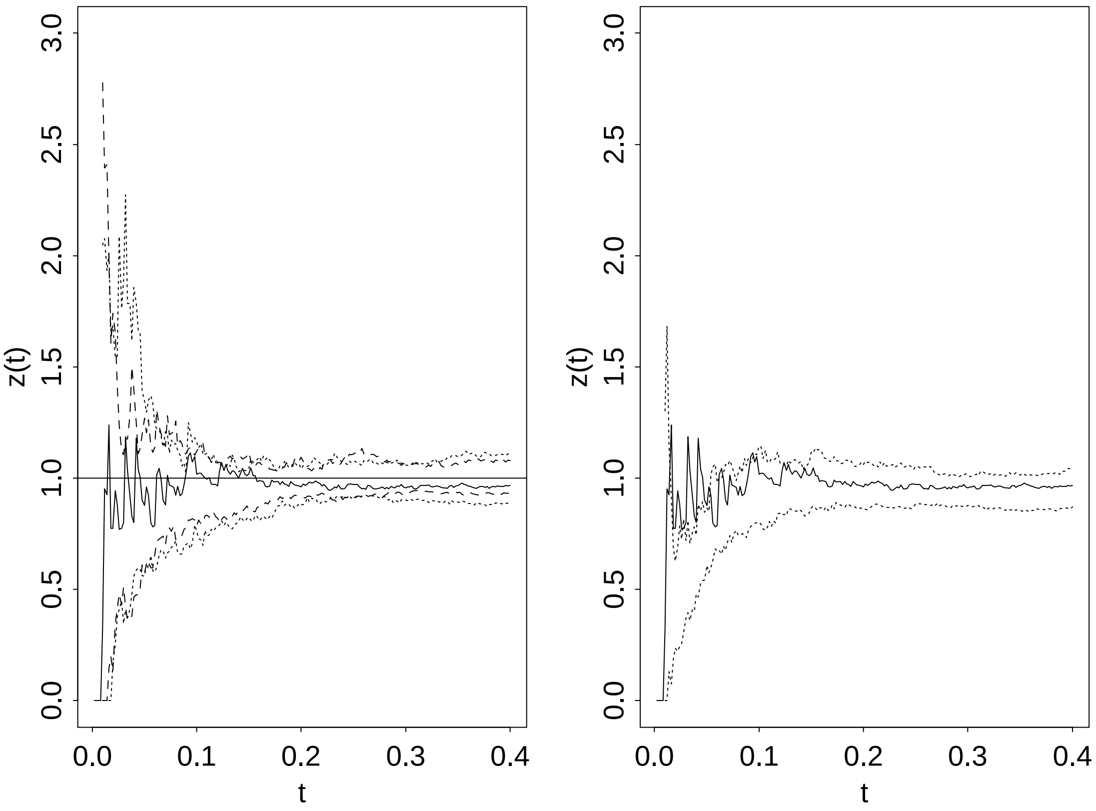

Figure 2.1: Example 1. Estimate of z based on the data (solid line) and 'conditional' envelopes (– – –) and 'unconditional' envelopes (- - - - -) based on 20 simulations. Left: Log Gaussian Cox process. Right: Matérn cluster process.

the discriminatory power of z we also calculated envelopes for the Matérn cluster process fitted by Stoyan & Stoyan (1994), see the right plot in Figure 2.1. The estimated z-function based on the data crosses the envelopes in an interval of t-values, so the plot raises serious doubt concerning the appropriateness of the Matérn cluster process as a model for the data.

2.2.4 Prediction and Bayesian inference

Suppose that x is an observation of a LGCP X within a bounded observation window W. Apart from estimation of the parameters μ, σ^2, and β, it may also be of interest to make inference about the unobserved intensity surface or Gaussian field. For this purpose it is natural to consider the conditional distribution of Λ or Y given $X \cap W = x$. This conditional distribution is not analytically tractable, but is well suited for evaluation by MCMC. Conditional simulations of Λ are also useful in connection with prediction of unobserved parts of the point process.

Suppose for specificity that $W = [0,1[^2$, and approximate $Y_W = \{Y_s\}_{s \in W}$ by the field $\tilde{Y}$ given by $\tilde{Y}_{(s_1,s_2)} = Y_{(i/l,j/l)}$ if $i/l \leq s_1 < (i+1)/l$ and $j/l \leq$

$s_2 < (j+1)/l$, $i, j \in I$, where $I = \{0, 1/l \ldots, (l-1)/l\}^2$, and $l \geq 1$ is a suitable value for the discretization. The random field $(\tilde{Y}(s))_{s \in I}$ is extended to a larger random field $\tilde{Y}_{ext}$ (see section 6 of Møller et al., 1996) defined on $I_{ext} = \{0, 1/l \ldots, (2(l-1)-1)/l\}^2$, where the marginal distribution of $(\tilde{Y}_{ext}(s))_{s \in I}$ is identical to the distribution of $(\tilde{Y}(s))_{s \in I}$, and the covariance matrix K of $\tilde{Y}_{ext}$ is a circulant matrix. This is advantageous since a circulant matrix is easy to diagonalize by means of the FFT. From the diagonalization we get that $Y_{ext} \stackrel{\mathcal{D}}{=} \Gamma Q + \mu_{ext}$ where $\Gamma \sim N_k(0, I)$, k is the rank of K, Q is a certain $k \times (2(l-1))^2$ matrix of rank k, and $\mu_{ext} = (\mu)_{s \in I_{ext}}$. The random field Γ is easily transformed into Y_{ext} by using the FFT, and we actually choose to obtain conditional simulations of $\tilde{Y}$ from conditional simulations of Γ given $X \cap W = x$.

The log conditional density of Γ is

$$\log f(\gamma|x) = \text{const}(x) - \frac{1}{2}\|\gamma\|^2 + \sum_{s \in I_{ext}} \Big(\tilde{y}_s n_s - \frac{\exp(\tilde{y}_s)}{l^2}\Big),$$

where $n_{(s_1,s_2)} = |x \cap [s_1, s_1 + 1/l[\times [s_2, s_2 + 1/l[|$, $(s_1, s_2) \in I_{ext}$ and $\tilde{y}_{ext} = \gamma Q + \mu_e xt$. The gradient of the log conditional density is

$$\bigtriangledown(\gamma) = -\gamma + (n_s - \frac{\exp(\tilde{y}_s)}{l^2})_{s \in I_{ext}} Q^*,$$

and it is easy to see that $\partial \bigtriangledown(\gamma)/\partial\gamma^*$ is strictly negative definite so that the conditional distribution is strictly log-concave. This is a nice property since multimodal target distributions are a major worry in applications of MCMC.

The conditional simulations are generated by a Metropolis adjusted Langevin algorithm (MALA) as suggested by Besag (1994) and further studied in Roberts & Tweedie (1997). This is a Metropolis-Hastings algorithm (see section 3.2) inspired by the Langevin diffusion, and the proposal kernel for the algorithm is given by $N_d(\gamma^{(m)} + (h/2) \bigtriangledown (\gamma^{(m)}), hI)$ where $h > 0$ is a user specified parameter, and $\gamma^{(m)}$ is the current state of the chain. The use of the gradient in the mean of the proposal kernel in general leads to faster convergence than when e.g. a random walk chain is used (Roberts & Rosenthal, 1995). Note that the MALA is different from typical MCMC algorithms in that all components in $\gamma^{(m)}$ are updated simultaneously.

The MALA is not geometrically ergodic (see section 3.1), but by replacing $\bigtriangledown(\gamma)$ in the proposal kernel with a truncated version

$$\bigtriangledown(\gamma)^{trunc} = -\gamma + \Big(n_s - \frac{\big(H \wedge \exp(\tilde{y}_s)\big)}{l^2}\Big)_{s \in I_{ext}} Q^*,$$

a geometrically ergodic chain is obtained whenever $0 < h < 2$ and $0 < H < \infty$ (Theorem 4 in Møller et al., 1996).

From conditional simulations of the model fitted in Example 1 in Møller et al. (1996) we estimated the posterior mean and variance of the Gaussian field, and the posterior mean of the intensity surface, see Figure 2.2 and Figure 2.3. For comparison also a nonparametric kernel estimate (Diggle, 1985) is shown. The band width of the kernel was chosen by minimizing an estimate of the mean square error. Diggle (1985)'s expression for the mean square error is derived for the kernel given by the uniform density on a disk, but can straightforwardly be generalized to arbitrary kernels by application of the Campbell formula. Maximum a posteriori estimates of the Gaussian field and the intensity surface are also calculated in Møller et al. (1996).

The conditional mean and the nonparametric estimate of the intensity surface for the scots pine data are quite different, since the conditional mean is much more peaked than the nonparametric estimate. In a simulation experiment in Møller et al. (1996) where the intensity surface was known, the conditional intensity was the best estimate, but further research is required to draw definite conclusions.

2.2.5 Bayesian estimation of the intensity surface of an inhomogeneous Poisson process

Heikkinen & Arjas (1996) propose a Bayesian method for estimation of the intensity surface of an inhomogeneous Poisson process. The realizations of the prior for the intensity surface are constant within Voronoi cells generated by a homogeneous Poisson process, and conditional on the Voronoi cells, the log intensity levels in each cell are modelled by a conditional autoregression (CAR) (Besag, 1974), i.e. a Gaussian MRF. MCMC is used to estimate posterior means and variances.

The individual step function realizations of the prior and the posterior are considered as rather crude approximations of realistic intensity surfaces, but Heikkinen & Arjas (1996) argue that this is not crucial in Bayesian inference where emphasis is on posterior probabilities and means. The authors e.g. point out that the posterior mean of the intensity surface is smoothly varying despite of the discontinuities of the individual posterior surfaces. One drawback of the method in Heikkinen & Arjas (1996) is that there does not seem to be a feasible method to choose the prior parameters on a data-driven basis. Estimation of prior parameters from data does not, strictly speaking, even make sense when the "true" intensity surface is not considered as a realization of the prior.

If the Gaussian field of a LGCP is considered as a prior model for Bayesian estimation of the intensity surface, then this prior has several advantages. First, there is a simple method for estimation of the prior parameters from the data, and secondly, the prior modelling can be checked by using the second and third order summary statistics described in section 2.2.3. This is valid only if a fine discretization of the Gaussian field is used, and the high dimensionality of the Gaussian field is for computational reasons a disadvantage when compared to Heikkinen & Arjas (1996)'s approach.

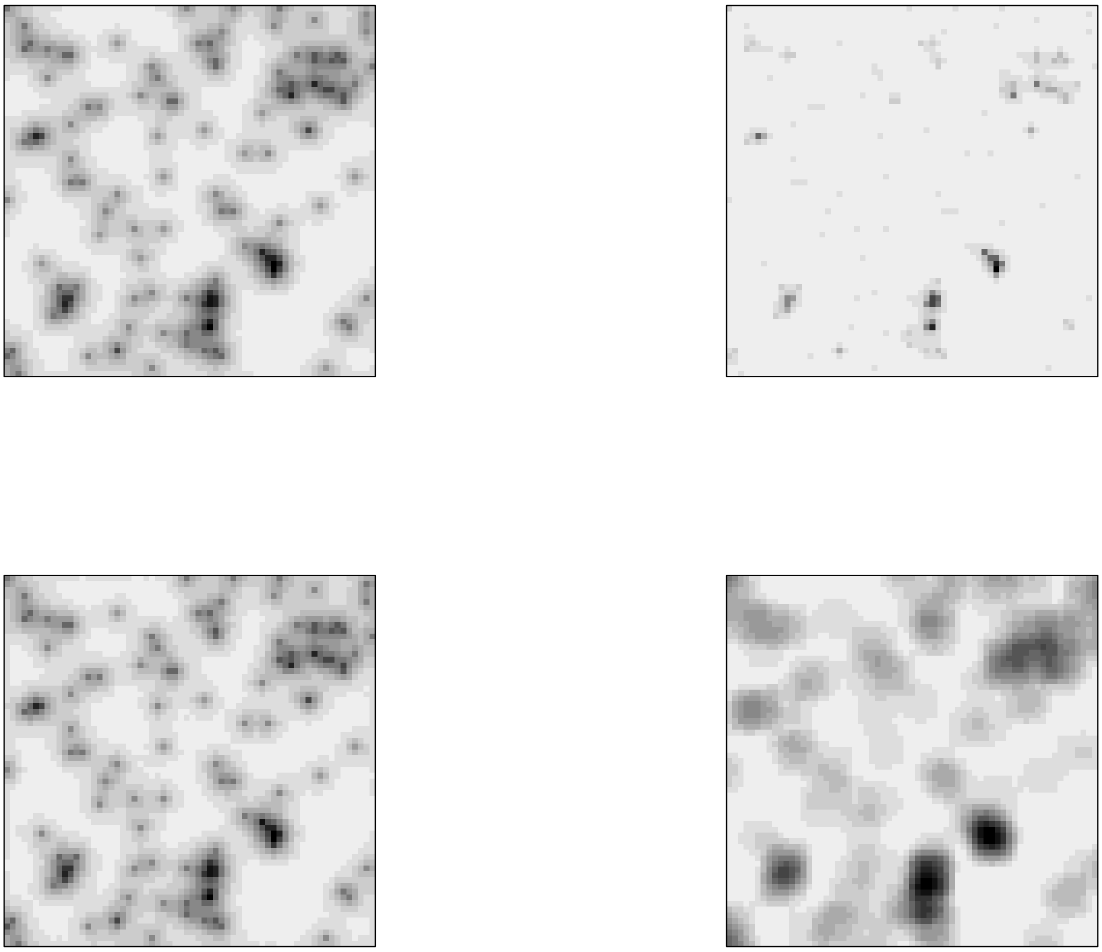

Figure 2.2: Scots pines data. Upper left plot: Monte Carlo posterior mean of the Gaussian field. Upper right plot: Monte Carlo posterior mean of the intensity surface. Lower left plot: Logarithm to the upper right plot. Lower right plot: The nonparametric kernel estimate of the intensity surface.

A third Bayesian approach is suggested in Wolpert & Ickstadt (1995). In this paper the random mean measure of the Cox process is given by a convolution of a smoothing kernel and a Gamma random field with independent increments. The smoothing kernel is needed in order to introduce spatial correlation. A prior is specified for the kernel and Gamma field parameters, and a Bayesian inference is carried out by MCMC. The prior parameters have an interpretation in relation to the data and this is a helpful support for the subjective choice of the prior parameters. The specification of the model and the MCMC scheme in Wolpert & Ickstadt (1995) is involved, and it is not clear whether their model is advantageous in terms of flexibility compared to a LGCP.

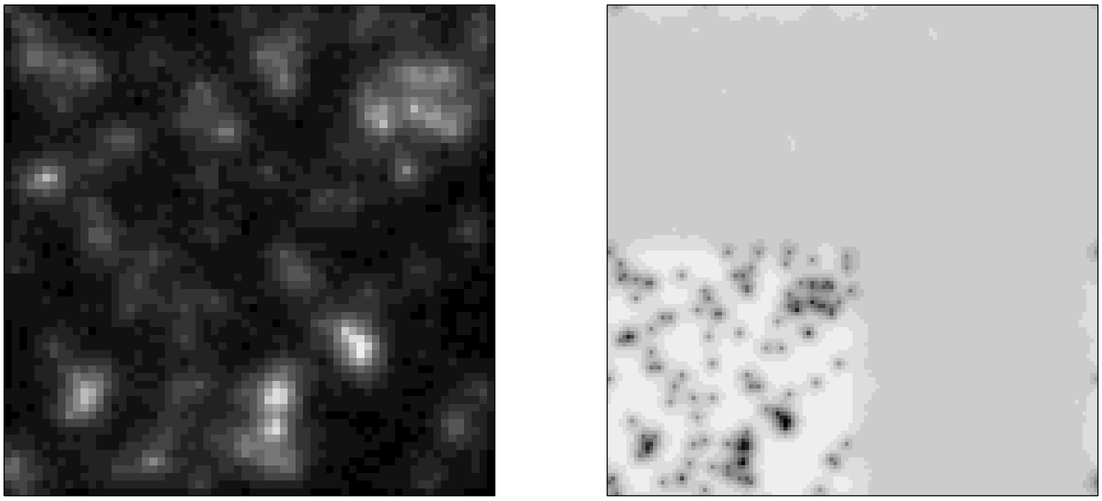

Figure 2.3: Left: Monte Carlo posterior variance of the Gaussian field on the original lattice. Right: Monte Carlo posterior mean of the Gaussian field on the extended lattice.

2.3 Thinned Markov point processes

Common hypotheses in statistics for spatial point patterns concern presence or absence of interaction between neighbouring points. Most methods of statistics for spatial point patterns assume stationarity, but it may also be of interest to be able to account for nonstationary variations in the point pattern.

One example is environmental epidemiology (see Diggle, 1993) where one is interested in possible spatial clustering of locations of disease cases, but where e.g. spatial fluctuations in population density also influence the incidence of disease cases. When the hypothesized source of variation is known, Diggle (1993) suggests applying a parametric model for an inhomogeneous Poisson process. In e.g. Diggle (1990) the intensity surface $\lambda(\cdot)$ of the Poisson process is given by a semiparametric model

$$\lambda(s) = \psi\lambda_0(s)w(s - s_0; \theta),\ s \in S,$$

where ψ and θ are parameters, λ_0 is a model for the underlying spatial heterogeneity, and $w(\cdot)$ models a possible raised incidence of disease cases near a source of pollution located at s_0. The surface λ_0 is in the likelihood replaced by a kernel estimate obtained from a point pattern of control locations. If the source of spatial variation is unknown or unobservable, Diggle (1993) suggests modelling the data by a stationary Cox process whereby the K-function can be used in the inference as in Diggle & Chetwynd (1991).

There are situations where the approaches proposed by Diggle (1993) are not applicable. Consider the point pattern in Figure 2.4 of "adult" longleaf-pines from the longleaf-pines data (Platt, Evans & Rathbun, 1988). The nonparametric

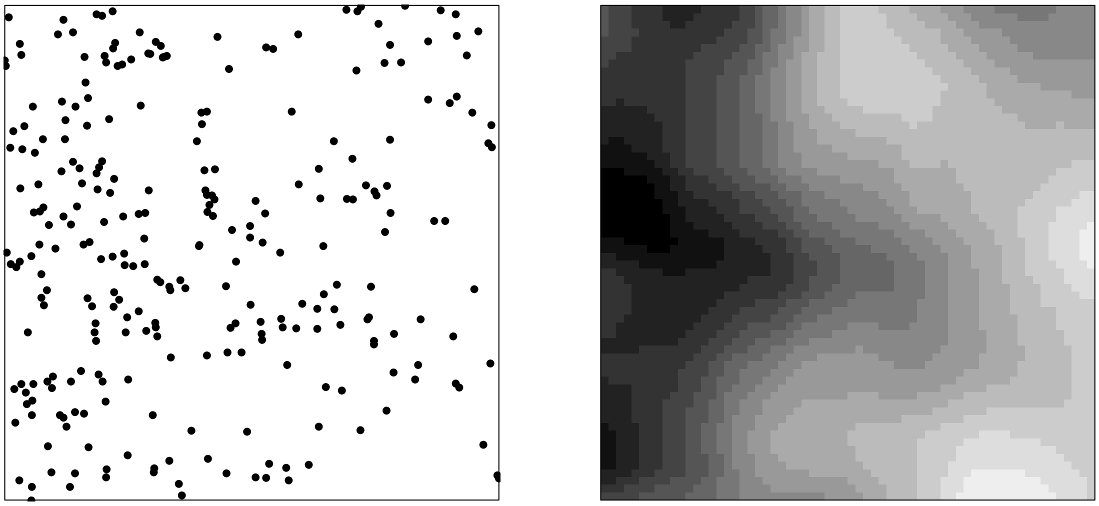

Figure 2.4: Left: the locations of the 271 adult longleaf pines. Right: nonparametric kernel estimate of the intensity surface.

estimate of the intensity surface in Figure 2.4 clearly exhibits a decreasing trend in the east-west direction. In the longleaf-pine data set there are no explanatory variables for this trend. The point pattern may, of course, in principle be regarded as a part of a realization of a stationary process, but this is not helpful since the observation window is too small, compared to the scale on which stationarity provides the replication needed for inference. The stands of the adult longleaf pines could possibly be modelled by an inhomogenous Poisson process with a log intensity surface given e.g. by a polynomial model

$$\log \lambda\Big((s_1, s_2)\Big) = \sum_{l=0}^{L} \sum_{k=0}^{K} a_{lk} {s_1}^l {s_2}^k,\ (s_1, s_2) \in S, \tag{2.13}$$

where $L,\ K \geq 0$, and $a_{lk} \in \mathbb{R}$, but this does not allow for modelling of interaction like e.g. repulsion due to competition among the trees.

One approach to modelling of nonstationary point patterns with interactions at a local scale is to use inhomogeneous Markov point processes or inhomogeneous Gibbs processes as in Ogata & Tanemura (1986). We restrict here attention to inhomogeneous Markov point processes which are specified by conditional densities of the following form:

$$f_A(x \cap A | x \setminus A; \theta) \propto \prod_{x_i \in A} \alpha(x_i; \theta) \prod_{\substack{C \in \mathcal{C}(x): \\ |C|>1 \text{ and } C \cap A \neq \emptyset}} \phi(C; \theta),\ A \in \mathcal{B}_0(\mathbb{R}^d). \tag{2.14}$$

Here θ is the unknown model parameter, the large scale variability is modelled by the activity field $\alpha = \Big(\alpha(s; \theta)\Big)_{s \in \mathbb{R}^2}$, and the interaction functions $\phi(\cdot; \theta)$ are typically assumed to be translation and rotation invariant. Ogata & Tanemura

(1986) actually apply a parametric model similar to (2.13) for the log activity field.

Another approach suggested in Baddeley et al. (1997) is to model a nonstationary point pattern as a realization of a thinned Markov point process X. Let $p = \big(p(s)\big)_{s\in\mathbb{R}^d}$ be the deterministic "thinning surface", where $p(\cdot) \in [0,1]$ is a measurable function. This is used for *independent* $p(\cdot)$-thinning of X: for a given realization x of X, a point x_i in x is removed with probability $1 - p(x_i)$ and retained with probability $p(x_i)$ independently of the other points in x. Let W be a bounded observation window. Then the density of the observed thinned point pattern $Y \cap W$ given $X = x$ is

$$f(y|x;p) = 1(y \subseteq x) \prod_{x_i \in y} p(x_i) \prod_{x_i \in (x\cap W)\setminus y} \Big(1 - p(x_i)\Big). \tag{2.15}$$

The intensity surface λ_Y of Y is given by

$$\lambda_Y(s) = \rho p(s),\ s \in \mathbb{R}^d, \tag{2.16}$$

where ρ is the intensity of X. The reduced second order measure (see e.g. Stoyan *et al.*, 1995) of Y is further absolutely continuous with respect to $\nu_d \times \nu_d$ and its density is $\rho^2 g(s_1 - s_2)p(s_1)p(s_2)$ (see Cressie, 1991, section 8.5.6), where g is the pair correlation function of X. Thus, g is the pair correlation function of Y too. This becomes useful for development of summary statistics and methods for model checking, see section 2.3.1. Inference for a thinned Markov point process is discussed in section 2.3.2.

2.3.1 Model checking

Suppose that $y = \{y_1, \ldots, y_n\}$ is a realization of $Y \cap W$, and that X is isotropic and second order stationary so that the pair correlation function $g = \rho^{(2)}/\rho^2$ is invariant under rigid motions in $\mathbb{R}^d$, i.e. $g(s_1, s_2) = g(||s_1 - s_2||)$. For all $t > 0$ with

$$\nu_d(\{s \in W \mid \partial b(s,r) \cap W \neq \emptyset\}) > 0 \text{ for all } r < t,$$

it is easily verified that an unbiased estimator of the K-function (see section 2.2.3) is given by

$$\hat{K}(t) = \frac{1}{\nu_d(W)} \sum_i \sum_{j\neq i} \frac{w_{y_i,y_j} 1(||y_i - y_j|| \leq t)}{\rho p(y_i) \rho p(y_j)} \tag{2.17}$$

where w_{s_1,s_2}, $s_1, s_2 \in W$, is the Ripley (1977) edge correction factor given by $2\pi/U_{s_1,s_2}$, where

$$U_{s_1,s_2} = \nu_1(\{v \in [0; 2\pi[:\ s_1 + ||s_2 - s_1||(\cos v, \sin v) \in W\}).$$

In practice, $\rho p(\cdot)$ must be replaced by an estimate of the intensity surface. From the Campbell formula it is straightforward to obtain an expression for the variance of $\hat{K}$, see Baddeley et al. (1997).

It is also possible to define analogues to the empty space distribution function and the nearest-neighbour distribution function. Assume for simplicity that the intensity surface $\lambda_Y(\cdot) > 0$ of Y is positive, and let for any $t > 0$ and $s \in \mathbb{R}^d$, the number $r(s,t)$ be determined by

$$t = \int_{b(s,r(s,t))} \lambda_Y(s')ds'.$$

The analogue of the empty space distribution function is

$$F_s(t) = P\Big(Y \cap b\big(s, r(s,t)\big) \neq \emptyset\Big),\ t > 0,$$

and the "nearest-neighbour" distribution function is

$$G_s(t) = P_s\Big(Y \cap b\big(s, r(s,t)\big) \neq \emptyset\Big),\ s > 0,$$

where P_s denotes the Palm-distribution of Y with respect to s. When Y is a Poisson process, $F_s(t)$ and $G_s(t)$ do not depend on s, and they are both equal to $1 - \exp(-t)$.

2.3.2 Semiparametric inference for a thinned Markov point process

Suppose to begin with that parametric models depending on parameters θ and ψ have been chosen for the Markov point process X and the thinning field p. The likelihood of θ and ψ is given by the following result in Baddeley et al. (1997):

Proposition 1 *Let y be an observation of the thinned process $Y \cap W$. The likelihood of the parameters θ and ψ given y is*

$$L(\theta,\psi) = \prod_{y_i \in y} p_\psi(y_i) E_\theta[\prod_{x_i \in X \cap W} \big(1 - p_\psi(x_i)\big)\lambda(y|X;\theta)] \tag{2.18}$$

Here $\lambda(\cdot|\cdot)$ is the *Papangelou conditional intensity* which for disjoint point patterns y and x is given by

$$\lambda(y|x;\theta) = \prod_{C \in \mathcal{C}(y \cup x;\theta): C \cap y \neq \emptyset} \phi(C;\theta), \tag{2.19}$$

where ϕ is the interaction function of X. The question of asymptotic normality is difficult to address for the likelihood (2.18), and bootstrap methods may therefore be required for inference concerning θ and ψ.

In practice, the mean value in (2.18) may be approximated by the mean value with respect to the finite point process $\tilde{X}$ given by the density

$$\tilde{f}(x;\theta) \propto \prod_{C \in \mathcal{C}(x)} \phi(C;\theta) \tag{2.20}$$

on an extended rectangular window $\tilde{W}$ where $W \subseteq \tilde{W}$. When a toroidal edgecorrection is used, i.e. $\tilde{W}$ is wrapped on a torus, the mean value with respect to X is hopefully well approximated by the mean value with respect to (2.20), when $\tilde{W}$ is chosen sufficiently large. The approximate likelihood is then

$$\tilde{L}(\theta,\psi) = \Big(\prod_{y_i \in y} p_\psi(y_i)\Big) \int \prod_{z_i \in z \cap W} \Big(1 - p_\psi(z_i)\Big) \tilde{f}(y \cup z;\theta)\mu_{\tilde{W}}(dz) =$$

$$\Big(\prod_{y_i \in y} p_\psi(y_i)\Big) \int \prod_{x_i \in (W \cap x)\setminus y} \Big(1 - p_\psi(x_i)\Big) \tilde{f}(x;\theta)\mu_{\tilde{W},y}(dx) \tag{2.21}$$

where $\mu_{\tilde{W}}$ is the unit rate Poisson process on $\tilde{W}$, and $\mu_{\tilde{W},y}$ is the Palm distribution of $\mu_{\tilde{W}}$ at y which is given by $Z \cup y \sim \mu_{\tilde{W},y}$ when $Z \sim \mu_{\tilde{W}}$.

The integral in (2.21) is not known in closed form, but may be estimated by MCMC methods as described in Gelfand & Carlin (1991) and Geyer (1994), see section 3.3.

In some situations it is difficult to suggest a suitable parsimonious parametric model for p, and it may then be desirable to estimate p nonparametrically while maximum likelihood estimation is used for θ. According to (2.16), p may be estimated up to the constant $1/\rho$ by an estimate of the intensity surface of Y. Suppose that $\hat{\lambda}_Y(\cdot)$ is a nonparametric kernel estimate of the intensity surface for Y, and that $\rho(\theta)$ is the intensity of X for a given θ. Following standard practice in semiparametric statistics, the corresponding estimate of p given by

$$\hat{p}(s) = \hat{\lambda}_Y(s)/\rho(\theta),\ s \in W,$$

may then be plugged into the likelihood in order to obtain a profile likelihood from which an estimate of θ can be obtained. One may run into problems of identifiability with this approach since the conditional density of X will typically contain an "activity" parameter which controls the intensity of X, and whose effect may in practice be difficult to distinguish from the effect of the thinning field: a model with a large value of ρ combined with small values of $p(\cdot)$ may produce similar realizations as a model with a small ρ but large $p(\cdot)$.

To solve the identifiability problem one may restrict p to the set $\{q : W \to [0,1] |\ \sup_s q(s) = 1\}$. When interest is focused on the local interactions this assumption should not affect the flexibility of the model seriously. A pragmatic estimate of p is then

$$\hat{p}(s) = \hat{\lambda}_Y(s)/\sup_{s \in S} \hat{\lambda}_Y(s) \tag{2.22}$$

which does not depend on θ. One thereby also avoids Monte Carlo estimation of $\rho(\theta)$ which is in general not known in closed form for a Markov point process. If p_ψ is replaced by the nonparametric estimate (2.22), the approximate likelihood simplifies to

$$\tilde{L}(\theta) \equiv \int \prod_{x_i \in (W \cap x) \setminus y} \Big(1 - \hat{p}(x_i)\Big) \tilde{f}(x;\theta) \mu_{\tilde{W},y}(dx) \tag{2.23}$$

which is easier to handle than (2.21) since the maximization is now only with respect to θ which is typically just two or threedimensional. The semiparametric approach is probably most useful if X is repulsive since variability due to clustering may be captured by small peaks in $\hat{\lambda}_Y$, unless some smoothness constraint is imposed on $\hat{\lambda}_Y$.

2.3.3 Discussion

In practice, realizations of an inhomogeneous Markov point process and a thinned stationary Markov point process may be impossible to distinguish. There are, however, several advantages in using the thinned Markov point process setup. Summary statistics may be defined and estimated, and there is a simple correspondence between the intensity surface of the thinned point process, and the thinning surface. The model thereby gives a clear distinction between the nonstationary large scale variation and the homogeneous local interactions. This is in contrast to the inhomogeneous Markov point process setup where summary statistics are usually not available, and there is not a simple correspondence between the intensity surface and the "activity" field. This for example means that a nonparametric estimate for the activity field is not available.

One drawback of using the thinned point process compared to the inhomogeneous Markov point process is that maximum likelihood estimation becomes computationally demanding, especially when a parametric model is chosen for the thinning surface.

The K-function has been fitted to the japanese black pine data (Numata, 1964) and the adult longleaf pines in Figure 2.4 as proposed in section 2.3.1. It turns out that the nonparametric kernel estimate of the intensity surface is not useful in this context due to overfitting. Better results can be obtained when a parametric model like (2.13) is fitted to the data under the Poisson assumption. We have not yet applied the semiparametric method suggested in section 2.3.2 to real data, and further research is needed in order to judge the usefulness of both the estimate (2.17) and the semiparametric approach.

2.4 An alternative to hierarchical Bayesian modelling

Heikkinen & Arjas (1996)'s approach to modelling of intensity surfaces (see section 2.2.5) is similar to the image models introduced in Nicholls (1996) (section 1.2.7) in that tesselations of a subset D of $\mathbb{R}^2$ are generated from a random set of points. In Heikkinen & Arjas (1996) the model is specified hierarchically. The points which generate the Voronoi tesselation are given by a homogeneous Poisson process of intensity $\lambda > 0$, and conditional on a point pattern $x = \{x_1, \dots, x_n\}$, the log intensity levels $l = (l_1, \dots, l_n)$ for each cell given by x are modelled by a CAR. The observed data is finally an inhomogeneous Poisson process where the intensities $\lambda(s)$, $s \in D$, are given by $\lambda(s) = l_i$ if s belongs to the i'th cell.

Let $g(x) \propto \lambda^{|x|}$ be the density of the Poisson process with intensity λ, and let $f(l|x) = c(x)h(l;x)$ be the conditional density of the levels, where $c(x)$ is the normalizing constant. When a state (l', x') is proposed given a current state (l, x) during the MCMC-sampling, the ratio $c(x')h(l';x')g(x')/(c(x)h(l;x)g(x))$ appears in the Metropolis-Hastings ratio. This is problematic when $x \neq x'$ since the normalizing constants $c(x')$ and $c(x)$ are in general not available in closed form. Heikkinen & Arjas (1996) apply a local approximation to $c(x')/c(x)$, but it is not clear how serious bias this approximation introduces.

Alternatively, one may proceed as in Nicholls (1996) and start by specifying a joint distribution for points and levels with respect to a measure defined similarly to (1.22), so that there is just one unknown normalizing constant which cancels out in the Metropolis-Hastings ratio. Specifically, one might for subsets A of the space of point configurations and levels, define a measure by

$$\chi(A) = \exp\Big(-\nu_2(D)\Big) \sum_{n=0}^{\infty} \frac{1}{n!} \int_{D^n} dx \int_{\mathbb{R}^n} 1\Big((\{x_1, \dots, x_n\}, l) \in A\Big) dl, \tag{2.24}$$

and model the joint distribution of points and levels by the density

$$g(x, l) = c\lambda^{|x|} h(l; x) \tag{2.25}$$

with respect to χ. A sufficient condition for this density to be well-defined is $c(x)^{-1} < K^{|x|}$ for some constant K, and all finite $x \subset D$. Under the model (2.25) the conditional density of the levels is again $f(l|x)$ while the marginal distribution of the points is a point process with density

$$g(x) = \frac{c\lambda^{|x|}}{c(x)} \tag{2.26}$$

with respect to μ_D.

A similar approach is used in Melas & Wilson (1997) for modelling of multitype textured images. Suppose that $p(x) = c_1 \exp\big(-U(x)\big)$, $x \in V^I$, is a model for the labelling of the image pixels $i \in I$ into different texture types in V, and that for a given labelling x, the observed image $y \in \mathbb{R}^I$ is modelled by an inhomogeneous CAR with density $\tilde{f}(y|x) = c_2(x)\exp\bigl(-V(y;x)\bigr)$. The joint density of the observed image and the underlying labelling process is then

$$\tilde{f}(x,y) = c_1 c_2(x)\exp\bigl(-U(x) - V(y;x)\bigr).$$

Also in this case the unknown normalizing constants $c_2(x)$, $x \in V^I$, cause problems when the label process is updated in the MCMC algorithm. In Melas & Wilson (1997), $U(\cdot)$ is the potential of a MRF, and the joint distribution of the observed image and the labelling process is modelled by a bivariate MRF given by the density

$$\tilde{g}(x,y) = c_3 \exp\bigl(-U(x) - V_x(y)\bigr),$$

where the unknown normalizing constant c_3 cancels out in the MCMC calculations. The marginal density

$$\tilde{g}(x) = c_3 \exp\bigl(-U(x)\bigr)/c_2(x) \tag{2.27}$$

of the labelling process under this model is different from $p(x)$ and is in particular not a MRF, while the conditional densities $\tilde{g}(y|x) = \tilde{g}(x,y)/\tilde{g}(x)$ and $\tilde{f}(y|x)$ are identical.

I think that it would be quite interesting to further compare the two approaches to Bayesian modelling. It might for example be of interest to study the prior models $g(\cdot)$ and $\tilde{g}(\cdot)$ given by (2.26) and (2.27) in more detail. Simulations of these models can easily be obtained from simulations of the joint distributions.

3 Markov chain Monte Carlo

Probability distributions which are intractable analytically occur frequently in spatial and Bayesian statistics. Also direct simulation is often not possible, but it is typically quite easy to simulate an ergodic Markov chain whose stationary distribution is the distribution of interest, and samples required e.g. for Monte Carlo approximations can thereby be obtained. Markov chain Monte Carlo (MCMC) is used in Møller & Waagepetersen (1996) to obtain MCMC maximum likelihood estimates, in Waagepetersen (1997*b*) for estimation of normalizing constants, and in Møller et al. (1996) for calculation of posterior means and variances. In Mase, Møller, Stoyan & Waagepetersen (1997) the MCMC method of simulated tempering (Marinari & Parisi, 1992; Geyer & Thompson, 1995) is combined with the Geyer & Møller (1994) algorithm to simulate hard core processes with high packing densities, see section 3.4.

An overview of basic notions of MCMC is given in section 3.1, and section 3.2 gives an account of the Metropolis-Hastings (Hastings, 1971; Green, 1995) algorithm. MCMC maximum likelihood is briefly described in section 3.3. Instructive references of Markov chain Monte Carlo are Tierney (1994), Besag, Green, Higdon & Mengersen (1995), and Geyer (1992).

3.1 Basic notions of Markov chain Monte Carlo

Consider a target distribution π defined on a measure space $(E, \mathcal{E})$, and a Markov chain $X = (X_n)_{n \in \mathbb{N}}$ given by an initial distribution π_0 and m-step transition kernels

$$P^m(x, A) = P(X_m \in A \mid X_0 = x) \tag{3.1}$$

for $m \geq 1$, $A \in \mathcal{E}$ and $x \in E$. Then

- π is an *invariant measure* for X if $\forall A \in \mathcal{E}$,

$$\int P(x, A)\pi(dx) = \pi(A).$$

- X is *aperiodic* if there does not exist a disjoint subdivision of E into subsets $A_0, \dots, A_{d-1}$, $d \geq 2$, such that

$$\forall x \in A_i : \; P(x, A_{(i+1) \bmod d}) = 1.$$

- X is *π-irreducible* if for all $x \in E$ and all $A \in \mathcal{E}$ with $\pi(A) > 0$ there exist an $m \geq 1$ such that $P^m(x, A) > 0$.

- X is *Harris recurrent* if

$$P(\exists m : X_m \in A \mid X_0 = x) = 1$$

  for all $x \in E$ and $A \in \mathcal{E}$ with $\pi(A) > 0$.

Define $\bar{f}_n = 1/n \sum_{i=1}^{n} f(X_i)$ for any function $f \in L_1(\pi)$. Suppose that X is irreducible and that π is an invariant measure for X. Then π is the unique stationary distribution of X (Tierney, 1994, Theorem 1) and conditional on $X_1 = x$,

$$\bar{f}_n \to E_\pi(f) = \int f d\pi \text{ for } \pi\text{-a.a } x \in E$$

(Geyer, 1995, Theorem 2). If X is Harris recurrent then by Theorem 17.1.7 in Meyn & Tweedie (1993) this law of large numbers holds for any initial condition

$x \in E$. If further X is aperiodic, then X becomes Harris ergodic and then by Theorem 1 in Tierney (1994),

$$\lim_{m\to\infty} \|P^m(x,\cdot) - \pi\| = 0 \text{ for all initial conditions } x \in E, \tag{3.2}$$

where $\|\cdot\|$ is the *total variation norm* which for a (signed) measure λ on $(E, \mathcal{E})$ is defined by

$$\|\lambda\| = \sup_{A\in\mathcal{E}} |\lambda(A)| - \inf_{A\in\mathcal{E}} |\lambda(A)|.$$

Some simple conditions for Harris recurrence are given by Corollary 1 and 2 in Tierney (1994), and Theorem 1 in Chan & Geyer (1994). If the chain is geometrically ergodic, i.e.

$$\|P^m(x,\cdot) - \pi\| \le M(x)\rho^m \ \forall x \in E \tag{3.3}$$

for some measurable function $M : E \to [0;\infty[$ in $L_1(\pi)$ and $0 < \rho < 1$, then a central limit theorem

$$\sqrt{n}\Big(\bar{f}_n - E_\pi(f)\Big) \to N\Big(0, \sigma^2(f)\Big)$$

holds for all f which are in $L_{2+\epsilon}(\pi)$ for some $\epsilon > 0$ (Chan & Geyer, 1994, Theorem 2), and if the chain is *uniformly ergodic*, i.e. $M(\cdot)$ is less or equal to a positive constant M, then the central limit theorem holds for all $f \in L_2(\pi)$ (Theorem 5, Tierney, 1994). The variance $\sigma^2(f)$ is given by

$$\sigma^2(f) = Var\Big(f(X_1)\Big) + 2\sum_{i=2}^{\infty} Cov(X_1, X_i)$$

for the stationary chain X, i.e. where $\pi_0 = \pi$.

The state space of the models considered in Møller & Waagepetersen (1996) is finite, and the Markov chains used for the simulations are then uniformly ergodic by Proposition 2 in Tierney (1994). In Møller et al. (1996) a chain for simulation of a posterior on $\mathbb{R}^k$ is constructed, and geometrical ergodicity is established using results in Roberts & Tweedie (1997) which are again based on a certain general drift condition (Theorem 15.0.1 in Meyn & Tweedie, 1993).

3.2 The Metropolis-Hastings kernel

This description of the Metropolis-Hastings kernel is based on Green (1995) where the original Metropolis-Hastings algorithm is generalized to state spaces of varying dimension.

The basic ingredients in the Metropolis-Hastings algorithm are a proposal kernel $Q : E \times \mathcal{E} \to [0,1]$, and a set of acceptance probabilities given by a function $a(\cdot,\cdot)$ defined on E^2. Suppose that there is a symmetric measure ξ on $(E \times E, \mathcal{E} \otimes \mathcal{E})$ such that $\forall A, B \in \mathcal{E}$,

$$\int_A Q(x,B)\pi(dx) = \int_{A\times B} f(x,y)\xi(dx,dy), \tag{3.4}$$

where f is a finite density for which

$$a(x,y)f(x,y) = a(y,x)f(y,x), \text{ for } \xi\text{-a.a } (x,y) \in E \times E. \tag{3.5}$$

It is then easy to check that the transition kernel P given by

$$P(x,A) = \int_A a(x,y)Q(x,dy) + 1(x \in A)\int(1-a(x,y))Q(x,dy),\ x \in E,\ A \in \mathcal{E},$$

is reversible, i.e.

$$\int_A P(x,B)\pi(dx) = \int_B P(x,A)\pi(dx)\ \forall\ A, B \in \mathcal{E},$$

whereby it follows that π is the invariant measure of P.

Given an $x \in E$, a value Z is sampled from $P(x,\cdot)$ by first generating a proposal Y from $Q(x,\cdot)$. With probability $a(x,Y)$ the proposal Y is accepted so that $Z = Y$, and otherwise $Z = x$. Under the condition (3.5), the acceptance probabilities are usually chosen to be maximal, so that

$$a(x,Y) = \min\{1, \frac{f(Y,x)}{f(x,Y)}\}.$$

Note that $f(x,Y) > 0$ for π-a.a. x by definition of f.

The advantage of the Metropolis-Hastings algorithm is that we are free to choose a proposal kernel which is easy to sample from, as long as the resulting chain becomes irreducible and aperiodic. In the simplest cases π and $Q(x,\cdot)$, $x \in E$, are absolutely continuous with respect to the same reference measure on E. In section 8 of Møller & Waagepetersen (1996), π is e.g. a posterior on $\mathbb{R}^k$, and $Q(x,\cdot)$ is a certain nondegenerate Gaussian distribution on $\mathbb{R}^k$ for each $x \in \mathbb{R}^k$, see also section 2.2.4. In this case $\xi = \nu_k \otimes \nu_k$ where ν_k is the Lebesgue measure on $\mathbb{R}^k$, and $f(x,y) = g(x)q(x,y)$ where g and q denotes the densities of π and $Q(x,\cdot)$, respectively.

In many applications, Q and π are actually singular. Such cases are discussed in the next two sections.

3.2.1 Combinations of Metropolis-Hastings kernels

Proposal kernels Q are often constructed so that $Q(x,\cdot)$ and π are singular for each $x \in E$. The corresponding kernel P is therefore not π-irreducible, but

Metropolis-Hastings kernels corresponding to different proposal kernels can be combined to obtain an irreducible chain.

Consider for example the case where E is a product space, $E = \times_{i=1}^{s} E_i$, $s \geq 1$, and assume that π has a density g with respect to $\mu = \otimes_{i=1}^{s} \mu_i$, where μ_i is a measure on E_i. Let for $i = 1, \ldots s$, the proposal kernel Q_i be defined by

$$Q_i(x, A) = \int_{\{z \in E_i | (z, x_{-i}) \in A\}} q_i(x, z) \mu_i(dz), \ \forall x \in E, \ A \in \mathcal{E},$$

where for each $x \in E$, $q_i(x, \cdot)$ is a density with respect to μ_i. Then for each $i = 1, \ldots, s$, (3.4) and (3.5) holds with ξ_i defined on sets $A \times B \in \mathcal{E} \times \mathcal{E}$ by

$$\xi_i(A \times B) = \int 1(x \in A, (z, x_{-i}) \in B) \mu_i(dz) \mu_1(dx_1) \cdots \mu_s(dx_s),$$

and f_i given by

$$f_i(x, y) = g(x) q_i(x, y_i), \ \forall x, y \in E.$$

For a given $x \in E$ the measure $Q_i(x, \cdot)$ is concentrated on $\{y \in E \mid y_{-i} = x_{-i}\}$, so the corresponding transition kernels P_i, $i = 1, \ldots, s$, are in general not irreducible. Irreducibility may be obtained by systematic or random combination of the kernels $P_1, \ldots, P_s$. A systematic combination is

$$P_{sys} = P_1 \cdots P_s,$$

and a random combination is

$$P_{ran} = \sum_{i=1}^{s} \delta_i P_i,$$

where $\sum_{i=1}^{s} \delta_i = 1$ and $\delta_i > 0, \ i = 1, \ldots, s$. Here

$$P_1 P_2(x, A) = \int P_2(z, A) P_1(x, dz)$$

The well-known Gibbs-sampler appears as the special case where $q_i(x, z) = g_i(z | x_{-i})$ is the conditional density for the i'th component given the rest. For the simulations in section 6 of Møller & Waagepetersen (1996) the proposal density was $q_i(x, z) = 1/2$, $z \in \{0, 1\}$, $x \in \{0, 1\}^I$, $i \in I$.

3.2.2 An algorithm for simulation of finite point processes

The Geyer & Møller (1994) algorithm used in Baddeley et al. (1997) and Mase et al. (1997) for simulation of finite point processes falls into the framework of Green (1995). Let S be a bounded subset of $\mathbb{R}^d$, and define $N(S)$ and $\mathcal{N}(S)$ as in section 2.1. The distribution μ_S of the unit rate Poisson process on S is given by

$$\mu_S(F) = \mathrm{e}^{-\nu_d(S)} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{S^n} 1(\{x_1, \ldots, x_n\} \in F) dx, \ F \in \mathcal{N}(S). \quad (3.6)$$

Suppose that $x = \{x_1, \ldots, x_n\}$ is the current point configuration and that the target point process π is given by a density g with respect to μ_S. The simplest version of the Geyer & Møller (1994) proposal kernel is given as follows: With probability 1/2 it is proposed to delete a point sampled from the uniform distribution on x (unless $x = \emptyset$, in which case nothing happens), and with probability 1/2 it is proposed to add a point sampled from the uniform distribution on S. The proposal kernel is thus $Q = (Q_{remove} + Q_{insert})/2$ where

$$Q_{insert}(x, F) = \int_{\{\eta \in S: x \cup \eta \in F\}} \frac{1}{\nu_d(S)} d\eta,$$

and

$$Q_{remove}(x, F) = \sum_{\eta \in x} 1(x \setminus \eta \in F) \frac{1}{|x|}.$$

The measure ξ is for $F, G \in \mathcal{N}(S)$ given by

$$\xi(F \times G) = \int_F \int_{\{\eta \in S: x \cup \eta \in G\}} d\eta \mu_S(dx) + \int_F \sum_{\eta \in x} 1(x \setminus \eta \in G) \mu_S(dx).$$

Let $F_n = \{x \in F : |x| = n\}$ and $G_{n-1} = \{x \in G : |x| = n - 1\}$. Then

$$\xi(F_n \times G_{n-1}) = \mathrm{e}^{-\nu_d(S)} \frac{1}{n!} \int_{S^n} \sum_{\eta \in x} 1(x \in F_n, x \setminus \eta \in G_{n-1}) dx =$$

$$\mathrm{e}^{-\nu_d(S)} \frac{1}{n!} \int_{S^n} n 1(\{x_1, \ldots, x_n\} \in F_n, \{x_1, \ldots, x_{n-1}\} \in G_{n-1}) dx =$$

$$\mathrm{e}^{-\nu_d(S)} \frac{1}{(n-1)!} \int_{S^{n-1}} \int_S 1(y \in G_{n-1}, y \cup \eta \in F_n) dy d\eta = \xi(G_{n-1} \times F_n),$$

whereby it follows that ξ is symmetric. The density f is given by

$$f(x, x \cup \eta) = \frac{g(x)}{2\nu_d(S)} \text{ and } f(x \cup \eta, x) = \frac{g(x \cup \eta)}{2(|x| + 1)}, \; x \in N, \; \eta \in S.$$

3.3 MCMC maximum likelihood

Suppose that z is an observation of $Z \sim f_\theta(\cdot) = c(\theta) h_\theta(\cdot)$, where f_θ, $\theta \in \Theta$, is a parametric family of densities, and $c(\theta)$ is an unknown normalizing constant. If $(Z_i)_{i \geq 1}$ is an ergodic MCMC sample from an importance sampling density $f(\cdot) = ch(\cdot)$ where $h_\theta(z') > 0 \Rightarrow h(z') > 0$, $z' \in E$, then

$$\frac{c}{c(\theta)} = \int \frac{h_\theta(z')}{h(z')} f(z') dz' = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(Z_i)}{h(Z_i)}$$

so that the log-likelihood may be approximated by

$$l(\theta) \equiv \log\Big(h_\theta(z)\Big) - \log(\frac{c}{c(\theta)}) \approx \log\Big(h_\theta(z)\Big) - \log(\frac{1}{n} \sum_{i=1}^{n} \frac{h_\theta(Z_i)}{h(Z_i)})$$

for some large n. MCMC maximum likelihood is treated in detail in Geyer & Thompson (1992) and Geyer (1994).

MCMC may also be useful in missing data situations, i.e. where $Z = (X, Y)$, but only an observation x of X is available. Let g_θ denote the marginal density of X, and let $g(x) = \int f(x,y)dy$ be the marginal density of the first component of the importance sampling distribution. The log-likelihood may then (Gelfand & Carlin, 1991; Geyer, 1994) be approximated by

$$\begin{aligned}\log(\frac{g_\theta(x)}{g(x)}) = \log\Big(\int \frac{f_\theta(x,y)}{f(x,y)}\frac{f(x,y)}{g(x)}dy\Big) = \\ \log\Big(\int \frac{h_\theta(x,y)}{h(x,y)} f(y|x)dy\Big) - \log(\frac{c}{c(\theta)}) \approx \\ \log\Big(\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{h_\theta(x,Y_i^x)}{h(x,Y_i^x)}\Big) - \log\Big(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{h_\theta(Z_i)}{h(Z_i)}\Big) \quad (3.8)\end{aligned}$$

where $(Y_i^x)_{i\geq 1}$ is an ergodic sample from the conditional density $f(\cdot|x)$.

In order to obtain precise Monte Carlo estimates it is required that the importance sampling distribution places appreciable mass on the support of f_θ. The importance sampling distribution can e.g. be obtained as a mixture of f_θ's for a suitable range of θ's, as suggested in Geyer (1991).

3.4 Simulated tempering

Suppose that densities $f_1, \ldots, f_n$, and corresponding Metropolis-Hastings kernels $P_1, \ldots, P_n$ are given, where the Markov chain for simulation of f_1 mixes well, while the chains become highly autocorrelated when i increases.

The idea of simulated tempering (Marinari & Parisi, 1992; Geyer & Thompson, 1995) is to sample efficiently a mixture of $f_1, \ldots, f_n$ by combining the kernels $P_1, \ldots, P_n$ to obtain a chain which inherits the good mixing properties of the chains given by P_i for small i. Specifically, simulated tempering is Metropolis-Hastings simulation of the pair (X, I) distributed according to the density

$$\tilde{f}(x,i) = f_i(x)\delta_i,\ x \in E,\ i \in \{1, \ldots n\}.$$

Here I is a so-called auxiliary variable where $P(I = i) = \delta_i > 0$, the marginal distribution of X is the mixture $\sum_{i=1}^n f_i\delta_i$, and f_i is the conditional density of $X|I = i$.

A proposal kernel Q on $\{1, \ldots, n\}$ is defined by $Q(i, i+1) = Q(i, i-1) = 1/2$ for $1 < i < n$, and $Q(1,2) = Q(n, n-1) = 1$, and given a current state (x, i), the two components are updated in turn using first the proposal kernel Q_i corresponding to P_i, and secondly Q. The Metropolis-Hastings ratio for the update $(x,i) \rightarrow (x', i)$ is identical to the ratio for the update $x \rightarrow x'$ for the chain given by P_i. The Metropolis-Hastings ratio for the update $(x',i) \to (x',i')$ is

$$\frac{f_{i'}(x')\delta_{i'}Q(i',i)}{f_i(x')\delta_i Q(i,i')} \tag{3.9}$$

which depends on the normalizing constants c_i and $c_{i'}$ of f_i and $f_{i'}$. Estimates $\hat{c}_i, i \in \{1,\dots,n\}$ can up to a constant of proportionality be obtained by stochastic approximation as described in Geyer & Thompson (1995), or by reverse logistic regression (Geyer, 1991). By choosing $\delta_i \propto \hat{c}_i/c_i$ we get an approximate uniform mixture and the ratio (3.9) can be evaluated.

Suppose e.g. that f_n is multimodal, and that the chain given by P_n gets stuck around the modes. One may then for $T_1 > T_2 > \ldots > T_n = 1$ take

$$f_i(x) \propto \left(f_n(x)\right)^{1/T_i},\ i = 1,\dots,n-1,$$

to be "flattened" versions of f_n so that the first component $(X_l)_{l\geq 1}$ of the simulated tempering chain $(X_l, I_l)_{l\geq 1}$ moves quickly through the state space of X when I_l is small. If the pairs T_i, T_{i+1}, $i = 1,\dots,n-1$, are sufficiently close so that reasonable acceptance rates (20%-40%) for transitions $(x,i) \leftrightarrow (x, i\pm 1)$ are obtained, the chain $(X_l)_{l\geq 1: I_l = n}$ yields a well-mixed sample from f_n.

In Mase et al. (1997) simulated tempering is used for simulation of the hard core point process given by the density

$$f(x) \propto \alpha^{|x|} 1(\forall x_i, x_j \in x : ||x_i - x_j|| > r), \tag{3.10}$$

where $r \geq 0$ is the hard core parameter, and α is the activity parameter. In Mase et al. (1997) it is shown that the hard core process converges to a uniform distribution on the set of point patterns with maximal packing density as $\alpha \to \infty$. The properties of the model for high packing densities (i.e. when α is large) is further studied through simulations. The hard core process can in principle be simulated by using the Geyer & Møller (1994) algorithm directly, but this algorithm mixes very slowly when α and the packing density becomes high. The algorithm gets stuck when a highly packed point pattern is reached, since a new point can basically only be inserted where another point has just been removed.

Simulated tempering is used to obtain an algorithm with better mixing properties as follows: Let

$$f_i(x) \propto \alpha_i^{|x|} D(x, \gamma_i, r),\ i = 1,\dots,n,$$

where D is a penalizing term such that $D(x,\gamma,r)$ is 1 if $\gamma = 1$ or x is a hard core point pattern, $D(x,\gamma,r) \to 0$ if x is not a hard core point pattern and $\gamma \to 0$, and $D(x,0,r) = 1(||x_i - x_j|| > r$ for all $x_i, x_j \in x)$. Decreasing values $1 > \gamma_1 > \cdots > \gamma_{n-1} > \gamma_n = 0$ are chosen so that f_i is close to the Poisson density and easy to simulate when i is small, f_i approaches the hard core density when i increases, and f_n is the hard core density.

In Mase et al. (1997) the Geyer & Møller (1994) proposal kernels Q_i, $i = 1, \ldots, n$, for simulation of f_i, $i = 1, \ldots, n$, are given by $Q_i = a_{1,i}Q_{insert} + a_{2,i}Q_{remove} + a_{3,i}Q_{move}$, where $a_{k,i} \geq 0$ and $a_{1,i} + a_{2,i} + a_{3,i} = 1$. The kernel $Q_{move,i}$ proposes to move a randomly chosen point to a new position which is chosen according to the uniform distribution on a square $\epsilon_i \times \epsilon_i$-neighbourhood centered in the old position.

References

Baddeley, A. J. & Møller, J. (1989), 'Nearest-neighbour Markov point processes and random sets', *Int. Statist. Rev.* **57**, 89–121.

Baddeley, A. J. & Van Lieshout, M. N. M. (1993), Stochastic geometry models in highlevel vision, *in* 'Statistics and images, Advances in Applied Statistics, a supplement to the Journal of Applied Statistics', Vol. 20, pp. 231–256.

Baddeley, A. J. & Van Lieshout, M. N. M. (1995), 'Area-interaction processes', *Ann. Inst. Statist. Math.* pp. 601–619.

Baddeley, A. J., Møller, J. & Waagepetersen, R. (1997), 'Estimation of local interaction in nonstationary point patterns'. In preparation.

Baddeley, A. J., Van Lieshout, M. N. M. & Møller, J. (1995), 'Markov properties of cluster processes', *Adv. Appl. Prob. (SGSA)* **28**, 346–355.

Bartlett, M. S. (1964), 'Spectral analysis of two-dimensional point processes', *Biometrika* **44**, 299–311.

Besag, J. E. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *J. Roy. Statist. Soc. B* **36**, 192–236.

Besag, J. E. (1994), 'Comment on "Representations of knowledge in complex systems" by Grenander and Miller', *J. Roy. Statist. Soc. B* **56**, 591–592.

Besag, J. E., Green, P. J., Higdon, D. & Mengersen, K. (1995), 'Bayesian computation and stochastic systems', *Statist. Sci.* **10**, 3–66.

Carstensen, J. M. (1992), Description and simulation of visual texture, Ph.D.-afhandling nr. 59, Technical University of Denmark, Lyngby.

Chan, K. S. & Geyer, C. J. (1994), 'Discussion of the paper "Markov chains for exploring posterior distributions" by Luke Tierney', *Ann. Statist.* **22**, 1747–1758.

Chen, F. & Kelly, P. A. (1992), Algorithms for generating and segmenting morphologically smooth binary images, *in* 'Proceedings of the 26th Conference on Information Sciences and Systems', Princeton, New Jersey.

Clifford, P. (1990), Markov random fields in statistics, *in* G. R. Grimmett & D. J. A. Welsh, eds, 'Disorder in physical systems, A volume in honour of J. M. Hammersley', Clarendon Press, Oxford.

Cressie, N. (1991), *Statistics for Spatial Data*, Wiley.

Daley, D. J. & Vere-Jones, D. (1988), *An introduction to the theory of point processes*, Springer-Verlag.

Diggle, P. J. (1983), *Statistical analysis of spatial point patterns*, Academic Press.

Diggle, P. J. (1985), 'A kernel method for smoothing point process data', *Applied Statistics* **34**, 138–147.

Diggle, P. J. (1990), 'A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point', *J. Roy. Statist. Soc. A* **153**, 349–362.

Diggle, P. J. (1993), Point process modelling in environmental epidemiology, *in* V. Barnett & K. F. Turkman, eds, 'Statistics for the environment', Wiley, pp. 89–110.

Diggle, P. J. & Chetwynd, A. G. (1991), 'Second-order analysis of spatial clustering for inhomogeneous populations', *Biometrics* **47**, 1155–1163.

Dinten, J. M., Guyon, X. & Yao, J. F. (1991), On the choice of the regularization parameter: The case of binary images in the Bayesian restauration framework, *in* 'Spatial statistics and imaging', Lect. Notes., Inst. Math. Stat., AMS-IMS-SIAM joint summer conf., Hayward, pp. 55–77.

Gelfand, A. E. & Carlin, B. P. (1991), Maximum likelihood estimation for constrained or missing data models, Research Report 91-002, Division of Biostatistics, University of Minnesota.

Georgii, H.-O. (1988), *Gibbs measures and phase transitions*, Walter de Gruyter, Berlin.

Geyer, C. J. (1991), Reweighting Monte Carlo mixtures, Technical Report 568, School of Statistis, University of Minnesota.

Geyer, C. J. (1992), 'Practical Markov chain Monte Carlo (with discussion)', *Statist. Sci.* **7**, 473–511.

Geyer, C. J. (1994), 'On the convergence of Monte Carlo maximum likelihood calculations', *J. Roy. Statist. Soc. B* **56**, 261–274.

Geyer, C. J. (1995), Lecture notes on Markov chain Monte Carlo, Unpublished.

Geyer, C. J. (1996), Likelihood inference for spatial point processes, *in* O. E. Barndorff-Nielsen, W. S. Kendall & M. N. M. van Lieshout, eds, 'Proceedings Seminaire Européen de Statistique, "Stochastic geometry, likelihood and computation"', Chapman and Hall.

Geyer, C. J. & Møller, J. (1994), 'Simulation procedures and likelihood inference for spatial point processes', *Scand. J. Statist.* **21**, 359–373.

Geyer, C. J. & Thompson, E. A. (1992), 'Constrained Monte Carlo maximum likelihood for dependent data (with discussion)', *J. Roy. Statist. Soc. B* **54**, 657–699.

Geyer, C. J. & Thompson, E. A. (1995), 'Annealing Markov chain Monte Carlo with applications to pedigree analysis', *J. Amer. Statist. Assoc.* **90**, 909–920.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* pp. 711–732.

Grenander, U. (1993), *General Pattern Theory*, Oxford University Press.

Grenander, U. & Miller, M. (1994), 'Representations of knowledge in complex systems (with discussion)', *J. Roy. Statist. Soc. B* **56**, 549–603.

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Heikkinen, J. & Arjas, E. (1996), Nonparametric Bayesian estimation of a spatial poisson intensity, Preprint 20, Department of Statistics, University of Jyväskylä. To appear in Scand. J. Statist.

Hurn, M. & Rue, H. (1997), Bayesian object recognition, Statistics 6/1997, Norwegian University of Science and Technology, Trondheim.

Kesten, H. (1982), *Percolation theory for mathematicians*, Birkhäuser.

Kindermann, R. & Snell, J. L. (1980), *Markov random fields and their applications*, American Mathematical Society, Providence, RI.

Künsch, H. R. (1994), 'Robust priors for smoothing and image restoration', *Ann. Inst. Statist. Math.* **46**, 1–19.

Künsch, H. R., Geman, S. & Kehagias, A. (1995), 'Hidden Markov random fields', *Ann. Appl. Prob.* **5**, 577–602.

Marinari, E. & Parisi, G. (1992), 'Simulated tempering: a new Monte Carlo scheme', *Europhys. Lett.* **19**, 451–458.

Mase, S., Møller, J., Stoyan, D. & Waagepetersen, R. (1997), 'Intensities of hard-core Gibbs processes and the closest packing density'. In preparation.

Melas, D. E. & Wilson, S. P. (1997), Texture based image segmentation using the double MRF model, *in* K. V. Mardia, C. A. Gill & R. G. Aykroyd, eds, 'Proceedings in the Art and Science of Bayesian Image Analysis', Leeds University Press.

Meyn, S. P. & Tweedie, R. L. (1993), *Markov chains and stochastic stability*, Springer-Verlag, London.

Møller, J. (1996), Markov chain Monte Carlo and spatial point processes, *in* O. E. Barndorff-Nielsen, W. S. Kendall & M. N. M. van Lieshout, eds, 'Proceedings Seminaire Européen de Statistique, "Stochastic geometry, likelihood and computation"', Chapman and Hall.

Møller, J. & Waagepetersen, R. (1996), Markov connected component fields, Research Report 341, Dept. of Theoretical Statistics, University of Aarhus. To appear in Adv. Appl. Prob.

Møller, J., Syversveen, A.-R. & Waagepetersen, R. (1996), Log Gaussian Cox processes, Research Report 357, Dept. of Theoretical Statistics, University of Aarhus. To appear in Scand. J. Statist.

Nicholls, G. (1996), 'Bayesian image analysis with Markov chain Monte Carlo and colored continuum triangulation mosaics', *J. Roy. Statist. Soc. B.* To appear.

Numata, M. (1964), 'Forest vegetation, particularly pine stems in the vicinity of Choshi-flora and vegetation at Choshi, Chiba prefecture IV', *Bull. Choshi Marine Laboratory* (6), 27–37.

Ogata, Y. & Tanemura, M. (1986), Likelihood estimation of interaction potentials and external fields of inhomogeneous spatial point patterns, *in* I. S. Francis, B. F. J. Manly & F. C. Lam, eds, 'Proc. Pacific Statistical Congress-1985', pp. 150–154.

Omre, H. & Tjelmeland, H. (1996), 'Petroleum geostatistics', Invited lecture at the Fifth Geostatistical Congress, Wollongong, Sept. 22.-27.

Peierls, R. E. (1936), 'On Ising's ferromagnet model', *Proc. Camb. Phil. Soc.* **32**, 477–481.

Penttinen, A., Stoyan, D. & Henttonen, H. M. (1992), 'Marked point processes in forest statistics', *Forest Science* **38**, 806–824.

Preston, C. (1976), *Random Fields*, Springer-Verlag, Berlin.

Quian, W. & Titterington, D. M. (1991), 'Estimation of parameters in hidden Markov models', *Philos. Trans. Roy. Soc. London Ser. A* pp. 407–428.

Ripley, B. D. (1977), 'Modelling spatial patterns (with discussion)', *J. Roy. Statist. Soc. B* **39**, 172–212.

Ripley, B. D. (1988), *Statistical inference for spatial processes*, Cambridge University Press, Cambridge.

Ripley, B. D. & Kelly, F. P. (1977), 'Markov point processes', *J. Lond. Math. Soc.* **15**, 188–192.

Roberts, G. O. & Rosenthal, J. S. (1995), Optimal scaling of discrete approximations to langevin diffusions, Research Report 95-11, Statistical Laboratory, Cambridge University.

Roberts, G. O. & Tweedie, R. L. (1997), 'Exponential convergence of Langevin diffusions and their discrete approximations', *Bernoulli* **2**, 314–363.

Rudemo, M. & Stryhn, H. (1994), 'Approximating the distribution of maximum likelihood contour estimators in two-region images', *Scand. J. Statist.* **21**, 41–55.

Rue, H. (1996), Image restoration with faithful residuals, Statistics 11/1996, Norwegian University of Science and Technology, Trondheim.

Serra, J. (1982), *Image analysis and mathematical morphology*, Academic Press, London.

Sivakumar, K. & Goutsias, J. (1997), Morphologically constrained discrete random sets, *in* D. Jeulin, ed., 'Advances in theory and applications of random sets', World Scientific Publishing Company. To appear.

Stoyan, D. & Stoyan, H. (1994), *Fractals, Random Shapes and Point Fields*, Wiley, Chichester.

Stoyan, D., Kendall, W. S. & Mecke, J. (1995), *Stochastic Geometry and its Applications*, second edn, Wiley, New York.

Swendsen, R. H. & Wang, J. S. (1987), 'Nonuniversal critical dynamics in Monte Carlo simulations', *Phys. Rev. Letters* **58**, 86–88.

Syversveen, A.-R. & Omre, H. (1996), Marked point models for facies units conditioned on well data, *in* 'Proceedings of the Fifth Geostatistical Congress, Sept. 22.-27.', Wollongong.

Thompson, A. M., Brown, J. C., Kay, J. W. & Titterington, D. M. (1991), 'A study of methods of choosing the smoothing parameter in image restoration by regularization', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 326–338.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *Ann. Statist.* **22**, 1701–1728.

Tjelmeland, H. & Besag, J. E. (1996), Markov random field models for objects against background, Statistics 1/1996, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Waagepetersen, R. (1997*a*), Analysis of residuals from segmentation of noisy images, Research Report 380, Dept. of Theoretical Statistics, University of Aarhus.

Waagepetersen, R. (1997*b*), Phase transition and simulation for a penalized Ising model with applications in image analysis, Research Report 381, Dept. of Theoretical Statistics, University of Aarhus.

Wolpert, R. L. & Ickstadt, K. (1995), Poisson/Gamma random field models for spatial statistics, Discussion paper 95-43, Institute of Statistics and Decision Sciences, Duke University.