VRIJE UNIVERSITEIT

STOCHASTIC GEOMETRY MODELS IN IMAGE ANALYSIS AND SPATIAL STATISTICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus
prof.dr E. Boeker,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der wiskunde en informatica
op donderdag 7 april 1994 te 13.30 uur
in het hoofdgebouw van de universiteit, De Boelelaan 1105

door

MARIA NICOLETTE MARGARETHA VAN LIESHOUT

geboren te Haarlemmermeer

Promotor	:	prof.dr A. J. Baddeley
Copromotor	:	prof.dr J. Oosterhoff
Referenten	:	prof.dr R. D. Gill
		prof.dr P. J. Green

Aan mijn ouders en Janine

ACKNOWLEDGEMENTS

It is a great pleasure to thank all those who contributed in one way or another to this thesis.

I am deeply indebted to my promotor Adrian Baddeley. I have profited in many ways by his fabulous expertise and wisdom and feel grateful for rewarding years of working under his supervision.

I am grateful to Kobus Oosterhoff for his constant support. He suggested to write a Master's thesis on image analysis and was always ready to offer assistance and advice.

Many thanks are due to Jesper Møller and Andrew Lawson for a pleasant collaboration and to Rein van den Boomgaard for some of the data in Chapter 4.

I would like to express my gratitude to the organisers and teachers of the 'aio-cursus stochastiek'; to Adri Steenbeek for solving many software problems and to Malcolm Hudson for his hospitality. Thanks are also due to my many room mates and colleagues at CWI and the Free University for a pleasant working environment and to all who encouraged me by their interest in my research.

Last but not least, I am truly grateful to my family for their moral support, love and encouragement throughout the years.

Table of Contents

Chapter 1
Introduction

The pioneering work by Geman & Geman [42] and Besag [17] stimulated a surge of interest in statistical approaches to image analysis. Until recently, most attention has been given to segmentation or classification tasks, i.e. dividing an image into relatively homogeneous regions of different type [71]. Following [17, 42], a Bayesian approach is usually taken, in which a Markov random field model is used as a prior distribution to impose smoothness on segmentations. Computational problems due to the high dimensionality of images are overcome by iterative algorithms relying only on the local characteristics of the probabilistic model. An efficient deterministic technique to find a locally optimal segmentation is Besag's Iterated Conditional Mode (ICM) algorithm. Realisations of the posterior distribution can be obtained by a Gibbs sampler and a simulated annealing schedule to approximate a globally optimal classification can be designed [42].

The goal of this monograph is to argue that the (continuous) Markov or Gibbs processes studied in stochastic geometry, spatial statistics and statistical physics [12, 97, 100, 101, 103] provide a rich collection of models usable in a broad range of problems in image analysis and spatial statistics.

Object recognition is the task of interpreting a noisy image to identify certain geometrical features. An object recognition algorithm must decide whether there are any objects of a specified kind in the scene,

and if so, determine the number of objects and their locations, shapes, sizes and spatial relationship. The image data are noisy and sometimes blurred. There is a wide field of applications, including industrial robot vision, document reading, interpretation of medical scans [21], automated cytology [82], classification of astronomical features [84, 104] and identification of grain structures in materials science.

It is increasingly acknowledged that discrete Markov random fields (MRFs) are not the appropriate prior models to use in object recognition. This is partly because geometrical shapes with smooth boundaries are unlikely to arise as realisations of a discrete MRF; but more importantly because the *procedures* that result from applying a discrete MRF model do not combine information 'globally' to identify geometrical shape.

This issue is familiar from the computer vision literature as the distinction between 'low-level' and 'high-level' vision. Low-level tasks such as segmentation, classification and tomographic reconstruction call for local (pixel neighbourhood) operations, converting the input image into another raster image. In high-level tasks such as object recognition and scene analysis we have to interpret the image globally, reducing it to a compact description (e.g. a vector graphics representation) of the scene.

Here we study the problem of detecting an unknown number of objects of (usually simple) shape in an unknown spatial arrangement, possibly overlapping each other. This requires a prior stochastic model for the spatial arrangement of the objects. We propose to use the Markov object processes [12, 103] which have a simple mathematical form, and for which there is a natural analogue of the Gibbs sampler (a spatial birth-and-death process [12, 83, 97]). Thus, analogues of the ICM and simulated annealing algorithms can be developed. Also, some existing techniques in computer vision turn out to be equivalent to maximum likelihood methods. The use of Markov point process models was also proposed by Ripley and coauthors [84, 102, 104].

Alternative approaches have been described in recent studies [29, 82, 84, 104] on recognising the shape of an interesting object (hand, galaxy, mitochondrion). The shape is described by a flexible template, typically a polygon, with edge lengths and angles governed by a joint prior distribution, typically a Markov chain.

We claim that the general framework described above is flexible enough to be easily adaptable to various other tasks in high-level vision and spatial statistics that concern the clustering of (image) features; like in object recognition, these problems involve the extraction of an underlying

pattern from a given set of data (that is not necessarily in image form). Applications include large-scale edge detection [88], the identification of cluster centres in a point pattern [14, 24, 20, 92], but also the analysis of geological faults in relation to earthquakes or the reconstruction of ancient roads using archeological finds [111]. Stochastic geometry is helpful in providing both the (conditional) modelling of the data given the underlying pattern of interest and the prior distribution.

Apart from finding the centres of clustering, it is also of interest to partition the data into meaningful groups. We describe how sibling information can be incorparated by means of auxiliary variables [18]. A connection with the classical k-means algorithm (see also [24]) is established and we note that under obvious conditions the nearest parent approximation [66, 67] is in fact a maximum likelihood estimator.

It is interesting to note that the discrete MRF priors used in image segmentation [17, 42] specify *positive* association between neighbouring pixel values, while the Markov object processes proposed here exhibit *negative* association or inhibition between neighbouring objects. In the case of spatial clustering Lawson [66] has used Poisson priors, in which the points have *zero* association, and the effect of the prior is simply to penalise configurations with large numbers of points.

We also study the Markov models themselves. Until recently attention focused on pairwise interaction models. These provide a flexible class for inhibition but they do not seem to be able to model clustered patterns and Møller [85] has argued that nearest-neighbour patterns are better suited to this task.

In support of this claim, we show that many cluster processes with bounded non-empty clusters fall within the class of nearest-neighbour Markov point processes. In particular, any Poisson cluster process with uniformly bounded clusters is Markov with respect to the connected component relation [12, p. 106] and if a Markov or nearest-neighbour Markov point process is used as the parent process in a cluster model, and the clusters are uniformly bounded and almost surely nonempty, then the cluster process is again nearest-neighbour Markov. These results suggest that nearest-neighbour Markov processes may be suitable multiple-generation cluster models, see Kingman [63] and help to explain why statistical inference for Poisson cluster processes based on interpoint distances (cf. Baddeley and Van Lieshout [10]) bears so close a resemblance to that for Markov point processes.

On the other hand, the Ripley-Kelly models also allow clustered patterns if interactions between more than two points or objects are per-

mitted. We discuss a model that can exhibit both clustering [119] and inhibition according to the value of a single parameter. The model has interactions of arbitrary high order and is closely related to the empty space function.

Much of the research presented in this thesis was performed in collaboration with others. Chapter 2 is based on

> A.J. Baddeley and M.N.M. van Lieshout. ICM for object recognition. In *Computational Statistics* Y. Dodge and J. Whittaker (Eds). Volume 2, pp. 271–286. Heidelberg-New York: Physica/Springer 1992.
>
> A.J. Baddeley and M.N.M. van Lieshout. Object recognition using Markov spatial processes. In *Proceedings 11th IAPR International Conference on Pattern Recognition* pp. B 136–139. Los Alamitos: IEEE Computer Society Press 1992.

The model in Chapter 3 is reported in

> A.J. Baddeley and M.N.M. van Lieshout. Area–interaction point processes. CWI Report BS-R9318, november 1993. Submitted to *Annals of the Institute of Statistical Mathematics*.

The deterministic algorithms in Chapter 4 can be found in

> A.J. Baddeley and M.N.M. van Lieshout. ICM for object recognition. In *Computational Statistics* Y. Dodge and J. Whittaker (Eds). Volume 2, pp. 271–286. Heidelberg-New York: Physica/Springer 1992.

while the stochastic annealing approach is in

> M.N.M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Advances in Applied Probability* **26**, 1994.

See also [72, 5, 73, 74, 3, 8].

The application in spatial statistics is based on

> A.J. Baddeley and M.N.M. van Lieshout. Stochastic geometry models in high-level vision. *Journal of Applied Statistics* **20**, pp. 233–258, 1993.

and the author's part in ongoing research

> A.B. Lawson, M.N.M. van Lieshout and A.J. Baddeley. Markov chain Monte Carlo methods for spatial cluster processes.

Finally, the last chapter is a revision of

> A.J. Baddeley, M.N.M. van Lieshout and J. Møller. Markov properties of cluster processes. Research Report 278, Department of Theoretical Statistics, University of Aarhus, 1994. Submitted to *Probability Theory and Related Fields*.

Chapter 2
Maximum likelihood object recognition

Object recognition can be formulated as a parameter estimation problem by direct analogy with the formulation of segmentation and classification [17, 40, 42] and in keeping with the general setup of Grenander [47, 48, 49]. In this Chapter we define the object recognition problem, develop a simple maximum likelihood treatment, and show that this is very similar to some existing techniques in computer vision.

2.1. Object recognition

Object recognition techniques are surveyed in [36, 95, 105]. They can be divided into methods such as region growing which detect an object of unspecified shape and size by characterising it as a region of homogeneous pixel intensity (etc.), and template matching methods which compare the data image with a translated and rotated copy of a reference shape and locate the optimal match [105]. Here we follow the template matching approach.

Suppose the experimental data consist of an image $\mathbf{y} = (y_t \; ; \; t \in T)$ where the *'image space'* T is an arbitrary finite set. Apart from the usual two-dimensional rectangular grids, T could be a pair of grids (carrying left and right stereo images), a temporal sequence, etc. The observed value y_t at pixel $t \in T$ ranges over a set V that is arbitrary. Examples include $\{0,1\}$ for *binary* images, $\{0,1,\ldots,255\}$ for 8-bit digitised *grey level* images or $\mathbb{R}$.

The class U of possible objects is an arbitrary set (*'object space'*). Typical examples would be the class of all polygons in $\mathbb{R}^2$ or all convex polyhedra in $\mathbb{R}^3$. However U need not be a class of subsets of $\mathbb{R}^d$ since the specification of an object may also include properties like colour or surface texture. Section 3.1.1 discusses this further. Here we assume only that each object $u \in U$ determines a subset $R(u) \subseteq T$ of image space 'occupied' by the object.

An object configuration is simply a finite set of objects

$$\mathbf{x} = \{x_1, \cdots, x_n\}$$

where $x_i \in U$, $i = 1, \cdots, n$, $n \geq 0$. The objects may be in any spatial relation to each other; the number of objects is variable and may be zero.

The goal then is to extract the unobserved underlying pattern $\mathbf{x}$ from a given data image $\mathbf{y}$.

A standard numerical criterion for the degree of match between $\mathbf{x}$ and $\mathbf{y}$ is the *Hough transform*, originally proposed by Hough [54] to detect straight lines in binary images [13, 25, 35, 62, 108]. A recent survey is [57]. The Hough transform is a real valued function of the object parameter vector u defined by

$$H_{\mathbf{y}}(u) = \sum_{t \in R(u)} y_t \, , u \in U \tag{2.1.1}$$

where $\mathbf{y}$ is the data image. This is often interpreted as a vote-counting operation: each pixel t votes with strength y_t for all the objects that contain that pixel. Objects are located typically by finding local maxima of the matching criterion, or by accepting all template positions where the match exceeds a threshold value.

An alternative approach using *mathematical morphology* is to perform an erosion with respect to the template. For example if $\mathbf{y}$ is a binary image and Y is the set of white pixels define the generalised erosion of Y by

$$\begin{aligned} E_R(Y) &= \{u : R(u) \subseteq Y\} \\ &= \{u : y_t = 1 \text{ for all } t \in R(u)\} \end{aligned} \tag{2.1.2}$$

i.e. accept only those objects u for which every pixel in the template $R(u)$ is white. This is a generalised erosion operator [106, 107]. On a discrete image lattice, the erosion is the set of u points where the Hough transform attains its maximum possible value.

2.2. NOISE MODELS

In the likelihood approach to image analysis the observed image $\mathbf{y}$ depends on the true object configuration $\mathbf{x}$ through a known conditional probability density $f(\mathbf{y} \mid \mathbf{x})$. This density describes the 'forward problem' of image formation and includes both the deterministic influence of $\mathbf{x}$ and the stochastic noise inherent in observing $\mathbf{y}$.

Following custom, we assume that the data pixel values y_t are conditionally independent given $\mathbf{x}$. This embraces additive and multiplicative random noise as well as Poisson distributed counts and more general exponential family models. Without loss of generality the conditional distributions of individual pixel values belong to a family of distributions with densities $\{g(\cdot \mid \theta) \ : \ \theta \in \Theta\}$ indexed by a parameter space Θ.

Definition 1 *An* independent noise model *is a stochastic model for* $\mathbf{y}$ *given* $\mathbf{x}$*, which assumes pixel values* y_t *are conditionally independent given* $\mathbf{x}$*, with joint probability density*

$$f(\mathbf{y} \mid \mathbf{x}) = \prod_{t \in T} g(y_t \mid \theta^{(\mathbf{x})}(t)) \tag{2.2.3}$$

where $\{g(\cdot \mid \theta) : \theta \in \Theta\}$ *is a family of probability densities on* V *and* $\theta^{(\mathbf{x})}(t)$ *is the parameter value of the conditional distribution of* y_t *given* $\mathbf{x}$*. Then* $\theta^{(\mathbf{x})}(\cdot)$ *is a* Θ*-valued image, deterministically derived from* $\mathbf{x}$*, which we call the* signal.

Note that no assertions are made about the way objects interact and that the model does not imply that the pixel values are (unconditionally) independent.

A simple example is the signal

$$\theta^{(\mathbf{x})}(t) = \begin{cases} \theta_1 & \text{if } t \in S(\mathbf{x}) \\ \theta_0 & \text{otherwise} \end{cases} \tag{2.2.4}$$

where θ_1, θ_0 are known parameters (*foreground* and *background* signal values) and $S(\mathbf{x})$ is the *silhouette*

$$S(\mathbf{x}) = \bigcup_{i=1}^{n} R(x_i)$$

formed by taking the union of all objects in the configuration. In other words, under this simple model, each of the objects in the configuration

$\mathbf{x}$ is 'painted' onto the scene, and independent random pixel noise is superimposed on the result.

We call this a *blur-free* independent noise model. It may seem oversimplified; yet we shall show that several standard techniques in computer vision are equivalent to assuming this model.

The following are examples of independent noise models. Strictly speaking, they represent a whole family of models; members are obtained by varying the signal function.

Model 1: additive Gaussian white noise

y_t is normally distributed with mean $\mu = \theta^{(\mathbf{x})}(t) \in \mathbb{R}$ and *fixed* variance $\sigma^2 > 0$:

$$g(y_t \mid \mu) = (2\pi\sigma^2)^{-1/2}\, e^{-(y_t-\mu)^2/(2\sigma^2)}.$$

Model 2: additive Laplacian noise

y_t is double exponentially distributed with mean $\mu = \theta^{(\mathbf{x})}(t) \in \mathbb{R}$ and *fixed* dispersion parameter $\lambda > 0$:

$$g(y_t \mid \mu) = \frac{\lambda}{2}\, e^{-\lambda|y_t-\mu|}.$$

Model 3: binary image

y_t has a Bernoulli distribution with success probability $\theta^{(\mathbf{x})}(t)$:

$$g(y_t \mid \theta) = \theta^{y_t}(1-\theta)^{(1-y_t)}.$$

Special cases are **salt-and-pepper noise** where

$$\theta^{(\mathbf{x})}(t) = \begin{cases} 1-p & \text{if } t \in S(\mathbf{x}) \\ p & \text{else} \end{cases}$$

and **pepper noise** where:

$$\theta^{(\mathbf{x})}(t) = \begin{cases} 1 & \text{if } t \in S(\mathbf{x}) \\ p & \text{else} \end{cases}.$$

Here $0 < p < 1$ is fixed.

Model 4: Poisson counts

y_t is integer valued and Poisson distributed with mean $\mu = \theta^{(\mathbf{x})}(t) \in \mathbb{R}^+$:

$$g(y_t \mid \mu) = e^{-\mu} \frac{\mu^{y_t}}{y_t!}.$$

Model 1 is widely used in image analysis. In Models 3a and 3b the silhouette is converted to a binary image and noise introduced by randomly flipping pixel values with probability p. In Model 3a all pixels are subject to change, while in Model 3b only background pixels are flipped. Model 4 is the usual model in emission tomography where a patient is injected with a radioactive isotope and emitted particles are recorded in a system of detectors placed around the patient.

2.3. MAXIMUM LIKELIHOOD ESTIMATION

Given observation of $\mathbf{y}$, the likelihood of a configuration $\mathbf{x}$ is defined to be $\ell(\mathbf{x};\mathbf{y}) = f(\mathbf{y} \mid \mathbf{x})$, and we seek 'the' maximum likelihood estimate of $\mathbf{x}$

$$\widehat{\mathbf{x}} = \text{argmax}_{\mathbf{x}}\, f(\mathbf{y} \mid \mathbf{x}) \tag{2.3.5}$$

which may be nonexistent or nonunique. Specifically, note that for any blur-free model the likelihood depends on $\mathbf{x}$ only through its silhouette $S(\mathbf{x})$, so configurations with the same silhouette cannot be distinguished in likelihood.

Since the log-likelihood is a sum of individual pixel error terms

$$L(\mathbf{x}\,;\,\mathbf{y}) = \log f(\mathbf{y} \mid \mathbf{x}) = \sum_{t\in T} \log\ g(y_t \mid \theta^{(\mathbf{x})}(t)),$$

maximum likelihood estimation is equivalent to regression of $\mathbf{y}$ on the class of signals $\theta^{(\mathbf{x})}(\cdot)$ for all possible $\mathbf{x}$, with pixelwise loss function $-\log g(y_t \mid \cdot)$.

Lemma 1 *An MLE for Model 1 is a solution of the least squares regression of* $\mathbf{y}$ *on the class of functions*

$$\left\{\theta^{(\mathbf{x})}(t) : \mathbf{x} = \{x_1, \ldots, x_n\}, x_i \in U, n \geq 0\right\}.$$

In Model 2 any MLE solves a least absolute deviation regression on the same class.

Proof: Writing $|T|$ for the area or number of pixels in T, the log-likelihood is

$$L(\mathbf{x}\,;\,\mathbf{y}) = -\frac{1}{2}\,|T|\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t\in T}\left(y_t - \theta^{(\mathbf{x})}(t)\right)^2$$

for Model 1 and

$$L(\mathbf{x}\,;\,\mathbf{y}) = |T|\log\frac{\lambda}{2} - \lambda\sum_{t\in T}|y_t - \theta^{(\mathbf{x})}(t)|$$

for Model 2 .

□

Typically, the equation (2.3.5) cannot be solved directly. This is clearly true for the regression above, because of the combinatorial and geometric complexity of the functions $\theta^{(\mathbf{x})}(t)$. We will return to this problem below.

Lemma 2 *For Model 3a with $0 < p < 1/2$, the ML equations are*

$$\hat{\mathbf{x}} = \mathit{argmin}_{\mathbf{x}} \, |S(\mathbf{x}) \triangle Y|$$

where $\triangle$ denotes the symmetric set difference ('exclusive-or') and $Y = \{t \in T : y_t = 1\}$ is the set of white pixels.

Proof:

$$\begin{aligned} L(\mathbf{x}\,;\,\mathbf{y}) &= |S(\mathbf{x}) \setminus Y| \log p \;+\; |S(\mathbf{x}) \cap Y| \log(1-p) \\ &+ |Y \setminus S(\mathbf{x})| \log p \;+\; (|T| - |Y \cup S(\mathbf{x})|) \log(1-p) \\ &= |T| \log(1-p) \;+\; |S(\mathbf{x}) \triangle Y| \log \frac{p}{1-p} \end{aligned}$$

and for $p < 1/2$ the coefficient of $|S(\mathbf{x}) \triangle Y|$ is negative. □

Squared error and L^1 error have been proposed as optimality criteria for object recognition in their own right but we see here that they are special cases of the maximum likelihood approach. These results confirm recent arguments in the literature [76, 78] in favour of using L^1 filtering, except when the noise is Gaussian.

2.3.1 Connection with mathematical morphology

The following result shows that standard morphological operators solve the ML equations (2.3.5) for a simple noise model.

Lemma 3 *A maximum likelihood estimator for Model 3b is the generalised erosion (2.1.2)*

$$\begin{aligned} \hat{\mathbf{x}}_{max} &= E_R(Y) \\ &= \{u \in U : R(u) \subseteq Y\}, \end{aligned}$$

where $Y = \{t \in T : y_t = 1\}$. This is the largest solution of the ML equations; the other solutions are the subsets $\hat{\mathbf{x}} \subseteq \hat{\mathbf{x}}_{\max}$ with the same silhouette,

$$S(\hat{\mathbf{x}}) = S(\hat{\mathbf{x}}_{max}).$$

Proof: The density is nonzero only if $S(\mathbf{x}) \subseteq Y$ and in this case equals

$$f(\mathbf{y}|\mathbf{x}) = \prod_{t \in T \setminus S(\mathbf{x})} p^{y_t}(1-p)^{(1-y_t)}.$$

The log likelihood is then

$$\begin{aligned} L(\mathbf{x}\,;\,\mathbf{y}) &= |Y \setminus S(\mathbf{x})| \log p + (|T| - |Y|) \log(1-p) \\ &= |T| \log(1-p) + |Y| \log \frac{p}{1-p} - |S(\mathbf{x})| \log p \end{aligned}$$

and the result follows. □

The 'classical' erosion operator $\ominus$ [106] is the special case where $U = T \subset \mathbb{R}^2$ and $R(u) = u + R$ where R is a fixed subset of T. Then

$$\hat{\mathbf{x}}_{\max} = Y \ominus \check{R} := \{u : (u+R) \subseteq Y\}.$$

Thus the erosion operator is the MLE for a simple noise model; its corresponding silhouette is the *opening* of Y by R [106].

The dual operator is the dilation $\oplus$

$$\begin{aligned} Y \oplus \check{R} &:= \{u : (u+R) \cap Y \neq \emptyset\} \\ &= \{u : (u+R) \not\subseteq Y^c\} \\ &= (Y^c \ominus \check{R})^c \end{aligned}$$

(see [106]). Thus, by exchanging foreground and background, i.e. taking $\theta^{(\mathbf{x})}(t) = 0$ in Lemma 3, one obtains a similar result for the dilation. A maximum likelihood estimator is

$$\hat{\mathbf{x}}_{\max} = D_R(Y)^c$$

where

$$D_R(Y) = \{u \in U : R(u) \cap Y \neq \emptyset\}$$

is the *generalised dilation* of Y. Other solutions of equations (2.3.5) are subsets with the same silhouette.

2.3.2 Relation to preprocessing

It is also interesting to note that popular 'pre-processing' techniques, such as change of scale, thresholding or gamma correction, amount to

simply modifying the noise model. If the pixel values y_t are subjected to an invertible, differentiable transformation ('anamorphosis' in morphology parlance) $y'_t = \phi(y_t)$ then the model (2.2.3) is transformed into another model of the same type with g replaced by another density g'

$$f(\mathbf{y}' \mid \mathbf{x}) = \prod_{t \in T} g'(y'_t \mid \theta^{(\mathbf{x})}(t))$$

where

$$g'(y'_t \mid \theta^{(\mathbf{x})}(t)) = g(\phi^{-1}(y'_t) \mid \theta^{(\mathbf{x})}(t))\ J(y'_t);$$

$J(y'_t)$ being 1 for grey level data, and $|\frac{d}{dz}\phi^{-1}(y'_t)|$ in the absolutely continuous case. In particular, for exponential families the transformed data is again an exponential family.

This is also true for transformations such as 'clipping' pixel values to an interval $[a, b]$

$$\mathrm{clip}(s, a, b) = \begin{cases} a & \text{if } s < a \\ s & \text{if } a \leq s \leq b. \\ b & \text{if } s > b \end{cases}$$

Treating clipped pixel data as if they arise from additive Gaussian noise is equivalent to assuming additive two-sided exponential noise with signal (= mean) values $\theta_0 = a$, $\theta_1 = b$.

Lemma 4 *For Model 2 with $\theta^{(\mathbf{x})}(t)$ defined by (2.2.4) with $\theta_0 < \theta_1$*

$$L(\mathbf{x} \cup \{u\}\ ;\ \mathbf{y}) - L(\mathbf{x}\ ;\ \mathbf{y}) =$$

$$2\lambda \left\{ \sum_{t \in R(u) \setminus S(\mathbf{x})} y'_t - \frac{\theta_1 + \theta_0}{2} |R(u) \setminus S(\mathbf{x})| \right\}$$

where $y'_t = \mathrm{clip}(y_t, \theta_0, \theta_1)$.

Proof: Observe that (with $a < b$)

$$\begin{aligned} |s-b| - |s-a| &= 1\{s < a\}(b-a) + 1\{s > b\}(a-b) \\ &+ 1\{a \leq s \leq b\}(a+b-2s) \\ &= a + b - 2s + 1\{s < a\}(-2a + 2s) \\ &+ 1\{s > b\}(-2b + 2s) \\ &= a + b - 2\mathrm{clip}(s, a, b). \end{aligned}$$

Hence

$$L(\mathbf{x} \cup \{u\}\,;\,\mathbf{y}) - L(\mathbf{x}\,;\,\mathbf{y}) =$$

$$= -\lambda \sum_{t \in R(u) \setminus S(\mathbf{x})} \{|y_t - \theta_1| - |y_t - \theta_0|\}$$

$$= -\lambda \sum_{t \in R(u) \setminus S(\mathbf{x})} \{\theta_0 + \theta_1 - 2\mathrm{clip}(y_t, \theta_0, \theta_1)\}$$

$$= 2\lambda \left\{ \sum_{t \in R(u) \setminus S(\mathbf{x})} y'_t - \frac{\theta_1 + \theta_0}{2} |R(u) \setminus S(\mathbf{x})| \right\}$$

and the result follows. □

These remarks do not hold for more complex pre-processing operations such as neighbourhood filtering, which interfere with the dependence structure of (2.2.3).

2.4. Iterative methods for MLE

It is usually impossible to solve the ML estimating equations (2.3.5) directly, and one has to resort to iterative approximation methods.

2.4.1 Add-and-delete algoritthms

The simplest form of iterative adjustment is to add or delete objects one-at-a-time. If the current configuration is $\mathbf{x}$ then we consider adding a new object $u \notin \mathbf{x}$, yielding configuration $\mathbf{x} \cup \{u\}$, if the log likelihood ratio

$$L(\mathbf{x} \cup \{u\} \; ; \; \mathbf{y}) - L(\mathbf{x} \; ; \; \mathbf{y}) = \log \frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})} \tag{2.4.6}$$

is sufficiently large or deleting an existing object $x_i \in \mathbf{x}$ yielding $\mathbf{x} \setminus \{x_i\}$ if

$$L(\mathbf{x} \setminus \{u\} \; ; \; \mathbf{y}) - L(\mathbf{x} \; ; \; \mathbf{y}) = \log \frac{f(\mathbf{y} \mid \mathbf{x} \setminus \{u\})}{f(\mathbf{y} \mid \mathbf{x})} \tag{2.4.7}$$

is sufficiently large. Two variations of this scheme are to visit the possible objects u sequentially (assuming U is discretised) applying the above rules at each step, or to scan the whole of U to find the object u whose addition or deletion would most increase the likelihood.

Algorithm 1 (Coordinatewise optimisation) *Initialize* $\mathbf{x}^{(0)} = \emptyset$ *or some other chosen initial state. When the current reconstruction is* $\mathbf{x}^{(k-1)}$, *visit every* $u \in U$ *sequentially in a predetermined order. If* $u \notin \mathbf{x}^{(k-1)}$ *and* $L(\mathbf{x}^{(k-1)} \cup \{u\} \; ; \; \mathbf{y}) - L(\mathbf{x}^{(k-1)} \; ; \; \mathbf{y}) \geq w$, *where* $w \geq 0$ *is a fixed threshold, then add* u *to the configuration, yielding* $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} \cup \{u\}$. *If* $u = x_i \in \mathbf{x}^{(k-1)}$ *and* $L(\mathbf{x}^{(k-1)} \setminus \{x_i\} \; ; \; \mathbf{y}) - L(\mathbf{x}^{(k-1)} \; ; \; \mathbf{y}) \geq w$, *then delete* x_i *yielding* $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} \setminus \{x_i\}$. *Update recursively until one complete scan of the image yields no changes.*

Algorithm 2 (Steepest ascent) *Initialize* $\mathbf{x}^{(0)} = \emptyset$ *or some other chosen initial state. Given* $\mathbf{x}^{(k-1)}$, *determine*

$$a = \max_{x_i \in \mathbf{x}^{(k-1)}} \left\{ L(\mathbf{x}^{(k-1)} \setminus \{x_i\} \; ; \; \mathbf{y}) - L(\mathbf{x}^{(k-1)} \; ; \; \mathbf{y}) \right\}$$

and

$$b = \sup_{u \in U} \left\{ L(\mathbf{x}^{(k-1)} \cup \{u\} \; ; \; \mathbf{y}) - L(\mathbf{x}^{(k-1)} \; ; \; \mathbf{y}) \right\}.$$

If $\max\{a, b\} < w$, *then stop. Otherwise, if* $b \geq a$, *add the corresponding object, while if* $a > b$, *delete the corresponding object.*

These algorithms bear a very strong resemblance to Besag's ICM algorithm [17] except for the lack of a prior distribution. The analogy will be explored in Chapter 4.

Clearly these algorithms increase the likelihood at each step,

$$f(\mathbf{y} \mid \mathbf{x}^{(k+1)}) \geq f(\mathbf{y} \mid \mathbf{x}^{(k)}).$$

As there are only a finite number of possible configurations, convergence of $f(\mathbf{y} \mid \mathbf{x}^{(k)})$ is guaranteed and (if $w = 0$) we reach a local maximum of the likelihood function. At worst there is cycling between images of equal likelihood. However the algorithms do not necessarily yield the global maximum likelihood solution, and the local maximum obtained will depend on the initial configuration $\mathbf{x}^{(0)}$ and for Algorithm 1 on the scanning order as well. We should therefore choose a sensible initial state, such as the empty list $\emptyset$, or the set of local maxima of

$$\frac{f(\mathbf{y} \mid \{u\})}{f(\mathbf{y} \mid \emptyset)}$$

wherever this ratio is larger than 1.

Complications arise if the model $f(\mathbf{y} \mid \mathbf{x})$ contains unknown parameters. We will return to this problem in Section 4.2.3.

2.4.2 General add-delete-shift algorithms

Another form of iterative adjustment is to change an existing object slightly by translation, rotation or expansion. The aim is to obtain methods that are more robust against imprecise information contained in the initial estimate. Another advantage is that convergence may be faster, as throwing away an incorrect object and replacing it by the right one can then be carried out in one single step.

Write

$$M(\mathbf{x}, x_i, u) = (\mathbf{x} \cup \{u\}) \setminus \{x_i\}$$

for the configuration obtained from $\mathbf{x}$ by moving the element $x_i \in \mathbf{x}$ to a new position u. Let $Q(\mathbf{x}, x_i)$ be the set of all object points u for which this operation is permitted. Typically u will be required to be close to x_i but not equal to any x_j, say $Q(\mathbf{x}, x_i) = N(x_i) \setminus \mathbf{x}$ where $N(x_i)$ is some neighbourhood of x_i. Then the criterion for a move from $x_i \in \mathbf{x}$ is

$$\max_{u \in Q(\mathbf{x}, x_i)} \left\{ L(M(\mathbf{x}^{(k-1)}, x_i, u) \; ; \; \mathbf{y}) - L(\mathbf{x}^{(k-1)} \; ; \; \mathbf{y}) \right\} \tag{2.4.8}$$

and the analogues of Algorithms 1 and 2 are as follows.

Algorithm 3 (Coordinatewise optimisation with shifts) *Assuming that the parameter space is finite, visit every $u \in U$ sequentially. Consider every possible transition involving u and select the maximum from (2.4.6), (2.4.7) and (2.4.8). If this maximum log likelihood ratio is larger than a given threshold w, update the reconstruction accordingly.*

Algorithm 4 (Steepest ascent with shifts) *Consider all possible transitions from the current state $\mathbf{x}$ and take that transition that has the maximum log likelihood ratio exceeding threshold w.*

The convergence properties are similar to Algorithms 1-2. Here, however, a 'local maximum' of the likelihood is a state $\mathbf{x}$ such that no neighbouring configuration $\mathbf{x} \cup \{u\}$ or $\mathbf{x} \setminus \{x_i\}$ or $M(\mathbf{x}, x_i, u)$ has larger likelihood. This is a more stringent definition than for the previous algorithms, and one expects the results to be better.

2.5. Relation to Hough transform

The Hough transform (2.1.1) is very similar to the log likelihood ratio (2.4.6) for a blur-free model:

$$L(\mathbf{x} \cup \{u\} \,;\, \mathbf{y}) - L(\mathbf{x} \,;\, \mathbf{y}) = \sum_{R(u) \setminus S(\mathbf{x})} z_t \tag{2.5.9}$$

where $z_t = \log g(y_t|\theta_1) - \log g(y_t|\theta_0)$ is the difference in goodness-of-fit at pixel t. Note that pixels can also cast fractional or negative votes, a modification that has been suggested ad hoc by several authors [13, 28, 116, 118] and [23, 27] respectively. In fact, (2.5.9) is identical to the (generalised) Hough transform of image $\mathbf{z}$ when the new object u does not overlap any existing object $x_i \in \mathbf{x}$. For example, *the Hough transform is the log likelihood ratio for comparing $\{u\}$, the scene consisting of a single object, against the empty scene $\emptyset$* [55]. When objects do overlap, the likelihood ratio (2.5.9) is a generalisation of the Hough transform, with domain of summation 'masked' by the silhouette of the current configuration. Equivalently (2.5.9) is the Hough transform of the masked image $w_t^{(\mathbf{x})} = z_t 1\{t \notin S(\mathbf{x})\}$.

The similarity between (2.1.1) and (2.5.9) is even stronger since z_t is linear in y_t for many exponential noise models. For additive Gaussian noise

$$z_t = \frac{\theta_1 - \theta_0}{\sigma^2}\left(y_t - \frac{\theta_0 + \theta_1}{2}\right)$$

and for Poisson noise

$$z_t = y_t \log \frac{\theta_1}{\theta_0} - (\theta_1 - \theta_0).$$

The likelihood ratio is then of the form

$$L(\mathbf{x} \cup \{u\} \,;\, \mathbf{y}) - L(\mathbf{x} \,;\, \mathbf{y}) = a \sum_{R(u) \setminus S(\mathbf{x})} y_t - b|R(u) \setminus S(\mathbf{x})|. \tag{2.5.10}$$

In particular when u does not overlap $\mathbf{x}$ this is a linear adjustment of the Hough transform of $\mathbf{y}$. More generally, let g be an exponential family of the form

$$g(y_t \mid \theta) = \exp\left[A(\theta) + B(y_t) + C(\theta)D(y_t)\right]$$

for some real-valued functions A, B, C and D. Then (2.5.10) holds if y_t is replaced by $D(y_t)$. The generalisation to vector-valued functions is straightforward.

Apart from illuminating the meaning of the Hough transform, these results show how to correctly interpret values of the Hough transform when the objects do not have equal area, e.g. in the presence of edge effects. For Gaussian or Poisson additive noise (say), the likelihood ratio is positive when the average value of y_t over $R(u) \setminus S(\mathbf{x})$ exceeds a critical value. The latter is simply the Neyman-Pearson critical value for classifying a single observation y_t as foreground or background ($\theta \in \{\theta_0, \theta_1\}$).

For the general 'blurred' model (2.2.3) the likelihood ratio equals

$$\sum_{t \in Z(\mathbf{x},u)} h(y_t, \theta^{(\mathbf{x})}(t), \theta^{(\mathbf{x} \cup \{u\})}(t)) \tag{2.5.11}$$

and thus is again similar to the Hough transform, where

$$h(y_t, \theta, \theta') = \log \frac{g(y_t|\theta')}{g(y_t|\theta)}$$

is as before the difference in goodness-of-fit at pixel t and

$$Z(\mathbf{x}, u) = \{t : \theta^{(\mathbf{x})}(t) \neq \theta^{(\mathbf{x} \cup \{u\})}(t)\} \tag{2.5.12}$$

is the set of pixels where the signal is affected by the addition of object u.

The log likelihood ratio for a transition from $\mathbf{x}$ to $M(\mathbf{x}, x_i, u)$ (Algorithms 3 and 4) can be represented as

$$[L(\mathbf{z} \cup \{u\}\ ;\ \mathbf{y}) - L(\mathbf{z}\ ;\ \mathbf{y})] - [L(\mathbf{z} \cup \{x_i\}\ ;\ \mathbf{y}) - L(\mathbf{z}\ ;\ \mathbf{y})]$$

where $\mathbf{z} = \mathbf{x} \setminus \{x_i\}$. This is a difference of two values of the generalised Hough transform (2.5.11) for the configuration with x_i deleted. In particular it depends only on data pixels within a region $Z(\mathbf{z}, u) \cup Z(\mathbf{z}, x_i)$.

2.6. EXAMPLE

Figure 2.1 shows a scanned 128×128 image ('pellets') taken from the Brodatz texture album [22]. This is a relatively easy dataset for object recognition but helps to illustrate the approach.

We treat the pellets as discs of fixed radius 4 pixels but with blurred boundaries. The grey-level histogram has two distinct peaks at value 8 and 172, suggesting that we can regard the background and foreground signal as roughly constant at these values. Assuming additive Gaussian noise, the noise variance was estimated by thresholding the image and taking the sample variance, giving an estimate of 83.1. Blurring was modelled by assuming that the original blur-free signal was subjected to a 3×3 averaging (linear) filter with relative weights 4 for the central pixel, 2 for horizontal and vertical neighbours and 1 for diagonal neighbours.

Figure 2.2 shows an approximate MLE computed by steepest ascent (Algorithm 2) from an empty initial configuration at threshold $w = 0$. Pellets are correctly identified but there is 'multiple response', i.e. the MLE sometimes contains clusters of objects around the position of each 'true' object.

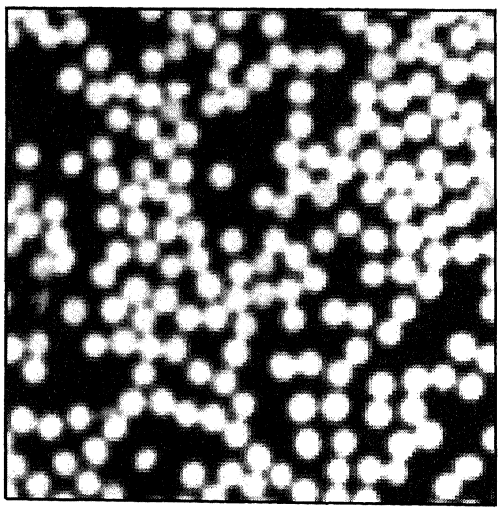


Figure 2.1: Pellets image taken from [22], digitised on a 128×128 square grid.

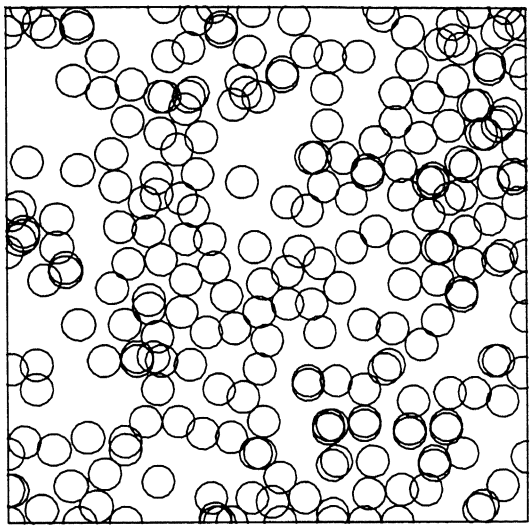


Figure 2.2: Approximate maximum likelihood reconstruction by steepest ascent of the pellets texture at threshold $w = 0$ from an empty initial state.

Chapter 3
Markov spatial processes

The first Section of this Chapter is an overview with some adaptations of the theory of random processes of geometrical objects [111] with emphasis on Markov object processes [103, 12]. These provide a flexible class of models for inhibition and we propose to use them as priors in the object recognition problem discussed in Chapter 2. In Sections 3.2 and 3.3 we introduce a Markov point process proposed by Baddeley and Van Lieshout [9] that exhibits both clustered and ordered pattern according to the value of a parameter. It has infinite order interactions and is related to the empty space function F as the Strauss model is to Ripley's K-function. Statistical inference for this model is studied in Section 3.4.

3.1. Survey of Markov spatial models

3.1.1 Objects

The 'objects' featuring in stochastic geometry range from simple geometrical figures (points, lines, discs) through plane polygons and convex compact sets to completely general closed sets. A given class of objects U is treated as a space in its own right, so that objects are regarded as points in U.

At one extreme, simple geometrical figures can be specified by the values of a few parameters (giving location, orientation etc.) so that U is isomorphic to a subset of $\mathbb{R}^k$. For example a disc in $\mathbb{R}^2$ can be specified by its centre (x, y) and radius r so that $U = \mathbb{R}^2 \times \mathbb{R}^+$. At

the other extreme, the space $\mathcal{F}$ of all closed subsets of $\mathbb{R}^d$ can be made into a locally compact, second countable Hausdorff space (l.c.s. space) so that a random closed set can be defined as a random element of $\mathcal{F}$ [81].

It is often useful to represent an object as a 'marked point', i.e. a pair (s, m) consisting of a point $s \in \mathbb{R}^d$ and a 'mark' $m \in \mathcal{M}$, where $\mathcal{M}$ is an arbitrary l.c.s. space. The point s fixes the location of the object and the mark m contains all other information such as size and shape. A disc in $\mathbb{R}^2$ can be regarded as a point (x, y) marked by a radius r. Objects with additional properties such as colour and surface texture can be represented as marked points by choosing an appropriate mark space $\mathcal{M}$. For example a grey-scale surface texture can be formalised as an upper-semicontinuous function $\mathbb{R}^d \to \mathbb{R}^+$, and the space of all such functions is l.c.s.

3.1.2 Markov object processes

Let U be the class of objects. As before, a configuration is a finite set $\mathbf{x} = \{x_1, \ldots, x_n\}$ of objects $x_i \in U$. Writing Ω for the set of all configurations, a random process of objects is a random element of Ω, or equivalently, a point process on U consisting of a finite number of 'points' with probability 1.

The basic reference model is the *Poisson object process* in U with intensity μ, where μ is a finite non-atomic measure on U. Under this model the total number of objects has a Poisson distribution with mean $\mu(U)$; given that exactly n objects are present, they are independent and identically distributed in U with probability distribution proportional to μ, i.e. $\mathbb{P}(x_i \in B) = Q(B) = \mu(B)/\mu(U)$ for $B \subseteq U$.

Further details can be consulted in [111].

Our interest is in constructing non-Poisson spatial processes exhibiting dependence between neighbouring objects. To do this we shall specify the probability density of the new process with respect to the Poisson process (thereby restricting attention to processes that are absolutely continuous with respect to the Poisson). The density is an integrable function $p : \Omega \to [0, \infty)$. For the new process, the distribution of the total number of objects is

$$\mathbb{P}(N = n) = \frac{e^{-\mu(U)}}{n!} \int_U \cdots \int_U p(\{x_1, \ldots, x_n\})\, d\mu(x_1) \ldots d\mu(x_n).$$

Writing $q_n = \mathbb{P}(N = n)$, given $N = n$, the n random objects have joint probability density

$$p_n(x_1, \ldots, x_n) = e^{-\mu(U)} \mu(U)^n p(\{x_1, \ldots, x_n\})/(n!\, q_n)$$

with respect to the distribution of n i.i.d. objects in U with distribution Q.

Provisionally define two objects u, v to be 'neighbours' if they overlap,

$$u \sim v \quad \Leftrightarrow \quad Z(u) \cap Z(v) \neq \emptyset. \tag{3.1.1}$$

This can be replaced by any symmetric ($u \sim v$ iff $v \sim u$), reflexive relation between elements of U. The *neighbourhood* $N(A)$ of a set $A \subseteq U$ is the set of all points in U neighbouring a point in A:

$$N(A) = \{u \in U : u \sim a \text{ for some } a \in A\}.$$

The simplest kind of spatial interaction is that which forbids objects to overlap. Consider a Poisson process of objects in T conditioned on the event that no pair of objects is overlapping. Its density with respect to the original Poisson process is simply

$$p(\mathbf{x}) = \begin{cases} 0 & \text{if } x_i \sim x_j \text{ for some } i \neq j \\ \alpha & \text{otherwise} \end{cases} \tag{3.1.2}$$

where $\alpha > 0$ is the normalising constant (= reciprocal of the probability that Poisson process has no overlapping objects). Call this a *hard object process* by analogy with the better-known hard core point process.

Next consider a *pairwise interaction*

$$p(\mathbf{x}) = \alpha\beta^{n(\mathbf{x})} \prod_{x_i \sim x_j} g(x_i, x_j) \tag{3.1.3}$$

where $\alpha, \beta > 0$ are constants, $n(\mathbf{x})$ is the number of points in $\mathbf{x}$, and $g : U \times U \to [0, \infty)$. The product is over all pairs of neighbouring objects $x_i \sim x_j$ with $i < j$.

If $g \equiv 1$ then (3.1.3) is simply a Poisson process with intensity $\beta\mu$; if $g \equiv 0$ it is the hard object process (3.1.2). If the (measurable) function $g \leq 1$, the process is *purely inhibitory*.

The special case $g \equiv \gamma$ for a constant $0 < \gamma < 1$ is called a *Strauss object process* and the density can be written

$$p(\mathbf{x}) = \alpha\beta^{n(\mathbf{x})}\gamma^{s(\mathbf{x})} \tag{3.1.4}$$

where

$$s(\mathbf{x}) = \sum_{i<j} 1\{x_i \sim x_j\}$$

is the number of pairs of neighbouring objects (e.g. number of overlaps) in the configuration. This process exhibits 'repulsion' or 'inhibition'

between objects, since $s(\mathbf{x})$ tends to be smaller than under the Poisson model. The density (3.1.4) is typically not integrable for $\gamma > 1$.

Just as discrete Markov random fields are closely connected with statistical physics, pairwise interaction processes (3.1.3) also occur as models of interacting particle systems. The log probability density of a particular configuration $\mathbf{x}$ can be interpreted as its physical 'energy': it is the sum of a ground potential $\log \alpha$, a potential $\log \beta$ for the presence of each object x_i, and an interaction potential $\log g(u, v)$ between each pair of neighbouring objects u, v.

The density (3.1.3) bears a close resemblance to the distribution of a discrete Markov random field with pairwise interaction. However, the number of terms appearing in the product in (3.1.3) depends on the realisation $\mathbf{x}$. Some configurations have more interaction than others.

Note that if $u \in U, u \notin \mathbf{x}$ with $p(\mathbf{x}) > 0$, the ratio

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} = \beta \prod_{x_i \sim u} g(u, x_i) \tag{3.1.5}$$

depends only on u and on the neighbours of u in $\mathbf{x}$. This important property signifies that all interaction is 'local'. In the statistical physics interpretation, $\log p(\mathbf{x} \cup \{u\}) - \log p(\mathbf{x})$ is the energy required to add a new point u to an existing configuration $\mathbf{x}$. In probabilistic terms $p(\mathbf{x} \cup \{u\})/p(\mathbf{x})$ is the Papangelou conditional intensity at u given the rest of the pattern $\mathbf{x}$ on $U \setminus \{u\}$, see [26].

Following are definitions and results of Ripley and Kelly [103] trivially generalised to random object processes [12, Section 3]. Let $\sim$ be any symmetric, reflexive relation on U.

Definition 2 *A random object process X with density p is called a* Markov object process *with respect to $\sim$ if for all $\mathbf{x} \in \Omega$*

(a) *$p(\mathbf{x}) > 0$ implies $p(\mathbf{y}) > 0$ for all $\mathbf{y} \subseteq \mathbf{x}$;*

(b) *if $p(\mathbf{x}) > 0$, then*

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \tag{3.1.6}$$

depends only on u and $N(\{u\}) \cap \mathbf{x} = \{x_i \in \mathbf{x} : u \sim x_i\}$.

The term 'Markov' is justified by the following spatial Markov property. Let A be a measurable subset of U. Then the conditional distribution of $X \cap A$ given $X \cap A^c$ depends only on X in the neighbourhood $N(A) \cap A^c = \{u \in A^c : u \sim a \text{ for some } a \in A\}$:

$$\mathcal{L}(X \cap A \mid X \cap A^c) = \mathcal{L}(X \cap A \mid X \cap N(A) \cap A^c).$$

Define a configuration $\mathbf{x} \in \Omega$ to be a *clique* if all members of $\mathbf{x}$ are neighbours ($x_i \sim x_j$ for all $i \neq j$). Configurations of 0 or 1 objects are cliques. Then the Ripley-Kelly analogue of the Hammersley-Clifford theorem [103] states that a process with density $p : \Omega \to [0, \infty)$ is Markov iff

$$p(\mathbf{x}) = \prod_{\text{cliques } \mathbf{y} \subseteq \mathbf{x}} q(\mathbf{y}) \tag{3.1.7}$$

for all $\mathbf{x} \in \Omega$, where the product is restricted to *cliques* $\mathbf{y} \subseteq \mathbf{x}$, and $q : \Omega \to [0, \infty)$ is an (arbitrary) function.

To conclude this Section, consider the following *area-interaction process*

$$p(\mathbf{x}) = \alpha \beta^{n(\mathbf{x})} \gamma^{-|S(\mathbf{x})|} \tag{3.1.8}$$

with parameters $\beta > 0$, $\gamma > 0$ and normalising constant $\alpha > 0$. As usual $|S(\mathbf{x})|$ is the area (or pixel count) of the silhouette. This is a *Markov overlapping object model* (i.e. Markov with respect to (3.1.1)) with interactions of infinite order. For $\gamma < 1$, configurations with relatively few overlapping objects are favoured, for $\gamma = 1$ it is a Poisson process and for $\gamma > 1$ the model encourages clustered patterns. This model is interesting in its own right and will be discussed in greater detail in Sections 3.2 and 3.4.

3.1.3 Nearest-neighbour Markov object processes

A further extension due to Baddeley and Møller [12] is to allow interaction behaviour to depend on the realisation of the process. For example, in a one-dimensional renewal process, each point can be said to interact with its nearest neighbours to the left and right, regardless of how far distant these neighbours may be. In two dimensions we would like to construct point processes exhibiting interaction between those pairs of points that are neighbours with respect to the Voronoi (Dirichlet) tessellation generated by the point pattern.

Assume that for each configuration $\mathbf{x}$ we have a symmetric reflexive relation $\underset{\mathbf{x}}{\sim}$ defined on $\mathbf{x}$. We might prefer to think of this as a finite graph whose vertices are the objects $x_i \in \mathbf{x}$.

Consider the following *pairwise interaction model* (cf. (3.1.3))

$$p(\mathbf{x}) = \alpha \beta^{n(\mathbf{x})} \prod_{i<j; x_i \underset{\mathbf{x}}{\sim} x_j} g(x_i, x_j).$$

Any generalisation of Definition 2 should at least embrace p. However

$$\frac{p(\mathbf{x}\cup\{u\})}{p(\mathbf{x})} = \beta \prod_{x_i \underset{\mathbf{x}\cup\{u\}}{\sim} u} g(x_i,u) \frac{\prod_{x_i \underset{\mathbf{x}\cup\{u\}}{\sim} x_j} g(x_i,x_j)}{\prod_{x_i \underset{\mathbf{x}}{\sim} x_j} g(x_i,x_j)}.$$

The extra factor arises, since $\underset{\mathbf{x}}{\sim}$ is depending on the pattern $\mathbf{x}$. If $\frac{p(\mathbf{x}\cup\{u\})}{p(\mathbf{x})}$ is to depend on 'local' information only, conditions must be imposed on $\underset{\mathbf{x}}{\sim}$.

The following definitions and results are taken from [12] to which we refer for further details. Define the $\mathbf{x}$–neighbourhood of a subset $\mathbf{z} \subseteq \mathbf{x}$ as

$$N(\mathbf{z} \mid \mathbf{x}) = \left\{\xi \in \mathbf{x} : \xi \underset{\mathbf{x}}{\sim} \eta \text{ for some } \eta \in \mathbf{z}\right\}.$$

Let $\mathbf{y} \subseteq \mathbf{z} \in \Omega$ and $u, v \in U$ with $u, v \notin \mathbf{z}$. Then require

(C1) $\chi(\mathbf{y} \mid \mathbf{z}) \neq \chi(\mathbf{y} \mid \mathbf{z}\cup\{u\})$ implies $\mathbf{y} \subseteq N(\{u\} \mid \mathbf{z}\cup\{u\})$;

(C2) if $u \not\sim_{\mathbf{x}} v$ where $\mathbf{x} = \mathbf{z}\cup\{u\}\cup\{v\}$ then

$$\chi(\mathbf{y} \mid \mathbf{z}\cup\{u\}) + \chi(\mathbf{y} \mid \mathbf{z}\cup\{v\}) = \chi(\mathbf{y} \mid \mathbf{z}) + \chi(\mathbf{y} \mid \mathbf{x}).$$

where χ is the clique indicator function.

Examples of relations satisfying these conditions are

- $x_i \underset{\mathbf{x}}{\sim} x_j$ iff $x_i \sim x_j$ where $\sim$ is any symmetric, reflexive relation on U (i.e. not depending on the configuration $\mathbf{x}$);
- for points or marked points in $\mathbb{R}^2$, $x_i \underset{\mathbf{x}}{\sim} x_j$ iff x_i, x_j are joined by an edge of the Delaunay triangulation generated by $\mathbf{x}$;
- for compact sets in $\mathbb{R}^d$, $x_i \underset{\mathbf{x}}{\sim} x_j$ iff x_i and x_j belong to the same connected component of the union of the objects.

The proofs can be found in [12, Appendix].

Definition 3 *A random object process with density p is called a* nearest-neighbour Markov object process (nnMp) *with respect to* $\{\underset{\mathbf{x}}{\sim} : \mathbf{x} \in \Omega\}$ *if, for all* $\mathbf{x}$ *with* $p(\mathbf{x}) > 0$

- $p(\mathbf{y}) > 0$ *for all* $\mathbf{y} \subseteq \mathbf{x}$;
- *the ratio* $\frac{p(\mathbf{x}\cup\{u\})}{p(\mathbf{x})}$ *depends only on* u, *on* $N(\{u\} \mid \mathbf{x}\cup\{u\}) \cap \mathbf{x} = \{x_i \in \mathbf{x} : x_i \underset{\mathbf{x}\cup\{u\}}{\sim} u\}$ *and on the relations* $\underset{\mathbf{x}}{\sim}, \underset{\mathbf{x}\cup\{u\}}{\sim}$ *restricted to* $N(\{u\} \mid \mathbf{x}\cup\{u\}) \cap \mathbf{x}$.

A subset $\mathbf{y} \subseteq \mathbf{x}$ is a *clique in* $\mathbf{x}$ if all members of $\mathbf{y}$ are $\mathbf{x}$-neighbours of one another ($u \underset{\mathbf{x}}{\sim} v$ for all $u, v \in \mathbf{y}$).

A generalised Hammersley-Clifford theorem holds [12]: a process with density p is nnMp iff

$$p(\mathbf{x}) = \begin{cases} \prod_{\text{cliques } \mathbf{y} \subseteq \mathbf{x}} q(\mathbf{y}) & \text{if } q(\mathbf{y}) > 0 \text{ for } \textit{all } \mathbf{y} \subseteq \mathbf{x} \\ 0 & \text{otherwise} \end{cases}$$

where $q : \Omega \to \mathbb{R}_+$ satisfies certain regularity conditions ((I1)–(I2) of [12]).

Kendall [61] proved a spatial Markov property for nnMp's.

3.2. AREA-INTERACTION PROCESSES

Pairwise interaction Markov processes (3.1.3) have been studied intensively. They provide a flexible class for modelling inhibition patterns and can be interpreted quite easily using the Hammersley-Clifford representation (3.1.7). However, they do not seem to be able to produce clustered patterns in sufficient variety. The original clustering model of Strauss [112] turned out [60] to be non-integrable for parameter values $\gamma > 1$ corresponding to the desired clustering; Gates and Westcott [39] showed that partly-attractive potentials may violate a stability condition, implying that they produce extremely clustered patterns with high probability; and recent simulation experiments by Møller [85, 86] suggest that the behaviour of the Strauss model with fixed n undergoes an abrupt transition from "Poisson-like" patterns to tightly clustered patterns rather than exhibiting intermediate, moderately clustered patterns.

The area-interaction model (3.1.8) in contrast is a Markov model that can yield both moderately clustered and moderately ordered patterns.

It has interactions of infinite order, and is similar in form to the pairwise interaction Strauss model (3.1.4). Both densities reduce to a Poisson process when $\gamma = 1$, and exhibit ordered patterns for $0 < \gamma < 1$; in contrast to the Strauss model, (3.1.8) is well-defined for all values of $\gamma > 0$ and produces clustering when $\gamma > 1$. The attractive case was proposed by Widom and Rowlinson [119] as the 'penetrable sphere model' for spherical molecules in the study of liquid-vapour phase transitions.

3.2.1 Definition of the process

As usual for Gibbs point processes we treat separately the cases of a finite point process (say, points in a bounded region $A \subseteq \mathbb{R}^d$) and a stationary point process on $\mathbb{R}^d$. The formal construction of finite Gibbs point processes is described in [26, p. 121 ff] or [96].

As before, let U be a locally compact complete separable metric space (typically $\mathbb{R}^d$ or a compact subset). The space of all possible realisations $\mathbf{x}$ shall be identified with the space N^f of all integer-valued measures on U which have finite total mass and are simple (do not have atoms of mass exceeding 1). Write $n(\mathbf{x})$ for the total number of points and $\mathbf{x}_B$ for $\mathbf{x}$ restricted to $B \subseteq U$. The σ-algebra $\mathcal{N}^f$ on N^f is the Borel σ-algebra of the weak topology, i.e. $\mathcal{N}^f$ is the smallest σ-algebra with respect to which the evaluation $\mathbf{x} \mapsto n(\mathbf{x}_B)$ is measurable for every (bounded) Borel set $B \subseteq U$.

Given a totally finite, non-atomic measure μ on U, construct the Poisson process of intensity μ as in Section 3.1.2 (typically μ is the restriction of Lebesgue measure to a compact window $A \subseteq \mathbb{R}^d$, yielding the unit

rate Poisson process restricted to A). Let π be its probability distribution on $(N^f, \mathcal{N}^f)$. Then we construct (Gibbs) point processes by specifying their density with respect to π.

Definition 4 ("Standard case") *The* area-interaction process *in a compact region $A \subseteq \mathbb{R}^d$ is the process with density*

$$p(\mathbf{x}) = \alpha\, \beta^{n(\mathbf{x})}\, \gamma^{-m(S_r(\mathbf{x}))} \tag{3.2.9}$$

with respect to the unit rate Poisson process on A, where $\beta, \gamma, r > 0$ are parameters and α is the normalising constant. Here m is Lebesgue measure and

$$S_r(\mathbf{x}) = \bigcup_{i=1}^{n} B(x_i, r)$$

is the union of spheres or discs of radius r centred at the points of the realisation, $B(x_i, r) = \left\{a \in \mathbb{R}^d : ||a - x_i|| \leq r\right\}$.

For $\gamma = 1$ this of course reduces to a Poisson process with intensity β. It is intuitively clear that for $0 < \gamma < 1$ the pattern will tend to be 'ordered' and for $\gamma > 1$ 'clustered'. The clustered model was introduced by Widom and Rowlinson [119].

It is often more convenient to use the parameter $\eta = \gamma^{-\pi r^2}$, since the addition of one point a to the configuration $\mathbf{x}$ alters $p(\mathbf{x})$ by a factor ranging from β to $\beta\eta$.

Various modifications are of interest, for example, one may wish to replace $S_r(\mathbf{x})$ by $A \cap S_r(\mathbf{x})$, or to allow the radii of the discs $B(x_i, r)$ to vary across the region [64].

Definition 5 ("General case") *Let ν be a totally finite, regular Borel measure on U and $Z : U \to \mathcal{K}$ a myopically continuous function [81, p. 12], assigning to each point $u \in U$ a set $Z(u) \subseteq U$ in the space of all compact subsets $\mathcal{K}$. Then the* general area-interaction process *is defined to have density*

$$p(\mathbf{x}) = \alpha\, \beta^{n(\mathbf{x})}\, \gamma^{-\nu(S(\mathbf{x}))} \tag{3.2.10}$$

with respect to π (the distribution of the finite Poisson process with intensity μ), where $S(\mathbf{x})$ is the compact set $\bigcup_{i=1}^{n} Z(x_i)$.

In a parametric statistical model the measure ν and the definition of $Z(\cdot)$ might also be allowed to depend on the parameter θ. Other generalisations are possible, for instance

$$p(\mathbf{x}) = a(\theta) \prod_{i=1}^{n} b(x_i;\theta) \exp\left(-\int_A f(d(\mathbf{x},u))\,du\right)$$

where $d(\mathbf{x},u) = \min_i ||x_i - u||$ and $f : [0,\infty] \to (-\infty,\infty]$. The model we will discuss in this Section is a special case, analogous to the Strauss process.

Lemma 5 *The density (3.2.10) is measurable and integrable for all values of $\beta, \gamma > 0$.*

Proof: Let $t > 0$ and consider $V = \left\{\mathbf{x} \in N^f : \nu(S(\mathbf{x})) < t\right\}$. We show that V is open in the weak topology.

Choose $\mathbf{x} \in V$. Since ν is regular, there is an open set $G \subseteq U$ containing $U(\mathbf{x})$ such that $\nu(G) < t$ too. Now $U(\mathbf{x}) \subseteq G$ iff $\mathbf{x}$ has no points in $H = \{u \in U : Z(u) \cap G^c \neq \emptyset\}$. But the class of all compact sets intersecting a given closed set is closed in the myopic topology on $\mathcal{K}$. Since $u \mapsto Z(u)$ is myopically continuous, H is closed in U. Moreover, $W = \left\{\mathbf{y} \in N^f : n(\mathbf{y} \cap H) = 0\right\}$ is open in the weak topology. To see this, note that for any $\mathbf{y} \in W$, the open ball in the Prohorov metric with radius $\epsilon < \min\{d(\mathbf{y},H), 1\}$ centred at $\mathbf{y}$ is an open environment of $\mathbf{y}$ contained in W. Thus, W is an open environment of $\mathbf{x}$ contained in V and V is weakly open.

In fact this shows that $\mathbf{x} \mapsto \nu(S(\mathbf{x}))$ is weakly upper semicontinous. It follows that the map $g : N^f \to [0,\infty)$ defined by $\mathbf{x} \mapsto \exp[-\nu(S(\mathbf{x})) \log\gamma]$ is weakly upper semicontinuous for $\gamma \in (0,1)$ and lower semicontinuous for $\gamma > 1$. Hence g is measurable. By definition of the weak topology, $\mathbf{x} \mapsto \beta^{n(\mathbf{x})}$ is measurable, and hence the density (3.2.10) is measurable.

To check integrability, observe that

$$0 \leq \nu(S(\mathbf{x})) \leq \nu(U) < \infty \tag{3.2.11}$$

Now the function $f(\mathbf{x}) = \beta^{n(\mathbf{x})}$ is integrable, yielding the Poisson process with intensity measure $\beta\mu$. Hence (3.2.10) is dominated by an integrable function, hence integrable. □

In fact (3.2.11) establishes a slightly stronger result.

Lemma 6 *The distribution $P_{\beta,\gamma}$ of the general area interaction process is uniformly absolutely continuous with respect to the distribution of the Poisson process π^β with intensity $\beta\mu$, that is its Radon-Nikodym derivative is uniformly bounded.*

Explicit bounds on the density $\alpha\gamma^{-\nu(S(\mathbf{x}))}$ with respect to a Poisson process of rate β are

$$\min\left\{\gamma^{\nu(U)}, \gamma^{-\nu(U)}\right\} \leq f \leq \max\left\{\gamma^{\nu(U)}, \gamma^{-\nu(U)}\right\}.$$

This suggests that the 'singularity' (highly clustered behaviour) of the Strauss model is unlikely.

As usual, the normalising constant α is difficult to compute, since

$$1/\alpha = \mathbb{E}\beta^{n(X)}\gamma^{-\nu(S(X))}$$

where the expectation is with respect to the reference Poisson process; this entails computing the moment generating function of $\nu(S(X))$, or equivalently, the vacancy distribution in the coverage problem [51]. A notable exception is the 1-dimensional penetrable sphere model [119].

Area-interaction seems a plausible model for some biological processes. For example the points x_i may represent plants or animals which consume food within a radius r of their current location. The total area of accessible food is then $S_r(\mathbf{x})$, and the herd will tend to maximise this area, so an area-interaction model with $\gamma < 1$ is plausible. Alternatively assume that the animals or plants are hunted by a predator which appears at a random position and catches any prey within a distance r. Then $S_r(\mathbf{x})$ is the area of vulnerability, and the herd as a whole should tend to minimise this [53] so an area-interaction model with $\gamma > 1$ is plausible. Area-interaction processes with $\gamma > 1$ are also used as a model for liquid-vapour equilibrium in chemical physics.

The process can be derived from Poisson processes.

Lemma 7 *Let X, Y be independent Poisson processes in A with intensity measures $\beta\mu$ and $|\log\gamma|\nu$ respectively. If $\gamma > 1$ then the conditional distribution of X given $\{Y \cap S(X) = \emptyset\}$ is an area-interaction process with parameter γ. If $\gamma < 1$ then the conditional distribution of X given $\{Y \subseteq S(X)\}$ is an area-interaction process with parameter γ.*

Proof: If $\gamma > 1$

$$\mathbb{P}(Y \cap S(X) = \emptyset | X) = e^{-\nu(S(X))\log\gamma} = \gamma^{-\nu(S(X))}$$

hence the conditional distribution of X given $Y \cap S(X) = \emptyset$ has a density proportional to the right hand side. Similarly if $\gamma < 1$

$$\begin{aligned}\mathbb{P}(Y \subseteq S(X)|X) &= \mathbb{P}(Y \cap (A \setminus S(X)) = \emptyset) \\ &= e^{\nu(A\setminus S(X))\log\gamma} = \gamma^{\nu(A)}\gamma^{-\nu(S(X))}.\end{aligned}$$

□

3.2.2 Limiting cases

Here we study the convergence of the area-interaction process as $\gamma \longrightarrow 0, \infty$.

Let $\nu^* = \max_{\mathbf{x}} \nu(S(\mathbf{x}))$, typically the measure of the observation window or its (generalised) dilation and

$$H = \{\mathbf{x} : \nu(S(\mathbf{x})) = \nu^*\}.$$

Further, write π^β for the distribution of the Poisson process of rate β in A and finally, in the standard case, denote

$$HC = \left\{\mathbf{x} : \ m(S(\mathbf{x})) = n(\mathbf{x})\pi r^2\right\}$$

for the set of configurations respecting a hard core distance r.

Lemma 8 *Let $P_{\beta,\gamma}$ be the distribution of the area interaction process with density (3.2.10).*

If $\gamma \to 0$ with β fixed, then $P_{\beta,\gamma}$ converges to a uniform process on H, i.e. $P_{\beta,\gamma}(A) \to \pi^\beta(A \cap H)/\pi^\beta(H)$.

In the standard case, if $\gamma \to 0$ and $\beta \to 0$ so that $\beta\gamma^{-\pi r^2} \to \zeta \in (0, \infty)$, then $P_{\beta,\gamma}$ converges to $P(A) = \pi^\zeta(A \cap HC)/\pi^\zeta(HC)$, a hard core process.

If $\gamma \to \infty$ with $\beta < \infty$ fixed, then $P_{\beta,\gamma}$ converges to a process that is empty with probability 1.

Proof: First consider $\gamma \to 0$. Then $\int \gamma^{\nu^* - \nu(S(\mathbf{x}))} d\pi^\beta(\mathbf{x}) \to \pi^\beta(H)$, hence

$$P_{\beta,\gamma}(\mathbf{x}) = \frac{\gamma^{\nu^* - \nu(S(\mathbf{x})}}{\int \gamma^{\nu^* - \nu(S(\mathbf{x}))} d\pi^\beta(\mathbf{x})} \to \frac{\mathbf{1}\{\mathbf{x} \in H\}}{\pi^\beta(H)}$$

from which the first assertion follows. To prove the second assertion, note that for $m(S_r(\mathbf{x})) < n(\mathbf{x})\pi r^2$, $\gamma^{n(\mathbf{x})\pi r^2 - m(S_r(\mathbf{x}))} \to 0$. Hence

$$P_{\beta,\gamma}(\mathbf{x}) \to \frac{\zeta^{n(\mathbf{x})}\mathbf{1}\{\mathbf{x} \in HC\}}{\int_{HC} \zeta^{n(\mathbf{x})} d\pi^1(\mathbf{x})}.$$

The third statement follows similarly, by noting that the density converges pointwise to zero unless the pattern is empty. □

3.2.3 Markov property

Define two points $a, b \in U$ to be neighbours whenever $Z(a) \cap Z(b) \neq \emptyset$, as in (3.1.1). In the standard case $a \sim b$ iff $||a - b|| \leq 2r$.

Lemma 9 *The area-interaction process (3.2.10) is a Markov point process with respect to (3.1.1) in the sense of Ripley and Kelly [103].*

Proof: The likelihood ratio

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} = \beta\gamma^{-\nu(Z(u)\setminus S(\mathbf{x}))} \tag{3.2.12}$$

is computable in terms of u and $\{x_i : x_i \sim u\}$, since

$$\begin{aligned} Z(u) \setminus S(\mathbf{x}) &= Z(u) \cap \left[\bigcup_i Z(x_i)\right]^c \\ &= Z(u) \cap \left[\bigcup_{x_i \sim u} Z(x_i)\right]^c . \end{aligned}$$

Hence (3.2.10) defines a Markov point process with respect to $\sim$. □

The Ripley-Kelly analogue of the Hammersley-Clifford theorem (3.1.7) then implies that the density p can be written as a product of clique interaction terms

$$p(\mathbf{x}) = \prod_{\mathbf{y} \subseteq \mathbf{x}} q(\mathbf{y})$$

where $q(\mathbf{y}) = 1$ unless $y_i \sim y_j$ for all elements of $\mathbf{y}$. To compute the interaction terms explicitly, invoke the inclusion-exclusion formula:

$$\begin{aligned} \nu(S(\mathbf{x})) \quad = \quad & \sum_{i=1}^{n} \nu(Z(x_i)) - \sum_{i<j} \nu(Z(x_i) \cap Z(x_j)) \; + \\ \cdots + \quad & (-1)^{n+1} \nu\left(\bigcap_{i=1}^{n} Z(x_i)\right) \end{aligned}$$

which gives

$$\begin{aligned} q(\emptyset) &= \alpha \\ q(\{a\}) &= \beta\gamma^{-\nu(Z(a))} \\ q(\{y_1, \ldots, y_k\}) &= \gamma^{(-1)^k \nu(\bigcap_{i=1}^{k} Z(y_i))} \end{aligned} \tag{3.2.13}$$

That is, the process exhibits *interactions of infinite order.*

The process satisfies a spatial Markov property (cf. [61, 103]). Define the dilation of a set $E \subseteq U$ by

$$D_Z(E) = \{u \in U : \exists e \in E \text{ such that } Z(u) \cap Z(e) \neq \emptyset\}; \tag{3.2.14}$$

in the standard case this becomes the classical dilation of mathematical morphology

$$D_Z(E) = \left\{ u \in \mathbb{R}^d : d(u, E) \leq 2r \right\} \tag{3.2.15}$$

where $d(u, E) = \inf \{||u - v|| : v \in E\}$. Then the spatial Markov property states that the restriction of the process to E is conditionally independent of the restriction to $D_Z(E)^c$ given the information in $D_Z(E) \setminus E$.

Figure 3.1 shows simulated realisations of (3.2.9). The number of points was fixed and the alternating birth/death technique of [98] used. This method entails alternating deletion of a random point and generation of a new point u from a probability density proportional to $p(\mathbf{x} \cup \{u\})$. The latter step can be implemented using rejection sampling, since the birth ratio (3.1.6) is dominated by a known constant by virtue of (3.2.11). In practice this will be a good bound for γ close to 1.

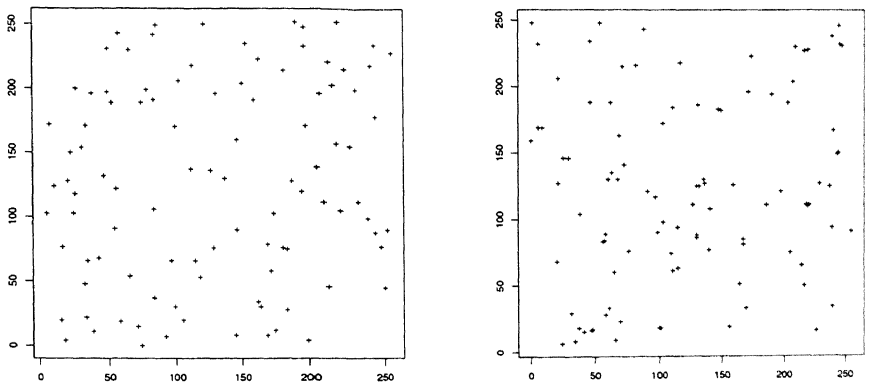


Figure 3.1: Simulated realisations of an area-interaction process conditional on $n = 100$ points, with $r = 5$ in a window of size 256×256. *Left:* ordered pattern, $\gamma = 0.9711$, $\gamma^{-25\pi} = 10$; *Right:* clustered pattern, $\gamma = 1.02975$, $\gamma^{-25\pi} = 0.1$.

3.3. STATIONARY AREA-INTERACTION PROCESS

Here we use the methods of Preston [96] to check that the area interaction model (standard case) can be considered as the restriction to a bounded sampling window of a stationary point process on the whole of $\mathbb{R}^d$.

Let N be the space of all locally finite counting measures (i.e. integer-valued Radon measures) on $\mathbb{R}^d$ with the vague topology, and $\mathcal{N}$ its Borel σ-algebra; that is, $\mathcal{N}$ is the smallest σ-algebra making $\mathbf{x} \mapsto n(\mathbf{x}_B)$ measurable for all bounded Borel sets B.

Write $\mathcal{C}$ for the class of all bounded Borel sets in $\mathbb{R}^d$. For every $B \in \mathcal{C}$ let N_B be the subspace of those $\mathbf{x} \in N$ contained in B (i.e. putting no mass outside B), $\mathcal{N}_B \subseteq \mathcal{N}$ the induced σ-algebra on N and π_B^β the distribution on $(N, \mathcal{N})$ of the homogeneous Poisson process of rate β on B. Note that any $\mathbf{x} \in N$ can be decomposed as $\mathbf{x} = \mathbf{x}_B \cup \mathbf{x}_{B^c}$. Define $f^B : N \to [0, \infty)$ by

$$f^B(\mathbf{x}) = \alpha_B(\mathbf{x}_{B^c})\, \gamma^{-m(S_r(\mathbf{x}) \cap B_{\oplus r})} \tag{3.3.16}$$

where $S_r(\mathbf{x}) = \bigcup_{x_i \in \mathbf{x}} B(x_i, r)$, $B_{\oplus r} = B \oplus B(0, r)$ and $\alpha_B(\mathbf{x}_{B^c})$ is the normalising constant

$$\alpha_B(\mathbf{x}_{B^c})^{-1} = \int_{N_B} \gamma^{-m(S_r(\mathbf{y} \cup \mathbf{x}_{B^c}) \cap B_{\oplus r})}\, d\pi_B^\beta(\mathbf{y}).$$

To check that this is well-defined, observe that

$$S_r(\mathbf{x}) \cap B_{\oplus r} = S_r(\mathbf{x}_{B_{\oplus 2r}}) \cap B_{\oplus r}$$

so that $\mathbf{x} \mapsto m(S_r(\mathbf{x}) \cap B_{\oplus r})$ is measurable with respect to $\mathcal{N}_{B_{\oplus 2r}}$ and a fortiori with respect to $\mathcal{N}$. It is clearly integrable. Hence, by Fubini's theorem, $\alpha_B(\cdot)^{-1}$ is $\mathcal{N}_{B_{\oplus 2r} \setminus B}$-measurable. Since α_B is uniformly bounded away from zero, α^{-1} is also measurable. Thus, (3.3.16) is $\mathcal{N}$-measurable and integrable, and we may define for $\mathbf{x} \in N, F \in \mathcal{N}$

$$\kappa_B(\mathbf{x}, F) = \int_{N_B} 1_F(\mathbf{x}_{B^c} \cup \mathbf{y}) f^B(\mathbf{x}_{B^c} \cup \mathbf{y})\, d\pi_B^\beta(\mathbf{y}). \tag{3.3.17}$$

Theorem 10 *There exists a stationary point process X on $\mathbb{R}^d$ such that*

$$\mathbb{P}(X \in F \mid X_{B^c}) = \kappa_B(X, F) \quad a.s.$$

for all $B \in \mathcal{C}$ and $F \in \mathcal{N}$. That is, (3.3.17) is a specification without forbidden states [96, p. 12] and the distribution of X is a stationary Gibbs state with this specification. The corresponding potential $V : N^f \to \mathbb{R}$,

$$V(\mathbf{x}) = (-\log \gamma) m(S_r(\mathbf{x}))$$

is stable [96, p. 96].

Note that this result does not exclude the possibility that the Gibbs state is not unique, i.e. there may be 'phase transition' [96, p. 46].

Proof: First we prove consistency condition (6.10) of [96, p. 91]. For any bounded Borel sets $A \subseteq B$

$$\frac{f^B(\mathbf{x})}{f^A(\mathbf{x})} = \frac{\alpha_B(\mathbf{x}_{B^c})}{\alpha_A(\mathbf{x}_{A^c})}\, \gamma^{-m(S_r(\mathbf{x}_{A^c}) \cap B_{\oplus r} \setminus A_{\oplus r})}.$$

On the other hand

$$\int_{N_A} f^B(\mathbf{x}_{A^c} \cup \mathbf{y})\, d\pi_A^\beta(\mathbf{y}) = \alpha_B(\mathbf{x}_{B^c}) \int_{N_A} \gamma^{-m(S_r(\mathbf{x}_{A^c} \cup \mathbf{y}) \cap B_{\oplus r})}\, d\pi_A^\beta(\mathbf{y}).$$

Since $m(S_r(\mathbf{x}_{A^c} \cup \mathbf{y}) \cap B_{\oplus r}) = m(S_r(\mathbf{x}_{A^c} \cup \mathbf{y}) \cap A_{\oplus r}) + m(S_r(\mathbf{x}_{A^c}) \cap B_{\oplus r} \setminus A_{\oplus r})$, $\int_{N_A} f^B(\mathbf{x}_{A^c} \cup \mathbf{y})\, d\pi_A^\beta(\mathbf{y}) = f^B(\mathbf{x})/f^A(\mathbf{x})$. It follows [96, pp. 90–91] that $\{\kappa_B : B \in \mathcal{C}\}$ is a specification in the sense of [96, p. 12].

Now we check the conditions of Theorem 4.3 of [96, p. 58]. Condition (3.6) of [96, p. 35] is trivially satisfied. Arguments similar to those used to derive Lemma 6 above yield that for any $K \in \mathcal{K}$ the family $\{\pi_K(\mathbf{y}, \cdot)\}_{\mathbf{y} \in N}$ considered as a class of measures on $(N, \mathcal{N}_K)$ is uniformly absolutely continuous with respect to π_K^β; hence Preston's condition (3.11) [96, p. 41] holds, which implies his (3.10). It remains to check (3.8) of [96, p. 35]. Let $\mathcal{K}$ be the class of all compact subsets of $\mathbb{R}^d$; then we claim that

> for any $B \in \mathcal{C}$ and $F \in \mathcal{N}_K$ where $K \in \mathcal{K}$, there exists $L \in \mathcal{K}$ such that $\kappa_B(\cdot, F)$ is measurable with respect to $\mathcal{N}_L$.

To check this, choose L to contain $K \cup B_{\oplus 2r}$ and observe that $1_F(\mathbf{x}_{B^c} \cup \mathbf{y})$ and $f^B(\mathbf{x}_B^c \cup \mathbf{y})$ are measurable with respect to $\mathcal{N}_L \otimes \mathcal{N}_B$, then apply Fubini's theorem. This proves the claim, which implies Preston's (3.8) and hence the conditions of his Theorem 4.3.

It is easy to see that V is the unique canonical potential corresponding to the densities f^B (cf. [96, p. 92]). Since $0 \le m(S_r(\mathbf{x})) \le \pi r^2 n(\mathbf{x})$, V is stable. □

3.4. Inference

This Section surveys parameter estimation techniques. We also derive a new identity from the Takacs-Fiksel formula that can be computed quite generally as an index for clustering and for estimating model parameters.

3.4.1 Sufficient statistics, exponential families

Consider a family of area-interaction processes (3.2.9) indexed by a parameter $\theta = (\beta, \gamma)$, with $A \subseteq \mathbb{R}^d$ fixed. This is an exponential family with canonical sufficient statistic

$$T(\mathbf{x}) = (n(\mathbf{x}), M(\mathbf{x}, \cdot))$$

where

$$M(\mathbf{x}, r) = m(S_r(\mathbf{x})) = m(\bigcup_{i=1}^{n} B(x_i, r)).$$

Modifying this slightly we obtain a connection with the 'empty space statistic' $F(t) = \mathbb{P}(X \cap B(0,t) \neq \emptyset)$ [31, 99]. Define

$$\widehat{F}(r) = \frac{m(A_{(-r)} \cap (\bigcup_{i=1}^{n} B(x_i, r)))}{m(A_{(-r)})} \tag{3.4.18}$$

where

$$A_{(-t)} = \{a \in A : B(a,t) \subseteq A\}.$$

This is the 'border method' [100, p. 25] or 'reduced sample' estimator [4] of the empty space function.

Lemma 11 *$(n(\mathbf{x}), \widehat{F})$ is a sufficient statistic for the area-interaction process (3.2.10) with parameters (β, γ) when $Z(u) = B(u, r)$ and the measure ν is Lebesgue measure restricted to $A_{(-r)}$.*

The canonical parameter is $-\log \gamma$ but we prefer to use γ to maintain the comparison with the Strauss process.

3.4.2 Maximum likelihood

As usual for Markov point processes, the likelihood (3.2.10) is easy to compute except for a normalising constant α that is not known analytically. Maximum likelihood estimation therefore rests on numerical or Monte Carlo approximations of α [89, 90, 91, 94] or recursive approximation methods [87]. For a more detailed review see [33, 45] or [100].

We will not explore this further here, except to note that the maximum likelihood estimating equations are as usual

$$n(\mathbf{x}) = \mathbb{E}_{\beta,\gamma} n(X) \tag{3.4.19}$$

$$\nu(S(\mathbf{x})) = \mathbb{E}_{\beta,\gamma} \nu(S(X)) \tag{3.4.20}$$

where $\mathbf{x}$ is the observed pattern and X is a random pattern with density (3.2.10). For the model conditioned on $n(\mathbf{x}) = n$ the ML estimating equation is analogous to (3.4.20) with β absent.

3.4.3 Takacs-Fiksel

The Takacs-Fiksel estimation method exploits the identity

$$\lambda \, \mathbb{E}_a^! f(X) = \mathbb{E}[\lambda(a; X) \, f(X)] \tag{3.4.21}$$

holding for any bounded measurable non-negative function $f : N \to \mathbb{R}_+$ and any stationary Gibbs process X on $\mathbb{R}^d$ with finite intensity λ [37, 38, 114, 115], see [100, p. 54–55], [33, §2.4] . The expectation on the left hand side of (3.4.21) is with respect to the reduced Palm distribution of X at $a \in \mathbb{R}^d$, and $\lambda(a; \mathbf{x})$ is the Papangelou conditional intensity of X at a.

A Takacs-Fiksel method [37, 38, 114, 115] is then to choose suitable functions and to estimate both sides in the above formula. The resulting set of equations is solved, yielding estimates for the parameters of the model. The idea was originally suggested by Takacs [114, 115] for a particular case and generalised by Fiksel [37, 38]. A variant using only nearest neighbour measurements was developed by Tomppo [117].

When X is a Gibbs point process the conditional intensity can be computed in terms of the potential [59]. For the standard area-interaction process

$$\lambda(a; \mathbf{x}) = \beta \gamma^{-m(B(a,r) \setminus S_r(\mathbf{x}))}$$

for $a \notin \mathbf{x}$; this depends only on $\mathbf{x}_{B(a,2r)}$.

One interesting instance of (3.4.21) is

$$f(\mathbf{x}) = \frac{\mathbf{1}\{\mathbf{x} \cap B(0,s) = \emptyset\}}{\lambda(0; \mathbf{x})}$$

using 0 as an arbitrary point of $\mathbb{R}^d$ (cf. [111, (5.5.18), p. 159]). Then $\mathbb{E}[\lambda(0; X) \, f(X)] = 1 - F(s)$ where $F(s) = \mathbb{P}(X \cap B(0,s) \neq \emptyset)$ is the empty space function of X. Now if $s > 2r$,

$$\begin{aligned} \lambda \, \mathbb{E}_0^! f(X) &= \lambda \int \mathbf{1}\{X \cap B(0,s) = \emptyset\} \, \beta^{-1} \gamma^{m(B(0,r) \setminus S_r(X))} \, dP_0^!(X) \\ &= \lambda \int \mathbf{1}\{X \cap B(0,s) = \emptyset\} \, \beta^{-1} \gamma^{m(B(0,r))} \, dP_0^!(\mathbf{x}) \\ &= \lambda \, \beta^{-1} \gamma^{\pi r^2} \, [1 - G(s)] \end{aligned}$$

where $G(s) = \mathbb{P}_0^!(X \cap B(0,s) \neq \emptyset)$ is the nearest neighbour distance distribution function of X. Equivalently

$$\lambda \frac{1-G(s)}{1-F(s)} = \beta\gamma^{-\pi r^2}. \tag{3.4.22}$$

for all $s > 2r$.

Thus parameter estimates for the area-interaction process can be extracted directly from the standard statistics F and G. Similar identities hold for any (Markov) model with finite range of interaction. The ratio can be computed in many cases, when F and G separately cannot, and the graph provides an estimate for the interaction range. For further development of this idea see [10].

Identity (3.4.22) also provides a further description of the typical pattern generated by the model; F and G are used in spatial statistics as measures of clustering versus regularity, with $1 - F(t) = 1 - G(t) = \exp\{-\lambda\pi t^2\}$ and hence $(1-G)/(1-F) \equiv 1$ in the case of a Poisson process. Clustering is sometimes characterised by elevated G and lowered F; regularity by elevated F and lowered G [31]. Hence $(1-G)/(1-F) < 1$ for clustering and > 1 for regularity.

3.4.4 Approximation by lattice processes

Besag, Milne and Zachary [19] proved that any purely inhibitory (or hard core) pairwise interaction point process is the weak limit of a sequence of lattice processes, and remark [19, p. 214] this is also true of general Gibbs point processes, again of purely inhibitory type. Here we extend the result to the area-interaction model, which is *not* inhibitory.

Consider a partition $\{C_1, \ldots, C_m\}$ of the observation window A and choose representatives $\xi_i \in C_i$. Denote the area of C_i by $A_i > 0$ and the set of all representatives by Ξ. We shall construct a $\{0,1\}$-valued stochastic process $\mathbf{n} = \{n_i : i = 1, \ldots, m\}$ which is auto-logistic,

$$\frac{\mathbb{P}(n_i = 1 \mid n_j, j \neq i)}{\mathbb{P}(n_i = 0 \mid n_j, j \neq i)} = \mu_i(C_i) \tag{3.4.23}$$

where

$$\mu_i(C_i) = A_i \prod q(\xi_i \cup \mathbf{y})^{\eta(\mathbf{y})};$$

the product ranges over all (possibly empty) subsets of $\Xi \setminus \{\xi_i\}$, the set function q is the clique interaction function (3.2.13) and $\eta(\mathbf{y}) = \prod_{\xi_j \in \mathbf{y}} n_j$ is either 0 or 1.

Given a realisation of $\mathbf{n}$, construct a realisation $\mathbf{x}$ of a point process such that if $n_i = 0$ then $\mathbf{x} \cap C_i$ is empty, while if $n_i = 1$ then $\mathbf{x} \cap C_i$ consists of one point uniformly distributed in C_i independently of other points.

Lemma 12 *The conditional distributions (3.4.23) specify a distribution for* $\mathbf{n}$ *and*

$$\frac{p(\mathbf{n})}{p(\mathbf{0})} = \prod_i A_i^{n_i} \prod_{\emptyset \neq \mathbf{y} \subseteq \Xi} q(\mathbf{y})^{\eta(\mathbf{y})}.$$

The point process X *is absolutely continuous with respect to a unit rate Poisson process on* A*, with density*

$$f(\mathbf{x}) = f(\emptyset) \prod_{\emptyset \neq \mathbf{y} \subseteq \Xi} q(\mathbf{y})^{\eta_{\mathbf{x}}(\mathbf{y})}.$$

Here $\eta_{\mathbf{x}}(\mathbf{y}) = \prod n(\mathbf{x} \cap C_j)$ *and the product ranges over all* j *such that* $\xi_j \in \mathbf{y}$.

Proof: Use Besag's factorisation theorem [15, p. 195]. □

Theorem 13 *Consider a sequence of partitions* $\mathcal{C}_r = \left\{C_{r,1}, \ldots, C_{r,m(r)}\right\}$ *such that* $\max_i \operatorname{diam}(C_{r,i}) \to 0$. *Then the corresponding point process* $\mathbf{x}^{(r)}$ *converges weakly and in total variation to the area-interaction process.*

Proof: Let f_r be the density of $\mathbf{x}^{(r)}$. For fixed $\mathbf{x}$

$$\frac{f_r(\mathbf{x})}{f_r(\emptyset)} \to \prod_{\emptyset \neq \mathbf{y} \subseteq \mathbf{x}} q(\mathbf{y}) = \beta^{n(\mathbf{x})}\, \gamma^{-\nu(S(\mathbf{x}))}$$

since all cells ultimately contain at most one point, and q is continuous in all its arguments. By dominated convergence,

$$\frac{1}{f_r(\emptyset)} = \int \frac{f_r(\mathbf{x})}{f_r(\emptyset)}\, d\pi(\mathbf{x}) \to \int \frac{1}{\alpha}\, p(\mathbf{x})\, d\pi(\mathbf{x}) = \frac{1}{\alpha};$$

thus,

$$f_r(\mathbf{x}) = \frac{f_r(\mathbf{x})}{f_r(\emptyset)}\, f_r(\emptyset) \to p(\mathbf{x})$$

pointwise and the theorem is proved. □

3.4.5 Pseudolikelihood estimation

Because of the lattice approximation, the model parameters can be estimated by pseudolikelihood [16, 19, 58]. Using the notation of Section 3.4.4, the pseudolikelihood for partition $\{C_i\}_{i=1}^m$ is

$$\begin{aligned} PL(\theta) \;=\; & \prod_{\mathbf{x} \cap C_i \neq \emptyset} P(n_i = 1 \mid n_j, j \neq i) \\ & \cdot \prod_{\mathbf{x} \cap C_i = \emptyset} \{1 - P(n_i = 1 \mid n_j, j \neq i)\}. \end{aligned}$$

Since

$$P(n_i = 1 \mid n_j, j \neq i) = \frac{A_i \lambda(\xi_i; \mathbf{x})}{1 + A_i \lambda(\xi_i; \mathbf{x})}$$

we have

$$\lim_{A_i \to 0} \frac{1}{A_i} P(n_i = 1 \mid n_j, j \neq i) = \lambda(\xi_i; \mathbf{x})$$

or $P(n_i = 1 \mid n_j, j \neq i) \approx A_i \lambda(\xi_i; \mathbf{x})$. Therefore

$$\begin{aligned} \log PL(\theta) - n(\mathbf{x}) \log A_i &= \sum_{\mathbf{x} \cap C_i \neq \emptyset} \log \left\{ \frac{1}{A_i} P(n_i = 1 \mid n_j, j \neq i) \right\} \\ &+ \sum_{\mathbf{x} \cap C_i = \emptyset} \log \{1 - P(n_i = 1 \mid n_j, j \neq i)\} \\ &\approx \sum_{\mathbf{x} \cap C_i \neq \emptyset} \log \lambda(\xi_i; \mathbf{x}) - \sum_{\mathbf{x} \cap C_i = \emptyset} A_i \lambda(\xi_i; \mathbf{x}) \\ &\approx \sum_{\xi \in \mathbf{x}} \log \lambda(\xi_i; \mathbf{x}) - \int_A \lambda(\xi_i; \mathbf{x}) \, d\xi. \end{aligned}$$

In the limit, the pseudolikelihood equations are

$$PL(\beta, \gamma; \mathbf{x}) = \exp \left\{ - \int_W \lambda(u; \mathbf{x}) du \right\} \prod_{i=1}^{n} \lambda(x_i; \mathbf{x}) \tag{3.4.24}$$

where $\lambda(\cdot; \mathbf{x})$ is the Papangelou conditional intensity. For the area-interaction model (3.2.9)

$$\lambda(u; \mathbf{x}) = \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x} \setminus \{u\})} = \beta \, \gamma^{-m(B(u,r) \setminus S_r(\mathbf{x} \setminus \{u\}))}.$$

Writing $t(u) = m(B(u,r) \setminus S_r(\mathbf{x} \setminus \{u\}))$ the maximum pseudolikelihood estimates of β and γ are the solutions of

$$n = \beta \int_A \gamma^{-t(u)} du \tag{3.4.25}$$

$$\sum_{i=1}^{n} t(x_i) = \beta \int_A t(u) \gamma^{-t(u)} du \tag{3.4.26}$$

Note that these have exactly the same form as the pseudolikelihood equations for the Strauss model [100, p. 53]; in that case $-t(u)$ is the number of points in $\mathbf{x}$ with $0 < ||u - x_i|| \leq r$.

For inhibitory pairwise interaction models it is known that pseudolikelihood estimation is a special case of the Takacs-Fiksel method when the interaction radius r is fixed [33, Section 2.4], [100, p. 54] or [109 Section 4]. The same is true for the area-interaction model.

Theorem 14 *For a stationary area-interaction process, the pseudolikelihood equations (3.4.25) and (3.4.26) are special cases of the Takacs-Fiksel method.*

Proof: Take f to be either of

$$\begin{aligned} f_\beta(\mathbf{x}) &= \beta^{-1}, \\ f_\gamma(\mathbf{x}) &= -\gamma^{-1} m(B(0,r) \setminus S_r(\mathbf{x} \setminus \{0\})). \end{aligned}$$

These are the partial derivatives of $\log \lambda(0;\mathbf{x})$ with respect to β and γ. When $f = f_\beta$ an unbiased estimator for the left hand side of (3.4.21) is

$$\frac{n}{m(A)} \frac{1}{\beta}$$

and, by stationarity, an unbiased estimator of the right hand side is

$$\frac{1}{m(A)} \int_A \frac{1}{\beta} \beta \gamma^{-t(u)} \, du.$$

When $f = f_\gamma$, the average over the observed events

$$\frac{n}{m(A)} \frac{1}{n} \sum_{i=1}^{n} \frac{-t(x_i)}{\gamma}$$

is an unbiased estimator of the left hand side of (3.4.26) by the Campbell-Mecke formula [111, p. 113], while an unbiased estimator for the right hand side is the window mean

$$\frac{1}{m(A)} \int_A \frac{-t(u)}{\gamma} \beta \gamma^{-t(u)} \, du.$$

These reduce to (3.4.25)–(3.4.26). □

Chapter 4
Bayesian object recognition

This Chapter develops a Bayesian approach to object recognition. We describe iterative optimisation schemes related to Besag's ICM [17] and discuss how spatial birth-and-death processes can be used for sampling and stochastic annealing [42]. The relative merits of the methods are investigated by means of a simple synthetic example.

4.1. General

A strong motivation for adopting a Bayesian approach to object recognition is that the MLE tends to contain clusters of almost identical objects, i.e. there is 'multiple response' to each true object, as noted in Section 2.6. This is related to the fact that the Hough transform has rather flat peaks around the correct object positions. Multiple response is undesirable if it is important to correctly determine the number of objects, and if it is believed that objects do not lie extremely close to one another. For instance in document reading it is known in advance that characters usually do not overlap. A standard approach in computer vision is to select one object per peak of the Hough transform; but this is very similar to a Bayesian approach using a prior model which assigns low probability to configurations in which objects are close to one another.

Natural prior models $p(\mathbf{x})$ belong to the class of nearest-neighbour Markov object processes described in Chapter 3. Given observation of image $\mathbf{y}$, the posterior probability density for scene $\mathbf{x}$ is then

$$p(\mathbf{x} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}). \tag{4.1.1}$$

The posterior distribution is often a nearest-neighbour object process too, though possibly with respect to a different neighbour relation. Specifically, for the overlapping objects relation (3.1.1) and a blur-free independent noise model with $g(\cdot \mid \cdot) > 0$, the posterior distribution is Markov with respect to (3.1.1) as well.

A *Maximum A Posteriori* (MAP) estimator of the true configuration solves

$$\begin{aligned} \tilde{\mathbf{x}} &= \operatorname{argmax}_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}) \\ &= \operatorname{argmax}_{\mathbf{x}} f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}). \end{aligned} \tag{4.1.2}$$

Again, the optimisation is over the space Ω of all object configurations. In decision theoretic terms, this corresponds to imposing a 0-1 loss function, according to whether the recognition is perfect or imperfect. Assuming $p(\cdot) > 0$ rewrite (4.1.2) as

$$\tilde{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} \left[\log f(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{x})\right]. \tag{4.1.3}$$

Because of this expression, $\tilde{\mathbf{x}}$ is also called a *penalised maximum likelihood estimator*. The factor $\log f(\mathbf{y} \mid \mathbf{x})$ is interpreted as a measure of goodness of fit to the data, and $-\log p(\mathbf{x})$ as a penalty for the complexity of the configuration $\mathbf{x}$. Typically, $p(\cdot)$ assigns low weight to configurations with many similar objects. Alternatively, P.J. Green noted that (4.1.3) can be interpreted as an optimisation of Euler-Lagrange type.

4.1.1 Prior models

A simple prior is the Strauss process (3.1.4) which results in a penalty of $-\log\beta$ for the presence of each object $x_i \in \mathbf{x}$ and a penalty of $-\log\gamma$ for each pair of neighbouring objects (e.g. overlapping objects). Modifications which might be useful in this application are

$$p(\mathbf{x}) = \alpha \prod_{i=1}^{n} \beta^{|R(x_i)|} \prod_{i<j} \gamma^{|R(x_i)\cap R(x_j)|} \tag{4.1.4}$$

and, for marked objects, to allow the interaction terms to depend on the marks.

Regarding the choice of parameter values, note that if the raster is made finer (say, quadrupling the number of pixels) then the log likelihood typically increases by the same factor. This suggests that to maintain the balance between f and p in (4.1.2)-(4.1.3) the parameters $\log\beta$ and $\log\gamma$ of a Strauss model should also be multiplied by this factor. Models such as (4.1.4) with interactions expressed in terms of pixel counts, do not require such adjustment.

Since the area-interaction model (3.1.8)

$$p(\mathbf{x}) = \alpha\beta^{n(\mathbf{x})}\gamma^{-|S(\mathbf{x})|}$$

is also defined through pixel counts, the same remark holds. To enforce inhibition, choose $\gamma < 1$. Then, with the number of objects fixed, the silhouette area $|S(\mathbf{x})|$ is maximal if there is as little overlap as possible.

4.2. ITERATIVE ALGORITHMS

Iterative methods are needed in order to find the MAP estimator (4.1.2). As in Section 2.4 we consider algorithms which add or delete objects one at a time. The log likelihood ratio criterion is now replaced by a log posterior probability ratio.

Algorithm 5 *Apply Algorithms 1, 2, 3 or 4 with* $f(\mathbf{y}|\mathbf{x})$ *replaced by the posterior probability* $p(\mathbf{x}|\mathbf{y})$. *Thus we iteratively add object* u *to list* $\mathbf{x}$ *iff*

$$\log \frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} \geq w,$$

delete existing object x_i *iff*

$$\log \frac{f(\mathbf{y} \mid \mathbf{x} \setminus \{x_i\})\, p(\mathbf{x} \setminus \{x_i\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} \geq w$$

and if shifts are permitted, we shift $x_i \in \mathbf{x}$ *to* u *iff* $u \in Q(\mathbf{x}, x_i)$ *and*

$$\log \frac{f(\mathbf{y} \mid M(\mathbf{x}, x_i, u))\, p(M(\mathbf{x}, x_i, u))}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} \geq w$$

where $w \geq 0$ *is a chosen threshold.*

The convergence properties of Algorithms 1–4 remain valid for this new objective function. An alternative description of Algorithm 5 is that the static threshold value used in the likelihood ratio algorithms is replaced by one that depends on the current reconstruction and on a smoothing parameter.

Algorithm 5 is a close analogue of Besag's ICM algorithm [17]. If the object space U is finite, consider labelling each $u_j \in U$ with value

$$v_j = \begin{cases} 1 & \text{if } u_j \in \mathbf{x} \\ 0 & \text{else} \end{cases} .$$

The ICM approach would be to visit each object sequentially and update its label in the light of current estimates for the other objects: when the labelling is $\hat{v}$, the label of the current object u_j is updated to

$$\text{argmax } \mathbb{P}(v_j = k \mid \mathbf{y}, (\hat{v}_i)_{i \neq j}).$$

But this is clearly equivalent to comparing

$$f(\mathbf{y} \mid \hat{\mathbf{x}} \setminus \{u_j\})\, p(\hat{\mathbf{x}} \setminus \{u_j\})$$

against

$$f(\mathbf{y} \mid (\hat{\mathbf{x}} \setminus \{u_j\}) \cup \{u_j\})\, p((\hat{\mathbf{x}} \setminus \{u_j\}) \cup \{u_j\})$$

where $\hat{\mathbf{x}}$ is the current object list; and this is Algorithm 5 ($w = 0$).

Algorithm 2 is a simple variant of ICM, at least in the discrete case. This algorithm is also defined when U is 'continuous' (any l.s.c. space) but the interpretation is more complex. We add a new object u at the position where the Papangelou conditional intensity of the posterior distribution, given the current configuration $\mathbf{x}$ on $U \setminus \{u\}$, is maximal, provided this is greater than e^w (relative to the reference measure μ).

4.2.1 Examples

Figure 4.1 shows the result of steepest ascent ICM for the pellet texture of Section 2.6. The prior was a Strauss process of overlapping discs (3.1.4) with $\log \beta = \log \gamma = -1000$.

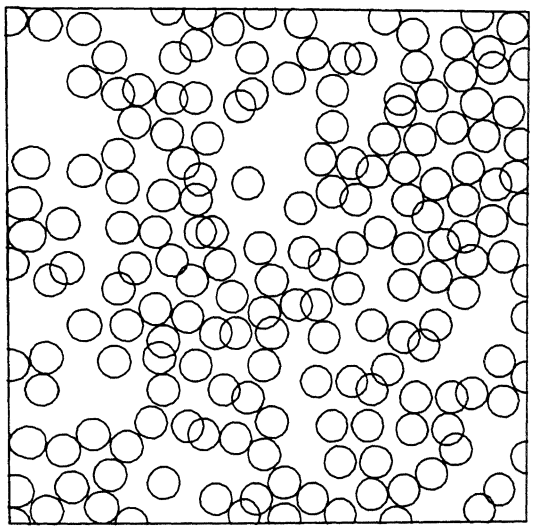


Figure 4.1: MAP reconstruction of the Brodatz pellets texture by steepest ascent from an empty initial state ($w = 0$) using a Strauss prior with $log\beta = \log \gamma = -1000$.

As another example, consider a noise corrupted image of a scene consisting of rectangles of various sizes (Figure 4.2). The object space is four-dimensional; any rectangle can be described by the coordinates of its upper left and lower right corner. The forward model is blur-free Gaussian white noise. Because the object size is variable, MLE solutions cannot distinguish between one large rectangle and a union of several small ones having the same silhouette. Therefore we introduce a penalty $-\log \beta = 5.0$ on addition of an object and obtain a (non-unique)

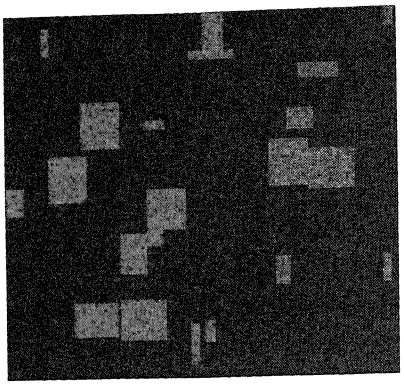


Figure 4.2: Realisation from a Gaussian blur-free independent noise model with $\sigma = 10$, $\theta_1 = 150$ and $\theta_0 = 100$, digitised on a 128×116 grid.

steepest ascent reconstruction given in Figure 4.3 (top). Fine tuning is possible by allowing the corners to move one pixel in every direction. The result is shown in Figure 4.3 (middle).

Note that the posterior distribution cannot distinguish between an equal number of objects having the same silhouette. If it is desirable to have as many overlap as possible, a penalty $-\log\gamma = 5.0$ on every pair of overlapping objects can be introduced, resulting in Figure 4.3 (bottom), or for every pixel in the intersection (resolving uncertainty in the two objects in the top centre of the image).

4.2.2 Relation to Hough transform

In many cases, computing the posterior log likelihood ratio is a local operation, related to the Hough transform. For example, taking the Strauss prior (3.1.4), $g(\cdot|\cdot) > 0$ and a blur-free signal,

$$\log \frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} = \log\beta + r(u;\mathbf{x}) \log\gamma + \sum_{R(u)\setminus S(\mathbf{x})} z_t$$

where $r(u;\mathbf{x}) = s(\mathbf{x} \cup \{u\}) - s(\mathbf{x})$ is the number of neighbours of u in $\mathbf{x}$ and $z_t = \log g(y_t|\theta_1) - g(y_t|\theta_0)$ (cf. Section 2.5). The new term in γ is a penalty against adding an object in the vicinity of existing objects. For the area-interaction prior (3.1.8) the term involving γ is replaced by $-|R(u) \setminus S(\mathbf{x})| \log\gamma$ so that we again obtain something very similar to the Hough transform.

In general, for any independent noise model (2.2.3) with $g(\cdot|\cdot) > 0$, and any nearest neighbour Markov object prior (Definition 3), the posterior log likelihood ratio depends only on data pixels in the update

zone $Z(\mathbf{x}, u)$ and on the configuration in the restricted neighbourhood $\mathcal{N}(u \mid \mathbf{x} \cup \{u\})$ of the added object.

4.2.3 Parameter estimation

Until now we assumed that the model parameters were known exactly. However, in realistic situations unknown physical variables have to be taken into account. In some applications where many similar images are available, these can be used for parameter estimation. In the absence of training data, estimation and recognition of objects must be carried out simultaneously.

Maximum likelihood can be applied to obtain estimates. This is feasible for many noise models; in particular for independent noise models the method is straightforward. The prior distribution, however, involves a normalising constant which cannot be evaluated, making maximum likelihood estimation intractable. An alternative, more efficient method is maximum pseudolikelihood estimation [15, 16, 58], which maximises the product of conditional densities. This product does not involve the normalising constant and is easy to compute for Markov processes. However, it is not a genuine likelihood, except in the case of spatially independent objects. (See also Section 3.4.5).

For the superficially similar case of image segmentation using Markov random fields, Besag [17] proposed the following procedure. Unknown parameters in f and p are denoted by ϕ and ψ respectively.

Algorithm 6

1. *obtain an initial estimate* $\hat{\mathbf{x}}$ *of the true pattern, with guesses for* ϕ *and* ψ *if necessary;*

2. *estimate* ϕ *by maximising* $f(\mathbf{y}|\hat{\mathbf{x}}; \phi)$*; optionally, estimate* ψ *by the maximum pseudolikelihood method;*

3. *carry out a single step of the reconstruction algorithm based on the current estimates* $\hat{\mathbf{x}}$, $\hat{\phi}$ *and* $\hat{\psi}$*, leading to a new estimate* $\hat{\mathbf{x}}$*, and return to step 2.*

An alternative, fully Bayesian approach is to specify a prior distribution for each model parameter and replace step 2 above by sampling from the posterior distributions.

As an illustration, Figure 4.4 displays a digitised piece of music, reproduced with kind permission of R. van den Boomgaard. The data was scanned by Peter Tax (University of Amsterdam). Here we will focus on finding the locations of notes. To a good approximation their shape is elliptical, the major axis making a 30 degree angle with the horizontal

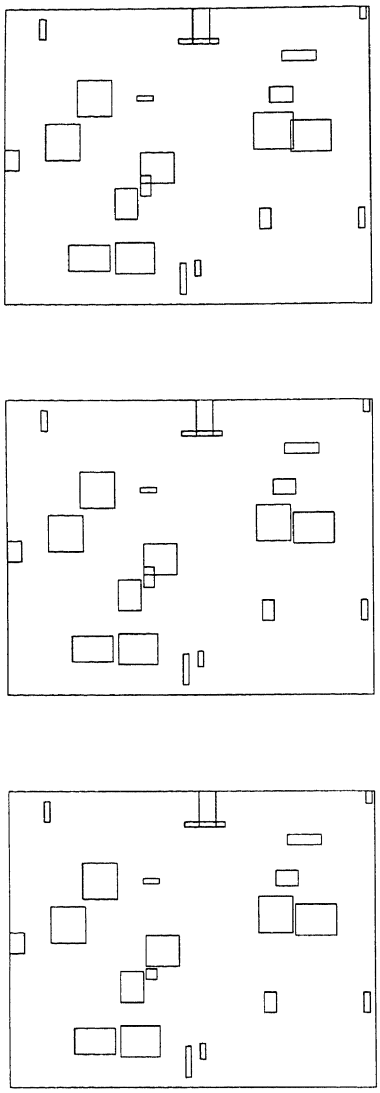


Figure 4.3: Steepest ascent reconstructions for the data in Figure 4.2 at threshold $w = 0$ using a Strauss prior. The initial state is empty. Top: only births and deaths, $\log \beta = -5$ and $\log \gamma = 0$. Middle: translations over 1 pixel in every direction, $\log \beta = -5$ and $\log \gamma = 0$. Bottom: translations over 1 pixel in every direction, $\log \beta = -5$ and $\log \gamma = -5$.

coordinate axis. As the shape is assumed to be constant, the parameter space and image space coincide, i.e. this is a translation model.

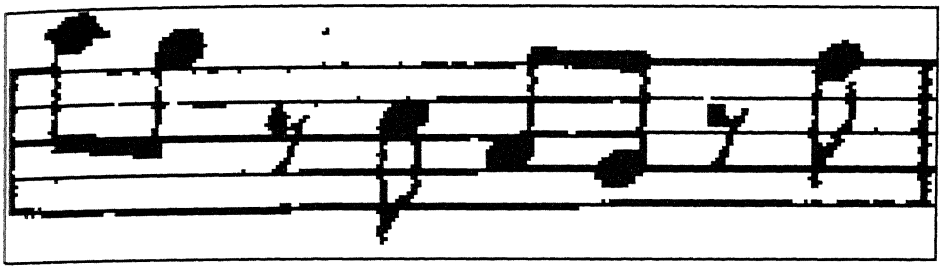


Figure 4.4: Binary image of a small piece of music digitised on a 250×65 grid.

For the noise distribution a salt-and-pepper model was chosen, with unknown error fraction p. As it is known in advance that notes cannot overlap each other, a suitable choice for the prior distribution is a hard core object process (3.1.2).

A steepest ascent reconstruction, starting from an empty list, for the data in Figure 4.4 is given in Figure 4.5. Here, $w = 0$ and $\log \beta = -50$. The noise error fraction p is estimated during iteration by its maximum likelihood estimator

$$\hat{p} = \frac{|S(\mathbf{x}) \Delta Y|}{|T|}.$$

Recall that Y is the set of pixels with value 1 and Δ denotes the symmetric set difference. To facilitate interpretation the data image was masked by the reconstruction and vice versa.

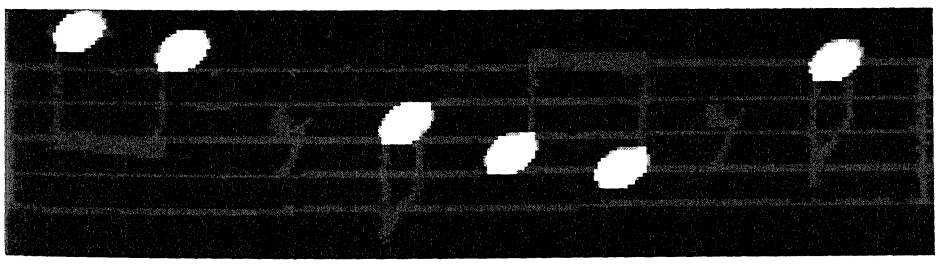


Figure 4.5: Steepest ascent reconstruction (white) 'masked' by the data. The initial configuration is empty, $w = 0$ and the prior distribution is hard core with $\log \beta = -50$.

The scene contains not only notes, but also other features such as horizontal and vertical lines, thick bars etc. However, their contribution to

the Hough transform is small in comparison to real notes, due to the fact that negative pixel votes serve to compensate the positive contributions. In other words, the objects $R(u)$ are large enough to accumulate strong evidence in the Hough transform and to suppress noise. By setting the penalty $-\log\beta$ for addition of a new object high enough, only desired objects are detected.

It is interesting to note that the parts in the image are strongly connected. Vertical line segments are usually linked with the basic notes (except to separate measures) and might be joined by bars. Once the basic ellipses have been found, one could add an extra stage to look for line segments and then bars, thus building a hierarchical method for automatic music reading. We do not pursue this issue further.

4.3. PERFORMANCE EVALUATION

In this Section we study in detail a simple synthetic example in which a pattern of discs of fixed radius (parameterised by the centre coordinates) has been observed after addition of Gaussian noise (Figure 4.6). As the true scene is known, it is possible to compare algorithms by computing numeric performance measures.

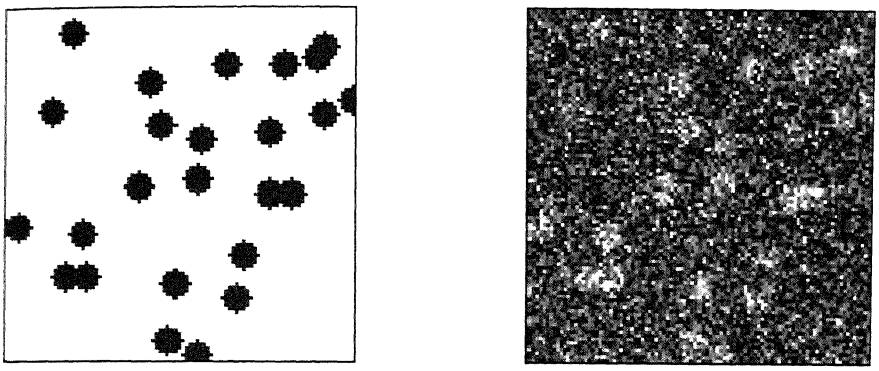


Figure 4.6: Binary silhouette scene of discs with radius 4 and realisation from Gaussian model with $\sigma = 50$, $\theta_1 = 150$ and $\theta_0 = 100$ digitised on a 98×98 square grid.

4.3.1 Reconstructions

Figure 4.7 shows maximum likelihood reconstructions obtained by the coordinatewise optimisation algorithm (Algorithms 1 and 3) taking threshold value $w = 0$. The pixels were scanned in row major order and for the initial state we took the local maxima of

$$\log \frac{f(\mathbf{y} \mid \{u\})}{f(\mathbf{y} \mid \emptyset)}$$

where the expression was non-negative. Furthermore, Algorithm 3 allowed translations over one pixel in every direction. Note the multiple response, especially where discs overlap each other.

The reconstructions obtained using steepest ascent (Algorithms 2 and 4) with empty initial state and threshold value $w = 0$ are given in figure 4.8. In this case it seems important to stop short of convergence. Typically, when all the objects that are really present have been detected, the method keeps adding spurious ones. This can be counteracted by taking a higher threshold value. In the present example choosing $w = 6$ will result in the 'best' reconstruction shown in Figure 4.9.

Algorithm 5 is illustrated in Figure 4.10 for coordinatewise optimisation. Again, the pixels were scanned in row major order choosing the

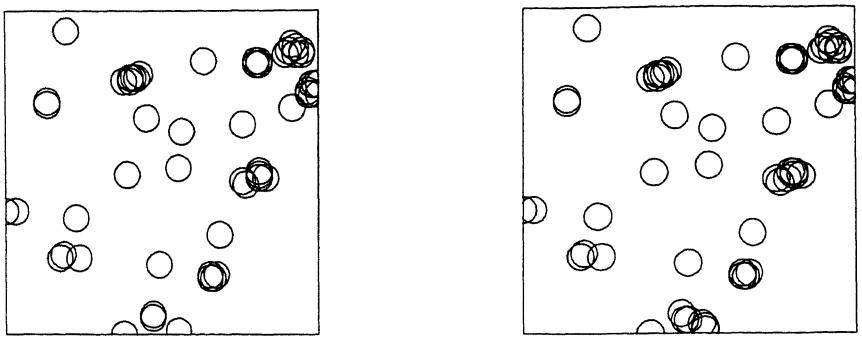


Figure 4.7: MLE reconstructions using coordinatewise ascent with the local extrema of the Hough transform as initial state. Left: only births and deaths; right: births, deaths and translation.

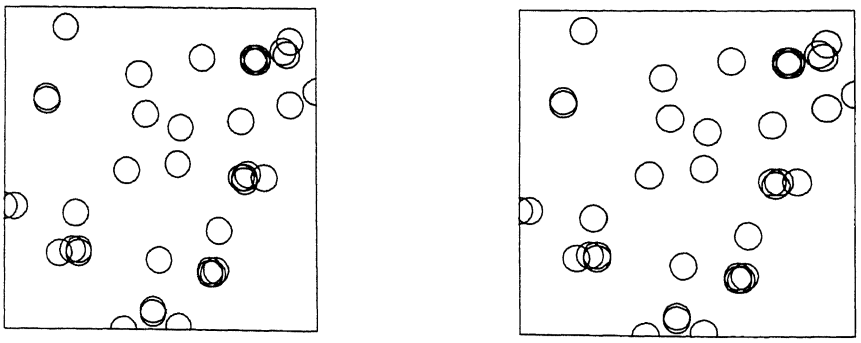


Figure 4.8: MLE reconstructions using steepest ascent with the empty list as initial state. Left: only births and deaths; right: births, deaths and translation.

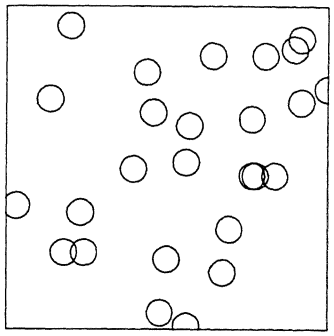


Figure 4.9: 'Best' intermediate MLE reconstruction using steepest ascent with the empty list as initial state (both Algorithms 2 and 4).

local extrema of the Hough transform as initial configuration. A Strauss prior model was used with parameters $\beta = .0025$ and $\gamma = .25$. In the steepest ascent case fewer spurious discs are added. Indeed the final result is the 'best' one for this particular example (Figure 4.11). Overall, MAP gives a clear improvement over MLE, successfully combating multiple response.

4.3.2 Typical behaviour

In the discussion below, we define one 'step' of each algorithm as a complete scan through the discretised parameter space. The number of 'transitions' (additions or deletions of objects) per scan may vary: steep-

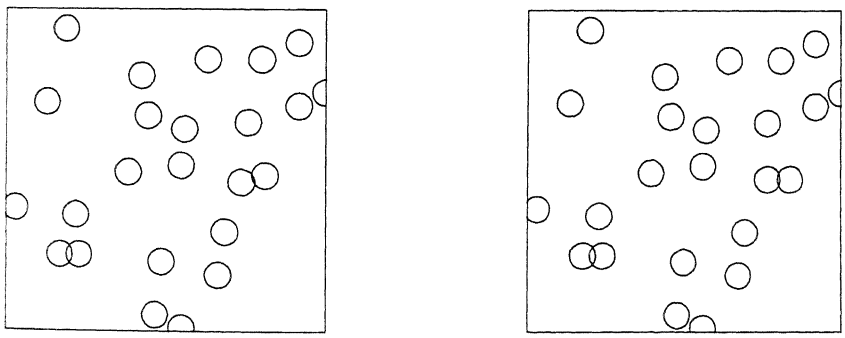


Figure 4.10: MAP reconstructions using coordinatewise ascent with the local extrema of the Hough transform as initial state. The prior distribution is a Strauss model with $\beta = .0025$ and $\gamma = .25$. Left: only births and deaths; right: births, deaths and translations.

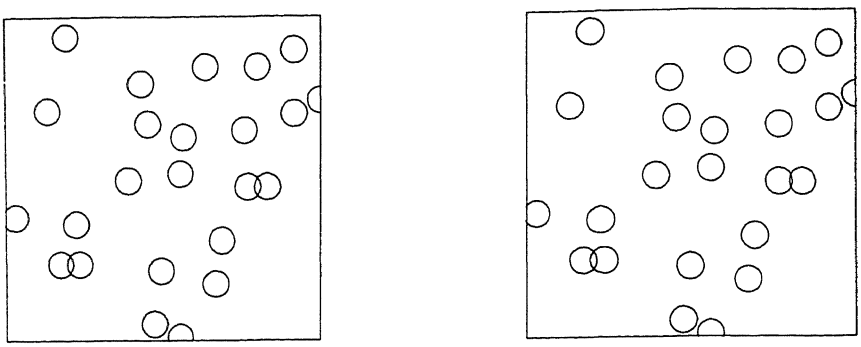


Figure 4.11: MAP reconstructions using steepest ascent with the empty list as initial state. The prior distribution is a Strauss model with $\beta = .0025$ and $\gamma = .25$. Left: only births and deaths; right: births, deaths and translations.

est ascent (and sampling techniques) generate one transition per scan, while coordinatewise optimisation yields a variable number depending on the data, the current reconstruction and the scanning order. 'Steps' are roughly proportional to total computer time, although this ignores the possibility of parallelism: steepest ascent could be implemented in parallel computation, but coordinatewise optimisation is inherently sequential.

The performance of an algorithm is measured by computing the log likelihood
$L(\mathbf{x}^{(k)} \; ; \; \mathbf{y})$ itself and two 'external' measures of fidelity, Pratt's figure of merit [1]

$$\frac{1}{\max\{n(\mathbf{x}), n(\mathbf{x}^{(k)})\}} \sum_{x_i' \in \mathbf{x}^{(k)}} \frac{1}{1 + \frac{1}{9} d(x_i', \mathbf{x})}$$

and Baddeley's Δ_2 metric [2]

$$\left\{ \frac{1}{n(T)} \sum_{t \in T} \left(d(t, \mathbf{x}) - d(t, \mathbf{x}^{(k)}) \right)^2 \right\}^{1/2}.$$

With a slight abuse of notation, the same symbol d is used to denote the distance from either a lattice or a configuration member to a given pattern. It can be justified by noting that in this particular example image space and object space coincide.

We computed the Δ_2 distance and the figure of merit for the reconstruction $\mathbf{x}^{(k)}$ at each iteration. Figure .1 graphs the performance of coordinatewise optimisation starting with the local extrema of the Hough

transform. Pixels were scanned in row major order. Again MAP shows a clear improvement over the MLE in terms of the external performance measures. Note that for Bayesian algorithms the log likelihood graph need not be monotone, as these methods aim at increasing the *posterior* likelihood.

For the steepest ascent algorithms the behaviour is qualitatively different, see Figure .2. While coordinatewise optimisation needs only a few scans through the image, steepest ascent from an empty initial image requires at least as many scans as there are objects in the image. New objects are added one-by-one, gradually improving the reconstruction quality, until all objects are detected; then the reconstructions deteriorate. This method can however yield more accurate reconstructions than coordinatewise optimisation algorithms, especially in the non-Bayesian case.

A widely used technique for object recognition is to compute the Hough transform and find its local extrema. In noisy images this procedure performs badly, as can be seen from the graphs in Figure .1 where the y intercept is the performance of the Hough extrema operator.

4.3.3 Initial state influence

Tables 1–12 show performance measures for the eight possible algorithms using various initial states.

First restrict attention to the methods based only upon addition and deletion. Using initial configurations other than the empty list does not change the overall pattern described above. Coordinatewise optimisation converges in a few steps; steepest ascent slowly improves the reconstruction by removing incorrect objects and replacing them by the right ones. MAP shows less sensitivity to the initial state than does MLE.

Not surprisingly, the best reconstructions were obtained when initialising with the true image, but a perfect match is not guaranteed. This reflects the fact that the truth is not necessarily a solution of the MAP equations.

One proposal is to use the local extrema of the Hough transform as the initial state. This appears sensible for the coordinatewise optimisation algorithms but not for steepest ascent. In most cases better reconstructions could be obtained with an empty image as initial state.

The same remarks hold if the initial state is a translation of the true state. A possible explanation is that extra effort is required to throw away incorrect estimates and replace them with better ones.

In summary, for the add-delete algorithms, it is best as a rule of thumb to use the empty list as a starting state for the steepest ascent

algorithms, unless additional information about the objects to be detected is available. This is especially so since time is wasted in throwing away incorrect objects (see the remarks in §7.2). Another advantage is that no preprocessing (e.g. computing the Hough transform) of the data is required.

Turning now to the add-delete-shift algorithms, we note that these exhibit less sensitivity to the choice of initial state. Another advantage is that the number of scans needed for steepest ascent decreases, as it is no longer necessary to throw away spurious objects first and then replace them by better ones. This is illustrated in Figure .2 where it must be noted that the plot for Algorithm 2 is cut off, as 71 transitions were needed. The graphs for the refined algorithms are smoother than those for the techniques based upon addition and deletion only, and the levels at convergence are better. The Bayesian method with only births and deaths is unable to reposition many discs, as the posterior log likelihood ratios are too small. When using Algorithm 2 more discs are thrown away and readded at their proper positions, but many more scans through the image are required.

4.3.4 Noise influence

It is also of interest to investigate the influence of signal-to-noise ratio. We generated ten independent realisations of Model 1 for several values of σ^2. Reconstructions were obtained and the average quality calculated for Algorithms 1–5. The results are depicted in Figure .3.

As could be expected, the reconstructions become poorer when more noise is added. Due to the extra term penalising undesirable configurations, the MAP solutions are less sensitive to the noise variance than ML estimates. Steepest ascent is more robust than coordinatewise optimisation.

4.4. FIXED TEMPERATURE SAMPLING

The natural analogue of the Gibbs sampler in this context is a *spatial birth-and-death process* [97, 83]. This is a continuous time, pure jump Markov process, whose states are configurations $\mathbf{x} \in \Omega$, and for which the only transitions are the birth of a new object (instantaneous transition from $\mathbf{x}$ to $\mathbf{x} \cup \{u\}$) or the death of an existing one (transition from $\mathbf{x}$ to $\mathbf{x} \setminus \{x_i\}$). Formally, write $\mathcal{B}$ for the (Borel) σ-algebra on U and let $D(\cdot,\cdot) : \Omega \times U \to [0,\infty)$ be a measurable function and $B(\cdot,\cdot) : \Omega \times \mathcal{B} \to [0,\infty)$ a finite kernel, i.e. $B(\mathbf{x},\cdot)$ is a finite measure on $(U,\mathcal{B})$ and $B(\cdot,F)$ is a measurable function on Ω. These are called the *death rate* and *birth rate*. The reason is clear from the following. Given the state $\mathbf{x}$ at time t,

- the probability of a death $\mathbf{x} \to \mathbf{x} \setminus \{x_i\}$ during a time interval $(t, t+h)$, $h \to 0$, is $D(\mathbf{x} \setminus \{x_i\}, x_i)h + o(h)$;
- the probability of a birth $\mathbf{x} \to \mathbf{x} \cup \{u\}$ during time $(t, t+h)$, where u lies in a given measurable subset $F \subseteq U$, is $B(\mathbf{x}, F)h + o(h)$;
- the probability of more than one transition during $(t, t+h)$ is $o(h)$.

We will assume that $B(\mathbf{x},\cdot)$ has a density $b(\mathbf{x},\cdot)$ with respect to μ on U, so that intuitively $b(\mathbf{x}, u)$ is the transition rate for a birth $\mathbf{x} \to \mathbf{x} \cup \{u\}$. Write

$$B(\mathbf{x}) = \int_U b(\mathbf{x}, u)\, d\mu(u)$$

for the total birth rate, and similarly define

$$D(\mathbf{x}) = \sum_{x_i \in \mathbf{x}} D(\mathbf{x} \setminus \{x_i\}, x_i).$$

To avoid explosion, i.e. an infinite number of transitions occurring in finite time, the rates have to satisfy certain assumptions. Preston [97, Prop. 5.1, Thm. 7.1] gave sufficient conditions under which there exists a unique spatial birth-and-death process with given rates solving Kolmogorov's backward equations

$$\begin{aligned} \frac{d}{dt}\mathbb{P}(X_t \in A | X_0 = \mathbf{x}) &= -[B(\mathbf{x}) + D(\mathbf{x})]\, \mathbb{P}(X_t \in A | X_0 = \mathbf{x}) \\ &\quad + \int_\Omega \mathbb{P}(X_t \in A | X_0 = \mathbf{z}) R(\mathbf{x}, d\mathbf{z}) \end{aligned}$$

with $R(\mathbf{x}, A) = B(\mathbf{x}, \{u \in U : \mathbf{x} \cup \{u\} \in A\}) + \sum_{x_i \in \mathbf{x}} 1\{\mathbf{x} \setminus \{x_i\} \in A\}$ $D(\mathbf{x} \setminus \{x_i\}, x_i)$ the total rate from pattern $\mathbf{x}$ into a measurable subset $A \subseteq \Omega$. For a given process (X_t) he also found conditions for the existence of a unique invariant probability measure and convergence in distribution (i.e. convergence of $\mathbb{P}(X_t \in F | X_0 = \mathbf{x})$).

Theorem 15 *For each* $n = 0, 1, \ldots$ *define* $\kappa_n = \sup_{n(\mathbf{x})=n} B(\mathbf{x})$ *and* $\delta_n = \inf_{n(\mathbf{x})=n} D(\mathbf{x})$. *Assume* $\delta_n > 0$ *for all* $n \geq 1$. *If* either (a) $\kappa_n = 0$ *for all sufficiently large* $n \geq 0$, or (b) $\kappa_n > 0$ *for all* $n \geq 0$ *and both the following hold:*

$$\sum_{n=1}^{\infty} \frac{\kappa_0 \cdots \kappa_{n-1}}{\delta_1 \cdots \delta_n} < \infty$$

$$\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\kappa_1 \cdots \kappa_n} = \infty$$

then there exists a unique spatial birth-and-death process for which $B(\cdot)$ *and* $D(\cdot)$ *are the transition rates; this process has a unique equilibrium distribution to which it converges in distribution from any initial state.*

A slightly stronger result given by Møller [83] includes the case $\kappa_0 = 0$, $\kappa_n > 0$ for all $n \geq 1$ and both $\sum_{n=2}^{\infty} \frac{\kappa_1 \cdots \kappa_{n-1}}{\delta_1 \cdots \delta_n} < \infty$ and $\sum_{n=1}^{\infty} \frac{\delta_1 \cdots \delta_n}{\kappa_1 \cdots \kappa_n} = \infty$, still assuming all δ_n positive for $n \geq 1$.

4.4.1 Construction

Suppose we want to sample from the *temperature modified posterior distribution*

$$p_H(\mathbf{x} \mid \mathbf{y}) \propto \{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\}^{1/H}. \tag{4.4.5}$$

The purpose of introducing a temperature parameter is to sharpen peaks in posterior probability. For small $H > 0$, configurations with large posterior density are favoured, while others are suppressed. Indeed, if object space U is discretised $p_H(\cdot \mid \mathbf{y})$ converges pointwise to a uniform distribution on the set of MAP solutions as H tends to zero.

Consider any blur-free independent noise model (2.2.3) with $g(\cdot|\cdot) > 0$ and a nearest-neighbour Markov object prior. The former assumption is needed so that the class $K = \{\mathbf{x} : f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) > 0\}$ is hereditary. For some fixed $k \in [0, 1]$ set

$$b_H(\mathbf{x}, u) = \begin{cases} \left(\frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}\right)^{k/H} & \text{if } \mathbf{x} \in K \\ 0 & \text{if } \mathbf{x} \notin K \end{cases} \tag{4.4.6}$$

for $u \notin \mathbf{x}$ and death rate

$$D_H(\mathbf{x} \setminus \{x_i\}, x_i) = \begin{cases} \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x} \setminus \{x_i\})\, p(\mathbf{x} \setminus \{x_i\})}\right)^{\frac{k-1}{H}} & \text{if } \mathbf{x} \in K \\ \delta_n'(\mathbf{x})/n(\mathbf{x}) & \text{if } \mathbf{x} \notin K \end{cases} \tag{4.4.7}$$

Here $\delta_n' = \inf \{\sum_{x_i \in \mathbf{x}} D_H(\mathbf{x} \setminus \{x_i\}, x_i) \mid f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) > 0,\ n(\mathbf{x}) = n\}$. By convention, the infimum of the empty set equals ∞. Note that by this

definition $\delta'_n = \delta_n$, where δ_n is defined as in Theorem 15. The boundary cases $k = 0$ ('constant birth rate') and $k = 1$ ('constant death rate') are well-known in spatial statistics, to obtain realisations of a point process. It is widely argued (e.g. [98]) that the constant death rate procedure should be preferred, as under the constant birth rate process there is a high probability that a newly-added object will have a large death rate and thus be rapidly deleted again.

For a nearest-neighbour Markov prior the above expressions are typically easy to evaluate, since the normalising constant α is eliminated. Moreover the 'detailed balance' equations

$$b_H(\mathbf{x}, u)\, p_H(\mathbf{x} \mid \mathbf{y}) = D_H(\mathbf{x}, u)\, p_H(\mathbf{x} \cup \{u\} \mid \mathbf{y}) \tag{4.4.8}$$

are satisfied whenever $f(\mathbf{y} \mid \mathbf{x} \cup \{u\})\, p(\mathbf{x} \cup \{u\}) > 0$. Intuitively, under the equilibrium distribution $p_H(\cdot \mid \mathbf{y}$ transitions from $\mathbf{x}$ into $\mathbf{x} \cup \{u\}$ are exactly matched by transitions from $\mathbf{x} \cup \{u\}$ into $\mathbf{x}$. Given a spatial birth-and-death process with rates satisfying (4.4.8), Ripley [98] remarked that the process is necessarily time reversible and $p_H(\cdot|\mathbf{y})$ is the density of its unique invariant probability measure. For each application, however, one should verify that the process just described is well-defined. For instance the following corollary of Theorem 15 holds.

Corollary 16 *Let* $\mathbf{y}$ *and* $H > 0$ *be fixed. For any blur-free independent noise model with* $g(\cdot|\cdot) > 0$, *and any nearest-neighbour Markov object process* $p(\cdot)$ *with uniformly bounded likelihood ratios*

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \leq \beta < \infty$$

there exists a unique spatial birth-and-death process for which (4.4.6) and (4.4.7) are the transition rates. The process has unique equilibrium distribution $p_H(\cdot \mid \mathbf{y})$ *and it converges in distribution to* $p_H(\cdot \mid \mathbf{y})$ *from any initial state.*

Proof: We will prove the following properties:

1. $\delta_n > 0$, for $n \geq 1$;
2. if $\kappa_{n_0} = 0$ for some $n_0 \geq 1$, then $\kappa_n = 0\ \forall n \geq n_0$;
3. if $\kappa_n > 0$ for all n, then condition (b) of Theorem 15 holds.

Property 1: Use the representation of the log likelihood ratio as a generalised Hough transform (2.1.1). Since T is finite, we have upper and lower bounds on the goodness of fit, say $\mid h(y_t, \theta_0, \theta_1) \mid \leq a$ for all t. Hence

$$| L(\mathbf{x} \cup \{u\} ; \mathbf{y}) - L(\mathbf{x} ; \mathbf{y}) | \leq a\, n(R(u)) \leq a\, n(T)$$

where n denotes the number of pixels. For $p(\cdot)$ we have by assumption $\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \leq \beta$. If $p(\mathbf{x}) > 0$ this implies that

$$D_H(\mathbf{x} \setminus \{u\}, u) = \left(\frac{f(\mathbf{y} \mid \mathbf{x})}{f(\mathbf{y} \mid \mathbf{x} \setminus \{u\})} \frac{p(\mathbf{x})}{p(\mathbf{x} \setminus \{u\})} \right)^{\frac{k-1}{H}}$$

$$\geq \exp\left[\frac{k-1}{H} (| L(\mathbf{x} \setminus \{u\} ; \mathbf{y}) - L(\mathbf{x} ; \mathbf{y}) | + \log \beta) \right]$$

$$\geq \exp\left[\frac{k-1}{H} (a\, n(T) + \log \beta) \right] =: \delta > 0.$$

Suppose $p(\mathbf{x}) = 0$. If $p(\mathbf{z}) = 0$ for all $\mathbf{z}$ with $n(\mathbf{z}) = n(\mathbf{x})$, then $D_H(\mathbf{x} \setminus \{u\}, u) = \infty \geq \delta$. Otherwise $n(\mathbf{x})\ D_H(\mathbf{x} \setminus \{u\}, u) = \inf \{D_H(\mathbf{z}) \mid n(\mathbf{z}) = n(\mathbf{x}),\ p(\mathbf{z}) > 0\}$. By the above argument, $D_H(\mathbf{z} \setminus \{z_i\}, z_i) \geq \delta$ for all such $\mathbf{z}$ and $z_i \in \mathbf{z}$. Hence $D_H(\mathbf{z}) \geq \delta\ n(\mathbf{x})$ and $D_H(\mathbf{x} \setminus \{u\}, u) \geq \delta$. Therefore $D_H(\mathbf{x}) \geq \delta\ n(\mathbf{x})$ for all patterns $\mathbf{x}$, and hence $\delta_n \geq \delta\ n > 0$ for $n \geq 1$.

Property 2: Use the fact that $K = \{\mathbf{x} \ : \ f(\mathbf{y} \mid \mathbf{x})\ p(\mathbf{x}) > 0\}$ is hereditary.

Property 3: The birth rates are also bounded. For $p(\mathbf{x}) > 0$

$$\begin{aligned} b_H(\mathbf{x}, u) &= \left(\frac{f(\mathbf{y} \mid \mathbf{x} \cup \{u\})}{f(\mathbf{y} \mid \mathbf{x})} \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \right)^{k/H} \\ &\leq \exp\left[\frac{k}{H} (a\, n(T) + \log \beta) \right] \\ &=: \kappa > 0. \end{aligned}$$

For $p(\mathbf{x}) = 0$, again $b_H(\mathbf{x}, u) = 0 \leq \kappa$. Hence $B_H(\mathbf{x}) \leq \kappa\ \mu(U)$ and so $\kappa_n \leq \kappa\ \mu(U)$.

Using these bounds, one obtains

$$\frac{\kappa_0 \cdots \kappa_{n-1}}{\delta_1 \cdots \delta_n} \leq \frac{\kappa^n\ \mu(U)^n}{n!\ \delta^n}.$$

Since $\mu(U)$ is finite by assumption, the first assertion follows. Similarly

$$\frac{\delta_1 \cdots \delta_n}{\kappa_1 \cdots \kappa_n} \geq \frac{n!\ \delta^n}{\kappa^n\ \mu(U)^n}$$

which does not converge to zero as $n \to \infty$. The Corollary is proved if we combine these properties with Theorem 15. □

If the state space is discretised, the situation is easier. Recall that any Markov chain on a finite state space is uniquely defined by its transition rates and if for an irreducible Markov chain with rates $R(i,j)$, $i \neq j$, π is a probability measure satisfying the detailed balance equations

$$R(i,j)\ \pi(i) = R(j,i)\ \pi(j) \tag{4.4.9}$$

then the chain is time reversible and has unique limit distribution π (see for instance [93, pp. 277–278]). Hence, if $f(\mathbf{y} \mid \mathbf{x}) > 0$ for all $\mathbf{x}$ the class K is hereditary, due to the Markov property of $p(\cdot)$. It follows that the birth-and-death process defined above restricted to K is irreducible.

To simulate the birth-and-death process we generate the successive states $X^{(k)}$ and the sojourn times $T^{(k)}$ as follows. Given $X^{(k)} = \mathbf{x}^{(k)}$, $T^{(k)}$ is exponentially distributed with mean $1/(D_H(\mathbf{x}^{(k)}) + B_H(\mathbf{x}^{(k)}))$, independent of other sojourn times and of past states. The next state transition is a death with probability $D_H(\mathbf{x}^{(k)})/(D_H(\mathbf{x}^{(k)}) + B_H(\mathbf{x}^{(k)}))$, obtained by deleting one of the existing points x_i with probability

$$\frac{D_H(\mathbf{x}^{(k)} \setminus \{x_i\}, x_i)}{D_H(\mathbf{x}^{(k)})}.$$

Otherwise the transition is a birth generated by choosing one of the points $u \notin \mathbf{x}^{(k)}$ with probability density

$$\frac{b_H(\mathbf{x}^{(k)}, u)}{B_H(\mathbf{x}^{(k)})}$$

with respect to μ and adding u to the state.

Algorithm 7 (Fixed temperature sampling) *Run the process described above for a 'large' time period C and take $\mathbf{x}^{(L)}$ where*

$$L = \min\{k = 0, 1, 2, \ldots \mid \sum_{i=0}^{k} t^{(i)} > C\}.$$

For a discussion of the rate of convergence see [83].

Figure 4.12 shows a single realisation from the posterior distribution for the pellets texture of Section 2.6 using the same Strauss prior as for Figure 4.1. The realisation was obtained by running the constant death rate process described above Algorihtm 7.

The main advantage of sampling techniques compared to deterministic methods is the ability to estimate any functional of the (modified) posterior distribution by taking a sufficient number of independent realisations. Examples of useful functionals are: the distribution (mean,

variance) of the number of objects; the probability that there is no object in a given subregion of the image; the distribution of the distance from a given reference point to the nearest object and the first-order intensity [111]. In the discrete case the first-order intensity at u is simply the (posterior) probability that u belongs to $\mathbf{x}$. It can be regarded as an alternative to the Hough transform. The posterior intensity is given in Figure 4.13.

Clearly, other sampling techniques [44] could be used as well. However, at small temperatures, those based on (standard) rejection sampling may have unacceptably long waiting times.

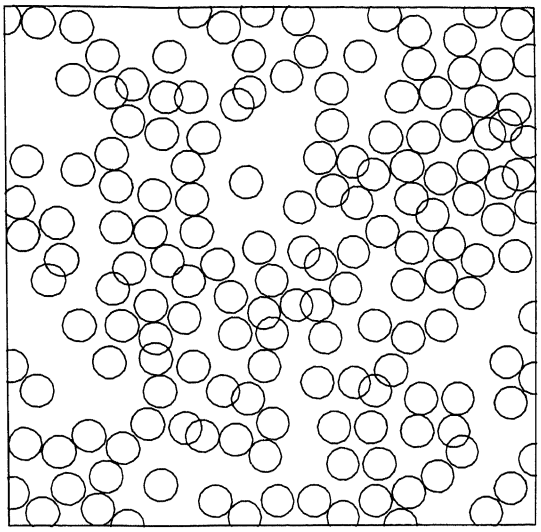


Figure 4.12: Realisation from the posterior distribution for the Brodatz pellets texture sampled at time 1. The prior distribution is a Strauss process with $\log \beta = \log \gamma = -1000$.

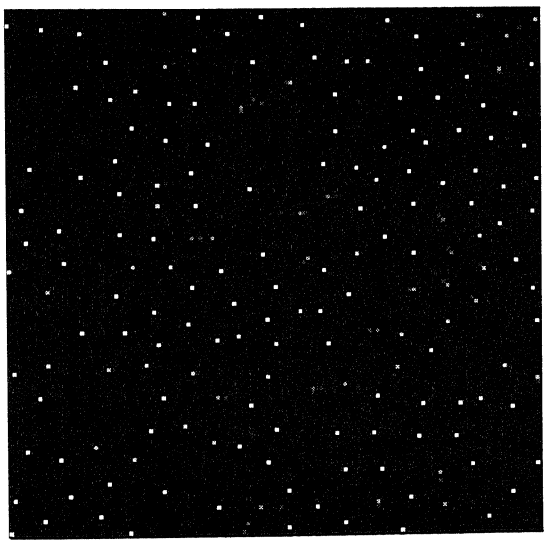


Figure 4.13: Posterior intensity estimated over 10 time units for the Brodatz pellets texture. The prior distribution is a Strauss process with $\log \beta = \log \gamma = -1000$.

4.5. Convergence of inhomogeneous Markov processes

Given a family of well-defined spatial birth-and-death processes indexed by temperature parameter H, each converging in distribution to the corresponding $p_H(\cdot \mid \mathbf{y})$, stochastic annealing involves an inhomogeneous Markov process with H gradually dropping to zero. To find an annealing schedule that ensures convergence to a mode of the posterior distribution, some results on inhomogenuous Markov processes are needed. The discrete time case has been studied in [50, 120].

4.5.1 Definitions

Let μ and ν be probability measures on a common measurable space $(\mathcal{S}, \mathcal{A})$. Their *total variation distance* is defined as the maximal difference in mass on measurable subsets $A \in \mathcal{A}$

$$\|\mu - \nu\| = \sup_{A \in \mathcal{A}} \mid \mu(A) - \nu(A) \mid.$$

If $|\mathcal{S}| < \infty$

$$\|\mu - \nu\| = \frac{1}{2} \sum_{i \in \mathcal{S}} \mid \mu(i) - \nu(i) \mid.$$

Similarly in the continuous case, if both μ and ν are absolutely continuous with respect to some measure m with Radon-Nikodym derivatives f_μ and f_ν,

$$\|\mu - \nu\| = \frac{1}{2} \int_{\mathcal{S}} \mid f_\mu(s) - f_\nu(s) \mid dm(s).$$

Definition 6 *For a transition probability function (stochastic matrix) $P(\cdot, \cdot)$ on $(\mathcal{S}, \mathcal{A})$,* **Dobrushin's contraction coefficient** *$c(P)$ is defined by*

$$c(P) = \sup_{x,y \in \mathcal{S}} \| P(x, \cdot) - P(y, \cdot) \|.$$

We list some properties that will be used in the sequel (see Dobrushin [34, Section 3]).

Lemma 17 *Let Λ be the set of all probability measures on $(\mathcal{S}, \mathcal{A})$. Then for all transition probability functions P and Q and for all $\mu, \nu \in \Lambda$ the following hold*

(i) $c(P) \leq 1$;

(ii) $c(P) = \sup_{\mu \neq \nu \in \Lambda} \frac{\|\mu P - \lambda P\|}{\|\mu - \lambda\|}$;

(iii) $\|\mu P - \nu P\| \leq c(P) \ \|\mu - \nu\|$;

(iv) $c(PQ) \leq c(P)\ c(Q)$.

For completeness, the proof is given below. Much can be found in [34].

Proof: Property (i) is trivial;

To prove (ii) first note that, since $\|\delta_x - \delta_y\| = 1$ for all $x \neq y$,

$$\begin{aligned} c(P) &= \sup_{x \neq y} \frac{\|\delta_x P - \delta_y P\|}{\|\delta_x - \delta_y\|} \\ &\leq \sup_{\lambda \neq \mu \in \Lambda} \frac{\|\mu P - \lambda P\|}{\|\mu - \lambda\|}. \end{aligned}$$

For the reverse inequality, choose $\lambda, \mu \in \Lambda$ and consider 'the' Jordan-Hahn decomposition [52] of $\mathcal{S}$ with respect to $\mu - \lambda$, i.e. find a disjoint measurable partition A^+, A^- such that

$$(\mu - \lambda)(A) \geq 0 \quad \forall A^+ \supseteq A \in \mathcal{A},$$

$$(\mu - \lambda)(A) \leq 0 \quad \forall A^- \supseteq A \in \mathcal{A}.$$

Then $\|\mu - \lambda\| = (\mu - \lambda)(A^+) \vee (\lambda - \mu)(A^-)$ and since $(\mu - \lambda)(A^+ \cup A^-) = 0$,

$$\|\mu - \lambda\| = (\mu - \lambda)(A^+) = (\lambda - \mu)(A^-).$$

Setting $\tau(A) := \inf_{y \in \mathcal{S}} P(y, A)$, for all $x \in \mathcal{S}$

$$P(x, A) - \tau(A) \leq \sup_{x,y,A} \mid P(x, A) - P(y, A) \mid = c(P).$$

Consider the case $\mu P(A) \geq \lambda P(A)$. The other case is similar. Then

$$\mu P(A) - \lambda P(A) = \int_{\mathcal{S}} P(x, A)(\mu - \lambda)(dx)$$

$$= \int_{A^+} P(x, A)(\mu - \lambda)(dx) + \int_{A^-} P(x, A)(\mu - \lambda)(dx)$$

$$\leq \int_{A^+} P(x, A)(\mu - \lambda)(dx) + \int_{A^-} \tau(A)(\mu - \lambda)(dx)$$

$$= \int_{A^+} (P(x, A) - \tau(A))\ (\mu - \lambda)(dx)$$

$$\leq c(P)\ (\mu - \lambda)(A^+) = c(P)\ \|\mu - \lambda\|.$$

Therefore

$$\|\mu P - \lambda P\| = \sup_{A \in \mathcal{A}} | \mu P(A) - \lambda P(A) | \leq c(P) \ \|\mu - \lambda\|$$

and hence

$$\sup_{\lambda \neq \mu \in \Lambda} \frac{\|\mu P - \lambda P\|}{\|\mu - \lambda\|} \leq c(P)$$

which proves (ii). Property (iii) is an easy consequence of property (ii); for (iv) note that

$$\begin{aligned} \|(PQ)(x,\cdot) - (PQ)(y,\cdot)\| &= \|P(x,\cdot)Q - P(y,\cdot)Q\| \\ &\leq c(Q) \ \|P(x,\cdot) - P(y,\cdot)\| . \end{aligned}$$

Therefore

$$\begin{aligned} c(PQ) &= \sup_{x,y} \|PQ(x,\cdot) - PQ(y,\cdot)\| \\ &\leq c(Q) \sup_{x,y} \|P(x,\cdot) - P(y,\cdot)\| \\ &= c(P)\, c(Q). \end{aligned}$$

□

4.5.2 Limit theorems

The main theorem of this Section states sufficient conditions under which a sequence of Markov processes converges in total variation to a well-defined limit.

Recall that the transition semi-group $(Q_t)_{t \geq 0}$ of a Markov process $(Y_t)_{t \geq 0}$ in continuous time is the semi-group of probability kernels representing its conditional distributions,

$$Q_t(y, F) = \mathbb{P}(Y_t \in F \mid Y_0 = y).$$

Theorem 18 *Let $(X_t)_{t \geq 0}$ be a non-stationary Markov process on a measurable space $(S, \mathcal{A})$, defined by a sequence of transition semi-groups $(Q_n)_{n \in \mathbb{N}}$. The process follows the transition rules Q_n in the time period $[t_n, t_{n+1})$, that is*

$$\mathbb{P}(X_s \in F \mid X_r = y) = (Q_n)_{s-r}(y, F)$$

for $t_n \leq r < s < t_{n+1}$. Here $t_n \nearrow \infty$ as $n \to \infty$. Assume that for each $n \in \mathbb{N}$, Q_n has an invariant measure μ_n, i.e.

$$\int_S (Q_n)_t(x, F)\, d\mu_n(x) = \mu_n(F)$$

for all $F \in \mathcal{A}$ and $t \geq 0$. Assume moreover that the following hold

$$(C) \quad \sum_{n=1}^{\infty} \|\mu_n - \mu_{n+1}\| < \infty$$

$$(D) \quad c(P_{tt'}) \to 0 \text{ as } t' \to \infty \text{ for all } t \geq 0$$

where $P_{tt'}(x, F) = \mathbb{P}(X_{t'} \in F \mid X_t = x)$. Then $\mu_\infty = \lim \mu_n$ exists and $\nu P_{0t} \to \mu_\infty$ in total variation as $t \to \infty$, uniformly in the initial distribution ν.

Proof: Condition (C) implies that (μ_n) is a Cauchy sequence in $\|\cdot\|$ and hence converges in total variation to μ_∞, say.
Define $n(t) = \sup\{n : t_n \leq t\}$ choose $0 \leq t < t_{n(t)+1} < t' < \infty$. Then

$$\begin{aligned} \mu_\infty P_{tt'} - \mu_\infty &= (\mu_\infty - \mu_{n(t)})P_{tt'} + \mu_{n(t)}P_{tt'} - \mu_\infty \\ &= (\mu_\infty - \mu_{n(t)})P_{tt'} + \mu_{n(t)}P_{tt_{n(t)+1}}P_{t_{n(t)+1}t'} - \mu_\infty. \end{aligned}$$

Since $\mu_{n(t)}$ is an invariant measure one sees that

$$\begin{aligned} \mu_\infty P_{tt'} - \mu_\infty &= (\mu_\infty - \mu_{n(t)})P_{tt'} + \mu_{n(t)}P_{t_{n(t)+1}t'} - \mu_\infty \\ &= (\mu_\infty - \mu_{n(t)})P_{tt'} + \sum_{k=n(t)}^{n(t')-1} (\mu_k - \mu_{k+1})P_{t_{k+1}t'} \\ &+ \mu_{n(t')} - \mu_\infty. \end{aligned}$$

Hence

$$\begin{aligned} \|\mu_\infty P_{tt'} - \mu_\infty\| &\leq \left\|\mu_\infty - \mu_{n(t)}\right\| c(P_{tt'}) \\ &+ \sum_{k=n(t)}^{n(t')-1} \|\mu_k - \mu_{k+1}\| \, c(P_{t_{k+1}t'}) + \left\|\mu_{n(t')} - \mu_\infty\right\| \\ &\leq 2 \sup_{k \geq n(t)} \|\mu_k - \mu_\infty\| + \sum_{k=n(t)}^{\infty} \|\mu_k - \mu_{k+1}\| \\ &\to 0 \ (t \to \infty). \end{aligned}$$

Let $\epsilon > 0$. Choose t such that $\|\mu_\infty P_{tt'} - \mu_\infty\| < \epsilon/2$ for all $t' > t_{n(t)+1}$.
Next observe that

$$\begin{aligned} \|\nu P_{0t'} - \mu_\infty\| &= \|(\nu P_{0t} - \mu_\infty)P_{tt'} + \mu_\infty P_{tt'} - \mu_\infty\| \\ &\leq \|\nu P_{0t} - \mu_\infty\| \, c(P_{tt'}) + \|\mu_\infty P_{tt'} - \mu_\infty\| \\ &\leq c(P_{tt'}) + \|\mu_\infty P_{tt'} - \mu_\infty\|. \end{aligned}$$

Use condition (D) to choose t' such that $c(P_{tt'}) < \epsilon/4$. Summarising, we obtain

$$\|\nu P_{0t} - \mu_\infty\| \to 0 \quad \text{uniformly in } \nu \ (t \to \infty).$$

□

A sufficient condition for (D) is given by the next result. It is easier to work with, since only stationary Markov chains have to be considered.

Lemma 19 *Use the same notation as in the previous Theorem. If* $c(P_{t_n t_{n+1}}) \leq 1 - 1/n$ *for all* $n \geq 2$, *the Dobrushin condition (D) holds.*

Proof: Write $P_n = P_{t_n t_{n+1}}$. Then

$$-\log c(P_n) \geq 1 - c(P_n) \geq \frac{1}{n}.$$

Thus

$$-\sum_{n=2}^{\infty} \log c(P_n) \geq \sum_{n=2}^{\infty} \frac{1}{n} = \infty$$

or equivalently

$$\prod_{n=2}^{N} c(P_n) \to 0 \ (N \to \infty).$$

Fix t. Then for $t' > t_{n(t)+2}$

$$\begin{aligned} c(P_{tt'}) &= c(P_{tt_{n(t)+1}} P_{t_{n(t)+1} t_{n(t)+2}} \cdots P_{t_{n(t')} t'}) \\ &\leq c(P_{tt_{n(t)+1}}) \left[\prod_{i=n(t)+1}^{n(t')-1} c(P_i) \right] c(P_{t_{n(t')} t'}) \\ &\leq \prod_{i=n(t)+1}^{n(t')-1} c(P_i) \to 0 \ (t' \to \infty). \end{aligned}$$

□

4.6. OBJECT RECOGNITION BY STOCHASTIC ANNEALING

An alternative method for solving the MAP equations (4.1.2) can be based on the results of Section 4.5. Assuming the conditions of Corollary 16, for each fixed H we can construct a spatial birth-and-death process with equilibrium distribution $p_H(\cdot \mid \mathbf{y})$. Our proposal therefore is to use a stochastic annealing algorithm that simulates these processes consecutively with H gradually dropping to zero. In contrast to ICM, if the temperature H decreases sufficiently slowly, stochastic annealing results in a global maximum.

In the superficially similar context of image segmentation, a simulated annealing algorithm was developed by Geman and Geman [42]. However, the Markov processes involved are rather different. Since in segmentation problems both object and image space are finite pixel grids, a discrete time Markov chain changing each pixel label in turn suffices.

4.6.1 The summability condition

First we consider the summability condition (C) in Theorem 18.

Lemma 20 *For fixed data* $\mathbf{y}$, *let* $H_n \searrow 0$ $(n \to \infty)$ *and consider the sequence of* H_n*-modified posterior distributions with densities*

$$p_{H_n}(\mathbf{x} \mid \mathbf{y}) \propto \{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\}^{1/H_n}$$

with respect to the reference measure π *on* Ω *(law of Poisson process or counting measure in the discretised case). Assume that* $\pi(\mathcal{M}) > 0$, *where* $\mathcal{M}$ *denotes the set of solutions to the MAP equations (4.1.2). Then the sequence* p_{H_n} *converges in total variation to a uniform distribution on* $\mathcal{M}$. *Moreover the sequence satisfies condition (C).*

Proof: Since $\pi(\mathcal{M}) > 0$, $\exists\, \mathbf{x}^{\#}$ attaining the maximum. Denote

$$l_n(\mathbf{x}) = \left(\frac{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{f(\mathbf{y} \mid \mathbf{x}^{\#})\, p(\mathbf{x}^{\#})} \right)^{1/H_n}, \quad Z_n = \int_{\Omega} l_n(\mathbf{x}) d\pi(\mathbf{x}).$$

It is easily seen that $l_n(\mathbf{x}) \to 1\{\mathbf{x} \in \mathcal{M}\}$, $(n \to \infty)$, and

$$\lim_{n\to\infty} Z_n = \int_{\Omega} \lim_{n\to\infty} l_n(\mathbf{x}) d\pi(\mathbf{x}) = \int_{\Omega} 1\{\mathbf{x} \in \mathcal{M}\} d\pi(\mathbf{x}) = \pi(\mathcal{M})$$

by the dominated convergence theorem. Moreover, $Z_n \downarrow \pi(\mathcal{M})$, $\sup Z_n = Z_1 \leq \pi(\Omega) < \infty$.

To prove condition (C), suppress the dependence on $\mathbf{x}$ and consider

$$\left| \frac{l_n}{Z_n} - \frac{l_{n+1}}{Z_{n+1}} \right| = (Z_n Z_{n+1})^{-1} \mid l_n Z_{n+1} - l_{n+1} Z_n \mid$$

$$\leq (Z_n Z_{n+1})^{-1} \{ l_{n+1} | Z_{n+1} - Z_n | + Z_{n+1} | l_{n+1} - l_n | \}$$

$$\leq \frac{1}{\pi(\mathcal{M})^2} \{ Z_{n+1}(l_n - l_{n+1}) + l_{n+1}(Z_n - Z_{n+1}) \}$$

$$\leq \frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \{ l_n - l_{n+1} + Z_n - Z_{n+1} \}$$

Therefore

$$\sum_{n=1}^{N-1} \int_\Omega \left| \frac{l_n(\mathbf{x})}{Z_n} - \frac{l_{n+1}(\mathbf{x})}{Z_{n+1}} \right| d\pi(\mathbf{x}) \leq$$

$$\leq \sum_{n=1}^{N-1} \frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \int_\Omega \{ l_n(\mathbf{x}) - l_{n+1}(\mathbf{x}) + Z_n - Z_{n+1} \} \, d\pi(\mathbf{x})$$

$$= \frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \int_\Omega \sum_{n=1}^{N-1} \{ l_n(\mathbf{x}) - l_{n+1}(\mathbf{x}) + Z_n - Z_{n+1} \} \, d\pi(\mathbf{x})$$

$$= \frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \int_\Omega \{ l_1(\mathbf{x}) - l_N(\mathbf{x}) + Z_1 - Z_N \} \, d\pi(\mathbf{x})$$

$$= \frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \, (1 + \pi(\Omega)) \, (Z_1 - Z_N)$$

Letting $N \to \infty$ the right hand side converges to $\frac{\pi(\Omega)}{\pi(\mathcal{M})^2} \, (1 + \pi(\Omega)) \, (Z_1 - \pi(\mathcal{M})) < \infty$. □

A more restricted version, $\pi_n \propto \exp[-f/H_n]$ for a bounded, measurable function f was proved in Theorem 3.3.a of [50]. The assumption $\pi(\mathcal{M}) > 0$ is needed; if $\pi(\mathcal{M}) = 0$ the sequence of modified posterior distributions will not converge in total variation (cf. [50, Theorem 3.3.b]).

4.6.2 The Dobrushin condition

From now on let f be a blur-free independent noise model with $g(\cdot|\cdot) > 0$. Again, let $H_n \searrow 0$ and consider the family $(X^{(n)})_{n \in \mathbb{N}}$ of spatial birth-and-death processes on $K = \{ \mathbf{x} : f(\mathbf{y} \mid \mathbf{x}) \, p(\mathbf{x}) > 0 \}$ defined by (4.4.6) and (4.4.7). As K is closed and irreducible, the processes are well-defined and converge to the unique limit $p_H(\cdot \mid \mathbf{y})$ (see Section 4.4).

Recall the following result by Møller [83], a generalisation of earlier work by Lotwick and Silverman [77].

Theorem 21 *Let X_t be a spatial birth-and-death process and define κ_n, δ_n as in Theorem 15. Assume moreover that $\delta_n > 0$ for all $n \geq 1$ and $\kappa_n = 0$ for all $n > n_0$ (condition (a)). Then for all fixed $t_0 > 0$*

$$\sup_{\mathbf{x},\mathbf{y}} \|P_t(\mathbf{x},\cdot) - P_t(\mathbf{y},\cdot)\| \leq 2\,(1 - K(t_0))^{\frac{t}{t_0}-1}$$

for all $t > t_0$. *The supremum is taken over all initial states* $\mathbf{x}, \mathbf{y}$ *containing at most* n_0 *objects.*

Here

$$\tilde{\delta}_m = \min_{i,j:i+j=m} \delta_i + \delta_j;\ \tilde{\kappa}_m = \max_{i,j:i+j=m} \kappa_i + \kappa_j;\ \tilde{\alpha}_m = \tilde{\delta}_m + \tilde{\kappa}_m;$$

$$K_1(j,t) = \left[1 - e^{-\tilde{\alpha}_j t}\right] \frac{\tilde{\delta}_j}{\tilde{\alpha}_j};\ K_0(n,t_0) = \prod_{j=1}^{n} K_1(j, \frac{t_0}{n});$$

$$K(t_0) = \min_{n \leq n_0} K_0(n,t_0). \tag{4.6.10}$$

Therefore, by Lemma 19 we can construct an annealing schedule satisfying condition (D) by requiring

$$2(1 - K(t_0))^{\frac{t}{t_0}-1} \leq 1 - \frac{1}{n}$$

or equivalently

$$t \geq t_0 \left(1 + \frac{\log(\frac{1}{2}\,(1 - \frac{1}{n}))}{\log(1 - K(t_0))}\right). \tag{4.6.11}$$

Under the assumption in Lemma 20, condition (C) also holds and by Theorem 18 the sequence of birth-and-death processes constructed this way converges in total variation to a uniform distribution on the set of global maxima of the posterior distribution, regardless of the initial state.

The proof of Theorem 21 is instructive and needed for the generalisations discussed below.

Proof: (Lotwick & Silverman, Møller)

Let X_t and Y_t be two independent spatial birth-and-death processes and assume that the number of objects is bounded by n_0. The initial distributions are λ, an arbitrary probability measure concentrated on configurations with at most n_0 points and the equilibrium measure ν respectively. Write $Z_t = (X_t, Y_t)$. Since state $(0,0)$ is discrete, coupling applies. Define

$$U_t = \begin{cases} X_t, & t < \tau \\ Y_t, & t \geq \tau \end{cases}$$

where $\tau = \inf\{t > 0 \ : \ X_t = Y_t = 0\}$.

Then τ is a stopping time and $U_t \stackrel{d}{=} X_t$. Hence for every measurable A,

$$\mathbb{P}_\lambda(X_t \in A) - \nu(A) = \mathbb{P}_\lambda(U_t \in A) - \nu(A)$$

$$= \mathbb{P}_\lambda(X_t \in A\ ;\ t < \tau) + \mathbb{P}_\lambda(Y_t \in A\ ;\ t \geq \tau) - \nu(A)$$

$$\leq \mathbb{P}_\lambda(X_t \in A\ ;\ t < \tau) + \mathbb{P}_\lambda(Y_t \in A) - \nu(A)$$

$$= \mathbb{P}_\lambda(X_t \in A\ ;\ t < \tau) \leq \mathbb{P}(t < \tau).$$

Similarly

$$\nu(A) - \mathbb{P}_\lambda(X_t \in A) = \nu(A) - \mathbb{P}_\lambda(U_t \in A)$$

$$= \nu(A) - \mathbb{P}_\lambda(X_t \in A\ ;\ t < \tau) - \mathbb{P}_\lambda(Y_t \in A\ ;\ t \geq \tau)$$

$$= \mathbb{P}_\lambda(Y_t \in A\ ;\ t < \tau) - \mathbb{P}_\lambda(X_t \in A\ ;\ t < \tau)$$

$$\leq \mathbb{P}_\lambda(Y_t \in A\ ;\ t < \tau) \leq \mathbb{P}_\lambda(t < \tau).$$

Thus

$$\|\nu - \mathbb{P}_\lambda(X_t \in \cdot)\| \leq \mathbb{P}_\lambda(t < \tau). \tag{4.6.12}$$

To find a bound on $\mathbb{P}_\lambda(t < \tau)$, again assume that X_t and Y_t are independent birth-and-death processes and fix an initial state $Z_0 = (\mathbf{x}, \mathbf{y})$ containing m and n objects respectively. Then $\mathbb{P}$(first transition in Z occurs before time t and is a death $\mid Z_0) \geq K_1(m+n, t)$. To see this, set

$$f(x,y) = [1 - e^{-x-y}]\ \frac{y}{x+y}$$

where $x = (B_H(\mathbf{x}) + B_H(\mathbf{y}))\ t \geq 0$ and $y = (D_H(\mathbf{x}) + D_H(\mathbf{y}))\ t > 0$. The partial derivatives of f are

$$\frac{\partial f}{\partial x}(x,y) = \frac{y}{(x+y)^2}\ [(1+x+y)e^{-x-y} - 1];$$

$$\frac{\partial f}{\partial y}(x,y) = \frac{y}{x+y}\ e^{-x-y} + \frac{x}{(x+y)^2}\ (1 - e^{-x-y}) > 0.$$

Set $g(z) = (1+z)e^{-z} - 1$ on $(0, \infty)$. Since $g'(z) = -ze^{-z} < 0$, $\frac{\partial f}{\partial x}(x,y) < 0$. Therefore a lower bound can be obtained by taking x as large and y as small as possible. It follows that

$$\mathbb{P}(\tau \leq t_0 \mid Z_0) \geq$$

$$\mathbb{P}(\text{ first } m+n \text{ transitions occur before } t_0 \text{ and all are deaths} \mid Z_0) \geq$$

$$\mathbb{P}(\text{ first } m+n \text{ sojourn times are all } \leq \frac{t_0}{m+n}$$

$$\text{and all these transitions are deaths } \mid Z_0) \geq K_0(m+n, t_0).$$

Since the number of objects is bounded above by n_0, $K(t_0) = \min_{n \leq n_0} K_0(n, t_0) > 0$ and $\mathbb{P}(\tau \leq t_0) \geq K(t_0)$ for all initial distributions. In fact,

$$\mathbb{P}(\tau > kt_0) \leq (1 - K(t_0))^k$$

for all $k \in \mathbb{N}$.

The proof is by induction. For $k = 1$ the assertion has been proved already and we may assume the statement is correct for some $k \in \mathbb{N}$. Then

$$\mathbb{P}(\tau > (k+1)t_0) =$$

$$\int \mathbb{P}(Z_t \neq (0,0) \text{ in } (t_0, (k+1)t_0) \mid Z_{t_0} = z)$$

$$\mathbb{P}(Z_t \neq (0,0) \text{ before } t_0 \text{ and } Z_{t_0} \in dz)$$

$$\leq (1 - K(t_0))^k \int \mathbb{P}(Z_t \neq (0,0) \text{ before } t_0 \text{ and } Z_{t_0} \in dz)$$

$$= (1 - K(t_0))^k \, \mathbb{P}(Z_t \neq (0,0) \text{ before } t_0) \leq (1 - K(t_0))^{k+1}$$

which completes the induction argument.

Now, for arbitrary $t > t_0$,

$$\mathbb{P}(\tau > t) \leq \mathbb{P}(\tau > \lfloor \frac{t}{t_0} \rfloor t_0) \leq (1 - K(t_0))^{\lfloor \frac{t}{t_0} \rfloor} \leq (1 - K(t_0))^{\frac{t}{t_0} - 1}.$$

Using (4.6.12), Theorem 21 is proved. □

4.6.3 Example

Consider the synthetic example studied in Section 4.3. We will use the same parameter values here, to enable comparison.

In practice, the theoretical temperature schedule (4.6.11) is too slow and one resorts to 'feasible' schemes. Here we chose a geometric cooling of rate $1/2$ and initial temperature $H = 4.0$. The log posterior likelihood as a function of time is given in Figure .4; Figure .5 graphs the number of objects against time. Finally a typical reconstruction sampled at $H = .25$ is given in Figure 4.14. The constant death rate method was used throughout.

In contrast to ICM, stochastic annealing results in a global maximum, regardless of the initial state. Experiments with several initial states are in accordance with the theory in that similar reconstructions were obtained. One has to be careful though, since a too fast cooling schedule was used. For a discussion on the implications of such ad hoc choices, see [46].

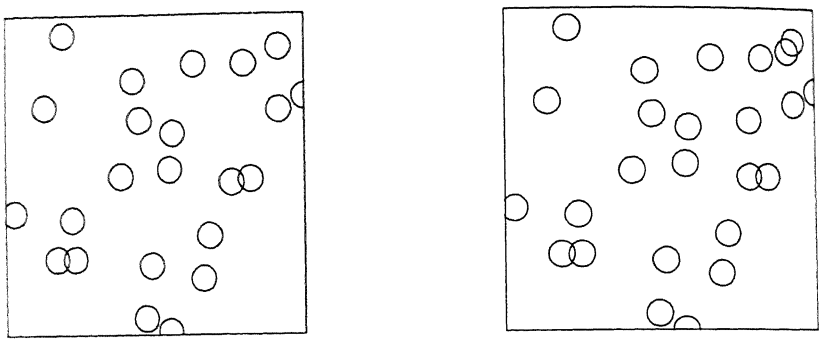


Figure 4.14: Sample taken at time 25 for a geometric cooling schedule starting at $H = 4.0$ of rate .5 (left) and from the posterior distribution after 2.0 time units (right) for a Strauss prior with $\beta = .0025$ and $\gamma = .25$

A good compromise between ICM and stochastic annealing is to sample at a fixed 'low' temperature (see also [41]). A reconstruction obtained by running the constant death rate procedure at $H = 1$ is given in Figure 4.14. Using constant birth rate instead was found to behave worse. The latter method tends to add an unlikely object and immediately deletes it again. This confirms experience reported in the literature ([83, 98]).

Estimates of the posterior intensity (Figure 4.15) suggest that the posterior distribution is rather peaked and can be used as an approximation to MAP estimation, apart from being interesting in its own right.

Typical runs of the constant death rate method are illustrated in Figure .6, where the Δ_2 distance [2] to the 'true' pattern is graphed against time. Starting from an empty scene, objects are immediately added to form a plausible reconstruction followed by deletion and immediate readding of one of the objects. Note that the results obtained this way are comparable to steepest ascent reconstructions.

For the Brodatz pellet texture of Figure 2.1, the results of simulated annealing are very similar to those of posterior sampling (Figure 4.12), as could be expected from Figure 4.13.

Figure 4.15: Posterior intensity at temperature 4.0, 1.0 and .25 (from top to bottom).

4.6.4 Remarks and extensions

Generalisations to diffusing objects are possible. In the finite case $|U| < \infty$, write

$$M(\mathbf{x}, x_i, u)$$

for the configuration obtained from $\mathbf{x}$ by replacing x_i by u. The set of $u \in U$ for which this operation is allowed is denoted by $Q(\mathbf{x}, x_i)$. Typically it consists of unoccupied objects close but not identical to x_i.

Suppose the diffusion rates are also powers of the log likelihood ratios

$$c_H(\mathbf{x}, x_i, u) = \left\{ \frac{f(\mathbf{y} \mid M(\mathbf{x}, x_i, u))\, p(M(\mathbf{x}, x_i, u))}{f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})} \right\}^{k/H} . \quad (4.6.13)$$

If detailed balance is required to hold as well, necessarily $k = \frac{1}{2}$.

Writing $C_H(\mathbf{x})$ for the total diffusion rate out of state $\mathbf{x}$, the expected sojourn time $\frac{1}{B_H(\mathbf{x})+C_H(\mathbf{x})+D_H(\mathbf{x})}$ is less than $1/\left(B_H(\mathbf{x}) + D_H(\mathbf{x})\right)$, the expected waiting time for the process with only births and deaths. New positions with a large increase in posterior likelihood are favoured.

The following additional notation is needed.

$$\gamma_i = \max_{n(\mathbf{x})=i} C_H(\mathbf{x});\ \tilde{\gamma}_m = \max_{i,j:i+j=m} \gamma_i + \gamma_j;\ \tilde{\alpha}_m = \tilde{\delta}_m + \tilde{\kappa}_m + \tilde{\gamma}_m$$

and K_1, K_0, K as in (4.6.10).

Lemma 22 *The Markov chain on K with rates given by (4.4.6), (4.4.7) and (4.6.13) is irreducible and satisfies the detailed balance equations. It is time reversible and has unique equilibrium distribution $p_H(\cdot|\mathbf{y})$. Moreover for all fixed $t_0 > 0$*

$$\max_{\mathbf{x},\mathbf{y}} || P_t(\mathbf{x}, \cdot) - P_t(\mathbf{y}, \cdot) || \leq 2\,(1 - K(t_0))^{\frac{t}{t_0}-1}$$

for all $t > t_0$.

Proof: Every configuration $\mathbf{x}$ can be reached from every other configuration $\mathbf{z}$ by first deleting all points in $\mathbf{x} \setminus \mathbf{z}$ and then adding $\mathbf{z} \setminus \mathbf{x}$, since all birth and death rates are strictly positive. Thus the Markov chain is irreducible. Since the detailed balance equations hold by construction, the first part of the result follows. The second part can be derived by coupling arguments similar to those in the proof of Theorem 21. The only change is that we now have to prove that $\mathbb{P}($ the first transition in Z occurs before time t and is a death $\mid Z_0) =$
$[1 - \exp\{-(B_H(\mathbf{x}) + B_H(\mathbf{y}) + C_H(\mathbf{x}) + C_H(\mathbf{y}) + D_H(\mathbf{x}) + D_H(\mathbf{y}))t\}]$
$\frac{D_H(\mathbf{x})+D_H(\mathbf{y})}{B_H(\mathbf{x})+B_H(\mathbf{y})+C_H(\mathbf{x})+C_H(\mathbf{y})+D_H(\mathbf{x})+D_H(\mathbf{y})} \geq K_1(m+n,t)$. To see this, let $f(x,y,z) = \left[1 - e^{-(x+y+z)}\right] \frac{z}{x+y+z}$. The partial derivatives are

$$\frac{\partial f}{\partial x}(x,y,z) = \frac{z}{(x+y+z)^2}\left[(x+y+z+1)\ e^{-(x+y+z)} - 1\right];$$

$$\frac{\partial f}{\partial y}(x,y,z) = \frac{z}{(x+y+z)^2}\left[(x+y+z+1)\ e^{-(x+y+z)} - 1\right];$$

and

$$\frac{\partial f}{\partial z}(x,y,z) = e^{-(x+y+z)}\frac{z}{x+y+z} + \left[1 - e^{-(x+y+z)}\right]\frac{x+y}{(x+y+z)^2}.$$

Since $\frac{\partial f}{\partial x}(x,y,z) < 0$, $\frac{\partial f}{\partial y}(x,y,z) < 0$ and $\frac{\partial f}{\partial z}(x,y,z) > 0$, the statement follows. □

In the continuous case, n objects can perform a diffusion on U^n

$$d\mathbf{x}_t = \nabla \log p_H(\mathbf{x} \mid \mathbf{y})\ dt + \sqrt{2H}\ dB_t$$

at least if $p_H(\cdot \mid \mathbf{y})$ is strictly positive and infinitely differentiable [98, p. 178]. Here B_t denotes Brownian motion. See also [43, 82].

4.7. IMPLEMENTATION AND COMPUTATIONAL COMPLEXITY

The building blocks of Algorithms 1-7 are the forward log likelihood ratios (2.4.6), (2.4.7) and (2.4.8) and the prior log likelihood ratios (3.1.5). The former is a function of the object parameter u, that is closely related to the Hough transform and involves only pixels in a relatively small zone $Z(\mathbf{x}, u)$ determined by u and the current pattern $\mathbf{x}$ (cf. (2.5.11), Section 2.5). Focusing on blur-free models, the (generalised) Hough transform (2.5.9) depends on the data image only through

$$z_t = h(y_t, \theta_0, \theta_1) = \log \frac{g(y_t \mid \theta_1)}{g(y_t \mid \theta_0)}$$

which can therefore be 'precomputed' in the initialisation step (in general, the data image is directly involved). Moreover, after adding or deleting a particular point u the log likelihood ratio (2.5.9) requires updating only for v in the region

$$V(u) = \{v \in U : R(v) \cap R(u) \neq \emptyset\}.$$

For example, in a translation model with $U = T = \mathbb{R}^d$ and $R(u) = R_0 + u$ this is the central symmetrisation $V(u) = R_0 \oplus \check{R}_0 + u$.

For a nearest-neighbour Markov prior, the latter component is 'local' too and depends only on those objects in the current reconstruction that are neighbours of u (cf. Section 3.1.3).

Computations are carried out on a logarithmic scale to avoid overflow. In fixed-temperature sampling or stochastic annealing, write a_H for the log birth rate (cf. formulas (4.4.6) and (4.4.7)),

$$a_H(\mathbf{x}, u) = \frac{k}{H}\left(L(\mathbf{x} \cup \{u\}\ ;\ \mathbf{y}) - L(\mathbf{x}\ ;\ \mathbf{y}) + \log \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})}\right).$$

One could first find

$$m(\mathbf{x}) = \max_u\ a_H(\mathbf{x}, u)$$

and then digitise to compute

$$B_H(\mathbf{x}) = e^{m(\mathbf{x})} \sum_j \exp\left[a_H(\mathbf{x}, u_j) - m(\mathbf{x})\right]$$

where the exponential terms lie between 0 and 1. The conditional likelihood of a birth at u is computable from the same summands

$$\frac{b_H(\mathbf{x}, u)}{B_H(\mathbf{x})} = \exp\left[a_H(\mathbf{x}, u) - m(\mathbf{x})\right] \exp\left[m(\mathbf{x}) - \log B_H(\mathbf{x})\right].$$

The death rates can be dealt with similarly.

We implemented the simulations and reconstructions in C++ (AT&T , [113]) using the CLIP library [110] for manipulating the image data.The C++ language is well suited, since it allows the user to define separate classes for each of the mathematical entities; changes in one class do not affect any other class. Here the model components are template shapes, object configurations, signal function, noise model and prior distribution. For each of these we defined a C++ class maintaining the relevant parameter settings, performing update operations and computing the likelihood ratios. Computations were carried out in single precision arithmetic; using double precision resulted in slightly different reconstructions.

4.7.1 Sampling

In fixed temperature sampling or stochastic annealing ($k \neq 1$), the computational effort is mainly in sampling from $b_H(\mathbf{x}, \cdot)/B_H(\mathbf{x})$. Since the birth rate $b_H(\mathbf{x}, u)$ is an exponential function of the Hough transform (2.5.11), it tends to have sharp peaks as a function of u when H is small or when $\mathbf{x}$ is suboptimal. There is then a high probability that the next transition will add a new object u at one of the locations that is close to maximal for $b_H(\mathbf{x}, u)$. This implies that rejection sampling methods that generate a putative new object uniformly become impractical. Many proposals would be rejected and the waiting times would be unacceptably long. This suggests using an algorithm which incorporates a search operation over U.

One simple algorithm of this kind is to find the global maximum of birth rate $b^* = \max_u b_H(\mathbf{x}, u)$, then locate all objects u satisfying $b_H(\mathbf{x}, u) \geq ab^*$ where $a < 1$, or larger than a given threshold value. Making a list of all such candidates we proceed as if these are the only objects in U, computing the total birth rate of the candidates and performing rejection sampling. After each transition the list of candidates has to be recomputed. This algorithm is an approximation to the desired birth-and-death process: larger values of a increase the speed but decrease accuracy of the approximation.

An exact algorithm can be obtained by incorporating a rejection sampling step. Denote the threshold on the birth rates defined above by $w_H(\mathbf{x})$ and, for $u \notin \mathbf{x}$, set

$$g_H(\mathbf{x}, u) = \begin{cases} b_H(\mathbf{x}, u) & \text{if } b_H(\mathbf{x}, u) \geq w_H(\mathbf{x}); \\ w_H(\mathbf{x}) & \text{else.} \end{cases}$$

Writing $G_H(\mathbf{x})$ for the integral of g_H, we generate a sequence $(X^{(k)}, T^{(k)})$ of states and sojourn times as follows.

Algorithm 8 *An initial state* $\mathbf{x}^{(0)}$ *is given. For each* $k = 0, 1, 2, \ldots$ *compute* $D = D_H(\mathbf{x}^{(k)})$, $G = G_H(\mathbf{x}^{(k)})$ *and set* $T^{(k)} = 0$.

- *Add an exponentially distributed time to* $T^{(k)}$ *with mean* $1/(D+G)$;
- *with probability* $\frac{D}{D+G}$ *generate a death* $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} \setminus x_I$ *by deleting one of the points in* $\mathbf{x}^{(k)}$ *at random according to distribution* $D_H(\cdot, \cdot)$ *and stop;*
- *else sample a point* u *from* $g_H(\mathbf{x}^{(k)}, u)/G$; *with probability* $\frac{b_H(\mathbf{x}^{(k)}, u)}{g_H(\mathbf{x}^{(k)}, u)}$ *accept a birth* $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} \cup \{u\}$ *and stop; otherwise repeat the whole procedure.*

It is important to note that the method does not require computation of the total birth rate.

To see the validity of the algorithm above, note that we generate a (1-dimensional) Poisson process of putative events with rate $D+G$. The acceptance probability p of an event is

$$\frac{D}{D+G} + \frac{G}{D+G} \int_U \frac{b_H(\mathbf{x}, u)}{g_H(\mathbf{x}, u)} \frac{g_H(\mathbf{x}, u)}{G} \, d\mu(u) = \frac{D + B_H(\mathbf{x})}{D+G}.$$

Due to the thinning property of Poisson processes [26, p. 241] the sojourn time is exponentially distributed with mean $1/(B_H(\mathbf{x}) + D_H(\mathbf{x})$. Next, consider the transition rate for the addition of an object u to pattern $\mathbf{x}$:

$$\sum_{k=0}^{\infty} (1-p)^k \frac{G}{D+G} \frac{b_H(\mathbf{x}, u)}{g_H(\mathbf{x}, u)} \frac{g_H(\mathbf{x}, u)}{G}$$

$$= \frac{1}{D+G} \, b_H(\mathbf{x}, u) \sum_{k=0}^{\infty} (1-p)^k = \frac{b_H(\mathbf{x}, u)}{D_H(\mathbf{x}) + B_H(\mathbf{x})}.$$

The death rates can be treated similarly.

4.7.2 Multiresolution techniques

Whatever strategy is adopted, there will be problems with the 'curse of dimension', i.e. as the dimension of the object space U increases, the cost of searching U increases exponentially. This problem is well-known in the context of Hough transforms; it is often named as the major limitation on their performance, and multiresolution strategies are usually recommended.

We propose using the following multiresolution algorithm. For each 'resolution level' $m = 0, 1, \ldots, M$ conceptually divide object space U into a partition $\mathcal{U}_m = \{U_{1,m}, \ldots, U_{k_m,m}\}$ such that each partition is a

refinement of the previous one: $U_{i,m} = \bigcup_j U_{\ell_j,m+1}$ for all i, m. 'Conceptually' means that the subdivision of a block U_{im} into smaller blocks is only carried out when needed. The standard example is the quad-tree in which the unit square is divided into $2^m \times 2^m$ smaller squares at stage m.

Interpret each partition $\mathcal{U}_m$ as a class of 'large objects' in T by defining for any $V \subseteq U$

$$R(V) = \bigcup_{u \in V} R(u)$$

(or $\cup_{u \in V} Z(\mathbf{x}, u)$ in the blurred case).

Define a Hough transform on $\mathcal{U}_m$

$$H_{\mathbf{w}}^{(m)}(V) = \sum_{t \in R(V)} w_t, \quad V \in \mathcal{U}_m$$

where $\mathbf{w}$ is any image. This provides an upper bound for the Hough transform on U, provided $w_t \geq 0$:

$$H_{\mathbf{w}}^{(m)}(V) \geq \max_{u \in V} H_{\mathbf{w}}(u).$$

Furthermore the Hough transform at level m is an upper bound for the Hough transform at level $m+1$

$$H_{\mathbf{w}}^{(m)}(U_{im}) \geq \max_{U_{j,m+1} \subset U_{im}} H_{\mathbf{w}}^{(m+1)}(U_{j,m+1}). \tag{4.7.14}$$

Suppose we can find an image $\mathbf{w}$, possibly depending on $\mathbf{x}$, such that

$$w_t \geq \max_{u \in U} h(y_t, \theta^{(\mathbf{x})}(t), \theta^{(\mathbf{x} \cup \{u\})}(t))$$

where h is as in (2.5.11). For example in the blur-free case we can take

$$w_t = 1\{t \notin S(\mathbf{x})\} \ (h(y_t, \theta_0, \theta_1))^+$$

where $a^+ = \max\{a, 0\}$. Assuming the conditions of Corollary 16 hold, we obtain a decreasing sequence of upper bounds on the birth rates of the stochastic annealing procedure:

$$b(\mathbf{x}, u) \leq b^{(M)}(\mathbf{x}, U_{i_M,M}) \leq \cdots \leq b^{(0)}(\mathbf{x}, U_{i_0,0}) \tag{4.7.15}$$

where $u \in U_{i_M,M} \subset \ldots \subset U_{i_0,0}$ and

$$b^{(m)}(\mathbf{x}, V) = \left[\beta \exp\{H_{\mathbf{w}}^{(m)}(V)\}\right]^{1/H}$$

with β is as in Corollary 16. The maximum birth rate, and the set of locations where the birth rate is near to maximum, can then be determined by searching the multiresolution space dynamically in the usual way. This method bears many resemblances to the adaptive Hough transform [56].

An application of this technique is shown in Figure 4.16. This is a synthetic example modelled on the problem of identifying linear features such as long crystals and fission tracks in micrographs of minerals. The objects are line segments of variable length, so that the object space U is 4-dimensional, making simple search methods computationally expensive.

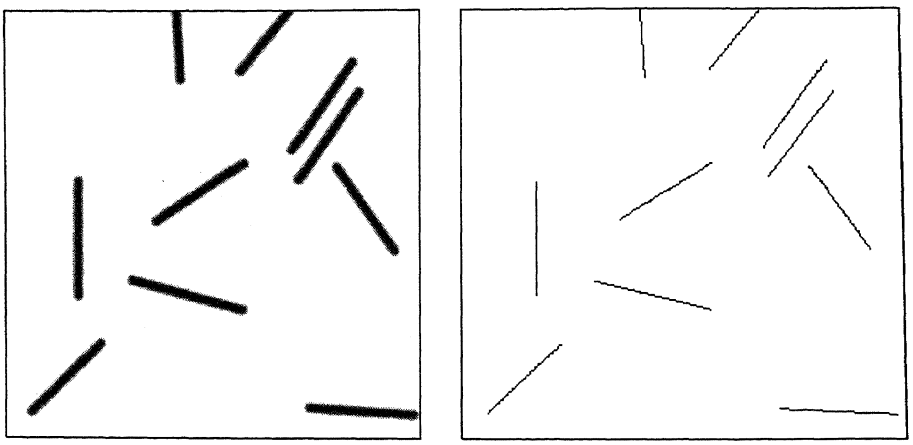


Figure 4.16: Synthetic 256×256 image of blurred line segments degraded by additive Gaussian noise with $\sigma^2 = 9$ (left). Reconstruction by steepest ascent using a Strauss prior with $\beta = .002$ and $\gamma = 0.25$ (right). The initial configuration is empty and $w = 0$.

The signal is assumed to be a function of the distance $d(t, \mathbf{x})$ from t to $S(\mathbf{x})$, in this case linear decay

$$\theta^{(\mathbf{x})}(t) = \begin{cases} \theta_1 + \frac{d(t,\mathbf{x})}{c}(\theta_0 - \theta_1) & \text{for } d(t,\mathbf{x}) \leq c; \\ \theta_0 & \text{otherwise.} \end{cases}$$

Figure 4.16 is a simulation of this model with (arbitrary) line segments of length ranging between 60 and 70, foreground brightness 100, background brightness 254, decay radius $c = 4$ pixel units, and additive Gaussian noise with variance 9.0.

The choice of parametrisation is important for the computational cost of the multiresolution algorithm. We choose to parametrise segments by their length ℓ, orientation ρ, and the coordinates (d, t) of the midpoint

of the segment *after rotation by* $-\rho$, so that d is the distance from the (arbitrary) origin O to the infinite line L containing the segment, and t is the distance from the midpoint of the segment to the foot of the perpendicular OL.

It is interesting to note that this is the standard parametrisation of lines in $\mathbb{R}^2$ in stochastic geometry, where this choice is dictated by properties of the 'invariant' measure for random lines. One may speculate that stochastic geometry can help suggest the right parametrisation for the Hough transform for other classes of geometrical objects, and in general, suggest the right way to formulate many problems in image analysis.

Our multiresolution algorithm splits U at level 0 into blocks of roughly constant (ρ, d) with a tolerance of 1 degree in ρ and 1 pixel unit in d. That is, we group together all those segments which lie (approximately) along a given infinite line of orientation ρ and distance d from the origin. The objects $R(V)$ at level 0 are thickened lines, whose Hough transform is relatively easy to compute.

We introduce a neighbourhood structure by defining two line segments to be neighbours $u \sim v$ iff their dilations by a ball of decay radius have a non-empty intersection. The prior model is a Strauss process with $\beta = 0.002$ and $\gamma = 0.25$. We applied the multiresolution algorithm with upper bounds (4.7.15) computed from $w_t = \frac{1}{2\sigma^2}\ (y_t - \theta^{(\mathbf{x})}(t))^2$. Note that the choice of parametrisation also means that the bounds (4.7.15) at level 0 are rather tight. The result of steepest ascent is given in Figure 4.16.

Chapter 5
Spatial clustering

Another high-level task in computer vision is clustering of image features. This can be formulated within the same general framework, but note that the data is now a list of features, instead of a digital image. Here, we study in detail the special case of clustering spatial point patterns, and indicate how the framework can be extended to other problems, such as edge detection.

5.1. Introduction

There are many similarities between object recognition (Chapter 4) and the problem of detecting clusters in a spatial pattern. In both problems, the goal is to extract an underlying pattern from a set of data that can be a digital image, a point pattern in Euclidean space or the output of an edge detector. In particular, the formal framework is the same and the techniques presented in Chapter 4 carry over with some adaptations.

The localisation of cluster centres is of interest in many areas of applications. Forestry is an obvious example, where seedlings tend to scatter around mature trees. While methods to test for clustering abound in the spatial statistics literature, see eg [31], these do not allow the estimation of cluster centre location or cluster membership.

An early work on locating centres is Baudin [14]. He derived a formula for the posterior intensity of a Neyman-Scott cluster process. However, it was found too complicated for practical application. The Geographical Analysis Machine by Openshaw et al. [92] is a grid based test using Monte Carlo testing in each disc centred at a grid point. The method

is computer intensive, suffers from multiple response and has unknown power. It was improved by Besag and Newell [20]. Their method centres around cases and is less computer intensive, but remains admittedly rather ad hoc.

Here we present a stochastic model in keeping with the general framework described in Chapter 4 and introduce a Markov prior to combat overestimation of the number of clusters and to improve robustness. These ideas in connection with object recognition were proposed by Baddeley and Van Lieshout [8]. Independently, a Gibbs sampler technique for detection of cluster centres in a Cox process was developed by Lawson [65, 66, 67, 69]. The presentation here is based on [8] and the author's contribution to ongoing research with A.B. Lawson and A.J. Baddeley. The suggestion to consider offspring labelling (Section 5.6) is due to A.J. Baddeley.

5.2. CLUSTER PROCESSES

The data consist of a set of points $\mathbf{y} = \{y_1, \ldots, y_m\} \subseteq T$, where $T \subseteq \mathbb{R}^2$ (say) is the window of observation and it is required to determine the locations of an unspecified number of cluster centres $x_1, \ldots, x_n \in T$. Typical applications include the analysis of spatial pattern in the occurrence of rare diseases and the estimation of the positions of ancestors of the current generation of trees in a wood.

In the likelihood approach, $\mathbf{y}$ depends on the unknown pattern $\mathbf{x}$ of cluster centres through a probability density $f(\mathbf{y} \mid \mathbf{x})$. Since the data is no longer a digital image but a list of points, f is the density of a point process with respect to the distribution π of a Poisson point process on T (as reviewed in Chapter 3).

5.2.1 Survey

The analogue of the conditional independence assumption in Section 2.2 is to assume that conditional on $\mathbf{x}$, the observed point process is a superposition of independent finite point processes $N^{(x_i)}$ of 'daughters' associated with each 'parent' $x_i \in \mathbf{x}$. We will denote the Radon-Nikodym density of daughter process $N^{(x_i)}$ by $g(\cdot \mid x_i)$. Models of this type are frequently considered in spatial statistics but the present problem is unusual in that we are trying to estimate the parent (cluster centre) process.

Lemma 23 *For independent clustering with daughter densities $g(\cdot \mid u)$, the data has density*

$$f(\mathbf{y} \mid \mathbf{x}) = e^{\mu(T)} \sum_{\phi} \prod_{i=1}^{n(\mathbf{x})} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right]. \tag{5.2.1}$$

with respect to π. The sum is over all ordered partitions $\phi : \{1, \ldots, n(\mathbf{y})\} \to \{1, \ldots, n(\mathbf{x})\}$, and for every $I \subseteq \{1, \ldots, n(\mathbf{y})\}$, $\mathbf{y}_I = \{y_i : i \in I\}$.

Proof: Set $m = n(\mathbf{y})$ and use the Janossy densities [26, p. 122]. For the total pattern

$$j_m(\mathbf{y} \mid \mathbf{x}) = e^{-\mu(T)} f(\mathbf{y} \mid \mathbf{x})$$

On the other hand, by partitioning in groups belonging to the same parent

$$j_m(\mathbf{y} \mid \mathbf{x}) = \sum_{\phi} \prod_{i=1}^{n(\mathbf{x})} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right].$$

Therefore

$$e^{-\mu(T)} f(\mathbf{y} \mid \mathbf{x}) = \sum_{\phi} \prod_{i=1}^{n(\mathbf{x})} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i)\, e^{-\mu(T)} \right].$$

and the assertion follows. □

Another common assumption in spatial statistics is that the daughter processes $N^{(x_i)}$ are identically distributed up to translation,

$$N^{(x_i)} \stackrel{D}{=} N^{(0)} + x_i,$$

in which case

$$g(\mathbf{z} \mid x_i) = g(\mathbf{z} - x_i \mid 0).$$

The interested reader can consult [26, 31] or [111] for further details.

A class of models that is fairly general, yet amenable to calculations, is that of *inhomogeneous Poisson processes* [111] (also called *general Poisson processes* [26]).

Definition 7 *An* independent inhomogeneous Poisson cluster process *is a point process on U having conditional density*

$$f(\mathbf{y} \mid \mathbf{x}) = \exp\left[\int_T (1 - \lambda(u \mid \mathbf{x})) d\mu(u) \right] \prod_{j=1}^{m} \lambda(y_j \mid \mathbf{x}). \tag{5.2.2}$$

with respect to the reference Poisson process π. The intensity function λ might be of the form

$$\lambda(u \mid \mathbf{x}) = \sum_{i=1}^{n} h(u - x_i), \tag{5.2.3}$$

where $h : U \to [0, \infty)$ is a measurable, integrable function.

In other words, the number of points $n(\mathbf{y})$ in $\mathbf{y}$ is Poisson distributed with mean $\int_T \lambda(u \mid \mathbf{x})\, d\mu(u)$ and conditionally on $n(\mathbf{y})$ the joint density is

$$f_m(\mathbf{y} \mid \mathbf{x}) = \frac{\prod_{j=1}^{m} \lambda(y_j \mid \mathbf{x})}{\frac{1}{\mu(T)^m} \left(\int_T \lambda(u \mid \mathbf{x})\, d\mu(u) \right)^m} \tag{5.2.4}$$

with respect to the distribution of m independent points distributed according to the 'uniform' measure $\mu(\cdot)/\mu(T)$ on T.

As the parent process is unknown, the intensity λ is a random measure. Doubly stochastic Poisson processes of this type are called *Cox processes*. If the parent distribution is that of a stationary Poisson process, $\mathbf{y}$ is

a *Neyman-Scott* process. Kingman [63] has argued that Cox processes are a natural framework in which to model the spatial pattern of a population of reproducing individuals. See also [11] or Chapter 6.

Note that the density is dependent on x only through the intensity function λ. Therefore λ takes over the role of the signal function in object recognition.

Background noise can be taken into account by superposition of a (stationary) Poisson process, adding one extra term to the expression (5.2.3).

Model 5: Matérn cluster process [80]

The daughter points are uniformly distributed in a disc of radius r around the cluster centre,

$$h(u - x_i) = \frac{\mu}{\pi r^2} 1\{||u - x_i|| \leq r\}.$$

Model 6: modified Thomas process

The positions follow a circular Gaussian distribution

$$h(u - x_i) = \frac{\mu}{2\pi\sigma^2} e^{-||u-x_i||^2/2\sigma^2}.$$

Strictly speaking, the original models assumed a Poisson parent process.

For certain applications, homogeneous clustering is inappropriate; in particular, the distribution of $N^{(x_i)}$ may well be dependent on x_i. The generalisation of Definition 7 is obtained by setting

$$\lambda(u \mid \mathbf{x}) = \sum_{i=1}^{n} h(u \mid x_i), \tag{5.2.5}$$

with $h(u \mid x_i)$ dependent on u and x_i. In epidemiology, information on population density is often taken into account as a modulating function $k(u)$ [32, 68]

$$\lambda(u \mid \mathbf{x}) = k(u) \sum_{i=1}^{n} h(u \mid x_i). \tag{5.2.6}$$

The function k can be estimated nonparametrically from census data.

The following corollary to Lemma 23 holds.

Corollary 24 *For Model (5.2.5), the forward model can be written as*

$$f(\mathbf{y} \mid \mathbf{x}) = e^{\mu(T)} \left(\prod_{i=1}^{n(\mathbf{x})} H(x_i) \right) \sum_{\phi} \left[\prod_{i=1}^{n(\mathbf{x})} \prod_{j \in \phi^{-1}(i)} h(y_j \mid x_i) \right].$$

The sum is over all ordered partitions $\phi : \{1, \ldots, n(\mathbf{y})\} \rightarrow \{1, \ldots, n(\mathbf{x})\}$ *and* $H(x_i) = \exp[-\int_T h(t \mid x_i)\, d\mu(t)]$.

Proof: Apply Lemma 23 with

$$g(m_i \mid x_i) = e^{\mu(T)} exp[-\int_T h(t \mid x_i)\, d\mu(t)] \prod_{t \in m_i} h(t \mid x_i).$$

To prove the result directly, remark that $\prod_{j=1}^{m} [\sum_{i=1}^{n} h(y_j \mid x_i)] = \sum_{\phi} \left[\prod_{j=1}^{m} h(y_j \mid x_{\phi(j)})\right]$. Hence

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{x}) &= e^{\mu(T)} \left(\prod_{i=1}^{n(\mathbf{x})} H(x_i)\right) \sum_{\phi} \prod_{j=1}^{n(\mathbf{y})} h(y_j \mid x_{\phi(j)}) \\ &= e^{\mu(T)} \left(\prod_{i=1}^{n(\mathbf{x})} H(x_i)\right) \sum_{\phi} \prod_{i=1}^{n(\mathbf{x})} \prod_{j \in \phi^{-1}(i)} h(y_j \mid x_i) \end{aligned}$$

which completes the proof. □

5.3. MAXIMUM LIKELIHOOD ESTIMATION

Given observation of a point pattern $\mathbf{y}$, the unknown pattern of cluster centres $\mathbf{x}$ is regarded as the parameter to be estimated. Thus, as before, the maximum likelihood equations are

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} f(\mathbf{y} \mid \mathbf{x}).$$

Here the optimisation is over all realisations $\mathbf{x} \in \Omega$. Again, these equations do not necessarily allow a (unique) solution.

Lemma 25 *For an inhomogeneous Poisson cluster model (5.2.2) the maximum likelihood and pseudolikelihood equations coincide and are*

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} \left[\sum_{j=1}^{m} \log \lambda(y_j \mid \mathbf{x}) - \int_T \lambda(t \mid \mathbf{x}) \, d\mu(t) \right]. \tag{5.3.7}$$

Proof: The log likelihood is

$$L(\mathbf{x} \, ; \, \mathbf{y}) = \sum_{j=1}^{m} \log \lambda(y_j \mid \mathbf{x}) + \int_T (1 - \lambda(t \mid \mathbf{x})) \, d\mu(t)$$

hence the maximum likelihood equations are (5.3.7). Since for any Poisson process the Papangelou conditional intensity is equal to the unconditional intensity

$$\lambda(t \, ; \, y \mid \mathbf{x}) = \frac{f(\mathbf{y} \cup \{t\} \mid \mathbf{x})}{f(\mathbf{y} \mid \mathbf{x})} = \lambda(t \mid \mathbf{x})$$

these are the pseudolikelihood equations for $\mathbf{x}$ as well. □

Solving the maximum likelihood equations is equivalent to solving a regression problem. The result should be compared to Lemma 1.

Lemma 26 *An MLE for Model (5.2.2) is a solution of the regression of* $\mathbf{y}$ *on the class of intensity functions* $\{\lambda(\cdot \mid \mathbf{x}) : \mathbf{x} \in \Omega\}$ *with pointwise loss*

$$\frac{1}{m} \int_T \lambda(t \mid \mathbf{x}) d\mu(t) - \log \lambda(y_j \mid \mathbf{x})$$

Proof: Immediate from (5.3.7). □

As an illustration, ignoring a constant $\log \frac{\mu}{2\pi\sigma^2}$, the loss function for Model 6 is

$$\frac{\mu}{2\pi\sigma^2 m} \sum_{i=1}^{n(\mathbf{x})} \int_T \exp[-\|t - x_i\|^2 / (2\sigma^2)] d\mu(t) -$$

$$\log \left\{ \sum_{i=1}^{n(\mathbf{x})} \exp[-\|y_j - x_i\|^2/(2\sigma^2)] \right\}.$$

Lemma 27 *For the Matérn cluster process (Model 5)*

$$\hat{\mathbf{x}} = \mathit{argmax} \left[\sum_{j=1}^{m} \log\, n(\mathbf{x} \cap B(y_j, r)) - \frac{\mu}{\pi r^2} \sum_{i=1}^{n(\mathbf{x})} \mu(T \cap B(x_i, r)). \right]$$

Here $B(u,r)$ denotes the closed ball with radius r centred at u and the maximum is taken over $\Omega(\mathbf{y}) = \{\mathbf{x} \in \Omega : \forall j, \mathbf{x} \cap B(y_j, r) \neq \emptyset\}$.

Proof: Note that the likelihood is positive only for $\mathbf{x} \in \Omega(\mathbf{y})$. In that case the log likelihood is well-defined,

$$\begin{aligned} L(\mathbf{x}\,;\,\mathbf{y}) &= \sum_{j=1}^{m} \log \left[\frac{\mu}{\pi r^2} n(\{i : \|x_i - y_j\| \leq r\} \right] \\ &- \frac{\mu}{\pi r^2} \sum_{i=1}^{n(\mathbf{x})} \int_T 1\{\|t - x_i\| \leq r\}\, d\mu(t) \\ &= m \log \frac{\mu}{\pi r^2} + \sum_{j=1}^{m} \log\, n(\mathbf{x} \cap B(y_j, r)) \\ &- \frac{\mu}{\pi r^2} \sum_{i=1}^{n(\mathbf{x})} \mu(T \cap B(x_i, r)). \end{aligned}$$

□

Restricting attention to configurations with a prespecified number n of points, the score functions for (5.2.2) are

$$\frac{\partial}{\partial x_i} L(\mathbf{x}\,;\,\mathbf{y}) = \sum_{j=1}^{m} \frac{1}{\lambda(y_j \mid \mathbf{x})} \frac{\partial}{\partial x_i} \lambda(y_j \mid \mathbf{x}) - \int_T \frac{\partial}{\partial x_i} \lambda(t \mid \mathbf{x})\, d\mu(t)$$

$(i = 1, \ldots, n)$. If λ is of the form (5.2.3) then the partial derivatives of λ can be replaced by $-$ the derivatives of h evaluated at points $y_j - x_i$.

5.4. THE BAYESIAN APPROACH

Generally a maximum likelihood estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$ may run into difficulties similar to those encountered in the context of object recognition. If h is smooth and almost flat near its maximum, and the data pattern is 'dense', the maximum likelihood estimate tends to contain multiple responses to each true cluster point. This is precisely what we wish to avoid.

To see where multiple response occurs, first consider a 'blur-free' inhomogeneous Poisson model with intensity function

$$\lambda(t \mid \mathbf{x}) = \lambda \, 1\{t \in S(\mathbf{x})\},$$

where $S(\mathbf{x}) = \bigcup B(x_i, r)$ denotes the 'silhouette' of pattern $\mathbf{x}$. This model cannot distinguish between configurations with the same silhouette (compare this to the situation in Chapter 2). As another example consider the Matérn model and assume a single isolated disc. Now add a parent to this configuration close to the true center, such that all daughters are in the overlap. This will be more likely if

$$m \log 2 > \mu$$

or equivalently $m > 1.4\ \mu$. Since the variance in the number of points is μ, this will happen frequently. If discs are overlapping, some points contribute a factor $\log 1.5$ instead of $\log 2$ and the effect is even stronger. Background scatter noise or incorrect parameter estimates may also cause multiple response.

As in Chapter 4 we introduce a nearest-neighbour Markov model to penalise scenes containing too many parents close together. Doing so, the MAP equations are formally given by (4.1.2)

$$\hat{\mathbf{x}} = \operatorname{argmax} f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}).$$

The difference is that f is the density of a point process instead of a pixel model.

Throughout the remainder of this session assume the inhomogeneous Poisson model (5.2.5). The MAP equations (4.1.2) cannot be solved explicitly due to the high dimensionality of the space.

5.4.1 Deterministic algorithms

The algorithms of Chapter 4 can be modified for use in this context. Typically, the observation window is two dimensional, and can easily be discretised and scanned.

The criterion for transitions are based on the posterior likelihood ratios

$$\frac{p(\mathbf{x} \cup \{u\} \mid \mathbf{y})}{p(\mathbf{x} \mid \mathbf{y})} = \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \prod_{j=1}^{m} \left[1 + \frac{h(y_j \mid u)}{\sum_{i=1}^{n} h(y_j \mid x_i)} \right]$$

$$\cdot \exp\left\{ - \int_T h(t \mid u)\, d\mu(t) \right\} \tag{5.4.8}$$

and

$$\frac{p(\mathbf{x} \setminus \{x_i\} \mid \mathbf{y})}{p(\mathbf{x} \mid \mathbf{y})} = \frac{p(\mathbf{x} \setminus \{x_i\})}{p(\mathbf{x})} \prod_{j=1}^{m} \left[1 - \frac{h(y_j \mid x_i)}{\sum_{k=1}^{n} h(y_j \mid x_k)} \right] \cdot \exp\left\{ \int_T h(t \mid x_i)\, d\mu(t) \right\}. \tag{5.4.9}$$

Algorithm 9 *Apply any of Algorithms 1–4 for the transition criteria above.*

The ratios (5.4.8) and (5.4.9) are easy to compute if p is a nearest-neighbour Markov point process (Definition 3). The term involving the data pattern is comparable to the Hough transform, in the sense that each point y_j votes with variable strength for a cluster centre at point u.

5.4.2 Stochastic algorithms

The posterior distribution for the parent pattern is too complicated to sample directly. Instead we construct a spatial birth-and-death process as in Section 4.4. An important feature is that the cluster models described in Section 5.2 all assume at least one parent. This causes difficulty, since the class $K = \{\mathbf{x} : f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) > 0\}$ ought to be *hereditary*. A similar remark holds for zero-likelihood configurations, as found e.g. in the Matérn cluster process (Model 5). To overcome these problems we propose to introduce a background stationary Poisson noise process with (small) intensity ϵ. An alternative in the absence of other zero-likelihood configurations is to set $f(\mathbf{y} \mid \emptyset) = \epsilon^m \exp[\mu(T)(1-\epsilon)]$. The resulting superposition is still an inhomogeneous Poisson process for any parent configuration.

Recalling the definition of temperature modified posterior distributions

$$p_H(\mathbf{x} \mid \mathbf{y}) \propto \{ f(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}) \}^H,$$

we want to construct a spatial birth-and-death process with rates given in (4.4.6) and (4.4.7) converging weakly to $p_H(\cdot \mid \mathbf{y})$.

Lemma 28 *Let* $\mathbf{y}$ *and* $H > 0$ *be fixed and assume an independent inhomogeneous Poisson cluster model superimposed on a background Poisson process of constant rate* ϵ. *Moreover let the prior model be a nearest-neighbour Markov process* $p(\cdot)$ *with uniformly bounded likelihood ratios*

$$\frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \leq \beta < \infty$$

If $0 \leq h(\cdot \mid \cdot) \leq h^* < \infty$, *then there exists a unique spatial birth-and-death process for which (4.4.6) and (4.4.7) are the transition rates. The process has unique equilibrium distribution* $p_H(\cdot \mid \mathbf{y})$ *and it converges in distribution to* $p_H(\cdot \mid \mathbf{y})$ *from any initial state.*

Proof: If $p(\mathbf{x}) > 0$, $b_H(\mathbf{x}, u) =$

$$\left(\prod_{j=1}^{m} \left[1 + \frac{h(y_j \mid u)}{\lambda(y_j \mid \mathbf{x})} \right] \exp\left[- \int_T h(t \mid u)\, d\mu(t) \right] \frac{p(\mathbf{x} \cup \{u\})}{p(\mathbf{x})} \right)^{k/H}$$

$$\leq \left(1 + \frac{h^*}{\epsilon} \right)^{mk/H} \beta^{k/H} =: \kappa > 0$$

and $D_H(\mathbf{x} \setminus \{u\}, u) =$

$$\left(\prod_{j=1}^{m} \left[1 + \frac{h(y_j \mid u)}{\lambda(y_j \mid \mathbf{x} \setminus \{u\})} \right] \exp\left[- \int_T h(t \mid u)\, d\mu(t) \right] \frac{p(\mathbf{x})}{p(\mathbf{x} \setminus \{u\})} \right)^{\frac{k-1}{H}}$$

$$\geq \left(1 + \frac{h^*}{\epsilon} \right)^{\frac{m(k-1)}{H}} \beta^{(k-1)/H} =: \delta > 0.$$

Once these bounds have been obtained, the proof follows the lines of the proof of Lemma 15. □

Clearly other sampling methods [44] could be applied too. Parameters ϕ in f and ψ in p can be estimated in advance or during iteration as in the following algorithm.

Algorithm 10 *Alternate estimation and sampling steps as follows*

1. *obtain an initial estimate* $\tilde{\mathbf{x}}$ *of the true pattern, with guesses for* ϕ *and* ψ *if necessary;*

2. *estimate* ϕ *by maximising* $f(\mathbf{y}|\tilde{\mathbf{x}}; \phi)$; *optionally, estimate* ψ *by the maximum pseudo-likelihood method;*

3. *sample from the posterior distribution for cluster centres given the current estimates* $\hat{\phi}$ *and* $\hat{\psi}$, *leading to a new estimate* $\tilde{\mathbf{x}}$, *and return to step 2.*

More generally, step 3 can be replaced by low temperature sampling and a sequence of these can be combined in a simulated annealing algorithm, as in Section 4.6. Furthermore, one can use Algorithm 10 to estimate functionals of the posterior, such as the distribution of the number of cluster centres, the probability that there is no cluster in a particular region, and the first-order intensity of cluster positions.

An alternative method is to specify a Gibbs sampler, by simply replacing the second step in Algorithm 10 by conditional sampling of the forward and prior parameters. The situation can be compared to the difference between profile and marginal likelihoods in statistical inference. Note that the Gibbs sampler requires a set of prior distributions, one for each parameter. For a Neyman-Scott model conditioned on m data points a (different) version of the Gibbs sampler was pursued by Lawson [66]. In that work, however, it was attempted to simulate $n(\mathbf{x})$ and $\mathbf{x}$ separately, necessitating severe approximations [67]. Here we treat $n(\mathbf{x})$ and $\mathbf{x}$ simultaneously as a realisation of a spatial point process.

5.5. EXAMPLE

Figure 5.1 shows the locations of 62 redwood seedlings in a square of side approximately 23 m. The data was extracted by Ripley [98] from a larger data set in Strauss [112]. The K-function for this data is given in [98] and suggests aggregation. As noted by Strauss this is caused by the presence of stumps known to exist in the plot, but whose position has not been recorded.

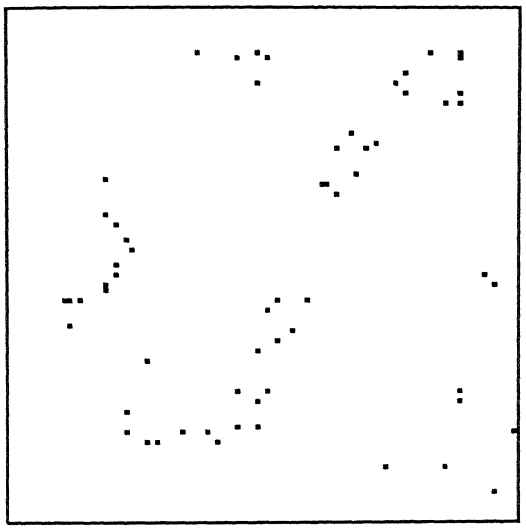


Figure 5.1: Positions of 62 redwood seedlings in a unit square (Ripley 1977).

Previous analyses of this data set include Strauss' who fitted a model later criticised by Kelly and Ripley [60]. Ripley [98] rejected the Poisson hypothesis and noted that there is both clustering and inhibition between clusters. Diggle [31] fitted a Poisson cluster process of Thomas type and reported least squares estimates for the parent intensity and sigma of (25.6 , .042). A goodness of fit test showed adequate fit, but from a biological point of view, a mean number of 26 stumps is implausible. In [30] Diggle fitted a Poisson cluster process of Matern type with similar results (radius .098).

None of the above have looked at cluster centre location. This was first studied by Lawson [66] who fitted a Poisson Thomas cluster process and reported 16 parents.

5.5.1 Model

Following [31, 66] we assume a (modified) Thomas model. Thus, the number of daughters per parent is Poisson and seedlings follow a radially

symmetric Gaussian distribution around their ancestor. In contrast to the aforementioned work, additionally a Strauss prior (3.1.4) with strict inhibition ($0 < \gamma < 1$) is introduced.

Lemma 29 *The maximum likelihood estimators for the model parameters are the solutions of*

$$\hat{\mu} = m / \int_T \frac{1}{2\pi\hat{\sigma}^2} \sum_{i=1}^{n} \exp\left[-\frac{1}{2\hat{\sigma}^2} \, ||t - x_i||^2 \right] d\mu(t)$$

and

$$\frac{1}{m} \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\hat{\sigma}^2} \, ||y_j - x_i||^2 \right] ||y_j - x_i||^2}{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\hat{\sigma}^2} \, ||y_j - x_i||^2 \right]} \right) =$$

$$\frac{\int_T \sum_{i=1}^{n} \exp\left[-\frac{1}{2\hat{\sigma}^2} \, ||t - x_i||^2 \right] ||t - x_i||^2 \, d\mu(t)}{\int_T \sum_{i=1}^{n} \exp\left[-\frac{1}{2\hat{\sigma}^2} \, ||t - x_i||^2 \right] d\mu(t)}.$$

Note that there is no good estimate for the background noise parameter ϵ. This is because the likelihood only depends on ϵ through the data pattern, and its estimate is just the number of daughters per unit volume (biased positively). The pseudolikelihood estimators for the Strauss model are well-known [100, Chapter 4, page 53].

Proof: The score functions are

$$\frac{\partial}{\partial \mu} L(\mathbf{x} \, ; \, \mathbf{y}) = \frac{m}{\mu} - \int_T \frac{1}{2\pi\sigma^2} \sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2} \, ||t - x_i||^2 \right] d\mu(t)$$

and

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} L(\mathbf{x} \, ; \, \mathbf{y}) &= \sum_{j=1}^{m} \frac{1}{\lambda(y_j \mid \mathbf{x})} \sum_{i=1}^{n} h(y_j - x_i) \left[-\frac{1}{\sigma^2} + \frac{||y_j - x_i||^2}{2\sigma^4} \right] \\
&- \int_T \sum_{i=1}^{n} h(t - x_i) \left[\frac{-1}{\sigma^2} + \frac{||t - x_i||^2}{2\sigma^4} \right] d\mu(t) \\
&= \frac{-m}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} h(y_j - x_i) \, ||y_j - x_i||^2}{\sum_{i=1}^{n} h(y_j - x_i)} \\
&+ \frac{1}{\sigma^2} \int_T \sum_{i=1}^{n} h(t - x_i) d\mu(t) \\
&- \frac{1}{2\sigma^4} \int_T \sum_{i=1}^{n} h(t - x_i) \, ||t - x_i||^2 \, d\mu(t).
\end{aligned}$$

Equating to zero completes the proof. □

With regard to the sampling step, note that $h \leq \frac{\mu}{2\pi\sigma^2}$ and $\frac{p(\mathbf{x}\cup\{u\})}{p(\mathbf{x})} \leq \beta$. Hence Lemma 15 applies.

5.5.2 Analysis

We analysed the redwood data using a Strauss prior with interaction distance .084 [31] and $\log\beta = \log\gamma = -10$. Throughout, a constant death rate spatial birth-and-death process was used. The initial parameter vales were $\mu = 7$ and $\sigma = .042$ and the initial list of cluster centres empty. Running the birth-and-death process for 2 time units, the maximum likelihood estimates were $\mu = 6.5$ and $\sigma = .05$. The posterior intensity for these parameter values, estimated over 50 time units is shown in Figure 5.3 and yielded parameter estimates $\mu = 6.4$ and $\sigma = .05$. For reasons of clarity, here *black* corresponds to high values. A sample taken after 2 time units is shown in Figure 5.2.

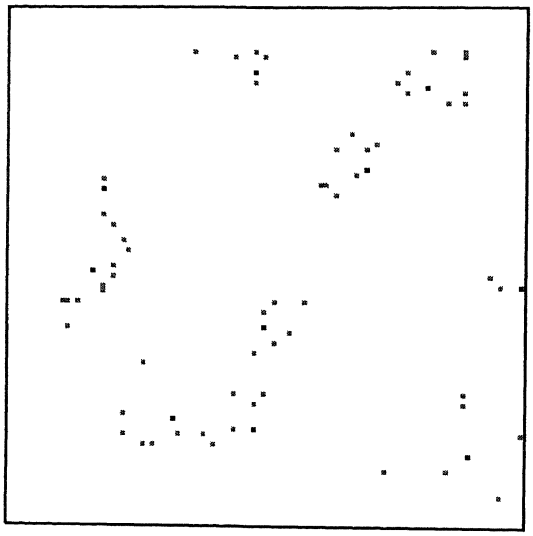


Figure 5.2: Realisation from the posterior distribution taken after 2 time units (black), for a Thomas model with $\mu = 6.5$, $\sigma = .05$ and a Strauss prior with $\log\beta = \log\gamma = -10$, $r = .084$. The data is displayed in grey.

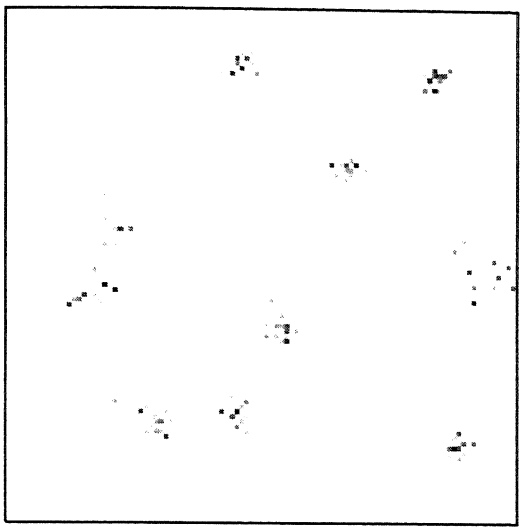


Figure 5.3: Posterior intensity of redwood seedlings (Ripley) estimated over 50 time units, for a Thomas model with $\mu = 6.5$, $\sigma = .05$ and a Strauss prior with $\log \beta = \log \gamma = -10$, $r = .084$.

Surprisingly, although the redwood data as extracted by Ripley [98] appears frequently in the spatial statistics literature, to the best of our knowledge the full data set [112] has not been reanalysed before. Thus we scanned region II of [112, Figure 1, p. 474], the roughly triangular area containing almost all the redwood stumps (Figure 5.4). The point coordinates on a range from 0 to 200 are listed in the Appendix.

We assumed a Thomas cluster model, a Strauss prior and computed the posterior intensity using the same (rescaled) model parameters as for the pattern in Figure 5.1. The estimate over 50 time units of the constant death rate Markov process (4.4.6) and (refbd:death) is displayed in Figure 5.5 while a typical realisation sampled at time 2 is shown in Figure 5.6.

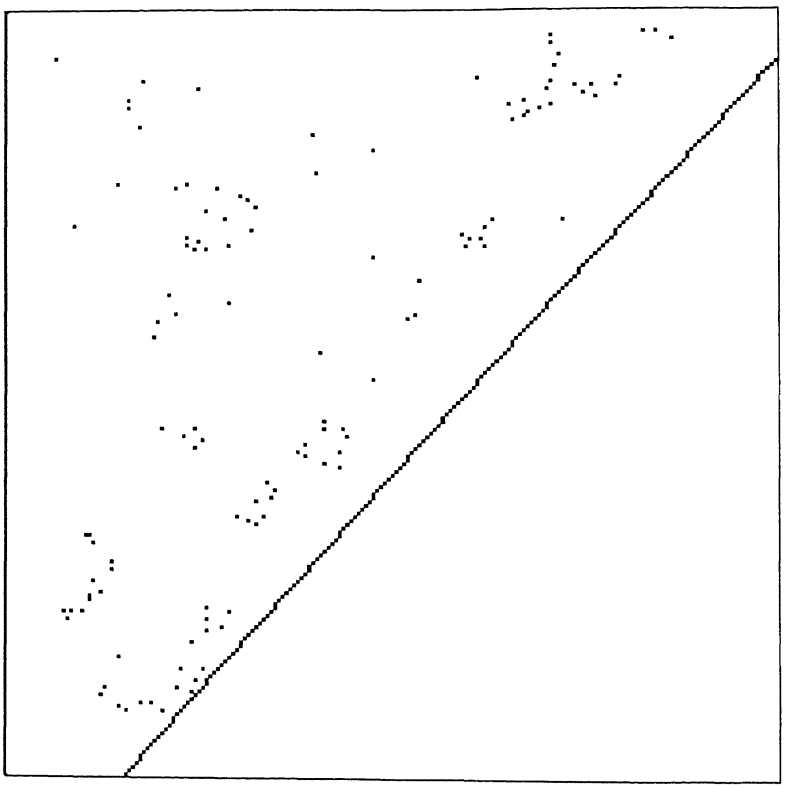


Figure 5.4: Positions of 123 redwood seedlings in a subset of the unit square (Strauss 1975).

Note that the results agree on the intersection of Figure 5.1 and Figure 5.4 and provide a plausible description of the clustered pattern. In particular, the Markov prior successfully combats the 'overestimation' (from a biological point of view) of the number of cluster centres.

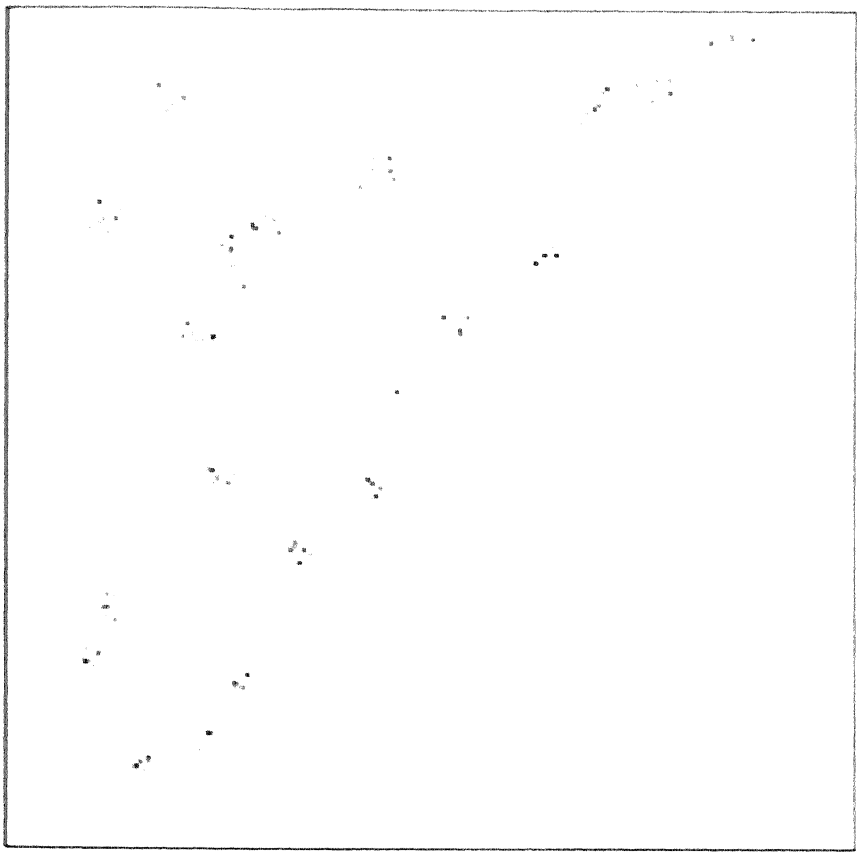


Figure 5.5: Posterior intensity of redwood seedlings in region II (Strauss) estimated over 50 time units, for a Thomas model with $\mu = 6.5$, $\sigma = .025$ and a Strauss prior with $\log \beta = \log \gamma = -10$, $r = .042$.

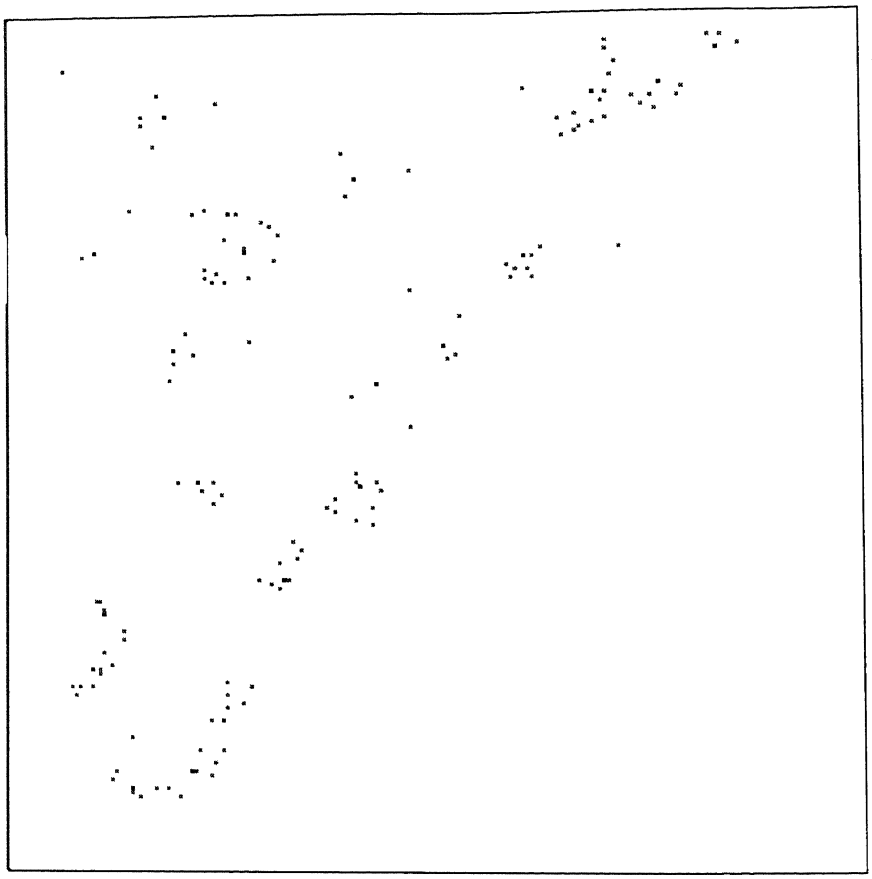


Figure 5.6: Realisation from the posterior distribution taken after 2 time units (black), for a Thomas model with $\mu = 6.5$, $\sigma = .025$ and a Strauss prior with $\log \beta = \log \gamma = -10$, $r = .042$. The data is displayed in grey.

5.6. Offspring Labelling

Apart from the cluster locations, the assignment of daughters to clusters is also of interest. Sibling information can be incorporated into the algorithm as auxiliary variables [18].

Formally, conditional on the data pattern $\mathbf{y}$, introduce the sibling set $\mathbf{z} = \{I_1, \ldots, I_k\}$, $k \in \{1, \ldots, m = n(\mathbf{y})\}$, where the sets $I_1, \ldots, I_k$ form an unordered partition of the data pattern. Points belonging to the same set I_j are called *siblings*, i.e. are offspring of the same parent. We need the following additional notation. Write $\phi_n(\mathbf{z})$ for the set of all 'label assignments'

$$\phi : \{1, \ldots, m\} \to \{1, \ldots, n\}$$

in agreement with the sibling variables $\mathbf{z}$, that is the members of $\mathbf{z}$ coincide with the sets $\phi^{-1}(i)$, $i = 1, \ldots n$.

In the (Boolean model) marked point process interpretation (Section 5.2.1) of cluster processes, write N^i for the offspring for parent i. We now proceed to derive the conditional distributions of $\mathbf{z}$ and $\mathbf{x}$.

Theorem 30 *For* $\mathbf{y} = \cup_{i=1}^n N^i$ *and* $n = n(\mathbf{x})$,

$$\mathbb{P}((N^i)_{i=1}^n \mid \mathbf{x}, \mathbf{y}) = \frac{\prod_{i=1}^n g(N^i \mid x_i)}{\sum_\phi \prod_{i=1}^n g(y_{\phi^{-1}(i)} \mid x_i)}.$$

Proof: Assume that $\mathbf{z}$ and $\mathbf{y}$ are compatible and write $m = n(\mathbf{y})$. Then we can identify the marks with a function $\phi : \{1, \ldots, m\} \to \{1, \ldots, n\}$. The Janossy densities with respect to $(\mu \times \pi)^n$

$$j_n(\{(x_1, N^1), \ldots, (x_n, N^n)\}) = e^{-\mu(T)}\, p(\{x_1, \ldots, x_n\}) \prod_{i=1}^n g(N^i \mid x_i)$$

can be rewritten in terms of $\mathbf{x}$, $\mathbf{y}$ and ϕ as

$$l_n(\mathbf{x}, \mathbf{y}, \phi) = e^{-\mu(T)}\, p(\{x_1, \ldots, x_n\}) \prod_{i=1}^n g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i)$$

with respect to $(\mu \times \pi)^n$ or equivalently by the family

$$l_{nm}(\mathbf{x}, \mathbf{y}, \phi) = e^{-\mu(T)}\, p(\{x_1, \ldots, x_n\}) \prod_{i=1}^n \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right]$$

with respect to μ^{n+m}. Therefore

$$\mathbb{P}((N^i)_{i=1}^n \mid \mathbf{x}, \mathbf{y}) = \frac{l_{nm}(\mathbf{x}, \mathbf{y}, \phi)}{e^{-\mu(T)} f(\mathbf{y} \mid \mathbf{x}) e^{-\mu(T)} p(\mathbf{x})}$$

$$= \frac{\prod_{i=1}^{n} \left[g(N^i \mid x_i) e^{-\mu(T)} \right]}{\sum_{\psi} \prod_{i=1}^{n} \left[g(\mathbf{y}_{\psi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right]}.$$

The last equality follows from Lemma 23. □

The following special case is of interest.

Corollary 31 *For Model (5.2.5),* $\mathbf{y} = \cup_{i=1}^{n} N^i$ *and* $n = n(\mathbf{x})$

$$\mathbb{P}((N^i)_{i=1}^{n} \mid \mathbf{x}, \mathbf{y}) = \frac{\prod_{i=1}^{n} \prod_{t \in N^i} h(t \mid x_i)}{\sum_{\phi} \prod_{j=1}^{m} h(y_j \mid x_{\phi(j)})}.$$

Proof: Apply the previous theorem and note that the term $H(x_i)$ cancels out. □

Corollary 31 can be used as a justification for the *nearest-parent assignment* proposed by Lawson [66]. Given a set of sites $\mathbf{y} = \{y_1, \ldots, y_m\}$ and a set of labels $\mathbf{x} = \{x_1, \ldots, x_n\}$, the task is to assign to each y_j a label from the collection $\mathbf{x}$. In Model (5.2.5) the labels are independent and hence a maximum likelihood classifier is

$$\hat{\phi}(j) = \max_{1,\ldots,n} h(y_j \mid x_i). \tag{5.6.10}$$

Lemma 32 *For Model (5.2.3), if* $h(\cdot)$ *is isotropic (i.e.* $h(t)$ *is a function of* $\|t\|$ *only) and decreasing in the norm of its argument, the maximum likelihood classifier (5.6.10) is the nearest parent classifier.*

Proof: Immediate from (5.6.10). □

An auxiliary variable sampler proceeds by alternatingly sampling from the conditional sibling distribution and from the conditional parent distribution. The required conditional probability laws are derived below.

Corollary 33 *For the setup of Theorem 30,*

$$\mathbb{P}(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = \frac{\sum_{\phi \in \phi_n(\mathbf{z})} \prod_{i=1}^{n} g(\mathbf{y}_{\phi^{-1}(i)} | x_i)}{\sum_{\phi} \prod_{i=1}^{n} g(\mathbf{y}_{\phi^{-1}(i)} | x_i).}$$

Proof: Use Theorem 30 and sum over the parent assignments. □

The conditional distribution of parents given a set of clusters is given in the following result.

Theorem 34 *For* $\mathbf{y} = \cup\mathbf{z}$, $n = n(\mathbf{x})$, *the conditional density with respect to the Poisson process of parents given an offspring division into clusters is*

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{y}) = \alpha\, p(\mathbf{x}) \sum_{\phi \in \phi_n(\mathbf{z})} \prod_{i=1}^{n} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right]$$

where $\alpha = \sum_{k=C(\mathbf{z})}^{\infty} \frac{e^{-\mu(T)}}{k!} \int_{T^k} p(\mathbf{x}) \sum_{\phi \in \phi_k(\mathbf{z})} \prod_{i=1}^{k} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right] d\mu(x_1) \ldots d\mu(x_k)$ *and we write* $C(\mathbf{z})$ *for the number of clusters in* $\mathbf{z}$.

Proof: The density with respect to the Poisson law is $e^{\mu(T)}$ times

$$\frac{\sum_{\phi \in \phi_n(\mathbf{z})} l_{nm}(\mathbf{x}, \mathbf{y}, \phi)}{\sum_{k=C(\mathbf{z})}^{\infty} \frac{1}{k!} \int_{T^k} \sum_{\phi \in \phi_k(\mathbf{z})} l_{nm}(\mathbf{x}, \mathbf{y}, \phi) d\mu(x_1) \ldots d\mu(x_k)}$$

for l_{nm} as in the proof of Theorem 30, yielding the formula. □

Offspring labelling can be combined with density estimation to assess the goodness-of-fit of the cluster distribution. Assuming Model (5.2.3), given a label assignment one has a set of vector differences $y_j - x_{\phi(j)}$ to which a density can be fitted. Comparison with the model h is an indication of the validity of the model.

5.6.1 Fixed number of points

Conditionally on the number of parents, one can consider *ordered* partitions instead of unordered. The distribution of offspring labels given data and parents is derived in Theorem 30. The conditional density of parents given the labelled data is treated below.

Theorem 35 *For any labelling* $\phi : \{1, \ldots, n(\mathbf{y})\} \to \{1, \ldots, n)\}$, *the conditional density of the* n *cluster centres with respect to* μ^n *is*

$$p_n(\mathbf{x} \mid \phi, \mathbf{y}) = \frac{p(\mathbf{x}) \prod_{i=1}^{n} g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i)}{\int_{T^n} p(\mathbf{x}) \prod_{i=1}^{n} g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) d\mu(x_1) \ldots d\mu(x_n)}$$

Proof: Since Janossy densities are not ordered, and μ^n is we have

$$p_n(\mathbf{x} \mid \phi, \mathbf{y}) = \frac{1}{n!} \frac{l_{nm}(\mathbf{x}, \mathbf{y}, \phi)}{\frac{1}{n!} \int_{T^n} l_{nm}(\mathbf{x}, \mathbf{y}, \phi) d\mu(x_1) \ldots d\mu(x_n)}$$

$$= \frac{\frac{1}{n!} e^{-\mu(T)} p(\mathbf{x}) \prod_{i=1}^{n} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right]}{\frac{1}{n!} e^{-\mu(T)} \int_{T^n} p(\mathbf{x}) \prod_{i=1}^{n} \left[g(\mathbf{y}_{\phi^{-1}(i)} \mid x_i) e^{-\mu(T)} \right] d\mu(x_1) \ldots d\mu(x_n)}$$

and the theorem follows. □

If the prior model is Poisson, the centres x_i are independent with density

$$\frac{g(N_i|x_i)}{\int_T g(N_i|t)d\mu(t)}$$

determined by their offspring.

A sample from the conditional distribution of n parents given the data is obtained by

- given a current parent configuration $\mathbf{x}$ assign labels according to $\mathbb{P}(\phi \mid \mathbf{x}, \mathbf{y})$;
- given a current labelling $\phi : \{1, \ldots, m\} \to \{1, \ldots n\}$ sample n parent positions from $p_n(\mathbf{x} \mid \phi, \mathbf{y})$.

The algorithm can be thought of as a stochastic version of the classical *k-means algorithm.* This deterministic technique iteratively assigns labels by means of the nearest parent classifier (Lemma 32) and then recomputes the parent positions as the centres of gravity of the current clusters. In the context of independent mixture models, see [24].

5.7. OTHER APPLICATIONS

We will briefly consider two other examples of identifying clusters in spatial patterns: fitting lines to point patterns and high-level edge detection.

5.7.1 Fitting curves to point patterns

In our first example, the data again consist of a point pattern $\mathbf{y} = \{y_1, \ldots, y_m\} \subseteq T$ with $T \subseteq \mathbb{R}^2$ bounded, but the points are believed to lie close to a curve or curves, possibly not contiguous, and the objective is to estimate the curves. An example is the image analysis task of joining a dot pattern into a curvilinear boundary. An application to spatial statistics is the identification of ancient roads or trade routes given information about the location of archaeological finds such as pottery or coins [111, p. 139], or the analysis of earthquake occurrences in relation to geographical fault patterns (Ogata, personal communication).

Let the curves be parametrised by a small number of real parameters and let U be the corresponding parameter space. The true curve pattern is then a configuration $\mathbf{x} = \{x_1, \ldots, x_n\} \subseteq U$. Again we can assume an independent cluster model in which each curve x_i gives rise to a point pattern $N^{(x_i)}$ and these daughter patterns are conditionally independent given $\mathbf{x}$.

It is no longer appropriate to assume that the daughter patterns $N^{(x_i)}$ are equivalent up to translation, since e.g. the expected number of points may well depend on the length of the curve. However we can assume $N^{(x_i)}$ is Poisson with intensity $h(\cdot \mid x_i)$, so that the observed point pattern is again a Cox process [111, p. 138]. Particular cases of interest would be the analogues of the Matérn and Thomas models in which distance to the cluster centre is replaced by distance to the nearest point on the cluster curve.

The treatment of this problem is formally equivalent to that in the previous Section, the only difference being that the cluster parents now belong to a general family of objects U instead of T. The general techniques of Section 5.4 apply.

5.7.2 High-level edge detection

Our final example is the high-level vision problem of identifying large scale edges in a scene using the output of a low-level edge detector. The 'data' $\mathbf{y}$ consist of a pattern of line segments and the objective is to cluster them around a small number of larger line segments.

Let W denote the set of possible outputs of the low-level edge detector. For example these may be line segments restricted to have unit length (= 1 pixel width) and orientation which is a multiple of 45 degrees. As

usual U denotes the space of objects we are looking for, which in this case are also line segments, but have unrestricted length and orientation.

Model $\mathbf{y}$ as a superposition of conditionally independent line segment processes $N^{(x_i)}$ associated with each true line segment x_i. Again it is not reasonable to suppose that all clusters are identically distributed up to translation, but we may assume they are all Poisson so that $\mathbf{y}$ is a Cox line segment process. Typically the expected number of line segments in $N^{(x_i)}$ will depend on the length of x_i.

The MLE and MAP estimators of $\mathbf{x}$ can then be determined using the techniques we have described above.

The possible benefits of a prior model for $\mathbf{x}$ include the ability to encourage long lines and continuity between lines, and to penalise lines that cross one another.

Chapter 6
Markov properties of cluster processes

The goal of this Chapter (a revision of Baddeley, Van Lieshout and Møller [11]) is to establish a theoretical link between Markov processes and cluster models. We prove that a Poisson cluster process with uniformly bounded clusters is a nearest-neighbour Markov point process with respect to the connected component relation introduced by Baddeley and Møller [12]; moreover, if a (fixed range) Markov or nearest-neighbour Markov point process is used as the parent process for a cluster process, and the clusters are uniformly bounded *and a.s. nonempty,* then the cluster process is again nearest-neighbour Markov. These results convincingly support Møller's claim [86] that nearest-neighbour Markov processes provide a rich class of models for (moderate) spatial clustering. The former result may also be helpful in understanding why statistical theory (as developed in Baddeley and Van Lieshout [10]) for Poisson cluster processes so closely resembles that for Markov point processes.

6.1. Setup

We consider finite, simple point processes X on a locally compact, complete separable metric space U (typically $\mathbb{R}^d$ or a compact subset). The metric is denoted d and the set of all configurations is identified with the space N^f of all simple, totally finite counting measures on U. For more details see Sections 3.1.2 and 3.2.1.

Let μ be a finite, non-atomic Borel measure on U. We will restrict attention to processes whose distributions are absolutely continuous with

respect to the law π of a Poisson process on U with intensity measure μ and denote the density by f.

Recall that f is called *hereditary* [103] if

$$f(\mathbf{x}) > 0 \quad \text{implies} \quad f(\mathbf{z}) > 0 \text{ for all } \mathbf{z} \subseteq \mathbf{x}. \tag{6.1.1}$$

We say that f is *hereditary excluding* $\emptyset$ if (6.1.1) holds except when $\mathbf{z} = \emptyset$.

6.1.1 Markov point processes

We are interested in Markov processes with respect to the following relation [103]. Define $u, v \in U$ to be *r-close* if

$$u \sim v \quad \text{iff} \quad 0 < d(u,v) \leq r$$

where d is the metric of U.

We also consider nearest-neighbour Markov point processes with respect to the following realisation dependent relation.

Definition 8 *For each* $\mathbf{x} \in N^f$, *define the* connected component relation *[12] between points of* $\mathbf{x}$ *by*

$$x_i \underset{\mathbf{x}}{\sim} x_j \quad \textit{iff} \quad x_i \sim z_1 \sim \cdots \sim z_n \sim x_j \quad \textit{for some } z_1, \ldots, z_n \in \mathbf{x}$$

In other words, two points of $\mathbf{x}$ are related under $\underset{\mathbf{x}}{\sim}$ if they are in the same connected component of the finite graph whose edges connect every pair of r-close points in $\mathbf{x}$.

The Hammersley-Clifford theorem [12, Theorem 4.13] specialises to the following result.

Lemma 36 *A point process* X *is nearest-neighbour Markov with respect to the connected component relation* $\underset{\mathbf{x}}{\sim}$ *iff*

$$f(\mathbf{x}) = \prod_{\textit{cliques } \mathbf{z} \subseteq \mathbf{x}} \varphi(\mathbf{z}) \tag{6.1.2}$$

where $\varphi(\cdot) \geq 0$ *is such that whenever* $\mathbf{z}$ *is a* $\underset{\mathbf{z}}{\sim}$*-clique with* $\varphi(\mathbf{z}) > 0$ *then* $\varphi(\mathbf{w}) > 0$ *for all* $\mathbf{w} \subseteq \mathbf{z}$.

Equivalently, X is nearest-neighbour Markov w.r.t. $\underset{\mathbf{x}}{\sim}$ iff

$$f(\mathbf{x}) = f(\emptyset) \prod_{k=1}^{K} \Phi(x_{D_k}) \tag{6.1.3}$$

where $x_{D_1}, \ldots, x_{D_K}$ are the connected components of $\mathbf{x}$ and $\Phi(\cdot) \geq 0$ is such that if $\mathbf{x}$ is a $\underset{\mathbf{x}}{\sim}$-clique and $\mathbf{z} \subseteq \mathbf{x}$ is a $\underset{\mathbf{z}}{\sim}$-clique then $\Phi(\mathbf{x}) > 0$ implies $\Phi(\mathbf{z}) > 0$.

6.1.2 Cluster processes

A cluster process is a two-stage point process in which each point ξ in an unobserved *parent* process $\mathbf{x}$ gives rise to a finite point process Z_ξ of *daughters*. It may or may not contain ξ. The data $\mathbf{y}$ is the superposition

$$\mathbf{y} = \cup_{x_i \in \mathbf{x}} Z_{x_i}.$$

It can be interpreted as a marked point process $\{(x_1, Z_{x_1}), \ldots, (x_n, Z_{x_n})\}$.

In the remainder of this Chapter we will assume independent clustering, that is the marks Z_{x_i} are independent.

For convenience we will sometimes consider densities with respect to $\rho = e^{\mu(U)}\pi$. Writing Q_ξ for the distribution of offspring of a parent at point ξ, we assume henceforth that

(A) Q_ξ is absolutely continuous with respect to ρ with density $q_\xi = g(\cdot \mid \xi)\, e^{-\mu(U)}$; equivalently, Q_ξ is absolutely continuous with respect to π with density $g(\cdot \mid \xi)$;

(B) $(\xi, \mathbf{z}) \mapsto q_\xi(\mathbf{z})$ is Borel measurable $U \times N^f \to \mathbb{R}_+$;

(C) **(uniform boundedness)** $Z_\xi \subseteq B(\xi, R)$ a.s. for some $R > 0$;

(D) q_ξ is hereditary excluding $\emptyset$.

Here $B(\xi, R)$ denotes the closed ball in the metric d with centre ξ and radius R.

Lemma 37 *The cluster process described above is absolutely continuous with respect to π with Radon-Nikodym derivative*

$$f(\mathbf{y}) = \mathbb{E}\left(e^{\mu(U)} \sum_{C_1, \ldots, C_{n(\mathbf{x})}} \prod_{i=1}^{n(\mathbf{x})} q_\xi(\mathbf{y}_{C_i}) \right) \tag{6.1.4}$$

where $\mathbb{E}$ denotes expectation with respect to the distribution of $\mathbf{x}$, and the sum is over all ordered partitions of $\mathbf{y} = \{y_1, \ldots, y_m\}$ into n subconfigurations $\mathbf{y}_{C_1}, \ldots, \mathbf{y}_{C_n}$ (allowing empty sets).

Proof: Conditionally given $\mathbf{x}$, the density with respect to ρ is

$$f(\mathbf{y} \mid \mathbf{x}) = e^{\mu(U)} \sum_{C_1, \ldots, C_n} \prod_{i=1}^{n} q_\xi(\mathbf{y}_{C_i})$$

by Lemma 23. Here the sum is over all ordered partitions of $\mathbf{y}$ into $n = n(\mathbf{x})$ clusters, allowing empty sets. The result follows by integrating over $\mathbf{x}$. □

6.2. STATEMENT OF RESULTS

Theorem 38 *Let* $\mathbf{x}$ *be a unit rate Poisson point process on* U *and* $\mathbf{y}$ *a cluster process with parent process* $\mathbf{x}$ *and clusters satisfying the assumptions (A)-(D) of Section 6.1.2. Then* $\mathbf{y}$ *is a nearest-neighbour Markov point process with respect to the connected component relation at distance* $2R$.

Any Poisson process with finite intensity measure λ can be treated in the same way by replacing the reference measure μ by $\lambda\mu$.

Consider the special case in which the offspring of each parent are Poisson. In $\mathbb{R}^d$, say, consider a parent process on a compact subset B and let μ be lebesgue measure restricted to the dilated set $B_{\oplus R}$. Assume that a parent at ξ has a Poisson number of offspring with mean ω, positioned i.i.d. with probability density $h(y_i - \xi)$ with respect to μ where h is supported on $B(0,R)$. For $\mathbf{y} \neq \emptyset$, the density is

$$f(\mathbf{y}) = \sum_{n=1}^{\infty} \{ \frac{e^{\mu(B_{\oplus R} \setminus B)} \omega^m e^{-\omega n}}{n!} \int_B \cdots \int_B \prod_{j=1}^{m} \left(\sum_{i=1}^{n} h(y_j - x_i) \right) d\mu(x_1) \ldots d\mu(x_n) \}.$$

This is easily (or via the proof of Theorem 38) factorised as

$$f(\mathbf{y}) = e^{\mu(B_{\oplus R})} \omega^m e^{-m\omega} e^{-\beta} \left[\sum_{C_1,\ldots,C_k} e^{\omega(m-k)} J(\mathbf{y}_{C_1}) \ldots J(\mathbf{y}_{C_k}) \right], \tag{6.2.5}$$

where $\beta = 1 - e^{-\omega}$, the sum is over all *unordered* partitions of $\mathbf{y}$ into disjoint *non-empty* subconfigurations, and

$$J(\mathbf{y}_C) = \int_B \prod_{y_j \in \mathbf{y}_C} h(y_j - \xi) d\mu(\xi). \tag{6.2.6}$$

Since $J(\mathbf{y}_C) = 0$ unless $\mathbf{y}_C$ is a $\underset{\mathbf{y}}{\sim}$–clique, the only nonzero terms in (6.2.5) are those for partitions which are refinements of the partition of $\mathbf{x}$ into connected components. Thus (6.2.5) factorises into terms associated with each connected component. According to (6.1.2) the process is nearest-neighbour Markov, provided the positivity condition stated below (6.1.2) is satisfied.

A special case is the Matérn cluster process in which $h \equiv$ constant on $B(0,R)$: then we have

$$J(\mathbf{y}_C) = \mu\left(\bigcap_{y_j \in \mathbf{y}_C} B(y_j, R) \cap B\right)$$

i.e. $J(\mathbf{y}_C)$ is the volume occupied within B by the intersection of the balls of radius R centred at the points of $\mathbf{y}_C$. In this case the positivity condition is clearly satisfied, so that the Matérn cluster process is nearest-neighbour Markov.

In Theorem 38, nearest-neighbour Markov cannot be replaced by Markov in the sense of Ripley and Kelly.

Counterexample 1 *A Neyman-Scott process with uniformly bounded clusters in general is* **not** *a Markov point process at any fixed range* $s < \infty$.

For, consider a configuration of three points y_1, y_2, y_3 such that $||y_1 - y_2|| < \min\{s, 2R\}$, $||y_2 - y_3|| < \min\{s, 2R\}$, but $||y_1 - y_3|| > \max\{s, 2R\}$. If f were a Markov function at range s then

$$f(\{y_1, y_2, y_3\})f(\{y_2\}) = f(\{y_1, y_2\})f(\{y_2, y_3\}).$$

Substituting (6.2.5) gives

$$[1 + e^{\omega}J(y_1, y_2) + e^{\omega}J(y_2, y_3)] = [1 + e^{\omega}J(y_1, y_2)]\,[1 + e^{\omega}J(y_2, y_3)]\,.$$

If we assume that $J(y_i, y_j) > 0$ whenever $||y_i - y_j|| < 2R$ (e.g. the Matérn model), this is clearly a contradiction. Hence f is not a Markov density in the Ripley–Kelly sense.

Next we turn to non-Poisson parent processes. The simplest generalisation is to take a Ripley-Kelly Markov parent process. However, in general, the cluster process constructed as in Section 6.1.2 is not nearest-neighbour Markov.

Heuristically, this is because parents without offspring can cause interaction by merging of disjoint $\underset{\mathbf{y}}{\sim}$ cliques. If we require each parent to have at least one daughter, the cluster process is nearest-neighbour Markov.

Theorem 39 *Let* $\mathbf{x}$ *be a Markov point process at range* r *and* $\mathbf{y}$ *the associated cluster process satisfying (A)-(D) of Section 6.1.2. If moreover*

(E) *the clusters are* **nonempty a.s.**

then $\mathbf{y}$ *is a nearest-neighbour Markov point process for the connected component relation at range* $2R + r$.

It is clear from the proof that the same result holds when the parent process is nearest-neighbour Markov. This suggests that a population of reproducing individuals can be modelled by the mechanism in Theorem 39. See also [63].

Theorem 40 *Let* $\mathbf{x}$ *be a nearest-neighbour Markov point process at range* r *and* $\mathbf{y}$ *the associated cluster process satisfying (A)-(D) of Section 6.1.2. If moreover*

(E) *the clusters are* **nonempty a.s.**

then $\mathbf{y}$ *is a nearest-neighbour Markov point process for the connected component relation at range* $2R + r$.

Proof: The following adaptation in the proof of Theorem 39 is needed.

Let $\epsilon : \{1, \ldots, m\} \to \{1, \ldots, n\}$ be a surjective mapping such that $d(y_j - x_{\epsilon(j)}) \leq R$ for all j. Suppose $\mathbf{z} \subseteq \mathbf{x}$ is a connected component with respect to $\underset{\mathbf{x}}{\sim}$ at range r such that $\epsilon^{-1}(\mathbf{z}) = \mathbf{w} \subseteq \mathbf{y}$. Then $\mathbf{w}$ is a connected component in $\mathbf{y}$ at range $2R + r$. To see this, choose any $w_1, w_2 \in \mathbf{w}$. There exists a $\mathbf{x}$–connecting path between the ϵ images $z_1, z_2 \in \mathbf{z}$, say

$$z_1 \sim x_{i_1} \cdots x_{i_m} \sim z_2.$$

As ϵ is surjective, x_{i_j} is the image of some $y_{k_j} \in \mathbf{y}$. Now apply the triangle inequality to obtain

$$d(w_1 - y_{k_1}) \leq d(w_1 - z_1) + d(z_1 - x_{i_1}) + d(x_{i_1} - y_{k_1})$$

$$\leq R + r + R = 2R + r.$$

Continuing in this manner proves the claim. □

Corollary 41 *Let* $\mathbf{x}$ *be a (nearest-neighbour) Markov point process at range* r *and* $\mathbf{y}$ *the associated cluster process obtained by translating each* $x_i \in \mathbf{x}$ *independently, where the displacements are absolutely continuous with respect to* μ *with density supported in* $B(0, R)$. *Then* $\mathbf{y}$ *is a nearest-neighbour Markov point process for the connected component relation at range* $2R + r$.

Proof: Apply Theorem 39 or Theorem 40. □

6.3. PROOFS

In this Section we give the proofs of the main results.

Proof: (Lemma 36)

Suppose that (6.1.2) holds. Let $\mathbf{x} \in N^f$, $\xi \in U$ and let $\mathbf{x}_{D_1}, \ldots, \mathbf{x}_{D_K}$ and $\mathbf{w} \cup \{\xi\}$ denote the connected components of $\mathbf{x} \cup \{\xi\}$. Then, if $\mathbf{x}_{D_{K+1}}, \ldots, \mathbf{x}_{D_L}$ are the connected components of $\mathbf{w}$, we have that $\mathbf{x}_{D_1}$, ..., $\mathbf{x}_{D_L}$ are the connected components of $\mathbf{x}$, and

$$f(\mathbf{x} \cup \{\xi\}) = \varphi(\emptyset) \left[\prod_{i=1}^{K} \prod_{\emptyset \neq \mathbf{z} \subseteq \mathbf{x}_{D_i}} \varphi(\mathbf{z}) \right] \prod_{\emptyset \neq \mathbf{z} \subseteq \mathbf{w} \cup \{\xi\}} \varphi(\mathbf{z})$$

while

$$f(\mathbf{x}) = \varphi(\emptyset) \left[\prod_{i=1}^{K} \prod_{\emptyset \neq \mathbf{z} \subseteq \mathbf{x}_{D_i}} \varphi(\mathbf{z}) \right] \prod_{j=K+1}^{L} \prod_{\emptyset \neq \mathbf{z} \subseteq \mathbf{x}_{D_j}} \varphi(\mathbf{z}).$$

Hence $f(\mathbf{x} \cup \{\xi\}) > 0$ implies $f(\mathbf{x}) > 0$ (as $\mathbf{z} \subseteq \mathbf{x}_{D_j}$ for $j > K$ implies that $\mathbf{z} \subseteq \mathbf{w}$) and $f(\mathbf{x} \cup \{\xi\})/f(\mathbf{x})$ satisfies the conditions of Definition 3. Thus X is nearest-neighbour Markov with respect to the connected component relation.

Conversely, suppose X is nearest-neighbour Markov. By the analogu of the Hammersley-Clifford theorem [12, Theorem 4.13],

$$f(\mathbf{x}) = \prod_{\mathbf{y} \subseteq \mathbf{x}} \varphi(\mathbf{y})^{\chi(\mathbf{y}|\mathbf{x})} \qquad \text{(taking } 0^0 = 0\text{)} \tag{6.3.7}$$

where $\chi(\mathbf{y} \mid \mathbf{x}) = 1$ if $\mathbf{y}$ is a $\underset{\mathbf{x}}{\sim}$-clique and 0 otherwise; and $\varphi : N^f \to \mathbb{R}_+$ satisfies

(I1) $\varphi(\mathbf{x}) > 0$ implies $\varphi(\mathbf{y}) > 0$ for all $\mathbf{y} \subseteq \mathbf{x}$

(I2) $\varphi(\mathbf{x}) > 0$ and $\varphi(N(\{\xi\} \mid \mathbf{x} \cup \{\xi\})) > 0$ imply $\varphi(\mathbf{x} \cup \{\xi\}) > 0$

where $N(\{\xi\} \mid \mathbf{x} \cup \{\xi\})$ denotes the neighbourhood of ξ in $\mathbf{x} \cup \{\xi\}$ (see Section 3.1.3). Note that, in the case of the connected component relation, $\xi \underset{\mathbf{y}}{\sim} \eta$ implies $\xi \underset{\mathbf{x}}{\sim} \eta$ for $\mathbf{x} \supseteq \mathbf{y}$, so that $\chi(\mathbf{y} \mid \mathbf{y}) = 1$ implies $\chi(\mathbf{y} \mid \mathbf{x})$ for any $\mathbf{x} \supseteq \mathbf{y}$.

To prove that (6.3.7) reduces to (6.1.2) we need to show that, if $\varphi(\mathbf{y}) > 0$ for all $\mathbf{y} \subseteq \mathbf{x}$ with $\chi(\mathbf{y} \mid \mathbf{x}) = 1$, then $\varphi(\mathbf{y}) > 0$ for all $\mathbf{y} \subseteq \mathbf{x}$.

To prove this, suppose $\mathbf{v}, \mathbf{w} \subseteq \mathbf{x}$ are disjoint connected subconfigurations of $\mathbf{x}$. If $\xi \in \mathbf{v}$ then $N(\{\xi\} \mid \mathbf{w} \cup \{\xi\}) = \{\xi\}$. By assumption $\varphi(\{\xi\}) > 0$, $\varphi(\mathbf{w}) > 0$ so (I2) gives $\varphi(\mathbf{w} \cup \{\xi\}) > 0$. Similarly, if $\{\xi, \eta\} \subseteq \mathbf{v}$ with $\eta \sim \xi$ then $N(\{\eta\} \mid \mathbf{w} \cup \{\xi, \eta\}) = \{\xi, \eta\}$, and by assumption $\varphi(\{\xi, \eta\}) > 0$, so (I2) gives $\varphi(\mathbf{w} \cup \{\xi, \eta\}) > 0$. Continuing in this way we obtain that $\varphi(\mathbf{y}) > 0$ for all $\mathbf{y} \subseteq \mathbf{x}$.

Hence if X is nearest-neighbour Markov then its density is of the form (6.1.2) where φ satisfies (I1) and hence the condition stated in the Lemma. □

Proof: (Theorem 38)
By (6.1.4), the density $f(\mathbf{y})$ of $\mathbf{y} \neq \emptyset$ with respect to π is

$$\sum_{n=1}^{\infty} \frac{e^{-\mu(U)}}{n!} \int_U \cdots \int_U e^{\mu(U)} \sum_{C_1,\ldots,C_n} \prod_{i=1}^{n} q_{x_i}(\mathbf{y}_{C_i})\, d\mu(x_1) \cdots d\mu(x_n)$$

$$= \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{C_1,\ldots,C_n} \prod_{i=1}^{n} \int_U q_\xi(\mathbf{y}_{C_i})\, d\mu(\xi); \tag{6.3.8}$$

here the inner sum is over all *ordered* partitions of $\mathbf{y}$ into n disjoint, possibly empty, sets. For $\mathbf{y} = \emptyset$,

$$f(\emptyset) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left[\int_U q_\xi(\emptyset)\, d\mu(\xi) \right]^n \tag{6.3.9}$$

$$= e^{\mu(U)-\beta} \tag{6.3.10}$$

if we define $\beta = \int_U (1 - q_\xi(\emptyset))\, d\mu(\xi)$. Note that since the parent process is Poisson, the number of non-empty clusters is Poisson distributed with mean β.

Now $q_\xi(\mathbf{z}) = 0$ whenever $\mathbf{z} \not\subseteq B(\xi, R)$; hence if $q_\xi(\mathbf{z}) \neq 0$ then all pairs of points in $\mathbf{z}$ are closer than $2R$ apart, i.e. $\mathbf{z}$ is a clique with respect to the finite range relation with distance $2R$. Hence the integral in (6.3.8) is nonzero only when the partition consists of $2R$-cliques.

For $\mathbf{y} \neq \emptyset$, let $\mathbf{y}_{D_1}, \ldots, \mathbf{y}_{D_{k_m}}$ be the connected components of $\mathbf{y}$ for the relation $\underset{\mathbf{y}}{\sim}$ with range $2R$. Then the integral in (6.3.8) is nonzero only when the partition is a refinement of $D_1, \ldots, D_{k_m}$. Let $C_1, \ldots, C_k$ be an *unordered* partition refining $D_1, \ldots, D_{k_m}$, consisting of *non-empty* sets. This contributes a term

$$\alpha \prod_{i=1}^{k} \int_U q_\xi(\mathbf{y}_{C_i})\, d\mu(\xi)$$

to the density. Since $\int_U q_\xi(\emptyset)\, d\mu(\xi) = \mu(U) - \beta$, the coefficient α is

$$\sum_{n=k}^{\infty} \frac{1}{n!} (\mu(U) - \beta)^{n-k}\, n(n-1)\cdots(n-k+1) = e^{\mu(U)-\beta}.$$

The class of all partitions that are refinements of $D_1, \ldots, D_{k_m}$ is the Cartesian product of the sets of partitions of each D_i. Hence, for $\mathbf{y} \neq \emptyset$,

$$f(\mathbf{y}) = e^{\mu(U)-\beta} \prod_{i=1}^{k_m} \Phi(\mathbf{y}_{D_i}) \tag{6.3.11}$$

where

$$\Phi(\mathbf{z}) = \sum_{k\geq 1} \sum_{\mathbf{z}_{C_1},\ldots,\mathbf{z}_{C_k}} \prod_{j=1}^{k} \int_U q_\xi(\mathbf{z}_{C_j})\, d\mu(\xi) \tag{6.3.12}$$

where $\mathbf{z}_{C_1}, \ldots, \mathbf{z}_{C_k}$ range over all unordered partitions of $\mathbf{z}$ into nonempty subconfigurations.

Since the offspring densities q_ξ are hereditary excluding $\emptyset$, clearly Φ is hereditary excluding $\emptyset$, and hence f is hereditary. According to (6.1.2) the density (6.3.11) is nearest-neighbour Markov with respect to the connected component relation at range $2R$.

□

Proof: (Theorem 39)

The density $p(\mathbf{x})$ of $\mathbf{x}$ can be factorised as in (6.1.2).

By (6.1.4), the density of $\mathbf{y}$ with respect to π is

$$f(\mathbf{y}) = e^{\mu(U)} \int_{N_f} p(\mathbf{x}) \sum_{C_1,\ldots,C_n(\mathbf{x})} \prod_{i=1}^{n(\mathbf{x})} q_{x_i}(\mathbf{y}_{C_i})\, d\pi(\mathbf{x})$$

where the inner sum ranges over all ordered partitions of $\mathbf{y}$ into disjoint, possibly empty subconfigurations.

First note that by assumption $q_\xi(\emptyset) = 0$, hence the integrand can be rewritten as

$$p(\mathbf{x}) \sum_{\epsilon} \prod_{i=1}^{n} q_{x_i}(\mathbf{y}_{\epsilon^{-1}(i)}) \tag{6.3.13}$$

where ϵ ranges over all surjective mappings of the points of $\mathbf{y}$ onto the points of $\mathbf{x}$, identified with mappings from $\{1,\ldots,m\}$ onto $\{1,\ldots,n\}$.

We can restrict attention to those ϵ such that

$$d(y_j - x_{\epsilon(j)}) \leq R \qquad \text{for all } j \tag{6.3.14}$$

since all other terms are zero. For such ϵ, if $\mathbf{z} \subseteq \mathbf{x}$ is an r-clique and $\epsilon^{-1}(\mathbf{z}) = \mathbf{w} \subseteq \mathbf{y}$, then $\mathbf{w}$ *must be a* $2R + r$*-clique.* To see this, take $y_i, y_j \in \mathbf{w}$ and apply the triangle inequality:

$$\begin{aligned} d(y_i - y_j) &\leq d(y_i - x_{\epsilon(i)}) + d(x_{\epsilon(i)} - x_{\epsilon(j)}) + d(x_{\epsilon(j)} - y_j) \\ &\leq R + r + R. \end{aligned}$$

Let $\mathbf{y}_{D_1}, \ldots, \mathbf{y}_{D_{k_m}}$ be the connected components of $\mathbf{y}$ at distance $2R+r$. Then we can rewrite (6.3.13) as

$$\prod_{\text{cliques } \mathbf{z}\subseteq\mathbf{x}} \varphi(\mathbf{z}) \sum_{\epsilon} \prod_{i=1}^{n} q_{x_i}(\mathbf{y}_{\epsilon^{-1}(i)}) =$$

$$\sum_{\epsilon} \left[\prod_{i=1}^{n} q_{x_i}(\mathbf{y}_{\epsilon^{-1}(i)}) \prod_{k=1}^{k_m} \prod_{\text{cliques } \mathbf{z}\subseteq\mathbf{x}:\epsilon^{-1}(\mathbf{z})\subseteq D_k} \varphi(\mathbf{z}) \right] =$$

$$\sum_{\epsilon} \prod_{k=1}^{k_m} \left[\prod_{i:\epsilon^{-1}(i)\subseteq D_k} q_{x_i}(\mathbf{y}_{\epsilon^{-1}(i)}) \prod_{\text{cliques } \mathbf{z}\subseteq\mathbf{x}:\epsilon^{-1}(\mathbf{z})\subseteq D_k} \varphi(\mathbf{z}) \right] . \quad (6.3.15)$$

Any ϵ of the type described above can be represented as a family of surjective mappings

$$\epsilon_k : D_k \to D'_k = \{i \mid d(x_i, y_j) \leq R \text{ for some } j \in D_i\}$$

automatically satisfying the norm condition (6.3.14). Note that $\mathbf{x}_{D'_k}$ form a disjoint partition of $\mathbf{x}$. Thus (6.3.15) is

$$\prod_{k=1}^{k_m} \sum_{\epsilon_k} \left[\prod_{i\in D'_k} q_{x_i}(\mathbf{y}_{\epsilon_k^{-1}(i)}) \prod_{\text{cliques } \mathbf{z}\subseteq\mathbf{x}_{D'_k}} \varphi(\mathbf{z}) \right] .$$

Integrating over $\mathbf{x}$ and exploiting the form of π yields $f(\mathbf{y}) =$

$$e^{\mu(U)} \prod_{k=1}^{k_m} \int_{N^f(\mathbf{y}_{D_k}\oplus R)} \sum_{\epsilon_k} \prod_{i=1}^{n(\mathbf{v})} q_{x_i}(\mathbf{y}_{\epsilon_k^{-1}(i)}) \prod_{\text{cliques } \mathbf{z}\subseteq\mathbf{v}} \varphi(\mathbf{z})\, d\pi_k(\mathbf{v})$$

where we write π_k for the Poisson process on the corresponding exponential space $N^f(\mathbf{y}_{D_k} \oplus R)$. Thus, f factorises as required by (6.1.2). The hereditary property follows as in the proof of the previous Theorem. □

REFERENCES

1. I.E. Abdou and W.K. Pratt. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, 67:753–763, 1979.

2. A.J. Baddeley. Errors in binary images and an L^p version of the Hausdorff metric. *Nieuw Archief voor Wiskunde*, 10:157–183, 1992.

3. A.J. Baddeley. Contribution to discussion of Grenander, U. and Miller, M.I.: Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B*, 1993.

4. A.J. Baddeley and R.D. Gill. Kaplan-Meier estimators for interpoint distance distributions of spatial point processes. Research Report 718, Mathematical Institute, University of Utrecht, march 1992.

5. A.J. Baddeley and M.N.M. van Lieshout. Recognition of overlapping objects using Markov spatial models. Technical Report BS-R9109, CWI, march 1991.

6. A.J. Baddeley and M.N.M. van Lieshout. ICM for object recognition. In Yadolah Dodge and Joe Whittaker, editors, *Computational statistics*, volume 2, pages 271–286, Heidelberg-New York, 1992. Physica/Springer.

7. A.J. Baddeley and M.N.M. van Lieshout. Object recognition using Markov spatial processes. In *Proceedings 11th IAPR International Conference on Pattern Recognition*, pages B 136–139, Los Alamitos, California, 1992. IEEE Computer Society Press.

8. A.J. Baddeley and M.N.M. van Lieshout. Stochastic geometry models in high-level vison. In K. Mardia and G.K. Kanji, editors, *Statistics and images*, *Journal of Applied Statistics*, 20:233–258, 1983.

9. A.J. Baddeley and M.N.M. van Lieshout. Area-interaction point processes. Technical Report BS9318, CWI, november 1993. Submitted to *Annals of the Institute of Statistical Mathematics*.

10. A.J. Baddeley and M.N.M. van Lieshout. A nonparametric measure of spatial interaction in point patterns. Technical Report WS-417, Faculteit der Wiskunde en Informatica, Vrije Universiteit Amsterdam, december 1993.

11. A.J. Baddeley, M.N.M. van Lieshout, and J. Møller. Markov properties of cluster processes. Research Report 278, Department of Theoretical Statistics, University of Aarhus, 1994. Submitted to *Probability Theory and Related Fields*.

12. A.J. Baddeley and J. Møller. Nearest-neighbour Markov point processes and random sets. *International Statistical Review*, 57:89–121, 1989.

13. D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13:111–122, 1981.

14. M. Baudin. Note on the determination of cluster centres from a realization of a multidimensional Poisson cluster process. *Journal of Applied Probability*, 20:136–143, 1983.

15. J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

16. J. Besag. Some methods of statistical analyisis for spatial data. *Bulletin of the International Statistical Institute*, 47:77–92, 1977.

17. J. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.

18. J. Besag and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, B 55:25–37, 1993.

19. J. Besag, R. Milne and S. Zachary. Point process limits of lattice processes. *Journal of Applied Probability*, 19, 1982.

20. J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154:143–155, 1991.

21. J.A.K. Blokland, A.M. Vossepoel, A.R. Bakker and E.K.J Pauwels. Automatic assignment of elliptical ROIs: First results in planar scintigrams of the left ventricle. *European Journal of Nuclear Medicine*, 15:87–92, 1989.

22. P Brodatz. *Texture: a photographic album for artists and designers.* Dover, New York, 1966.

23. C.M. Brown. Inherent bias and noise in the Hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:493–505, 1983.

24. G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

25. M. Cohen and G.T. Toussaint. On the detection of structures in noisy pictures. *Pattern Recognition*, 9:95–98, 1977.

26. D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes.* Springer Verlag, New York, 1988.

27. E.R. Davies. Improved localization in a generalized Hough scheme for the detection of straight edges. *Image and Vision Computing*, 5:279–286, 1987.

28. E.R. Davies. A new framework for analyzing the properties of the generalized Hough transform. *Pattern Recognition Letters*, 6:1–8, 1987.

29. J. Dengler and M. Guckes. Estimating a global shape model for objects with badly-defined boundaries. In W. Förstner and S. Ruwiedel, editors, *Robust Computer Vision*, pages 122–136, Karlsruhe, 1992. Wichman.

30. P.J. Diggle. On parameter estimation for spatial point processes. *Journal of the Royal Statistical Society, Series B*, 40:178–181, 1978.

31. P.J. Diggle. *Statistical analysis of spatial point patterns.* Academic Press, London, 1983.

32. P.J. Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 153:349–362, 1990.

33. P.J. Diggle, T. Fiksel, Y. Ogata, D. Stoyan and M. Tanemura. On parameter estimation for pairwise interaction processes. *International Statistical Review*, 1993. to appear.

34. R.L. Dobrushin. Central limit theorem for non-stationary Markov chains I, II. *Theory of Probability and its Applications*, 1:65–80, 329–383, 1956.

35. R.O. Duda and P.E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15:11–15, 1972.

36. R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* John Wiley and Sons, New York, 1973.

37. T. Fiksel. Estimation of parametrized pair potentials of marked and non-marked Gibbsian point processes. *Elektronische Informationsverarbeitung und Kybernetik*, 20:270–278, 1984.

38. T. Fiksel. Estimation of interaction potentials of Gibbsian point processes. *Statistics*, 19:77–86, 1988.

39. D.J. Gates and M. Westcott. Clustering estimates for spatial point distributions with unstable potentials. *Annals of the Institute of Statistical Mathematics*, 38:123–135, 1986.

40. D. Geman. *Ecole d'ete de probabilites de Saint-Flour XVIII - 1988*, volume 1427 of *Lecture Notes in Mathematics*, chapter Random fields and inverse problems in imaging. Springer-Verlag, Berlin, 1990.

41. D. Geman, S. Geman, C. Graffigne and P. Dong. Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:609–628, 1990.

42. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

43. S. Geman and C.R. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24:1031–1043, 1986.

44. C.J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. Technical Report 260, University of Aarhus, 1993.

45. C.J. Geyer and E.A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B*, 54:657–699, 1992.

46. D.M. Greig, B.T. Porteous and A.H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, 51:271–279, 1989.

47. U. Grenander. *Lectures on Pattern Theory, Vol. 1: Pattern Synthesis.* Applied Mathematical Sciences vol. 18. Springer-Verlag, New York-Berlin, 1976.

48. U. Grenander. *Lectures on Pattern Theory, Vol. 2: Pattern Analysis.* Applied Mathematical Sciences vol. 24. Springer-Verlag, New York-Berlin, 1978.

49. U. Grenander. *Lectures on Pattern Theory, Vol. 3: Regular Structures.* Applied Mathematical Sciences vol. 33. Springer-Verlag, New York-Berlin, 1981.

50. H. Haario and E. Saksman. Simulated annealing process in general state space. *Advances in Applied Probability*, 23:886–893, 1991.

51. P. Hall. *Introduction to the theory of coverage processes.* John Wiley and Sons, New York, 1988.

52. P.R. Halmos. *Measure theory.* Springer-Verlag, New York, 1950.

53. W.D. Hamilton. Geometry for the selfish herd. *Journal of Theoretical Biology*, 31:295–311, 1971.

54. P.V.C. Hough. Method and means for recognizing complex patterns. US Patent 3069654, 1962.

55. D.J. Hunt, L.W. Nolte and W.H. Ruedger. Performance of the Hough transform and its relationship to statistical signal detection theory. *Computer Vision, Graphics and Image Processing*, 43:221–238, 1988.

56. J. Illingworth and J. Kittler. The adaptive Hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:690–698, 1987.

57. J. Illingworth and J. Kittler. A survey of the Hough transform. *Computer Vision, Graphics and Image Processing*, 44:87–116, 1988.

58. J.L. Jensen and J. Møller. Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, 1991.

59. O. Kallenberg. An informal guide to the theory of conditioning in point processes. *International Statistical Review*, 52:151–164, 1984.

60. F.P. Kelly and B.D. Ripley. On Strauss's model for clustering. *Biometrika*, 63:357–360, 1976.

61. W.S. Kendall. A spatial Markov property for nearest-neighbour Markov point processes. *Journal of Applied Probability*, 28:767–778, 1990.

62. C. Kimme, D. Ballard and J. Sklansky. Finding circles by an array of accumulators. *Communications of the ACM*, 18:120–122, 1975.

63. J.F.C. Kingman. Remarks on the spatial distribution of a reproducing population. *Journal of Applied Probability*, 14:577–583, 1977.

64. A.B. Lawson. Gibbs sampling putative pollution sources. Technical report, Dundee Institute of Technology, 1993. In preparation.

65. A.B. Lawson. On fitting non-stationary Markov point process models on GLIM. In Yadolah Dodge and Joe Whittaker, editors, *Computational statistics*, volume 1, pages 35–40, Heidelberg-New York, 1992. Physica/Springer.

66. A.B. Lawson. Discussion contribution on The Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, B 55:61–62, 1993.

67. A.B. Lawson. Rejection methods for spatial cluster processes. Technical report, Dundee Institute of Technology, 1993.

68. A.B. Lawson. Statistical models in spatial epidemiology: a review and proposal. Technical Report MACS 93/02, Dundee Institute of Technology, 1993.

69. A.B. Lawson. Gibbs sampling a spatial Cox process. Technical report, Dundee Institute of Technology, 1993.

70. A.B. Lawson, M.N.M. van Lieshout and A.J. Baddeley. Markov chain Monte Carlo methods for spatial cluster processes, 1993. In preparation.

71. M.N.M. van Lieshout. Noise removal and segmentation methods in statistical image analysis. Master's thesis, Vrije Universiteit Amsterdam, 1990.

72. M.N.M. van Lieshout. A Bayesian approach to object recognition. In U. Eckhardt, A. Hűbler, W. Nagel, and G. Werner, editors, *Geometrical problems of image processing*, volume 4 of *Research in Informatics*, pages 185–190, Berlin, 1991. Akademie Verlag. Proceedings Geobild'91.

73. M.N.M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. Technical Report BS9306, CWI, march 1993.

74. M.N.M. van Lieshout. Contribution to discussion of Grenander, U. and Miller, M.I.: Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B*, 1993.

75. M.N.M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Advances in Applied Probability*, 26, 1994.

76. J.-H. Lin, T. M. Sellke and E. J. Coyle. Adaptive stack filtering under the mean absolute error criterion. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38:938–954, 1990.

77. H.W. Lotwick and B.W. Silverman. Convergence of spatial birth-and-death processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 90:155–165, 1981.

78. P. Maragos. Optimal morphological approaches to image matching and object detection. In *Proceedings of the IEEE International Conference on Computer Vision 1988, Florida*, pages 695–699, 1988.

79. R.J. Marshall. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society, Series A*, 154:421–441, 1991.

80. Matérn, B. Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut*, 49:1-144, 1960.

81. G. Matheron. *Random sets and integral geometry.* John Wiley and Sons, New York, 1975.

82. M. Miller et al. Automated segmentation of biological shapes in electron microscopic autoradiography. In F. Davidson and J. Goutsias, editors, *25th Annual Conference on Information Sciences and Systems*, 1991.

83. J. Møller. On the rate of convergence of spatial birth-and-death processes. *Annals of the Institute of Statistical Mathematics*, 41:565–581, 1989.

84. R. Molina and B.D. Ripley. Using spatial models as priors in astronomical image analysis. *Journal of Applied Statistics*, 16:193–206, 1989.

85. J. Møller. Lecture at SPA93.

86. J. Møller. Markov chain Monte Carlo and spatial point processes. Manuscript in preparation, 1994.

87. R.A. Moyeed and A.J. Baddeley. Stochastic approximation of the MLE for a spatial point pattern. *Scandinavian Journal of Statistics*, 18:39–50, 1991.

88. P. Nacken. A metric for line segments. Technical Report BS-R9126, CWI, 1991.

89. Y. Ogata and M. Tanemura. Estimation for interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, B 33:315–338, 1981.

90. Y. Ogata and M. Tanemura. Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society, Series B*, 46:496–518, 1984.

91. Y. Ogata and M. Tanemura. Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns. *Annals of the Institute of Statistical Mathematics*, 41:583–600, 1989.

92. S. Openshaw, M.G. Charlton, C. Wymer and A.W. Craft. A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal on Geographical Information Systems*, 1:335–358, 1987.

93. E. Parzen. *Stochastic processes.* Holden-Day, San Francisco, 1962.

94. A. Penttinen. Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jyvaskyla Studies in Computer Science, Economics and Statistics*, 7:1–105, 1984.

95. W.K. Pratt. *Digital image processing.* John Wiley and Sons, New York, 1977.

96. C.J. Preston. *Random fields.* Springer Verlag, Berlin-Heidelberg-New York, 1976.

97. C.J. Preston. Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, 46:371–391, 1977.

98. B.D. Ripley. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:172–212, 1977.

99. B.D. Ripley. *Spatial statistics.* John Wiley and Sons, New York, 1981.

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

101. B.D. Ripley. Gibbsian interaction models. In D.A. Griffiths, editor, *Spatial statistics: past, present and future*, pages 1–19. Image, New York, 1989.

102. B.D. Ripley. *Statistics in the environmental and earth sciences*, chapter Stochastic models for the distribution of rock types in petroleum reservoirs. 1991.

103. B.D. Ripley and F.P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 15:188–192, 1977.

104. B.D. Ripley and A.I. Sutherland. Finding spiral structures in images of galaxies. *Philosophical Transactions of the Royal Society of London, Series A*, 332:477–485, 1990.

105. A. Rosenfeld and A.C. Kak. *Digital picture processing*, volume II. Academic Press, New York, second edition, 1982.

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

107. J. Serra, editor. *Image analysis and mathematical morphology, volume 2: Theoretical advances.* Academic Press, London, 1988.

108. S.D. Shapiro. Feature space transforms for curve detection. *Pattern Recognition*, 10:129–143, 1978.

109. A. Särkkä. On parameter estimation of Gibbs point processes through the pseudo-likelihood method. Technical Report 4, De-

partment of statistics, University of Jyväskylä, 1989.

110. A.G. Steenbeek. *CLIP: a C++ library for image processing*, 1992. Release 1.1.

111. D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. Springer-Verlag, Berlin, 1987.

112. D.J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.

113. B. Stroustrup. *The C++ programming language*. Addision-Wesley, Reading, Mass., 2 edition, 1991.

114. R. Takacs. Estimator for the pair-potential of a Gibbsian point process. Institutsbericht 238, Institut für Mathematik, Johannes Kepler Universität Linz, Austria, 1983.

115. R. Takacs. Estimator for the pair potential of a Gibbsian point process. *Statistics*, 17:429–433, 1986.

116. P.R. Thrift and S.M. Dunn. Approximating point set images by line segments using a variation of the Hough transform. *Computer Vision, Graphics and Image Processing*, 21:383–394, 1983.

117. E. Tomppo. Models and methods for analysing spatial patterns of trees. *Communicationes Instituti Forestalis Fenniae*, 138:1–65, 1986.

118. T.M. Van Veen and F.C.A. Groen. Discretization errors in the Hough transform. *Pattern Recognition*, 14:137–145, 1981.

119. B. Widom and J.S. Rowlinson. New model for the study of liquid-vapor phase transitions. *The Journal of Chemical Physics*, 52:1670–1684, 1970.

120. G. Winkler. An ergodic L^2 theorem for simulated annealing in Bayesian image reconstruction. *Journal of Applied Probability*, 28:779–791, 1990.

Appendix

Delta distance

Initial pattern	MLE (coordinatewise)	MAP (coordinatewise)
empty	.546	.404
true	.230	.099
shifted	.367	.335
Hough extrema	.314	.170

Table .1: Delta distance between the true pattern and the reconstruction obtained using coordinatewise ascent. The only transitions are births and deaths.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.210	.140
true	.235	.099
shifted	.260	.291
Hough extrema	.213	.182

Table .2: Delta distance between the true pattern and the reconstruction obtained using steepest ascent. The only transitions are births and deaths.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.093	.140
true	.000	.000
shifted	.192	.291
Hough extrema	.178	.182

Table .3: Optimal delta distance between the true pattern and intermediate reconstructions using steepest ascent. The only transitions are births and deaths.

Initial pattern	MLE (coordinatewise)	MAP (coordinatewise)
empty	.528	.207
true	.236	.132
shifted	.344	.132
Hough extrema	.335	.132

Table .4: Delta distance between the true pattern and the reconstruction obtained using coordinatewise ascent. The transitions are births, deaths and translations.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.205	.132
true	.205	.132
shifted	.210	.132
Hough extrema	.220	.132

Table .5: Delta distance between the true pattern and the reconstruction obtained using steepest ascent. The transitions are births, deaths and translations.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.093	.132
true	.000	.000
shifted	.049	.099
Hough extrema	.187	.132

Table .6: Optimal delta distance between the true pattern and intermediate reconstructions using steepest ascent. The transitions are births, deaths and translations.

Figure of merit

Initial pattern	MLE (coordinatewise)	MAP (coordinatewise)
empty	.627	.733
true	.883	.960
shifted	.794	.815
Hough extrema	.802	.929

Table .7: Figure of merit between the true pattern and the reconstruction obtained using coordinatewise ascent. The only transitions are births and deaths.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.900	.944
true	.879	.960
shifted	.865	.859
Hough extrema	.890	.926

Table .8: Figure of merit between the true pattern and the reconstruction obtained using steepest ascent. The only transitions are births and deaths.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.985	.944
true	1.00	1.00
shifted	.906	.860
Hough extrema	.907	.931

Table .9: Optimal figure of merit between the true pattern and intermediate reconstructions using steepest ascent. The only transitions are births and deaths.

Initial pattern	MLE (coordinatewise)	MAP (coordinatewise)
empty	.632	.908
true	.881	.948
shifted	.809	.948
Hough extrema	.787	.948

Table .10: Figure of merit between the true pattern and the reconstruction obtained using coordinatewise ascent. The transitions are births, deaths and translations.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.902	.948
true	.902	.948
shifted	.900	.948
Hough extrema	.890	.948

Table .11: Figure of merit between the true pattern and the reconstruction obtained using steepest ascent. The transitions are births, deaths and translations.

Initial pattern	MLE (steepest ascent)	MAP (steepest ascent)
empty	.985	.948
true	1.00	1.00
shifted	.993	.960
Hough extrema	.906	.948

Table .12: Optimal figure of merit between the true pattern and intermediate reconstructions using steepest ascent. The transitions are births, deaths and translations.

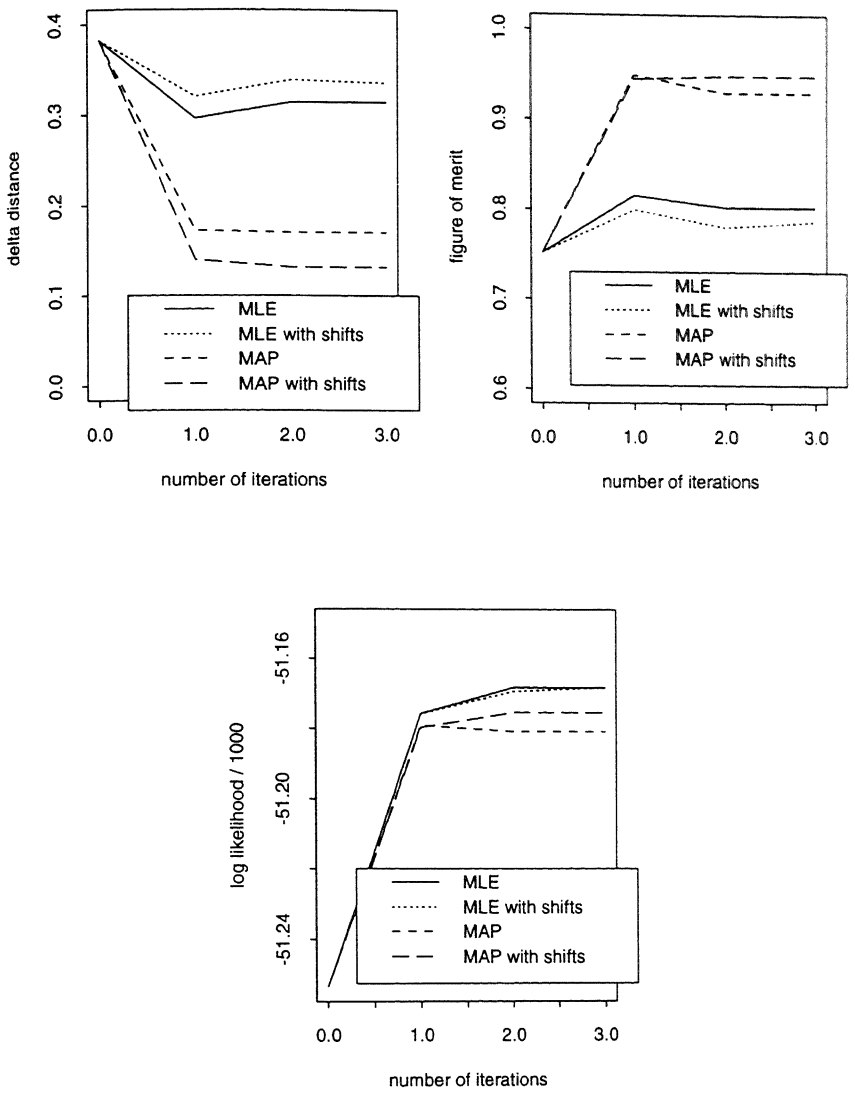
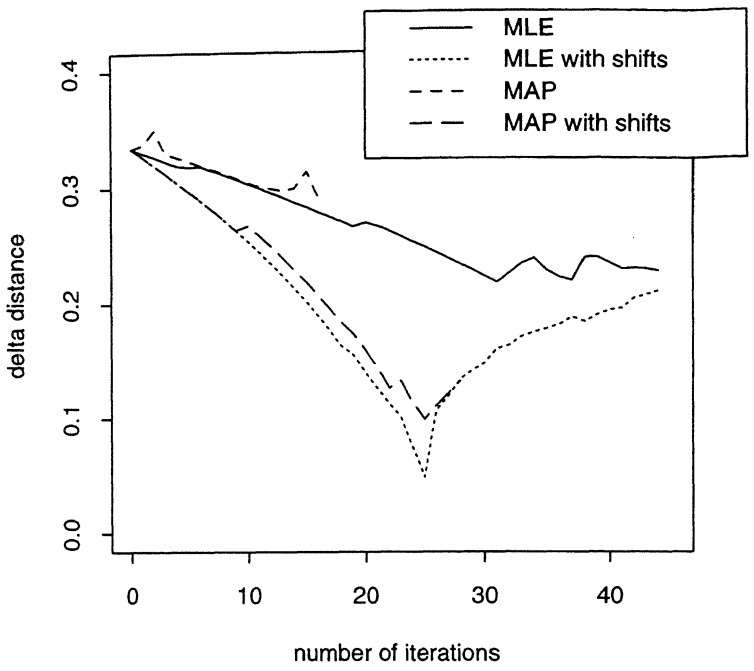


Figure .1: Reconstruction quality at successive steps of coordinatewise ascent with the local maxima of the conditional Hough transform as initial state.


MLE
MLE with shifts
MAP
MAP with shifts
delta distance
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
number of iterations

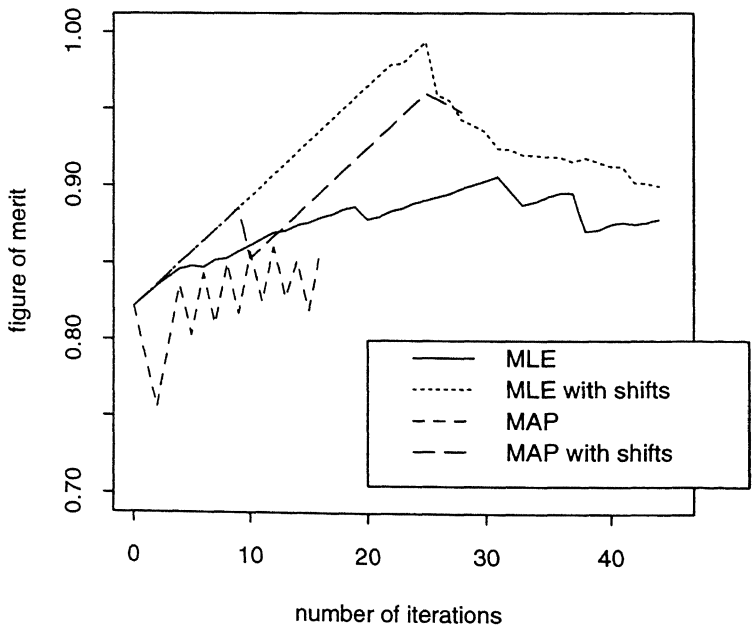


figure of merit
0.70
0.80
0.90
1.00
0
10
20
30
40
number of iterations
MLE
MLE with shifts
MAP
MAP with shifts

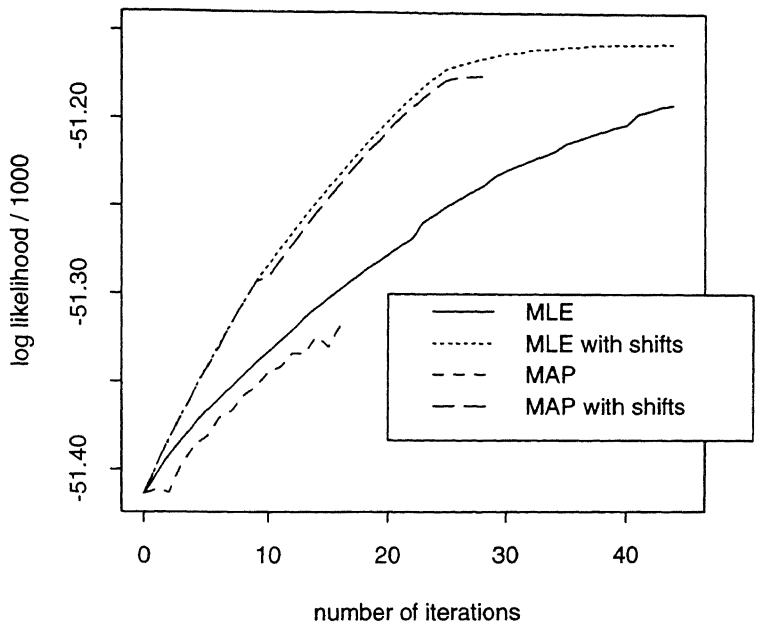


Figure .2: Reconstruction quality for the successive reconstructions obtained using steepest ascent starting with a diagonally translated copy of the true configuration.

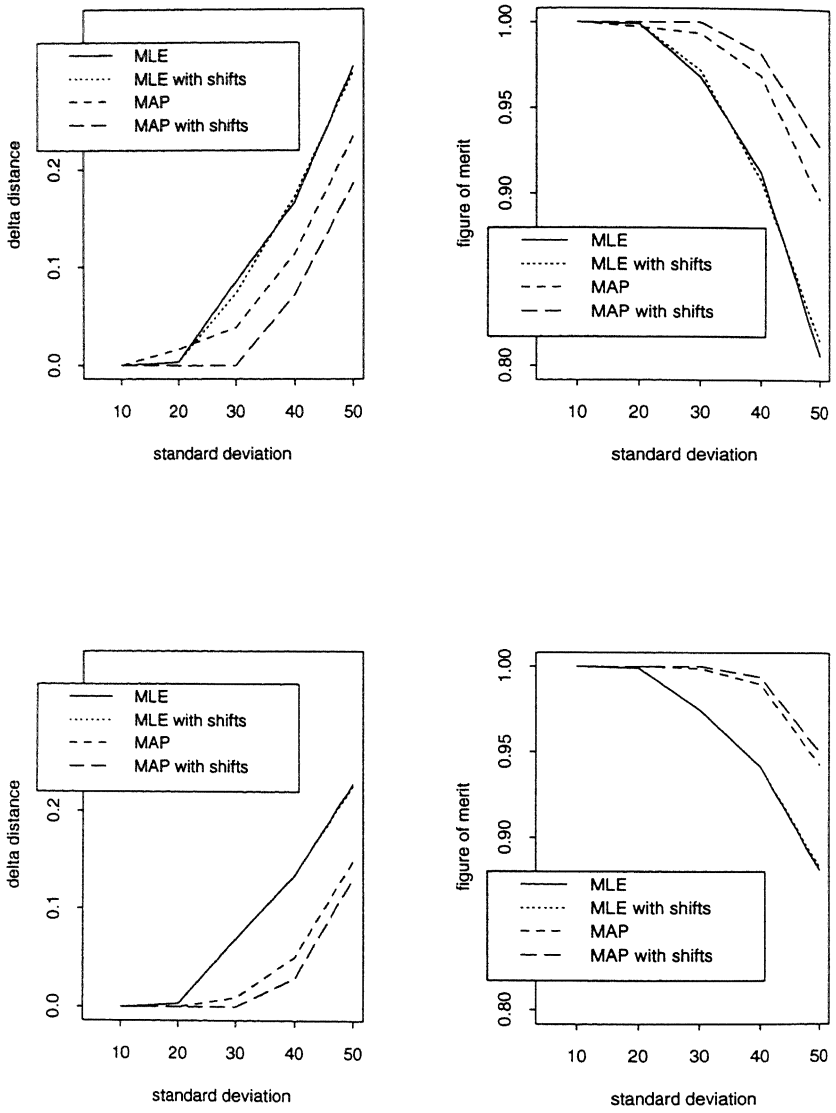


Figure .3: Average reconstruction quality for 10 independent realisations at different noise levels. Top: coordinatewise ascent; bottom: steepest ascent.

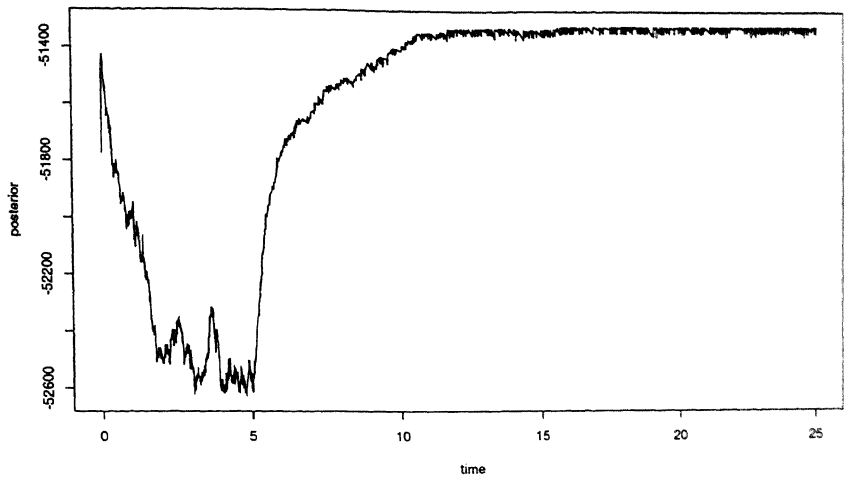


Figure .4: Log posterior likelihood against time for a geometric cooling schedule starting at $H = 4.0$ of rate .5 and a Strauss prior with $\beta = .0025$ and $\gamma = .25$

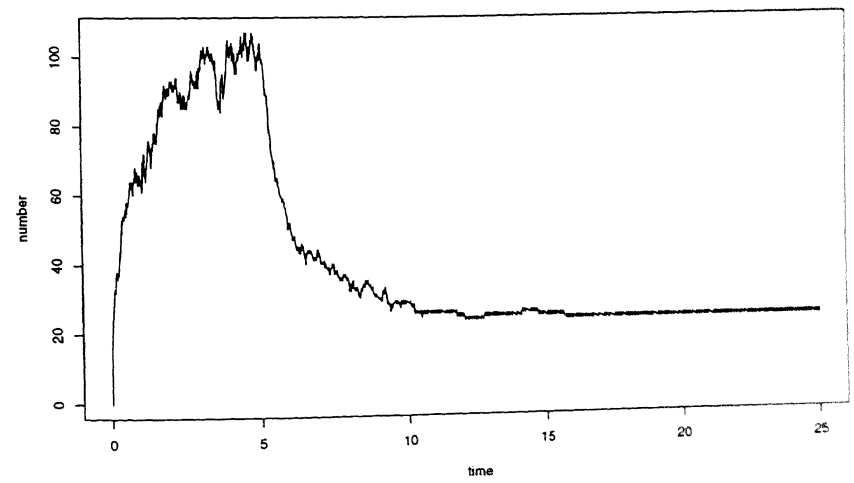


Figure .5: Number of objects against time for a geometric cooling schedule starting at $H = 4.0$ of rate .5 and a Strauss prior with $\beta = .0025$ and $\gamma = .25$

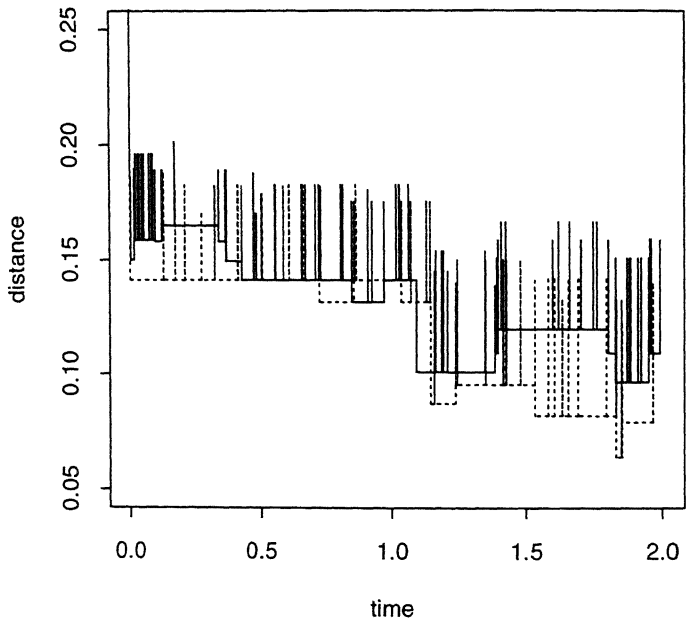


Figure .6: Delta-2 distance between reconstructed and true pattern as a function of time. The cutoff value is 4.

Coordinates of points in region II of Strauss' redwood data

167	5	170	5
144	6	174	7
144	8	146	11
13	12	145	14
125	17	161	17
37	18	144	18
150	19	154	19
160	19	52	20
143	20	152	21
155	22	33	23
137	23	133	24
144	24	33	25
141	25	138	26
137	27	134	28
36	30	82	32
98	36	83	42
30	45	49	45
46	46	57	46
63	48	65	49
67	51	54	52
59	54	129	54
147	54	18	56
127	56	66	57
121	58	49	59
123	59	126	59
52	60	49	61
60	61	122	61
127	61	51	62
54	62	98	64
110	70	44	74
60	76	46	79
109	79	107	80
41	81	40	85
84	89	98	96
85	107	42	109
51	109	85	109
90	109	48	111
91	111	53	112
80	113	51	114
78	115	89	115
80	116	85	118

89	119	70	123
72	125	71	127
67	128	62	132
69	132	65	133
67	134	21	137
22	137	23	139
28	144	28	146
23	149	25	152
22	153	22	154
54	156	15	157
17	157	20	157
60	157	16	159
54	159	58	161
54	162	50	16
30	169	47	172
53	172	51	175
26	177	46	177
50	178	25	179
36	181	39	181
30	182	32	183
42	183		

Index

Samenvatting

Het doel van dit proefschrift is te beargumenteren dat de stochastische meetkunde een rijke collectie modellen biedt, die toepasbaar zijn binnen de beeldanalyse en de ruimtelijke statistiek.

Objectherkenning is het deelgebied van de beeldanalyse dat zich bezighoudt met het bepalen of er objecten (van een gegeven type) in een ruizig beeld aanwezig zijn, en indien dit het geval is, het localiseren en typeren ervan. Er zijn vele toepassingen, zoals het automatisch lezen van postcodes of andere documenten, robotvisie, het interpreteren van scintigrammen of magnetische resonantiebeelden voor medische doeleinden, de classificatie van sterrenstelsels in astronomie of van cellen in cytologie en de detectie van de componenten in ertsen.

In hoofdstuk 2 bespreken wij eenvoudige statistische technieken en laten zien dat er vele connecties zijn met populaire, klassieke methoden zoals de Houghtransformatie en de mathematische morfologie. Deze methoden hebben als belangrijkste bezwaar, dat ze lijden aan meervoudige respons, dat wil zeggen: er worden te veel bijna identieke objecten gevonden. Om dit probleem aan te pakken, stellen wij een Bayesiaanse aanpak voor, waarin de a priori verdeling een kleine kans toekent aan configuraties met veel overlappende objecten. Een goede keuze voor deze a priori verdelingen is de klasse van ruimtelijke Markovprocessen uit de stochastische meetkunde. Deze modellen hebben de eigenschap dat zij eenvoudig te interpreteren zijn in termen van interacties tussen overlappende objecten en dat quotiënten van het type $p(J\mathbf{x})/p(\mathbf{x})$ gemakkelijk te berekenen zijn voor eenvoudige transformaties J, zoals het toevoegen of wegnemen van een object.

In hoofdstuk 3 geven wij een overzicht van Markov objectprocessen en introduceren een model (ten dele reeds bekend in de chemische fysica onder de naam 'penetrating spheres'). In tegenstelling tot de meer gangbare modellen met uitsluitend interacties tussen *twee* objecten, is dit model geschikt voor zowel het modelleren van afstoting (objectherkenning) als van aantrekking (vloeistofmodellen in de chemische fysica). Wij tonen aan dat dit model welgedefinieerd is en gezien kan worden als de restrictie van een stationair proces op de hele Euclidische ruimte. Het model kent interacties tussen willekeurig veel overlappende objecten en heeft als afdoende statistische grootheid de bekende geschatte lege ruimtefunctie. De parameters van het model kunnen worden geschat middels de gebruikelijke methoden, maar ook direct via de lege ruimtefunctie en de verdelingsfunctie van de afstand tussen punten van het proces. Het model kan gezien worden als de zwakke limiet van autologistieke roosterprocessen.

In het midden van de jaren '80 ontwikkelden D. en S. Geman, J. Besag en anderen Bayesiaanse modellen en technieken voor segmentatie, het in relatief homogene stukken opdelen van een beeld. Hier is zowel de data als de gewenste segmentatie een gedigitaliseerd, eindig beeld en de a priori verdeling een discreet Markovveld.

Hierdoor geinspireerd ontwikkelen wij in hoofdstuk 4 een Bayesiaanse aanpak van objectherkenning. Wegens het eenvoudige locale gedrag van ruimtelijke Markov objectprocessen is het mogelijk zowel iteratieve methoden verwant aan Besag's ICM-algorithme als Monte Carlo-technieken gebaseerd op simulaties te ontwikkelen. In plaats van sequentieel elk pixel (beeldelementje) een nieuwe stochastische waarde te geven, worden geboorte-sterfteprocessen gebruikt. Dit zijn continue-tijds Markovprocessen, waarin de transities bestaan uit het toevoegen (geboorte) of wegnemen (sterfte) van een object. Hoewel de nadruk ligt op wiskundig begrip, worden de methoden geillustreerd aan de hand van enkele eenvoudige voorbeelden.

Het is belangrijk op te merken dat het niet noodzakelijk is te 'geloven' in de a priori verdeling in de zin dat een goede reconstructie een kleine a priori kans kan hebben. Dit is door Besag naar voren gebracht als een argument ten gunste van ICM; de globale eigenschappen van het a priori model zijn vaak ongewenst, maar alleen de geschikte locale eigenschappen worden gebruikt.

In hoofdstuk 5 bespreken wij een andere toepassing, namelijk het bepalen van de centra van clustering. De data kan hier vele vormen aannemen, bijvoorbeeld een patroon van punten in het platte vlak, maar ook de uitvoer van een randdetector. Toepassingen zijn onder meer het

reconstrueren van de vorige generatie bomen in een woud, het localiseren van wegen uit de Romeinse tijd aan de hand van archeologische vondsten en het bestuderen van breuklijnen in de aardkorst in relatie tot het optreden van aardbevingen. De stochastische meetkunde is hier behulpzaam bij zowel het (voorwaardelijk) modelleren van de data als bij de a priori verdeling. Wij behandelen in detail het eenvoudigste geval van ruimtelijke puntpatronen en geven aan hoe dit kan worden uitgebreid. Het is ook van belang de data op te delen in zinvolle groepen. Wij tonen een verband aan met het klassieke 'k-means' algorithme en laten zien dat onder voor de hand liggende voorwaarden de naaste oudertoewijzing een meest aannemelijke schatter is. De methode wordt geillustreerd aan de hand van een bekend voorbeeld uit de literatuur.

In hoofdstuk 6 tenslotte laten wij zien dat vele clustermechanismen Markoveigenschappen bezitten met betrekking tot de samenhangende componentenrelatie. Als de ouders een naaste buur Markovproces vormen en de verzameling nakomelingen niet-leeg en begrensd is, is het totale nageslacht weer een naaste buur Markovproces. Dit suggereert dat deze klasse modellen een geschikte keuze is voor het modelleren van een populatie zich voortplantende individuen. De resultaten zijn een sterk argument voor het vermoeden van Møller dat naaste buur Markovprocessen zeer geschikt zijn voor het modelleren van gematigd geclusterde patronen, dit in tegenstelling tot de beter bekende Markovmodellen met uitsluitend paarsgewijze interacties.