Likelihood Devices in Spatial Statistics

Aannemelijkheids Listen in de Ruimtelijke Statistiek
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR AAN DE UNIVERSITEIT UTRECHT OP GEZAG VAN DE RECTOR MAGNIFICUS, PROF. DR H.O. VOORMA INGEVOLGE HET BESLUIT VAN HET COLLEGE VOOR PROMOTIES IN HET OPENBAAR TE VERDEDIGEN OP VRIJDAG 3 SEPTEMBER 1999 DES OCHTENDS TE 10.30 UUR

door
Erik Willem van Zwet
geboren op 10 november 1970, te 's-Gravenhage

Promotor: Prof. dr R.D. Gill

Faculteit der Wiskunde en Informatica

1991 Mathematics Subject Classification: 62M30

---

van Zwet, Erik Willem

Likelihood Devices in Spatial Statistics
Erik Willem van Zwet – Utrecht:
Universiteit Utrecht, Faculteit Wiskunde en Informatica
Proefschrift Universiteit Utrecht. – Met lit. opg.
Met samenvatting in het Nederlands.
ISBN 90-393-2154-X

Contents

Chapter 1

Introduction

1.1 Context and scope

This thesis deals with spatial statistics, or more precisely, with the statistical analysis of spatial data. The spatial data we consider are random configurations of points, line segments or discs in the two dimensional plane. Such random 'processes' could serve, and have served, to model a wide variety of real life phenomena. The points may represent the locations of trees in a forest, the line segments could be fractures in a rock surface and the discs could be cells under a microscope.

Usually, we observe these spatial processes through a bounded window while we imagine they extend throughout the entire plane. As a result, we often have to deal with 'edge-effects'. For instance, observing a collection of line segments through a bounded window some of the segments will be clipped at the window's boundaries. Such *censoring* has been studied extensively in the one dimensional case. The line segments then represent lifetimes which are sometimes not completely observed. Estimation of the lifetime distribution from incomplete observations is the main subject of the field of survival analysis. Laslett (1982a, b) was the first to point out the analogy of censoring in survival analysis and edge effects in spatial statistics.

Our goal in this thesis is always to estimate the probability distribution of the (partly) observed random spatial process. Sometimes this distribution is determined by a *finite* dimensional parameter, such as the expected number of points in a given set. At other times this will not be the case and then we speak of a semi- or nonparametric problem or of a problem with a 'large parameter space'. In recent years estimation of infinite dimensional parameters has received much attention and has made great headway. Bickel et al. (1993) is a standard reference for modern semi-parametric theory. In-

complete observation, such as due to edge effects, is well incorporated in the theory.

One of the main themes of this thesis is the application to spatial data of semi-parametric methods. Another, closely related theme is maximum likelihood estimation from spatial data. Maximum likelihood estimation is not common practice in spatial statistics as the method of moments and minimum contrast methods are more often applied (Stoyan et al. 1987). One might view our effort in the light of a development which has been described—somewhat provocatively—by Geyer (1999):

"Exotic areas of statistics, such as spatial statistics in general and spatial point processes in particular often recapitulate the history of statistics. The first formal inference is nonparametric, using method of moments estimators. At this stage there is no modeling. Ordinary statistics was in this phase a century ago with Pearson families of curves fit by method of moments. Time series was in this phase in the 1950s with spectral analysis. Spatial statistics was in this phase in the 1970s. Ripley's K-functions are an example. So is spatial autocorrelation analysis of lattice processes.

The next phase involves the introduction of parametric statistical models, efficient estimation methods, and hypothesis tests, which happened in ordinary statistics in the 1920s and 1930s, in time series in the 1960s, and is only now happening in spatial statistics. "

The thesis is organized around three problems: Laslett's line segment problem; estimation of the chord length distribution; and estimation for the bombing model. We describe these problems in section 1.2 of this introduction and return to each of them in the next three chapters. In the remainder of this chapter we gather some theory and tools for later reference. In particular, we briefly discuss the issue of independence in spatial processes and introduce the Poisson process and the related Boolean model. Next, we discuss some topics in semi-parametrics and finally consider Markov Chain Monte Carlo methods and perfect sampling. We want to draw the reader's special attention to two topics to which we contribute something new. The first is a general discussion of consistency of maximum likelihood estimators in section 1.4.2. The second is a modification of a stochastic version of the well-known EM algorithm (Dempster et al. 1977, Celeux and Diebolt 1986) in section 1.5.3.

1.2 Three problems

In this thesis we analyze three spatial missing data problems. In all of these we are unable to write down the likelihood of the data, let alone maximize it. This seems typical of spatial statistical estimation problems and explains why maximum likelihood is usually not the preferred approach.

In each of the first two of our problems we manage to find 'analogous' problems, which are simpler yet similar. For these analogous problems we *can* perform maximum likelihood estimation. Thus we obtain estimators that are MLEs—only not for the actual problems at hand! Still, if the simplifications are minor we may expect that the nice properties that an MLE typically possesses will be preserved, at least to a certain degree. Efficiency will generally fail, but consistency and asymptotic normality may not. We shall study consistency only, but we feel confident that it would be also possible to tackle asymptotic normality.

Although the first two problems both concern line segments and our approach to them is similar, we should point out that they really are quite different. In the first problem we have to deal with spatial dependence among the data. The difficulty with the second problem is that the data are in the form of random functions on random sets.

In the third problem, concerning the Boolean model of randomly placed discs, our difficulties are of a different nature still. The likelihood is easy enough, except for an intractable normalizing constant. No simplification will help and so the emphasis shifts from analytic methods to a Monte Carlo approach. In particular, we present a new algorithm to sample from the distribution of a wide variety of point processes.

We now briefly state our three problems and point out their particular difficulties. We do not as yet strive for mathematical rigor, but merely wish to indicate where we are going—and hopefully whet the reader's appetite.

Laslett's line segments

Figure 1.1 is a map of an area of about 160 by 160 meters of the Canadian Shield near Manitoba. The black parts are the rock surface, the white parts are soil, vegetation or water. The white lines through the black regions indicate fractures and the problem is to estimate the distribution of their lengths. This is a missing data problem because the fractures are unobservable in the white regions and outside the 160 by 160 meters area of the map. The particular difficulty here lies in the fact that the (black) observation region is not convex. This means that of a single fracture we might observe several fragments. Because the resolution of the map is rather poor, it is impossible

to know if two fragments belong to the same fracture.

The fractures are conveniently modeled by a location and an associated line segment. By 'location' we mean, for instance, the location of a preferred point on the segment, such as the mid point or the left-hand end-point.

In earlier contributions, the case of a convex observation window was considered. Assuming the locations and line segments to be a Poisson marked point process one finds that the fractures are independent. Standard methods can then be applied to compute the (non-parametric) maximum likelihood estimator (Gill 1994, Wijers 1995a and b). For us there is no gain in assuming that the locations of the fractures are Poisson, because even if the *fractures* are independent some of the *fragments* are certainly dependent.

Chord Length Estimation

Consider a random closed set $\mathcal{W} \subset \mathbb{R}^2$. With each point in $\mathcal{W}$ we can associate a *chord* in a given direction which is the longest line segment through that point which lies wholly within $\mathcal{W}$. Also, with each point in $\mathcal{W}^c = \mathbb{R}^2 \setminus \mathcal{W}$ we can associate a *contact segment* (our terminology) in a given direction which is the longest line segment starting at that point which is completely contained in $\mathcal{W}^c$. Of course these two concepts are closely related.

The problem is estimation of the distribution function of a typical chord and a typical contact segment, when we observe $\mathcal{W}$ through a compact set B. This is a missing data problem because some chords and contact segments are clipped at the boundaries of B. The particular difficulty here is that for *every* point in $\mathcal{W}$, respectively $\mathcal{W}^c$ we have a chord or a contact segment. Such continuous data is awkward and we are unable to apply standard methods for missing data problems.

The Bombing Model

Consider the so-called Boolean model in $\mathbb{R}^2$ where at each point of a Poisson point process of intensity λ, a disc of random radius is placed. This model is also known as the bombing model. We observe only the union of the discs through a compact window W. In this example data is missing as some discs are not observed because they are covered by others. Also there is the usual censoring at the observation window's boundaries. For our present purpose, estimation of the intensity parameter λ, the latter censoring is not the main difficulty. The difficulty is that the conditional distribution of the complete data, given the observed data involves a normalizing constant which cannot possibly be evaluated. Again, standard methods for missing data problems fail.

Figure 1.1: Fractures in 160 m. by 160 m. granitic rock of the Lac du Bonnet batholith in Manitoba. From Stone, Kamineni and Brown (1984). Digitized and post-processed by professor A.J. Baddeley of the University of Western Australia. The irregular black region is the rock's surface. The white parts represent areas where the rock cannot be observed due to soil, vegetation or water. The white lines through the black regions indicate fractures.

1.3 Independence, Poisson processes and Boolean models

This section borrows its title from section 1.5 of Baddeley (1999a). It contains some basic facts concerning the Poisson process in more general spaces than $\mathbb{R}^+$, and introduces the Boolean model. The Poisson process and the Boolean model—which is derived from it—are perhaps the simplest of all spatial models. They have been studied in detail, for instance in Daley and Vere-Jones (1988) and Stoyan et al. (1987). The content of this section can be said to be part of the folklore of stochastic geometry.

Classical statistics deals typically with independent, identically distributed random variables. An essential element of spatial statistics is a lack of independence. Suppose that X is a random spatial process of, say, randomly placed points. If there is any spatial structure or regularity, then typically we fail to have independence of $X \cap A$ and $X \cap B$, even when A and B are disjoint. The prime example of a *completely random* point process is the Poisson process. We now give a rather general definition.

Let $(S, \mathcal{B}, \Lambda)$ denote a measure space with a finite diffuse Radon measure Λ. A typical example is the case where S is a compact subset of $\mathbb{R}^2$ with the Borel sigma-algebra and $\Lambda(\mathrm{d}s) = \lambda \mathrm{d}s$.

Consider the 'Carter–Prenter exponential space' Ω of all finite subsets x of elements of S with a sigma-algebra $\mathcal{F}$ generated by sets $\{x \in \Omega : n(x \cap B) = n\}$ $(B \in \mathcal{B})$, where $n(x)$ denotes the number of elements in x.

The *Poisson* process Φ of *intensity* Λ can be seen as a random (counting) measure on $(S, \mathcal{B})$ or as a measure on $(\Omega, \mathcal{F})$. By $\Phi(B)$ we denote the (random) number of points of Φ that fall in a $\mathcal{B}$-measurable set B. The Poisson process is determined by the following two properties

1. $P(\Phi(B) = n) = \Lambda(B)^n e^{-\Lambda(B)}/n!, \qquad$ for $n = 0, 1, 2, \ldots,$ for all $B \in \mathcal{B}$
2. $\Phi(B_1), \Phi(B_2), \ldots, \Phi(B_k)$ are independent for disjoint $B_1, B_2, \ldots, B_k$.

Note that $E\Phi(B) = \Lambda(B) = \int_B \Lambda(\mathrm{d}s)$.

The expected number of points in a set B has the Poisson distribution with parameter $\Lambda(B)$. Conditionally on $\Phi(B) = n$, the points of $\Phi \cap B$ are distributed as a sample of size n from the 'normalized intensity' $\Lambda(S)^{-1}\Lambda(\mathrm{d}s)$.

Poisson processes have a nice 'superposition' property: The superposition (union) of independent Poisson processes is again a Poisson process, whose intensity is the sum of the intensities of the individual processes.

It is often useful to associate with each point of a point process a characteristic or 'mark'. For instance, when the points indicate locations (of the

midpoint) of fractures in a rock surface, an extra mark could indicate each fracture's length or direction. When the points represent (the center of) trees in a forest the mark could be their stem's diameter.

Now suppose we have a point process $\Phi = \{S_i\}$ on a space S. Suppose with each point S_i is associated a random mark M_i in some 'mark space' M, equipped with a probability measure P. Then the 'Marking Theorem' states that the following are equivalent

- The point process Φ is Poisson with intensity Λ and conditional on Φ, the marks are i.i.d. according to P.
- The marked point process $\{(S_i, M_i)\}$ is a Poisson process on $S \times M$ with intensity measure $\Lambda \times P$.

As a consequence of the Marking Theorem we have the 'thinning' property of Poisson processes: Suppose we have a Poisson process of intensity Λ. Now if we delete each point, independently of the others, with probability $1 - p$, the result will be a Poisson process with intensity $p\Lambda$.

Above we gave two examples of marked point processes; one representing a random collection of line segments (fractures), the other of circles (trees with specified stems). When the underlying point process is Poisson, such collections of randomly placed objects are called *Boolean models* (Matheron 1975, Stoyan et al. 1987). More precisely, suppose that $\Phi = \{S_i\}$ is a Poisson process on $\mathbb{R}^d$ and suppose that $\Xi_1, \Xi_2, \ldots$ are a sequence of independent, identically distributed random closed sets in $\mathbb{R}^d$. Then the Boolean model is defined as $\cup_i (S_i \oplus \Xi_i)$. The symbol $\oplus$ means Minkowski addition defined as $A \oplus B = \{a + b,\ a \in A,\ b \in B\}$.

1.4 Semi-parametrics

In this section we introduce the maximum likelihood estimator (MLE) $\hat{P}_n$ of a probability measure P_0 among a class $\mathcal{P}$ which we allow to be infinite dimensional. In this situation, the MLE is often called the nonparametric MLE (NPMLE). Also, we discuss differentiable, parametric sub-models through $\mathcal{P}$. Each such submodel will have a score function. With these scores, we can consider solutions of score equations. We discuss consistency of the NPMLE and of solutions of score equations. Other desirable properties we might look for in estimators, such as efficiency and asymptotic normality will not be touched upon in this thesis and are therefore omitted in the present discussion. Finally, we also consider so-called missing data problems, which go also by the name 'mixture models'.

A standard reference on semi-parametrics is the monograph by Bickel et al. (1993). Our exposition here owes much to the very readable chapter 25 on semi-parametrics of a recent book by van der Vaart (1998).

1.4.1 Nonparametric maximum likelihood estimation

In this section we define the so-called nonparametric maximum likelihood estimator (NPMLE). Consider random variables $X_1, X_2, \ldots, X_n$ which are independent and identically distributed according to some distribution P_0 in a class $\mathcal{P}$ of probability measures.

When $\mathcal{P}$ is not dominated by a single measure, the ordinary definition of the maximum likelihood estimator fails because there really is no likelihood. Kiefer and Wolfowitz (1956) proposed a way out of this difficulty by considering only pairwise comparisons. Now let $\mathbb{P}_n$ denote the empirical measure.

Definition 1.1. *A maximum likelihood estimator $\hat{P}_n$ of $P_0 \in \mathcal{P}$ is any element of $\mathcal{P}$ such that*

$$\int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}(\hat{P}_n + P)} \mathrm{d}\mathbb{P}_n \geq \int \log \frac{\mathrm{d}P}{\mathrm{d}(\hat{P}_n + P)} \mathrm{d}\mathbb{P}_n \quad \forall P \in \mathcal{P}, \tag{1.1}$$

while for every $P \in \mathcal{P}$, the left hand side is greater than minus infinity.

The latter requirement is a technicality. Its meaning will become clear in Lemma 1.1 below.

The definition says that $\hat{P}_n$ should 'beat' any other candidate on the data in pairwise comparison. Note that as we only compare two measures at a time, $\mathcal{P}$ need not be dominated. Note also that if $\mathcal{P}$ is dominated the definition reduces to the ordinary definition of maximum likelihood.

The definition depends on which versions of the Radon-Nikodym derivatives are used. These versions are only determined up to $(P_1 + P_2)$ null sets (when comparing P_1 and P_2 in $\mathcal{P}$), which are not necessarily $\mathbb{P}_n$ null sets. This problem has nothing to do with the fact that $\mathcal{P}$ may be undominated or infinite dimensional. It arises in parametric models just the same. It is generally felt that there should not be a problem—certainly not asymptotically—if we assume "what are known in some quarters as the British regularity conditions", to quote Cox (2000) (on a different matter). For a more thorough discussion see Scholz (1980).

For definiteness, we imagine that we have some mechanism to choose a version of each Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}(P+Q)$ for all $P, Q \in \mathcal{P}$ in advance, thus making sure that our choices do not depend on the data.

Moreover, we do this in such a way that $0 \leq \mathrm{d}P/\mathrm{d}(P+Q) \leq 1$ and $\mathrm{d}P/\mathrm{d}(P+Q) + \mathrm{d}Q/\mathrm{d}(P+Q) = 1$.

We find the definition of the NPMLE in its present form (1.1) slightly awkward and prefer to have a definition in terms of 'log likelihood ratios', see (1.2) and (1.3) below. Let us now define the likelihood ratio and log likelihood ratio.

Consider two probability measures P_1 and P_2 in $\mathcal{P}$, where P_2 need not dominate P_1. Define the *likelihood ratio* to be

$$\frac{\mathrm{d}P_1}{\mathrm{d}P_2}(x) = \frac{(\mathrm{d}P_1/\mathrm{d}(P_1+P_2))(x)}{(\mathrm{d}P_2/\mathrm{d}(P_1+P_2))(x)}.$$

This likelihood ratio should of course not be confused with the Radon–Nikodym derivative, which conventionally is defined only when $P_1 \ll P_2$. Because we can choose versions of the numerator and denominator which add up to 1, there are no problems with dividing zero by zero. If the numerator is positive while the denominator is zero, the ratio is defined to be infinite.

Lemma 1.1. *The defining property (1.1) of the NPMLE $\hat{P}_n$ is equivalent to*

$$\int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P} \mathrm{d}\mathbb{P}_n \geq 0 \quad \forall P \in \mathcal{P}, \tag{1.2}$$

and also to

$$\int \log \frac{\mathrm{d}P}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \leq 0 \quad \forall P \in \mathcal{P}. \tag{1.3}$$

Proof This lemma is immediate from our definition of likelihood ratio. The assumption that the left hand side of (1.1) is greater than minus infinity is equivalent to (1.2) and (1.3) being well-defined. □

We should mention that an NPMLE need not exist and if it exists it need not be unique. A modification called the method of *sieves* due to Grenander (1981) is sometimes useful. We choose submodels $\mathcal{P}_n \subseteq \mathcal{P}$ such that $\cup_n \mathcal{P}_n$ will almost surely eventually include the true P_0. Then we define a 'sieved' NPMLE $\hat{P}_n^s$ to be such that

$$\int \log \frac{\mathrm{d}\hat{P}_n^s}{\mathrm{d}(\hat{P}_n^s + P)} \mathrm{d}\mathbb{P}_n \geq \int \log \frac{\mathrm{d}P}{\mathrm{d}(\hat{P}_n^s + P)} \mathrm{d}\mathbb{P}_n \quad \forall P \in \mathcal{P}_n, \tag{1.4}$$

while for every $P \in \mathcal{P}_n$, the left hand side is greater than minus infinity.

The $\mathcal{P}_n$ should, if possible, be chosen in such a way that $\hat{P}_n^s$ exists and is easily computed.

It is often useful to choose $\mathcal{P}_n$ in a way that is dependent on the data. For instance, we can take $\mathcal{P}_n = \{P \in \mathcal{P} : P \ll \mathbb{P}_n\}$. The resulting sieved NPMLE we should perhaps call *data* sieved NPMLE. However, we shall generally omit *both* modifiers 'data' and 'sieved'.

1.4.2 Consistency of the (NP)MLE

In this section we argue that the 'natural' approach to proving consistency of an MLE leads to *Hellinger* consistency. This usually overshoots the mark. Hellinger consistency—being equivalent to consistency in total variation—is too strong for most purposes. In many situations, especially when the model $\mathcal{P}$ is infinite dimensional, it just does not hold. For instance, when the NPMLE $\hat{P}_n$ is discrete while P_0 is absolutely continuous (with respect to Lebesgue measure) the Hellinger distance between $\hat{P}_n$ and P_0 is never less than 1.

One way to escape the rigidity of Hellinger consistency is to try to find a sequence P_n such that P_n tends to P_0 in some reasonable sense (say weak convergence), while at the same time the Hellinger distance between P_n and $\hat{P}_n$ tends to zero. Such a comparative sequence has successfully been used by Murphy (1994), Wijers (1995b) and Gill (1994) and it will also be used several times in this thesis.

Now we define two concepts of 'distance' between two probability measures. The first, Hellinger distance, is a true distance in the mathematical sense; the second, Kullback–Leibler information is not. Kullback–Leibler information and the Hellinger metric play an important role in proving consistency of maximum likelihood estimators as they fit in well with the MLE definition as we will see later on.

Definition 1.2. *The* Hellinger distance $H(P, Q)$ *between two probability measures P and Q is defined to be the square root of*

$$\begin{aligned} H^2(P,Q) &= \int \left| \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2} - \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2} \right|^2 \mathrm{d}\mu \\ &= 2 - 2\int \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2} \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2} \mathrm{d}\mu, \end{aligned}$$

for μ dominating both P and Q

This definition does not depend on the choice of μ. It can be shown that the topologies induced by Hellinger metric and the—perhaps more familiar—total variation metric are the same. In fact, we have

$$\frac{1}{2}H^2(P,Q) \leq \|P-Q\|_{TV} \leq H(P,Q),$$

where $\|.\|_{TV}$ denotes the total variation norm. The variational metric is always between 0 and 1 and the Hellinger metric is between 0 and $\sqrt{2}$.

Definition 1.3. *The* Kullback–Leibler information *of a probability measure P with respect to Q is defined to be*

$$K(P:Q) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P. \tag{1.5}$$

Conventionally, $\mathrm{d}P/\mathrm{d}Q$ is only defined when Q dominates P and the Kulback–Leibler information is defined to be infinite when that is not the case. Since we defined $\mathrm{d}P/\mathrm{d}Q$ even when Q does not dominate P, (1.5) completely specifies the Kullback–Leibler information. One can check that, also according to our definition (1.5), $K(P:Q) = \infty$ when Q does not dominate P. Note that $K(P:Q)$ can well be infinite even when Q dominates P.

Even though the Kullback–Leibler information is often used to quantify the difference between two measures, it is not a distance in the mathematical sense. However, as a simple consequence of Jensen's inequality we do have that $K(P:Q) \geq 0$ with equality if and only if $P = Q$.

The following lemma relates Kullback–Leibler information to Hellinger distance.

Lemma 1.2. *For any two probability measures P and Q we have*

$$K(P:Q) \geq H^2(P,Q).$$

Proof For all $x > 0$ we have $\log x \leq x - 1$. Hence,

$$\begin{aligned}
K(P:Q) &= \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}P \\
&= -2\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{1/2}\mathrm{d}P \\
&\geq 2\int 1 - \left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{1/2}\mathrm{d}P \\
&= 2 - 2\int \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1/2}\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{1/2}\mathrm{d}\mu \\
&= H^2(P,Q).
\end{aligned}$$

where μ is some measure dominating both P and Q. □

We now turn to consistency. Suppose for now that

$$K(P_0 : \hat{P}_n) = \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}P_0 < \infty.$$

We can then rewrite (1.3) as

$$\begin{aligned} 0 &\geq \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \\ &= \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) + \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}P_0 \\ &= \int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) + K(P_0 : \hat{P}_n), \end{aligned}$$

so if

$$\int \log \frac{\mathrm{d}P_0}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_0) \to 0, \qquad P_0\text{-almost surely} \tag{1.6}$$

then $K(P_0 : \hat{P}_n) \to 0$, which by the above lemma implies $H(P_0 : \hat{P}_n) \to 0$. In other words, (1.6) implies the almost sure Hellinger consistency of the NPMLE $\hat{P}_n$.

Condition (1.6) is an obvious consequence of a uniform law of large numbers over a class of functions. Such uniform LLN's are the object of much study in modern empirical process theory. In 'empirical process parlance', (1.6) follows directly from saying that the random function $\log(\mathrm{d}P_0/\mathrm{d}\hat{P}_n)$ is, for n large enough, in a so-called Glivenko–Cantelli class.

The convergence of (1.6) establishes 'Kullback–Leibler' consistency, which by lemma 1.2 implies Hellinger consistency. We can also get at Hellinger consistency directly. The equality we will now prove is due to van de Geer (1993), though she considered only the case where $\mathcal{P}$ is dominated by a single sigma-finite measure. We make no assumptions on $\mathcal{P}$ here. Taking $P = P_0$

in (1.2) and then using the fact that $\log x \leq x - 1$ $(x > 0)$ have

$$\begin{aligned}
0 &\leq \int \log \frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \mathrm{d}\mathbb{P}_n \\
&= 2 \int \log \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}\mathbb{P}_n \\
&\leq 2 \int \left[\left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} - 1 \right] \mathrm{d}\mathbb{P}_n \\
&= 2 \int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) + 2 \int \left[\left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} - 1 \right] \mathrm{d}P_0 \\
&= 2 \int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) - 2H^2(P_0 : \hat{P}_n).
\end{aligned}$$

Because $H^2(P_0; \hat{P}_n)$ is always between 0 and 2 it is not a problem to add and delete it as we did in the one-but-last line above. Now if

$$\int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_0} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_0) \to 0 \tag{1.7}$$

we can conclude that $H(P_0, \hat{P}_n) \to 0$, which is to say that $\hat{P}_n$ is Hellinger consistent. Again is suffices to verify a uniform law of large numbers. Van de Geer (1993) proceeds to use empirical process theory to investigate the rate of convergence of $\hat{P}_n$ to P_0 (when indeed they converge).

In case $\mathcal{P}$ is convex, she suggests comparing $\hat{P}_n$ not to P_0 but to a convex combination of $\hat{P}_n$ and P_0, say P_n. The idea is that it may be easier to establish $H(\hat{P}_n, P_n) \to 0$ than $H(\hat{P}_n, P_0) \to 0$ when the two are actually equivalent.

The two methods we presented for proving consistency of the NPMLE yield Hellinger consistency, which—being equivalent to consistency in total variation norm—is a *very* strong form of consistency. It may be too much to ask for. It may also be more than we want as we are often only interested in weak consistency (i.e. consistency in the Prohorov metric). In many non– and semi-parametric models the NPMLE is discrete whereas the true P_0 could be absolutely continuous. As the total variation distance between singular measures is 1, the Hellinger distance between $\hat{P}_n$ and P_0 cannot be less than 1.

We propose the use of a sequence $\{P_n\}$ which tends to P_0 in some desired sense, say weakly, while $H(\hat{P}_n, P_n)$ tends to zero. Such a sequence was successfully applied by Murphy (1994) in a case where Hellinger consistency indeed did not hold.

We cast this idea in the form of a theorem. Suppose that d is a metric on $\mathcal{P}$ which weaker than (or equivalent to) Hellinger metric.

Theorem 1.1. *Let $\hat{P}_n$ denote the NPMLE based on independent observations $X_1, X_2, \ldots, X_n$ which are all distributed according to $P_0 \in \mathcal{P}$ and let $\mathbb{P}_n$ denote the empirical measure. If there exists a sequence $\{P_n\} \in \mathcal{P}$ such that $d(P_n, P_0) \to 0$ (P_0-a.s.) while either*

$$K(P_n : \hat{P}_n) < \infty \quad \textit{and} \quad \int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) \to 0, \qquad P_0\textit{-almost surely}$$

or

$$\int \left(\frac{\mathrm{d}\hat{P}_n}{\mathrm{d}P_n} \right)^{1/2} \mathrm{d}(\mathbb{P}_n - P_n) \to 0, \qquad P_0\textit{-almost surely}$$

then $d(\hat{P}_n, P_0) \to 0$ (P_0-almost surely)

Proof Replacing P_0 with P_n in the preceding discussion, we see that either condition on P_n allows us to conclude that $H(\hat{P}_n, P_n) \to 0$. This theorem's claim now follows from the assumption that $d(P_n, P_0) \to 0$. □

1.4.3 Score equations

In the parametric case the 'score function', i.e. the derivative of the log likelihood, is a familiar object. Often the maximum likelihood estimator solves the score equations: sum of scores (is less than or) equals zero. This sometimes enables us to calculate the MLE and it also helps us to prove efficiency and asymptotic normality. As it turns out, the notion of a score can be extended very neatly to the infinite dimensional case. Hence we can also write down 'non-parametric' score equations. In the infinite dimensional case the score equations usually do not characterize the NPMLE. However, solutions of the score equations are often relatively easy to obtain, when the real NPMLE is much more elusive. Besides, we tend to prove favorable properties of maximum likelihood estimators by considering them as solutions of score equations. Why then not simply consider such solutions from the outset? We shall at times be sloppy and even call any solution of the score equations an NPMLE.

Scores are defined through what are called 'differentiable submodels'. Following van der Vaart (1998), we consider maps $t \mapsto P_t$ from a neighborhood of 0 in $[0, \infty)$ to $\mathcal{P}$ such that there exists a measurable function $g : \mathcal{X} \to \mathbb{R}$ for which

$$\int \left[\frac{p_t^{1/2} - p^{1/2}}{t} - \frac{1}{2} g p^{1/2} \right]^2 \mathrm{d}\mu \to 0. \tag{1.8}$$

where p_t and p are Radon–Nikodym derivatives of P_t and P with respect to a sigma-finite measure $\mu(P, P_t)$, which dominates both P_t and P.

We call $t \mapsto P_t$ a differentiable submodel through P. If the integrand converges pointwise then $g = \frac{\mathrm{d}}{\mathrm{d}t}|_{t=0} \log \mathrm{d}P_t$ which is just the familiar score function. The collection of scores associated with all differential submodels is called the *tangent set* of the model $\mathcal{P}$ at P and denoted $\dot{\mathcal{P}}(P)$. Lemma 25.14 in van der Vaart (1998) asserts that from (1.8) it follows that $\int g \mathrm{d}P = 0$ and $\int g^2 \mathrm{d}P < \infty$. Hence, $\dot{\mathcal{P}}(P)$ can be identified (up to equivalence) with a subset of $L_2^0(P)$.

Here a clear distinction can be made between semi- and nonparametric models: a nonparametric model has a tangent set equal to the whole of $L_2^0(P)$, while the tangent set of a semi-parametric model is a true subset of $L_2^0(P)$. Moreover, the tangent set of a semi-parametric model is not contained in a finite dimensional subspace of $L_2^0(P)$. In other words, it is infinite dimensional but not everything. Indeed, if the model consists of all probability measures (on a given measurable space) then $\dot{\mathcal{P}}(P) = L_2^0(P)$.

Now consider the 'score inequalities'

$$\int g \mathrm{d}\mathbb{P}_n = \frac{1}{n} \sum_i g(X_i) \leq 0 \qquad \forall g \in \dot{\mathcal{P}}(\hat{P}_n) \tag{1.9}$$

An NPMLE as defined in the previous section is a solution of these inequalities. The converse does not hold. Still, we will call a solution of these equations an NPMLE.

The most familiar NPMLE, the empirical measure, solves a the score inequalities when the model $\mathcal{P}$ is non-parametric so that $\dot{\mathcal{P}}(\hat{P}_n) = L_2^0(\hat{P}_n)$. Taking scores $g(x) = \mathbf{1}_F(x) - \hat{P}_n(F)$, for all $F \in \mathcal{F}$, we see that the empirical $\hat{P}_n = \mathbb{P}_n$ indeed arises as a solution of (1.9).

We now discuss how the methods of the previous section can be used to prove consistency of solutions of the score equations. To this end we must choose particular submodels. Let $\hat{P}_n$ be a solution of the score equations and consider a sequence $\{P_n\} \in \mathcal{P}$ such that $P_n \ll \hat{P}_n$.

Now suppose that $\log(\mathrm{d}P_n / \mathrm{d}\hat{P}_n) \in L_2(\hat{P}_n)$ and that $\hat{P}_n^{(1-t)} P_n^t \in \mathcal{P}$ for t in an open neighborhood of 0 in $[0, \infty)$. Then $t \mapsto \hat{P}_n^{(1-t)} P_n^t$ is a differentiable

submodel with score

$$\log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \in \dot{\mathcal{P}}(\hat{P}_n).$$

Because $\hat{P}_n$ solves all the score equations it follows that

$$\int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n \leq 0,$$

which coincides with (1.3). Now we can apply the reasoning of the previous section to try to prove that a particular $\hat{P}_n$ is consistent.

Unfortunately, it is not usually true that $\hat{P}_n^{(1-t)} P_n^t \in \mathcal{P}$. However, it *is* often true that $\mathcal{P}$ contains the straight line $(1-t)P + tQ$ for all P and Q in $\mathcal{P}$ (i.e. $\mathcal{P}$ is convex). If (and only if) $P_n \ll \hat{P}_n$ and $\mathrm{d}P_n/\mathrm{d}\hat{P}_n \in L_2(\hat{P}_n)$ then $t \mapsto (1-t)\hat{P}_n + tP_n$ is a differentiable submodel with score

$$\frac{\mathrm{d}P_n - \mathrm{d}\hat{P}_n}{\mathrm{d}\hat{P}_n} = \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \in \dot{\mathcal{P}}(\hat{P}_n)$$

and since $\hat{P}_n$ solves all score equations, it follows in particular that

$$\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}\mathbb{P}_n \leq 0. \tag{1.10}$$

We are assuming that $\mathrm{d}P_n/\mathrm{d}\hat{P}_n \in L_2(\hat{P}_n)$, so in other words

$$\int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}P_n < \infty.$$

Hence, we can rewrite (1.10) as

$$\begin{aligned} 0 &\geq \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}\mathbb{P}_n \\ &= \int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) + \int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}P_n. \end{aligned}$$

Now since $x - 1 > \log x$ for all $x > 0$ we have

$$\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1 \right) \mathrm{d}P_n > \int \log \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}P_n = K(\hat{P}_n, P_n) \geq 0.$$

Hence if

$$\int \frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} \mathrm{d}(\mathbb{P}_n - P_n) \to 0 \qquad P_0 - \text{almost surely} \tag{1.11}$$

then $\int \left(\frac{\mathrm{d}P_n}{\mathrm{d}\hat{P}_n} - 1\right) \mathrm{d}P_n$ tends to zero, which implies that $K(\hat{P}_n, P_n)$ tends to zero which in turn implies that $H(\hat{P}_n, P_n)$ tends to zero. Thus we have a theorem which is very similar to Theorem 1.1.

Let d be a metric on $\mathcal{P}$ which is weaker than (or equivalent to) Hellinger metric.

Theorem 1.2. *Let $\hat{P}_n$ denote a solution of the score equations (1.9) based on independent observations $X_1, X_2, \ldots, X_n$ which are all distributed according to $P_0 \in \mathcal{P}$ and let $\mathbb{P}_n$ denote the empirical measure. Suppose also that $\mathcal{P}$ is convex. If there exists a sequence $\{P_n\} \in \mathcal{P}$ such that*

1. *$d(P_n, P_0) \to 0$ (P_0-a.s.)*
2. *The submodel $t \mapsto (1-t)\hat{P}_n + tP_n$ is Hellinger differentiable at $\hat{P}_n$*
3. *(1.11) holds*

then $d(\hat{P}_n, P_0) \to 0$ (P_0-a.s.)

1.4.4 Missing data problems and the EM algorithm

In this section we study missing data models which are also known as information loss models or mixture models. The issue is that for some reason we do not (completely) observe the random variables whose common distribution $Q_0 \in \mathcal{Q}$ is of interest.

Let X be a generic random variable on some (measurable) space $\mathcal{X}$ with distribution $Q_0 \in \mathcal{Q}$ and let C be another random variable whose conditional distribution given X is known to us. Let $Y = \Phi(X, C)$, where Φ is a known many-to-one mapping. Y takes values in a measurable space $\mathcal{Y}$. Now suppose we observe independent $Y_1, \ldots, Y_n$ which are distributed as Y and from these we want to estimate Q_0. We call the X_i the complete data and the Y_i the incomplete or observed data. Estimation of $Q_0 \in \mathcal{Q}$ (or functionals thereof) from the observations Y_i is called a missing data problem.

The setup we have described is so general that one could maintain that *every* estimation problem is a missing data problem. Still, it will prove a convenient way to think about all sorts of censoring mechanisms.

We now give another way to describe the situation. Consider a *Markov kernel* $K(\mathrm{d}y; x)$. This means that $K(\mathrm{d}y; x)$ is a probability measure for every $x \in \mathcal{X}$ and $K(A; x)$ is a measurable function of x when $A \subseteq \mathcal{Y}$ is a measurable set. Now for all $Q \in \mathcal{Q}$ define the 'mixture' $P = KQ$ by

$$P(\mathrm{d}y) = KQ(\mathrm{d}y) = \int_{\mathcal{X}} K(\mathrm{d}y; x)\mathrm{d}Q(x). \tag{1.12}$$

Clearly P is a probability measure on the space $\mathcal{Y}$. Q is called the 'mixing distribution'. We suppose K is known and we observe an i.i.d. sample $Y_1, \ldots, Y_n$ with common distribution $P_0 = KQ_0$, where Q_0 is known to lie in a class $\mathcal{Q}$. The model for the distribution of a generic observation Y is of course $\mathcal{P} = K\mathcal{Q} = \{KQ : Q \in \mathcal{Q}\}$. The objective, again, is to estimate Q_0.

We have explained what we mean by a 'missing data model' and by 'a mixture model'. In fact, these two descriptions are the same. A missing data model can be described as a mixture model by defining $K(\mathrm{d}y; x)$ to be the conditional distribution of $Y = \Phi(X, C)$ given $X = x$. Conversely, a mixture model can be described as a missing data model as follows. First, define a random variable C which, conditional on $X = x$, has distribution $K(\mathrm{d}y; x)$. Next, define a random variable $Y = \Phi(X, C) = C$.

We shall now give two useful properties of $\mathcal{P} = K\mathcal{Q}$ which follow directly from properties of $\mathcal{Q}$.

Lemma 1.3. *If $\mathcal{Q}$ is convex then so is $\mathcal{P} = K\mathcal{Q}$.*

Proof Take two arbitrary elements $P, P' \in \mathcal{P}$. Then there exist $Q, Q' \in \mathcal{Q}$ such that $P = KQ$ and $P' = KQ'$. Now for $0 \leq t \leq 1$

$$\begin{aligned} tP + (1-t)P' &= t \int K(\mathrm{d}y; x)\mathrm{d}Q(x) + (1-t) \int K(\mathrm{d}y; x)\mathrm{d}Q'(x) \\ &= \int K(\mathrm{d}y; x)\mathrm{d}(tQ(x) + (1-t)Q'(x)) \in \mathcal{P} \end{aligned}$$

because $tQ + (1-t)Q' \in \mathcal{Q}$. □

Even more importantly,

Lemma 1.4. *If $t \mapsto Q_t$ is a differentiable submodel through $Q \in \mathcal{Q}$ with score function g , then $t \mapsto P_t = KQ_t$ is a differentiable submodel through $P = KQ \in \mathcal{P}$ with score $E_Q(g(X)|Y)$.*

Proof For a proof refer to Bickel et al. (1993) proposition A.5.5 or Le Cam and Yang (1988) proposition 4. □

We shall concern ourselves exclusively with completely *nonparametric* missing data / mixture problems, which means that $\dot{\mathcal{Q}}(Q) = L_2^0(Q)$ for all $Q \in \mathcal{Q}$. In other words, every square integrable function, which integrates to zero is a score function belonging to a differentiable submodel. The fact that scores are so plentiful is useful when we want to estimate Q_0 by solving score equations. The nonparametric situation typically arises when we have

no prior information about Q_0 so that we must take our model $\mathcal{Q}$ to consist of all probability measures on $\mathcal{X}$. In passing, we note that if $\mathcal{Q}$ consists of all probability measures on a given space, then $\mathcal{Q}$ is convex. By the first of the above lemmas, convexity of $\mathcal{Q}$ implies convexity of $\mathcal{P} = K\mathcal{Q}$. This will be very useful as this puts us in the realm of Theorem 1.2.

The fact that we only observe Y and not X may cause a lack of *identifiability*. We may well have that $KQ = KQ'$ while $Q \neq Q'$. This is unfortunate as it means that from the data we can not tell Q and Q' apart. We can of course reduce $\mathcal{Q}$ in such a way that we do have identifiability, but then our missing data problem will most likely fail to be completely non-parametric.

In the problems we study in this thesis we can construct a *different* missing data problem which *is* identifiable, but also completely non-parametric. It features a different model $\tilde{\mathcal{Q}}$ of probability measures on a different space $\tilde{\mathcal{X}}$ and also a different Markov kernel $\tilde{K}$. Of course we make sure that $\tilde{K}\tilde{\mathcal{Q}} = K\mathcal{Q}$ so that the model for the data is the same. For the remainder of this section we simply assume identifiability.

Turning to maximum likelihood estimation, the NPMLE $\hat{Q}_n$ of Q_0 is defined to satisfy

$$\int \log \frac{\mathrm{d}P}{\mathrm{d}\hat{P}_n} \mathrm{d}\mathbb{P}_n := \int \log \frac{\mathrm{d}KQ}{\mathrm{d}K\hat{Q}_n} \mathrm{d}\mathbb{P}_n \leq 0, \qquad \forall Q \in \mathcal{Q},$$

where $\mathbb{P}_n$ is the empirical distribution of the *observed* data $Y_1, \ldots, Y_n$.

Instead of trying to find the NPMLE, we might be already be content with a solution $\hat{Q}_n$ of the score equations

$$\int E_{\hat{Q}_n}(g(X)|y)\mathrm{d}\mathbb{P}_n(y) = \frac{1}{n}\sum_i E_{\hat{Q}_n}(g(X_i)|Y_i) = 0, \quad \forall g \in \dot{\mathcal{Q}}(\hat{Q}_n). \tag{1.13}$$

If we have a completely nonparametric missing data problem every function g in $L_2^0(\hat{P}_n)$ is a score. Thus we can take $g(x) = \mathbf{1}_A(x) - \hat{Q}_n(A)$ for any measurable set A to obtain the so-called *self consistency equations*

$$\hat{Q}_n(A) = \int E_{\hat{Q}_n}(\mathbf{1}_A(X)|y)\mathrm{d}\mathbb{P}_n(y)) = \frac{1}{n}\sum \hat{Q}_n(X_i \in A|Y_i). \tag{1.14}$$

It is usually assumed that for all x the kernels $K(\mathrm{d}y; x)$ have densities with respect to a single sigma-finite measure. This may not hold, and in particular it does not hold in the problems that are considered in this thesis. Fortunately, to solve the score equations we only need to compute integrals with respect to the conditional distribution under $\hat{Q}_n$ of X given Y. Hence,

it suffices if the $K(\mathrm{d}y; x)$ are dominated for $\hat{Q}_n$-almost all x. Thus suppose that a sigma-finite measure μ_n dominates $K(\mathrm{d}y; x)$ for $\hat{Q}_n$-almost all x. The measure μ_n is allowed to depend on the data $Y_1, \ldots, Y_n$. We denote $K(\mathrm{d}y; x)/\mathrm{d}\mu_n(y) = k_n(y; x)$. Then the conditional distribution of X given $Y = y$ under $X \sim \hat{Q}_n$ is

$$\frac{k_n(y; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(y; x')\mathrm{d}\hat{Q}_n(x')} \tag{1.15}$$

and the score equations (1.13) become

$$\sum_i \int \frac{g(x)k_n(Y_i; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(Y_i; x')\mathrm{d}\hat{Q}_n(x')} = 0, \quad \forall g \in \dot{\mathcal{Q}}(\hat{Q}_n). \tag{1.16}$$

Again, if a missing data problem is completely nonparametric then we can take $g(x) = \mathbf{1}_A(x) - \hat{Q}_n(A)$ for any measurable set A and (1.14) becomes

$$\hat{Q}_n(A) = 1/n \sum_i \int_A \frac{k_n(Y_i; x)\mathrm{d}\hat{Q}_n(x)}{\int_{x'} k_n(Y_i; x')\mathrm{d}\hat{Q}_n(x')}. \tag{1.17}$$

These equations are the well known self-consistency equations. Solving them iteratively is an instance of the EM algorithm. When we start the iterations with an initial guess $Q_n^{(0)}$ then all subsequent iterates will be dominated by it. Hence we can never do better than compute a 'sieved' NPMLE over the model $\{Q \in \mathcal{Q} : Q \ll Q_n^{(0)}\}$.

There has been much discussion of the merits and flaws of the EM algorithm (Dempster et al. 1977). Vardi and Lee (1993) point out the wide applicability of EM, but—obviously—blind application of the algorithm to every problem one meets is unwise. However, it does seem to perform rather well in 'nice' cases, i.e., parametric– and $\sqrt{n}$-nonparametric problems. EM is wholly unsuitable to determine the support of a distribution.

The rate at which EM converges is painfully slow, and has even been described as 'lethargic'. On the other hand, only infantile computer skills are required to implement it. A clever mathematician who is good with computers can generally do much better (cf. Meilijson, 1989).

1.5 Perfect Simulation

The distribution of a spatial process is often very complicated. Typically, if there is any spatial dependence we have a normalizing constant that is utterly intractable. To appreciate this fact, consider a Poisson point process

of intensity $\lambda > 0$ on the unit square. Suppose we are interested in the conditional distribution of the process, given that no two points are closer than at distance $r > 0$. This is called the hard core model. The normalizing constant here is an integral over all finite configurations of points that are all further apart than r. Performing this integration analytically appears impossible. Stochastic integration would be feasible if we can efficiently obtain realizations of the hard core model. Rejection sampling, i.e. generating Poisson processes and discarding those that do not meet the requirement, may not be efficient depending on λ and r.

Coupling-from-the-past (CFTP) (Propp and Wilson, 1996) is a clever method to obtain samples from complicated distributions which arise as the stationary distribution of a Markov chain—as quite often happens in spatial models. For instance, the hard core model arises as the stationary distribution of a so-called spatial birth and death process, where births that violate the condition are rejected.

In chapter 4 we consider estimation of the intensity of the Boolean model of discs (also known as the 'bombing model'). As it turns out we need to sample from a conditional Poisson point process. We have implemented a CFTP algorithm which is extremely quick.

This Boolean discs problem is a missing data problem, and we can use a stochastic version of the EM algorithm (StEM) to estimate the intensity. This algorithm produces an ergodic Markov chain on the parameter space. The StEM estimator is a sample (or an average of samples) from the stationary distribution of this chain. We can again use CFTP to actually obtain such samples. This we call the *perfect* StEM algorithm.

This section is based on van Zwet (2000). It is organized as follows. Below we first describe coupling-from-the-past, then the stochastic EM and finally the perfect stochastic EM algorithm.

1.5.1 Coupling From The Past

Consider an ergodic (i.e., irreducible and aperiodic) Markov chain X_t on a state space S and suppose we want to simulate its equilibrium distribution. Starting the chain from some arbitrary initial state and then running it for a very long, but finite time will generally not ensure that samples are from the stationary distribution. Recently, Propp and Wilson (1996) devised a method, called *coupling from the past* (CFTP), to produce perfect or exact samples. We closely follow Kendall and Thönnes (1998) to explain how it works.

For now, let us assume that the state space S is finite. A Markov chain X_t on S can be described by means of i.i.d. 'random transition maps' H_t :

$S \to S$. Such a random transition map H_t determines for each state $i \in S$ that the chain X might be in at time $t-1$ where the chain would jump to next. If p_{ij} are the transition probabilities of the Markov chain to move from state i to j, then the common distribution of the H_t should be such that $P(H_t(i) = j) = p_{ij}$. Then a realization of the Markov chain X is generated by fixing $X(0)$ at some X_0 and setting

$$X(t) = H_t(X(t-1)).$$

Coupling from the past now works as follows. We select a time $-T < 0$ in the past and run chains starting from each state of S from time $-T$ to time 0. The chains are coupled by using the *same* realizations of the transition maps for all the chains. Hence, it follows that if two chains started at different initial states coalesce, they will forever after stay together. Now we check if all chains have coalesced at time 0. If so, then the state at time zero must be a sample from the stationary distribution. This is understood as follows. Imagine that at some time long before $-T$ we also started a chain from an initial state selected according to the stationary distribution. This chain will remain in equilibrium, so in particular its state at time zero is distributed according to the stationary distribution. However, we have arranged it so that *all* chains, no matter which state they were in at time $-T$, are at time zero in the same state.

If not all chains have met, then we run chains from time $-2T$ to time 0, making sure that we use the same realizations of $H_{-T+1}, H_{-T+2}, \ldots, H_0$. If the paths still have not coalesced, we run chains from time $-4T$ and so on.

If the state space S has more than just a few elements it will not be feasible to run chains starting from all possible states.

Let us now consider finite, countable or even uncountable S. Suppose that S admits a partial ordering $\preceq$, and that there are a minimal and a maximal element, $\underline{s}$ and $\overline{s}$ such that

$$\underline{s} \preceq s \preceq \overline{s}, \qquad \text{for all } s \in S.$$

Also suppose that the chain is 'monotone' in that it respects the ordering.

$$H_t(s) \preceq H_t(s'), \qquad \text{for all } s \preceq s' \text{ a.s.}$$

Now we can imagine that we run coupled chains starting at all possible states, while in practice we only run chains from states $\underline{s}$ and $\overline{s}$. Because all paths are ordered, it follows that if the latter two have met then all paths starting from intermediate states would have met as well. Of course it remains to verify in each application that the algorithm will almost surely terminate in finite time.

1.5.2 Stochastic EM

The stochastic EM (StEM) algorithm (Celeux and Diebolt (1986), Wei and Tanner (1990)), which we describe here, is an iterative procedure to estimate a parameter in a missing data problem. It produces an ergodic Markov Chain on the parameter space. The equilibrium distribution of this chain has certain desirable properties (Nielsen (1997a and b). In the next section we will investigate if and when the ideas of the previous section can be applied to actually obtain a sample from this equilibrium distribution.

Suppose that X is distributed according to a probability measure P_{θ_0}. Suppose we can observe only the result of a many-to-one mapping $Y(C, X)$, where the conditional distribution of C given X is known. The goal is to estimate θ_0, where θ_0 is assumed to be in some general set Θ, from observing $Y = y$. This is sometimes called a missing data problem (see section 1.4.4). Often the EM algorithm (Dempster, Laird and Rubin (1977)) provides a method to find the maximum likelihood estimator of θ_0. There are two drawbacks. The first is that it is not known how many iteration steps are needed to bring one close enough to convergence. The other is that sometimes the E-step, computation of the conditional expectation of the likelihood given the data, is impossible.

In this latter case, the stochastic version of the EM algorithm (StEM) may be a viable alternative. For a review and large sample results see Nielsen (1997a and b). The algorithm works as follows. Suppose the complete data maximum likelihood estimator is readily computable: $\hat{\theta}_{MLE} = M(X)$.

1. Fix a $\theta(0)$ in Θ;
2. Sample $X(1)$ from $P_{\theta(0)}(.|Y = y)$;
3. Set $\theta(1) = M(X(1))$.

By iterating this procedure, we obtain a sequence $\theta(0), \theta(1), \theta(2), \ldots$. If step 2 is carried out using independent Unif(0,1) variables for each t, the sequence $\theta(t)$ is a time homogeneous Markov chain. Under certain conditions it is also ergodic. If so, the algorithm converges in that the $\theta(t)$ converge in distribution to a random variable, say $\hat{\theta}$, which is distributed according to the stationary distribution of the Markov chain. Then $\hat{\theta}$ is the StEM estimate. In other words, a StEM estimate is a sample from the stationary distribution of the 'StEM Markov chain'.

It is not clear for how long we should run the StEM chain to allow it to reach equilibrium. This lack of a termination rule seems to be something StEM and EM have in common. However, we shall find that in some cases

we can apply coupling-from-the-past to obtain a sample that is guaranteed to come from the stationary distribution of the StEM chain.

1.5.3 Perfect Stochastic EM

In this section we combine the ideas from the previous two sections. We use the set-up and notation of the preceding section. Let us suppose that the parameter space Θ admits a partial ordering $\preceq$. Fix a time $-T < 0$ in the past. Suppose that for $t = 0, -1, -2, \ldots$ we can construct independent collections of random variables $\{X_\theta(t),\ \theta \in \Theta\}$ such that

1. $X_\theta(t) \ \sim \ P_\theta(.|Y = y)$
2. $\theta \preceq \theta' \Rightarrow M(X_\theta(t)) \preceq M(X_{\theta'}(t))$ almost surely.

Recall that $M(X)$ is the complete data maximum likelihood estimator of θ.

To simulate a Markov chain $\theta_{-T}(-T), \theta_{-T}(-T+1), \ldots, \theta_{-T}(0)$ we define random transition maps

$$H_t(\theta(t-1)) = M(X_{\theta(t-1)}(t)), \qquad t = -T+1, -T+2, \ldots, 0.$$

Now suppose that there are 'minimal' and 'maximal' elements $\underline{\theta}$ and $\overline{\theta}$ such that $\underline{\theta} \preceq \theta \preceq \overline{\theta}$ for all $\theta \in \Theta$. Consider two coupled paths $\theta^L_{-T}(.)$ and $\theta^U_{-T}(.)$ starting at $\theta^L_{-T}(-T) = \underline{\theta}$ and $\theta^U_{-T}(-T) = \overline{\theta}$. It now suffices to check if these paths have coalesced at time zero, i.e. if $\theta^L_{-T}(0) = \theta^U_{-T}(0)$. If so, then we have a perfect StEM estimate $\hat{\theta} = \theta^L_{-T}(0)$. If not we have to go back further in time.

We demonstrate the Perfect StEM algorithm in a simple example. We also apply it to a more involved problem in section 4.4.2.

An Example

Suppose $X = (X_1, X_2, \ldots, X_n)$ is an i.i.d. sample from $P_{\theta_0} = \text{Exp}(\theta_0)$; the exponential distribution with intensity (reciprocal of the mean) θ_0. We wish to estimate the intensity θ_0. Suppose that θ_0 is known to be bigger than some $\underline{\theta} > 0$, that is $\theta_0 \in \Theta = [\underline{\theta}, \infty)$ Of course the maximum likelihood estimator of θ_0 is $\hat{\theta}_{MLE} = M(X) = (n/\sum X_i) \vee \underline{\theta}$. Suppose we only observe

$$\tilde{X}_i = X_i \wedge C \qquad \Delta_i = \mathbf{1}_{\{X_i > C\}},$$

for some fixed positive constant C. Write $Y_i = (\tilde{X}_i, \Delta_i)$ for the observed data.

The maximum likelihood estimator of θ_0 based on the observed data is known to be the 'occurrence' divided by the 'exposure'.

$$\frac{n - \sum \Delta_i}{\sum \tilde{X}_i}.$$

So there is really no need to apply the StEM algorithm here. The purpose of this example is strictly illustrative.

We now describe how the StEM algorithm works here. When multiplying vectors we mean coordinate wise multiplication (mapping two vectors to one vector).

1. Fix $\theta(0) > 0$
2. Generate $X(1) \sim P_{\theta(0)}(X|Y = y)$:

$$X(1) = \tilde{X} + \Delta E, \quad \text{where } E = (E_1, \dots, E_n),\ E_i \sim \text{Exp}(\theta(0))$$

3. $\theta(1) = M(X(1)) = \frac{n}{\sum X_i(1)} \vee \underline{\theta}$.

Repeating this procedure, we obtain an ergodic Markov chain $\theta(0), \theta(1), \dots$.

Now we demonstrate how we can apply perfect sampling to this chain. Recall the usual ordering on $\mathbb{R}^n$: $x \leq y$ if $x_1 \leq y_1$ and $x_2 \leq y_2 \dots$ and $x_n \leq y_n$. Note that if $x \geq y$ then $M(x) \leq M(y)$. Set $\overline{\theta} = M(\tilde{X})$. Since, for all t, $X(t) \geq \tilde{X}$ it follows that $\theta(t) = M(X(t))$ will always be less than $\overline{\theta}$. Hence, we may reduce our parameter-space Θ to $[\underline{\theta}, \overline{\theta}]$. Of course $\underline{\theta}$ and $\overline{\theta}$ are minimal and maximal elements of Θ with the usual ordering.

We now construct collections $\{X_\theta(t),\ \theta \in \Theta\}$ as required at the beginning of this section. For $t = -T+1, -T+2, \dots 0$ and $i = 1, 2, \dots, n$ generate independent

$$E_{\underline{\theta},i}(t) \sim \text{Exp}(\underline{\theta}) \quad \text{and} \quad E_i(t) \sim \text{Exp}(\overline{\theta} - \underline{\theta})$$

and define

$$E_{\theta,i}(t) = E_{\underline{\theta},i}(t) \wedge \frac{\overline{\theta} - \underline{\theta}}{\theta - \underline{\theta}} E_i(t).$$

Then, evidently, $E_{\theta,i}(t) \sim \text{Exp}(\theta)$, and $\theta \leq \theta'$ implies $E_{\theta,i}(t) \geq E_{\theta',i}(t)$. For $t = 0, -1, -2, \dots$ and all $\theta \in \Theta$ we define

$$X_{\theta,i}(t) = \widetilde{X}_i + \Delta_i E_{\theta,i}(t),$$

Thus, for $t = 0, -1, -2, \dots$ we have collections of random variables $\{X_\theta(t),\ \theta \in \Theta\}$ such that

1. $X_\theta(t) \ \sim \ P_\theta(.|Y = y)$

2. $(\theta \leq \theta' \Rightarrow M(X_\theta(t)) \leq M(X_{\theta'}(t)))$ a.s.

as required. We can now run a 'lower' chain $\theta^L_{-T}(-T) = \underline{\theta}, \theta^L_{-T}(-T+1), \ldots, \theta^L_{-T}(0)$ and an 'upper' chain $\theta^U_{-T}(-T) = \overline{\theta}, \theta^U_{-T}(-T+1), \ldots, \theta^U_{-T}(0)$ and check if $\theta^L_{-T}(0) = \theta^U_{-T}(0)$. We do need to make sure that the algorithm will terminate in finite time. It is enough to check that for some fixed $-T$ the event $\theta^L_{-T}(0) = \theta^U_{-T}(0)$ has positive probability. Well,

$$\begin{aligned} \Pr(\theta^L_{-1}(0) = \theta^U_{-1}(0)) &\geq \Pr(X_{\underline{\theta},i}(0) = X_{\overline{\theta},i}(0), \quad \forall i) \\ &\geq \Pr(\mathrm{Exp}(\underline{\theta}) \leq \mathrm{Exp}(\overline{\theta} - \underline{\theta}))^n > 0. \end{aligned}$$

This concludes the example.

The difficulty in making the StEM algorithm perfect lies in the construction of collections of random variables satisfying points 1 and 2 at the beginning of this section. There probably is no general recipe to achieve this. Here we ask ourselves when such a construction is at all possible. It turns out that the concept of *realizable monotonicity* (Fill and Machida (1998)) is essentially what we need.

Realizable monotonicity is very close to stochastic monotonicity, which is a more familiar concept and which is also easier to check. Suppose that the complete data are random elements in some space E, and that the complete data maximum likelihood estimator of a parameter $\theta \in \Theta$ is given by a function M on E. We assume that the parameter space Θ admits a partial order $\preceq_1$.

Suppose that E admits a partial ordering $\preceq_2$ such that, for any $x_1, x_2 \in E$

$$x_1 \preceq_2 x_2 \;\Rightarrow\; M(x_1) \preceq_1 M(x_2).$$

Then we need to be able to construct a collection $\{X_\theta,\ \theta \in \Theta\}$ such that

1. $X_\theta \sim P_\theta$,

2. $\theta \preceq_1 \theta' \Rightarrow X_\theta \preceq_2 X_{\theta'}$, almost surely.

We shall have a look at two notions of monotonicity for a collection $\{P_\theta\ ,\ \theta \in \Theta\}$ of probability measures: realizable and stochastic monotonicity.

Definition 1.4. *The collection* $\{P_\theta\ ,\ \theta \in \Theta\}$ *is called* realizably monotone *if there exists a collection of random variables* $\{X_\theta\ ,\ \theta \in \Theta\}$ *such that*

1. $X_\theta \sim P_\theta$,

2. $\theta \preceq_1 \theta' \Rightarrow X_\theta \preceq_2 X_{\theta'}$, *almost surely.*

A subset U of E is said to be an *up-set* in $(E, \preceq_2)$ if $y \in U$ whenever $x \in U$ and $x \preceq_2 y$. If P_1 and P_2 are probability measures on E then P_1 is *stochastically smaller* than P_2 if $P_1(U) \leq P_2(U)$ for all up-sets U in $(E, \preceq_2)$. we then write $P_1 \preceq_2^D P_2$.

Definition 1.5. *The collection* $\{P_\theta\ ,\ \theta \in \Theta\}$ *is called* stochastically monotone *if*

$$\theta \preceq_1 \theta' \Rightarrow P_\theta \preceq_2^D P_{\theta'},\ a.s.$$

Realizable monotonicity is clearly what we need, while stochastic monotonicity is usually easier to check. It is easily seen that realizable monotonicity implies stochastic monotonicity. That the converse is not always true is demonstrated by an example in Ross (1993). However, for various finite classes of $(E, \preceq_2)$, Fill and Machida (1998) give conditions on finite index sets $(\Theta, \preceq_1)$ such that stochastic monotonicity *does* imply realizable monotonicity. For instance, realizable and stochastic monotonicity are equivalent when $(E, \preceq_2)$ or $(\Theta, \preceq_1)$ is a finite linearly ordered set. Recall that a set is linearly ordered if each pair of elements is comparable.

Fill and Machida (1998) coined the term 'realizable monotonicity' and noted its relevance for perfect sampling. Their results for finite sets are all the more useful because of the following unpublished result by Ross (1993).

Theorem 1.3. *Suppose that* $(\Theta, \preceq_1)$ *is a partially ordered set and* $(E, \preceq_2)$ *is a complete separable metric space with closed partial order. Then* $\{P_\theta,\ \theta \in \Theta\}$ *is realizably monotone if and only if for every finite* $\Psi \subseteq \Theta$ $\{P_\theta,\ \theta \in \Psi\}$ *is realizably monotone.*

Now if for some separable set with a closed partial order we can use the results of Fill and Machida (1998) to check realizable monotonicity for all its finite subsets then Ross's theorem allows us to conclude realizable monotonicity for the entire infinite set.

Chapter 2

Laslett's line segments

2.1 Introduction

Almost to decades ago a study was to be made into the hazards of nuclear fuel waste disposal in underground excavations in selected plutonic rock masses of the Canadian Shield. Experiments had to be done related to thermal heating and hydraulic conductivity along fractures in the rock. As part of this research the construction was planned of an underground research laboratory within the granitic rock of the Lac du Bonnet batholith in southeastern Manitoba. For more information we refer to Stone et al. (1984). Figure 1.1 in the introduction to this thesis shows part of a map from that report of roughly 160 by 160 meters of fractures in the rock at the Lac du Bonnet site.

This particular data set enters the statistical literature with Chung (1989a and b). The statistical problem is to estimate from figure 1.1 the distribution of the lengths of the fractures. Estimation of the length distribution of line segments observed through a bounded window is sometimes called *Laslett's line segment problem* after Laslett (1982a and b).

We are not sure if the underground laboratory was ever built, but quite apart from the original motivation the estimation problem presents a very interesting statistical challenge. We encounter three main difficulties. First, we have to deal with censoring, since most of the fractures are only partly observed as the rock is only partly exposed due to vegetation, soil and water. Secondly, the sample of (partly) observed cracks is biased, because longer cracks stand a better chance of being observed than shorter ones. Thirdly, the area of exposed rock where we observe the cracks is not convex. This means that we might observe several fragments of a single crack. A single glance at Figure 1.1 will convince the reader that it would be very difficult to assess if two observed fragments belong to the same underlying fracture.

The first two problems, censoring and length bias, have been studied rather extensively Laslett (1982a and b), van der Laan (1993), Gill (1994) and Wijers (1995). The third difficulty, non-convexity of the observation window, is treated here for the first time. Also, in the above contributions the assumption is made that the locations of the fractures are distributed according to a Poisson process. We will obtain our results under far more general ergodicity assumptions.

A line segment process can be modeled as a marked point process, where the points indicate the locations of the segments and the marks their lengths and orientations. Assuming that the locations are scattered as a Poisson process greatly facilitates any statistical analysis because then, conditional on their number, the segments are independent. We will argue that an analysis under the Poisson assumption is still relevant when that assumption is not fulfilled. We feel that this is an important point and we do not want it to be snowed under by details. Therefore we give a general—albeit somewhat heuristic—discussion in the next section, which will then serve as a blueprint for the remainder of the chapter.

The reader will search in vain for the actual estimate of the length distribution of the Canadian fractures. As this thesis is on its way to the printer, we have not yet implemented our estimator and extracted the necessary data from Figure 1.1.

2.1.1 Inference for ergodic point processes

Consider a point process (see section 1.3) $\Phi = \{(T_i, X_i)\}$ on $\mathbb{R} \times \mathcal{X}$ with intensity measure

$$\mathrm{d}\Lambda_{\lambda,F}(t,x) = \lambda \mathrm{d}t \mathrm{d}F(x),$$

where F is known to belong to some class $\mathcal{F}$.

We think of Φ as a marked point process on $\mathbb{R}$ with marks X_i in a mark space $\mathcal{X}$. For instance, the T_i could be the locations of cars parked along a street and the X_i could be their make. For another example, the X_i could determine the length of a line segment starting at T_i. Evidently, this is a very general set up and in fact our restriction to T_i taking values in $\mathbb{R}$ is not at all necessary for our discussion.

We are primarily interested in estimating the mark distribution F_0 which is supposed to be in some collection $\mathcal{F}$. The 'nuisance' parameter λ is also assumed to be unknown to us.

We do not observe Φ entirely. Suppose that of points (T_i, X_i) outside a set $W \subseteq \mathbb{R} \times \mathcal{X}$ we observe nothing, while if $(T_i, X_i) \in W$ we only observe the result of some known function $Y_i = g(T_i, X_i)$. The mapping g may depend

on W. For instance, if the (T_i, X_i) represent line segments $[T_i, T_i + X_i]$ then $Y_i = g(T_i, X_i)$ could be $|[T_i, T_i + X_i] \cap [0, 1]|$, the length of the intersection with the unit interval. Then $W = \{(t, x) : [t, t + x] \cap [0, 1] \neq \emptyset\}$.

We now define a new 'point process' of observations

$$\Psi = \{Y_i\} = \{g(T_i, X_i) : (T_i, X_i) \in W\}.$$

It is convenient to think of the Y_i as a point process because the observed data will often be of similar structure as the incompletely observed model Φ.

Now suppose that Φ is *Poisson* (cf. section 1.3 on Poisson point processes). Then the intensity $\lambda \mathrm{d}t\mathrm{d}F(x)$ completely determines the distribution of Φ. The distribution of Ψ is also determined and can be described as follows.

Define, for all $F \in \mathcal{F}$

$$\mu_F = \int \int_W \mathrm{d}t\mathrm{d}F(x),$$

and assume it to be finite. The reader should mistake μ_F for the mean of F. Now,

1. let N be distributed as a Poisson random variable with mean $\lambda\mu_F$;
2. conditional on $N = n$ draw an i.i.d. sample (T_i, X_i) of size n from the 'normalized intensity' measure

$$\mathbf{1}_W(t, x)\frac{1}{\mu_F}\mathrm{d}t\mathrm{d}F(x); \tag{2.1}$$

3. compute $Y_i = g(X_i, T_i)$;
4. identify samples $Y_1, \ldots, Y_n$ that are equal up to ordering.

It is easy to write down the likelihood for the pair (λ, F) at the data $(N, Y_1, \ldots, Y_N)$.

$$\frac{e^{-\lambda\mu_F}}{N!}(\lambda\mu_F)^N \prod_{i=1}^{N} \iint_{(t,x)\in g^{-1}(Y_i)\cap W} \frac{1}{\mu_F}\mathrm{d}t\mathrm{d}F(x)N!.$$

We find the *profile likelihood* for estimating F by first fixing F and replacing λ by its maximum likelihood estimator N/μ_F. We find

$$e^{-N}N^N \prod_{i=1}^{N} \iint_{(t,x)\in g^{-1}(Y_i)\cap W} \frac{1}{\mu_F}\mathrm{d}t\mathrm{d}F(x).$$

This profile likelihood can be interpreted as the likelihood of a missing data problem (see section 1.4.4.) To make this clear we re-parameterize the problem. Define for all x

$$W(x) = \{t \in \mathbb{R} : (t, x) \in W\}.$$

Also define for all $F \in \mathcal{F}$ a transformation

$$\mathrm{d}V_F(x) = \frac{|W(x)|}{\mu_F}\mathrm{d}F(x). \tag{2.2}$$

For later use we have the inverse transformation

$$\mathrm{d}F_V(x) = \left(\int_{x'} \frac{1}{|W(x')|}\mathrm{d}V(x')\right)^{-1} \frac{1}{|W(x)|}\mathrm{d}V(x). \tag{2.3}$$

so that $F_{V_F} = F$. Define $\mathcal{V} = \{V_F,\ F \in \mathcal{F}\}$.

We can rewrite the distribution of (T, X) as

$$\mathbf{1}_W(t, x)\frac{\mathrm{d}t}{|W(x)|}\mathrm{d}V(x). \tag{2.4}$$

Our missing data problem takes the form

1. Sample X from $V \in \mathcal{V}$
2. Given $X = x$ sample T from the uniform distribution on $W(x)$
3. Set $Y = g(T, X)$

In line with section 1.4.4 we can also describe the situation as follows. Defining a Markov kernel

$$K(\mathrm{d}y; x) = \int_{W(x)} \delta_{g(t,x)}(\mathrm{d}y)\frac{\mathrm{d}t}{|W(x)|},$$

we have that

$$KV(\mathrm{d}y) = \int K(\mathrm{d}y; x)\mathrm{d}V(x)$$

is the distribution of $Y = g(T, X)$. The model for the observations Y_i is thus

$$K\mathcal{V} = \{KV : V \in \mathcal{V}\}.$$

Conditionally on $N = n$, the maximum likelihood estimator $\hat{V}_n$ satisfies

$$\int \log \frac{\mathrm{d}KV}{\mathrm{d}K\hat{V}_n}\mathrm{d}\mathbb{P}_n \geq 0 \qquad \forall V \in \mathcal{V}.$$

Suppose that F_0 is the 'true' parameter. The MLE $\hat{V}_n$ is of course supposed to estimate $V_0 = V_{F_0}$ (cf. (2.2)). We define $\hat{F}_n = F_{\hat{V}_n}$ (cf. (2.3)) as the MLE of F_0.

Now if the model $\mathcal{V}$—and hence $K\mathcal{V}$—is convex, we can attempt to use Theorem 1.2 to prove the convergence of $K\hat{V}_n$ to KV_0. To conclude convergence of $\hat{V}_n$ to V_0 we of course need to have identifiability: If $V \neq V'$ then $KV \neq KV'$. Ultimately, convergence of $\hat{V}_n$ to V_0 hopefully implies that $\hat{F}_n = F_{\hat{V}_n}$ tends to $F_{V_{F_0}} = F_0$.

To use Theorem 1.2 we need a sequence V_n such that KV_n tends to KV_0 and prove

$$\int \frac{\mathrm{d}KV_n}{\mathrm{d}K\hat{V}_n} \mathrm{d}(\mathbb{P}_n - KV_n) \to 0. \tag{2.5}$$

Such a proof, and indeed the choice of V_n, will rely on the fact that $\mathbb{P}_n$ converges to KV_0.

But what if Φ is *not* Poisson? Then we can no longer condition on the number of observations and treat them as an i.i.d. sample. We cannot write down the likelihood and maximize it. Also, the above asymptotics involving the number of observations tending to infinity become meaningless.

In spite of all this, we propose to use exactly the same estimator. Though no longer the maximum likelihood estimator, one still expects it to have nice properties. We explain.

First, some notation. We can consider Ψ as a random set of points but also as a random measure. By $\Psi(A)$ we mean the number of points of Ψ that fall in some measurable set A. In fact, we write $\Psi(A) = \int_A \mathrm{d}\Psi(y)$. Upon normalization, we obtain a random probability measure

$$\frac{\Psi(\mathrm{d}y)}{\Psi(g^{-1}(W))}.$$

Now we can define a 'maximum likelihood estimator' $\hat{V}$ as satisfying

$$\int \left(\log \frac{\mathrm{d}KV}{\mathrm{d}K\hat{V}} \right) \mathrm{d} \frac{\Psi(y)}{\Psi(g^{-1}(W))} \geq 0 \qquad \forall V \in \mathcal{V},$$

and we define $\hat{F} = F_{\hat{V}}$. Of course, all this is just notation; the estimator we just defined coincides exactly with the maximum likelihood estimator we defined earlier for the case where Φ is Poisson.

We will study the asymptotics of the present estimator as more and more of the underlying process Φ is revealed. Suppose we have a sequence $W = W_1 \subseteq W_2 \subseteq, \ldots$. Define

$$\Psi_i = \{g_i(T_i, X_i) : (T_i, X_i) \in W_i\}.$$

Note that the g_i may depend on W_i. Since we defined $W = W_1$ we have $\Psi = \Psi_1$.

Define

$$\mu_{i,F} = \int \int_{W_i} \mathrm{d}t \mathrm{d}F(x),$$

and

$$W_i(x) = \{t \in \mathbb{R} : (t, x) \in W_i\}.$$

Similarly to (2.2) and (2.3), define

$$\mathrm{d}V_{i,F}(x) = \frac{|W_i(x)|}{\mu_{i,F}} \mathrm{d}F(x) \tag{2.6}$$

and

$$\mathrm{d}F_{i,V}(x) = \left(\int_{x'} \frac{1}{|W_i(x')|} \mathrm{d}V(x') \right)^{-1} \frac{1}{|W_i(x)|} \mathrm{d}V(x). \tag{2.7}$$

Define

$$\mathcal{V}_i = \{V_{i,F},\ F \in \mathcal{F}\}.$$

Finally, define a Markov kernel

$$K_i(\mathrm{d}y; x) = \int_{W_i(x)} \delta_{g_i(t,x)}(\mathrm{d}y) \frac{\mathrm{d}t}{|W_i(x)|}.$$

We now have a sequence of statistical experiments: observation of Ψ_i. The model for the distribution of the data *under the Poisson assumption* changes with each experiment as the mapping g_i may depend on the 'window' W_i. Thus we have a sequence of models

$$K_i \mathcal{V}_i = \{K_i V : V \in \mathcal{V}_i\} = \{K_i V_{i,F} : F \in \mathcal{F}\}.$$

We can define a sequence of estimators $\hat{V}_i$ by requiring

$$\int \left(\log \frac{\mathrm{d}K_i V}{\mathrm{d}K_i \hat{V}_i} \right) \mathrm{d} \frac{\Psi_i(y)}{\Psi_i(g_i^{-1}(W_i))} \geq 0 \qquad \forall V \in \mathcal{V}_i.$$

We set $\hat{F}_i = F_{i,\hat{V}_i}$

We would like to use the ideas behind theorem 1.2 to prove that the $\hat{F}_i$ tend to F_0. However, first we must try to use ergodic properties of Φ and the sequence W_i to show that for all $F \in \mathcal{F}$ there exist $V_{\infty,F}$ and $K_\infty V_{\infty,F}$ such that

$$V_{i,F} \to V_{\infty,F} \quad \text{and} \quad K_i V_{i,F} \to K_\infty V_{\infty,F}. \tag{2.8}$$

These convergences are established for Laslett's problem in section 2.5.3. There it is also made clear in which sense they hold.

Next, we need a sequence V_i such that $K_i V_i$ tends to $K_\infty V_{\infty,F_0}$. Then we show that

$$\int \left(\frac{\mathrm{d}K_i V_i}{\mathrm{d}K_i \hat{V}_i} \right) \mathrm{d} \left(\frac{\Psi_i(y)}{\Psi_i(g^{-1}(W))} - K_i V_i \right) \to 0. \tag{2.9}$$

From this we can conclude that $K_i \hat{V}_i$ converges to $K_\infty V_{\infty,F_0}$. Identifiability will then imply that also $\hat{V}_i$ tends to V_{∞,F_0}. This will (at least for Laslett's problem) imply that $\hat{F}_i = F_{i,\hat{V}_i}$ tends to $F_{\infty,V_{\infty,F_0}} = F_0$.

For the convergence of (2.9), we will need that

$$\frac{\Psi_i(\mathrm{d}y)}{\Psi_i(g^{-1}(W))} \to K_\infty V_{\infty,F_0}(\mathrm{d}y).$$

For Laslett's problem this is shown in section 2.5.3.

2.1.2 Laslett's line segment problem

A line segment process is conveniently modeled by a marked point-process in the plane, where the points indicate the location of, say, the left endpoints and the marks indicate orientation and length. We consider a point process $\Phi = \{(\vec{S}_i, X_i, \Theta_i)\}$ on $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$. We could call $\mathbb{R}^+ \times (-\pi/2, \pi/2)$ the mark-space. Take the $\vec{S}_i$ to be left endpoints of the line segments and let X_i and Θ_i be their lengths and orientations. Let us use square brackets, writing $[s, x, \vartheta]$, to denote a line segment in $\mathbb{R}^2$, rather than a point in $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$.

Suppose that Φ is stationary (with respect to shifts on $\mathbb{R}^2$). Also, suppose that each segment's length and orientation are independent, in the sense that Φ has an intensity measure of the form $\lambda \mathrm{d}\vec{s}\mathrm{d}F(x)\mathrm{d}K(\vartheta)$, where $\lambda > 0$ and F and K are distribution functions on $\mathbb{R}^+$ and $(-\pi/2, \pi/2)$, respectively. Let μ denote the mean of F, and suppose that it is finite. Note that by giving this intensity we have not completely described the distribution of Φ.

Now we assume that K is known. Our goal is nonparametric estimation of the length distribution F. This means that F belongs to a *model*, say $\mathcal{F}$, which consists of all probability distributions on the $\mathbb{R}^+$. Lok (1994) considers just the opposite situation: she assumes F to be known, and estimates K. If both F and K are unknown we could alternate estimation of F as presented here and estimation of K as presented by Lok (1994).

As a technical aside we point out the following. For reasons of mathematical convenience we decided to let the orientations range in $(-\pi/2, \pi/2)$,

excluding vertical line segments with orientation $\pi/2$. However, there is no loss of generality in doing so, because we can choose the orientation of the entire process so that K does not have a jump at $\pi/2$.

We have defined a stationary line segment process and stated our aim to estimate the distribution of the lengths of the line segments. We now describe the data that are available to us. Let $\mathcal{W} \subset \mathbb{R}^2$ be a random closed set and let B be the unit square; $B = [0,1] \times [0,1]$. Suppose we can only observe the intersections of the line segments with $W = \mathcal{W} \cap B$. In the Canadian dataset, c.f. Figure 1.1, B corresponds to the 160×160 meters square area and W is the irregular black region through which we observe the fractures.

After we come up with an estimator of F, shall study its asymptotic properties. In spatial statistics basically two types of asymptotics are usually considered (Cressie (1991) p. 100). The first is 'infill' asymptotics where the observation region is kept fixed, but within that region the number of observations increases. The other is 'increasing-domain' asymptotics, where the observation region is expanded. In fact, we shall define $B_n = [0,n] \times [0,n]$ and consider observation of the line segment process Φ through $\mathcal{W} \cap B_n$ as n tends to infinity. When doing so, we shall have to impose some ergodicity assumptions on Φ and $\mathcal{W}$.

2.1.3 Some history

We now present a brief overview of the work that has already been done by other authors. Much effort has been put into the one dimensional case where line segments are scattered according to a Poisson process on the real line and the segments are observed (without occlusion) through an interval. Laslett (1982 a) showed how the EM algorithm can be used to obtain the (sieved) nonparametric maximum likelihood estimator. Wijers (1995b) has shown it to be consistent. Gill (1994), van der Laan (1995) and Wijers (1995a) have (jointly) established its asymptotic normality and efficiency. Many of the methods that have been developed for the one-dimensional case carry over to the two dimensional case, as long as the observation window is convex. The trick is to subdivide the plane into parallel strips of infinitesimal width and doing so in every direction. Then through each strip we observe—as it were—a one-dimensional line segment process and then we integrate over all strips and all directions. We shall perform such calculations later on. At that time this approach will be made more precise.

Two things should be noted about assuming Φ to be Poisson and W to be convex. First, the Poisson assumption allows us to condition on the number of observations, after which we have an i.i.d. sample and we can apply standard estimation techniques. Secondly, the fact that the observation window W is

convex ensures that we can not observe more than one fragment of a single underlying line segment.

2.1.4 Scope of our contribution

We study the two dimensional line segment process observed through a non-convex window. This is prompted by the fact that the exposed rock surface of the Canadian data is obviously non-convex.

If the observation window is non-convex we might observe several fragments of a single line segment. It may occur, as it does with the Canadian data, that it is near impossible to decide if two fragments belong to the same fracture. As a consequence the observed fragments may be dependent in a way unknown to us.

In this situation it is of no use to assume that the segments are scattered as a Poisson process. Even if we could condition on the number of different segments that are observed, the fragments are not necessarily distributed as an i.i.d. sample.

However, we propose that the statistician analyze the data "as if" the observed fragments are scattered according to a Poisson process. In this simpler situation we find the non-parametric maximum likelihood estimator of the length distribution of the fractures. We shall show consistency of the estimator without using the Poisson assumption.

Under the Poisson assumption the estimator may well be efficient, though we have not tried to prove this. Without the Poisson assumption, it will certainly not be efficient because we effectively ignore dependencies among the data, thus throwing away information. However, we feel confident that this loss of information is minor.

2.2 A re-parameterization

Above we briefly explained how to extend the one-dimensional case to the two-dimensional case. We now demonstrate in detail how this works. The main step is to re-parameterize the location of each line segment relative to the unique line in the plane on which the segment lies.

Recall that we are studying a point-process $\Phi = \{(\vec{S}_i, X_i, \Theta_i)\}$ on $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$, with finite intensity $\lambda \mathrm{d}\vec{s}\mathrm{d}F(x)\mathrm{d}K(\vartheta)$. The $\vec{S}_i$ denote left endpoints of line segments of length X_i and orientation Θ_i. It is straightforward to represent the locations of the left endpoints relative to rotated

coordinate axes. Define two functions

$$t_1(s_1, s_2, \vartheta) = s_1 \cos(\vartheta) + s_2 \sin(\vartheta)$$
$$t_2(s_1, s_2, \vartheta) = -s_1 \sin(\vartheta) + s_2 \cos(\vartheta).$$

For fixed ϑ this map represents $(s_1, s_2) \in \mathbb{R}^2$, with respect to the coordinate axes tilted counter-clockwise over an angle ϑ, cf. Figure 2.1. Now consider the

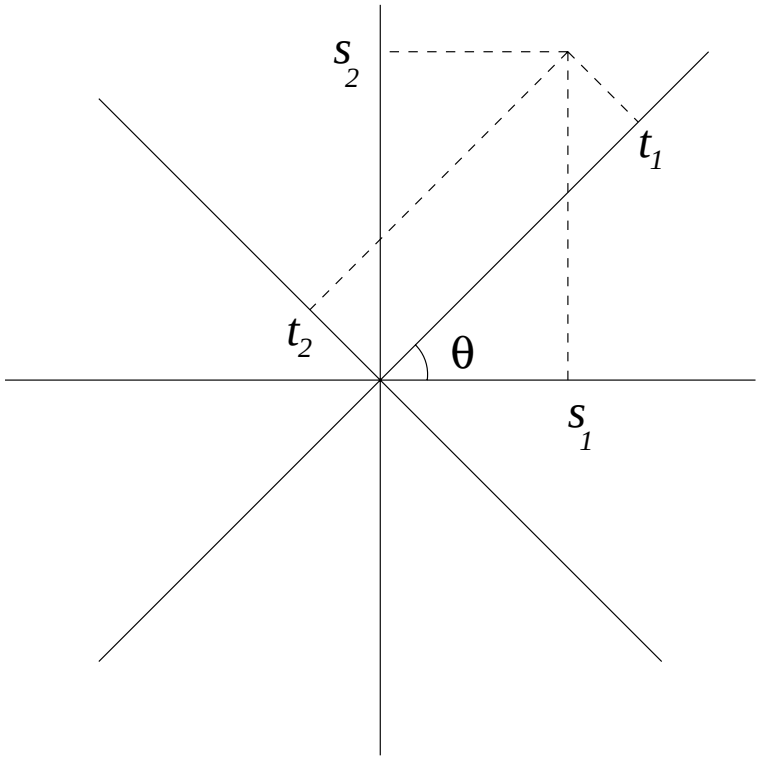

Figure 2.1: Re-parameterization of locations.

map that assigns $(t_1(s_1, s_2, \vartheta), t_2(s_1, s_2, \vartheta), x, \vartheta)$ to (s_1, s_2, x, ϑ). The image of Φ under this mapping is a new point-process $\Phi' = \{(\vec{T}_i, X_i, \Theta_i)\}$ on $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$ with intensity

$$\lambda \mathrm{d}\vec{t}\mathrm{d}F(x)\mathrm{d}K(\vartheta).$$

This follows since Lebesgue measure is invariant under rotation.

Write $\ell(d, \vartheta)$ for the line with orientation ϑ at (signed) distance d from the origin. In our new parameterization, a line segment $[\vec{t}, x, \vartheta]$ lies exactly on the line $\ell(t_2, \vartheta)$.

Denote $\mathcal{W}(d,\vartheta) = \mathcal{W} \cap \ell(d,\vartheta)$. Let us assume that $\mathcal{W}(d,\vartheta)$ is always the union of at most a countable number of intervals: $\mathcal{W}(d,\vartheta) = \cup_m \mathcal{W}(d,\vartheta,m)$. The situation is depicted in Figure 2.2.

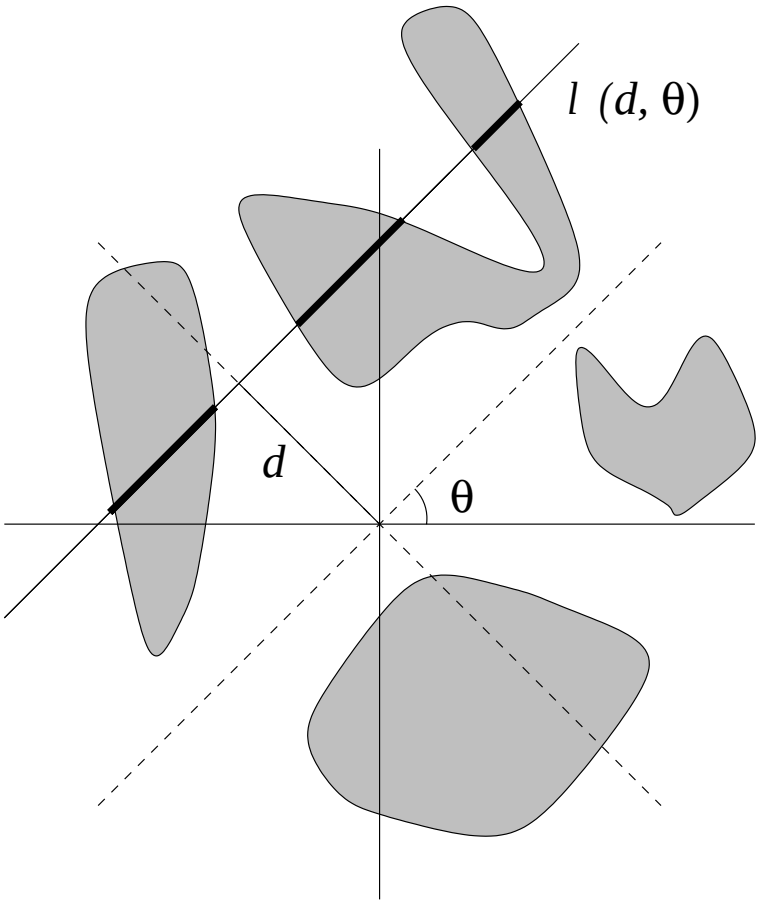

Figure 2.2: The line $\ell(d,\vartheta)$ intersecting (part of) the random set $\mathcal{W}$ (grey) producing closed intervals $\mathcal{W}(d,\vartheta,m)$ (bold).

Since $\mathcal{W}$ is non-convex, the intersection of a line segment $[\vec{t},x,\vartheta]$ with $\mathcal{W}$ might consist of several *fragments*. In other words, we may have a non-empty intersection of $[\vec{t},x,\vartheta]$ with $\mathcal{W}(t_2,\vartheta,m)$ for several values of m. Conversely, every fragment corresponds to a line segment $[\vec{t},x,\vartheta]$ and a positive number m. Our use of the words 'fragment' and 'segment' will have always precisely this meaning.

Consider countably many identical copies of $\Phi' = \{(\vec{T}_i, X_i, \Theta_i)\}$, one for each m. The result, $\Phi'' = \Phi' \times \mathbb{N}$ is a stationary point-process on $\mathbb{R}^2 \times \mathbb{R}^+ \times$

$(-\pi/2, \pi/2) \times \mathbb{N}$ with intensity

$$\lambda \mathrm{d}\vec{t}\mathrm{d}F(x)\mathrm{d}K(\vartheta)\mathrm{d}\mathcal{C}(m),$$

where $\mathcal{C}$ denotes counting measure on $\mathbb{N}$. This new point-process Φ'' can represent any fragment that could possible arise from the line segments of Φ'. A fragment $[\vec{t}, x, \vartheta, m]$ is the intersection of a line segment $[\vec{t}, x, \vartheta]$ with $\mathcal{W}(t_2, \vartheta, m)$. Of course such an intersection may very well be empty.

We must now deal with further edge effects because we observe the line segments only through the intersection of $\mathcal{W}$ with the unit square $B = [0,1] \times [0,1]$. For all $(d, \vartheta, m) \in \mathbb{R} \times (-\pi/2, \pi/2) \times \mathbb{N}$ let

$$W(d, \vartheta, m) = \mathcal{W}(d, \vartheta, m) \cap B$$

Define

$$\mathcal{D} = \{(\vec{t}, x, \vartheta, m) \in \mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2) \times \mathbb{N} : W(t_2, \vartheta, m) \neq \emptyset\}.$$

Only points in $\mathcal{D}$ can correspond to fragments that could be observed through W.

We define two functions on $\mathcal{D}$, $t : \mathcal{D} \to \mathbb{R}$ and $l : \mathcal{D} \to \mathbb{R}^+$, as follows. Let $t(\vec{t}, x, \vartheta, m)$ be the distance from the left endpoint of the line segment $[\vec{t}, x, \vartheta]$ to the left endpoint of $W(t_2, \vartheta, m)$, which we take negative if the first is to the left of the latter. Let $l(\vec{t}, x, \vartheta, m)$ be the (strictly positive) length of the interval $W(t_2, \vartheta, m)$.

Now consider the map that assigns $(t(\vec{t}, x, \vartheta, m), x, l(\vec{t}, x, \vartheta, m))$ to each 'potential fragment' $(\vec{t}, x, \vartheta, m) \in \mathcal{D}$. The image under this map of $\Phi'' \cap \mathcal{D}$ is a new point-process $\Psi = \{(T_i, X_i, L_i)\}$ on $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$. The points of Ψ can be interpreted as segments $[T_i, T_i + X_i]$ on the real line which can be observed through intervals $[0, L_i]$. These intersections $[T_i, T_i + X_i] \cap [0, L_i]$ are the observed fragments.

We have now accomplished what we set out to do; we are now in the one-dimensional case. Wijers (1995) studied the case where the L_i are constant almost surely and the T_i follow a homogeneous Poisson process. We shall be able to make much use of his methods. However, we wish to stress that the Poisson assumption certainly does not hold here because some of the $[T_i, T_i + X_i]$ are actually different representations of the same line segment!

We now demonstrate that the intensity of Ψ is given by

$$\lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)$$

where

$$\mathrm{d}\nu(l) = \int_\vartheta \int_t \sum_m \delta_{l(t,\vartheta,m)}(\mathrm{d}l)\mathrm{d}t\mathrm{d}K(\vartheta).$$

Here $\delta_{l(t,\vartheta,m)}(\mathrm{d}l)$ is Dirac measure putting point mass at $l(t,\vartheta,m)$.

We must show that $\mathrm{d}t_1\mathrm{d}\nu(l)$ is the image measure of $\mathrm{d}\vec{t}\mathrm{d}K(\vartheta)\mathrm{d}\mathcal{C}(m)$ under the transformation, described above, that turned $\Phi'' \cap \mathcal{D}$ into Ψ. First we point out that $t(.)$ does not depend on x and that $l(.)$ does not depend on either t_1 or x. Abusing notation we write

$$t(\vec{t},x,\vartheta,m) = t(\vec{t},\vartheta,m).$$
$$l(\vec{t},x,\vartheta,m) = l(t_2,\vartheta,m).$$

Next, we notice that $t(\vec{t},x,\vartheta,m) = t_1 + t(0,t_2,x,\vartheta,m)$. It now follows that

$$\begin{aligned}
&\int_{\vec{t}}\int_{\vartheta}\sum_m \mathbf{1}_{\{a<t(\vec{t},x,\vartheta,m)\leq b\}}\mathbf{1}_{\{c<l(\vec{t},x,\vartheta,m)\leq d\}}\mathrm{d}\vec{t}\mathrm{d}K(\vartheta)\\
&= \int_{t_1}\int_{t_2}\int_{\vartheta}\sum_m \mathbf{1}_{\{a<t_1+t(0,t_2,\vartheta,m)\leq b\}}\mathbf{1}_{\{c<l(t_2,\vartheta,m)\leq d\}}\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}K(\vartheta)\\
&= \int_a^b \mathrm{d}t_1 \int_{t_2}\int_{\vartheta}\sum_m(\mathbf{1}_{\{c<l(t_2,\vartheta,m)\leq d\}}\mathrm{d}K(\vartheta)\mathrm{d}t_2\\
&= \int_a^b \mathrm{d}t_1 \int_c^d \mathrm{d}\nu(l).
\end{aligned}$$

The measure ν on $\mathbb{R}^+$ can be interpreted geometrically. Recall that we defined $\ell(d,\vartheta)$ to be a line with orientation ϑ at (signed) distance d from the origin. We define a random line L as follows. First, consider a circle $B(c,r)$ with center c and radius R such that $W \subset C(c,R)$. Next, draw a random variable Θ from K and a random variable D from the uniform distribution on $[-R,R]$. Finally, define L to be the line at an angle Θ with respect to the positive x-axis and at (signed) distance D from the origin. Now we condition on the event that L hits W. Intersecting L with W we obtain a collection of intervals and $\nu(a,b) = \int_a^b \mathrm{d}\nu(l)$ is the expected number of intervals of length between a and b.

We note that

$$\begin{aligned}
\int_l l\mathrm{d}\nu(l) &= \int_l\int_{\vartheta}\int_t\sum_m l\mathbf{1}_{l(t,\vartheta,m)}(\mathrm{d}l)\mathrm{d}t\mathrm{d}K(\vartheta)\\
&= \int_{\vartheta}\int_t\sum_m l(t,\vartheta,m)\mathrm{d}t\mathrm{d}K(\vartheta)\\
&= \int_{\vartheta}|W|\mathrm{d}K(\vartheta)\\
&= |W|.
\end{aligned}$$

We shall denote

$$\int_l \mathrm{d}\nu(l) = \kappa.$$

Integrals with respect to ν of such simple functions as above, are easily estimated using stereological methods. One would for instance sample random lines (or even a random grid of lines) hitting W and estimate the integrals by empirical averages. A nice introduction into the rudiments of stereology is given in Baddeley (1999a). There one will also find that if K is the uniform distribution, then $\kappa = \pi|\partial W|$, where $|\partial W|$ denotes the length of the boundary of W.

The points of $\Psi = \{(T_i, X_i, L_i)\}$ represent segments $[T_i, T_i + X_i]$ on the real line that *could* be observed through intervals $[0, L_i]$. Define

$$A = \{(t, x, l) \ : \ [t, t + x] \cap [0, l] \neq \emptyset\} = \{(t, x, l) \ : \ -x \leq t \leq l\}.$$

If a point of Ψ falls in A, the corresponding fragment is (partly) observed. Let us calculate the expected number of observed fragments, which equals the expected number of points of Ψ in A.

$$\begin{aligned} \int_A \lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l) &= \int_l \int_x \int_{-x}^{l} \lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l) \\ &= \int_l \int_x (l + x)\lambda \mathrm{d}F(x)\mathrm{d}\nu(l) \\ &= \int_l (l + \mu)\lambda \mathrm{d}\nu(l) = \lambda(|W| + \mu\kappa) \qquad (2.10) \end{aligned}$$

2.3 Poisson

We have a stationary point process $\Psi = \{T_i, X_i, L_i\}$ on $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ with intensity $\lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)$. These points represent line segments $[T_i, T_i + X_i]$ that might be observed through associated intervals $[0, L_i]$. The situation is very much akin to the set-up of Wijers (1995a and b). There are two differences. First he supposed the L_i to be constant. Secondly—and much more importantly—he assumed that the fragments $\{(T_i, X_i)\}$ are a Poisson process. This is certainly not the case here, even if the original line segments of Φ were Poisson. The problem is that two different fragments could have come from a single segment.

However, we may as well analyze the data *as if* Ψ were Poisson, as long as we make sure that our eventual results hold regardless of the dependencies between the points of Ψ. This is exactly what we shall do. Under the Poisson assumption we shall be able to derive the non-parametric maximum

likelihood estimator of the distribution function F. But whether or not the observations are independent, the algebraic properties of the maximum likelihood estimator remain valid. To prove its consistency when the observations are dependent we can basically copy the proof for the independent case. We only need to replace the strong law of large numbers by some ergodic theorem. We discussed these ideas in section 2.1.1.

So let us suppose Ψ is a Poisson point process. Then, given $\Psi(A) = n$, the points (T_i, X_i, L_i) in A are distributed as the set of values in an i.i.d. sample of size n from the normalized intensity, which by (2.10) equals

$$\mathbf{1}_A(t,x,l)\frac{\lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)}{\int_A \lambda \mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)} = \mathbf{1}_A(t,x,l)\frac{\mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)}{|W|+\mu\kappa}. \tag{2.11}$$

We must now address the problem of 'length bias': Because longer line segments stand a better chance of hitting W than shorter ones, the lengths X_i of the line segments that hit W are not a sample from F. Hence, even if we observed the X_i, their empirical distribution would not estimate F. We therefore introduce the distribution function, say V, of the length X of a line segment $[T, T+X]$, given its observation through $[0, L]$. We calculate just like we did to derive (2.10)

$$\begin{aligned} V(x) &= \int_{y=0}^{x}\int_l\int_{t=-x}^{l} \frac{\mathrm{d}t\mathrm{d}F(x)\mathrm{d}\nu(l)}{|W|+\mu\kappa} \\ &= \int_{y=0}^{x} \frac{|W|+y\kappa}{|W|+\mu\kappa}\mathrm{d}F(y). \end{aligned} \tag{2.12}$$

One can show that as F ranges over all possible distributions (with finite mean μ), then V too varies over all distributions. Note that

$$\int \frac{1}{|W|+x\kappa}\mathrm{d}V(x) = \int \frac{1}{|W|+\mu\kappa}\mathrm{d}F(x) = \frac{1}{|W|+\mu\kappa}$$

so that one can recover μ from V and hence also recover F from V.

Formula (2.11) for the conditional distribution of (T, X, L) given observation becomes in terms of V

$$\mathbf{1}_A(t,x,l)\frac{\mathrm{d}t\mathrm{d}V(x)\mathrm{d}\nu(l)}{|W|+x\kappa} = \mathbf{1}_A(t,x,l)\frac{\mathrm{d}t}{l+x}\mathrm{d}V(x)\frac{l+x}{|W|+x\kappa}\mathrm{d}\nu(l). \tag{2.13}$$

Define, for all x, measures

$$\mathrm{d}\rho(l|x) = \frac{l+x}{|W|+x\kappa}\mathrm{d}\nu(l).$$

and note that these are probability measures for all x. Hence, by inspection of (2.13) we can interpret the distribution of (T, X, L) as

1. $X \sim \mathrm{d}V$
2. Given $X = x$, $L \sim \mathrm{d}\rho(l|x)$
3. Given $X = x$ and $L = l$, $T \sim \mathrm{Unif}[-x, l]$

We call the collection (T_i, X_i, L_i) the complete data. We observe only a many-to-one mapping of it when we observe the intervals $[0, L_i]$ and their intersection with the $[T_i, T_i + X_i]$. This reduction, together with the fact that the distribution of the complete data has the above cascade-like structure, casts the problem of estimating V as a non-parametric missing data problem, cf. section 1.4.4. In this special model the parameter space is convex and the distribution of a single observation is linear in the parameter.

We now derive the distribution of the (observed) data. Define $Y_i = |[T_i, T_i + X_i] \cap [0, L_i]|$, the length of the i-th fragment. Let Δ_i be the number of endpoints of $[T_i, T_i + X_i]$ that fall outside of $[0, L_i]$. Also we observe L_i, the length of the interval in which a fragment is observed. Under the assumption that the $\{T_i\}$ are Poisson it will turn out that (Y, Δ) is sufficient for estimating V. At this point we introduce two functions that will help describe the distribution of the data.

$$g(x) = \int_{[x,\infty)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) \tag{2.14}$$

$$h(x) = \int_{[x,\infty)} \frac{y - x}{|W| + \kappa y} \mathrm{d}V(y). \tag{2.15}$$

For the case $\Delta = 1$, V will enter the distribution of Y only through g, and for the case $\Delta = 2$ only through h. There is a *very* useful relation between V, g and h which is easy to check using the above two definitions. For all $x > 0$

$$1 = \kappa h(x) + (|W| + \kappa x) g(x) + V(x^-). \tag{2.16}$$

where $V(x^-) = \int_{[0,x)} \mathrm{d}V(y)$.

Suppose that $\mathrm{d}\rho(.|x)$ (the distribution of L given $X = x$) is degenerate at some fixed l. We are then *exactly* in the situation studied by Wijers (1995). We follow his derivation. First consider the region

$$A_l = \{(t, x) \ : \ [t, t + x] \cap [0, l] \neq \emptyset\} = \{(t, x) \ : \ -x \leq t \leq l\}.$$

which can be subdivided into regions *lc, rc, uc* and *dc* as in figure 2.3. For a point (t, x) in *lc* the intersection of $[t, t + x]$ with $[0, l]$ will be left censored. Similarly, we distinguish right censored (*rc*), uncensored (*uc*) and doubly

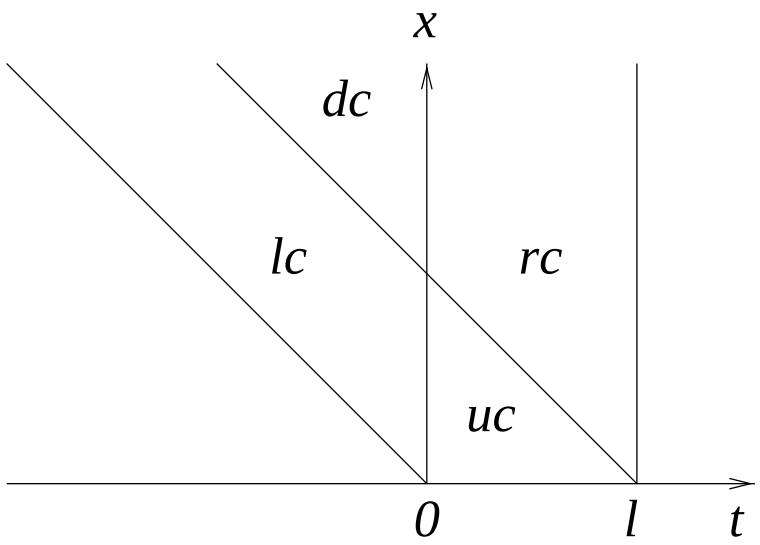

Figure 2.3: Subdividing A_l into different censoring types.

censored (dc). Left and right censored observations are taken together as censoring 'type' $\Delta = 1$; observation of one endpoint. Because $L = l$ a.s., the density of (t, x) becomes, cf. (2.13)

$$\mathbf{1}_{A_l}(t,x)\frac{\mathrm{d}t}{l+x}\mathrm{d}V(x).$$

The sub-distributions for the various censoring types can be computed by integrating, over the shaded regions in Figure 2.3. We take right and left censored observations together. We find

$$\begin{aligned} P^l(\mathrm{d}y, 0) &= \frac{l-y}{l+y}\mathrm{d}V(y) \\ P^l(\mathrm{d}y, 1) &= 2\int_y^\infty \frac{1}{l+x}\mathrm{d}V(x)\mathrm{d}y \\ P^l(\mathrm{d}y, 2) &= \int_y^\infty \frac{x-y}{l+x}\mathrm{d}V(x)\delta_l(\mathrm{d}y). \end{aligned}$$

These formulas are easily modified to accommodate general $\mathrm{d}\rho(l|x)$. Re-

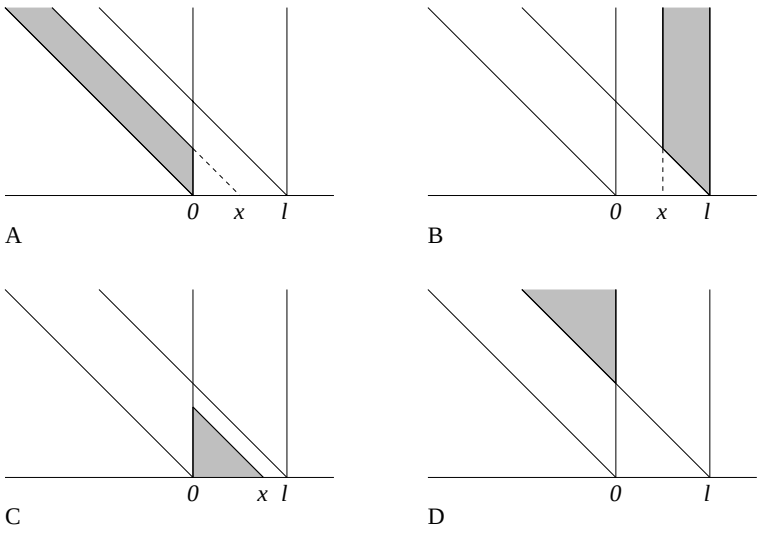

Figure 2.4: Integration regions for various sub-distribution functions.

calling that $\mathrm{d}\rho(l|x) = \frac{l+x}{|W|+x\kappa}\mathrm{d}\nu(l)$ we have

$$\begin{aligned}
P(\mathrm{d}y, \mathrm{d}l, 0) &= \mathbf{1}_{\{l\geq y\}}\frac{l-y}{l+y}\mathrm{d}\rho(l|y)\mathrm{d}V(y) = \mathbf{1}_{\{l\geq y\}}\frac{l-y}{|W|+\kappa y}\mathrm{d}\nu(l)\mathrm{d}V(y)\\
P(\mathrm{d}y, \mathrm{d}l, 1) &= 2\mathbf{1}_{\{l\geq y\}}\int_y^\infty \frac{1}{l+x}\mathrm{d}\rho(l|x)\mathrm{d}V(x)\mathrm{d}y\\
&= 2\mathbf{1}_{\{l\geq y\}}\mathrm{d}\nu(l)\int_y^\infty \frac{1}{|W|+\kappa x}\mathrm{d}V(x)\mathrm{d}y\\
&= 2\mathbf{1}_{\{l\geq y\}}\mathrm{d}\nu(l)g(y)\mathrm{d}y\\
P(\mathrm{d}y, \mathrm{d}l, 2) &= \int_y^\infty \frac{x-y}{l+x}\mathrm{d}\rho(l|x)\mathrm{d}V(x)\delta_l(\mathrm{d}y)\\
&= \mathrm{d}\nu(l)\int_y^\infty \frac{x-y}{|W|+\kappa x}\mathrm{d}V(x)\delta_l(\mathrm{d}y)\\
&= \mathrm{d}\nu(l)h(y)\delta_l(\mathrm{d}y).
\end{aligned}$$

We see that the conditional distribution of L, given Y and Δ does not depend on V. Hence (Y, Δ) is sufficient for estimation of V (at least under the present

Poisson assumption). The joint distribution of (Y, Δ) is

$$P(\mathrm{d}y, 0) = \int_l P(\mathrm{d}y, \mathrm{d}l, 0) = \frac{\int_{l \geq y} (l - y) \mathrm{d}\nu(l)}{|W| + \kappa y} \mathrm{d}V(y) \tag{2.17}$$

$$P(\mathrm{d}y, 1) = \int_l P(\mathrm{d}y, \mathrm{d}l, 1) = 2\nu([y, \infty)) g(y) \mathrm{d}y \tag{2.18}$$

$$P(\mathrm{d}y, 2) = \int_l P(\mathrm{d}y, \mathrm{d}l, 2) = \mathrm{d}\nu(y) h(y) \tag{2.19}$$

Suppose that τ is the supremum of the support of $\rho(.|x)$ over all x. Then τ is an upper bound for the length of a fragment that can be observed in W. That is, $P(Y > \tau) = 0$. Note that

$$\begin{aligned}
g(x) &= \int_{[x,\infty)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) \\
&= \int_{[x,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) + \int_{[\tau,\infty)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) \\
&= \int_{[x,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) + g(\tau) \\
h(x) &= \int_{[x,\infty)} \frac{y - x}{|W| + \kappa y} \mathrm{d}V(y) \\
&= \int_{[x,\tau)} \frac{y - x}{|W| + \kappa y} \mathrm{d}V(y) + \int_{[\tau,\infty)} \frac{y - \tau + \tau - x}{|W| + \kappa y} \mathrm{d}V(y) \\
&= \int_{[x,\tau)} \frac{y - x}{|W| + \kappa y} \mathrm{d}V(y) + h(\tau) + (\tau - x) g(\tau).
\end{aligned}$$

We define

$$H = \kappa h(\tau) \qquad G = (|W| + \kappa\tau) g(\tau).$$

and recall that relation (2.16) states that $G + H + V(\tau^-) = 1$. Later on we shall interpret G and H as probabilities. We can express the distribution of

(Y, Δ) in terms of V restricted to $[0, \tau)$, G and H. For $y < \tau$

$$
\begin{aligned}
P(\mathrm{d}y, 0) &= \frac{\int_{l \geq y} (l - y) \mathrm{d}\nu(l)}{|W| + \kappa y} \mathrm{d}V(y) && (2.20) \\
P(\mathrm{d}y, 1) &= 2\nu([y, \infty)) \mathrm{d}y g(y) = 2\nu([y, \infty)) \mathrm{d}y \int_y^\infty \frac{1}{|W| + \kappa x} \mathrm{d}V(x) \\
&= 2\nu([y, \infty)) \mathrm{d}y \Big(\int_y^{\tau-} \frac{1}{|W| + \kappa x} \mathrm{d}V(x) + \int_\tau^\infty \frac{1}{|W| + \kappa x} \mathrm{d}V(x) \Big) \\
&= 2\nu([y, \infty)) \mathrm{d}y \int_y^{\tau-} \frac{1}{|W| + \kappa x} \mathrm{d}V(x) + \frac{2\nu([y, \infty)) \mathrm{d}y}{|W| + \kappa \tau} G && (2.21) \\
P(\mathrm{d}y, 2) &= \mathrm{d}\nu(y) h(y) = \mathrm{d}\nu(y) \int_y^\infty \frac{x - y}{|W| + \kappa x} \mathrm{d}V(x) \\
&= \mathrm{d}\nu(y) \Big(\int_y^{\tau-} \frac{x - y}{|W| + \kappa x} \mathrm{d}V(x) + \int_\tau^\infty \frac{x - \tau}{|W| + \kappa x} \mathrm{d}V(x) \\
&+ \int_\tau^\infty \frac{\tau - y}{|W| + \kappa x} \mathrm{d}V(x) \Big) \\
&= \mathrm{d}\nu(y) \int_y^{\tau-} \frac{x - y}{|W| + \kappa x} \mathrm{d}V(x) + \frac{\mathrm{d}\nu(y)}{\kappa} H + \frac{(\tau - y) \mathrm{d}\nu(y)}{|W| + \kappa \tau} G. && (2.22)
\end{aligned}
$$

Because (2.16) expresses G in terms of $V(\tau^-)$ and H, the distribution of the data is fully parameterized by

$$(V|_{[0,\tau)}, H),$$

where $V|_{[0,\tau)}$ ranges over all (possibly defective) distribution functions and H is any positive real such that $V(\tau^-) + H \leq 1$. Below we demonstrate that there is a 1-1 correspondence between $(V|_{[0,\tau)}, H)$ and $(F|_{[0,\tau)}, \mu)$, where $F|_{[0,\tau)}$ ranges over all (possibly defective) distribution functions and μ is any positive real exceeding $\int_0^{\tau-} x \mathrm{d}F(x)$.

By the way, it is interesting that without any observations exceeding τ, the mean of F can be estimated.

To express $F|_{[0,\tau)}$ and μ in terms of $V|_{[0,\tau)}$ and H, we first note that $g(\tau) = G/(|W| + \kappa\tau)$ is determined by $V(\tau^-)$ and H through (2.16). We have

$$
\begin{aligned}
\int_{[0,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) + g(\tau) &= \int_0^\infty \frac{1}{|W| + \kappa y} \mathrm{d}V(y) \\
&= \int_0^\infty \frac{1}{|W| + \kappa \mu} \mathrm{d}F(y) = \frac{1}{|W| + \kappa \mu}
\end{aligned}
$$

so that

$$\mu = \frac{1}{\kappa}\left(\frac{1}{\int_{[0,\tau)} \frac{1}{|W|+\kappa y}\mathrm{d}V(y) + g(\tau)} - |W|\right) \tag{2.23}$$

and

$$\begin{aligned} F(x) &= \int_0^x \frac{|W|+\kappa\mu}{|W|+\kappa y}\mathrm{d}V(y) \\ &= \frac{1}{\int_{[0,\tau)} \frac{1}{|W|+\kappa y}\mathrm{d}V(y) + g(\tau)} \int_0^x \frac{1}{|W|+\kappa y}\mathrm{d}V(y). \end{aligned} \tag{2.24}$$

2.4 EM

Under the assumption that the fragments we observe are independent, we have a nonparametric missing data problem (see section 1.4.4).

1. $X \sim V, \;\; V \in \mathcal{V} = \{\text{all distributions on } \mathbb{R}^+\}$
2. Given $X = x, \;\; L \sim \mathrm{d}\rho(l|x)$
3. Given $X = x$ and $L = l$, $T \sim \text{Unif}[-x, l]$
4. Observe $(Y, \Delta) = g(X, T, L)$

where $Y = |[T, T+X] \cap [0, L]|$ and Δ is the number of unobserved endpoints.

As we have seen, the distribution of the data (Y, Δ) depends on V only through its restriction to $[0, \tau)$ and a functional H. This means that the model $\mathcal{V}$ is not identifiable from the data. If two distribution functions V and V' agree on $[0, \tau)$ and have the same H's, then they can never be told apart on the basis of observing (Y_i, Δ_i).

We could reduce $\mathcal{V}$ in such a way that it becomes identifiable, but then the model is no longer completely nonparametric. However, we can construct a different model consisting of all probability measures on a different space, and a new missing data problem such that model *is* identifiable. Of course we must make sure that the model for the observed data remains the same.

We now state a nonparametric, identifiable missing data problem, such that the distribution of the observed data satisfies (2.20) to (2.22). We again use the symbol $\mathcal{V}$ to denote the model for the complete data. We hope this does not confuse too much.

The (new) model $\mathcal{V}$ for the complete data, say X, consists of *all* probability distributions on the space $[0, \tau) \cup \{\dagger, \ddagger\}$. We write $V(\{\dagger\}) = P(X = \dagger) = G$ and $V(\{\ddagger\}) = P(X = \ddagger) = H$. We must have $V(\tau^-) + G + H = 1$.

Sample X from $V \in \mathcal{V}$. Next,

- If $X = x \in [0, \tau)$ we draw an L from $\mathrm{d}\rho(.|x)$ and given $L = l$ we draw a uniform T on $[-x, l]$. We set $Y = |[T, T + X] \cap [0, L]|$ and Δ is the number of endpoints of $[T, T + X]$ outside of $[0, L]$.
- If $X = \dagger$ we sample (Y, Δ) from

$$\mathbf{1}_1(\delta)\frac{2\nu([y,\infty))\mathrm{d}y}{|W| + \kappa\tau} + \mathbf{1}_2(\delta)\frac{(\tau - y)\mathrm{d}\nu(y)}{|W| + \kappa\tau}.$$

- If $X = \ddagger$ we sample (Y, Δ) from

$$\mathbf{1}_2(\delta)\frac{\mathrm{d}\nu(y)}{\kappa}.$$

It is certainly not hard to check that the distribution of (Y, Δ) under this new scheme is indeed again given by (2.20) to (2.22).

Now, because we have a completely nonparametric missing data model we can use every square integrable function which integrates to zero as a score function. In particular, we can use indicator functions of measurable sets in $[0, \tau) \cup \{\dagger, \ddagger\}$ minus their expectations. As in (1.14) we find the following system of score equations

$$\hat{V}_n(A) = \frac{1}{n}\sum_1^n P_{\hat{V}_n}(X_i \in A | Y_i, \Delta_i), \qquad A \subseteq [0, \tau) \cup \{\dagger, \ddagger\} \tag{2.25}$$

or in other words

$$\hat{V}(\mathrm{d}x) = \frac{1}{n}\sum_1^n P_{\hat{V}_n}(X_i \in \mathrm{d}x | Y_i, \Delta_i), \qquad x \in [0, \tau) \tag{2.26}$$

$$\hat{V}_n(\{\ddagger\}) = \hat{H} = \frac{1}{n}\sum_1^n P_{\hat{V}_n}(X = \ddagger | Y_i, \Delta_i) \tag{2.27}$$

and by (2.16)

$$\hat{V}_n(\{\dagger\}) = 1 - \hat{V}_n(\{\ddagger\}) - \hat{V}_n(\tau^-). \tag{2.28}$$

Solving the above equations iteratively is an instance of the EM algorithm. From one iteration step to the next the support will not increase. Hence, one should first decide on the support of $\hat{V}_n$ and then start the iterations with an initial guess which does not have smaller support. The support of the true NPMLE (as defined in (1.1)) is unknown to us. We *choose* to start the iterations with a distribution V on $[0, \tau) \cup \{\dagger, \ddagger\}$ with mass only at the

observed uncensored observations and at the points † and ‡. Thus we obtain a 'data sieved' NPMLE (see section 1.4.1).

To find explicit expressions for the above score equations, note that the distribution of the data, specified in (2.20) to (2.22), is of the form

$$P(\mathrm{d}y, \delta) = KV(\mathrm{d}y, \delta) = \int K(\mathrm{d}y, \delta|x)\mathrm{d}V(x).$$

We recognize, for $y \in [0, \tau)$ and $\delta \in \{0, 1, 2\}$,

$$\begin{aligned} K(\mathrm{d}y, \delta|x) = {} & \mathbf{1}_0(\delta)\frac{\int_{l\geq x}(l-x)\mathrm{d}\nu(l)}{|W|+\kappa x}\delta_x(\mathrm{d}y) \\ & + \mathbf{1}_1(\delta)2\nu([y,\infty))\left(\frac{\mathbf{1}_{(y,\tau)}(x)}{|W|+\kappa x} + \frac{\mathbf{1}_{\{\dagger\}}(x)}{|W|+\kappa\tau}\right)\mathrm{d}y \\ & + \mathbf{1}_2(\delta)\left(\frac{\mathbf{1}_{(y,\tau)}(x)(x-y)}{|W|+\kappa x} + \frac{\mathbf{1}_{\{\dagger\}}(x)(\tau-y)}{|W|+\kappa\tau} + \frac{\mathbf{1}_{\{\ddagger\}}(x)}{\kappa}\right)\mathrm{d}\nu(y). \end{aligned}$$

Given a sample $(y_1, \delta_1), \ldots, (y_n, \delta_n)$ suppose that $x_1, \ldots, x_m$ are the lengths of all different uncensored observations. Consider the (random) measure

$$\begin{aligned} \mu(\mathrm{d}y, \delta) = {} & \mathbf{1}_0(\delta)\sum_{i=1}^{m}\frac{\int_{l\geq x_i}(l-x_i)\mathrm{d}\nu(l)}{|W|+\kappa X_i}\delta_{x_i}(\mathrm{d}y) \\ & + \mathbf{1}_1(\delta)\mathbf{1}_{[0,\tau)}(y)\nu([y,\infty))\mathrm{d}y \\ & + \mathbf{1}_2(\delta)\mathbf{1}_{[0,\infty)}(y)\nu(\mathrm{d}y). \end{aligned}$$

This sigma-finite measure μ dominates $K(\mathrm{d}y, \delta|x)$ for all $x \in \{x_1, \ldots, x_m\} \cup \{\dagger, \ddagger\}$. The Radon–Nikodym derivative of $K(\mathrm{d}y, \delta|x)$ with respect to $\mu(\mathrm{d}y, \delta)$ is, for all $x \in \{x_1, \ldots, x_m\} \cup \{\dagger, \ddagger\}$

$$\begin{aligned} k(y, \delta; x) = {} & \mathbf{1}_0(\delta)\mathbf{1}_{[0,\tau)}(x)\mathbf{1}_x(y) \\ & + \mathbf{1}_1(\delta)\left(\frac{\mathbf{1}_{(y,\tau)}(x)}{|W|+\kappa x} + \frac{\mathbf{1}_{\{\dagger\}}(x)}{|W|+\kappa\tau}\right) \\ & + \mathbf{1}_2(\delta)\left(\frac{\mathbf{1}_{(y,\tau)}(x)(x-y)}{|W|+\kappa x} + \frac{\mathbf{1}_{\{\dagger\}}(x)(\tau-y)}{|W|+\kappa\tau} + \frac{\mathbf{1}_{\{\ddagger\}}(x)}{\kappa}\right). \end{aligned}$$

We arrange it so that the data sieved NPMLE $\hat{V}_n$ puts all its mass on $\{x_1, \ldots, x_m\} \cup \{\dagger, \ddagger\}$. Thus, for $\hat{V}_n$ almost all x the mixing kernel K is dominated by a single sigma-finite measure as was required in section 1.4.4. In that section we also specified what the score equations for a nonparametric missing data problem look like, cf. (1.17). We find

$$\hat{V}_n(A) = \int \frac{\int_A k(y, \delta; x)\mathrm{d}V(x)}{\int_{x'} k(y, \delta; x')\mathrm{d}V(x)}\mathrm{d}\mathbb{P}_n(y, \delta)$$

where $\mathbb{P}_n$ is the empirical distribution of the data putting mass $1/n$ at the observed $(y_1, \delta_1), \ldots, (y_n, \delta_n)$. The equations work out to

$$\begin{aligned}\hat{V}_n(\{x_i\}) &= \mathbb{P}_n(x_i, 0) \\ &\quad + \frac{\hat{V}_n(\{x_i\})}{|W| + \kappa x_i} \int_{y=0}^{x_i} \frac{1}{\hat{g}_n(y)} \mathbb{P}_n(\mathrm{d}y, 1) \\ &\quad + \frac{\hat{V}_n(\{x_i\})}{|W| + \kappa x_i} \int_{y=0}^{x_i} \frac{x_i - y}{\hat{h}_n(y)} \mathbb{P}_n(\mathrm{d}y, 2), \qquad i = 1, 2, \ldots, m \\ \hat{V}_n(\{\ddagger\}) = \hat{H}_n &= \hat{H}_n \int_y \frac{1}{\hat{h}_n(y)} \mathbb{P}_n(\mathrm{d}y, 2).\end{aligned}$$

where, cf. (2.16)

$$\hat{V}_n(\{\dagger\}) = \hat{G}_n = 1 - \hat{H}_n - \hat{V}_n(\tau^-)$$

$$\hat{g}_n(\tau) = \hat{G}_n \;/\; (|W| + \kappa\tau)) \quad \text{and} \quad \hat{h}_n(\tau) = \hat{H}_n/\kappa$$

and, for $x \in [0, \tau)$,

$$\begin{aligned}\hat{g}_n(x) &= \int_{[x,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}\hat{V}_n(y) + \hat{g}_n(\tau) \\ \hat{h}_n(x) &= \int_{[x,\tau)} \frac{y - x}{|W| + \kappa y} \mathrm{d}\hat{V}_n(y) + \hat{h}_n(\tau) + (\tau - x)\hat{g}_n(\tau).\end{aligned}$$

Recall that $(V_{[0,\tau)}, H)$ and $(F_{[0,\tau)}, \mu)$ stand in a one-to-one relation. In the previous section we established formulas (2.23) and (2.24), expressing $(F_{[0,\tau)}, \mu)$ in terms of $(V_{[0,\tau)}, H)$. With $g(\tau) = (1 - V(\tau^-) - H)/(|W| + \kappa\tau)$, we have

$$\mu = \frac{1}{\kappa} \left(\frac{1}{\int_{[0,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) + g(\tau)} - |W| \right)$$

and

$$F(x) = \frac{1}{\int_{[0,\tau)} \frac{1}{|W| + \kappa y} \mathrm{d}V(y) + g(\tau)} \int_0^x \frac{1}{|W| + \kappa y} \mathrm{d}V(y).$$

It is now obvious how we define $\hat{F}_n$ and $\hat{\mu}_n$ as transformations of $\hat{V}_n$ and $\hat{H}_n$.

There really is not much point in proving desirable asymptotic properties, such as consistency, of $\hat{V}_n$ as n tends to infinity. Any result we obtain here will only hold under the Poisson assumption stated at the beginning of the

previous section. However, it will turn out that the *method* of proving consistency will—with minor modifications—work without the Poisson assumption. When we no longer have the Poisson assumption we shall have to use different asymptotics, which will lead to some complications. To give an (admittedly sketchy) proof of consistency of $\hat{V}_n$ here, allows us a more clean presentation. We shall use Theorem 1.2, but we need a *working hypothesis*:

working hypothesis
If $g(\tau) > 0$ then $\liminf \hat{g}_n(\tau) > 0$, almost surely.

Whether or not this is true, we do not know. When, in the next section, we prove consistency without the Poisson assumption this hypothesis will no longer be needed. By that time, to get around other difficulties, we shall have introduced some extra, artificial censoring. This will have the added benefit of making our present working hypothesis superfluous.

First, we note that the model $\mathcal{V}$ (of all probability measures on $[0,\tau) \cup \{\dagger, \ddagger\}$) is convex, which by Lemma 1.3 implies that also the model for the distribution of the data $K\mathcal{V} = \{KV : V \in \mathcal{V}\}$ is convex. This places us in the realm of Theorem 1.2. We can use this theorem to prove consistency (under our working hypothesis) of the NPMLE $K\hat{V}_n$ in the sense that, almost surely,

$$\|\hat{V}_n - V_0\|_{[0,\tau)} \to 0 \quad \text{and} \quad |\hat{H}_n - H_0| \to 0. \tag{2.29}$$

Here $\|.\|_{[0,\tau)}$ denotes the uniform distance on the set $[0,\tau)$. From (2.29) it is not too difficult (using empirical process theory) to conclude that also

$$\|\hat{F}_n - F_0\|_{[0,\tau)} \to 0 \quad \text{and} \quad |\hat{\mu}_n - \mu_0| \to 0, \tag{2.30}$$

almost surely.

To apply Theorem 1.2 we need to compare $K\hat{V}_n$ to a sequence KV_n for which we can easily establish consistency. We choose

$$V_n(x) = \int_0^x \frac{|W| + \kappa y}{\int_{l \geq y}(l-y)\mathrm{d}\nu(l)} \mathbb{P}_n(\mathrm{d}y, 0) \qquad x \in [0,\tau).$$

We define $V_n(\tau^-)$ in the obvious manner and note that by the law of large numbers it is consistent. Hence, for $V_n(\{\ddagger\}) = H_n$ we can take a sequence tending to H while $H_n \leq V_n(\tau)$. We define G_n, g_n and h_n in terms of $V_n|_{[0,\tau)}$ and H_n.

From (2.20) we see that $KV_n(\mathrm{d}y, 0) = \mathbb{P}_n(\mathrm{d}y, 0)$.

First we show convergence of (V_n, H_n) to satisfy the first requirement of Theorem 1.2. The collection

$$\{\mathbf{1}_{[0,x]}(y)\frac{|W|+\kappa y}{\int_{l\geq y}(l-y)\mathrm{d}\nu(l)},\ x\in[0,\tau)\}$$

is a *Glivenko–Cantelli* class of functions over which the law of large numbers holds uniformly. In other words, we have

$$\|V_n - V_0\|_{[0,\tau)} \to 0 \ \text{ and } \ |H_n - H_0| \to 0 \tag{2.31}$$

almost surely. For empirical process theory see, for instance, van der Vaart and Wellner (1996). That the above collection of functions is a Glivenko–Cantelli class follows from the fact that the functions are all the product of an indicator function (of intervals) and the same monotone function and hence they are uniformly of bounded variation: the variation is bounded uniformly over the collection of functions. Also, there is an integrable 'envelop' (majorant) for the entire collection.

Next, we need that the straight line between KV_n and $K\hat{V}_n$ is Hellinger differentiable at $K\hat{V}_n$. By Lemma 1.4 it suffices to show that the straight line between V_n and $\hat{V}_n$ is Hellinger differentiable at $\hat{V}_n$. Because $V_n \ll \hat{V}_n$ and also $\mathrm{d}V_n/\mathrm{d}\hat{V}_n \in L_2(\hat{V}_n)$ this is indeed true.

Now, if we establish the final, third condition of Theorem 1.2

$$\sum_{\delta=0}^{2}\int_{[0,\tau)}\left(\frac{\mathrm{d}KV_n}{\mathrm{d}K\hat{V}_n}\right)\mathrm{d}(\mathbb{P}_n(y,\delta) - KV_n(y,\delta)) \to 0 \tag{2.32}$$

then we can conclude that $K\hat{V}_n$ converges to KV_0 in the same sense as KV_n does. In other words, (2.32) implies (2.29), which in turn implies (2.30).

From formulas (2.17) to (2.19) we see that (2.32) reads

$$\begin{aligned}
&\sum_{\delta=0}^{2}\int_{[0,\tau)}\left(\frac{\mathrm{d}KV_n}{\mathrm{d}K\hat{V}_n}\right)\mathrm{d}(\mathbb{P}_n(y,\delta) - KV_n(y,\delta))\\
&= \int_{[0,\tau)}\left(\frac{\mathrm{d}V_n(y)}{\mathrm{d}\hat{V}_n(y)}\right)\mathrm{d}(\mathbb{P}_n(y,0) - KV_n(y,0))\\
&+ \int_{[0,\tau)}\left(\frac{g_n(y)}{\hat{g}_n(y)}\right)\mathrm{d}(\mathbb{P}_n(y,1) - KV_n(y,1))\\
&+ \int_{[0,\tau)}\left(\frac{h_n(y)}{\hat{h}_n(y)}\right)\mathrm{d}(\mathbb{P}_n(y,2) - KV_n(y,2))\\
&\to 0.
\end{aligned} \tag{2.33}$$

Since $KV_n(\mathrm{d}y, 0) = \mathbb{P}_n(\mathrm{d}y, 0)$ we get convergence of the first term of (2.33) for free. To prove convergence of the other two terms, we again apply empirical process theory to assert that not only for $\delta = 0$, but also for $\delta = 1, 2$

$$\| \int_0^{\cdot} \mathrm{d}\mathbb{P}_n(y, \delta) - \int_0^{\cdot} \mathrm{d}KV_n(y, \delta) \|_{[0,\tau)} \to 0, \tag{2.34}$$

almost surely. Now we need some elementary calculus

Lemma 2.1. *Suppose that H_n ($n = 1, 2, \dots$) and H are real valued functions on an interval $[a, b)$ which are of bounded variation. Suppose that G_n are real valued functions on $[a, b)$ which are bounded and of bounded variation uniformly in n. Suppose that the H_n tend to H uniformly on $[a, b)$. Then*

$$\int_{[a,b)} G_n \mathrm{d}(H_n - H) \to 0.$$

Proof Applying integration by parts, we rewrite the integral as

$$G_n(b^-)(H_n(b^-) - H(b^-)) - G_n(a)(H_n(a) - H(a)) + \int_{[a,b)} (H_n - H)\mathrm{d}G_n$$

The first two terms tend to zero because the G_n are uniformly bounded and the H_n converge to H. The third term is in absolute value not more than $||H_n - H||_\infty \int_{[a,b)} |\mathrm{d}G_n|$. This term tends to zero because the H_n tend uniformly to H and the G_n are uniformly of bounded variation. □

We need to show convergence of the remaining two terms of (2.33). With the above lemma and (2.34), this boils down to showing that $g_n/\hat{g}_n$ and $h_n/\hat{h}_n$ are uniformly bounded and of bounded variation (almost surely for n large enough). Under our 'working hypothesis' this can be accomplished without much difficulty.

2.5 Without the Poisson assumption

Recall that we started out with a marked point-process $\Phi = \{(\vec{S}_i, X_i, \Theta_i)\}$ on $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$ to model a collection of line segments with left endpoint $\vec{S}_i$, length X_i and orientation Θ_i. We then transformed Φ into $\Phi' = \{(\vec{T}_i, X_i, \Theta_i)\}$, representing the left endpoints with respect to the unique straight line on which the associated line segment lies. Using square brackets, we write $[\vec{t}, x, \vartheta]$ to denote a line segment (i.e. a random closed subset of $\mathbb{R}^2$), and $(\vec{t}, x, \vartheta)$ to denote a point in $\mathbb{R}^2 \times \mathbb{R}^+ \times (-\pi/2, \pi/2)$. We use

$\Phi' \times \mathbb{N}$ to represent all possible fragments arising from intersecting the line segments with a random closed set $\mathcal{W}$. Recall we introduced lines $\ell(d, \vartheta)$ with orientation ϑ at (signed) distance d from the origin and defined $\mathcal{W}(d, \vartheta) = \ell(d, \vartheta) \cap \mathcal{W}$. We assumed that the $\mathcal{W}(t, \theta)$ are always the union of (at most) countably many closed intervals: $\mathcal{W}(t, \vartheta) = \cup_m \mathcal{W}(t, \vartheta, m)$. Conditionally on $\mathcal{W}$, a point $(\vec{t}, x, \vartheta, m)$ corresponds to a (possibly empty) random closed set $[\vec{t}, x, \vartheta, m] = [\vec{t}, x, \vartheta] \cap \mathcal{W}(t_2, \vartheta, m)$, which we called a fragment. Finally we represented all observed fragments as a point process $\Psi = \{(T_i, X_i, L_i)\}$ on $\mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ corresponding to the situation where we observe segments $[T_i, T_i + X_i]$ through intervals $[0, L_i]$.

In the preceding section we assumed that Ψ is a Poisson process. We conditioned on the number of fragments that are observed, so that the fragments could be interpreted as an independent, identically distributed sample. We were then able to interpret the observations as 'missing data' (see section 1.4.4) in a very convenient way. This led us to a set of nonparametric score equations. The nonparametric maximum likelihood estimator (see section 1.4.1) in the sense of Kiefer and Wolfowitz (1956) is a solution of these equations. We proceeded to apply general methods (Theorem 1.2) to prove consistency of the estimator as the number of observations tends to infinity.

As we explained earlier, the assumption that the fragments are Poisson is simply not justified. They are not even independent, because several fragments might belong to a single underlying segment. However, as it turns out, without the Poisson assumption it is still sensible to use the *same* estimator. Also, we can still use Theorem 1.2 to prove its consistency.

Without the Poisson assumption we no longer have i.i.d. observations, the number of which we can send to infinity. We shall have to resort to different asymptotics. Indeed, we shall consider observation of $\mathcal{W} \cap \Phi$ through expanding sets B_n. Instead of empirical distribution functions we shall have empirical averages and instead of Glivenko–Cantelli we use an ergodic theorem to obtain the necessary convergence. We discussed this approach in some generality in section 2.1.1.

2.5.1 Expanding domain asymptotics

Let

$$B_n = nB = [0, n] \times [0, n] \quad \text{and} \quad B_\infty = \mathbb{R}^2$$

and suppose we observe $\mathcal{W} \cap \Phi$ through these B_n. We must now modify some earlier definitions to accommodate the varying n. Our original set-up corresponds exactly to the case $n = 1$.

We define, for all $d \in \mathbb{R}$, $\vartheta \in (-\pi/2, \pi/2)$ and $m \in \mathbb{N}$

$$W_n(d, \vartheta, m) = \mathcal{W}(d, \vartheta, m) \cap B_n \text{ and } W_\infty(d, \vartheta, m) = \mathcal{W}(d, \vartheta, m)$$

Now consider sets

$$\mathcal{D}_n = \{(\vec{t}, x, \vartheta, m) \ : \ W_n(t_2, \vartheta, m) \neq \emptyset\}$$

and note that even the set $\mathcal{D}_\infty$ is well defined.

As before we define on $\mathcal{D}_n$ two functions: $t_n(\vec{t}, \vartheta, m)$ and $l_n(\vec{t}, \vartheta, m)$. $t_n(\vec{t}, x, \vartheta, m)$ is the distance from the left endpoint of the line segment $[\vec{t}, x, \vartheta]$ to the left-most endpoint of $W(t_2, \vartheta, m)$, which we take negative if the first is to the left of the latter. Let $l_n(\vec{t}, x, \vartheta, m)$ be the length of $W_n(t_2, \vartheta, m)$. Consider the map that assigns $(t_n(\vec{t}, x, \vartheta, m), x, l_n(\vec{t}, x, \vartheta, m))$ to $(\vec{t}, x, \vartheta, m)$ for all $(\vec{t}, x, \vartheta, m) \in \mathcal{D}_n$. The image under this map of $\Phi'' \cap \mathcal{D}$ is a point-process $\Psi_n = \{(T_i, X_i, L_i)\}$ on $\mathbb{R} \times \mathbb{R}^+ \times [0, \infty)$ with intensity

$$\lambda \mathrm{d}t \mathrm{d}F(x) \mathrm{d}\nu_n(l),$$

where

$$\mathrm{d}\nu_n(l) = \int_t \int_\vartheta \sum_m \delta_{l_n(t,\vartheta,m)}(\mathrm{d}l) \mathrm{d}t \mathrm{d}K(\vartheta).$$

The Ψ_n represent the fragments arising from intersecting $\Phi \cap \mathcal{W}$ with B_n.

We note that

$$\int_l l \mathrm{d}\nu_n(l) = |W_n|$$

and define

$$\int_l \mathrm{d}\nu_n(l) = \kappa_n.$$

As before $A = \{(t, x, l) \ : \ [t, t + x] \cap [0, l] \neq \emptyset\}$, and the expected number of fragments we observe in $W_n = \mathcal{W} \cap B_n$ is

$$\int_A \lambda \mathrm{d}t \mathrm{d}F(x) \mathrm{d}\nu_n(l) = \lambda(|W_n| + \mu\kappa_n).$$

Finally, define

$$V_n(x) = \int_0^x \frac{|W_n| + y\kappa_n}{|W_n| + \mu\kappa_n} \mathrm{d}F(y)$$

and set

$$\begin{aligned} g_n(x) &= \int_{[x,\infty)} \frac{1}{|W_n| + \kappa_n y} \mathrm{d}V_n(y) \\ h_n(x) &= \int_{[x,\infty)} \frac{y - x}{|W_n| + \kappa_n y} \mathrm{d}V_n(y). \end{aligned}$$

2.5.2 Some extra censoring

We have a sequence of statistical problems: observation of $\Phi \cap \mathcal{W}$ through sets W_n. For each fixed n we can proceed as before and, pretending that the observed fragments are independent, write down the score equations. This is exactly what we do in this section, but there is one slight complication. To overcome it, we shall introduce some artificial censoring.

The normalized version of the measure $\mathrm{d}\nu_n$ for a given x is

$$\mathrm{d}\rho_n(l|x) = \frac{l+x}{|W_n| + x\kappa_n}\mathrm{d}\nu_n(l)$$

which may be interpreted as the distribution of the length of the interval through which a fragment is observed, given that the underlying line segment has length x. Now, if we define τ_n to be the supremum of the support of $\mathrm{d}\rho_n(l|x)$ over all x, then τ_n is an upper bound for the length of a fragment that can be observed through W_n. The distribution of the data from observing $\Phi \cap W_n$ is determined by the restriction of V_n to $[0, \tau_n)$ and $H_n = \kappa_n h_n(\tau_n)$. This is changing with n in a very awkward way. We also believe that estimation of F at τ_n from fragments seen through W_n is unstable, because it is typically very unlikely to observe fragments of length near τ_n.

To regularize our problem we introduce some extra censoring. We fix a value τ such that with probability 1 (under the distribution of $\mathcal{W}$), the $\sup_x \operatorname{supp} \mathrm{d}\rho_n(l|x)$ will exceed τ. We group together all observations of length equal to, or exceeding τ, irrespective of their censoring type. In practice it would be wise to choose τ in such a way that a certain percentage of the observed fragments are longer than τ.

We write P_n for the distribution, under the parameter $(V_n|_{[0,\tau)}, H_n)$, of the fragments observed through B_n. The probability that the length of a fragment observed through W_n exceeds τ is

$$\begin{aligned}
P_n(Y > \tau) &= \int_{x>\tau}\int_{l>\tau}\int_{t=\tau-x}^{0} \frac{\mathrm{d}t\mathrm{d}\nu_n(l)\mathrm{d}V_n(x)}{|W_n| + \kappa_n\tau} \\
&+ \int_{x>\tau}\int_{l>\tau}\int_{t=0}^{l-\tau} \frac{\mathrm{d}t\mathrm{d}\nu_n(l)\mathrm{d}V_n(x)}{|W_n| + \kappa_n\tau} \\
&= \int_{l>\tau}(l-\tau)\mathrm{d}\nu_n(l)\int_{x>\tau}\frac{1}{|W_n| + \kappa_n\tau}\mathrm{d}V_n(x) \\
&+ \int_{l>\tau}\mathrm{d}\nu_n(l)\int_{x>\tau}\frac{x-\tau}{|W_n| + \kappa_n\tau}\mathrm{d}V_n(x) \\
&= \int_{l>\tau}(l-\tau)\mathrm{d}\nu_n(l)g_n(\tau) + \int_{l>\tau}\mathrm{d}\nu_n(l)h_n(\tau) \\
&= a_nG_n + b_nH_n
\end{aligned}$$

where

$$\begin{aligned} a_n &= \frac{1}{|W_n| + \kappa_n \tau} \int_{l>\tau} (l - \tau) \mathrm{d}\nu_n(l) \\ b_n &= \frac{1}{\kappa_n} \int_{l>\tau} \mathrm{d}\nu_n(l) \end{aligned}$$

The estimating equations of V_n are similar to what we found earlier, for $x < \tau$

$$\begin{aligned} \mathrm{d}\hat{V}_n(x) &= \mathrm{d}\mathbb{P}_n(x, 0) \\ &+ \frac{\mathrm{d}\hat{V}_n(x)}{|W_n| + \kappa_n x} \int_{y=0}^{x} \frac{1}{\hat{g_n}(y)} \mathrm{d}\mathbb{P}_n(y, 1) \\ &+ \frac{\mathrm{d}\hat{V}_n(x)}{|W_n| + \kappa_n x} \int_{y=0}^{x} \frac{x - y}{\hat{h_n}(y)} \mathrm{d}\mathbb{P}_n(y, 2) \end{aligned} \tag{2.35}$$

The extra censoring does yield an extra term in the estimating equation for H

$$\hat{H}_n = \hat{H}_n \int_{y=0}^{\tau-} \frac{1}{\hat{h}_n(y)} \mathrm{d}\mathbb{P}_n(y, 2) + \frac{b_n \hat{H}_n}{a_n \hat{G}_n + b_n \hat{H}_n} \mathbb{P}_n(Y > \tau)) \tag{2.36}$$

where, cf. (2.16)

$$\hat{G}_n = 1 - \hat{H}_n - \hat{V}_n(\tau^-)$$

and

$$\hat{g_n}(\tau) = \hat{G}_n \ / \ (|W_n| + \kappa_n \tau)) \quad \text{and} \quad \hat{h_n}(\tau) = \hat{H}_n / \kappa$$

and, for $x \in [0, \tau)$,

$$\begin{aligned} \hat{g_n}(x) &= \int_{[x,\tau)} \frac{1}{|W_n| + \kappa_n y} \mathrm{d}\hat{V}_n(y) + \hat{g_n}(\tau) \\ \hat{h_n}(x) &= \int_{[x,\tau)} \frac{y - x}{|W_n| + \kappa_n y} \mathrm{d}\hat{V}_n(y) + \hat{h_n}(\tau) + (\tau - x)\hat{g_n}(\tau). \end{aligned}$$

2.5.3 ergodic lemmas

In this subsection we first work through a number of results establishing that various (properly normalized) functionals of $\mathcal{W} \cap B_n$ converge as n tends to infinity. Then we consider convergence of functionals of $\Phi \cap \mathcal{W} \cap B_n$. We must assume:

Assumption 2.1. $(\Phi, \mathcal{W})$ *is jointly stationary ergodic.*

This assumption actually holds if and only if Φ is weakly mixing and $\mathcal{W}$ is ergodic or vice versa. Recall that weakly mixing implies ergodicity so that in particular we are assuming that $\mathcal{W}$ is ergodic.

Observing the line segments of Φ through $\mathcal{W} \cap B_n$ we have censoring at the boundaries of $\mathcal{W}$ and of B_n. Censoring at the boundaries of B_n is a negligible effect as n tends to infinity because the expected total number of fragments observed through $\mathcal{W} \cap B_n$ is of order n^2, while the expected number of fragments hitting the boundary of B_n is only of order n. In a nutshell, this is what we make precise in this section.

Recall our transformation $\Phi' = \{(\vec{T}_i, X_i, \Theta_i)\}$ of $\Phi = \{(\vec{S}_i, X_i, \Theta_i)\}$ to represent the location of the left endpoint of each line segment relative to the unique straight line on which it lies. Observing $\Phi \cap \mathcal{W} \cap B_n$, we see non-empty intersections $[\vec{T}_i, X_i, \Theta_i, m] = [\vec{T}_i, X_i, \Theta_i] \cap \mathcal{W}_n(T_{2,i}, X_i, \Theta_i, m) \cap B_n = [\vec{T}_i, X_i, \Theta_i] \cap W_n(T_{2,i}, X_i, \Theta_i, m)$ for all m. We named these non-empty intersections 'fragments'. Now consider a modification. Suppose we observe all non-empty intersections $[\vec{T}_i, X_i, \Theta_i, m]^* = [\vec{T}_i, X_i, \Theta_i] \cap \mathcal{W}(T_{2,i}, X_i, \Theta_i, m)$, but only if the left endpoint of $[\vec{T}_i, X_i, \Theta_i, m]^* \in B_n$. In other words, we observe all fragments with left endpoint in the interior of B_n, without censoring at the boundary of B_n. If we assume that these new fragments are a Poisson process, we can condition of their observed number and consider their lengths and censoring types as an independent sample. To describe their common distribution we need to introduce modifications ν_n^* of the measures ν_n. Define

$$l^*(t, \vartheta, m) = \begin{cases} |\mathcal{W}(t, \vartheta, m)| & \text{if the left endpoint of } \mathcal{W}(t, \vartheta, m) \text{ falls in } B_n. \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathrm{d}\nu_n^*(l) = \int_t \int_\vartheta \sum_m \delta_{l_n^*(t,\vartheta,m)}(\mathrm{d}l)\mathrm{d}t\mathrm{d}K(\vartheta).$$

Conditionally on $\mathcal{W}$, the distribution of the lengths and censoring types of the fragments without censoring at ∂B_n is the same as that of the original fragments, but with ν replaced by ν^* throughout.

We shall now establish the convergence of various functions of ν_n. We shall make extensive use of an ergodic theorem for spatial processes due to Nguyen and Zessin (1979, corollary 4.20).

A spatial process on a set S is a family $\{X_G : G \in \mathcal{G}\}$ where $\mathcal{G}$ is the collection of all bounded Borel sets in S. The spatial process is said to be *covariant* if for any $G \in \mathcal{G}$ and $s \in S$

$$X_{G+s}(T_s) = X_G(.)$$

It is said to be *additive* if

$$X_{H\cup G} = X_H + X_G, \quad \text{if } H, G \in \mathcal{G}, \text{ and } \quad H \cap G = \varnothing.$$

We denote by $\mathcal{K}$ the collection of bounded and convex subsets of S and by B the unit square. We have the following theorem (Nguyen and Zessin (1979),

Theorem 2.1. *If a spatial process $\{X_G : G \in \mathcal{G}\}$ is covariant and additive and if there exists a non-negative random variable $Y \in L^1(P)$ such that*

$$|X_G| \leq Y, \text{ a.s. for each } \quad G \in B \cap \mathcal{K}$$

then the limit

$$\lim_{n\to\infty} \frac{1}{|G_n|} X_{G_n} = E(X_B|\mathcal{J})$$

exists almost surely for each regular countable sequence G_n of sets in $\mathcal{K}$.

Here $\mathcal{J}$ denotes the invariant sigma-algebra. If $\{X_G : G \in \mathcal{G}\}$ is ergodic, i.e. $\mathcal{J}$ is trivial, then the limit equals $E(X_B)$.

We want to establish convergence of V_n and the sub-distribution functions $F_n(y,\delta) = \int_0^y P_n(\mathrm{d}y', \delta)$. To this end, we must first prove convergence—upon normalization—of $|W_n| = \int_l l\mathrm{d}\nu_n(l)$, $\kappa_n = \int_l \mathrm{d}\nu_n(l)$, $a_n = 1/(|W_n| + \kappa_n\tau)\int_{l>\tau}(l-\tau)\mathrm{d}\nu_n(l)$ and $b_n = (1/\kappa_n)\int_{l>\tau}\mathrm{d}\nu_n(l)$. We define $|W_n|^*$, κ_n^* a_n^* and B_n^* similarly as integrals with respect to ν_n^*. We can think of these as random quantities indexed by sets B_n. As such they are additive, covariant and ergodic. The sequence B_n is a regular, countable sequence of convex bounded sets, and $|B_n| = n^2$. By Theorem 2.1 we have $|W_n|^*/n^2 \to E_{\mathcal{W})}|W_1|^*$, $\kappa_n^*/n^2 \to E_{\mathcal{W})}\kappa_1^*$, $a_n^*/n^2 \to E_{\mathcal{W})}a_1^*$ and $b_n^*/n^2 \to E_{\mathcal{W})}\kappa_1^*$. It is not difficult to show that $(|W_n|^* - |W_n|)/n^2 \to 0$, $(\kappa_n - \kappa_n^*)/n^2 \to 0$, $(a_n - a_n^*)/n^2 \to 0$ and $(b_n - b_n^*)/n^2 \to 0$.

We can now show convergence of V_n and its relatives. Recall our definition of V_n,

$$V_n(x) = \int_0^x \frac{|W_n| + y\kappa_n}{|W_n| + \mu\kappa_n}\mathrm{d}F(y)$$

Proposition 2.1. *As n tends to infinity V_n tends almost surely, uniformly to*

$$V_\infty(x) = \int_0^x \frac{E_{\mathcal{W}}|W_1|^* + yE_{\mathcal{W}}\kappa_1^*}{E_{\mathcal{W}}|W_1|^* + \mu E_{\mathcal{W}}\kappa_1^*}\mathrm{d}F(y).$$

Proof We divide numerator and denominator by n^2 and then use the convergence of $|W_n|/n^2$ and κ_n/n^2 to obtain pointwise convergence. Uniform convergence follows from the usual arguments as V_n and V_∞ are distribution functions. □

Of course g_n, G_n, h_n and H_n converge to limits g_∞, G_∞, h_∞ and H_∞, which are defined in the obvious manner as integrals with respect to V_∞.

In section 2.3, we derived the distribution $P(\mathrm{d}y, \delta)$ of the length and censoring type of the fragments observed through $W = W_1$, under the assumption that they are independent. We noted that for a certain Markov kernel $K(\mathrm{d}y, \delta; x)$ we can write $P(\mathrm{d}y, \delta) = KV(\mathrm{d}y, \delta) = \int_x K(\mathrm{d}y, \delta; x)\mathrm{d}V(x)$. For the fragments observed through W_n we find, for $y \in [0, \tau)$,

$$\begin{aligned}
P_n(\mathrm{d}y, 0) &= \frac{\int_{l \geq y}(l - y)\mathrm{d}\nu_n(l)}{|W_n| + \kappa_n y}\mathrm{d}V_n(y) \\
P_n(\mathrm{d}y, 1) &= 2\nu_n([y, \infty))g_n(y)\mathrm{d}y \\
P_n(\mathrm{d}y, 2) &= \mathrm{d}\nu_n(y)h_n(y) \\
P_n(y \geq \tau) &= a_n G_n + b_n H_n.
\end{aligned}$$

Again, P_n can be written as $\int_x K_n(\mathrm{d}y, \delta; x)\mathrm{d}V_n(x)$, where

$$\begin{aligned}
K_n(\mathrm{d}y, \delta; x) &= \mathbf{1}_0(\delta)\mathbf{1}_{[0,\tau)}(y)\frac{\int_{l \geq x}(l - x)\mathrm{d}\nu_n(l)}{|W_n| + \kappa_n x}\delta_x(\mathrm{d}y) \\
&\quad + \mathbf{1}_1(\delta)\mathbf{1}_{[0,\tau)}(y)2\nu_n([y, \infty))\left(\frac{\mathbf{1}_{(y,\tau)}(x)}{|W_n| + \kappa_n x} + \frac{\mathbf{1}_{\{\dagger\}}(x)}{|W_n| + \kappa_n \tau}\right)\mathrm{d}y \\
&\quad + \mathbf{1}_2(\delta)\mathbf{1}_{[0,\tau)}(y)\left(\frac{\mathbf{1}_{(y,\tau)}(x)(x - y)}{|W_n| + \kappa_n x} + \frac{\mathbf{1}_{\{\dagger\}}(x)(\tau - y)}{|W_n| + \kappa_n \tau} + \frac{\mathbf{1}_{\{\ddagger\}}(x)}{\kappa_n}\right)\mathrm{d}\nu_n(y) \\
&\quad + \mathbf{1}_3(\delta)\delta_\tau(\mathrm{d}y)(a_n\mathbf{1}_{\{\dagger\}}(x) + b_n\mathbf{1}_{\{\ddagger\}}(x)).
\end{aligned}$$

Let K_∞ be defined by

$$\begin{aligned}
K_\infty(\mathrm{d}y, \delta; x) &= \mathbf{1}_0(\delta)\mathbf{1}_{[0,\tau)}(y)\frac{\int_{l \geq x}(l - x)\mathrm{d}\nu_1^*(l)}{|W_1|^* + \kappa_1^* x}\delta_x(\mathrm{d}y) \\
&\quad + \mathbf{1}_1(\delta)\mathbf{1}_{[0,\tau)}(y)2\nu_1^*([y, \infty))\left(\frac{\mathbf{1}_{(y,\tau)}(x)}{|W_1|^* + \kappa_1^* x} + \frac{\mathbf{1}_{\{\dagger\}}(x)}{|W_1|^* + \kappa_1^* \tau}\right)\mathrm{d}y \\
&\quad + \mathbf{1}_2(\delta)\mathbf{1}_{[0,\tau)}(y)\left(\frac{\mathbf{1}_{(y,\tau)}(x)(x - y)}{|W_1|^* + \kappa_1^* x} + \frac{\mathbf{1}_{\{\dagger\}}(x)(\tau - y)}{|W_1|^* + \kappa_1^* \tau} + \frac{\mathbf{1}_{\{\ddagger\}}(x)}{\kappa_1^*}\right)\mathrm{d}\nu_1^*(y) \\
&\quad + \mathbf{1}_3(\delta)(a_1^*\mathbf{1}_{\{\dagger\}}(x) + b_1^*\mathbf{1}_{\{\ddagger\}}(x)).
\end{aligned}$$

Proposition 2.2. *$K_n V_n(\mathrm{d}y, \delta)$ tends to $K_\infty V_\infty(\mathrm{d}y, \delta)$ almost surely as n tends to infinity, in the sense of uniform convergence of the associated (sub)distribution functions.*

Proof The proof is similar to that of the previous proposition. □

So far in this section we have only considered functionals of the random set $\mathcal{W}$ as observed through increasing sets B_n. We now turn to the actual data, which are fragments arising from observation of $\Phi \cap \mathcal{W}$ through the B_n. We introduce a sequence of point processes, representing the data observed through B_n.

Recall that we defined

$$\mathcal{D}_n = \{(\vec{t}, x, \vartheta, m) : \ W_n(t_2, \vartheta, m) \neq \emptyset\}.$$

Now define

$$\mathcal{E}_n = \{(\vec{t}, x, \vartheta, m) \in \mathcal{D}_n \ : \ [\vec{t}, x, \vartheta] \cap W_n(t_2, \vartheta, m) \neq \emptyset\}.$$

$\mathcal{E}_n$ is well defined for $n = \infty$, if we take $B_\infty = \mathbb{R}^2$ and $W_\infty(t_2, \vartheta, m) = \mathcal{W}(t_2, \vartheta, m)$. On the sets $\mathcal{E}_n$ we define three functions, p_n, y_n and δ_n.

$$\begin{aligned}
p_n(\vec{t}, x, \vartheta, m) &= \text{left-most endpoint of } [\vec{t}, x, \vartheta] \cap W_n(t_2, \vartheta, m) \\
y_n(\vec{t}, x, \vartheta, m) &= \min(\tau, |[\vec{t}, x, \vartheta] \cap W_n(t_2, \vartheta, m)| \\
\delta_n(\vec{t}, x, \vartheta, m) &= \#\text{ endpoints of } [\vec{t}, x, \vartheta] \text{ outside of } W_n(t_2, \vartheta, m) \\
&\qquad \text{or 3 when } y_n(\vec{t}, x, \vartheta, m) = \tau.
\end{aligned}$$

These definitions are also valid for $n = \infty$. Define, for $n = 1, 2, \ldots, \infty$, point processes Π_n by applying these functions to $\Phi'' \cap \mathcal{E}_n$. Thus, Π_n are marked point processes on B_n with mark space $\mathcal{M} = (0, \tau] \times \{0, 1, 2, 3\}$ representing the fragments observed through B_n. Π_∞ is of special interest. It is a marked point process on $\mathbb{R}^2 \times \mathcal{M}$ representing all fragments arising from intersecting Φ with $\mathcal{W}$. At the beginning of the present section, we claimed that as n tends to infinity, the censoring at the boundary of B_n becomes negligible as compared to the censoring at the boundary of $\mathcal{W}$. The next lemma makes this claim precise.

Lemma 2.2. *For any measurable set $M \subseteq \mathcal{M}$,*

$$\frac{1}{n^2} E_\Phi(|\Pi_n(B_n \times M) - \Pi_\infty(B_n \times M)||\mathcal{W}) \to 0$$

almost surely, as n tends to infinity

Proof The difference between $\Pi_n(B_n \times M)$ and $\Pi_\infty(B_n \times M)$ is less than the number of fragments that cross the boundary of B_n. This, in turn, is less than the number of line segments that cross the boundary of B_n. The expected number of line segments that cross a fixed line segment of length

n (such as one side of B_n) is no more than $n\lambda\mu$, where μ is the (finite) mean of F. This is seen by noting that for a line segment of length x and orientation ϑ to cross a given object line segment, its left-endpoint must lie in a parallelogram with the object line segment for its base, and sides of length x. The area of such a parallelogram is always less than nx. Integrating this with respect to $\lambda \mathrm{d}\vec{t}\mathrm{d}F(x)\mathrm{d}K(\vartheta)$ we find our upper bound, $n\lambda\mu$. We conclude that the expected difference between $\Pi_n(B_n \times M)$ and $\Pi_\infty(B_n \times M)$ is less than $4n\lambda\mu$ and the stated result follows. □

For a given measurable set $M \subseteq \mathcal{M}$, we have

$$\int_M P_n(\mathrm{d}y, \delta) = P_n(M) = \frac{E_\Phi(\Pi_n(B_n \times M)|\mathcal{W})}{E_\Phi(\Pi_n(B_n \times \mathcal{M})|\mathcal{W})}.$$

We define

$$P_n^*(M) = \frac{E_\Phi(\Pi_\infty(B_n \times M)|\mathcal{W})}{E_\Phi(\Pi_\infty(B_n \times \mathcal{M})|\mathcal{W})}$$

and

$$\mathbb{P}_n(M) = \frac{\Pi_n(B_n \times M)}{\Pi_n(B_n \times \mathcal{M})}.$$

We draw the reader's attention to the fact that the subscript n refers to the fact that $\Phi \cap \mathcal{W}$ is observed through B_n and *not* to the number of observed fragments. The number of observed fragments, in fact, equals $\Pi_n(B_n \times \mathcal{M})$.

Everything is now in place for the final result of this section.

Proposition 2.3. *For every measurable set $M \subseteq \mathcal{M}$,*

$$\mathbb{P}_n(M) \to P_\infty(M),$$

in probability, as n tends to infinity.

Proof First, note that $\Pi_\infty(B_n \times M)$ is an additive covariant function indexed by the sets B_n. Since we assume that Φ and $\mathcal{W}$ are jointly ergodic it follows by Theorem 2.1 that $\Pi_\infty(B_n \times M) \to E\Pi_\infty(B_n \times M)$, almost surely. Hence,

$$\frac{\Pi_\infty(B_n \times M)}{\Pi_\infty(B_n \times \mathcal{M})} \to \frac{E\Pi_\infty(B_n \times M)}{E\Pi_\infty(B_n \times \mathcal{M})}.$$

It immediately follows that $P_n^*(M)$ converges to the same limit.

By Lemma 2.2, $P_n^*(M)$ and $P_n(M)$ have the same limit. In Proposition 2.2 we found that $P_n(M)$ converges to $P_\infty(M)$, so we may now conclude that

$$\frac{\Pi_\infty(B_n \times M)}{\Pi_\infty(B_n \times \mathcal{M})} \to P_\infty(M).$$

The reasoning of Lemma 2.2 allows us also to conclude that

$$\left| \frac{\Pi_n(B_n \times M)}{\Pi_n(B_n \times \mathcal{M})} - \frac{\Pi_\infty(B_n \times M)}{\Pi_\infty(B_n \times \mathcal{M})} \right| \to 0$$

in probability. Hence, our claim follows. □

Of course the above convergence is uniform over sets M of the form $[0, y] \times \{\delta\}$ ($y < \tau$, $\delta = 0, 1, 2, 3$), corresponding to 'empirical sub-distribution functions'. Still, the convergence is weaker than when we were working under the Poisson assumption. The Glivenko–Cantelli theorem yields *almost sure* uniform convergence, and here we only have uniform convergence in probability. Fortunately, this is not much of a problem. We simply apply the Skorohod–Dudley–Wichura *almost sure representation theorem* as stated, for instance in Gill (1989). For us, this boils down to pretending that we do have almost sure convergence and going ahead and prove almost sure uniform convergence of our estimator. In the end, we must then weaken the result to uniform convergence in probability.

2.6 An estimator for the line segments

With $\mathbb{P}_n$ as defined in the previous section, we consider estimators

$$\begin{aligned} \mathrm{d}\hat{V}_n(x) &= \mathrm{d}\mathbb{P}_n(x, 0) \\ &+ \frac{\mathrm{d}\hat{V}_n(x)}{|W_n| + \kappa_n x} \int_{y=0}^{x} \frac{1}{\hat{g}_n(y)} \mathrm{d}\mathbb{P}_n(y, 1) \\ &+ \frac{\mathrm{d}\hat{V}_n(x)}{|W_n| + \kappa_n x} \int_{y=0}^{x} \frac{x - y}{\hat{h}_n(y)} \mathrm{d}\mathbb{P}_n(y, 2) \qquad (2.37) \\ \hat{H}_n &= \hat{H}_n \int_{y=0}^{\tau-} \frac{1}{\hat{h}_n(y)} \mathrm{d}\mathbb{P}_n(y, 2) + \frac{b_n \hat{H}_n}{a_n \hat{G}_n + b_n \hat{H}_n} \mathbb{P}_n(Y > \tau)) \qquad (2.38) \end{aligned}$$

where, cf. (2.16)

$$\hat{G}_n = 1 - \hat{H}_n - \hat{V}_n(\tau^-).$$

the functions $\hat{h}_n$ and $\hat{g}_n$ are defined as usual. These estimators are of course inspired by (2.35) and (2.36), but it should be noted that the $\mathbb{P}_n$ have a different meaning here. The subscript n here refers to observation through $W_n = \mathcal{W} \cap B_n$, while the subscript n in (2.35) and (2.36) referred to n observed fragments.

We estimate $(F|_{0,\tau)}, \mu)$ by

$$\hat{\mu}_n = \frac{1}{\kappa_n}\left(\frac{1}{\int_{[0,\tau)} \frac{1}{|W_n|+\kappa_n y}\mathrm{d}\hat{V}_n(y) + \hat{g}_n(\tau)} - |W_n|\right)$$

and

$$\hat{F}_n(x) = \frac{1}{\int_{[0,\tau)} \frac{1}{|W_n|+\kappa_n y}\mathrm{d}\hat{V}_n(y) + \hat{g}_n(\tau)} \int_0^x \frac{1}{|W_n|+\kappa_n y}\mathrm{d}\hat{V}_n(y).$$

We now prove the consistency of our estimator.

Theorem 2.2. *Suppose that $\mathcal{W}$ and Φ are jointly ergodic. Then*

$$\|\hat{F}_n - F_0\|_{[0,\tau)} \to 0 \quad \textit{and} \quad |\hat{\mu}_n - \mu_0| \to 0$$

in probability

Proof The proof will follow the outline we gave at the end of section 2.1.1. Also, we use elements of an earlier consistency proof we gave at the end of section 2.4. However, we no longer need the 'working hypothesis' of that section. There we had difficulty proving that the mle of $g(\tau)$ is bounded away from zero. Since the function g is non-decreasing, it suffices to show that for n large enough $\hat{G}_n(\tau)$ stays away from zero if $G_n(\tau)$ is positive. Because of the extra censoring, this is not a problem anymore.

Proposition 2.1 established that for all $F \in \mathcal{F}$ (and in particular for F_0) $V_{n,F}$ tends to $V_{\infty,F}$ almost surely, uniformly on $[0,\tau)$. Also, Proposition 2.2 yielded that $K_n V_{n,F}$ tends to $K_\infty V_{\infty,F}$ almost surely, in the sense of uniform convergence over sets of the form $[0,y] \times \{\delta\}$ $(y < \tau,\ \delta = 0,1,2,3)$, corresponding to 'empirical sub-distribution functions'.

To apply Theorem 1.2, we need to compare the sequence $K_n\hat{V}_n$ with another sequence $K_n\tilde{V}_n$, say, which converges to $K_\infty V_{\infty,V_{F_0}} = K_\infty V_{\infty,V_0}$. For $\tilde{V}_n$ we choose an appropriately scaled version of the 'empirical average' of uncensored fragments:

$$\tilde{V}_n(x) = \int_0^x \frac{|W_n| + \kappa_n y}{\int_{l\geq y}(l-y)\mathrm{d}\nu(l)}\mathbb{P}_n(\mathrm{d}y, 0), \qquad x \in [0,\tau).$$

We define $\tilde{V}_n(\tau^-)$ in the obvious way. An ergodic theorem can be used to show its convergence of $\tilde{V}_n(\tau^-)$ to $V_{\infty,F_0}(\tau^-)$. Hence, we can take a sequence $\tilde{V}_n(\{\ddagger\}) = \tilde{H}_n$ (such that $\tilde{H}_n \leq \tilde{V}_n(\tau^-)$) converging to $V_{\infty,F_0}(\{\ddagger\}) = \tilde{H}_{\infty,F_0}$. Much like we did in section 2.4, we can apply empirical process theory to

obtain almost sure uniform convergence of $\tilde{V}_n$ to V_{∞,F_0}. Again using empirical process theory, we obtain almost sure convergence of $K_n\tilde{V}_n$ to $K_\infty V_{\infty,F_0}$, uniformly over sets of the form $[0,y]\times\{\delta\}$.

Proposition 2.3 tells us that $\mathbb{P}_n$ also converges to $P_\infty = K_\infty V_{\infty,F_0}$, uniformly over sets of the form $[0,y]\times\{\delta\}$. However, this convergence is 'in probability'. For the remainder of this proof we assume that the convergence is 'almost sure'. This will lead to almost sure consistency of our estimator, which we then modify to consistency in probability, as stated in the theorem. We may do so by the Skorohod–Dudley–Wichura 'almost sure representation theorem', which we discussed above.

Using the fact that $K_n\tilde{V}_n(\mathrm{d}y,0) = \mathbb{P}_n(\mathrm{d}y,o)$ and applying Lemma 2.1, we can now conclude that

$$\int \left(\frac{\mathrm{d}K_n\tilde{V}_n}{\mathrm{d}K_n\hat{V}_n}\right) \mathrm{d}(\mathbb{P}_n - K_n\tilde{V}_n) \to 0$$

almost surely. Theorem 1.2 now allows us to conclude that $K_n\hat{V}_n$ converges to $K_\infty V_{\infty,F_0}$ in the same sense as $K_n\tilde{V}_n$ does. Since we have made sure that our parameter is identifiable it follows that $\hat{V}_n$ converges to V_{∞,F_0}, whence it follows that $\hat{F}_n = F_{n,\hat{V}_n}$ converges to $F_{\infty,V_{\infty,F_0}} = F_0$. □

Chapter 3

Chord Length Estimation

3.1 Introduction

Consider a random closed set $\mathcal{W} \subset \mathbb{R}^2$ which we observe through a bounded 'window' B. Important characteristics of the probability distribution of a random set are the chord length distribution (CLD) and the—perhaps more familiar—linear contact distribution (LCD) (Stoyan et al. 1987 p. 178–188). These characteristics can give us some insight in the 'architectural' features of the random set, such as the shape, size and dispersion of its various connected parts. Also, the CLD and LCD can help us when we want to test for stationarity (translation invariance) or for isotropy (rotation invariance).

The CLD and LCD are closely related as is established in formula (3.24) below. With this relation, estimates for the one can be transformed into estimates for the other. For instance, in Hansen et al. (1996) an estimator for the LCD is developed and it is noted that an estimator for the CLD can be obtained from it. Our primary interest is in estimation of the CLD and we shall argue that it is better to estimate the CLD and transform it into an estimate of the LCD than to do it the other way around.

The 'classical' definition of a chord or an intercept (Weibel, 1979, p. 315) is a "line segment contained fully within an object and extending between two points on the object's surface".

We can informally define the chord length distribution in the direction e as the distribution of the length of the longest line segment through the origin in the direction e which is fully contained within $\mathcal{W}$, conditionally on the event that the origin lies in $\mathcal{W}$. If $\mathcal{W}$ is stationary and rotation invariant then this distribution does not depend on the choice of the reference point (the origin) or on the direction e. Hence, if we should want to test whether $\mathcal{W}$ is isotropic, we could estimate the CLD in various directions and see if the

estimates differ. Note that our definition introduces a length bias; the origin is more likely to fall in a 'large' part of $\mathcal{W}$ and hence the chord through the origin will have a tendency to be unusually long. If we should want, we could easily correct for this to obtain an 'unweighted' CLD (see formula (3.3)) below. For further discussion on chord lengths cf. Serra (1982, Chapter X), Matheron (1975, p. 53) and Stoyan et al. (1987, pp. 178–182).

The CLD has been applied in biology to study certain features of tissue, especially to characterize lung airspace dimensions, see Rosenthal (1989), Lum et al. (1990) and Oldmixon et al. (1994). Estimation of the CLD can be used also for inferential purposes in the statistical analysis of random set models (Stoyan et al. 1987, Hansen et al. 1996).

Estimation of certain characteristics of the chord length distribution can be done from stereological estimates of the surface to volume ratio and the volume fraction. Such stereological estimates are obtained by straightforward intersection and point counting methods (Stoyan et al. 1987 p. 180). These methods are simple, fast and easy to deal with, but they do have their limitations. Stereological mean value formulas can only be used to obtain certain low-dimensional characteristics of the CLD, whereas knowledge of the shape of the CLD may give important additional information about the structure of the random set under consideration.

Estimation is, of course, influenced by edge effects as the random set $\mathcal{W}$ is only observed in a bounded observation window B. More specifically, when a certain point x is used as a reference point, the chord through x could extend beyond B. A given chord is called uncensored (u.c.) when both endpoints are in B; singly censored (s.c.) when one of its endpoints is outside B; and doubly censored (d.c.) when both endpoints are outside B.

Estimation from spatial data in the presence of censoring is often dealt with by means of 'minus sampling' which is also known as the 'border method'. In our situation this means that when we want to estimate the probability that the length of a chord is less than r, we restrict attention to those reference points which are further than r away from the boundary of B. The resulting estimator can be termed the 'reduced sample estimator' (Baddeley and Gill, 1997). The obvious disadvantage of minus sampling is that much information is discarded, especially for large values of r.

Oldmixon et al. (1994) suggest to make two separate estimates; one based only on the uncensored chords and another one based only on the singly censored chords. These two estimates could then be combined by taking some convex combination of them. Olmixon et al. do not discuss to do so optimally. Also, they discard doubly censored chords.

Estimation from censored observations belongs traditionally to the field of survival analysis, where it has—among other things—resulted in the de-

velopment and study of the famous Kaplan–Meier estimator. Laslett (1982a and b) first noted the similarity of censoring in survival studies and edge effects in spatial statistics. Subsequently, this analogy has been exploited by Wijers (1995), van der Laan (1996,1998), Hansen et al. (1996), Baddeley and Gill (1997) and by ourselves in the previous chapter of this thesis. In particular, in Hansen et al. (1996) a Kaplan–Meier type estimator was proposed for the linear contact distribution. The unique correspondence between the LCD and the CLD suggests a transformation of the estimator of the former to obtain an estimator of the latter. Unfortunately, doing so involves a numerically unstable differentiation, which is believed to cause loss of efficiency.

We want to use semi-parametric methods and missing data theory but cannot do so directly. Much like we did in the previous section, we first consider an analogous problem. We derive an NPMLE for *that* problem and then show that the estimator can also be applied in the real problem. The estimator will not be the maximum likelihood estimator in the real problem. However, it does utilize *all* the available data—though not in the absolutely most efficient way.

The organization of this chapter is as follows. First, we give a more precise definition of the chord length distribution and discuss why it is difficult to estimate it. Then we consider an analogous, but similar problem and derive the nonparametric maximum likelihood estimator. We proceed to show that this NPMLE is consistent for the original problem. We report a modest simulation experiment to try out our new estimator. Finally, we shall consider the linear contact distribution and study its relation to the chord length distribution.

3.2 The chord length distribution

Let $\mathcal{W}$ be a stationary random closed set in $\mathbb{R}^2$ and e a unit vector in $\mathbb{R}^2$. The *chord* through a point of $\mathcal{W}$ in the direction e is the longest line segment through the given point with orientation e which is fully contained within $\mathcal{W}$. We are interested in estimation of the distribution of the length of the chord through the origin, given that the origin belongs to $\mathcal{W}$. Since $\mathcal{W}$ is stationary, the particular choice of reference point (in this case the origin) is irrelevant. The 'chord length distribution' (CLD) is an interesting feature of the distribution of $\mathcal{W}$. The data on which we shall base our estimation will consist of n independent realizations of $\mathcal{W}$ which are all observed through a fixed compact set B.

Let $\xi_e(s,t)$ denote the chord in the direction e through the point $(s,t) \in$

$\mathcal{W}$. The situation is clarified in figure 3.1. A formal definition of the chord

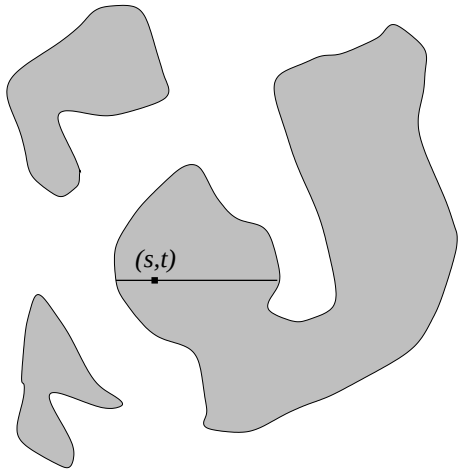

Figure 3.1: The chord $\xi_{(1,0)}(s,t)$. The shaded region is (part of) the random set $\mathcal{W}$.

length distribution is as follows. Let $W = \mathcal{W} \cap B$ where B is some compact set of positive area. Define, for all positive x

$$W(x) = \{(s,t) \in W : |\xi_e(s,t)| \leq x\} \tag{3.1}$$

where $|\xi_e(s,t)|$ denotes the length of the chord $\xi_e(s,t)$. Now define the (length weighted) chord length distribution function as

$$L_e(x) = \frac{E|W(x)|}{E|W|} \tag{3.2}$$

where the expectation is with respect to the distribution of $\mathcal{W}$ and $|.|$ denotes area. It is easily seen that this definition does not depend on the choice of B.

For simplicity we fix $e = (1,0)$ and omit e from our notation.

We call L the *length weighted* chord length distribution for a reason. Recall that L may be interpreted as the distribution of the length of the chord though the origin, given that the origin lies in $\mathcal{W}$. Informally, the origin is more likely to fall in a big part of $\mathcal{W}$ than in a small part. Hence L is biased towards greater chord lengths. We can easily correct for this. The unweighted chord length distribution is

$$L_0(x) = \frac{\int_0^x \frac{1}{t} \mathrm{d}L(t)}{\int_0^\infty \frac{1}{t} \mathrm{d}L(t)}. \tag{3.3}$$

The subscript '0' indicating 'unweighted' should not be confused with the subscript 'e' we used earlier.

Of a chord $\xi(s,t)$ through a point $(s,t) \in W = \mathcal{W} \cap B$ we observe only $\xi(s,t) \cap B$. Hence, from observation of $W = \mathcal{W} \cap B$ we can not in general infer $W(x)$ as a chord through a given point in $\mathcal{W} \cap B$ might well extend beyond B. However, we do observe the length of the intersection of each chord with B and we observe how many of each chord's endpoints are outside of B. Define two functions

$$y(s,t) = |\xi(s,t) \cap B| \tag{3.4}$$

$$\delta(s,t) = \#\text{ endpoints of } \xi(s,t) \text{ outside of } B. \tag{3.5}$$

From observation of W we know, for all positive y and $\delta = 0, 1$ or 2

$$A(y,\delta) = \{(s,t) \in W : y(s,t) \leq y,\ \delta(s,t) = \delta\}. \tag{3.6}$$

We define

$$F(y,\delta) = \frac{E|A(y,\delta)|}{E|W|}, \tag{3.7}$$

where again the expectation is with respect to the distribution of $\mathcal{W}$. For $\delta = 0, 1, 2$, $F(y,\delta)$ are sub-distribution functions.

We can interpret L and F in the following way which will be very useful in the next section.

Lemma 3.1. *Let (S,T) be uniformly distributed on W. Then*

1. *L is the distribution of $|\xi(S,T)|$;*
2. *F is the distribution of $(y(S,T), \delta(S,T))$;*
3. *conditionally on $\xi(S,T)$, (S,T) is uniformly distributed on $\xi(S,T)$.*

Proof The first two statements are immediate from the definitions of L and F. To prove the third, define, for all t, $W_t = \{s : (s,t) \in W\}$. Conditionally on $\xi(S,T)$, T is degenerate. Also, conditional on T, it is clear that S is uniformly distributed on W_T. Now W_T consists of one and possibly several line segments, $\xi(S,T)$ being one of these. Hence, conditionally on $\xi(S,T)$, (S,T) is uniformly distributed on $\xi(S,T)$. □

We now turn to the statistical problem of estimating L from independent observations. Let $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n$ be independent and identically distributed as $\mathcal{W}$. For simplicity, let $B = [0,\tau] \times [0,1]$ ($\tau > 0$). Note that doubly censored chords ($\delta = 2$) for which neither endpoint is in B, will always have length τ.

Define $W_i = \mathcal{W}_i \cap B$ and let $\xi(i,s,t)$ be the chord through (s,t) in $\mathcal{W}_i$. Like before,

$$W_i(x) = \{(s,t) \in W_i : |\xi(i,s,t)| \leq x\} \tag{3.8}$$
$$y(i,s,t) = |\xi(i,s,t) \cap B| \tag{3.9}$$
$$\delta(i,s,t) = \#\text{ endpoints of } \xi(i,s,t) \text{ outside of } B. \tag{3.10}$$
$$A_i(y,\delta) = \{(s,t) \in W_i : y(i,s,t) \leq y,\ \delta(i,s,t) = \delta\} \tag{3.11}$$

We really need only one result about the observed chords:

Lemma 3.2.
$$\frac{\sum_i |A_i(y,\delta)|}{\sum_i |W_i|} \to F(y,\delta)$$
almost surely, uniformly.

Proof Pointwise convergence follows from the strong law of large numbers. By the usual 'Glivenko–Cantelli argument', uniform convergence follows as usual from the fact that F is monotone and bounded. □

3.3 Digression: A related problem

We can think of chord length estimation as a *missing data* problem. The complete data consists of the W_i and $\{\xi(i,s,t) : (s,t) \in W_i,\ i = 1,\ldots n\}$. The observed data are the W_i and $\{(y(i,s,t),\ \delta(i,s,t)) : (s,t) \in W_i,\ i = 1,\ldots n\}$. Application of the EM algorithm springs to mind, but we do not see how to how to find the conditional distribution of the complete data given the observed data, as required by that algorithm.

In this section we consider a *different* missing data problem which is easier to handle as the observations will be numbers instead of functions on random sets. However, this easier problem will bear so much resemblance to chord length estimation that the analysis will be very useful. In fact, our findings in this section will guide us to an estimator which we can use in the chords problem. The reader should keep in mind that in this section there are no random sets $\mathcal{W}_i$ involved. However, we shall re-use many of the symbols of the first section to point out the similarities between the two problems. We hope this will clarify and not confuse.

Let $B = [0,\tau] \times [0,1]$ and let (S,T) be a uniformly distributed point in B. Associate with (S,T) a random variable X which is distributed according to L as defined in (3.2). Conditional on S and X, let R be a random variable which is uniformly distributed on the interval $[S_i - X_i, S_i]$. Now let ξ be the line-segment from (R,T) to $(R+X,T)$. Note that, given ξ, (S,T) is

uniformly distributed on ξ. Define $Y = |\xi \cap B| = |[R, R+X] \cap [0,\tau]|$ and Δ is the number of endpoints of ξ that fall outside of B which is the same as the number of endpoints of $[R, R+X]$ that fall outside of $[0,\tau]$. We see that our specific choices of $B = [0,\tau] \times [0,1]$ and $e = (1,0)$ have made T irrelevant; from (S, X, R) we can derive (Y, Δ)

Lemma 3.3. *The joint distribution of Y and Δ is given by F as defined in (3.7).*

Proof This is an immediate consequence of Lemma 3.1. □

The joint distribution of (S, X, R) is of course given by

$$\mathbf{1}_A(s,x,r)\frac{\mathrm{d}s}{\tau}L(\mathrm{d}x)\frac{\mathrm{d}r}{x} \tag{3.12}$$

where $A = \{(s,x,r) : [r, r+x] \cap [0,\tau] \neq \emptyset\}$.

From our description of the model we can explicitly derive the distribution of the observed data (Y, Δ). We introduce very two useful quantities

$$G = \int_{[\tau,\infty)} \frac{\tau}{x}\mathrm{d}L(x) \quad \text{and} \quad H = \int_{[\tau,\infty)} \frac{x-\tau}{x}\mathrm{d}L(x) \tag{3.13}$$

and note that

$$L(\tau^-) + G + H = 1 \tag{3.14}$$

where $L(\tau^-) := \int_0^{\tau-} \mathrm{d}L(x) := \int_{(0,\tau)} \mathrm{d}L(x)$.

We can write down the distribution of (Y, Δ) in terms of L restricted to $[0,\tau)$, G and H. (Re)define functions $y(s,x,r) = |[r, r+x] \cap [0,\tau]|$ and $\delta(s,x,r) =$ the number of endpoints of $[r, r+x]$ outside $[0,\tau]$. Now consider sets

$$\begin{aligned}
S(y,0) &= \{(s,x,r) : y(s,x,r) = y,\ \delta(s,x,r) = 0\} \\
&= \{(s,x,r) : x = y,\ 0 < r < \tau - x\} \\
S(y,1) &= \{(s,x,r) : y(s,x,r) = y,\ \delta(s,x,r) = 1\} = S(y,lc) \cup S(y,rc) \\
S(y,lc) &= \{(s,x,r) : x > y,\ s < y,\ r = y - x\} \\
S(y,rc) &= \{(s,x,r) : x > y,\ s > \tau - y,\ r = \tau - y\} \\
S(\tau,2) &= \{(s,x,r) : y(s,x,r) = \tau,\ \delta(s,x,r) = 2\} \\
&= \{(s,x,r) : x > \tau,\ \tau - x < r < 0\}.
\end{aligned}$$

Integrating the distribution of (S, X, R) given in (3.12) over these sets we find, for $0 \leq y < \tau$,

$$\begin{aligned} P(\mathrm{d}y, 0) &= \int_{A \cap S(y,0)} \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= \int_{x=0}^{\infty} \int_{s=0}^{\tau} \int_{r=s-x}^{s} \mathbf{1}_y(x) \mathbf{1}_{\{0<r<\tau-x\}}(r) \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= \int_{s=0}^{\tau} \int_{r=s-y}^{s} \mathbf{1}_{\{0<r<\tau-y\}}(r) \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{y} L(\mathrm{d}y) \end{aligned} \tag{3.15}$$

There is no need to further evaluate this expression, but should be it noted that it depends on L only through its restriction to $[0, \tau)$. For the distribution of the singly censored observations we have by symmetry of right and left censored observations

$$P(\mathrm{d}y, 1) = \int_{A \cap S(y,1)} \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) = 2 \int_{A \cap S(y,rc)} \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x).$$

Hence, for $0 \leq y < \tau$,

$$\begin{aligned} P(\mathrm{d}y, 1) &= 2 \int_{x=0}^{\infty} \int_{s=0}^{\tau} \int_{r=s-x}^{s} \mathbf{1}_{\{x>y\}}(x) \mathbf{1}_{\{s>\tau-y\}}(s) \mathbf{1}_{\{r=\tau-y\}}(r) \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= 2 \int_{x=y}^{\infty} \int_{s=\tau-y}^{\tau} \frac{\mathrm{d}s}{x} \frac{\mathrm{d}(\tau - y)}{\tau} \mathrm{d}L(x) \\ &= 2 \int_{x=y}^{\infty} \frac{y}{x\tau} \mathrm{d}y \mathrm{d}L(x) \\ &= 2 \left(\int_{x=y}^{\tau-} \frac{y}{x\tau} \mathrm{d}y \mathrm{d}L(x) + \frac{y}{\tau^2} \mathrm{d}y G \right), \end{aligned} \tag{3.16}$$

where we used that a random variable U (say) is uniformly distributed on $[0, \tau)$ if and only if $\tau - U$ has the same distribution.

The doubly censored observations are always of length τ.

$$\begin{aligned} P(Y = \tau, \Delta = 2) &= \int_{A \cap S(y,2)} \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= \int_{x=0}^{\infty} \int_{s=0}^{\tau} \int_{r=s-x}^{s} \mathbf{1}_{\{x>\tau\}}(x) \mathbf{1}_{\{\tau-x<r<0\}}(r) \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= \int_{x=\tau-}^{\infty} \int_{s=0}^{\tau} \int_{r=\tau-x}^{0} \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \mathrm{d}L(x) \\ &= \int_{x=\tau-}^{\infty} \frac{x - \tau}{\tau} \mathrm{d}L(x) \\ &= H \end{aligned} \tag{3.17}$$

Because G can be expressed in terms of $L|_{[0,\tau)}$ and H, we see that the distribution of the data is fully parameterized by $L|_{[0,\tau)}$ and H. This means that the model $\mathcal{L} = \{\text{all distributions on } \mathbb{R}^+\}$ is not identified. We could reduce $\mathcal{L}$ so that it is, but then the missing data problem will no longer be completely nonparametric. We now give a different model, that is both nonparametric and identified, while the distribution of the observed (Y, Δ) remains the same.

Consider a new model $\mathcal{L}$ for X of all distributions on a new space $[0, \tau) \cup \{\dagger, \ddagger\}$. We denote $L(\{\dagger\}) = P(X = \dagger) = G$ and $L(\{\ddagger\}) = P(X = \ddagger) = H$. We have the following missing data problem which is both completely nonparametric and identified:

Draw X from $L \in \mathcal{L}$

- if $L = l \in [0, \tau)$, then sample S uniformly on $[0, \tau]$ and given $S = s$ sample R uniformly on $[s - x, s]$. Finally, set $Y = [R, R + X] \cap [0, \tau]$ and Δ equal to the number of unobserved endpoints.
- if $X = \dagger$, Sample Y from $\mathbf{1}_{[0,\tau)}(y)2y/\tau^2$ and set $\Delta = 1$
- if $X = \ddagger$, set $(Y, \Delta) = (\tau, 2)$

It is easily verified that the distribution of the data (Y, Δ) is indeed given by formulas (3.15) to (3.15). $P(\mathrm{d}y, \delta)$ is of the form

$$P(\mathrm{d}y, \delta) = \int_{x \in [0,\tau) \cup \{\dagger,\ddagger\}} K(\mathrm{d}y, \delta; x) \mathrm{d}L(x)$$

where

$$\begin{aligned} K(\mathrm{d}y, \delta; x) = \mathbf{1}_0(\delta) \int_{s=0}^{\tau} \int_{r=s-x}^{s} \mathbf{1}_{\{0<r<\tau-x\}}(r) \frac{\mathrm{d}s}{\tau} \frac{\mathrm{d}r}{x} \delta_x(\mathrm{d}y) \\ + \mathbf{1}_1(\delta) 2 \left(\mathbf{1}_{(y,\tau)}(x) \frac{y}{x\tau} + \mathbf{1}_{\{\dagger\}}(x) \frac{y}{\tau^2} \right) \mathrm{d}y \\ + \mathbf{1}_2(\delta) \mathbf{1}_{\{\ddagger\}}(x) \delta_\tau(\mathrm{d}y) \end{aligned}$$

Now suppose we have n i.i.d. copies (S_i, T_i, X_i, R_i) of (S, T, X, R). Let ξ_i denote the line-segment from (R_i, T_i) to $(R_i + X_i, T_i)$ and $Y = |\xi_i \cap B| = |[R_i, R_i + X_i] \cap [0, \tau]|$ and Δ_i is the number of endpoints of ξ_i that fall outside of B. Let $\mathbb{F}_n$ denote the empirical distribution of the (Y_i, Δ_i). The EM equations are, cf. (1.17)

$$\begin{aligned} \widehat{L}_n(\mathrm{d}x) &= \mathbb{F}_n(\mathrm{d}x, 0) \\ &\quad + \int_{y=0}^{x} \frac{(y/x)\mathrm{d}\widehat{L}_n(x)}{\int_{x=y}^{\tau} (y/x)\mathrm{d}\widehat{L}_n(x) + (y/\tau)\widehat{G}_n} \mathbb{F}_n(\mathrm{d}y, 1) \\ \widehat{H}_n &= \mathbb{P}_n(\Delta = 2), \end{aligned}$$

where

$$\widehat{G}_n = 1 - \widehat{L}_n(\tau^-) - \widehat{H}_n.$$

3.4 Back to the chord lengths

We now return to our original problem of estimating the chord length distribution of a random set $\mathcal{W}$ observed through the set $B = [0,\tau] \times [0,1]$. Inspired by the previous section, we propose as an estimator $\hat{L}_n$ satisfying

$$\begin{aligned}\widehat{L}_n(\mathrm{d}x) &= \frac{\sum_i |A_i(\mathrm{d}x, 0)|}{\sum_i |W_i|} \\ &\quad + \int_{y=0}^{x} \frac{(y/x)\mathrm{d}\widehat{L}_n(x)}{\int_{x=y}^{\tau}(y/x)\mathrm{d}\widehat{L}_n(x) + (y/\tau)\widehat{G}_n} \frac{\sum_i |A_i(\mathrm{d}y, 1)|}{\sum_i |W_i|} \\ \widehat{H}_n &= \frac{\sum_i |A_i(\tau, 2)|}{\sum_i |W_i|},\end{aligned}$$

where

$$\widehat{G}_n = 1 - \widehat{L}_n(\tau^-) - \widehat{H}_n.$$

We could prove consistency of this estimator using Theorem 1.2. Such a proof would be strikingly similar to our proof of section 2.4. Instead of empirical (sub)distribution functions we have empirical averages of areas. This is not a problem, as we have already established the uniform convergence of these averages in lemma 3.2. We can again compare the maximum likelihood estimator to a comparison sequence of based on a simple transformation of the empirical distribution (empirical averages of areas) of the uncensored chords.

Unfortunately, we expect that we will again have difficulty estimating the chord length distribution at the point τ^-. With that in mind, it might well be worth-while to introduce some extra artificial censoring, similar to what we did for Laslett's problem.

3.5 A simulation

In this section we compare our estimator to a simple alternative. Recall our definition of the chord length distribution in the direction $e = (1,0)$. $L(x) = E|W(x)|/E|W|$ where $W(x) = \{(s,t) \in W = \mathcal{W} \cap B : |\xi_e(s,t)| \leq x\}$. It is clear that this definition does not depend on B. From observing $\mathcal{W} \cap B$ we cannot in general infer $|W(x)|$.

Define the line segment $C = \{te : -1 \leq t \leq 1\}$ which, since $e = (1, 0)$ is just the horizontal line segment of length 2 though the origin. Let $xC = \{te : -x \leq t \leq x\}$. Finally, we define 'Minkowski' subtraction of two sets A_1 and A_2 as $A_1 \ominus A_2 = (A_1^c \oplus A_2)^c$. We shall consider $B \ominus xC = [x, \tau - x] \times [0, 1]$, for positive $x \leq \tau$. This set is called the erosion of B by xC.

Note that $W(x) \cap (B \ominus xC) = \{(s, t) \in \mathcal{W} \cap (B \ominus xC) : |\xi_e(s, t)| \leq x\}$ and that $|W(x) \cap (B \ominus xC)|$ can be inferred from observing W. It is natural to define the following 'reduced sample' estimator based on an i.i.d. sample of n copies $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n$ of $\mathcal{W}$ all observed through B

$$\tilde{L}_n(x) = \frac{\sum_i |W_i(x) \cap (B \ominus xC)|}{\sum_i |W_i \cap (B \ominus xC)|}.$$

This estimator is 'ratio unbiased', meaning that it is the ratio of unbiased estimators of the numerator and denominator appearing in the definition of $L(x)$. It is clearly (uniformly, almost surely) consistent.

In Figure 3.2(a) we see a realization of a Boolean model Ξ observed in the unit square, $B = [0, 1]^2$. The grains are circular discs with a constant radius of 0.04. The underlying intensity of the Poisson process is 33.0.

We take the random set $\mathcal{W}$ of interest to be the closure of the void of the Boolean model $\mathcal{W} = \overline{\Xi^c}$. The reason for doing this, is that we know how to obtain a closed form expression for the chord length distribution of $\overline{\Xi^c}$. We use the so called linear contact distribution function of Ξ (Stoyan et al., 1987)

$$H(x) = 1 - \exp(-2\mu R x),$$

where $R = 0.04$ is the radius and $\mu = 33$ is the intensity. The following equation relates the linear contact distribution function H of Ξ and the chord length distribution L of $\overline{\Xi}^c$ (Matheron, 1975, p. 53)

$$L(r) = H(r) - r\frac{\mathrm{d}H(r)}{\mathrm{d}r}.$$

We have generated 100 realizations $\Xi_1, \Xi_2, \ldots, \Xi_{100}$ of our Boolean model. On the basis of, say, Ξ_k we have derived the 'NPMLE' estimator $\hat{L}_1^{(k)}$ and the reduced sample estimator $\tilde{L}_1^{(k)}$. The subscript "1" indicates that the estimator is based on a single observation. In Figure 3.2(b) we show $\hat{L}_1^{(k)}$ and $\tilde{L}_1^{(k)}$ for some value of k. In Figure 3.3 we see a comparison of the estimated root means square errors of the two estimators. That is, the figure shows

$$\left(\sum_{k=1}^{100}(\tilde{L}_1^{(k)}(x) - L(x))^2/100\right)^{1/2} \quad \text{and} \quad \left(\sum_{k=1}^{100}(\hat{L}_1^{(k)}(x) - L(x))^2/100\right)^{1/2}.$$

The latter seems uniformly more efficient. We should point out that comparing our estimator to this very simple reduced sample estimator is not the last word. One could surely think of more sophisticated alternatives (such as various edge correction estimators) that would present a sterner test.

3.6 The linear contact distribution

Recall that $\mathcal{W}$ is a stationary random closed set in $\mathbb{R}^2$ and e is a unit vector in $\mathbb{R}^2$. The *contact segment* through a point outside $\mathcal{W}$ in the direction e, is the longest line-segment starting at that point that does not hit $\mathcal{W}$. The distribution of the length of this contact segment is called the linear contact distribution. Let $\chi_e(s,t)$ denote the contact segment originating at (s,t) and let $|\chi_e(s,t)$ denote its length. The situation is illustrated in figure 3.6.

We now formally define the linear contact distribution function. Let B be any compact set of positive area and set $W = \mathcal{W} \cap B$ and $W^c = B \setminus W$. Now define

$$V(x) = \{(s,t) \in W^c : |\chi_e(s,t)| \leq x\}. \tag{3.18}$$

The linear contact distribution function is

$$H_e(x) = \frac{E|V(x)|}{E|B \setminus W|} \tag{3.19}$$

The expectations are with respect to the distribution of $\mathcal{W}$. This definition does not depend on B. Again, for simplicity we take $e = (1,0)$ and drop it from our notation. Also, we take $B = [0,\tau] \times [0,1]$. We now proceed in a similar way as before. In fact, as we are running out of useful letters we shall now *re-define* the functions $y(s,t)$ and $\delta(s,t)$ and the sets $A(s,t)$. Define

$$y(s,t) = |\chi(s,t) \cap B| \tag{3.20}$$

$$\delta(s,t) = \#\text{ endpoints of } \chi(s,t) \text{ outside of } B \tag{3.21}$$

to indicate a contact segment's length and censoring type. Previously, these From observation of W we ascertain for all positive y and $\delta = 0$ or 1

$$A(y,\delta) = \{(s,t) \in W^c : y(s,t) \leq y,\ \delta(s,t) = \delta\}. \tag{3.22}$$

We define

$$G(y,\delta) = \frac{E|A(y,\delta)|}{E|W^c|}, \tag{3.23}$$

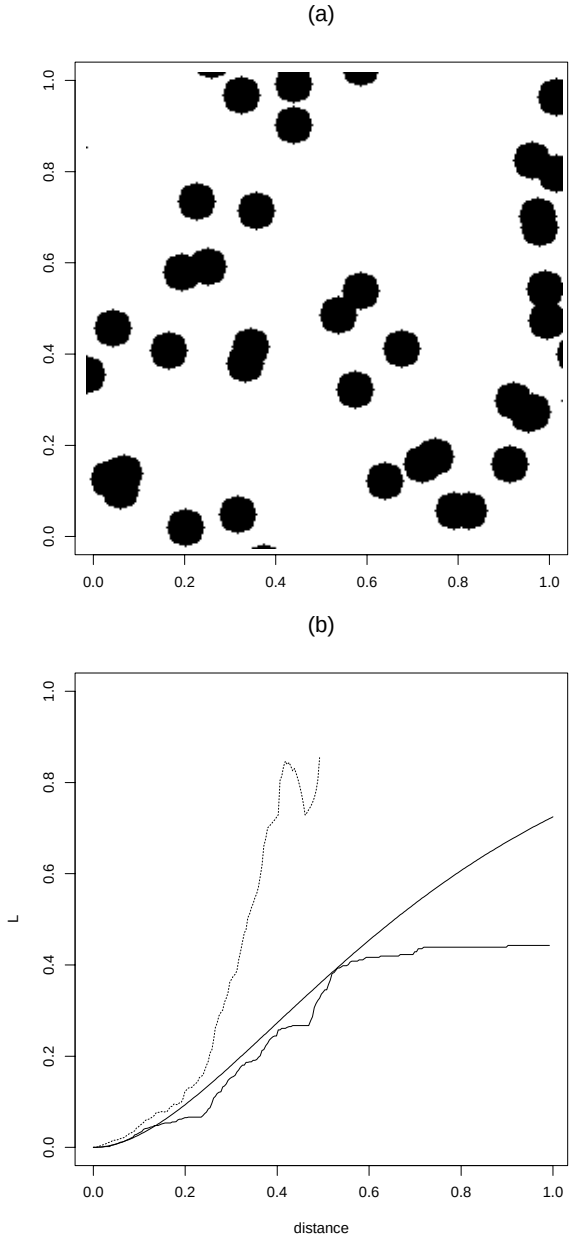

Figure 3.2: (a) Realization of a Boolean model; (b) Corresponding "NPMLE"-estimator $\hat{L}_1$ (solid, ragged line), Reduced sample estimator $\tilde{L}_1$ (dotted line) and the estimand L (solid, smooth line).

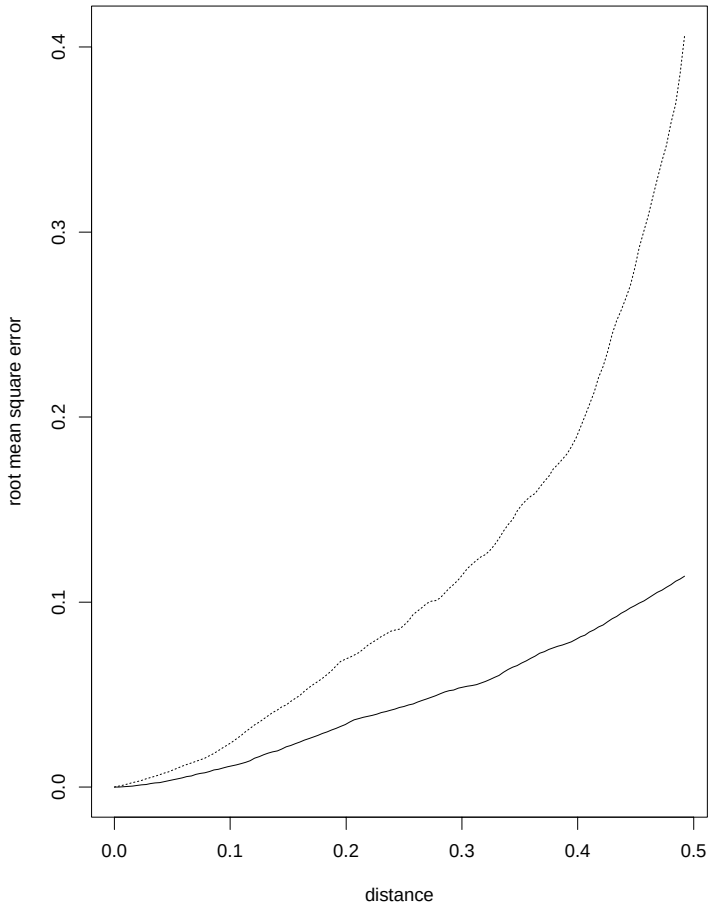

Figure 3.3: Root mean square error comparison of the NPMLE-estimator (solid line) and reduced sample estimator (dotted line).

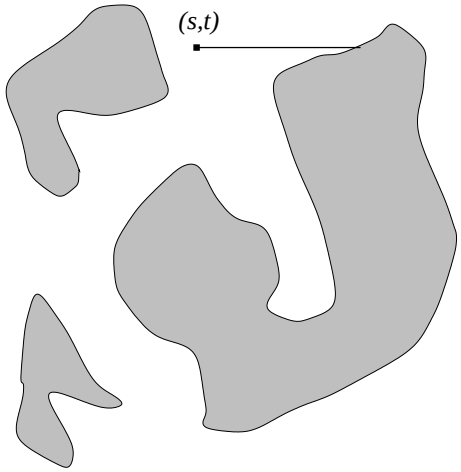

Figure 3.4: The contact segment $\chi_{(1,0)}(s,t)$. The shaded region is (part of) the random set $\mathcal{W}$.

where the expectation is with respect to the distribution of $\mathcal{W}$. This definition should not be confused with (3.7). The sets $A(y,\delta)$ have a different meaning here.

Note that if (S,T) is a uniform point in W^c then H is the distribution function of $|\chi(S,T)|$ and G is the joint distribution function of $y(S,T) = |\chi(S,T) \cap B|$ and $\delta(S,T)$, which is the number (0 or 1) of endpoints of $\chi(S,T)$ that fall outside of B.

Now consider the following related experiment. Let (S,T) be uniformly distributed on B and let X be independently distributed according to H. Define χ to be the line-segment between (S,T) and $(S+X,T)$. Define $Y = |\chi \cap B| = |[S,S+X] \cap [0,\tau]|$ and let Δ be the number of endpoints of χ outside of B. This set-up is the familiar random censorship model. The maximum likelihood estimator of H from observing n i.i.d. copies (Y_i,Δ_i) of (Y,Δ) is the well-known Kaplan–Meier estimator. The Kaplan–Meier estimator is a function of the empirical distribution of the (Y_i,Δ_i). In the original problem—estimation the linear contact distribution H from observing i.i.d. copies W_i of W—we replace these empirical distribution functions with empirical ratios of areas of the form $\sum_i |A_i(y,\delta)| / \sum_i |B \setminus W_i|$. The resulting estimator was introduced and studied in Hansen et al. (1996). There it is also compared to a reduced sample estimator.

There is a *very* useful connection between the chord length distribution and the contact distribution given in Stoyan et al. (1987 p. 180). Let $\overline{\mathcal{W}^c}$ denote the closure of the complement of $\mathcal{W}$, which is again a stationary

random closed set. Let L_0 denote the unweighted chord length distribution function of $\overline{\mathcal{W}^c}$ and let H be the linear contact distribution function of $\mathcal{W}$ itself. Then

$$H(x) = \int_0^x (1 - L_0(t))\mathrm{d}t/m \tag{3.24}$$

where m is the mean typical chord length

$$m = \int x \mathrm{d}L_0(x).$$

First of all we note from (3.24) that H is absolutely continuous with density $h(x) = (1 - L_0(x))/m$. We easily deduce

$$L_0(x) = 1 - h(x)/h(0). \tag{3.25}$$

Now from (3.24) it also follows that h is *monotone* or—equivalently—that H is concave. Non-parametric maximum likelihood estimation of a concave distribution function from right-censored observations is studied in Huang and Zhang (1994) and Huang and Wellner (1995). Using the NPMLE derived in these papers instead of the Kaplan–Meier estimator (and replacing empirical distribution functions by ratios of areas) we can improve the estimator of Hansen et al. (1996). Replacing the Kaplan–Meier estimator with its least concave majorant will also improve the estimator.

To estimate the chord length distribution of $\mathcal{W}$ we can first estimate the density of the contact distribution of $\overline{W^c}$ and then apply (3.25). However, it seems that this estimator will not improve the chord length estimator we obtained earlier, as density estimation is generally awkward. It is more fruitful to work in the other direction: To estimate the contact distribution of $\mathcal{W}$, first estimate the chord length distribution of $\overline{\mathcal{W}^c}$ and then use (3.24).

Chapter 4

The Bombing Model

4.1 Introduction

Many images found in microscopy, materials science and biology can be described by means of a random set. Perhaps the best known model is the *Boolean model* formalizing a configuration of independent, randomly placed particles. It is formed by replacing the points of a Poisson process by random closed sets. The points of the Poisson process are sometimes called the *germs*, the associated random sets the *grains* or *particles*. If the process is defined on the plane and the grains are discs, then it is also know as the 'bombing model'. Notwithstanding the strong independence assumptions, inference for Boolean models is far from trivial (Molchanov, 1997). The difficulty lies in the occlusion arising from the fact that only the union of all particles is observed, not the individual germ–grain pairs.

One may distinguish between two types of parameters of a Boolean model: aggregate (or macroscopic) parameters and individual (or microscopic) ones (Molchanov, 1997). Typical examples of aggregate parameters are the area fraction and the set-covariance, which can be easily estimated by their observed image counterparts. The resulting estimators are unbiased, and expressions for the variance can be obtained from Robbins' theorem (e.g., Stoyan et al. 1987). Under mild ergodicity assumptions they are strongly consistent (Molchanov, 1997) as the observation window expands to the entire plane. Aggregate functionals such as the contact distribution and pair correlation function are of interest when fitting the Boolean model to a data image. Usually, estimation is hampered by edge effects, but minus sampling ideas (Ripley, 1988, Stoyan et al. 1987) are generally applicable as are Horvitz–Thompson style estimators including the Kaplan–Meier (Baddeley and Gill, 1995) and Hanisch style estimators (Hanisch, 1984). Unbiased-

ness follows from the Campbell–Mecke theorem (Stoyan et al. 1987) and asymptotic results are available (Molchanov, 1997).

Individual parameters, including the intensity of the germ process, are much harder to estimate. Minimum contrast methods (Dupač, 1980, Serra, 1982) for the intensity are based on minimizing the distance between an estimated aggregate parameter (e.g. the contact distribution) and an approximation expressed in terms of the intensity. Some asymptotic results are available, but the expressions for the asymptotic variance are too complicated to be useful in practice. An alternative is the method of moments based on coverage fraction, mean area and boundary length, and the Euler–Poincaré characteristic. This method is computationally easy but leads to a biased estimator (Weil, 1988). In the tangent point approach, the Euler–Poincaré characteristic is replaced by the specific connectivity number, resulting in easier asymptotics (Molchanov and Stoyan, 1994). Further details can be found in (Molchanov, 1977) and the references therein.

In this chapter we shall take a likelihood based approach using Monte Carlo methods to perform the necessary computations. To do so, we need to be able to sample from the conditional distribution of a Boolean model given an observation of the union of its particles. In the next section we show that this distribution is straightforward if the grains are balls. However, due to an intractable normalizing constant, direct sampling is not possible. In section 4.3, we use 'coupling from the past' (Propp and Wilson, 1996) to design an algorithm yielding *exact* or *perfect* samples. We discussed coupling from the past in section 1.5.1 of this thesis. Also, we provide a simple modification of our CFTP algorithm which speeds it up considerably.

Section 4.4 is devoted to two approaches to maximum likelihood estimation through simulation. The first method is based on a Monte Carlo approximation of the likelihood ratio with respect to a fixed parameter value (Geyer, 1998). The other approach is a stochastic version of the EM-algorithm (StEM algorithm, cf. section 1.5.2) (Celeux and Diebolt, 1986). This iterative algorithm works as follows. We start by choosing some initial parameter value. At each E-step one or more samples are drawn from the conditional distribution, under the current parameter value, of the complete data given the observed data. These samples are used to estimate the conditional expectation of the likelihood, given the data. This estimated likelihood is then maximized in the M-step to obtain a new parameter value. Repeating these steps a sequence of parameter values is obtained. In fact, under certain conditions, this sequence is an ergodic Markov chain on the parameter space. The stochastic EM (StEM) estimator is defined to be a sample, or an average of samples, from the stationary distribution of this Markov chain. See (Nielsen, 1997a and b) for asymptotic results. A drawback of the algorithm is

that it is not clear when the chain has reached stationarity. However, we can make the algorithm 'perfect' by applying coupling from the past to obtain perfect samples from the 'StEM' chain's stationary distribution. This issue was explored in section 1.5.3.

In section 4.5 we present results of a modest simulation experiment we conducted to compare the two methods.

In section 4.6 we note that our simulation algorithm can be applied to sample from any point process for which the so-called Papangelou conditional intensity is bounded away from zero. In this sense, our work is complementary to an algorithm of Kendall and Møller (1999) to sample from point processes for which the Papangelou conditional intensity is bounded away from infinity. We also present a generalization of the algorithm of Kendall and Møller making it faster.

Between them, the two algorithms will enable us to sample from a very wide variety of point processes. Especially sampling from Boolean models under certain constraints could be of considerable practical use. A typical example comes from the oil industry (Chessa, 1995), where a Boolean model is employed to represent the reservoir geometry. If it is known from test drilling or geological surveying that the reservoir has certain characteristics, simulation studies into its further properties then amount to sampling from a Boolean model conditional on these characteristics.

4.2 The conditional Boolean model

Suppose a realization of a Boolean model is observed with the aim to estimate the intensity of the underlying point process of germs. Since the conditional distribution of the germs given the observed union of particles involves an intractable normalizing constant depending on the intensity parameter of interest, direct maximum likelihood estimation seems very hard. However, in certain situations, it is possible to sample from the conditional distribution, so that Monte Carlo-based maximum likelihood estimation is a viable alternative.

In this chapter, we focus our attention on the case where the particles are (random) balls. Then, the location of a germ is identified upon observation of any part of the boundary of its associated ball. The remaining, not directly identifiable germs turn out to be distributed as a Poisson process conditioned to satisfy a coverage condition. The following makes this claim more precise.

Definition 4.1. *A* Boolean model *of balls is defined as follows. Let X be a stationary Poisson process with intensity $\lambda > 0$ on $\mathbb{R}^d$, and $B = B(0,1)$ the*

d-dimensional closed unit ball centered at the origin. Then, writing $A \oplus B = \{a + b : a \in A,\ b \in B\}$*, the random set*

$$\mathcal{B}(X) = \cup_{x_i \in X}(x_i \oplus rB)$$

is a Boolean model of balls with radius $r > 0$ *on* $\mathbb{R}^d$*.*

For $d = 2$, the process of Definition 4.1 is sometimes referred to as the *bombing model*. The underlying points X are called the *germs*. The set B is called the *primary grain*. More general Boolean models are obtained by letting the germs be scattered according to a non-stationary Poisson process or letting the grains be arbitrary random closed sets.

Our goal in this chapter is to perform statistical inference for the intensity parameter λ based on an observation of $\mathcal{B}(X)$ in a compact sampling window W with non-empty interior. Because B is symmetric about the origin we have for all $x \in \mathbb{R}^2$

$$x \in W \oplus B \Leftrightarrow x \oplus B \cap W \neq \emptyset,$$

and hence

$$\mathcal{B}(X \cap (W \oplus rB)) \cap W = \mathcal{B}(X) \cap W$$

So inference may be based on the conditional distribution of the germ process $X \cap (W \oplus rB)$ given the data $Y = \mathcal{B}(X) \cap W$.

The unconditional distribution of $X \cap (W \oplus rB)$ is absolutely continuous with respect to that of a unit rate Poisson process on $(W \oplus rB)$ with Radon-Nikodym derivative at a configuration $\mathbf{x}$ given by $\exp\{(1 - \lambda)|W \oplus rB|\}\lambda^{n(\mathbf{x})}$. This density is defined on the space of all finite, unordered sets $\mathbf{x} = \{x_1, \ldots, x_n\}$ of points in $W \oplus rB$. Here $|W \oplus rB|$ denotes the area of the set $W \oplus rB$, and $n(\mathbf{x})$ is the number of points in configuration $\mathbf{x}$.

We return to the conditional distribution of $X \cap W \oplus rB$ given $Y = \mathcal{B}(X) \cap W$. Since the primary grains are balls, the location of a germ is identified whenever a part of its associated grain's boundary is exposed. Therefore, the conditional distribution of $X \cap (W \oplus rB)$ can be decomposed into a deterministic 'exposed boundary' part X^b and a stochastic 'interior' X^i of germs that cannot be uniquely identified. Indeed we write $X \cap (W \oplus rB) = X^i \cup X^b$. The conditional distribution given Y of the exposed boundary part X^b is of course degenerate at some configuration $\mathbf{x}^b$. The distribution of the unobserved germs X^i is only slightly more complicated.

Define

$$\mathcal{C} = Y \setminus \mathcal{B}(X^b) \tag{4.1}$$

$$\mathcal{D} = \{y \in W \oplus rB : (y \oplus rB) \cap W \subseteq Y\}. \tag{4.2}$$

In words, $\mathcal{C}$ is the part of Y which is not covered by exposed grains, and must therefore be covered by the interior grains. The set $\mathcal{D}$ describes the locations where interior points may fall such that their associated grains are not outside of Y.

Lemma 4.1. *The conditional distribution of X^i given Y is that of a Poisson process on $\mathcal{D}$ with intensity λ, conditional on coverage of $\mathcal{C}$.*

We write π_λ for this conditional distribution of X^i given Y. Its density is

$$f_\lambda(\mathbf{x}^i) = \frac{\mathbf{1}_{\{\mathbf{x}^i \text{covers } \mathcal{C}\}} e^{(1-\lambda)|\mathcal{D}|}\, \lambda^{n(\mathbf{x}^i)}}{P_\lambda(\mathcal{C} \text{ is covered})} \tag{4.3}$$

where $P_\lambda(\cdot)$ denotes the distribution of a Poisson process with intensity λ on $\mathcal{D}$.

Proof The Janossy density (Daley and Vere–Jones, 1988) at any $\mathbf{x}^i \cup \mathbf{x}^b$ satisfying $\mathcal{B}(\mathbf{x}^i \cup \mathbf{x}^b) \cap W = Y$ is given by

$$\begin{aligned} \frac{e^{-|W \oplus rB|}}{(n(\mathbf{x}^i)+n(\mathbf{x}^b))!} p(\mathbf{x}^i \cup \mathbf{x}^b) &= \frac{e^{-|W \oplus rB|}}{(n(\mathbf{x}^i)+n(\mathbf{x}^b))!} e^{(1-\lambda)|W \oplus rB|} \lambda^{n(\mathbf{x}^i)+n(\mathbf{x}^b)} \\ &= \frac{e^{-\lambda|W \oplus rB|}}{(n(\mathbf{x}^i)+n(\mathbf{x}^b))!} \lambda^{n(\mathbf{x}^i)}\, \lambda^{n(\mathbf{x}^b)}. \end{aligned}$$

Hence, the Radon–Nikodym density is proportional to $\lambda^{n(\mathbf{x}_i)}$ for any $\mathbf{x}_i \subseteq \mathcal{D}$ covering $\mathcal{C}$. Since $\mathbf{x}_i \subseteq \mathcal{D}$, it is natural to replace the reference distribution of a unit rate Poisson process on $W \oplus rB$ by that of a unit rate Poisson process on $\mathcal{D}$. Noting that the probability of covering $\mathcal{C}$ by balls centered at the points of a Poisson process on $\mathcal{D}$ with intensity λ equals

$$P_\lambda(\mathcal{C} \text{ is covered}) = \sum_{n=0}^{\infty} \frac{e^{-\lambda|\mathcal{D}|}}{n!} \lambda^n \int_{\mathcal{D}} \cdots \int_{\mathcal{D}} \mathbf{1}_{\{\{e_1,\ldots,e_n\} \text{covers } \mathcal{C}\}}(e) \mathrm{d}e_1 \ldots \mathrm{d}e_n$$

it follows that the normalizing constant is $\frac{e^{(1-\lambda)|\mathcal{D}|}}{P_\lambda(\mathcal{C} \text{ is covered})}$. □

The covering probability $P_\lambda(\mathcal{C}$ is covered) is typically impossible to compute, see Hall (1988) on covering problems. Because of this, direct sampling from π_λ is also impossible. An alternative is of course rejection sampling: simply generating independent Poisson processes of intensity λ until one of them satisfies the covering condition. Unfortunately, depending on λ and $\mathcal{C}$ the probability of success, $P_\lambda(\mathcal{C}$ is covered), may be too small.

It is important for us to note that $P_\lambda(\mathcal{C}$ is covered) is strictly increasing in λ. This follows from the fact that a Poisson process of intensity $\kappa > \lambda$

can be written as the independent superposition of Poisson processes with intensity λ and $\kappa - \lambda$ respectively. So we see that if we choose κ large enough we can quite easily obtain a sample from π_κ. In the next section we use this fact to construct a method to sample from π_λ by thinning a sample from π_κ for an arbitrary $\kappa > \lambda$.

4.3 Coupling from the past for the conditional Boolean model

Recall that an ordinary stationary Poisson point process of intensity $\lambda > 0$ on a compact set S arises as the stationary distribution of a spatial birth-and-death process (Preston 1977). We start with any finite collection of points on S. Then each point is deleted after an exponential lifetime with rate 1. Also, after exponential waiting times with mean $1/\lambda$ new points are added which are uniformly distributed on S. The sequence of point configurations thus obtained converges weakly to a Poisson point process on S of intensity λ. Lantuéjoul (1997) demonstrates that similarly we can obtain a Poisson process point process, conditioned on a certain event $\mathcal{E}$ of positive probability, as the limit of a birth and death process. Informally, if whenever a point is added or deleted according to the scheme described above, we make sure never to enforce a transition violating $\mathcal{E}$, the stationary distribution of the resulting process is the conditional version of the Poisson process.

In our case, $S = \mathcal{D}$ and $\mathcal{E}$ is the event that $\mathcal{C}$ is covered. Let $X(t)$ be a birth and death process with birth rate λ and individual death rate 1. A birth is understood to be the addition of a uniform point on $\mathcal{D}$. Let $X(t^-)$ denote the configuration at time t^- just before t. If a point z_i is born at time t it is added: $X(t) = X(t^-) \cup \{x\}$. If $x \in X(t^-)$ dies at time t it is removed only if that does not cause part of $\mathcal{C}$ to become uncovered, that is if

$$B(x, r) \cap \mathcal{C} \subseteq \mathcal{B}(X(t^-) \setminus \{x\}) \cap \mathcal{C}. \tag{4.4}$$

If x cannot be removed, then it is granted an extra exponential lifetime after which its removal is reconsidered. Following the terminology in (Kendall and Thönnes, 1999) x is then called *perpetuated*.

Following Lantuéjoul (1997) it would not be difficult to show that the stationary distribution of $X(t)$ is indeed π_λ. However, $X(t)$ is *not* the birth and death process we shall use to sample from π_λ. We want to apply coupling from the past (section 1.5.1) to obtain perfect samples from π_λ but $X(t)$ does not appear amenable to this approach.

The difficulty is that for perfect sampling from the stationary distribution of a Markov chain it is convenient if the state-space admits a partial order

and there exist maximal and minimal elements majorizing and minorizing all other elements with respect to this partial order. The state space of $X(t)$ consists of all finite collections of points on $\mathcal{D}$. The obvious (partial) ordering of this space is the inclusion ordering, but then there is certainly no maximal state, because $\mathcal{D}$ itself is infinite.

We shall construct a different birth-and-death process on a state space consisting of all subsets of a finite (but random) set of points which also converges to π_λ. We shall be able to apply coupling from the past to this process.

It should be remarked that for the special case where $\mathcal{E}$ is the event that a certain finite collection of points is covered, a perfect sampling algorithm is available due to Kendall and Thönnes (1998). Unfortunately, their method does not seem to extend to our situation: coverage of a uncountable set. However, some of the ideas in (Kendall and Thönnes, 1998) play a role in our construction also.

As noted previously, rejection sampling from (4.3) is possible for large intensity parameters. Using this observation, the first step in our algorithm to sample from a given π_λ is to choose a $\kappa > \lambda$ and to generate a sample, say $D = \{z_1, z_2, \ldots, z_n\} \subset \mathcal{D}$, from π_κ. D will serve as maximal state. All configurations obtained when running the birth-and-death processes will be subsets of this finite set D. Now D is thinned, independently retaining each point with probability λ/κ.

The key result is the following.

Proposition 4.1. *Let D have distribution π_κ and define E_0 to be the independent thinning of D with retention probability λ/κ. Conditionally on coverage of $\mathcal{C}$, E_0 is distributed according to π_λ (cf. (4.3)).*

Proof Note that the conditional distribution of E_0 given coverage of $\mathcal{C}$ is concentrated almost surely on configurations of points of $\mathcal{D}$. For any such configuration $\mathbf{x}$, the Janossy density is given by

$$\begin{aligned} & j_n(\mathbf{x}) \\ &= \sum_{n=0}^{\infty} \frac{e^{-|\mathcal{D}|}}{n!} \int \ldots \int_{\mathcal{D}^n} f_\kappa(\mathbf{x} \cup \{y_1, \ldots, y_n\}) \left(\frac{\lambda}{\kappa}\right)^{n(\mathbf{x})} \left(1 - \frac{\lambda}{\kappa}\right)^n \mathrm{d}y_1 \ldots \mathrm{d}y_n \\ &= \sum_{n=0}^{\infty} \frac{e^{-|\mathcal{D}|}}{n!} \alpha_\kappa e^{(1-\kappa)|\mathcal{D}|} \lambda^{n(\mathbf{x})} \left[\kappa(1 - \frac{\lambda}{\kappa})\right]^n |\mathcal{D}|^n = \alpha_\kappa \lambda^{n(\mathbf{x})} e^{-\lambda|\mathcal{D}|}, \end{aligned}$$

where $\alpha_\kappa^{-1} = P_\kappa(\mathcal{C}$ is covered) is the probability that balls of radius r around the points of a Poisson process with intensity κ cover $\mathcal{C}$. The claim follows upon normalization and comparison to (4.3). □

We shall construct a coupling-from-the-past (CFTP) algorithm that selects a configuration Z between E_0 and D ($E_0 \subseteq Z \subseteq D$) such that Z is distributed according to π_λ.

First we construct a spatial birth-and-death process $E = \{E(t),\ t \geq 0\}$, such that $E(t) \subseteq D = \{z_1, z_2, \ldots, z_n\}$ for all t. We set $E(0) = E_0$. The dynamics of the process are such that each point is removed from $E(\cdot)$ after an exponentially distributed lifetime with mean 1. Whenever a point is removed, it is added again after an exponentially distributed waiting time of mean $(\kappa - \lambda)/lambda$. All life– and waiting times are independent of each other and everything else. Points $z_i \in E(0)$ are treated as if they were added at time 0, those $z_i \notin E(0)$ as removed at time 0.

Lemma 4.2. *Conditionally on D, the spatial birth-and-death process $E(t)$, $t \geq 0$, is irreducible, homogeneous, and positive recurrent. Moreover, $E(t)$ is in equilibrium and time-reversible.*

In particular $E(t)$ is distributed as $E(0) = E_0$ for all $t \leq 0$.

Proof As $E(\cdot)$ is defined conditionally on $D_0 = \{z_1, z_2, \ldots, z_n\}$, at any time there are only a finite number of points. Clearly, the transition rates do not change in time, hence $E(\cdot)$ is homogeneous. Moreover, any state $\mathbf{x} \subseteq D$ can be reached from any other state $\mathbf{x}'$ (say) by successively deleting the points in $\mathbf{x}'$ followed by addition of the points in $\mathbf{x}$. Therefore, the birth-and-death process is well-defined and possesses a stationary distribution (Parzen, 1962, Chapter 7). Moreover, since the state space is finite, not all stationary probabilities can be zero, hence they are all positive and the $E(\cdot)$ is positive recurrent.

The process $E(\cdot)$ is in equilibrium and time reversible because it satisfies 'detailed balance' with respect to the distribution (given D) of $E(0) = E_0$

$$\frac{\lambda}{\kappa - \lambda} P(E_0 = \mathbf{x}|D) = P(E_0 = \mathbf{x} \cup \{z_i\}|D),$$

where $z_i \notin \mathbf{x} \subseteq D$. □

We now define a spatial birth-and-death process $Z(t)$, $t \geq 0$, as a conditional version of $E(\cdot)$ in the spirit of Lantuéjoul (1997). The equilibrium distribution of $Z(\cdot)$ will be π_λ.

Start with $Z(0)$ being any subset of D_0. The dynamics of the process are such that a point z_i *might be* removed from $Z(\cdot)$ after an exponentially distributed lifetime of mean 1. The point z_i is *actually* removed only if

$$B(z_i, r) \cap \mathcal{C} \subseteq \mathcal{B}(Z(t^-) \setminus \{z_i\}) \cap \mathcal{C}. \tag{4.5}$$

If z_i cannot be removed, then it is granted an extra exponential lifetime with mean 1 after which its removal is reconsidered. Whenever a point *is* removed, it is added again after an exponentially distributed waiting time of mean $(\kappa - \lambda)/\lambda$. All life– and waiting times are independent of each other and everything else. Points $z_i \in Z(0)$ are treated as if they were added at time 0, those $z_i \notin Z(0)$ as removed at time 0.

Proposition 4.2. *Conditionally on D, the spatial birth-and-death process $Z(t)$, $t \geq 0$, is homogeneous, and has a single positive recurrent class consisting of those subsets $\mathbf{z}$ of D for which $\mathcal{B}(\mathbf{z})$ covers $\mathcal{C}$. It tends in distribution to π_λ as $t \to \infty$.*

Proof We work conditionally on $D = \{z_1, z_2, \ldots, z_n\}$. First note that almost surely $\mathcal{B}(Z(t))$ will cover $\mathcal{C}$ for some $t \geq 0$. The transition mechanism ensures that $\mathcal{B}(Z(s)) \supseteq \mathcal{C}$ for all $s \geq t$. Moreover, the class C of configurations $\mathbf{z} \subseteq D$ whose associated Boolean model covers $\mathcal{C}$ is irreducible, since any state $\mathbf{z}$ can be reached from any other state $\mathbf{z\prime}$ by successively adding all points of $D \setminus \mathbf{z\prime}$, then deleting those of $D \setminus \mathbf{z}$. Thus $Z(\cdot)$ is well-defined, with a stationary distribution concentrated on C (Parzen, 1962). Again, since the state space is finite, not all stationary probabilities can be zero, hence they are all positive and the C is positive recurrent.

Consider a move at time t of $Z(\cdot)$ from a configuration $\mathbf{x} \cup z_i \subseteq D$ to $\mathbf{x}$, for which both $\mathcal{B}(\mathbf{x})$ and $\mathcal{B}(\mathbf{x} \cup z_i)$ cover $\mathcal{C}$ and $z_i \notin \mathbf{x}$. The birth rate of $Z(\cdot)$ is the same as the birth rate of $E(\cdot)$. The death rate for both perpetuated and non-perpetuated points is 1, as it is for E, provided the coverage condition is not violated. Thus, the detailed balance conditions for $E(\cdot)$ and $Z(\cdot)$ coincide on C, from which it follows that $Z(t)$ tends to the distribution of $E(0)$ restricted to C as $t \to \infty$. Using lemma 4.1 the result follows. □

We shall now describe how to apply coupling from the past to obtain a sample from the stationary distribution of $Z(\cdot)$. We already have a maximal state D.

Fix a time $-T < 0$. By lemma 4.2, $E(\cdot)$ is time-reversible, hence can easily be extended backwards from $E(0)$ until time $-T$. $E(\cdot)$ on the interval $[-T, 0]$ will be the minimal state.

It is straightforward to define a coupled process $Z_{-T}(t)$, $-T \leq t \leq 0$ which has the same dynamics as $Z(\cdot)$ while $E(t) \subseteq Z(t) \subseteq D$. First set $Z_{-T}(-T) = E(-T)$. Now with each $z_i \in D$ associate a unit rate Poisson process Ξ_i on the set $\{-T \leq t \leq 0 : z_i \notin E(t)\}$. These Ξ_i will govern possible deaths of perpetuated points. At the (forward) birth in the process $E(\cdot)$ at time t of a point z_i, add z_i to $Z(t^-)$ (if not already present). At the (forward)

death in the process $E(\cdot)$ at time t of a point z_i, remove z_i from $Z(t^-)$ but only if (cf. 4.5)

$$B(z_i, r) \cap \mathcal{C} \subseteq \mathcal{B}(Z_{-T}(t^-) \setminus \{z_i\}) \cap \mathcal{C}. \tag{4.6}$$

If z_i cannot be removed at time t, then its removal is reconsidered at the next event time after t of the process Ξ_i.

Lemma 4.3. *$Z_{-T}(0)$ tends in distribution to π_λ as $T \to \infty$.*

Proof $Z_{-T}(\cdot)$ has the same dynamics and hence the same stochastic properties as the process $Z(\cdot)$ considered in the beginning of this section. Hence, this lemma follows directly from Proposition 4.2. □

To check if $Z_{-T}(0)$ has reached equilibrium—so to say—we need processes $L_{-T}(\cdot)$ and $U_{-T}(\cdot)$ such that (cf. Kendall and Møller (1999) formulas (3.5) to (3.7))

1. ('sandwiching') $E(t) \subseteq L_{-T}(t) \subseteq U_{-T}(t) \subseteq D_0$ for all $-T \leq t \leq 0$;
2. ('funneling') $L_{-T}(t) \subseteq L_{-S}(t) \subseteq U_{-S}(t) \subseteq U_{-T}(t)$, for all $-S \leq -T \leq t \leq 0$;
3. ('coalescing') if, for some s, $L_{-T}(s) = U_{-T}(s)$ then $L_{-T}(t) = U_{-T}(t)$, for all $t \geq s$.

Moreover, we need that $L_{-T}(0) = U_{-T}(0)$ almost surely for finite (preferably small) T. If the above inclusions hold and the latter event occurs then we can conclude that $L_{-T}(0) = U_{-T}(0)$ is a perfect sample from π_λ. This is roughly the content of Theorem 4.1 below.

Now the construction of the processes $L_{-T}(\cdot)$ and $U_{-T}(\cdot)$ is very similar to that of $Z_{-T}(\cdot)$. The only difference is that removal (if present) of a point z_i at time t from $L_{-T}(t^-)$ is enforced only if

$$B(z_i, r) \cap \mathcal{C} \subseteq \mathcal{B}(U_{-T}(t^-) \setminus \{z_i\}) \cap \mathcal{C}. \tag{4.7}$$

and the same removal from $U_{-T}(t^-)$ is enforced only if

$$B(z_i, r) \cap \mathcal{C} \subseteq \mathcal{B}(L_{-T}(t^-) \setminus \{z_i\}) \cap \mathcal{C}. \tag{4.8}$$

The—at first perhaps mystifying—fact that these processes pay attention to each other is necessary to ensure the above inclusions. This 'cross-over' trick has also been applied by Häggström and Nelander (1998) and Kendall (1997).

Summarizing, we propose the following algorithm.

Algorithm 4.1. *Generate a random sample D from π_κ for some $\kappa > \lambda$ and delete each point independently with probability $1-(\lambda/\kappa)$ to obtain $E(0)$. Set $T = 1$ and write $[\frac{T}{2}]$ for the integer part of $\frac{T}{2}$.*

- *extend $E(\cdot)$ backwards on $[-T, -[\frac{T}{2}])$ with birth rate $\lambda/(\kappa-\lambda)$ and death rate 1;*
- *extend independent unit rate Poisson processes Ξ_i backwards on $\{-T \leq t < [\frac{T}{2}] : E_i(t) = 0\}$;*
- *set $L_{-T}(-T) = E(-T)$ and $U_{-T}(-T) = D$;*
- *at a birth transition $E(t^+) = E(t) \cup \{z_i\}$, add z_i to $L_{-T}(t)$ and $U_{-T}(t)$.*
- *at a death transition $E(t^+) = E(t) \setminus \{z_i\}$ or an event time of Ξ_i*
  - *delete (if present) z_i from $L_{-T}(t)$ provided that does not cause the Boolean model associated with $U_{-T}(t)$ to uncover part of $\mathcal{C}$, i.e. z_i may be deleted only if (4.8) holds;*
  - *delete (if present) z_i from $U_{-T}(t)$ provided that does not cause the Boolean model associated with $L_{-T}(t)$ to uncover part of $\mathcal{C}$, i.e. z_i may be deleted only if (4.7) holds;*
- *if $L_{-T}(0) = U_{-T}(0)$ exit; otherwise double T and repeat.*

Algorithm 4.1 is designed in so that the inclusion relations 1, 2 and 3 hold.

Lemma 4.4. *The processes $E(t)$, $L_{-T}(t)$, $U_{-T}(t)$ ($t \leq 0$) and D satisfy the inclusion relations 1, 2 and 3.*

Proof By definition, $E(-T) - L_{-T}(-T) \subseteq U_{-T}(-T) = D$, hence the sandwiching property holds for $t = -T$. Also D is a supset of all other sets. Since births in $E(\cdot)$ are reciprocated in $L_{-T}(\cdot)$ and $U_{-T}(\cdot)$, the inclusion relationship is preserved under birth transitions. Next, consider the death at some time $t \in [-T, 0]$, say $E(t^+) = E(t) \setminus \{z_i\}$ or an event time t of Ξ_i. Suppose that $E(t) \subseteq L_{-T}(t) \subseteq U_{-T}(t)$. Since Ξ_i is restricted to the set $\{t \leq 0 : E_i(t) = 0\}$, $E(t^+)$ is a subset of $L_{-T}(t^+)$ and $U_{-T}(t^+)$. Furthermore, if z_i dies in the upper process,

$$B(z_i, r) \cap \mathcal{C} \subseteq \mathcal{B}(L_{-T}(t) \setminus \{z_i\}) \cap C \subseteq \mathcal{B}(U_{-T}(t) \setminus \{z_i\}) \cap C$$

and consequently z_i also dies in the lower process.

Turning to the funneling property, we have to show that $L_{-T}(t) \subseteq L_{-S}(t)$ and that $U_{-S}(t) \subseteq U_{-T}(t)$. Now, by definition $E(-S) = L_{-S}(-S)$. Since the dynamics of algorithm 4.1 preserve inclusion, it follows that $L_{-T}(-T) = E(-T) \subseteq L_{-S}(-T)$ and, more generally, $L_{-T}(t) \subseteq L_{-S}(t)$ for any $t \geq -T$. Regarding the upper process, $U_{-T}(-T) = D \supseteq U_{-S}(-T)$. Applying once more the fact that the algorithm preserves the inclusion ordering yields $U_{-T}(t) \supseteq U_{-S}(t)$ for any $t \geq -T$.

Finally, suppose that the upper and lower processes meet at some time $s \leq 0$. Then, as they are coupled by the same E– and Ξ_i-processes, they proceed as one. □

We are now ready to state the main result of this section.

Theorem 4.1. *Algorithm 4.1 almost surely terminates in finite time; its output has density $f_\lambda(\cdot)$ (cf. (4.3)) with respect to a unit rate Boolean model on $\mathcal{D}$.*

Proof Note that $P(E(0) = D|D) = \left(\frac{\lambda}{\kappa}\right)^{n(D)}$, hence

$$P(E(0) = D) = E\left[\left(\frac{\lambda}{\kappa}\right)^{n(D)}\right]$$

where $n(D)$ is Poisson distributed with mean $\kappa|\mathcal{D}|$. Consequently, the event $\{E(0) = D\}$ has strictly positive probability. By lemma 4.1, $E(-T) = D$ will occur for some T almost surely. Hence, using Lemma 4.4, the algorithm terminates almost surely in finite time.

As stated by Lemma 4.3 $Z_{-T}(0)$ tends in distribution to π_λ as $T \to \infty$. Moreover, using the fact that Algorithm 4.1 preserves the inclusions ordering, $L_{-T}(t) \subseteq Z_{-T}(t) \subseteq U_{-T}(t)$ for all $-T \leq t \leq 0$.

Suppose T_0 is a (random) time such that $L_{-T_0}(0) = U_{-T_0}(0)$. It follows that $Z_{-T}(0) = L_{-T_0}(0) = U_{-T_0}(0)$ for all $-T \leq T_0$. Hence, $L_{-T_0}(0) = U_{-T_0}(0)$ is a sample from π_λ.

This completes the proof. □

To conclude this section we present a modification of algorithm 4.1, which not only speeds it up but also facilitates and even makes possible some interesting estimation methods. These estimation methods will be discussed in the following section.

First, we note that the algorithm typically takes a long time to terminate when the difference between the dominating pattern D and its thinning E_0 is large. For example, suppose we want to sample from π_1 while rejection

sampling only allows us to sample from π_{100} to obtain D. This is clearly a 'hard' problem. E_0 would on average contain only one hundredth of the points of D so the algorithm has many configurations between D and $E(0)$ to choose from—so to speak.

The idea is to first use algorithm 4.1 to reduce $D \sim \pi_{100}$ to a sample, say X_{99}, from π_{99}. Then use algorithm 4.1 again with $D = X_{99}$ to obtain a sample X_{98} from π_{98}. Repeating this, we obtain a *nested* sequence of samples $X_{99} \supseteq X_{98}, \cdots \supseteq X_1$ from $\pi_{99},\ \pi_{98}, \ldots, \pi_1$.

This will be much faster than trying to sample from π_1 straight away. In fact, for small steps of λ it might well happen that D and E_0 coincide, in which case there is clearly no need to run the algorithm. Also, it might happen that there is no pattern X with $E_0 \subseteq X \subset D$ which meets the coverage requirement. Again, there would be no need to run the algorithm at because the output must be D itself.

With an estimation method called 'perfect stochastic EM' (section 1.5.3) in mind, we now present an algorithm that produces a nested continuum of samples $\{X_\lambda \sim \pi_\lambda : \underline{\lambda} \leq \lambda \leq \overline{\lambda}\}$ for any two values $0 < \underline{\lambda} \leq \overline{\lambda}$. By 'nested' we mean that $X_\lambda \subset X_{\lambda'}$ if $\lambda \leq \lambda'$. The idea is to slowly thin an initial sample form $\pi_{\overline{\lambda}}$, removing one point at a time.

We suppose (without any loss of generality) that it is feasible to sample from $\pi_{\overline{\lambda}}$. Denote $\overline{\lambda} = \lambda(0)$. Let $X_{\lambda(0)} = \{x_1, \ldots, x_n\}$ be a sample from $\pi_{\lambda(0)}$. Now associate with each x_i a standard uniform random variable U_i^1; all the U_i^1 being independent. Define

$$E_\lambda^1 = \{x_i \in X_{\lambda(0)} : U_i^1 < \lambda/\lambda(0)\} \qquad \lambda \leq \lambda(1)$$

Note that for each λ, E_λ^1 is a thinning of $X_{\lambda(0)}$ with retention probability $\lambda/\lambda(0)$. Conceptually, we could—for every λ—run algorithm 4.1 with $D = X_{\lambda(0)}$ and $E_0 = E_\lambda^1$ to obtain a sample from π_λ.

Define $\lambda(1) = \lambda(0) \max_i U_i^1$. Then $E_\lambda^1 = X_{\lambda(0)}$ for all $\lambda(1) < \lambda \leq \lambda(0)$, while $E_{\lambda(1)}^1 = X_{\lambda(0)} \setminus \{x_i\}$ for some i.

Now set

$$X_\lambda = X_{\lambda(0)} \qquad \lambda(1) < \lambda \leq \lambda(0).$$

It is clear that these X_λ are distributed according to π_λ; application of algorithm 4.1 with $D = X_{\lambda(0)}$ and $E_0 = E_\lambda^1$ must output $X_{\lambda(0)}$ because $D = E_0$.

Next, run algorithm 4.1 with $D = X_{\lambda(0)}$ and $E_0 = E_{\lambda(1)}^1 = X_{\lambda(0)} \setminus \{x_i\}$ to obtain a sample $X_{\lambda(1)}$ from $\pi_{\lambda(1)}$. Note that when $E_{\lambda(1)}^1$ does not cover, the algorithm must output $X_{\lambda(0)}$ and hence need not be run.

Now repeat the entire procedure, starting with $X_{\lambda(1)}$ instead of $X_{\lambda(0)}$. We must associate new uniform random variables U_i^2 with the points of $X_{\lambda(1)}$ and

define new thinnings

$$E^2_\lambda = \{x_i \in X_{\lambda(1)} : U^2_i < \lambda/\lambda(1)\} \qquad \lambda \le \lambda(1)$$

Note that the distributions of E^1_λ and E^2_λ are the same for all $\lambda \le \lambda(1)$. Hence we are indeed starting afresh.

Repeating again and again in this manner, we obtain a sequence $\lambda(0) > \lambda(1) > \lambda(2) \dots$ and after n repetitions we have a nested continuum of samples $\{X_\lambda \sim \pi_\lambda : \lambda(n) \le \lambda \le \lambda(0) = \overline{\lambda}\}$.

Define N to be the smallest number such that $\lambda(N) < \underline{\lambda}$. It is not difficult to see that N is almost surely finite. Hence the above method provides, almost surely in finite time, nested samples from π_λ for all λ between $\underline{\lambda}$ and $\overline{\lambda}$.

4.4 Maximum likelihood

In general, direct maximum likelihood estimation of the intensity parameter in a Boolean model (Definition 4.1) seems very hard. In this section, we will describe two alternative techniques.

4.4.1 MCMC approach

Suppose a Boolean model $\mathcal{B}(X)$ of discs with radius r is observed though a non-empty compact set W. We write $Y = \mathcal{B}(X) \cap W$. The goal is to estimate the intensity parameter λ of the underlying Poisson germ process X. Although the likelihood of Y is known, it involves a normalizing constant which we cannot compute. A solution is to use the approach in (Geyer and Thompson, 1992, Geyer, 1998) and consider the likelihood ratio

$$\frac{p(Y|\lambda)}{p(Y|\kappa)} = e^{(\kappa-\lambda)|W \oplus rB|} \frac{c(\lambda|Y)}{c(\kappa|Y)}$$

with respect to a fixed parameter $\kappa > 0$. Note that the normalizing constant $c(\lambda|Y) = E\left[\lambda^{n(X)} \mathbf{1}_{\{\mathcal{B}(X)\cap W = Y\}}\right]$ where the expectation is taken with respect to a unit rate Poisson process. Note that

$$\frac{c(\lambda|Y)}{c(\kappa|Y)} = E_\kappa \left[\left(\frac{\lambda}{\kappa} \right)^{n(X)} \middle| \mathcal{B}(X) = Y \right].$$

Hence the log likelihood ratio can be rewritten as

$$\ell(\lambda) = \log \frac{p(Y|\lambda)}{p(Y|\kappa)} = (\kappa - \lambda)|W \oplus rB| + \log E_\kappa \left[\left(\frac{\lambda}{\kappa} \right)^{n(X)} \middle| \mathcal{B}(X) \cap W = Y \right]. \tag{4.9}$$

The expectation can be estimated by an average over independent realizations of the conditional distribution of X given $\mathcal{B}(X) \cap W = Y$ (under parameter κ) (see section 4.3).

The approximation works well if the reference value κ is not too far from the true value. To try to make sure of this, we could use a pilot estimate as κ. Alternatively, we can take a sequence of different parameter values $\lambda < \lambda_1 <, \ldots, < \lambda_n < \kappa$. Noting that

$$\ell(\lambda) = \log \frac{p(Y|\lambda)}{p(Y|\lambda_1)} + \log \frac{p(Y|\lambda_1)}{p(Y|\lambda_2)} + \cdots + \log \frac{p(Y|\lambda_{n-1})}{p(Y|\lambda_n)} + \log \frac{p(Y|\lambda_n)}{p(Y|\kappa)}$$

we need samples from $\pi_{\lambda(i)}$ for $i = 1, 2, \ldots, n$. We could (perhaps should) produce these samples independently, but we can also start with a sample from $\pi_{\lambda(n)}$ and then successively work our way down. When sampling from $\pi_{\lambda(i)}$ we can then take the sample from $\pi_{\lambda(i+1)}$ as dominating state D.

4.4.2 EM

The EM-algorithm (Dempster et al. 1977) is an iterative technique to approximately solve the likelihood equations for missing data problems by alternating expectation and maximization steps. In general, let X be the unobserved (complete) data, and Y the observed data obtained by a many-to-one mapping $Y = Y(X)$ and assume that X has a density $p(\mathbf{x}; \lambda)$ depending on a parameter $\lambda > 0$. Starting from any initial value $\lambda(0)$, construct a sequence $\lambda(0), \lambda(1), \ldots$ by repeating the following.

Expectation-step compute as the conditional expectation under $\lambda(k)$ of the complete data log-likelihood, given the observed data

$$E_{\lambda(k)} \left[\log p(\lambda; X) \mid Y \right]; \tag{4.10}$$

Maximization-step find $\lambda(k+1)$ by maximizing (4.10) with respect to λ.

In our set-up, $Y = \mathcal{B}(X) \cap W$ and $p(\mathbf{x}; \lambda)$ is the Poisson(λ) density. The (complete data) log-likelihood is linear in the sufficient statistic $n(X)$. Hence the expectation and maximization steps amount to computing the conditional expectation under $\lambda(k)$ of the complete data maximum likelihood estimator given the observed data, that is

$$\lambda(k+1) = E_{\lambda(k)} \left[\frac{n(X)}{|W \oplus rB|} \mid Y \right] = E_{\lambda(k)} \left[\frac{n(X^i) + n(X^b)}{|W \oplus rB|} \mid Y \right] \tag{4.11}$$

We cannot compute (4.11), as required by the EM-algorithm. However, since we can sample from the conditional distribution of X given Y, we can use Monte Carlo methods to approximate it. Suppose that at the k-th iteration step we use m samples to do so. The average of these estimates $E_{\lambda(k)}[n(X)/|W \oplus rB| \mid Y]$. This estimate is the next parameter value, $\lambda(k+1)$. At the next iteration we need a new set of m independent samples under $\lambda(k+1)$. This algorithm is known as the Monte Carlo EM or, if $m = 1$, the stochastic EM-algorithm (StEM). We discussed this algorithm in section 1.5.2.

We now specialize to StEM, i.e. the case $m = 1$. Usually, the algorithm is run for a great number of steps to allow it to approach equilibrium (burn-in). After that, the chain is continued for many more iterations and its steps are averaged to bring down the variance of the estimator. Since the discovery of CFTP, the arbitrary burn-in is unsatisfactory. Here we demonstrate how we can use CFTP to generate a StEM chain that is actually guaranteed to be in equilibrium. We need one mild assumption, namely that the true λ is known to be less than some finite $\overline{\lambda}$. The complete data maximum likelihood estimator is slightly modified to become $n(X)/|W \oplus rB| \wedge \overline{\lambda}$.

We might as well (and do) assume that $\overline{\lambda}$ is so big that we can easily sample from $\pi_{\overline{\lambda}}$. Of course $\overline{\lambda}$ will be our 'maximal state'. As $n(X^i)$ is obviously bounded below by $|\mathcal{C}|/(\pi r^2)$, there is a very natural minimal state:

$$\underline{\lambda} = \frac{|\mathcal{C}|/(\pi r^2) + n(X^b)}{|W \oplus rB|}.$$

Thus we may restrict the parameter space to $[\underline{\lambda}, \overline{\lambda}]$.

Using the sampling scheme described at the end of section 4.3 we can obtain *nested* samples X_λ for *all* $\lambda \in [\underline{\lambda}, \overline{\lambda}]$. With such a collection of samples we can define a random transition maps (cf. section 1.5.1) $H : [\underline{\lambda}, \overline{\lambda}] \to [\underline{\lambda}, \overline{\lambda}]$ by

$$H(\lambda) = \frac{n(X_\lambda) + n(X^b)}{|W \oplus rB|} \wedge \overline{\lambda}.$$

Since the X_λ are nested, $n(X_\lambda)$ is ('surely') monotone increasing as a function of λ and hence so is H. Coupling-from-the-past works as follows here. Fix an integer $-T < 0$. For $t = -T+1, -T+2, \ldots, 0$ generate independent copies N_t of N. Define an upper chain $U_{-T} = \{U_{-T}(t) : t = -T, \ldots, 0\}$ by setting $U_{-T}(t) = H_t(U_{-T}(t-1))$ and $U_{-T}(-T) = \overline{\lambda}$. Similarly, define a lower chain L_{-T} by $L_{-T}(t) = H_t(L_{-T}(t-1))$ and $L_{-T}(-T) = \underline{\lambda}$.

At time 0, check if $U_{-T}(0) = L_{-T}(0)$. If so terminate; if not repeat, starting at time $-2T$ and re-using the $\{H_t : t = -T+1, \ldots, 0\}$.

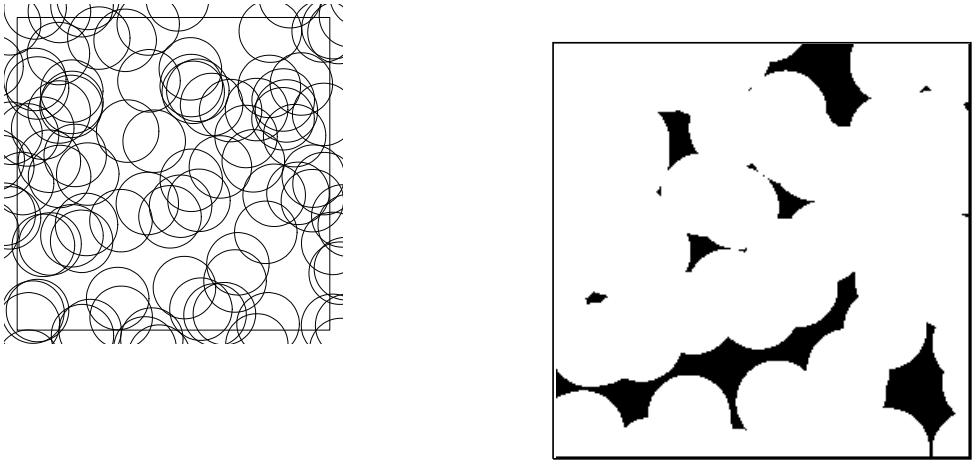

Figure 4.1: (a) Realization of a Poisson point process of intensity 75 on the set $[0,1]^2 \oplus rB$. There are 101 points. With each point a circle with radius 0.1 is associated. (b) The observed Boolean model $Y = \mathcal{B}(X) \cap [0,1]^2$

4.5 An example

In Figure 4.1 (a) we have a realization of a Boolean model of discs of radius .10 as observed through the set $[0,1]^2 \oplus (0.1)B$ (B is the unit disc). The intensity of the underlying germ process is 75. There were 101 points in the set $[0,1]^2 \oplus (0.1)B$, of which 56 could be identified from observing the Boolean model through $[0,1]^2$ as depicted in Figure 4.1 (b). Figure 4.2 shows the region $\mathcal{D}$ where the interior points 'live' and the region $\mathcal{C}$ to be covered.

We have applied the Monte Carlo likelihood ratio approach and the StEM algorithm to the data in Figure 4.1 (b). First, we estimated the log likelihood ratio (4.9) with respect to $\kappa = 80$ for $\lambda \in [50, 100]$. The estimate, shown in Figure 4.3, is based on 50 independent samples from π_{80}. To obtain such a sample we first generated a dominating pattern D. This was done by rejection sampling, where we gently increased the intensity until success. The first, rejection sampling attempts were done at intensity 80. After 50 failures, the intensity was increased by $10/|[0,1]^2 \oplus (0.1)B|$. Generally, success occurred at intensities around 110.

From figure 4.3 the log likelihood ratio is readily maximized. We find 69 as an estimate of the true intensity (75).

Of course the choice $\kappa = 80$ as reference value for the log likelihood ratio is quite arbitrary. Instead, we could have used some pilot estimate.

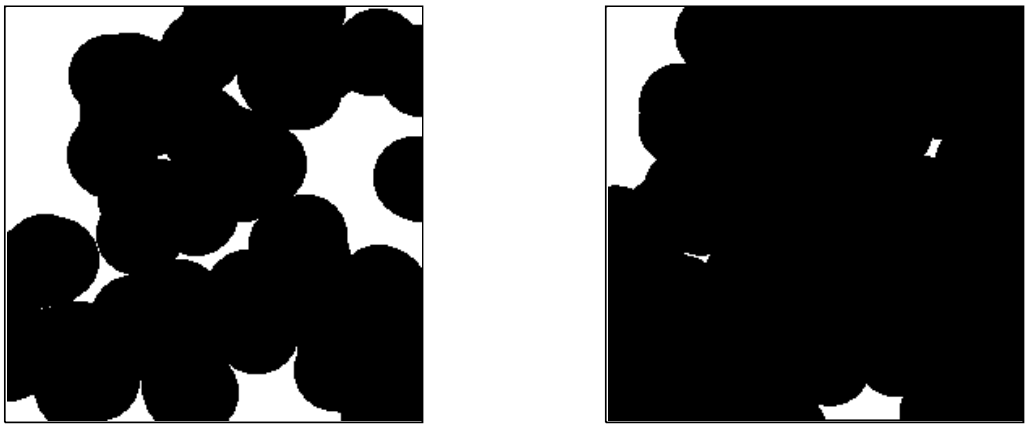

Figure 4.2: (a) The set $\mathcal{D}$ where the interior points 'live'. (b) the set $\mathcal{C}$ which must be covered by the Boolean model associated with the interior points.

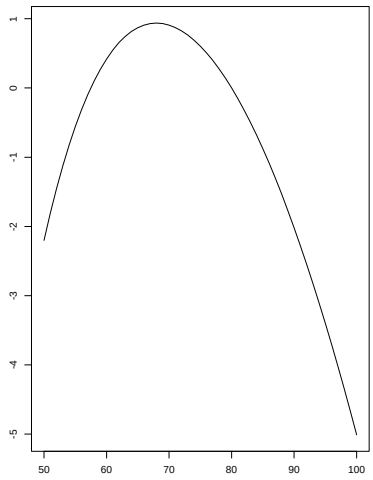

Figure 4.3: Estimated log likelihood ratio as a function of λ with respect to $\kappa = 80$.

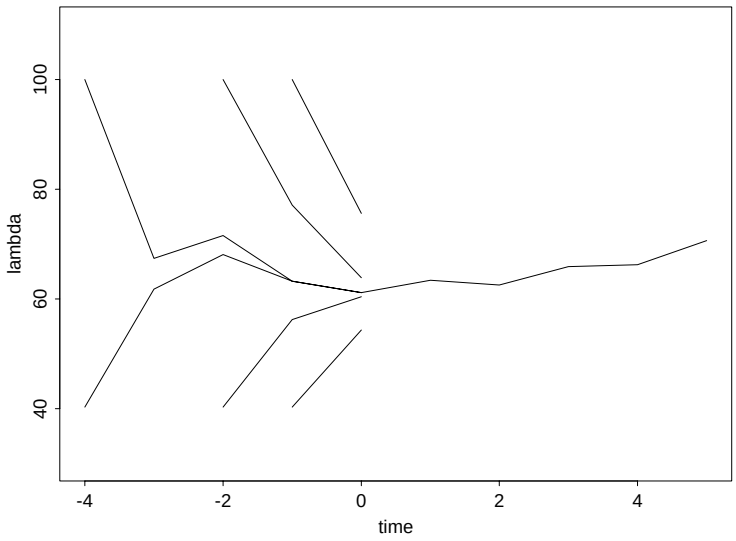

Figure 4.4: The upper and lower processes in the perfect stochastic EM algorithm, starting at times -1, -2 and -4. After the upper and lower processes have coalesced, the (then stationary) chain is continued.

A run of the perfect StEM algorithm is shown in figure 4.4. We see upper and lower processes $U_{-1}(t)$ and $L_{-1}(t)$ and $U_{-2}(t)$ and $L_{-2}(t)$ which have not met at time zero, and processes $U_{-4}(t)$ and $L_{-4}(t)$ which have. We have chosen to start the upper processes at $\overline{\lambda} = 100$, pretending that we know for a fact that the true intensity is below 100. Note how quickly the algorithm has terminated. The value at time zero (after termination) is 61. We could now continue to run the StEM chain starting at 61 at time zero and average the result to bring down the variance.

We should mention that it took us longer to generate Figure 4.4 than Figure 4.3. Although fewer samples were needed for the former, they were of lower intensity (40) than the samples required for the latter (80).

An advantage of the likelihood ratio approach is that it allows us to do much more than just estimate the intensity. Indeed, it is easy to estimate the expectation and variance of any function ϕ of the complete data. We use importance sampling ideas (Geyer, 1994). To this end, 50 independent samples were generated from the conditional distribution of interior points with at intensities $\kappa_i = 60.0 + 10i$ $(i = 0, 1, 2, 3)$. Hence, in all, 200 independent

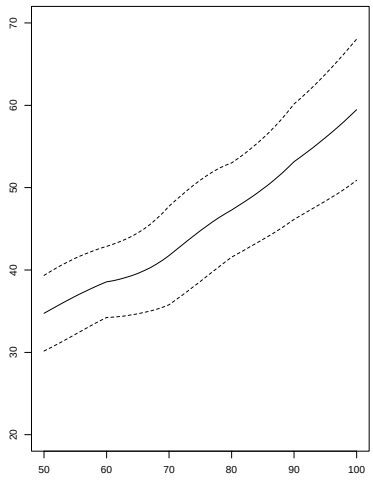

Figure 4.5: The solid line denotes a Monte Carlo estimate of $E_\lambda n(X^i)$ as a function of λ. The dashed lines show the (pointwise) estimated standard deviation.

samples were generated. An estimate for $E_\lambda \phi$ given by

$$s\,\frac{\sum_{j=1}^{50}\phi(X_{i,j})(\frac{\lambda}{\kappa_i})^{n(X_{i,j})}}{\sum_{j=1}^{50}(\frac{\lambda}{\kappa_i})^{n(X_{i,j})}} + (1-s)\,\frac{\sum_{j=1}^{50}\phi(X_{i+1,j})(\frac{\lambda}{\kappa_{i+1}})^{n(X_{i=1,j})}}{\sum_{j=1}^{50}(\frac{\lambda}{\kappa_{i+1}})^{n(X_{i=1,j})}}$$

if $\lambda = s\kappa_i + (1-s)\kappa_{i+1}, 0 \le s < 1$, and where $X_{i,j}$ is the jth sample from π_{κ_i}.

In our case, the expectation of $n(X^i)$ may be the most interesting. In Figure 4.5 we provide Monte Carlo estimates of the expected number of interior points under π_λ, as a function of λ, together with the (pointwise) standard deviation envelopes.

4.6 Sampling from general point processes

In a recent paper Kendall and Møller (1999) presented two perfect sampling algorithms for so-called hereditary (or more precisely: locally stable) point processes. One based on spatial birth-and-death processes and the other a Metropolis–Hastings algorithm. In this section we show that our algorithm 4.1 will generally allow us to sample from what one might call anti-hereditary point processes. We show that the modification we applied to algorithm 4.1 can also be used in this more general setup. Finally, we show that there

exists a similar modification of the birth-and-death algorithm of Kendall and Møller.

Following the notation and setup of Kendall and Møller (1999), let $(S, \mathcal{B}, \lambda)$ be a measure space. For us, λ will always be the uniform measure on S. Consider the 'Carter–Prenter exponential space' Ω of all finite collections of elements of S with its natural sigma-algebra $\mathcal{F}$ generated by sets $\{x \in \Omega : n(x \cap B) = n\}$ $(B \in \mathcal{B})$.

In particular we can take S to be a compact set in $\mathbb{R}^2$ and let $\mathrm{d}\lambda = \lambda \mathrm{d}s$, where $\mathrm{d}s$ is Lebesgue measure on S. Let Po_λ denote the probability measure on $(\Omega, \mathcal{F})$ corresponding to a Poisson process of intensity λ on S. The Poisson process of intensity λ is of course absolutely continuous with respect to the Poisson process of intensity 1. The density is $p_\lambda : \Omega \to [0, \infty)$ given by

$$p_\lambda(x) = e^{(1-S)|S|} \lambda^{n(x)} \tag{4.12}$$

Consider a point process X on S, whose distribution π_λ on $(\Omega, \mathcal{F})$ is absolutely continuous with respect to Po_λ with density $f : \Omega \to [0, \infty)$

$$\mathrm{d}\pi_\lambda = f \mathrm{dPo}_\lambda \tag{4.13}$$

In the previous sections we considered the special case $S = \mathcal{D}$ and

$$f(x) = \frac{\mathbf{1}_{\{x \text{ covers } \mathcal{C}\}}(x)}{\int \mathbf{1}_{\{x \text{ covers } \mathcal{C}\}}(x) \mathrm{dPo}_\lambda(x)}.$$

Defining

$$f_\lambda(x) = f(x) e^{(1-\lambda)|S|} \lambda^{n(x)} \tag{4.14}$$

we have

$$\mathrm{d}\pi_\lambda = f_\lambda \mathrm{dPo}_1.$$

Now consider the problem of obtaining a sample from π_λ. If f is bounded by, say, M we can apply *rejection sampling*. First we take a sample $X \sim \mathrm{Po}_\lambda$ and a uniform random variable U on $[0, M]$. We 'accept' X as a sample from π_λ if $U \leq f(X)$ and otherwise we repeat the procedure. Unfortunately, it often happens that the acceptance probability is so small that this sampling method would take forever. We shall now investigate various other approaches.

We consider two cases:

hereditary There is a constant $K > 0$ such that

$$f(x \cup \{\xi\}) \leq K f(x), \qquad \forall x \in \Omega,\ \xi \in S\ (\xi \notin x) \tag{4.15}$$

To be precise, the hereditary property usually (Kendall and Møller 1999, Geyer 1999) means: if $x \subseteq y$ then $f(x) > 0$ implies $f(y) > 0$. Condition (4.15) is really a 'local stability condition' which *implies* the hereditary property. The ratio $f(x \cup \{\xi\})/f(x)$ is sometimes called the 'Papangelou conditional intensity' and is denoted $\ell^*(x;\xi)$. Perfect sampling for locally stable point processes was considered by Kendall and Møller (1999).

anti-hereditary There is a constant $K > 0$ such that

$$f(x) \leq K f(x \cup \{\xi\}), \qquad \forall x \in \Omega,\ \xi \in S\ (\xi \notin x) \tag{4.16}$$

Straightforward rejection sampling is feasible in the hereditary case when λ is small. In fact, it works trivially for $\lambda = 0$ when we would (almost) always obtain the empty configuration. In the anti-hereditary case rejection sampling typically is feasible when λ is large. Therefore we shall assume throughout that we can easily obtain a sample from π_κ when we choose κ large enough.

4.6.1 spatial birth-and-death processes

We now discuss so-called spatial birth-and-death (b&d) processes as they were introduced by Preston (1975). We closely follow a brief review in Kendall and Møller (1999).

A spatial birth-and-death process X_t $(t \geq 0)$ is a random process taking its values in Ω, the collection of finite sets of points of S. The process is Markov and (hardly surprising) it has two kinds of transitions: births (addition of a point) and deaths (deletion of a point). Now suppose $X(\cdot)$ has 'birth rate' b and death 'rate' d. These are non-negative functions on $\Omega \times S$, while b must satisfy $B(x) = \int b(x,\xi)\mathrm{d}\lambda(\xi) < \infty$. The birth rate $b(x,\xi)$ determines the rate at which $X(\cdot)$ jumps from a state x to $x \cup \{\xi\}$ and the death rate $d(x,\xi)$ determines the rate at which it jumps from $x \cup \{\xi\}$ to x. More precisely, the dynamics of $X(\cdot)$ are as follows.

Suppose that $X_t = x = \{x_1, \ldots, x_n\}$. Let $E_0, E_1, \ldots, E_n$ denote independent exponential random variables with respective means $1/B(x), 1/d(x \setminus \{x_1\}, x_1), \ldots, 1/d(x \setminus \{x_n\}, x_n)$. The first transition after time t happens at time $t + E$, where E is the minimum of $E_0, \ldots, E_n$. If $E = E_0$ then the transition is a birth and a point ξ is added. The point ξ is selected according to the density $b(x,\xi)/B(x)$. If $E = E_i$, $i \neq 0$, the transition is a death and the point x_i is removed.

The following important lemma is from Preston (1975)

Lemma 4.5. *If there exists a density $g : \Omega \to [0, \infty)$ satisfying the so-called* detailed balance equations

$$g(x)b(x,\xi) = g(x \cup \{\xi\})d(x,\xi) > 0, \text{ whenever } g(x \cup \{\xi\}) > 0, \tag{4.17}$$

then X_t is time reversible and it has g as density of its unique stationary distribution.

Suppose that we have—as we do—a density f and wish to construct a b&d process with f as its equilibrium, then we can take $b(x,\xi) = f(x \cup \{\xi\})/f(x)$ and $d(x,\xi) = 1$ or alternatively $b(x,\xi) = 1$ and $d(x,\xi) = f(x)/f(x \cup \{\xi\})$. Of course these birth and death rates should be well-defined, and in this light conditions (4.15) and (4.16) above are not surprising.

4.6.2 the anti-hereditary case

Suppose we have a sample from an anti-hereditary point process with distribution π_κ. In this section we show how to obtain a sample from π_λ for any $\lambda \leq \kappa$ by means of thinning. This section generalizes our earlier work on sampling from the bombing model under a covering condition. The following lemma is a generalization of Proposition 4.1.

Lemma 4.6. *Suppose condition (4.16) holds. Let $\kappa > \lambda > 0$. Suppose Y is a sample from π_κ and let Y' be an independent thinning of Y with retention probability λ/κ. Then*

$$\int_y \frac{f(x)}{f(x \cup y)} P(Y' = x | Y = x \cup y)\pi_\kappa(x \cup \mathrm{d}y) = e^{-|S|} f_\lambda(x) \tag{4.18}$$

Proof By condition (4.16) we have that $f(x)/f(x \cup y)$ is bounded. Now it follows from (4.14) that

$$f_\kappa(x \cup y) = f(x \cup y)e^{(1-\kappa)|S|}\kappa^{n(x)+n(y)}$$

and

$$f(x) = f_\lambda(x)e^{(\lambda-1)|S|}\lambda^{-n(x)}.$$

Hence

$$\begin{aligned}
&\int_y \frac{f(x)}{f(x \cup y)} P(Y' = x | Y = x \cup y) \pi_\kappa(x \cup \mathrm{d}y) \\
&= \sum_{n=0}^{\infty} \frac{e^{-|S|}}{n!} \int \cdots \int_{S^n} \frac{f(x)}{f(x \cup \{y_1, \ldots, y_n\})} \left(\frac{\lambda}{\kappa}\right)^{n(x)} \left(1 - \frac{\lambda}{\kappa}\right)^n \\
&\times f_\kappa(x \cup \{y_1, \ldots, y_n\}) \mathrm{d}y_1 \ldots \mathrm{d}y_n \\
&= \sum_{n=0}^{\infty} \frac{e^{-|S|}}{n!} \int \cdots \int_{S^n} \frac{f_\lambda(x)}{f(x \cup \{y_1, \ldots, y_n\})} e^{(\lambda-1)|S|} \lambda^{-n(x)} \left(\frac{\lambda}{\kappa}\right)^{n(x)} \left(1 - \frac{\lambda}{\kappa}\right)^n \\
&\times f(x \cup \{y_1, \ldots, y_n\}) e^{(1-\kappa)|S|} \kappa^{n(x)+n} \mathrm{d}y_1 \ldots \mathrm{d}y_n \\
&= f_\lambda(x) e^{-|S|} e^{(\lambda-\kappa)|S|} \sum_{n=0}^{\infty} \frac{1}{n!} \kappa^n \left(1 - \frac{\lambda}{\kappa}\right)^n |S|^n \\
&= f_\lambda(x) e^{-|S|}
\end{aligned}$$

□

Suppose we have a sample $D = y = \{y_1, \ldots, y_n\}$ from π_κ and require a sample from π_λ for some $\lambda < \kappa$. Consider a spatial b&d process, $X(t)$, $t \geq 0$, on the powerset of y (denoted $\mathcal{P}(y)$) with birthrate $b(x, \xi) = 1_{\{\xi \in y \setminus x\}} \lambda/(\kappa - \lambda)$ and death rate $d(x, \xi) = f(x)/f(x \cup \{\xi\})$. Let D' denote an independent thinning of D with retention probability λ/κ. We have the following identity, for all $x \subset x \cup \{\xi\} \subseteq y$

$$\frac{\lambda}{\kappa - \lambda} \frac{f(x)}{f(y)} P(D' = x | D = y) = \frac{f(x)}{f(x \cup \{\xi\})} \frac{f(x \cup \{\xi\})}{f(y)} P(D' = x \cup \{\xi\} | D = y),$$

In other words, $X(t)$ satisfies detailed balance with respect to

$$\frac{f(x)}{f(y)} P(D' = x | D = y)$$

which is therefore the stationary distribution of $X(t)$, conditionally on $D = y$. Since D is distributed according to π_κ it follows from lemma 4.6 that *unconditionally* the stationary distribution of $X(t)$ is π_λ.

Now consider another spatial birth and death process, $E(t)$, $t \geq 0$, on $\mathcal{P}(y)$ with birth rate $b(x, \xi) = 1_{\{\xi \in y \setminus x\}} \lambda/(\kappa - \lambda)$ and death rate K. Let E_0 be an independent thinning of D with retention probability

$$\frac{\lambda/(K(\kappa - \lambda))}{1 + \lambda/(K(\kappa - \lambda))}.$$

This time we have, for all $x \subset x \cup \{\xi\} \subseteq y$

$$\frac{\lambda}{\kappa - \lambda} P(E_0 = x | D = y) = KP(E_0 = x \cup \{\xi\} | D = y),$$

We see that $E(t)$ satisfies detailed balance with respect to the conditional distribution of E_0 given D. It follows that $E(t)$ is time-reversible and in equilibrium if we set $E(0) = E_0$.

$X(t)$ and $E(t)$ have identical birth rates while by (4.16) the death rate of $E(t)$ is greater than that of $X(t)$. Setting $X(0) = E(0)$ it is possible to generate coupled realizations of $X(t)$ and $E(t)$ such that $E(t) \subseteq X(t)$ for all t. This works as follows. Suppose $X(t) = x$ and $E(t) = e$, $e \subseteq x \subseteq y$. Then associate with every point of $x_i \in x$ an exponential lifetime L_i with mean $1/K$ and a uniform random variable U_i on $[0, K]$. Associate with all points $y_i \in y \setminus e$ an exponential waiting time W_i with mean $(\kappa - \lambda)/\lambda$. The first transition after time t takes place at time $t + m$ where $m = \min(L_1, \ldots, L_{n(x)}, W_1, \ldots, W_{n(y)-n(e)})$. If $m = L_i$ then the point x_i is removed (if present) from the E process and x_i is removed from the X process, *but only if* $U_i \leq f(x \setminus \{x_i\})/f(x)$. If $m = W_i$ the point e_i is added to the E process and (if not already present) to the X process.

We now explain how we can apply coupling-from-the-past (CFTP) to obtain a perfect sample from the π_λ, the stationary distribution of $X(t)$. Fix a time $-T < 0$ in the past. Since the E process is time reversible, we can extend it from $E(0) = E_0$ into the past until time $-T$. Now define a process X_{-T} on $[-T, 0]$ starting at $X_{-T}(-T) = E(-T)$. We can arrange it so that X_{-T} is a birth and death process with the same dynamics as X, while $E(t) \subseteq X_{-T}(t) \subseteq D$ for all $t \in [-T, 0]$. Because X_{-T} has the same dynamics as X, it follows that $X_{-T}(0)$ tends in distribution to π_λ as $T \to \infty$.

To check if $X_{-T}(0)$ has reached equilibrium—so to say—we need a 'lower bound process' $L_{-T}(\cdot)$ with $L_{-T}(-T) = E(-T)$ and an 'upper bound process' $U_{-T}(\cdot)$ with $U_{-T}(-T) = D$ such that (cf. Kendall and Møller (1999) formulas (3.5) to (3.7))

1. ('sandwiching') $E(t) \subseteq L_{-T}(t) \subseteq X_{-T}(t) \subseteq U_{-T}(t) \subseteq D$ for all $-T \leq t \leq 0$;

2. ('funneling') $L_{-T}(t) \subseteq L_{-S}(t) \subseteq X_{-S}(t) \subseteq U_{-S}(t) \subseteq U_{-T}(t)$, for all $-S \leq -T \leq t \leq 0$;

3. ('coalescing') if, for some s, $L_{-T}(s) = U_{-T}(s)$ then $L_{-T}(t) = U_{-T}(t)$, for all $t \geq s$.

Moreover, we need that the event $\{L_{-T}(0) = U_{-T}(0)\}$ almost surely occurs for finite T. If the above 3 inclusion properties hold and the latter event

occurs then we can conclude that $L_{-T}(0) = U_{-T}(0)$ is a perfect sample from π_λ. This is roughly the content of Theorem 3.1 of Kendall and Møller (1999). The difference is that they consider a varying 'maximal state' $D(t)$ but a fixed minimal state $E(t) = \emptyset$, whereas we consider a fixed maximal state $D(t) = D$, but a varying minimal state.

In practice, the CFTP algorithm would consist of successively choosing increasing T and running $E(t)$, $L_{-T}(t)$ and $U_{-T}(t)$ for $t \in [-T, 0]$ until $L_{-T}(0) = U_{-T}(0)$.

The algorithm takes a long time to terminate when the difference between D and $E(0)$ is large. This occurs typically if the difference between κ and λ is large. Therefore it is a good idea to choose a decreasing sequence $\kappa = \lambda(0) > \lambda(1) > \cdots > \lambda(n) = \lambda$ and successively sample from $\pi_{\lambda(0)}, \pi_{\lambda(1)}, \ldots, \pi_{\lambda(n)}$. At each step we can use the sample obtained in the previous step as dominating pattern D. In fact, in section 4.3 we showed how to choose the $\lambda(i)$ in such a way that the difference between D and $E(0)$ is never more than a single point. This allowed us to obtain samples from π_λ for *all* λ in a given interval.

4.6.3 the hereditary case

Recall our definition of an hereditary point process at (4.15). Kendall and Møller (1999) give two algorithms to obtain samples from a hereditary point process with distribution π_λ, one of which is based on birth-and-death processes. They construct a b&d process $X(t)$ with π_λ as its stationary distribution and a dominating process $D(t)$ which is in equilibrium and time reversible. In fact, $D(t)$ is at all times t a spatial Poisson process of intensity $K\lambda$. Kendall and Møller give a coupling of $D(t)$ and $X(t)$ such that $X(t) \subseteq D(t)$ for all t. They then demonstrate a coupling-from-the-past algorithm with $D(t)$ ($t \in [-T, 0]$) as a (variable) maximal state and $E(t) = \emptyset$ as a trivial minimal state. The sample from π_λ which the algorithm produces lies between $D(0)$ and $E = \emptyset$. When λ is large the difference between $D(0)$ and E will typically be large. If this is the case, then the algorithm will take very long to terminate because it has many different configurations to choose from.

We give a generalization of the work of Kendall and Møller which allows us to take as a fixed minimal state a sample $E \sim \pi_\kappa$ for an arbitrary $\kappa < \lambda$. The maximal state $D(t)$ will be a birth-and-death process which has the superposition of E and a Poisson process of intensity $K(\lambda - \kappa)$ as its equilibrium distribution. Noting that the empty set is trivially a sample from π_0 we see that our generalization reduces to the Kendall and Møller algorithm when we take $\kappa = 0$.

The use of the generalization is that it will allow us to choose an increasing

sequence $0 = \lambda(0) < \lambda(1) < \cdots < \lambda(n) = \lambda$ and successively sample from $\pi_{\lambda(0)}, \pi_{\lambda(1)}, \ldots, \pi_{\lambda(n)}$. At each step we can use the sample obtained in the previous step as minimal pattern E to which we add a Poisson point process of intensity $K(\lambda(i+1) - \lambda(i))$ to obtain the maximal pattern $D(0)$. If we choose the steps from $\lambda(i)$ to $\lambda(i+1)$ small enough, then the difference between $D(0)$ and E will always be small too.

The next lemma shows how we can obtain a sample from π_λ by adding points to a sample from π_κ for an arbitrary $\kappa < \lambda$.

Lemma 4.7. *Suppose condition (4.15) holds. Let $0 \le \kappa < \lambda$. Then for all $x \in \Omega$*

$$\sum_{y \subseteq x} \frac{f(x)}{f(y)} f_\kappa(y) p_{\lambda-\kappa}(x \setminus y) = f_\lambda(x). \tag{4.19}$$

Proof This lemma basically follows from the superposition property of Poisson processes. By condition (4.15) we have that $f(x)/f(y)$ is bounded. Now it follows from (4.14) that

$$f_\kappa(y) = f(y) e^{(1-\kappa)|S|} \kappa^{n(y)}$$

and

$$f(x) = f_\lambda(x) e^{(\lambda-1)|S|} \lambda^{-n(x)}.$$

Hence

$$\begin{aligned}
&\sum_{y \subseteq x} \frac{f(x)}{f(y)} p_{\lambda-\kappa}(x \setminus y) f_\kappa(y) \\
&= \sum_{y \subseteq x} \frac{f_\lambda(x)}{f(y)} e^{(\lambda-1)|S|} \lambda^{-n(x)} e^{-(\lambda-\kappa)|S|} (\lambda - \kappa)^{n(x \setminus y)} f(y) e^{(1-\kappa)|S|} \kappa^{n(y)} \\
&= f_\lambda(x) \lambda^{-n(x)} \sum_{y \subseteq x} (\lambda - \kappa)^{n(x)-n(y)} \kappa^{n(y)} \\
&= f_\lambda(x).
\end{aligned}$$

□

Suppose we have a sample $E = y = \{y_1, \ldots, y_n\}$ from π_κ for some $\kappa < \lambda$. Now consider a spatial birth-and-death process $X(t)$, $t \ge 0$ on the space $\{y \cup x : x \in \Omega\}$ with birth rate $b(y \cup x, \xi) = (\lambda - \kappa) f(y \cup x \cup \{\xi\}) / f(y \cup x)$ and death rate $d(y \cup x, \xi) = 1$. One easily checks that

$$(\lambda - \kappa) \frac{f(y \cup x \cup \{\xi\})}{f(y \cup x)} \frac{f(y \cup x)}{f(y)} p_{\lambda-\kappa}(x) = 1 \frac{f(y \cup x \cup \{\xi\})}{f(y)} p_{\lambda-\kappa}(x \cup \{\xi\}).$$

So $X(t)$ satisfies detailed balance with respect to

$$\frac{f(y \cup x)}{f(y)} p_{\lambda-\kappa}(x),$$

which is therefore the density of the stationary distribution of $X(t)$, conditionally on $E = y$. Hence, by lemma 4.7 the unconditional stationary distribution of $X(t)$ is π_λ.

Consider another b&d process, $D(t)$, $t \geq 0$, on $\{y \cup x : x \in \Omega\}$ with birthrate $b(y \cup x, \xi) = K(\lambda - \kappa)$ and death rate $d(y \cup x, \xi) = 1$. Let D_0 be the union of y and be a sample from $\mathrm{Po}_{K(\lambda-\kappa)}$. We have for all $x \in \Omega$

$$K(\lambda - \kappa) p_{K(\lambda-\kappa)}(x) = 1 p_{K(\lambda-\kappa)}(x \cup \{\xi\}).$$

Hence if we set $D(0) = D_0$ then $D(t)$ is time reversible and in equilibrium.

$X(t)$ and $D(t)$ have the same death rates, while by (4.15) the birth rate of $D(t)$ is greater than the birth rate of $X(t)$. Setting $X(0) = D(0)$ it is not difficult to couple $D(t)$ and $X(t)$ such that $X(t) \subseteq D(t)$ for all t. This could be done much like we indicated in the previous subsection for the anti-hereditary case. In fact, a detailed description of the coupling is given in Kendall and Møller (1999) for the particular—but in no sense special—case $\kappa = 0$. They also explain how to apply CFTP to obtain a perfect sample from the stationary distribution of $X(t)$.

References

A.J. Baddeley and R.D Gill (1997). Kaplan-Meier estimators for interpoint distance distributions of spatial point processes. *Ann. Statist.* **25** 263–292.

A.J. Baddeley (1999a). A crash course in stochastic geometry. In: *Proceedings Seminaire Européen de Statistique, 'Stochastic geometry, likelihood, and computation'*, O.E. Barndorff–Nielsen, W.S. Kendall and M.N.M. van Lieshout (editors). CRC Press (Chapman and Hall).

A.J. Baddeley (1999b). Spatial sampling and censoring. In: *Proceedings Seminaire Européen de Statistique, 'Stochastic geometry, likelihood, and computation'*, O.E. Barndorff–Nielsen, W.S. Kendall and M.N.M. van Lieshout (editors). CRC Press (Chapman and Hall).

P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

G. Celeux and J. Diebolt (1986). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2** 73–82.

A. Chessa (1995). *Conditional Simulation of Spatial Stochastic Models for Reservoir Heterogeneity.* PhD dissertation, Technische Universiteit Delft.

C.F. Chung (1989a). Confidence bands for residual lifetime under random censorship model. *J. Multiv. Anal.* **29** 94–126.

C.F. Chung (1989b). Confidence bands for quantile function under random censorship. *Ann. Inst. Statist. Math.* **42** 21–36.

D.R. Cox (2000). Some Remarks on Likelihood Factorization. In: *State of the Art in Probability and Statistics.* M.C.M. de Gunst, C.A.J. Klaassen and A.W. van der Vaart (editors). IMS Lecture notes (to appear).

N.A.C. Cressie (1991). *Statistics for Spatial Data*, John Wiley & Sons.

D. J. Daley and D. Vere-Jones (1988). *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.

A.P. Dempster, N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

S.A. van de Geer (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

J.A. Fill and M. Machida (1998). Stochastic monotonicity and realizable monotonicity. Technical Report # 573, Department of Mathematical Sciences, The Johns Hopkins University. Submitted for publication.
`http://www.mts.jhu.edu/~fill/`

C.J. Geyer (1999). Likelihood inference for spatial point processes, *Proceedings Seminaire Européen de Statistique, 'Stochastic geometry, likelihood, and computation'*, O.E. Barndorff–Nielsen, W.S. Kendall and M.N.M. van Lieshout (editos), CRC Press (Chapman and Hall).

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

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

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

R.D. Gill (1989). Non- and semiparametric maximum likelihood estimators and the von Mises method (part I). *Scand. J. Statist.* **16** 97–128.

R.D. Gill (1994). *Lectures on Survival Analysis.* In: D. Bakry, R.D. Gill and S. Molchanov, École d'Été de Probabilités de Saint Flour XXII-1992, P. Bernard (editor). Springer Lecture Notes in Mathematics 1581 115–241.

P. Groeneboom and J.A. Wellner (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Birkhäuser Verlag, Basel.

M.B. Hansen, A.J. Baddeley and R.D. Gill (1996). Kaplan-Meier type estimators for linear contact distributions. *Scand. J. Statist.* **23** 129–155.

M.B. Hansen, A.J. Baddeley and R.D. Gill (1999). First contact distributions for spatial patterns: Regularity and estimation. *Adv. Appl. Probab.* **31** 15–33.

Y. Huang and C.-H. Zhang (1994). Estimating a monotone density from censored observations. *Ann. Statist.* **22** 1256–1274.

J. Huang and J.A. Wellner (1995). Estimation of a monotone density or monotone hazard under random censorship. *Scand. J. Statist.* **22** 3–33.

W.S. Kendall (1997). On some weighted Boolean models. In: *Proceedings of the International Symposium on Advances in Theory and Applications of Random Sets*, D. Jeulin (editor), pp. 105–120, World Scientific Publishing.

W.S. Kendall and E. Thönnes (1998). Perfect simulation in stochastic geometry. University of Warwick Research Report 323.
`http://www.warwick.ac.uk/statsdept/Staff/WSK/ppt.html`

W.S. Kendall and J. Møller (1999). Perfect Metropolis-Hastings simulation of locally stable spatial point processes. University of Warwick Research Report 347.
`http://www.warwick.ac.uk/statsdept/Staff/WSK/ppt.html`

J. Kiefer and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

M.J. van der Laan (1993). *Efficient and Inefficient Estimation in Semiparametric models* PhD dissertation, University of Utrecht. Also: CWI tract 114 (1995).

M.J. van der Laan (1995). Efficiency of the NPMLE in the line-segment problem. *Scand. J. Statist.* **23** 527–550.

M.J. van der Laan (1998). The two-interval line-segment problem. *Scand. J. Statist.* **25** 163–186.

C. Lantuéjoul (1997). Conditional simulation of object-based models. In: *Advances in Theory and applications of Random Sets*, D. Jeulin (editor). World Scientific Publishing Company, Singapore, 271–288.

G.M. Laslett (1982a). The survival curve under monotone density constraints with applications to two-dimensional line-segment processes. *Biometrika* **69**, 153–160.

G.M. Laslett (1982b). Censoring and edge effects in areal and line transect sampling of rock joint traces. *Math. Geol.* **14**, 125–140.

L. Le Cam and G.L. Yang (1988). On the preservation of local asymptotic normality under information loss. *Ann. Statist.* **16** 483–520.

M.N.M. van Lieshout (1997). On likelihoods for Markov random sets and Boolean models. In: *Proceedings of the International Symposium on*

Advances in Theory and Applications of Random Sets, D. Jeulin (editor). World Scientific Publishing, Singapore.

T. Lindvall (1992). *Lectures on the Coupling Method.* John Wiley & Sons.

J. Lok (1994). *The NPMLE of the Orientation Distribution and Comparison of Estimators of the Length Distribution in the Line Segment Problem.* Master's dissertation, University of Utrecht.

H. Lum, I. Huang and W. Mitzner (1994). Morphological evidence for alveolar recruitment during inflation at high transpulmonary pressure. *J. Appl. Physiol.* **68** 2280–2286.

G. Matheron (1975). *Random Sets and Integral Geometry.* John Wiley & Sons.

I. Meilijson (1989). A fast improvement to the EM-algorithm. *J. Roy. Statist. Soc. Ser. B* **51** 127–138.

I.S. Molchanov (1995). Statistics of the Boolean model: From the estimation of means to the estimation of distributions. *Adv. in Appl. Probab.* **27** 63–86.

I.S. Molchanov (1997). *Statistics of the Boolean model for Practitioners and Mathematicians.* John Wiley & Sons.

J. Møller (1998). Markov chain Monte Carlo and spatial point processes. In: *Proceedings Seminaire Européen de Statistique, 'Stochastic geometry, likelihood, and computation'*, O.E. Barndorff–Nielsen, W.S. Kendall and M.N.M. van Lieshout (editors). CRC Press (Chapman and Hall).

S.A. Murphy (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22** 712–731.

X.X. Nguyen and H. Zessin (1979). Ergodic theorems for spatial processes. *Z. Wahrsch. Verw. Gebiete* **48** 133–158.

S.F. Nielsen (1997a). The imputation maximization algorithm. Preprint 2, Institute of Mathematical Statistics, University of Copenhagen.
`http://www.math.ku.dk/~feodor/publications/index.html`

S.F. Nielsen (1997b). On simulated EM algorithms. Preprint 3, Institute of Mathematical Statistics, University of Copenhagen.
`http://www.math.ku.dk/~feodor/publications/index.html`

E.H. Oldmixon, J.P. Butler and F.G. Hoppin (1994). Semi-automated measurement of true chord length distributions and moments by video microscopy and image analysis. *J. Microsc.* **175** 60–69.

C.J. Preston (1977). Spatial birth-and-death processes. *Bull. Int. Statist. Inst.* **46** (2) 371–391

J.G. Propp and D.B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9** 223–252.

B.D. Ripley (1981). *Spatial Statistics*, John Wiley & Sons.

B.D. Ripley (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press.

F.S. Rosenthal (1989). Aerosol recovery following breathholding derived from the distribution of chordlengths in pulmonary tissue. *J. Aerosol. Sci.* **20** 267–277.

D.A. Ross (1993). A coherence theorem for ordered families of probability measures on a partially ordered space. Unpublished manuscript.

J. Serra (1982). *Image Analysis and Mathematical Morphology.* Academic Press, London.

F.W. Scholz (1980). Towards a unified definition of maximum likelihood. *Canad. J. Statist.* **8** 193–203.

D. Stone, D.C. Kamineni and A. Brown (1984). Geology and fracture characteristics of the underground research laboratory near Lac du Bonnet, Manitoba, Tech. Rep. 243, Atomic Energy of Canada Ltd. Research corporation.

D. Stoyan, W.S. Kendall, and J. Mecke (1987). *Stochastic Geometry and its Applications.* Wiley, Chichester.

A.W. van der Vaart (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24** 862–878.

A.W. van der Vaart and J.A. Wellener (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

A.W. van der Vaart (1998). *Asymptotic Statistics.* Cambridge University Press.

Y. Vardi (1982). Nonparametric estimation in renewal processes. *Ann. Statist.* **10**, 772–785.

Y. Vardi and D. Lee (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 569–612.

G.C.G. Wei and M.A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Ass.* **85** 699–704.

E.W. Weibel (1979). *Stereological Methods*, Volume 2. Academic Press, London.

B.J. Wijers (1995a). *Nonparametric Estimation for a Windowed Line-Segment Process.* PhD dissertation, University of Utrecht. Also: CWI tract 121 (1997).

B.J. Wijers (1995b). Consistent non-parametric estimation for a one-dimensional line segment process observed in an interval. *Scand. J. Statist.* **22** 335–360.

E.W. van Zwet (2000). Perfect stochastic EM. In: *State of the Art in Probability and Statistics.* M.C.M. de Gunst, C.A.J. Klaassen and A.W. van der Vaart (editors). IMS Lecture notes (to appear).

Acknowledgements

First of all, I would like to express my gratitude to Richard Gill, my supervisor, for supporting me mathematically while allowing me great freedom. Chapter 3 is based on joint work with Martin Hansen (Aalborg) and Chapter 4 is joint with Marie-Colette van Lieshout (CWI, Amsterdam). I'm looking forward to turning these chapters into papers with them.

I am grateful for the hospitality of the maths and stats departments of the University of Western Australia (Perth) and the University of Washington (Seattle) where I enjoyed stays of six and three months respectively. These travels have been supported by the Dutch national science foundation (NWO grants SIR 13-4457 and SIR 13-4641) and by Shell.

Despite being head of an understaffed computing support group, André de Meijer has always found the time to help me out with computer problems. These sessions were both illuminating and enjoyable.

It seems to have become a tradition in our department to devote at least a full page to thanking one's friends and family. As much as I enjoy reading such displays of affection, I shall not include one here. I trust my friends and family will not take its absence as a sign of ingratitude.

Samenvatting

Op vele gebieden, zoals geologie, biologie of medicijnen worden gegevens verzameld bestaande uit een beeld van een proces van geometrische objecten in de ruimte. Lijnsegmenten kunnen breukvlakken in een rotsformatie voorstellen, waargenomen als rechte lijnen op het rotsoppervlak. Punten kunnen de locaties van bomen in een bos weergeven. Cirkelschijfjes kunnen cellen onder een microscoop zijn. Altijd is ons beeld begrensd, ook al strekt het proces zich ver buiten het beeld uit. Er ontstaat een soort censurering van de gegevens, namelijk dat van sommige objecten alleen een gedeelte waarneembaar is zodat de werkelijke omvang onbekend blijft.

Dit proefschrift gaat over ruimtelijke statistiek, of preciezer: de mathematisch statistische analyse van ruimtelijke data. We observeren een of meerdere beelden, zoals hierboven beschreven, en stellen ons voor dat deze zijn ontstaan als trekkingen uit een kansverdeling. Het doel is om uit onze waarnemingen deze kansverdeling—of belangrijke eigenschappen ervan—zo goed mogelijk te achterhalen. Wellicht verheldert het volgende voorbeeld wat we bedoelen.

Figuur 1.1 in dit proefschrift is een geologische kaart van een gebied van zo'n 160 bij 160 meter van het Canadese schild. In het zwarte deelgebied kunnen we het granieten rotsoppervlak waarnemen. Buiten dit gebied kunnen we het graniet niet zien vanwege begroeiing of water. De witte lijnen in het zwarte deelgebied zijn barsten in de rots. In hoofdstuk 2 geven we een methode om de kansverdeling van de lengte van de barsten te schatten op basis van de kaart.

Men zou kunnen zeggen dat ruimtelijke statistiek moeilijker is dan 'gewone' statistiek. Een typische bijkomende moeilijkheid is ruimtelijke afhankelijkheid: wat men op verschillende locaties in een beeld waarneemt is zelden onafhankelijk van elkaar. Onafhankelijkheid is een standaard aanname in de statistiek, die in de ruimtelijke problemen zelden kan worden gehandhaafd. Andere moeilijkheden komen door censurering en andere rand-effecten. 'Randeffecten' is een verzamelnaam voor wiskundige complicaties die ontstaan als een ruimtelijk proces slechts in een beperkt gebied wordt waargenomen. Als

gevolg van al deze moeilijkheden is de ruimtelijke statistiek enigzins een achtergebleven gebied. In mijn proefschrift passen we echter moderne 'mainstream' statistische methoden toe in enige ruimtelijke problemen, waarbij we bovengenoemde moeilijkheden omzeilen. De wijze waarop wij dit doen zijn de 'listen' uit de titel.

In hoofdstuk 1 leggen we ons gereedschap klaar. Nieuw in dit hoofdstuk is een algemene verhandeling over consistentie van de zogeheten meest aannemelijke schatter. Dit resulteert onder andere in Stelling 1.2. Deze stelling geeft voorwaarden waaronder de (niet parametrische) meest aannemelijke schatter in convexe modellen consistent is. In hoofdstuk 2 en 3 kunnen we deze stelling toepassen, hoewel we dit slechts in hoofdstuk 2 daadwerkelijk uitwerken. Ook presenteren we in hoofdstuk 1 een modificatie van de stochastische versie van het bekende EM algorithme. Dit algorithme passen we toe in hoofdstuk 4.

In hoofdstuk 2 beschouwen we 'Laslett's lijnsegmenten probleem'. Het doel is de verdeling van de lengte van lijnsegmenten, waargenomen door een stochastische verzameling, te schatten. Dit probleem heeft een een ruime geschiedenis en werd onder meer bestudeerd in twee eerdere Utrechtse proefschriften, die van Mark van der Laan en Bart Wijers. Wij beschouwen het probleem in grotere algemeenheid en laten de aanname vallen dat het waarnemingsgebied convex is. Als gevolg hiervan kunnen we verschillende fragmenten van één lijnsegment waarnemen. In de praktijk is het onmogelijk te onderscheiden welke fragmenten bij elkaar horen, en zodoende zijn ze op een onontwarbare wijze afhankelijk. Onze 'list' is een eenvoudiger, maar vergelijkbaar probleem te beschouwen zonder deze afhankelijkheid. Voor dit probleem kunnen we de meest aannemelijke schatter afleiden. We tonen vervolgens aan dat de schatter ook goede eigenschappen heeft in het oorspronkelijke probleem.

In hoofdstuk 3 schatten we de kansverdeling van de lengte van een 'typische koorde' van een stochastische verzameling. Een koorde is het langste lijnsegment door een gegeven punt van de verzameling in een gegeven richting, dat geheel binnen de verzameling past. De moeilijkheid is dat we de kansverdeling van de stochastische verzamling in termen van de koordelengte verdeling zouden moeten kennen om de meest aannemelijke schatter uit te kunnen rekenen. Dit is helaas zelden het geval. Wederom beschouwen we een eenvoudiger variant van het probleem, leiden de meest aannemelijke schatter af voor *dat* probleem en laten zien dat de schatter ook goed is in de oorspronkelijke situatie.

In hoofdstuk 4 bestuderen we het 'bommen' model. We nemen de vereniging waar van enige toevallig rondgestrooide cirkelschijfjes. Men zou zich kunnen voorstellen dat deze schijfjes bom-kraters zijn, hetgeen de naam van dit

model verklaart. Op basis van onze waarnemingen willen we het verwachte aantal schijfjes schatten, waarvan het centrum zich in een verzameling van gegeven oppervlakte bevindt. Voor de kenner: We schatten de intensiteit van het Poisson process dat het waargenomen Boolese model genereerde. Het probleem is dat sommige schijfjes geheel bedekt kunnen worden door andere en zodoende volledig onzichtbaar zijn. Hier bestaat onze list uit een algorithme dat ons in staat stelt een steekproef te nemen uit de voorwaardelijke verdeling van de niet-geobserveerde schijfjes, gegeven wat we wel observeren. Met behulp van dergelijke steekproeven kunnen we de meest aannemelijke schatter schatten, zo gezegd. Het algorithme is in grotere algemeenheid toe te passen om puntprocessen te genereren waarvan de zogeheten Papangelou intensiteit van nul is wegbegrensd.

Curriculum Vitæ

De schrijver van dit proefschrift werd geboren te 's-Gravenhage op 10 november 1970. Van 1983 tot 1989 doorliep hij met goed gevolg het curriculum van het Stedelijk Gymnasium te Leiden. Vanaf 1989 studeerde hij wiskunde aan de Universiteit van Leiden, alwaar hij in 1995 het doctoraal behaalde onder het judicium 'met genoegen'. Tijdens zijn studie te Leiden bezocht hij de Universiteit van North Carolina te Chapel Hill en het Institut für Theoretische Nachrichten Technik und Informations Verarbeitung van de Universteit van Hannover. Van 1995 tot 1999 was hij onder begeleiding van Prof. dr Gill als AiO verbonden aan het Mathematisch Instituut van de Universiteit van Utrecht. In deze periode bezocht hij de Universiteit van West-Australië te Perth en de Universiteit van Washington te Seattle. Naast het schrijven van dit proefschrift gaf hij—immer met veel plezier—het nodige onderwijs. Per 1 oktober zal hij werkzaam zijn als post-doc aan de Universiteit van Californië te Berkeley.