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

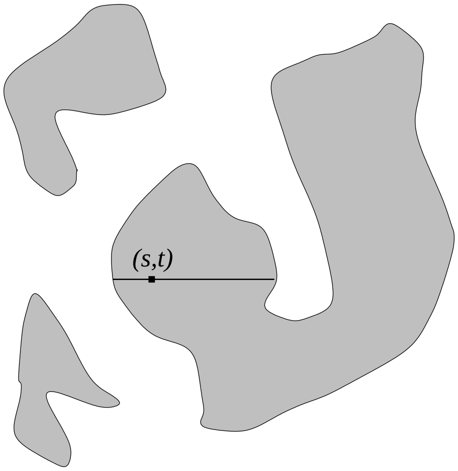

Figure 3.1: The chord $\xi_{(1,0)}(s,t)$. The shaded region is (part of) the random set $\mathcal{W}$.

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

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 $\tau$.

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 $\{\xi(i,s,t) : (s,t) \in W_i,\ i = 1, \dots n\}$. The observed data are the $W_i$ and $\{(y(i,s,t),\ \delta(i,s,t)) : (s,t) \in W_i,\ i = 1, \dots n\}$. 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 $\xi$ be the line-segment from $(R,T)$ to $(R+X,T)$. Note that, given $\xi$, $(S,T)$ is uniformly distributed on $\xi$. Define $Y = |\xi \cap B| = |[R, R+X] \cap [0,\tau]|$ and $\Delta$ is the number of endpoints of $\xi$ 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 $(Y, \Delta)$

**Lemma 3.3.** *The joint distribution of $Y$ and $\Delta$ 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 $(Y, \Delta)$. 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 $(Y, \Delta)$ 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 $0 \leq y < \tau$,

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

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 $0 \leq y < \tau$,

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

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 $\tau$.

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

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 $(Y, \Delta)$ 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 $\Delta$ 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 $(Y, \Delta)$ 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 $\xi_i$ 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 $\Delta_i$ is the number of endpoints of $\xi_i$ that fall outside of $B$. Let $\mathbb{F}_n$ denote the empirical distribution of the $(Y_i, \Delta_i)$. 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 $\tau^-$. 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 $\Xi$ 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 $\Xi$ (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 $\Xi$ 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, $\Xi_k$ 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

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

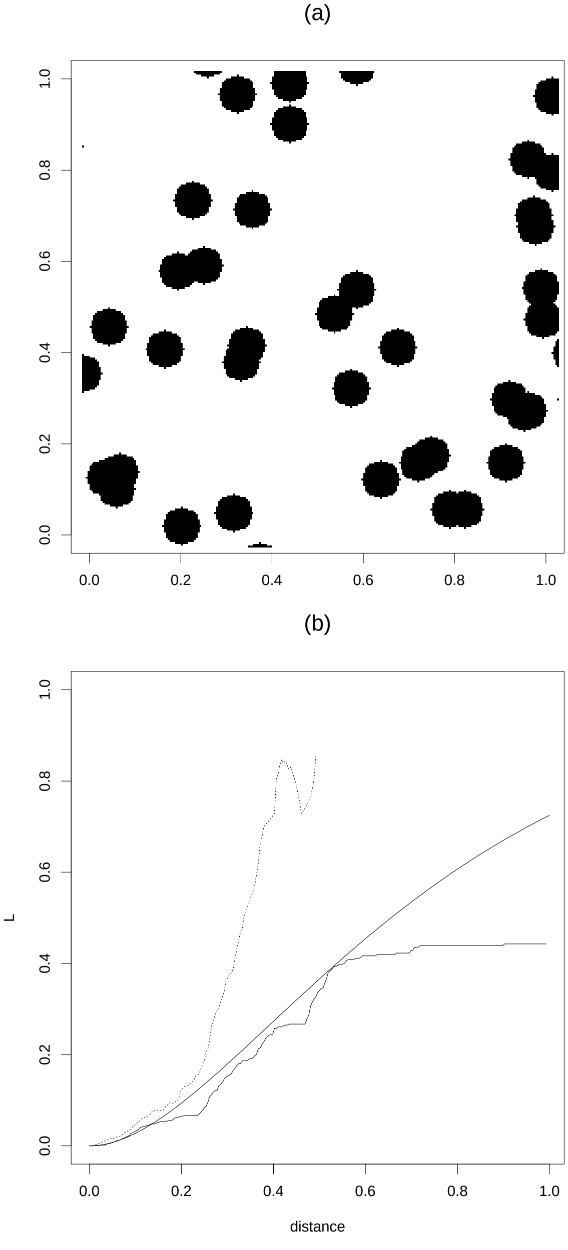

Figure 3.2: (a) Realization of a Boolean model; (b) Corresponding "NPMLE"-estimator $\hat{L}_1$ (solid, ragged line), Reduced sample estimator $\tilde{L}_1$ (dotted line) and the estimand $L$ (solid, smooth line).

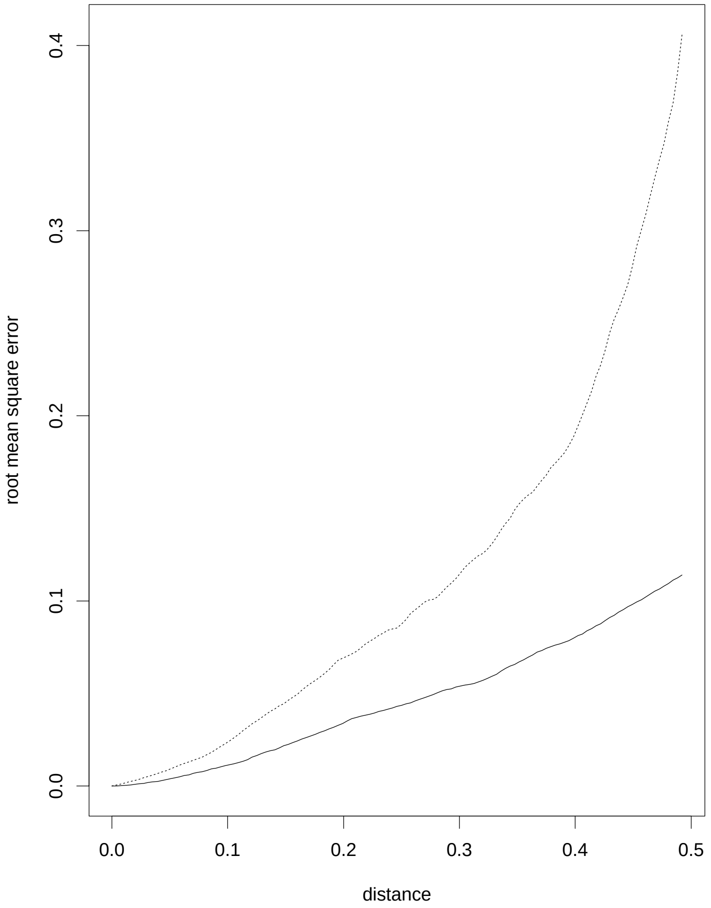

Figure 3.3: Root mean square error comparison of the NPMLE-estimator (solid line) and reduced sample estimator (dotted line).

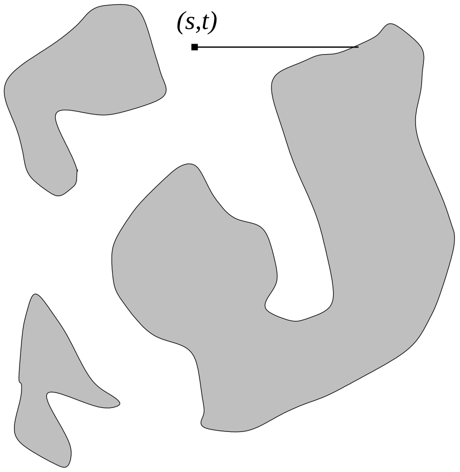

Figure 3.4: The contact segment $\chi_{(1,0)}(s,t)$. The shaded region is (part of) the random set $\mathcal{W}$.

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 $\chi$ 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 $\Delta$ be the number of endpoints of $\chi$ 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 $(Y_i, \Delta_i)$ of $(Y, \Delta)$ is the well-known Kaplan–Meier estimator. The Kaplan–Meier estimator is a function of the empirical distribution of the $(Y_i, \Delta_i)$. 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).