doi:10.5566/ias.1889

A COMPARISON OF NONPARAMETRIC ESTIMATORS FOR LENGTH DISTRIBUTION IN LINE SEGMENT PROCESSES

ZBYNĚK PAWLAS✉,[1] AND MARKÉTA ZIKMUNDOVÁ[2]

[1]Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Prague, Czech Republic; [2]Department of Mathematics, Faculty of Chemical Engineering, University of Chemistry and Technology, Technická 5, 16628 Prague, Czech Republic
e-mail: pawlas@karlin.mff.cuni.cz, marketa.zikmundova@vscht.cz
(Received January 14, 2018; revised May 25, 2019; accepted May 27, 2019)

ABSTRACT

We study nonparametric estimation of the length distribution for stationary line segment processes in the d-dimensional Euclidean space. Several methods have been proposed in the literature. We review different approaches (Horvitz-Thompson type estimator, reduced-sample estimator, Kaplan-Meier estimator, nonparametric maximum likelihood estimator, stochastic restoration estimation) and compare the finite sample behaviour by means of a simulation study for stationary line segment processes in 2D and 3D. Several data generating processes (Poisson point process, Matérn cluster process and Matérn hard-core process II) are considered with both independent and dependent segments. Our finite sample comparison reveals that the nonparametric likelihood estimator provides the most preferable method which works reasonably also if its assumptions are not satisfied.

Keywords: Horvitz-Thompson estimator, Kaplan-Meier estimator, line segment process, nonparametric maximum likelihood estimator, reduced-sample estimator, SRE algorithm.

INTRODUCTION

Germ-grain processes are one of the most important models in stochastic geometry. They are defined as marked point processes with the mark space formed by a family of nonempty compact sets, for details see Schneider and Weil (2008). We focus on the special case where the grains are line segments in the d-dimensional Euclidean space $\mathbb{R}^d$. Such germ-grain processes will be referred to as line segment processes. An important first order numerical characteristics of every stationary line segment process is its length intensity (mean total length of segments per unit volume). Different nonparametric unbiased length intensity estimators were compared in Pawlas and Honzl (2010). The aim of this paper is to study nonparametric estimators of segment length distribution. This problem is of interest for applications in several areas. Here, we mention three examples. In geology it is relevant to study geological faults, the data example analyzed in Laslett (1982a) comes from a sedimentary rock environment in Zambia. An application in forestry is contained in Svensson *et al.* (2006) where the authors find length distribution of standing trees from a sample area in northern Sweden. Kuhlmann and Redenbach (2015) estimate length distribution in fibre-reinforced composite materials providing an example of microstructure characteristics investigated in material science.

We assume that a single realization of a stationary line segment process is available within a bounded observation window. We assume that the individual segments can be identified. The difficulties arise due to the edge effects. Their role in the analysis of spatial processes is well clarified in Baddeley (1999). There are different strategies how to deal with edge effects. In the problem of estimating the distribution of segment lengths, edge effects were treated already in Laslett (1982a). An optimal estimator in the sense of maximizing the likelihood function was found in Wijers (1997) for a stationary planar Poisson line segment process observed through a convex window. This estimator is the nonparametric maximum likelihood estimator (NPMLE).

Sometimes the segments are not fully observed within the observation window. If the irregular part of the window is covered, then only several pieces of the segments are observed. In this case van Zwet (2004) derives the NPMLE of the length distribution. In some applications the determination of individual segment lengths could be very demanding. Kuhlmann and Redenbach (2015) consider a line segment process in $d = 3$ and propose an estimator of the length distribution based on segment endpoints only.

Another approach for estimating the parameters of a line segment process is studied in Chadœuf *et al.* (2000). The censored segments are treated as incomplete data and an iterative Monte Carlo procedure is considered. It is based on the iteration of two steps, restoration of the unobserved parts of segments and updating of estimates. This procedure is called stochastic restoration estimation (SRE).

In this paper we briefly introduce several nonparametric estimators of line segment length distribution function. We compare their finite sample properties through a Monte Carlo simulation study on different processes generating planar and spatial segment patterns. The performance is measured by the Kolmogorov–Smirnov and Cramér–von Mises statistics (Stephens, 1992). Our aim is to find out whether simple and natural methods (Horvitz-Thompson type estimator, reduced-sample estimator, Kaplan-Meier type estimator) can compete with computationally more demanding procedures (NPMLE, SRE). The former three estimators were chosen as examples of usual way how the standard empirical distribution is modified to handle the edge effects. The latter two estimators (NPMLE, SRE) were chosen as examples of more advanced estimators proposed in the literature, they require computation of many iterations of the algorithm.

The finite sample performance is tested on three basic types of point patterns (complete spatial randomness, regularity and clustering), the segment directions are attached independently and for the segment lengths we assume either independence or certain correlation structure. We consider only the most common cases for the dimension, *i.e.*, $d = 2$ and $d = 3$.

The paper is organized as follows. First we introduce line segment processes. Then we define different nonparametric estimators of typical length distribution of stationary line segment processes. Their quality is compared by an extensive Monte Carlo study.

Let $\mathscr{S}$ be the system of all nondegenerate line segments in $\mathbb{R}^d$. Each segment $S \in \mathscr{S}$ can be uniquely represented by its reference point $c(S)$, positive length $L(S)$ and direction $\theta(S) \in \mathscr{L}_1$, where $\mathscr{L}_1$ is the space of one-dimensional linear subspaces in $\mathbb{R}^d$. We require that the mapping $c : \mathscr{S} \to \mathbb{R}^d$ is measurable and equivariant under translations, *i.e.*, $c(S+z) = c(S)+z$ for all $S \in \mathscr{S}$ and $z \in \mathbb{R}^d$. Note that c is called a center function in Schneider and Weil (2008). We always choose $c(S)$ as either lexicographic minimum or lexicographic maximum point and by $e(S)$ we denote the other endpoint of S, that is distinct from $c(S)$. Let $\mathscr{S}_0 = \{S \in \mathscr{S} : c(S) = o\}$ be the set of segments with the reference point at the origin $o \in \mathbb{R}^d$. This space is isomorphic to the space $(0,\infty) \times \mathscr{L}_1$.

We can view a *line segment process* as a special case of germ-grain process, see Heinrich and Pawlas (2008) or Schneider and Weil (2008),

$$\Phi = \{X_i + \Xi_i, i \geq 1\} .$$

The points $\{X_i, i \geq 1\}$ create a point process in $\mathbb{R}^d$ and the grains Ξ_i are random line segments with values in $\mathscr{S}_0$, *i.e.*, the point X_i serves as a reference point of the segment $X_i + \Xi_i$. Fig. 1 shows two realizations of different models for Φ, observed through a square window in $\mathbb{R}^2$.

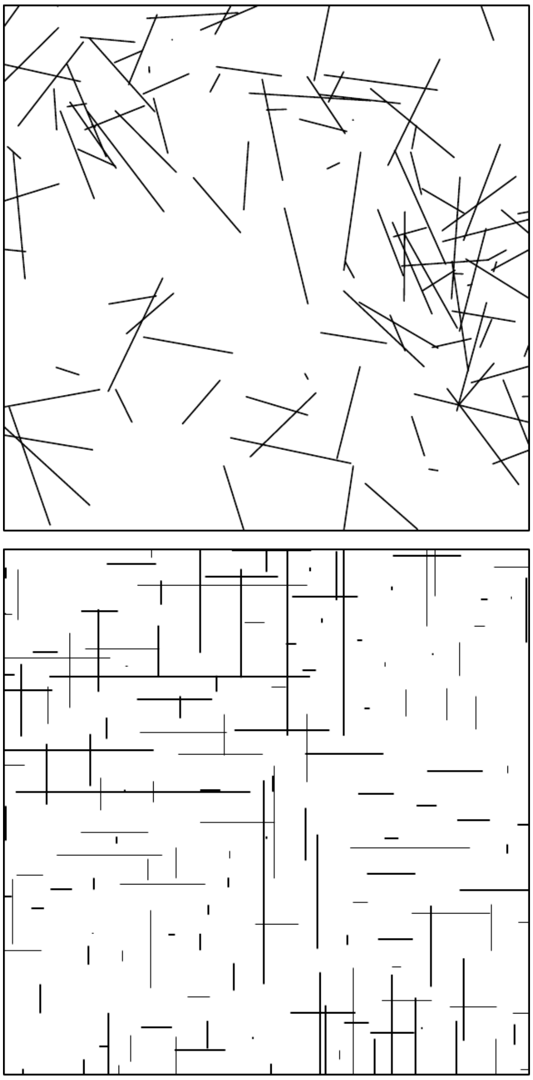

Fig. 1. *Two illustrations of realizations of planar line segment processes.*

If Φ is a stationary line segment process with intensity λ, then there exists a probability distribution $\mathbb{Q}$ (so called *typical segment distribution*) on $\mathscr{S}_0$ such that

$$\mathbb{E}\sum_{i\geq 1} f(X_i + \Xi_i) = \lambda \int_{\mathbb{R}^d}\int_{\mathscr{S}_0} f(x+S)\,\mathbb{Q}(\mathrm{d}S)\,\mathrm{d}x\,, \quad (1)$$

where f is an arbitrary nonnegative measurable function on $\mathscr{S}$. With only a slight abuse of notation, we write $\mathbb{Q}$ also for the image of $\mathbb{Q}$ under the isomorphism between $\mathscr{S}_0$ and $(0,\infty)\times\mathscr{L}_1$. Let $\mathscr{D}(\cdot)=\mathbb{Q}(\cdot\times\mathscr{L}_1)$ and $\rho=\mathbb{Q}((0,\infty)\times\cdot)$ be the distributions of typical length and typical direction, respectively. The distributions $\mathscr{D}$ and ρ need not to be independent. In what follows we call $\mathscr{D}$ the length distribution. It is given by the cumulative distribution function

$$F(t)=\mathbb{Q}(\{S:L(S)\le t\})=\mathscr{D}([0,t])\,,\quad t>0.$$

This paper deals with the estimation of F. We consider five existing nonparametric estimation methods.

If a line segment process Φ is defined by an independently marked point process (Illian *et al.*, 2008, Section 5.1.3), *i.e.*, $\{\Xi_i\}$ is a sequence of independent and identically distributed (= i.i.d.) random segments, independent of $\{X_i\}$, then it is called an *independent line segment process*. A *Poisson line segment process* $\Phi=\{X_i+\Xi_i, i\ge 1\}$ is an independent line segment process such that the germ process $\{X_i, i\ge 1\}$ is the Poisson point process (Illian *et al.*, 2008, Chapter 2) in $\mathbb{R}^d$. It follows from Theorem 3.5.7 in Schneider and Weil (2008) that the Poisson segment process Φ is the Poisson process in $\mathscr{S}$ (this space is isomorphic to $\mathbb{R}^d\times\mathscr{S}_0$).

A tractable class of models allowing dependencies among segments is obtained by geostatistical (or external) marking, see Illian *et al.* (2008, Section 5.1.3) for the definition in context of marked point processes. Let $\{\Xi(x):x\in\mathbb{R}^d\}$ be a stationary random field with values in $\mathscr{S}_0$, independent of $\{X_i\}$. We say that $\{X_i+\Xi(X_i), i\ge 1\}$ is a *geostatistically marked line segment process*. The typical segment distribution $\mathbb{Q}$ coincides with the distribution of $\Xi(o)$.

MATERIAL AND METHODS

Usually we observe a single realization of the line segment process Φ through a compact and convex window $W\subseteq\mathbb{R}^d$. Thus, the estimation is hampered by edge effects which introduce spatial sampling bias. For example, if we consider all segments which intersect W (with their true lengths), the final estimator of the length distribution will be biased because longer segments have a greater chance to be included in the sample. On the other hand, if we consider just those segments which are completely inside the window, segments longer than the diameter of W cannot be sampled.

We can divide the line segments hitting W into four groups: $\mathscr{Y}_0=\{i:X_i+\Xi_i\subseteq W\}$, $\mathscr{Y}_1=\{i:X_i\in W, e(X_i+\Xi_i)\notin W\}$, $\mathscr{Y}_2=\{i:X_i\notin W, e(X_i+\Xi_i)\in W\}$, $\mathscr{Y}_3=\{i:X_i\notin W, e(X_i+\Xi_i)\notin W, (X_i+\Xi_i)\cap W\neq\emptyset\}$. Only $\mathscr{Y}_0$ provides complete information about segment lengths, in other cases the segments are not totally observed. Directions $\theta(\Xi_i)$ are observable for all line segments hitting W. Segments corresponding to $\mathscr{Y}_0$ are *uncensored*, segments from $\mathscr{Y}_1$ and $\mathscr{Y}_2$ may be called *single end censored* and segments from $\mathscr{Y}_3$ are called *double censored*, see also Wijers (1997), p. 6.

Horvitz-Thompson type estimator

The sampling bias which is the result of the edge effects can be corrected by changing the sampling rule or by an appropriate weighting of the observations. This leads us to the *Horvitz-Thompson type estimator*,

$$\widehat{F}_{\mathrm{HT}}(t)=\frac{1}{\widehat{\lambda}_{\mathrm{HT}}}\sum_{i:X_i+\Xi_i\in\mathrm{sample}}\frac{1}{\tau(\Xi_i)}\mathbf{1}\{L(\Xi_i)\le t\}\,,$$

where

$$\widehat{\lambda}_{\mathrm{HT}}=\sum_{i:X_i+\Xi_i\in\mathrm{sample}}\frac{1}{\tau(\Xi_i)}\,,$$

and τ is a suitable weighting function, *i.e.*, $\tau(\Xi_i)=\int\mathbf{1}\{x+\Xi_i\in\mathrm{sample}\}\,\mathrm{d}x$, see Baddeley (1999). As a consequence of Eq. 1, $\widehat{\lambda}_{\mathrm{HT}}\widehat{F}_{\mathrm{HT}}(t)$ is an unbiased estimator of $\lambda F(t)$. Thus, $\widehat{F}_{\mathrm{HT}}(t)$ is a ratio of two random variables so that the ratio of their expectations is $F(t)$. It means that $\widehat{F}_{\mathrm{HT}}(t)$ is so-called ratio-unbiased estimator of $F(t)$. We will consider following three basic sampling rules:

1) *minus sampling* – the sample consists of fully observable segments ($X_i+\Xi_i$ is sampled if and only if $X_i+\Xi_i\subseteq W$),
2) *unbiased sampling* – the sample consists of segments with reference point inside the window ($X_i+\Xi_i$ is sampled if and only if $X_i\in W$),
3) *plus sampling* – the sample consists of all segments hitting the window ($X_i+\Xi_i$ is sampled if and only if $(X_i+\Xi_i)\cap W\neq\emptyset$).

It means that minus sampling uses $\mathscr{Y}_0$, unbiased sampling $\mathscr{Y}_0\cup\mathscr{Y}_1$, and plus sampling $\mathscr{Y}_0\cup\mathscr{Y}_1\cup\mathscr{Y}_2\cup\mathscr{Y}_3$. The weights $\tau(\Xi_i)$ become $|W\ominus\Xi_i|$ for minus sampling, $|W|$ for unbiased sampling, and $|W\oplus\Xi_i|$ for plus sampling. Here, $|B|$ is the d-dimensional Lebesgue measure of the set B, $B\ominus S_0=\{x:x+S_0\subseteq B\}$ is the erosion of B by the line segment S_0 and $B\oplus S_0=\{x:(x+S_0)\cap B\neq\emptyset\}$ is the dilation of B by the line segment S_0. When applying unbiased sampling or plus sampling rule, we need also some information outside the sampling window W in order to determine $\widehat{F}_{\mathrm{HT}}$. For independent line segment processes, asymptotic properties of $\widehat{F}_{\mathrm{HT}}$ (as the window W increases) follow from the results of Heinrich and Pawlas (2008).

Reduced-sample estimator

For $S \in \mathscr{S}_0$ and $t > 0$, let $\tilde{S}^{(t)} = \frac{t}{L(S)} S$ be the line segment in the same direction as S with length t. When estimating $F(t)$, a simple approach is to reduce the sample and consider only those pairs $(X_i, \theta(\Xi_i))$ for which the line segment $X_i + \tilde{\Xi}_i^{(t)}$ with reference point X_i, direction $\theta(\Xi_i)$ and length t would lie completely inside the window W. Then the *reduced-sample estimator* of $F(t)$ can be defined by

$$\widehat{F}_{\mathrm{rs}}(t) = \frac{\sum_{i:X_i \in W} \mathbf{1}\{L(\Xi_i) \le t, X_i + \tilde{\Xi}_i^{(t)} \subseteq W\}}{\sum_{i:X_i \in W} \mathbf{1}\{X_i + \tilde{\Xi}_i^{(t)} \subseteq W\}}, \quad t > 0. \tag{2}$$

This estimator takes into account only segments from $\mathscr{Y}_0$ and $\mathscr{Y}_1$. Moreover, only the length of the visible part is required for segments from $\mathscr{Y}_1$.

Since from Eq. 1,

$$\mathbb{E} \sum_{i:X_i \in W} \mathbf{1}\{L(\Xi_i) \le t, X_i + \tilde{\Xi}_i^{(t)} \subseteq W\} = \lambda \int_{\mathscr{S}_0} |W \ominus \tilde{S}^{(t)}| \mathbf{1}\{L(S) \le t\} \mathbb{Q}(\mathrm{d}S),$$

the reduced-sample estimator is ratio-unbiased provided that $\mathbb{Q} = \mathscr{D} \times \rho$. For larger t, it may discard a lot of information given by data. Note that it is not necessarily nondecreasing. The estimators of this type are often used in spatial statistics in order to deal with edge effects caused by the bounded observation window, see, *e.g.* Baddeley (1999). This approach is sometimes also called the *border method* for edge correction.

Kaplan-Meier estimator

Random censoring and survival theory provide us another look at the edge effects. Let $\{T_i\}$ be i.i.d. positive random variables (survival data) with distribution function H. Instead of them we observe only censored data $T_i' = \min(T_i, C_i)$ and indicators of non-censoring $D_i = \mathbf{1}\{T_i < C_i\}$. If we assume that $\{C_i\}$ are i.i.d. random variables, independent of $\{T_i\}$, then the product-limit estimator defined as

$$\widehat{H}(t) = 1 - \prod_{s \le t} \left(1 - \frac{\sum_{i \ge 1} \mathbf{1}\{T_i' = s, D_i = 1\}}{\sum_{i \ge 1} \mathbf{1}\{T_i' \ge s\}} \right)$$

is the nonparametric maximum likelihood estimate of $H(t)$. It is known as the Kaplan-Meier estimator (Kaplan and Meier, 1958). In our context, $T_i = L(\Xi_i)$ is the true segment length and C_i is the distance from the reference point X_i to the boundary of W in direction $\theta(\Xi_i)$ of the line segment $X_i + \Xi_i$. It means that the line segment $X_i + \Xi_i$ is not censored (*i.e.*, $D_i = 1$) if $X_i + \Xi_i \subseteq W$ ($i \in \mathscr{Y}_0$). To avoid the sampling bias, let us consider only those segments with reference point inside the window W (*i.e.*, $i \in \mathscr{Y}_0 \cup \mathscr{Y}_1$). Then the *Kaplan-Meier estimator* of F is given by

$$\widehat{F}_{\mathrm{KM}}(t) = 1 - \prod_{s \le t} \left(1 - \frac{\sum_{i \ge 1} \mathbf{1}\{X_i \in W, L(\Xi_i) = s, X_i + \Xi_i \subseteq W\}}{\sum_{i \ge 1} \mathbf{1}\{X_i \in W, L((X_i + \Xi_i) \cap W) \ge s\}} \right).$$

This estimator for general germ-grain processes was introduced in Pawlas (2006). A related estimator is used in Laslett (1982a) for line segment processes. Since the independence assumptions are no longer satisfied, the optimality (in the sense of nonparametric maximum likelihood) of the Kaplan-Meier estimator is destroyed in our setting. An analogous situation happens in Baddeley and Gill (1997) where the empty space function and the nearest neighbour distance distribution function of spatial point processes are estimated. Baddeley and Gill (1997) show that the Kaplan-Meier technique provides reasonable estimators. Similarly, $\widehat{F}_{\mathrm{KM}}(t)$ should yield an estimator of $F(t)$ that is more efficient than the reduced-sample estimator $\widehat{F}_{\mathrm{rs}}(t)$.

Nonparametric maximum likelihood estimator

A natural question is how the NPMLE of $F(t)$ looks like. Laslett (1982b) noted that it is not the Kaplan-Meier estimator and proposed a method of estimating the length distribution for stationary Poisson segment processes in $\mathbb{R}^2$. The NPMLE of the length distribution in a stationary planar Poisson line segment process was found by Wijers (1997) as the solution of the self-consistency equations that can be solved numerically by the expectation-maximization (EM) algorithm. We shortly describe the procedure for the case $W = [0, a]^2$, for details see Wijers (1997). Define

$$V(t) = \int_0^t \frac{a^2 + auh(\rho)}{a^2 + a\mu h(\rho)} \mathrm{d}F(u), \quad t > 0,$$

where $\mu = \int_0^\infty t \,\mathscr{D}(\mathrm{d}t)$ is the mean length of a typical segment and $h(\rho) = \int_{-\pi/2}^{\pi/2} (\cos\theta + |\sin\theta|) \rho(\mathrm{d}\theta)$. Here, we identify $\mathscr{L}_1$ with $(-\pi/2, \pi/2]$. In an isotropic case (ρ is the uniform distribution), $h(\rho) = \frac{4}{\pi}$. Let $n = \sum_i \mathbf{1}\{(X_i + \Xi_i) \cap W \neq \emptyset\}$ be the number of observations (number of line segments hitting W). We introduce the empirical subdistribution functions

$$F_n^{(0)}(t) = \frac{1}{n} \sum_{i:i\in\mathscr{Y}_0} \mathbf{1}\{L(\Xi_i) \le t\},$$
$$F_n^{(1,2)}(t) = \frac{1}{n} \sum_{i:i\in\mathscr{Y}_1\cup\mathscr{Y}_2} \mathbf{1}\{L((X_i+\Xi_i)\cap W) \le t\},$$
$$F_n^{(3)}(t) = \frac{1}{n} \sum_{i:i\in\mathscr{Y}_3} \mathbf{1}\{L((X_i+\Xi_i)\cap W) \le t\},$$

corresponding to uncensored, single end censored and double censored observations, respectively. The NPMLE $\hat{V}_n$ of reparametrization V satisfies the self-consistency equation

$$\begin{aligned} \mathrm{d}\hat{V}_n(t) = {} & \mathrm{d}F_n^{(0)}(t) \\ & + \int_0^t \frac{1}{\hat{g}_n(u)} \mathrm{d}F_n^{(1,2)}(u) \frac{1}{a^2+ath(\rho)} \mathrm{d}\hat{V}_n(t) \\ & + \int_0^t \frac{t-u}{\hat{d}_n(u,u)} \mathrm{d}F_n^{(3)}(u) \frac{1}{a^2+ath(\rho)} \mathrm{d}\hat{V}_n(t), \end{aligned} \tag{3}$$

where

$$\hat{g}_n(u) = \int_u^\infty \frac{1}{a^2+awh(\rho)} \mathrm{d}\hat{V}_n(w),$$
$$\hat{d}_n(u,u) = \int_u^\infty \frac{w-u}{a^2+awh(\rho)} \mathrm{d}\hat{V}_n(w).$$

If ρ is assumed to be known, then a solution of Eq. 3 can be found using the EM algorithm. We start with an initial estimator $\hat{V}_n^0$ which sets positive mass to all observation points. The iterative scheme of the EM algorithm is obtained by replacing $\hat{V}_n$ with $\hat{V}_n^{k+1}$ on the left hand side of Eq. 3 and by replacing $\hat{V}_n$ with $\hat{V}_n^k$ on the right hand side of Eq. 3. Now suppose that the distribution ρ of segment directions is unknown. The NPMLE $\hat{\rho}_n$ of ρ for given F can be expressed as

$$\hat{\rho}_n(\eta) = \frac{\sum_{i=1}^n (a^2+a\mu(\cos\theta(\Xi_i)+|\sin\theta(\Xi_i)|))^{-1}\mathbf{1}\{\theta(\Xi_i)\le\eta\}}{\sum_{i=1}^n (a^2+a\mu(\cos\theta(\Xi_i)+|\sin\theta(\Xi_i)|))^{-1}}. \tag{4}$$

This leads us to a natural iterative scheme. For given $\hat{\rho}_n^k$ we determine $\hat{F}_n^{k+1}$ using the step of the EM algorithm described above and for given $\hat{F}_n^{k+1}$ we determine $\hat{\rho}_n^{k+1}$ from Eq. 4. The EM algorithm for estimation of the length distribution is also used in Svensson *et al.* (2006). We stress that Eq. 3 and Eq. 4 are derived specifically for $d=2$.

Stochastic restoration estimation

Besides nonparametric estimators, one can also consider a parametric approach, where the length distribution is known up to a vector of unknown parameters. Parametric estimation procedures were studied, for example, by Chadœuf *et al.* (2000) who propose so called *stochastic restoration estimation (SRE) algorithm.* As it is mentioned in Chadœuf *et al.* (2000), this Monte Carlo algorithm can be applied also in the nonparametric setting. In order to avoid the sampling bias, we again take into account only the line segments with reference point within W. Denote their number by $m = |\mathscr{Y}_0| + |\mathscr{Y}_1|$ and consider the empirical distribution function of observable lengths given by

$$\hat{F}_0(t) = F_m^{(0,1)}(t) = \frac{1}{m} \sum_{i:X_i\in W} \mathbf{1}\{L((X_i+\Xi_i)\cap W) \le t\}\,.$$

At iteration p two main steps are performed. The first step (R-step) is the restoration of lengths of censored segments ($i \in \mathscr{Y}_1$). It is made by simulation from the conditional distribution with current estimate $\hat{F}_p$, we obtain lengths $\tilde{L}_i$. In the second step (E-step) the estimate $\hat{F}$ is updated by taking $\hat{F}_{p+1}$ as the empirical distribution function of uncensored segment lengths $L(\Xi_i)$, $i \in \mathscr{Y}_0$, and restored lengths $\tilde{L}_i$, $i \in \mathscr{Y}_1$. The result of this algorithm is a homogeneous Markov chain $\{\hat{F}_p\}$. In practice, we define the SRE estimator as $\hat{F}_p$ for some prescribed large p.

Simulations

A simulation study is conducted to compare the behaviour of individual estimators. Simulations and computations are performed using R (R Core Team, 2018) and its contributed package *spatstat* (Baddeley and Turner, 2005). We generate six types of stationary line segment processes in $\mathbb{R}^d$ (with $d=2$ or $d=3$). The point process of reference points $\{X_i\}$ has intensity $\lambda > 0$ and is chosen as one of the following three processes:

1) Poisson point process,

2) Matérn cluster process with mean number of points per cluster $\mu = 5$ and radius of clusters $R = 0.1$,

3) Matérn hard-core process II with hard-core distance $h = 0.05(d-1)$.

These three processes provide models for random, clustered and regular patterns, respectively. For the definition of Matérn cluster process we refer to Illian *et al.* (2008, Section 6.3.2) or Schneider and Weil (2008, p. 93). The definition of Matérn hard-core process II can be found in Illian *et al.* (2008, Section 6.5.2) or (Schneider and Weil, 2008, p. 94). Once we have a pattern of reference points, the segments are attached according to either independent or geostatistical marking. The length and direction of typical segment are assumed to be independent, *i.e.*, $\mathbb{Q} = \mathscr{D} \times \rho$. First we consider the isotropic case – directions are uniformly distributed. Then we keep the locations of reference points and lengths of segments unchanged and only change the directions to get an anisotropic segment process. In particular, the uniform distribution on d perpendicular directions parallel to the axes is used, *i.e.*, each canonical direction with probability $1/d$. This procedure gives us realizations of six segment processes, which are denoted by 1i, 1a, 2i, 2a, 3i, 3a, where the numbers stand for Poisson (1), clustered (2), and regular (3) pattern while the letters stand for isotropic case (i) and anisotropic case (a). Furthermore, in order to have dependent lengths we use geostatistical marking as follows: we consider a Gaussian random field $\{Z(x)\}$ with zero mean and exponential covariance function $\operatorname{cov}(Z(x),Z(y)) = \mathrm{e}^{-4\|x-y\|}$. We set $L(\Xi_i) = 0.25\Phi(Z(X_i))$, where Φ is the cumulative distribution function of the standard normal distribution. It means that to the point X_i we assign a segment of length $L(\Xi_i)$ and direction that is independent of $\{Z(x)\}$ and $\{X_i\}$, *i.e.*, the geostatistical marking is only applied to the length while the independent marking is applied to the direction. The length distribution is then uniform on $(0,0.25)$.

The procedure is repeated 10 000 times for each simulation experiment. We study the influence of intensity λ (values 25, 50, 75, 100 and 125 are used), length distribution (either uniform on $(0,0.25)$ or exponential with mean 0.125) and dimension ($d = 2$ or $d = 3$).

We observe a realization of each process in the unit square or unit cube window $W = [0,1]^d$. However, also the information about all segments hitting W is recorded so that we can evaluate estimators based on plus sampling as well. Fig. 1 shows two examples of segment processes, a realization of the process 2i with geostatistically marked uniform lengths (left) and a realization of the process 3a with independently marked exponential lengths (right).

For every simulated segment pattern we determine all mentioned estimators:

(a) Horvitz-Thompson type estimator $\widehat{F}_{\mathrm{HT}}$ using minus sampling,

(b) Horvitz-Thompson type estimator $\widehat{F}_{\mathrm{HT}}$ using unbiased sampling,

(c) Horvitz-Thompson type estimator $\widehat{F}_{\mathrm{HT}}$ using plus sampling,

(d) reduced-sample estimator $\widehat{F}_{\mathrm{rs}}$,

(e) Kaplan-Meier estimator $\widehat{F}_{\mathrm{KM}}$,

(f) nonparametric maximum likelihood estimator assuming that the directions are uniformly distributed (only for $d = 2$),

(g) nonparametric maximum likelihood estimator for unknown distribution of directions (only for $d = 2$),

(h) estimator obtained after $p = 100000$ steps of stochastic restoration estimation algorithm.

The estimators (b), (d), (e) and (h) depend on the choice of a reference point. Two natural choices are lexicographic minimum point $c(S)$ and lexicographic maximum point $e(S)$. Thus, for each type of estimator we can obtain the estimators $\hat{F}_c$ and $\hat{F}_e$ corresponding to these choices as a reference point. We do not consider both estimators separately but we improve them by taking the average $\frac{1}{2}(\hat{F}_c + \hat{F}_e)$. Since the estimators (b) and (c) require information from outside W, we include in the comparison with other estimators only the minus sampling estimator (a). The estimators (f) and (g) were almost identical in all our experiments. So in what follows we only deal with (g). We run 10^6 iterations of the EM algorithm to evaluate this estimator.

To measure the quality of the estimator $\widehat{F}$ we use two criterion functions that are used for goodness of fit tests (Stephens, 1992), the Kolmogorov–Smirnov statistic

$$d_{KS}(\widehat{F},F) = \sup_{t\in\mathbb{R}^+} |\widehat{F}(t) - F(t)| ,$$

and the Cramér–von Mises statistic

$$d_{CvM}(\widehat{F},F) = \int_0^\infty (\widehat{F}(t) - F(t))^2 \,\mathrm{d}F(t) .$$

These deviation measures are computed for each simulated realization of line segment process for all estimators. In the forthcoming figures we present their sample means obtained from 10 000 repetitions for each process and each experiment.

RESULTS

First we consider the case of independent line segment process in $\mathbb{R}^2$. The best performance in most scenarios was found for NPMLE. Only for the smallest intensity ($\lambda = 25$) and uniform length distribution it was outperformed by the reduced-sample estimator. For $\lambda = 50$ and uniform lengths, $\widehat{F}_{\mathrm{rs}}$ has slightly smaller values of mean d_{CvM} for some of the models. The results of the comparison are shown in Fig. 2 (left for $\lambda = 50$ and uniform length distribution and right for $\lambda = 100$ and exponential length distribution). The estimation error is lower for random and regular patterns if the lengths were uniformly distributed while for exponential length distribution clustered

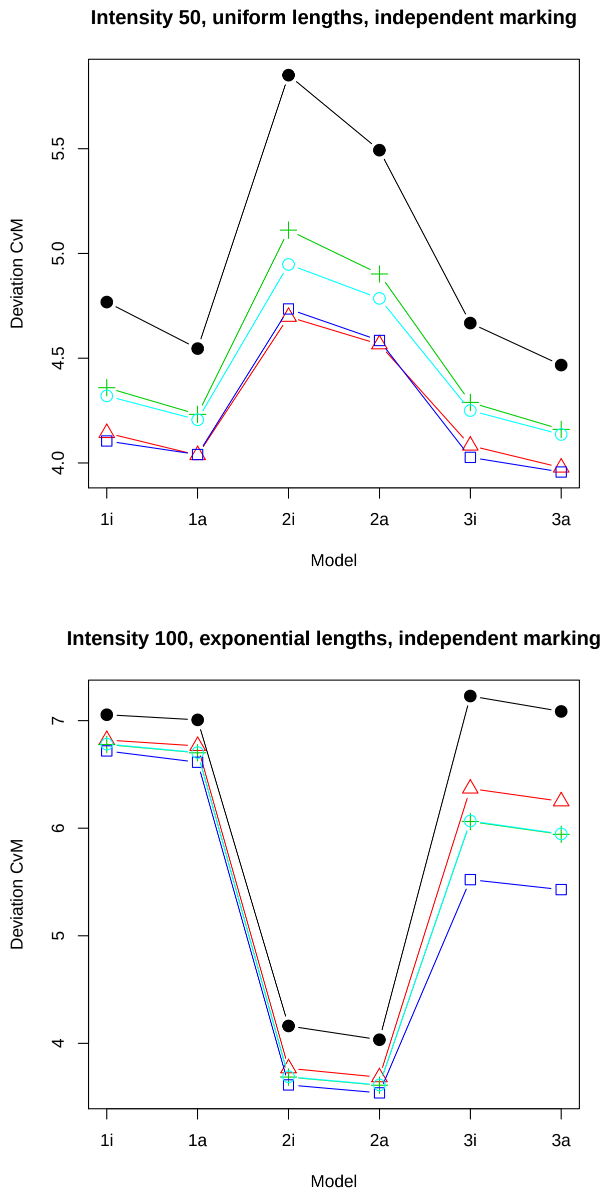

Fig. 2. *The values of* $1000 \cdot d_{CvM}$ *for six considered models in case of independent marking,* $d = 2$*, uniformly distributed lengths and* $\lambda = 50$ *(top) and exponentially distributed lengths and* $\lambda = 100$ *(bottom). Horvitz-Thompson (black, bullets), reduced-sample (red, triangles), Kaplan-Meier (green, crosses), NPMLE (blue, squares), and SRE (cyan, circles) estimators are compared.*

patterns lead to lower values of d_{KS} and d_{CvM}. Since the NPMLE was derived under the assumption of Poisson segment process, it is not surprising that it has the smallest deviation from the true distribution function in that case. Under independent marking, we observed that the influence of underlying configuration of reference points is quite negligible. The NPMLE works very well also for clustered and regular patterns. Comparing the remaining estimators, the reduced-sample estimator performed well for uniform length and smaller intensities while in other cases (uniform length and larger intensity, exponential length and arbitrary intensity) SRE and Kaplan-Meier estimator were better. Both these estimators gave very similar values, in particular for exponential lengths. The Horvitz-Thompson type estimator has the poorest behaviour among all studied estimators.

Similar conclusions could be made also for geostatistically marked line segment process. In this case the segments are dependent. However, the NPMLE still resulted in the lowest mean deviation measures. SRE and Kaplan-Meier estimator behave very similarly. The reduced-sample estimator was to some degree worse than in independent case. In some situations (especially with larger intensity) it was even beaten by the Horvitz-Thompson type estimator. The comparison for two different intensities is depicted in Fig. 3.

Finally, we have investigated the estimation for independent line segment processes in $\mathbb{R}^3$. Since the calculation of NPMLE is designed only for the planar case, it was not taken into account. The reduced-sample estimator shows the best behaviour for smaller intensities. For larger intensities reduced-sample estimator, Kaplan-Meier estimator, and SRE provide comparable results. The Horvitz-Thompson type estimator has again the largest mean deviation from true distribution function. Fig. 4 shows the comparison of results for two different intensities and uniformly distributed lengths.

Minus sampling version of the Horvitz-Thompson estimator disregards information given by partially observed segments. More information is used in unbiased sampling or plus sampling. Obviously, the corresponding estimators are more precise. We compare them in Fig. 5 where we consider independent line segment processes of intensity 75 and with uniformly distributed lengths. For unbiased sampling two different estimators could be distinguished depending on the choice of reference point (lexicographic minimum or maximum point). We also present their average which improves the quality of individual estimators. As we already noticed, averaging of the estimators based on two different endpoints is used also for estimators (d), (e) and (h) from the list above.

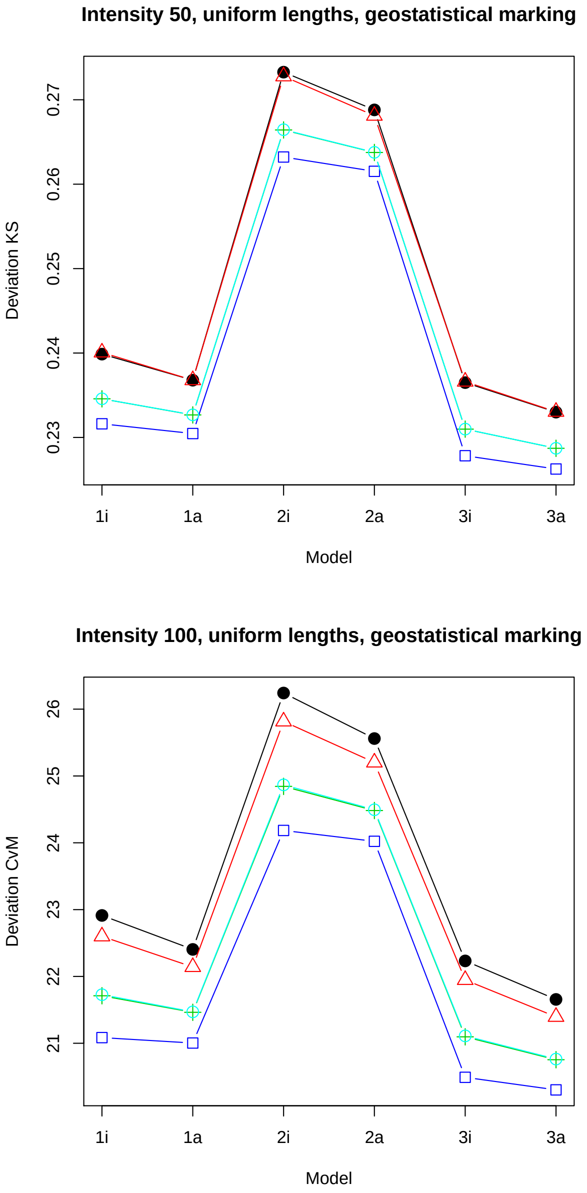

Fig. 3. *The comparison of results for models with geostatistical marking and uniformly distributed lengths. The values of* d_{KS} *(top) and* $1000 \cdot d_{CvM}$ *(bottom) are shown for six considered models with* $\lambda = 50$ *(left) and* $\lambda = 100$ *(right). Horvitz-Thompson (black, bullets), reduced-sample (red, triangles), Kaplan-Meier (green, crosses), NPMLE (blue, squares) and SRE (cyan, circles) estimators are considered.*

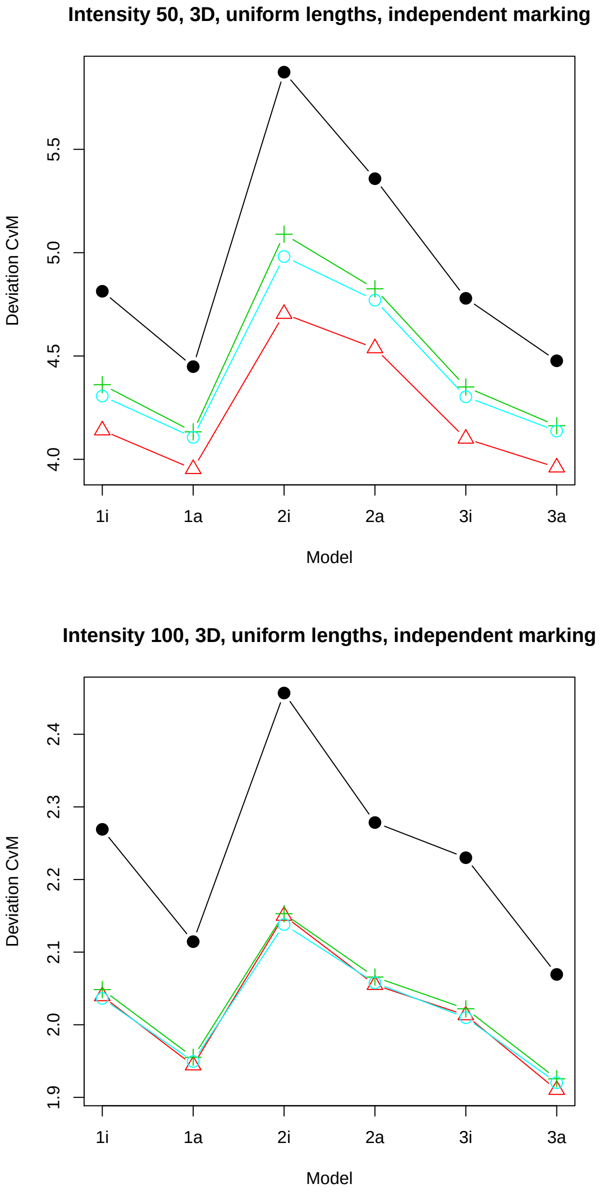

Fig. 4. *The values of* $1000 \cdot d_{CvM}$ *for six considered models in case of independent marking,* $d = 3$*, uniformly distributed lengths and* $\lambda = 50$ *(top) and* $\lambda = 100$ *(bottom). Horvitz-Thompson (black, bullets), reduced-sample (red, triangles), Kaplan-Meier (green, crosses), and SRE (cyan, circles) estimators are compared.*

With increasing intensity we have more data and the estimators are more accurate. It is demonstrated in Fig. 6 where independent marking and exponentially distributed lengths are considered. The underlying point process is Matérn hard-core process II.

DISCUSSION

We have reviewed several nonparametric estimators of the length distribution and compared their performance based on Monte Carlo experiments.

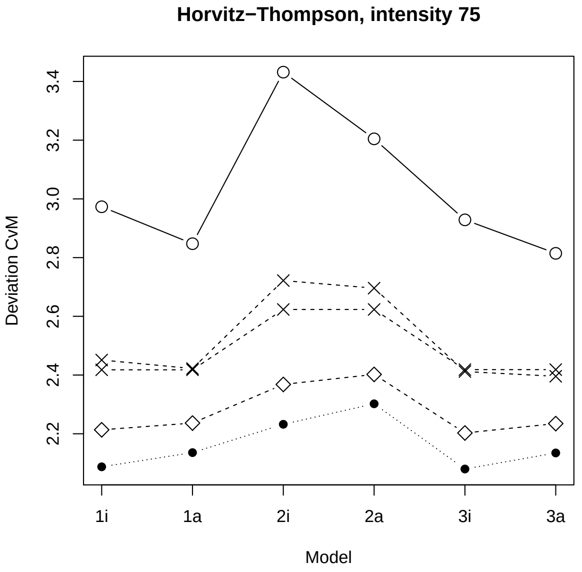

Fig. 5. *The values of* $1000 \cdot d_{CvM}$ *for six considered models in case of independent marking,* $d = 2$*, uniformly distributed lengths and* $\lambda = 75$*. Different Horvitz-Thompson estimators are compared: minus sampling (full line, circles), unbiased sampling (dashed line, crosses for individual estimators and rhombi for average), and plus sampling (dotted line, bullets).*

In the planar case, the nonparametric maximum likelihood estimator (NPMLE) is based on the assumption of Poisson process. Our simulation study revealed that this estimator is quite robust and preserves its superior behaviour also if the underlying point process is not Poisson and if the independence of segments is not satisfied. This estimator is computed using the EM algorithm that requires many iterations. However, its calculation is still quite fast (few seconds). Stochastic restoration estimation (SRE) requires an iterative numerical procedure as well. The computation time depends on the number of steps. In our experiments, the rate of convergence was quite good. The results for $p = 1\,000$ steps were almost the same as for $p = 100\,000$ steps. The computation time for $p = 1\,000$ was comparable with NPMLE (few seconds). For larger number of steps the calculation of SRE becomes more time demanding (few minutes). Our aim was to find out whether simpler estimators can compete with NPMLE and SRE. These simpler estimators are very easy to implement and they are computed almost immediately. The Kaplan-Meier estimator gave results which are comparable with SRE in most scenarios. The reduced-sample estimator also provides a simple and reasonable alternative. It worked particularly well for lower values of intensity. This estimator is less precise for larger t where more information from data is discarded. Therefore, it behaves worse for exponential length in comparison with uniform length. Furthermore, the reduced-sample estimator, Eq. 2, is not necessarily monotonic. In that case, we suggest to use a natural modification

$$\widehat{F}_{\mathrm{rs,m}} = \sup_{s \le t} \widehat{F}_{\mathrm{rs}}(s), \quad t > 0.$$

Horvitz-Thompson type estimator was the least efficient estimator in our simulation study. It was expected because it uses only uncensored segments, single end censored segments are ignored.

In conclusion, we can recommend both Kaplan-Meier and reduced-sample estimators when a computationally simple method is required. They are also convenient in higher dimensions since the equations for the NPMLE are derived for the planar case.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their constructive suggestions and comments. The work is supported by the Grant Agency of the Czech Republic, project 16-03708S.

REFERENCES

Baddeley AJ (1999). Spatial sampling and censoring. In: Barndorff-Nielsen OE, Kendall WS, van Lieshout MNM, eds. Stochastic Geometry: Likelihood and Computation. London: Chapman and Hall, 37–78.

Baddeley AJ, Gill RD (1997). Kaplan-Meier estimators of distance distributions for spatial point processes. Ann Statist 25:263–92.

Baddeley AJ, Turner R (2005). Spatstat: an R package for analyzing spatial point patterns. J Stat Softw 12:1–42.

Chadœuf J, Senoussi R, Yao JF (2000). Parametric estimation of a Boolean segment process with stochastic restoration estimation. J Comput Graph Statist 9:390–402.

Heinrich L, Pawlas Z (2008). Weak and strong convergence of empirical distribution functions from germ-grain processes. Statistics 42:49–65.

Illian J, Penttinen A, Stoyan H, Stoyan D (2008). Statistical Analysis and Modelling of Spatial Point Patterns. Chichester: John Wiley & Sons.

Kaplan EL, Meier P (1958). Nonparametric estimation from incomplete observations. J Amer Statist Assoc 53:457–81.

Kuhlmann M, Redenbach C (2015). Estimation of fibre length distributions from fibre endpoints. Scand J Statist 42:1010–22.

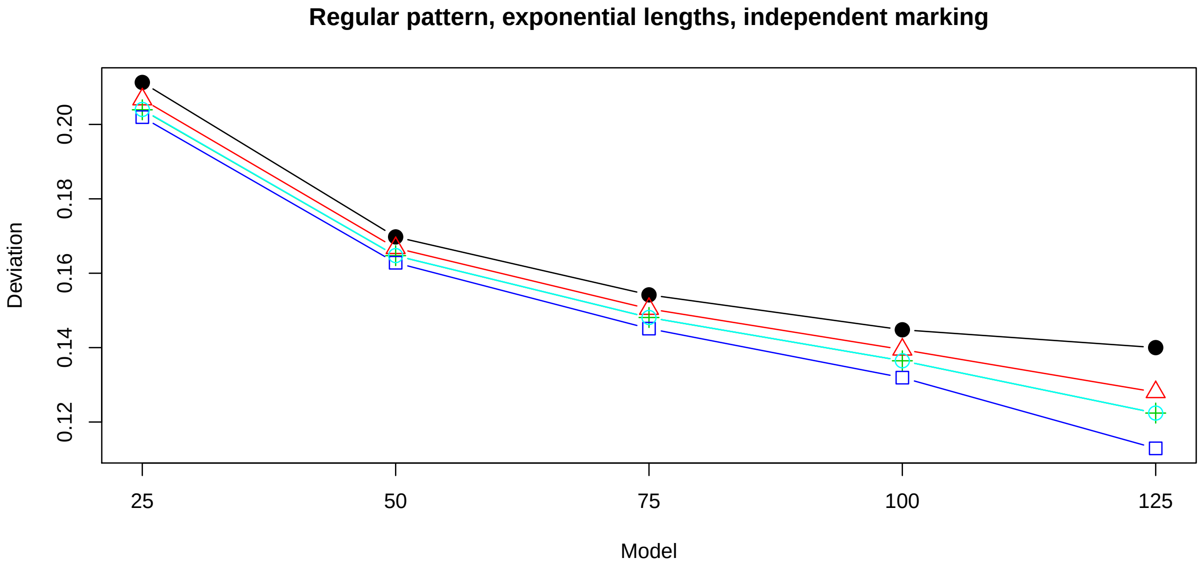

Fig. 6. *The values of d_{KS} for different choices of λ in the case of independent line segment process in the plane with regular pattern of reference points. Typical segment distribution is composed from two independent components: exponential length and isotropic direction. Horvitz-Thompson (black, bullets), reduced-sample (red, triangles), Kaplan-Meier (green, crosses), NPMLE (blue, squares), and SRE (cyan, circles) estimators are compared.*

Laslett GM (1982a). Censoring and edge effects in areal and line transect sampling of rock joint traces. Math Geol 14:125–40.

Laslett GM (1982b). The survival curve under monotone density constraints with applications to two-dimensional line segment processes. Biometrika 69:153–60.

Pawlas Z (2006). Estimation of the distribution function in germ-grain models. In: Hušková M, Janžura M, eds. Proceedings Prague Stochastics 2006. Prague: Matfyzpress, 579–89.

Pawlas Z, Honzl O (2010). Comparison of length-intensity estimators for segment processes. Statist Probab Lett 80:825–33.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schneider R, Weil W (2008). Stochastic and Integral Geometry. Berlin: Springer-Verlag.

Stephens MA (1992). Introduction to Kolmogorov (1933) On the empirical determination of a distribution. In: Kotz S, Johnson NL, eds. Breakthroughs in Statistics: Methodology and Distribution. New York: Springer-Verlag, 93–105.

Svensson I, Sjöstedt-de Luna S, Bondesson L (2006). Estimation of wood fibre length distributions from censored data through an EM algorithm. Scand J Statist 33:503–22.

van Zwet EW (2004). Laslett's line segment problem. Bernoulli 10:377–96.

Wijers BJ (1997). Nonparametric estimation for a windowed line-segment process. Amsterdam: Stichting Mathematisch Centrum.