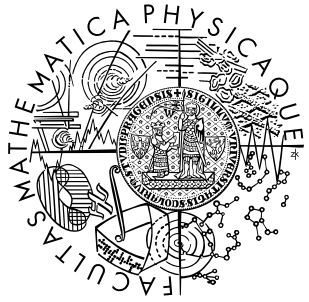
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Martin Schenk

Náhodné procesy indexované množinami

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.
Studijní program: Matematika
Studijní plán: Teorie pravděpodobnosti a náhodné procesy

Rád bych na tomto místě poděkoval vedoucímu práce RNDr. Zbyňku Pawlasovi, Ph.D. za neocenitelnou pomoc během celého procesu tvorby práce a za čas strávený diskusemi nad jejím obsahem. Za podporu během celého studia bych rád poděkoval mým rodičům a Zdeňce Linkové.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Berouně dne 14. dubna 2008 Martin Schenk

Název práce: Náhodné procesy indexované množinami
Autor: Martin Schenk
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.
e-mail vedoucího: pawlas@karlin.mff.cuni.cz

Abstrakt: Tato práce se zabývá řešením problému odhadu sdruženého pravděpodobnostního rozdělení parametrů značkovaného bodového procesu pro cenzorovaná data. Nejprve je uveden jednorozměrný případ, pro nějž je zkonstruován Nelsonův-Aalenův odhad kumulativní rizikové funkce, který je následně pomocí jádrové funkce vyhlazen. Následně je uveden Kaplanův-Meierův odhad funkce přežití. Je vybudována teorie náhodných procesů indexovaných množinami, na jejímž základě je zkonstruován zobecněný Nelsonův-Aalenův odhad kumulativní rizikové funkce, který je poté opět vyhlazen. Pro speciální případ je zkonstruován též zobecněný Kaplanův-Meierův odhad vícerozměrné funkce přežití. Na konkrétním příkladě jsou ukázány aplikace zmíněných zobecněných odhadů. Tyto odhady jsou poté použity na simulovaná data.

Klíčová slova: proces částic, riziková funkce, funkce přežití, Nelsonův-Aalenův odhad, Kaplanův-Meierův odhad

Title: Set-indexed random processes
Author: Martin Schenk
Department: Department of probability and mathematical statistics
Supervisor: RNDr. Zbyněk Pawlas, Ph.D.
Supervisor's e-mail address: pawlas@karlin.mff.cuni.cz

Abstract: This thesis deals with the problem of estimating the joint probability distribution of a marked process' parameters from a censored data. First, a Nelson-Aalen estimator of the cumulative hazard rate for one-dimensional case is constructed. This estimator is then smoothed by using a kernel function estimator. Then, a Kaplan-Meier estimator of the survival function is brought in. Further, a theory of set-indexed random processes is built up to be a base for the construction of a generalized Nelson-Aalen estimator of the cumulative hazard rate, which is then again smoothed. For a special case, a generalized Kaplan-Meier estimator of the multi-dimensional survival function is constructed. The application of the mentioned generalized estimators is shown on a particular case. These estimators are then used on simulated data.

Keywords: particle process, hazard rate, survival function, Nelson-Aalen estimator, Kaplan-Meier estimator

Contents

Introduction

Set-indexed random processes are a generalization of classic random processes in $\mathbb{R}^1$. This generalization is somehow a delicate point. The reason is as follows: if we have a "time-indexed" random process $\{X_t, t \geq 0\}$, we can for each X_t and X_s decide whether $t < s$ or vice-versa. On the other hand, for two general sets in $\mathbb{R}^d_+$ we cannot explicitly decide which one is "greater", because we do not have an unique ordering in $\mathbb{R}^d_+$. Therefore, we must develop a suitable structure of the indexing sets, which will allow us to work with set-indexed random processes in a way which will be as much as possible analogous to the one-dimensional case.

In this thesis, the set-indexed random processes will provide us a theoretical background to the solution of the following problem. We consider a point process of some particles in $\mathbb{R}^n$, which we observe in a bounded observation window. We associate a vector of d parameters with each particle. For a given realization we assume that the parameters can be determined for the particle lying completely in the window. However, for only partially observable particles the information about these parameters is incomplete. Our aim is to estimate the distribution of these parameters. Since these parameters should describe some geometrical qualities of the observed particles (e.g. surface, volume, diameter), it is natural to assume that these parameters are non-negative.

We will use the following mathematical construction to represent the described problem. For every particle of the realization of our particle process, we create a d-dimensional vector describing the particle. Naturally, each of these vectors can be represented as a point in $\mathbb{R}^d_+$, therefore it belongs to some suitably chosen set in $\mathbb{R}^d_+$. We will use a theory of set-indexed random processes in $\mathbb{R}^d_+$ to construct a d-dimensional estimator for our vector.

Since we are dealing with censored data, there is an analogy with survival data analysis. We will be mainly interested in the estimation of the cumulative hazard rate, the hazard rate, and the survival function. For this purpose, we will apply the theory of set-indexed survival analysis, which is presented in [7].

This thesis will be divided into three chapters. In Chapter 1, we will consider the one-dimensional case and we will develop a Nelson-Aalen estimator of the

cumulative hazard rate. Then we will smooth this estimator to get an estimator of the hazard rate. At the end of Chapter 1, we will develop a Kaplan-Meier estimator of the survival function.

In Chapter 2, we will introduce the issue of set-indexed random processes, which is taken from [6], and based on this concept, we will develop an analogy to the Nelson-Aalen estimator for the cumulative hazard rate in d dimensions. The analogy between the one-dimensional and the multi-dimensional case will be studied. We will also introduce a generalization of the smoothed Nelson-Aalen estimator and a multi-dimensional Kaplan-Meier estimator, for which the distinction between the one-dimensional and the multi-dimensional estimator will be explained.

In Chapter 3, we will apply our theoretical achievements to simulated data and we will estimate the hazard rate and the survival function in some particular cases. Based on these results, the influence of the selected parameters will be discussed. The Kaplan-Meier estimator will be compared with a Horvitz-Thompson estimator which uses only information about the completely observed particles.

The thesis concludes with a brief discussion of the problem and suggestions for further research.

Chapter 1

One-dimensional case

Let us consider the situation which was described in Introduction. As we have mentioned, we will work with a point process of particles. Hence, it is necessary to define the particle process and its desired properties exactly. This will be done in the following section.

1.1 Particle processes

In this section, we give basic definitions for point processes of compact sets. We follow [9] and [10], where more details can be found.

Let $(\mathcal{K}', \varrho)$ be the family of non-empty compact subsets of $\mathbb{R}^n$ endowed with the Hausdorff metric

$$\varrho(K, L) = \max\left\{\sup_{x\in K} d(x, L), \sup_{y\in L} d(y, L)\right\}, \quad K, L \in \mathcal{K}',$$

where $d(x, L) = \inf_{z\in L} \|x - z\|$ is the distance of the point x to the set L. Denote by $\mathcal{N}$ the set of locally finite counting measures on $\mathcal{K}'$. We equip $\mathcal{N}$ with a σ-algebra $\mathfrak{N}$ which is defined as the smallest σ-algebra on $\mathcal{N}$ making the mappings $\psi \mapsto \psi(U)$ measurable for all Borel sets $U \in \mathcal{B}(\mathcal{K}')$.

Definition 1.1.1 (Particle process) A measurable mapping $\Psi : (\Omega, \mathcal{F}, P) \longrightarrow (\mathcal{N}, \mathfrak{N})$ is called a *point process of compact sets* or a *particle process.*

We will consider stationary particle processes.

Definition 1.1.2 (Stationary point process) Let t_z be a shift operator on $\mathcal{N}$ defined as

$$t_z\psi(U) = \psi(\{K - z : K \in U\}), \quad U \in \mathcal{B}(\mathcal{K}').$$

A point process Ψ on $\mathcal{K}'$ is *stationary* if its distribution is invariant with respect to t_z, i.e. Ψ and $t_z\Psi$ have the same distribution for all $z \in \mathbb{R}^n$.

Definition 1.1.3 (Reference point) Let $c : \mathcal{K}' \longrightarrow \mathbb{R}^n$ be a measurable mapping that is equivariant under translations, i.e. $c(K+x) = c(K) + x$ for all $x \in \mathbb{R}^n$ and $K \in \mathcal{K}'$. A point $c(K)$ will be referred as the *reference point* of the particle K. Furthermore, denote $\mathcal{K}'_0 = \{K \in \mathcal{K}' : c(K) = 0\}$ the family of non-empty compact subsets of $\mathbb{R}^n$ with the reference point in the origin.

There are many possibilities how to choose the mapping c. In this thesis, we will assume (if not stated otherwise) that the reference point is the minimum point with respect to the lexicographic order.

A measurable mapping $\phi : (x, K_0) \mapsto x + K_0$ is a bijection of $\mathbb{R}^n \times \mathcal{K}'_0$ onto $\mathcal{K}'$. Obviously, the inverse mapping is $\phi^{-1} : K \mapsto (c(K), K - c(K))$. Using this bijection, we can identify each stationary particle process Ψ with a stationary marked point process $\tilde{\Psi}$ having the mark space $\mathcal{K}'_0$:

$$\Psi(\{K\}) > 0 \Longleftrightarrow \tilde{\Psi}(\{\phi^{-1}(K)\}) > 0.$$

In the remainder of the thesis, we do not distinguish between Ψ and $\tilde{\Psi}$ and use the same symbol Ψ for both of them. It means that we can use the representation

$$\Psi = \sum_i \delta_{(X_i, \Xi_i)}, \tag{1.1}$$

where the points $\{X_i\}$ form a stationary point process

$$\Phi = \sum_i \delta_{X_i}$$

on $\mathbb{R}^n$ with an intensity α and the marks Ξ_i belong to $\mathcal{K}'_0$. The process (1.1) is also called a *germ-grain process*, see [5] or [10].

We will assume that (1.1) is an *independently marked point process*, i.e. Φ and $\{\Xi_i\}$ are independent, and $\{\Xi_i\}$ is a sequence of independent identically distributed random elements of $\mathcal{K}'_0$. The common distribution of the particles Ξ_i coincides with the mark distribution of the marked point process Ψ. Let Ξ_0 be a random compact set having this distribution; it is called a *typical particle*.

Later, we will often consider stationary Poisson particle processes. Every stationary Poisson particle process is automatically independently marked point process (see

[10], 4.3.3), and the corresponding point process of reference points is a stationary Poisson point process.

It was mentioned in Introduction that we will consider a d-dimensional vector of non-negative parameters associated with the particles. Our aim is to get the estimates of the distributional properties of this vector based on a single realization of the particle process in a bounded observation window W. We suppose that for completely observed particles, we have an information about the vector. But since the window W is bounded, edge effects cause that some of the particles may be only partially observable. This can be considered as a type of random censoring. We will exploit the methods from survival analysis to deal with the censoring effects in our problem.

In addition to the censoring effects, another type of edge effects is spatial sampling bias, see [2]. The probability of observing a particle depends on its size or shape. It is intuitively clear that larger particles have greater chance to hit the observation window. In order to avoid sampling bias, we will take into account only those particles whose reference points lie in W.

In this chapter, we concern ourselves with the case $d = 1$. For concreteness, we consider the process of discs in $\mathbb{R}^2$, and we are interested in two geometrical parameters for each disc: its area and perimeter. Since both parameters depend only on the radius of the disc, our two-dimensional vector can be calculated from a single one-dimensional variable. Therefore, we will construct the estimators in one dimension. This construction does not require a theory of set-indexed random processes, but it gives us a basic principle which is then generalized in the multi-dimensional case as will be apparent in the following chapters. We denote the radius of the disc Ξ_i by Y_i.

It is useful to note that if the window is rectangular, we can determine the radius of each disc with the lexicographic minimum point in the window even if we are not able to observe the full extent of the disc. Thus, there are in fact no censoring effects in this situation. In this way, we obtain an additional information, and it is possible to construct a better (uncensored) estimator.

In the following section we will use some theory about jump processes and martingales to get an estimator of the hazard rate of the one-dimensional vector formed by the radius of the typical disc Ξ_0.

1.2 Nelson-Aalen estimator

We will write down a few definitions and propositions which are needed to make the terminology clear and unified.

Definition 1.2.1 (Cadlag process) Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(X(t), t \in \mathcal{T})$, where $\mathcal{T} = [0, T]$ for some $T \in \mathbb{R}_+$, be a stochastic process on this space. Then X is called a *cadlag process* if its sample paths $(X(t, \omega), t \in \mathcal{T})$ are right-continuous with left-hand limits P-almost surely (a.s.).

Definition 1.2.2 (Multivariate counting process) Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $(\mathcal{F}_t, t \geq 0)$ that satisfies the *usual conditions,* i.e. it is increasing, right continuous and complete. A *multivariate counting process*

$$\boldsymbol{N} = (N_1, \ldots, N_m)$$

is a vector of m adapted cadlag processes, all zero at time zero and with paths that are piecewise constant and non-decreasing and have jumps of size +1 only. No two components may jump simultaneously a.s.

Remark 1.2.1 Since the components of a counting process $\boldsymbol{N}$ are adapted, cadlag, locally bounded, and non-decreasing, they are local submartingales and therefore have compensators $\Lambda_i, i = 1, \ldots, m$. Thus, $M_i = N_i - \Lambda_i$ is a local martingale for every $i = 1, \ldots, m$.

Definition 1.2.3 (Intensity process) Let $\boldsymbol{N} = (N_1, \ldots, N_m)$ be a multivariate counting process. We say that N_i has *an intensity process* λ_i if

$$\Lambda_i(t) = \int_0^t \lambda_i(s)\, \mathrm{d}s,$$

where Λ_i is a compensator of N_i. If N_i has an intensity process λ_i for every $i = 1, \ldots, m$, we say that the process $\boldsymbol{N}$ has an intensity process $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$.

Definition 1.2.4 (Multiplicative intensity model) Let $\boldsymbol{N} = (N_1, \ldots, N_m)$ be a multivariate counting process with an intensity process $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$. We say that $\boldsymbol{N}$ satisfies the *multiplicative intensity model* if λ_i can be written in the form

$$\lambda_i(t) = a_i(t) Z_i(t), \qquad i = 1, \ldots, m,$$

where $a_i(t)$ is a non-negative deterministic function, and $Z_i(t)$ is a predictable process.

Let us consider a multivariate counting process $\boldsymbol{N} = (N_1, \ldots, N_m)$ with an intensity process $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ satisfying the multiplicative intensity model

$$\lambda_i(t) = a_i(t) Z_i(t), \qquad i = 1, \ldots, m.$$

We want to get an estimator for

$$A_i(t) = \int_0^t a_i(s)\,\mathrm{d}s, \qquad i = 1, \ldots, m.$$

Since

$$M_i(t) = N_i(t) - \int_0^t a_i(s) Z_i(s)\,\mathrm{d}s$$

is a local martingale, we can write symbolically

$$\mathrm{d}N_i(t) = a_i(t) Z_i(t)\mathrm{d}t + \mathrm{d}M_i(t),$$

where $\mathrm{d}M_i(t)$ can be considered as a "random noise" component. By this, we get an estimator of $A_i(t)$ as

$$\hat{A}_i(t) = \int_0^t Z_i(s)^{-1}\,\mathrm{d}N_i(s).$$

Remark 1.2.2 Some problems could arise at this point if there would exist some $t \in \mathbb{R}_+$ for which $Z_i(t) = 0$. At this point, we will assume that $Z_i(t)$ is positive, and we will return to this problem later when we will have particular process $Z_i(t)$.

Let $T_{i1} < T_{i2} < \ldots$ denote the successive jump times of N_i. Then N_i gives mass 1 to each of these jump times and mass 0 elsewhere, i.e.

$$N_i = \sum_j \delta_{T_{ij}}.$$

Thus, it follows that we may write $\hat{A}_i(t)$ as a simple sum

$$\hat{A}_i(t) = \sum_{\{j:T_{ij} \le t\}} Z_i(T_{ij})^{-1}, \qquad i = 1, \ldots, n.$$

The function $\hat{A}_i(t)$ is called a *Nelson-Aalen estimator*.

Now, we will work with a special counting process $\boldsymbol{N}(t)$ to get the results which will be applicable to our situation. First, we define a *univariate counting process* N by

$$N(t) = I(Y \le t), \tag{1.2}$$

where Y is a non-negative random variable with an absolutely continuous distribution function F, the survival function $S = 1 - F$, and the density f. The *hazard rate* of Y is defined as

$$h(t) = \frac{f(t)}{S(t)}.$$

This random variable Y will stand for the radius of the typical disc Ξ_0, so for clarity, we will denote the variable by r instead of t.

Along with the hazard rate, we define the *cumulative hazard rate* by

$$H(r) = \int_0^r h(s)\,\mathrm{d}s.$$

We will now prove the following key proposition, which gives us the form of a compensator of N. Both Proposition 1.2.1 and its proof are taken from [1].

Proposition 1.2.1 *The counting process N defined in* (1.2) *has a compensator Λ_0 given by*

$$\Lambda_0(r) = \int_0^r h(s)Z_0(s)\,\mathrm{d}s$$

and, hence, N has an intensity process λ_0 in the form

$$\lambda_0(r) = h(r)Z_0(r),$$

where $Z_0(r)$ is a left-continuous adapted process defined as $Z_0(r) = I(Y \geq r)$, and $h(r)$ is the hazard rate of Y.

Proof: Let τ_F be the upper limit of the support of F and

$$\mathcal{N}_r = \sigma\{N(s) : s \leq r\} = \sigma\{Y \wedge r, I(Y \leq r)\}$$

be the filtration generated by the process N. We note that Λ_0 is predictable (it is continuous and adapted), so we need only to verify that $M = N - \Lambda_0$ is a local martingale (and therefore a square integrable martingale, as can be found in [1]). To show the martingale property, it suffices to verify that

$$\mathbb{E}[M(\infty)|\mathcal{N}_r] = M(r),$$

because this implies, for $s < r$,

$$\begin{aligned}\mathbb{E}[M(r)|\mathcal{N}_s] &= \mathbb{E}[\mathbb{E}[M(\infty)|\mathcal{N}_r]|\mathcal{N}_s] \\ &= \mathbb{E}[M(\infty)|\mathcal{N}_s] \\ &= M(s).\end{aligned}$$

First, we will consider the case $r = 0$. Because $\mathcal{N}_0$ is trivial, we must show that $\mathbb{E}[M(\infty)] = 0$. We have $N(\infty) = 1$ and

$$\begin{aligned}\mathbb{E}[\Lambda_0(\infty)] &= \mathbb{E}\int_0^{\tau_F} h(s)Z_0(s)\,\mathrm{d}s \\ &= \int_0^{\tau_F} P(Y \geq s)h(s)\,\mathrm{d}s \\ &= \int_0^{\tau_F} S(s)\frac{f(s)}{S(s)}\,\mathrm{d}s = \int_0^{\tau_F} f(s)\,\mathrm{d}s = 1.\end{aligned}$$

Next, we will show that the result for general r follows from that for $r = 0$. When conditioning on $\mathcal{N}_r = \sigma\{Y \wedge r, I(Y \leq r)\}$, we have to consider two separate cases: conditioning on $Y = s \leq r$ for some $s \leq r$ and conditioning on $Y > r$. In the first case, $M(\infty) = M(r) = M(s)$ and there is nothing to prove. In the second case,

$$M(\infty) - M(t) = 1 - \int_t^{\tau_F} I(Y \geq s)h(s)\,\mathrm{d}s$$

and we must show that the expectation of this random variable, given $Y > r$, is zero. But conditionally on $Y > r$, Y has the hazard rate $hI_{(r,\tau_F)}$. So this is just the same as the case $r = 0$ only with a different hazard rate.

□

Now, we can define the process $\boldsymbol{N}(r)$. It will be based on the radii of the discs $X_i + \Xi_i$ such that $X_i \in W$. Let $\boldsymbol{N}_0 = (N_1, \ldots, N_{\Phi(W)})$ be a counting process with

$$N_i(r) = I(Y_i \leq r),$$

where Y_i are iid non-negative random variables (the radii of the discs Ξ_i) with the same distribution as Y above. We note that since no two components of $\boldsymbol{N}_0$ jump simultaneously a.s.,

$$\boldsymbol{N}(r) = \sum_i N_i(r) = \sum_i I(Y_i \leq r)$$

is also a counting process. Since the process Ψ is independently marked, then from Proposition 1.2.1 it immediately follows that $\boldsymbol{N}$ has a compensator

$$\Lambda(r) = \int_0^r h(s)Z(s)\,\mathrm{d}s,$$

where

$$Z(r) = \sum_i I(Y_i \geq r).$$

Using the previous results, we can now construct a Nelson-Aalen estimator of the cumulative hazard rate:

$$\hat{H}(r) = \sum_{\{j:Y_j \leq r\}} Z(Y_j)^{-1}. \tag{1.3}$$

Remark 1.2.3 If we return to Remark 1.2.2, we see that if there exists some $r_1 > 0$ such that $Z(r_1) = 0$, then there will be no observed disc with radius greater than r_1. But in that case, it would be natural to set $\hat{S}(r_1) = 0$, and the estimation of $H(r_1)$ would make no sense. Therefore, we will construct an estimator only for $r \in \left(0, \max_{i=1,\ldots,\Phi(W)} Y_i\right)$

The last constructed estimate (1.3) is however an uncensored one, i.e. it could be used if we had a complete observation, as it is, for example, in the case of the process of discs observed in some rectangular window. Generally, if some of the variables Y_i are censored, we only know some lower boundary which they certainly exceed. From the observation of the process in the window W we get the censored sample $(\widetilde{Y}_i, D_i)$, where $\widetilde{Y}_i = Y_i \wedge C_i$, $D_i = I(\widetilde{Y}_i = Y_i)$ and C_i are the censoring random variables. According to the definition of the process of discs, X_i is the lexicographic minimum point of the disc $X_i + \Xi_i$ which has the radius Y_i. If we omit the possibility to determine the radii exactly, we can define the right-censoring random variables as

$$C_i = d(X_i, \partial W) = \inf_{w \in \partial W} \|X_i - w\|.$$

The C_i form a right-censoring process $\boldsymbol{C}$.

Definition 1.2.5 (Independent censoring) Let $\boldsymbol{N}$ be a multivariate counting process with a compensator $\boldsymbol{\Lambda}$ with respect to a given filtration $(\mathcal{F}_t)$. Let $\boldsymbol{C}$ be a right-censoring process which is adapted to a filtration $(\mathcal{G}_t) \supseteq (\mathcal{F}_t)$. Then we call the right-censoring of $\boldsymbol{N}$ generated by $\boldsymbol{C}$ *independent* if the compensator of $\boldsymbol{N}$ with respect to $(\mathcal{G}_t)$ is also $\boldsymbol{\Lambda}$.

Analogously to Proposition 1.2.1, it can be proved that if the censoring is independent, the univariate counting process

$$\boldsymbol{N}_C(r) = \sum_i I(\widetilde{Y}_i \leq r, D_i = 1)$$

satisfies the multiplicative intensity model with $\lambda(r) = h(r)Z_C(r)$, where

$$Z_C(r) = \sum_i I(\widetilde{Y}_i \geq r).$$

But from the definition of C_i and from the assumptions we made on Ψ it is obvious that our censoring is independent. Now, it is easy to see that in this situation the Nelson-Aalen estimator will be

$$\hat{H}_C(r) = \sum_{\{j:\widetilde{Y}_j \leq r\}} \frac{D_j}{Z_C(\widetilde{Y}_j)}.$$

1.3 Smoothing of Nelson-Aalen estimator

In the previous section, we constructed a censored version of the Nelson-Aalen estimator of the cumulative hazard rate. But the subject of our interest is mainly the hazard rate $h(r)$, so we will now smooth the Nelson-Aalen estimator to get a censored-case estimator $\hat{h}_C(r)$ of $h(r)$. To achieve this, we will use a *kernel function estimator*.

Definition 1.3.1 (Kernel function) The function $k : \mathbb{R} \longrightarrow \mathbb{R}$ is called a *kernel function* if k is bounded, non-negative, vanishes outside $[-1, 1]$, and has an integral equal to 1.

Now, we construct a *kernel function estimator* of $h(r)$:

$$\hat{h}_C(r) = b^{-1} \int_0^\infty k\left(\frac{r-s}{b}\right) \mathrm{d}\hat{H}_C(s),$$

where the *bandwidth* b is a positive parameter.
As $\hat{H}_C$ is a step-function, the last integral can be again written as a sum:

$$\hat{h}_C(r) = b^{-1} \sum_j k\left(\frac{r-\widetilde{Y}_j}{b}\right) \cdot \frac{D_j}{Z_C(\widetilde{Y}_j)}.$$

First, we must note that since k vanishes outside $[-1, 1]$, only those indices j contribute to the sum for which $r - b \leq \widetilde{Y}_j \leq r + b$. Second, it is obvious that particular values of this estimator will depend on the choice of k and b. We will discuss this problem in Chapter 3.

1.4 Kaplan-Meier estimator

In this section, we will derive a Kaplan-Meier estimator of the survival function S. To see how this estimator is created, we use a different formulation of the survival function S, which can be found in [1]. In order to write down the formula, we need the following definition.

Definition 1.4.1 (Product integral) Let $X(t), t \in \mathcal{T}$ be a cadlag function. We define a *product integral* of X over intervals of the form $[0, t], t \in \mathcal{T}$ as

$$\prod_{s \leq t} (1 + X(\mathrm{d}s)) = \lim_{\max |t_i - t_{i-1}| \to 0} \prod (1 + X(t_i) - X(t_{i-1})),$$

where $0 = t_0 < t_1 < \ldots < t_k = t$ is a partition of $[0, t]$.

The survival function can be written in the following form:

$$S(r) = \prod_{s \leq r} (1 - \mathrm{d}H(s)), \qquad r \in \mathcal{T} = [0, \rho) \text{ with } \rho = \sup\{r : S(r) > 0\},$$

where $H(r)$ is the cumulative hazard rate.

But now, since we have the Nelson-Aalen estimator of $H(r)$, it is straightforward to define a *Kaplan-Meier estimator* $\hat{S}_C(r)$:

$$\hat{S}_C(r) = \prod_{s \leq r} (1 - \mathrm{d}\hat{H}_C(s)).$$

Because $\hat{H}$ is a step-function, we may write

$$\hat{S}_C(r) = \prod_{s \leq r} (1 - \Delta\hat{H}_C(s)) = \prod_{\{j : \widetilde{Y}_j \leq r\}} \left(1 - \frac{D_j}{Z_C(\widetilde{Y}_j)}\right),$$

where

$$\Delta\hat{H}_C(s) = \hat{H}_C(\Delta s) = \hat{H}_C(s) - \hat{H}_C(s-).$$

Chapter 2

Multi-dimensional case

2.1 Set-indexed random processes

In this chapter, we will consider our problem in a general way. At the beginning, a framework of set-indexed processes will be established, making the terminology analogous to the terminology in the one-dimensional case as much as possible in order to make the analogy between these two cases clear. The framework of set-indexed processes is taken from [6].

Let T be a locally compact Hausdorff space and ν a measure on $\mathcal{B}$ (the Borel sets of T) which is finite on compact sets. All processes will be indexed by a class $\mathcal{A}$ of compact connected subsets of T. We will now define some properties of such class, which are needed for consistent definition of set-indexed random processes.

In what follows, $\mathcal{A}(u)$ will denote the class of finite unions of sets from $\mathcal{A}$, $\overline{(\cdot)}$ and $(\cdot)^\circ$ will denote, respectively, the closure and the interior of a set.

Definition 2.1.1 (Indexing collection) A non-empty class $\mathcal{A}$ of compact, connected subsets of T is called an *indexing collection* if it satisfies the following conditions:

(i) $\emptyset \in \mathcal{A}$ and $\forall A \in \mathcal{A}, A \neq A^\circ$ if $A \neq \emptyset$ or T. In addition, there is an increasing sequence (B_n) of sets in $\mathcal{A}(u)$ such that $T = \bigcup_{n=1}^{\infty} B_n^\circ$.

(ii) $\mathcal{A}$ is closed under arbitrary intersections and if $A, B \in \mathcal{A}$ are non-empty, then $A \cap B$ is non-empty. If (A_i) is an increasing sequence in $\mathcal{A}$ and there exists n such that $A_i \subseteq B_n$ for every i, then $\overline{\bigcup_i A_i} \in \mathcal{A}$. (Such a sequence (A_i) is called bounded.)

(iii) For $\sigma(\mathcal{A})$ (the σ-algebra generated by $\mathcal{A}$) it holds $\sigma(\mathcal{A}) = \mathcal{B}$.

(iv) *Separability from above:* There exists an increasing sequence of finite subclasses $\mathcal{A}_n = \{A_1^n, \ldots, A_{k_n}^n\}$ of $\mathcal{A}$ closed under intersections and satisfying $\emptyset, B_n \in \mathcal{A}(u)$ (B_n is defined in (i) above) and a sequence of functions $g_n : \mathcal{A} \longrightarrow \mathcal{A}(u) \cup \{T\}$ such that:

  (a) g_n preserves arbitrary intersections and finite unions, i.e. $g_n(\bigcap_{A\in\mathcal{A}'} A) = \bigcap_{A\in\mathcal{A}'} g_n(A)$ for any $\mathcal{A}' \subseteq \mathcal{A}$, and $\bigcup_{i=1}^k g_n(A_i) = \bigcup_{j=1}^m g_n(A'_j)$ whenever $\bigcup_{i=1}^k A_i = \bigcup_{j=1}^m A'_j$,

  (b) for each $A \in \mathcal{A}$, $A \subseteq (g_n(A))^\circ$,

  (c) $g_n(A) \subseteq g_m(A)$ if $n \geq m$,

  (d) for each $A \in \mathcal{A}$, $A = \bigcap_n g_n(A)$,

  (e) if $A, A' \in \mathcal{A}$, then for every n, $g_n(A) \cap A' \in \mathcal{A}$, and if $A' \in \mathcal{A}_n$, then $g_n(A) \cap A' \in \mathcal{A}_n$.

  (f) $g_n(\emptyset) = \emptyset \ \forall n$.

(v) Every countable intersection of sets in $\mathcal{A}(u)$ may be expressed as a closure of a countable union of sets in $\mathcal{A}$.

Further, for $t \in T$ we define the following sets:

- The "past" of t: $A_t = \bigcap\{A, A \in \mathcal{A}, t \in A\}$.
- The "future" of t: $E_t = \bigcap\{B^c, B \in \mathcal{A}, t \notin B\}$.

We shall also define a class $\mathcal{C}$ of all subsets of T of the form

$$C = A \setminus B, \quad A \in \mathcal{A},\ B \in \mathcal{A}(u).$$

At this point, the purpose of defining the class $\mathcal{C}$ may not be clear, but it will be necessary for the definition of a set-indexed martingale.

Remark 2.1.1 When developing the multi-dimensional analogy to the Nelson-Aalen estimator, we will use the following particular T and $\mathcal{A}$:

$$\begin{aligned} T &= \mathbb{R}_+^d, \\ \mathcal{A} &= \{[0,t] : t \in \mathbb{R}_+^d\}, \end{aligned} \tag{2.1}$$

where $[0,t] = [0,t_1] \times \cdots \times [0,t_d]$, $t = (t_1, \ldots, t_d)$. It is obvious that this particular $\mathcal{A}$ satisfies the conditions to be an indexing collection on T. Furthermore, the class $\mathcal{C}(u)$ (the class of finite unions of sets from $\mathcal{C}$) consists of all finite unions of disjoint rectangles of the form

$$(s,t], \quad s,t \in \mathbb{R}_+^d.$$

It is also easily seen that $A_t = [0,t]$ and $E_t = [t,\infty), t \in \mathbb{R}_+^d$.

Since we want to define a random process on $\mathcal{A}$ and later a set-indexed martingale, we have to define an $\mathcal{A}$-indexed filtration.

Definition 2.1.2 (Set-indexed filtration) Let $(\Omega, \mathcal{F}, P)$ be any complete probability space. An *$\mathcal{A}$-indexed filtration* is a class $\{\mathcal{F}_A, A \in \mathcal{A}\}$ of complete sub-σ-fields of $\mathcal{F}$ which satisfies the following conditions:

(i) $\forall A, B \in \mathcal{A}, \mathcal{F}_A \subseteq \mathcal{F}_B$, if $A \subseteq B$,

(ii) *monotone outer-continuity:* $\mathcal{F}_{\cap A_i} = \bigcap \mathcal{F}_{A_i}$ for any decreasing sequence (A_i) in $\mathcal{A}$,

(iii) if $T \notin \mathcal{A}$, then $\mathcal{F}_T = \mathcal{F}$.

We are now able to define a set-indexed stochastic process.

Definition 2.1.3 (Set-indexed stochastic process) For an indexing collection $\mathcal{A}$ we define an *$\mathcal{A}$-indexed stochastic process* $X = \{X_A, A \in \mathcal{A}\}$ as a collection of random variables indexed by $\mathcal{A}$. It is said to be *adapted* if X_A is $\mathcal{F}_A$-measurable for every $A \in \mathcal{A}$. The process X is said to be *integrable* if $\mathbb{E}|X_A| < \infty$ for every $A \in \mathcal{A}$. The process $X : \Omega \longrightarrow \mathbb{R}$ is *increasing* if X can be extended to a finitely additive process on $\mathcal{C}$, $X_\emptyset = 0$ and $X_C \geq 0$ $\forall C \in \mathcal{C}$ a.s.

At this point, we will construct the σ-algebra $\mathcal{G}^*_C$, which will allow us to define a strong martingale. If $B \in \mathcal{A}(u)$, then

$$\mathcal{F}^0_B = \bigvee_{A \in \mathcal{A}, A \subseteq B} \mathcal{F}_A,$$

where $\bigvee \mathcal{F}_A$ is the smallest σ-algebra generated by $\bigcup \mathcal{F}_A$. The σ-algebras $\{\mathcal{F}^0_B : B \in \mathcal{A}(u)\}$ are complete and increasing, but not necessarily monotone outer-continuous. Thus, we define

$$\begin{aligned}
\mathcal{F}_B &= \bigcap_{n \in \mathbb{N}} \mathcal{F}^0_{g_n(B)}, \qquad B \in \mathcal{A}(u), \\
\mathcal{G}^*_C &= \bigvee_{B \in \mathcal{A}(u), B \cap C = \emptyset} \mathcal{F}_B, \qquad C \in \mathcal{C}(u) \backslash \mathcal{A}, \\
\mathcal{G}^*_A &= \mathcal{F}_\emptyset, \qquad A \in \mathcal{A}.
\end{aligned}$$

Definition 2.1.4 (Set-indexed martingales) Let $X = \{X_A, A \in \mathcal{A}\}$ be an adapted and integrable additive process.

(i) X is called a *strong martingale* if for all $C \in \mathcal{C}$ we have $E[X_C|\mathcal{G}^*_C] = 0$.

(ii) If the process X satisfies the condition of strong martingale, but it is not adapted, we call it a *pseudo-strong martingale.*

Definition 2.1.5 (*-compensator) A process $\overline{X}$ is called a **-compensator* of the process X if it is increasing ($\overline{X}_C \geq 0$ for all $C \in \mathcal{C}$), and the difference $X - \overline{X}$ is a pseudo-strong martingale. The *-compensator is not necessarily unique.

We are now in a position to define the *multi-dimensional hazard rate.* Let $(\Omega, \mathcal{F}, P)$ be a complete probability space and $Y : \Omega \longrightarrow T$ be a T-valued random variable. We denote by F the distribution function of Y: $F(B) = P[Y \in B]$. The *survival function* associated with Y is

$$S(t) = F(E_t), \tag{2.2}$$

(but, in contrary to the one-dimensional case, $F(A_t) \neq 1 - S(t)$). We assume that F is absolutely continuous with respect to ν and denote by f the Radon-Nikodym derivative of F with respect to ν on the Borel sets of T.

Under these assumptions, we define the *hazard rate* of Y as

$$h(t) = \frac{f(t)}{S(t)}, \qquad t \in T.$$

Further, we define the *cumulative hazard rate* as

$$H_A = \int_A h(u)\, \nu(\mathrm{d}u), \qquad A \in \mathcal{A}.$$

Moreover, the hazard rate may be also defined as

$$h(t) = \lim_{n\to\infty} \frac{P(Y \in g_n(A_t) \mid Y \in E_t)}{\nu(g_n(A_t) \cap E_t)}, \tag{2.3}$$

when this limit exists.

Now, using the random variable Y from the previous paragraph, we will introduce a single jump process N as follows. Let

$$N = \{N_A, A \in \mathcal{A}\} = \{I(Y \in A), A \in \mathcal{A}\}$$

be a single jump process associated with Y and $\{\mathcal{F}^Y_A, A \in \mathcal{A}(u)\}$ its minimal filtration: $\mathcal{F}^Y_A = \sigma\{N_B : B \in \mathcal{A}, B \subseteq A\} \cup \{P_0\}$, where P_0 is the class of P-null sets. It can be easily shown that $\mathcal{F}^Y$ is monotone outer-continuous. We also note that the process N can be extended to an additive process on the more general index sets $\mathcal{A}(u)$ and $\mathcal{C}$.

The following proposition will give us a form of the *-compensator of our jump process N.

Proposition 2.1.1 *The process* Λ *defined by*

$$\begin{aligned}\Lambda_A &= \int_{A\cap A_Y} F(E_u)^{-1}\, F(\mathrm{d}u) \\ &= \int_{A\cap A_Y} h(u)\,\nu(\mathrm{d}u) \\ &= \int_A I(Y\in E_u)h(u)\,\nu(\mathrm{d}u)\end{aligned}$$

*is a *-compensator of the process* N *with respect to its minimal filtration, where* $A_Y(\omega) = A_{Y(\omega)}$.

The proof of this proposition can be found in [7].

We have enough information to be able to construct the multi-dimensional analogy to the Nelson-Aalen estimator. But this construction is analogical to that in one dimension, so we will leave it out, and instead we will straightly develop the censored estimator. To be able to do this, we must first introduce a multi-dimensional censoring mechanism by defining a *stopping set* ξ.

Definition 2.1.6 (Stopping set) A random variable $\xi : \Omega \longrightarrow \mathcal{A}(u)$ is a *stopping set* with respect to a filtration $\mathcal{F}$ if for any $A \in \mathcal{A}$, $\{\omega : A \subseteq \xi(\omega)\} \in \mathcal{F}_A$ and $\{\omega : \emptyset = \xi(\omega)\} \in \mathcal{F}_\emptyset$.

Generally, the observation of the random variable Y is "right-censored" or in our setting "outer-censored"; that is, Y is observed not on A but only on a subset of the form $A\cap\xi$. Consequently, it will be necessary to assume a type of independence between the censoring mechanism and the random variable being observed.

Definition 2.1.7 (Weak independence) Let Y be a T-valued random variable and let $\mathcal{F}^Y$ be the minimal filtration generated by its associated jump process N. Let $\mathcal{F}$ be a filtration such that $\mathcal{F}^Y_A \subseteq \mathcal{F}_A\ \forall A \in \mathcal{A}(u)$, and let ξ be a $\mathcal{F}$-stopping set. Then ξ is *weakly independent* of Y if the *-compensator of N with respect to $\mathcal{F}$ is the same as the *-compensator with respect to $\mathcal{F}^Y$.

When it exists, we may define the *hazard rate of the censored random variable* Y as

$$h^\xi(t) = \lim_{n\to\infty} \frac{P(Y \in g_n(A_t) \mid Y \in E_t, \xi \not\subseteq E_t^c)}{\nu(g_n(A_t)\cap E_t)}.$$

If we look at this definition and at (2.3), it is obvious that if ξ is weakly independent of Y, then $h^\xi = h$.

Now, let Y be a T-valued random variable whose associated jump process N is adapted to a filtration $\mathcal{F}$ and let ξ be an $\mathcal{F}$-stopping set. We define a *censored jump process* N^ξ as

$$N^\xi = \{N^\xi_A, A \in \mathcal{A}\} = \{I(Y \in A \cap \xi), A \in \mathcal{A}\}.$$

The following proposition, which is an analogoue of Proposition 2.1.1, will give us the *-compensator of this jump process N^ξ.

Proposition 2.1.2 *Assume that ξ is a stopping set, weakly independent of Y. Then the stopped process Λ^ξ defined by*

$$\begin{aligned}\Lambda^\xi_A &= \int_{A\cap\xi\cap A_Y} F(E_u)^{-1}\, F(\mathrm{d}u) \\ &= \int_{A\cap\xi\cap A_Y} h(u)\,\nu(\mathrm{d}u) \\ &= \int_A I(u\in\xi) I(Y\in E_u) h(u)\,\nu(\mathrm{d}u)\end{aligned}$$

*is a *-compensator of N^ξ.*

At this point, we have enough information to be able to construct the set-indexed version of the Nelson-Aalen estimator of the cumulative hazard rate

$$H_A = \int_A h(u)\,\nu(\mathrm{d}u)$$

using censored data. As in the one-dimensional case, let (Y_i) be a sequence of iid T-valued random variables with the same distribution as Y, and let (ξ_i) be a sequence of stopping sets. We shall assume that for every i and j, ξ_i is an $\mathcal{F}$-stopping set, weakly independent of Y_j. Next, we define

$$\boldsymbol{N}^\xi_A = \sum_i I(Y_i \in A \cap \xi_i).$$

Then, by independence and Proposition 2.1.2, $\boldsymbol{N}^\xi_A$ has a *-compensator in the form

$$\boldsymbol{\Lambda}_A = \int_A Z(t) h(t)\,\nu(\mathrm{d}t),$$

where

$$Z(t) = \sum_i I(Y_i \in E_t) I(t \in \xi_i). \tag{2.4}$$

Therefore, the process

$$M_A = \boldsymbol{N}^\xi_A - \int_A Z(t) h(t)\,\nu(\mathrm{d}t)$$

is a pseudo-strong martingale with respect to $\mathcal{F}$, and, exactly as in the classical case, we have

$$\boldsymbol{N}^{\xi}(\mathrm{d}t) = Z(t)h(t)\nu(\mathrm{d}t) + M(\mathrm{d}t).$$

Regarding M as noise, we are led to the set-indexed version of the Nelson-Aalen estimator of H_A:

$$\hat{H}_A = \int_A \frac{\boldsymbol{N}^{\xi}(\mathrm{d}t)}{Z(t)} = \sum_{\{j:Y_j \in A \cap \xi_j\}} (Z(Y_j))^{-1}. \tag{2.5}$$

2.2 Application to a process of rectangles

Now, we will apply the developed estimator to a particular case, for which we will later in Chapter 3 explicitly calculate the estimator from simulated data.

Our case will be as follows: Let us have a process of rectangles in the plane and assume that we can observe this process only in a rectangular window W. Let for simplicity the rectangles be parallel to the window W in the manner which is shown in Figure 2.1.

The observed variables of these rectangles will be their area and perimeter as well as in Chapter 1. But now, the number of parameters cannot be lessened, so we have to use the multi-dimensional theory. The reference point is the lexicographic minimum point of the rectangle (the left bottom corner). For each rectangle Ξ_i, we create a vector $Y_i = (Y_i^1, Y_i^2)$, where Y_i^1 denotes the area of Ξ_i and Y_i^2 the perimeter of Ξ_i. According to Remark 2.1.1, we put

$$\begin{aligned} T &= \mathbb{R}^2_+, \\ \mathcal{A} &= \{[0,t] : t \in \mathbb{R}^2_+\}. \end{aligned}$$

We will regard censoring in this example in the following way. For each rectangle $X_i + \Xi_i$ we need to measure the perimeter and area, or in the case that $X_i + \Xi_i$ is not fully observed, we need to get lower boundaries for these values. For this purpose, we will measure the distance between the left bottom corner X_i and the right and upper boundary of the rectangular window W. These distances will be denoted, respectively, by $d_1(X_i, W)$ and $d_2(X_i, W)$. Then we use these values to construct the censoring set ξ_i. Explicitly,

$$\begin{aligned} \xi_i &= [0, \xi_i^1] \times [0, \xi_i^2], \\ \xi_i^1 &= \varphi_i^1 \cdot \varphi_i^2, \\ \xi_i^2 &= 2(\varphi_i^1 + \varphi_i^2), \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \varphi_i^1 &= \min\{d_1(X_i, W), R_i^1\}, \\ \varphi_i^2 &= \min\{d_2(X_i, W), R_i^2\} \end{aligned}$$

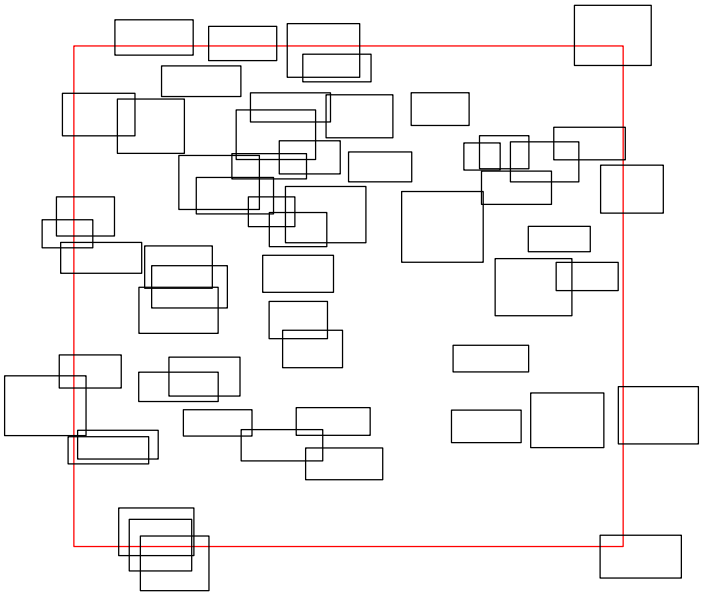


Figure 2.1: An example of a realization of the particle process Ψ in a rectangular window W.

with R_i^1, R_i^2 denoting the lengths of the two sides of Ξ_i, respectively. It is obvious that if $X_i + \Xi_i$ hits the boundary of W, then we do not know either R_i^1 or R_i^2, but we know that they are at least equal to $d_1(X_i, W)$ or $d_2(X_i, W)$, respectively, so both φ_i^1 and φ_i^2 are correctly defined.

Since we have defined the censoring sets ξ_i, we can also put the random vectors $Y_i = (Y_i^1, Y_i^2)$ into our scheme. For the partially observed particles we only know that $Y_i \in E_{\xi_i}$. On the other hand, if the particle $X_i + \Xi_i$ is contained in W, we have

$$Y_i = \xi_i \cap E_{\xi_i}$$

and

$$D_i = I(R_i^1 < d_1(x_i, W)) \cdot I(R_i^2 < d_2(x_i, W)) = 1.$$

Hence, for these particular random variables we will use the formula (2.5) to get

the estimator in our particular case. As mentioned above, the exact calculation will be done in Chapter 3.

Note that since we are dealing with non-negative variables we are in fact in the position of two-dimensional survival analysis, see e.g. [8]. However, the set-indexed framework enables applications for more general situations.

Remark 2.2.1 In [7], the consistency and asymptotic normality of the estimator (2.5) was shown under the assumption that the censoring sets ξ_i are iid and independent of a sequence of iid random vectors Y_i (see [8] for the same results in the two-dimensional setting). If we want to use the estimator (2.5) in our bivariate case, we meet the following problem. The censoring mechanism defined in (2.6) is not independent of the vectors Y_i. Nevertheless, we still think that it is reasonable to consider the corresponding two-dimensional Nelson-Aalen estimator. Of course, the statistical properties of the estimator can be destroyed. The open problem is whether the assumption of independence can be weakened, and how it influences the quality of the estimator. In Chapter 3, we make small simulation study which shows that the influence of the dependence between the censoring and the data is not so severe. However, this issue still requires further investigation.

2.3 Smoothing of the multi-dimensional Nelson-Aalen estimator

Similarly as in the one-dimensional case, we will smooth our multi-dimensional Nelson-Aalen estimator of the cumulative hazard rate by using a multi-dimensional kernel function. We will develop this smoothed estimator for a general multi-dimensional case and then apply it to our example with the rectangles.

Definition 2.3.1 (Multi-dimensional kernel function) We call the function $k : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ a *d-dimensional kernel function* if k is bounded, non-negative, vanishes outside $[-1, 1]^d$, and has integral equal to 1.

Let now

$$N = \{N_{A\cap\xi}, A \in \mathcal{A}\}$$

be a counting process in $\mathbb{R}^d_+$ (and therefore $\mathcal{A} = \{[0, t] : t \in \mathbb{R}^d_+\}$ according to (2.1)) with jump points $Y_i \in \mathbb{R}^d_+$. Using the notation from the previous sections, we will assume that this counting process has a compensator in the form

$$\int_A h(t)Z_0(t)\,\nu(\mathrm{d}t).$$

This assumption is strong, but we know that in our example we have the explicit form of the compensator, so we will be able to use the theory under such assumption.

Since we know the compensator of N, we can now, based on the theory developed earlier, construct an estimator of the cumulative hazard rate:

$$\hat{H}_A = \sum_{\{j:Y_j \in A \cap \xi_j\}} (Z_0(Y_j))^{-1}.$$

According to (2.1), we will take $A = [0, s]$, $s \in \mathbb{R}^d_+$, and we will formally write

$$\hat{H}_A = \hat{H}(s).$$

Using the multi-dimensional kernel function, we get a *kernel function estimator* $\hat{h}(t)$ of $h(t)$ as

$$\hat{h}(t) = b^{-d} \int_{\mathbb{R}^d_+} k\left(\frac{t-s}{b}\right) \mathrm{d}\hat{H}(s), \quad t \in \mathbb{R}^d_+,$$

where the bandwidth b is a positive real parameter. Since $\hat{H}$ is a jump process, the last integral can be written as a sum:

$$\hat{h}(t) = b^{-d} \sum_{\{j:Y_j \in [t-b,t+b] \cap \xi_j\}} k\left(\frac{t-Y_j}{b}\right) (Z_0(Y_j))^{-1}, \tag{2.7}$$

where

$$[t-b, t+b] = [t_1 - b, t_1 + b] \times \cdots \times [t_d - b, t_d + b]\,, \; t = (t_1, \ldots, t_d).$$

The application to our case is straightforward, we only set $d = 2$ and instead of the general process Z_0, we will use our particular process Z defined in (2.4).

2.4 Kaplan-Meier estimator on the plane

In this section, we will first develop a general theory to get a Kaplan-Meier estimator of the survival function in the two-dimensional space. Then we will apply this estimator to our example with the rectangles. The reasons, why we do not develop the theory for general multi-dimensional case as we have done with the Nelson-Aalen estimator, are at least two. First, although the generalization to the multi-dimensional case is possible and intuitive, in some points it can lose clarity, and also it can become uselessly complicated. Second, since in our example we work with two-dimensional vector of parameters, this theory will be sufficient to our case, and also the analogy to the one-dimensional case can be discussed more easily.

The definition of the Kaplan-Meier estimator on the plane is according to [4]. There are also different approaches to the estimation of the survival function, see e.g. [11].

2.4.1 Theory

First, we will define the bivariate cumulative hazard rate.

Definition 2.4.1 Let $Y = (Y^1, Y^2)$ be a pair of non-negative random variables on a probability space $(\Omega, \mathcal{F}, P)$, and let $S(s,t) = P(Y^1 > s, Y^2 > t)$ be the corresponding joint survival function. By a *bivariate cumulative hazard rate*, we mean a vector function $H(s,t) = (H_{10}(s,t), H_{01}(s,t), H_{11}(s,t))$, where

$$\begin{aligned} H_{11}(\mathrm{d}s, \mathrm{d}t) &= \frac{P(Y^1 \in \mathrm{d}s, Y^2 \in \mathrm{d}t)}{P(Y^1 \geq s, Y^2 \geq t)} = \frac{S(\mathrm{d}s, \mathrm{d}t)}{S(s-, t-)}, \\ H_{10}(\mathrm{d}s, t) &= \frac{P(Y^1 \in \mathrm{d}s, Y^2 > t)}{P(Y^1 \geq s, Y^2 > t)} = \frac{-S(\mathrm{d}s, t)}{S(s-, t)}, \\ H_{01}(s, \mathrm{d}t) &= \frac{P(Y^1 > s, Y^2 \in \mathrm{d}t)}{P(Y^1 > s, Y^2 \geq t)} = \frac{-S(s, \mathrm{d}t)}{S(s, t-)} \end{aligned}$$

and

$$H_{11}(0,0) = H_{10}(0,t) = H_{01}(s,0) = 0.$$

In what follows, for any bivariate function $\phi(s,t)$ we will use the following notation:

$$\begin{aligned} \phi(\Delta s, t) &= \phi(s,t) - \phi(s-,t), \\ \phi(s, \Delta t) &= \phi(s,t) - \phi(s,t-), \end{aligned}$$

and

$$\phi(\Delta s, \Delta t) = \phi(s,t) - \phi(s-,t) - \phi(s,t-) + \phi(s-,t-).$$

Since in our example we have to work with incomplete observation, we now need to derive an estimator of the bivariate survival function from censored data. For this purpose, we will introduce the random variables $Y = (Y^1, Y^2)$ and censoring random variables $C = (C^1, C^2)$, which are defined on a common probability space $(\Omega, \mathcal{F}, P)$. Let $S(s,t)$ and $G(s,t)$ denote, respectively, the survival function of Y and C. The observable random variables will be $\widetilde{Y} = (\widetilde{Y}^1, \widetilde{Y}^2)$ and $D = (D^1, D^2)$, where $\widetilde{Y}^j = \min\{Y^j, C^j\}$ and $D^j = I(\widetilde{Y}^j = Y^j)$, $j = 1, 2$.

Now we will introduce the following functions:

$$\begin{aligned} J(s,t) &= P(\widetilde{Y}^1 > s, \widetilde{Y}^2 > t), \\ K_1(s,t) &= P(\widetilde{Y}^1 > s, \widetilde{Y}^2 > t, D^1 = 1, D^2 = 1), \\ K_2(s,t) &= P(\widetilde{Y}^1 > s, \widetilde{Y}^2 > t, D^1 = 1), \\ K_3(s,t) &= P(\widetilde{Y}^1 > s, \widetilde{Y}^2 > t, D^2 = 1). \end{aligned} \tag{2.8}$$

Let us assume that $Y = (Y^1, Y^2)$ and $C = (C^1, C^2)$ are independent. Then, for (s,t) such that $J(s,t) > 0$, we have

$$\begin{aligned} J(s,t) &= G(s,t) \cdot S(s,t), \\ K_1(\mathrm{d}s, \mathrm{d}t) &= G(s-,t-) \cdot S(\mathrm{d}s, \mathrm{d}t), \\ K_2(\mathrm{d}s, t) &= G(s-,t) \cdot S(\mathrm{d}s, t), \\ K_3(s, \mathrm{d}t) &= G(s,t-) \cdot S(s, \mathrm{d}t), \end{aligned} \tag{2.9}$$

and therefore from Definition 2.4.1, we get the following equalities:

$$\begin{aligned} H_{11}(s,t) &= \int_0^s \int_0^t K_1(\mathrm{d}u, \mathrm{d}v)/J(u-,v-), \\ H_{10}(s,t) &= -\int_0^s K_2(\mathrm{d}u, t)/J(u-,t), \\ H_{01}(s,t) &= -\int_0^t K_3(s, \mathrm{d}v)/J(s,v-). \end{aligned} \tag{2.10}$$

To be able to construct the estimator, we need to express the survival function $S(s,t)$ in the form that will be suitable for estimating. For this purpose, we define a function $L(s,t)$ by

$$L(\mathrm{d}u, \mathrm{d}v) = \frac{H_{10}(\mathrm{d}u, v-)H_{01}(u-, \mathrm{d}v) - H_{11}(\mathrm{d}u, \mathrm{d}v)}{\{1 - H_{10}(\Delta u, v-)\}\{1 - H_{01}(u-, \Delta v)\}}.$$

Then, as shown in [4], the survival function $S(s,t)$ can be written as a product integral

$$S(s,t) = \prod_{u \le s} (1 - H_{10}(\mathrm{d}u, 0)) \times \prod_{v \le t} (1 - H_{01}(0, \mathrm{d}v)) \times \prod_{\substack{u \le s \\ v \le t}} (1 - L(\mathrm{d}u, \mathrm{d}v)), \tag{2.11}$$

where the last factor on the right-hand side is defined by

$$\prod_{\substack{u \le s \\ v \le t}} (1 - L(\mathrm{d}u, \mathrm{d}v)) = \lim_{\substack{\max|u_i - u_{i-1}| \to 0 \\ \max|v_i - v_{i-1}| \to 0}} \prod_{i,j} (1 - L((u_{i-1}, u_i] \times (v_{j-1}, v_j]))$$

with $0 = u_0 < \cdots < u_m = s$, $0 = v_0 < \cdots < v_n = t$ being, respectively, a partition of intervals $[0,s]$ and $[0,t]$ and

$$L((u_{i-1}, u_i] \times (v_{j-1}, v_j]) = L(u_i, v_j) - L(u_{i-1}, v_j) - L(u_i, v_{j-1}) + L(u_{i-1}, v_{j-1}),$$

cf. with Definition 1.4.1. We will now derive an estimator of the survival function $S(s,t)$ from (2.11) by using the estimator of the bivariate cumulative hazard rate $\hat{H}(s,t) = (\hat{H}_{10}(s,t), \hat{H}_{01}(s,t), \hat{H}_{11}(s,t))$. This estimator will be constructed in the following way, see [4] for details.

Let $Y_i = (Y_i^1, Y_i^2)$, $i = 1, \ldots, m$ and $C_i = (C_i^1, C_i^2)$, $i = 1, \ldots, m$ be iid random samples, where $(\widetilde{Y}_i, D_i)$ is defined analogously to $(\widetilde{Y}, D)$ and has the same distribution

as $(\widetilde{Y}, D)$. Since we want an estimator for the bivariate cumulative hazard rate, according to (2.9) and (2.10), we need to develop estimators for the functions $J(s,t)$, $K_1(s,t)$, $K_2(s,t)$ and $K_3(s,t)$ from the random sample $(\widetilde{Y}_i, D_i)$, $i = 1, \ldots, m$. This can be done in a natural way:

$$\begin{aligned}
\hat{J}(s,t) &= m^{-1} \sum_{i=1}^{m} I(\widetilde{Y}_i^1 > s, \widetilde{Y}_i^2 > t), \\
\hat{K}_1(s,t) &= m^{-1} \sum_{i=1}^{m} I(\widetilde{Y}_i^1 > s, \widetilde{Y}_i^2 > t, D_i^1 = 1, D_i^2 = 1), \\
\hat{K}_2(s,t) &= m^{-1} \sum_{i=1}^{m} I(\widetilde{Y}_i^1 > s, \widetilde{Y}_i^2 > t, D_i^1 = 1), \\
\hat{K}_3(s,t) &= m^{-1} \sum_{i=1}^{m} I(\widetilde{Y}_i^1 > s, \widetilde{Y}_i^2 > t, D_i^2 = 1).
\end{aligned}$$

Now we can use these estimators instead of the original functions in (2.10) and get the estimator for the bivariate cumulative hazard rate:

$$\begin{aligned}
\hat{H}_{11}(s,t) &= \int_0^s \int_0^t \hat{K}_1(\mathrm{d}u, \mathrm{d}v) / \hat{J}(u-, v-), \\
\hat{H}_{10}(s,t) &= -\int_0^s \hat{K}_2(\mathrm{d}u, t) / \hat{J}(u-, t), \\
\hat{H}_{01}(s,t) &= -\int_0^t \hat{K}_3(s, \mathrm{d}v) / \hat{J}(s, v-).
\end{aligned}$$

As mentioned above, we will use these estimators to get the estimator of the survival function $S(s,t)$:

$$\hat{S}_{KM}(s,t) = \hat{S}(s,0)\hat{S}(0,t) \prod_{\substack{0<u\leq s \\ 0<v\leq t}} [1 - \hat{L}(\Delta u, \Delta v)],$$

where

$$\hat{L}(\Delta u, \Delta v) = \frac{\hat{H}_{10}(\Delta u, v-)\hat{H}_{01}(u-, \Delta v) - \hat{H}_{11}(\Delta u, \Delta v)}{\{1 - \hat{H}_{10}(\Delta u, v-)\}\{1 - \hat{H}_{01}(u-, \Delta v)\}}$$

with $\hat{S}(s,0)$ and $\hat{S}(0,t)$ being the usual Kaplan-Meier estimators, i.e.

$$\begin{aligned}
\hat{S}(s,0) &= \prod_{u \leq s} (1 - \hat{H}_{10}(\Delta u, 0)), \\
\hat{S}(0,t) &= \prod_{v \leq t} (1 - \hat{H}_{01}(0, \Delta v)).
\end{aligned}$$

2.4.2 Application

We will now apply the theoretical results achieved in the previous section to our case, which is described in Section 2.2. We will again work with random variables

$Y_i = (Y_i^1, Y_i^2)$, where Y_i^1 denotes the area and Y_i^2 the perimeter of the rectangle Ξ_i. As opposed to Section 2.4.1, the sample size is now random. Nevertheless, from our assumption on Ψ, the random number $\Phi(W)$ is independent of the sequence (Y_i). Therefore, the construction of the Kaplan-Meier estimator can be done in the same way. Using the notation from Section 2.2, we put

$$\begin{aligned} C_i^1 &= \xi_i^1, \\ C_i^2 &= \xi_i^2. \end{aligned}$$

The random variable D_i will have the same form as in the theoretical part:

$$D_i^j = I(\widetilde{Y}_i^j = Y_i^j), \qquad j = 1, 2. \tag{2.12}$$

For those i for which $D_i = (0, 0)$, we do not know the values of Y_i, but we know that they are greater or equal to C_i, and therefore $\widetilde{Y}_i$ is correctly defined.

As in Section 2.2, the problem which arises in this particular case is the required independence of Y_i and C_i. We see from the definition of Y_i and C_i that this condition is not satisfied in our situation. Moreover, the C_i are not independent if Φ is not the Poisson point process. However, the application of the developed estimator does make sense in our case. For the calculations with particular data, the exact form of the censoring mechanism is not required. In Chapter 3 we study the quality of the estimator by simulation experiments. The problem of losing independence is common while working with spatial data. For example, see [3], where the Kaplan-Meier estimators of the nearest neighbour and the contact distribution function for point processes are introduced. There is no satisfying solution up to now and still further investigation is needed.

There is one thing which is worth noticing in our example. Since the two observed parameters are the area and the perimeter of the rectangles, either both or none of them are censored. It cannot happen that D_i is either $(1, 0)$ or $(0, 1)$. Because of this fact, all the functions K_1, K_2 and K_3 defined in (2.8) coincide into one function. Moreover, in some sense of words, we can say that our case is "properly" doubly parametric, because the two parameters are somehow "connected" in the sense that they are observed at once. The following example will show the difference between our "proper" doubly parametric case and the case in which we also have two parameters, the number of parameters cannot be lessened, and though the independence of observations and censoring can be achieved.

Let us now consider the following setup. We will again observe the same process of rectangles as in the previous case, but the observed parameters of this rectangles will now be the lengths of their sides. Using the same notation as in Section 2.2 and denoting $Y = (Y_i^1, Y_i^2)$ the observed values, where Y_i^1, Y_i^2 are the lengths of the two sides of the rectangle Ξ_i, we set

$$\begin{aligned} C_i^1 &= d_1(X_i, W), \\ C_i^2 &= d_2(X_i, W). \end{aligned}$$

In this case, the random variable D_i will have the same form as in (2.12). However, since Ψ is assumed to be independently marked point process, it is obvious that now the censoring mechanism is independent of the random variables Y_i.

The difference between our two examples is in the fact that in the second one, two parameters can be observed separately.

2.4.3 Differences compared to one dimension

We will now explain the differences between the one-dimensional and the multi-dimensional Kaplan-Meier estimator. If we look at the classical Kaplan-Meier estimator in a less formal way, we can see that it is computed by the product of some values (in fact the values are (1 − size of jump of the Nelson-Aalen estimator)) taken in those times the Nelson-Aalen estimator jumps (in our case we work with the radii instead of time). In the multi-dimensional case, the situation is to a certain extent analogical, but there is one important difference. If the dimension is greater than one, then we have to be careful while defining the jump time. Speaking of the two-dimensional case, as a jump time of the two-dimensional random vector has to be taken each time in which at least one of the components of the random vector jumps. However, it can happen that both of these components jump in one time. Intuitively, this "double jump" has to be treated in another way than the "single jumps". In our case, this was solved by introducing the functions H_{11}, H_{10}, H_{01}, that, roughly speaking, stand for the instantaneous risk of "double jump" (represented by H_{11}) or both components' "single jumps" (represented by H_{10}, H_{01}). If the dimension would be $d > 2$, the situation would become even more complicated, because the d-variate cumulative hazard rate would consist of $2^d - 1$ functions representing the instantaneous risk of "q-tuple jumps" of all d components. As we mentioned above, this was one of the reasons why we developed the multi-dimensional Kaplan-Meier estimator only for $d = 2$.

Chapter 3

Simulations

We will apply the multi-dimensional estimators on simulated data in this chapter. In particular, we will observe a stationary Poisson process of rectangles. As the observation window we choose

$$W = [0, 1]^2.$$

For simplicity, we will assume that the rectangles are transparent, i.e. that if two rectangles of the process overlap, then we still can see both of them. The point process Φ is the Poisson point process with an intensity α. The reference points of the rectangles will be the lexicographic minimum points, the sides of the rectangles will be parallel to the coordinate axes, and the joint distribution of their lengths (A, B) will have a two-dimensional probability density $f(a, b)$.

The realization of the process Φ will be simulated in a greater window

$$W^+ = [-a_{max}, 1]^2,$$

where a_{max} is the maximal side length of the rectangle. The choice of this window ensures that we get all rectangles that hit the window W. Whenever X_i lies outside W^+, the corresponding particle $X_i + \Xi_i$ does not hit W. If the coordinates of X_i are (X_i^1, X_i^2), then $d_1(X_i, W) = 1 - X_i^1$ and $d_2(X_i, W) = 1 - X_i^2$, and the censoring random vectors are given by (2.6).

Since we want to compare our Kaplan-Meier estimator with the theoretical survival function $S(x, y)$, we need to compute the density function $g(x, y)$ of the random vector $Y_i = (Y_i^1, Y_i^2)$, and then for each (x, y) we have

$$S(x, y) = \int_x^\infty \int_y^\infty g(u, v)\, \mathrm{d}v\, \mathrm{d}u.$$

The choice of the density function $f(a, b)$ is crucial, because we need the density function $g(x, y)$ to be as simple as possible so that we are able to express $S(x, y)$.

For the computation of $g(x,y)$, we will use the change of variables theorem. If we transform the random vector (A,B) with a density function $f(a,b)$ to the random vector $(X,Y) = (A \cdot B, 2(A+B))$, then, according to the change of variables theorem, the density $g(x,y)$ has the form

$$\begin{aligned} g(x,y) &= f(a,b)\frac{1}{\sqrt{y^2-16x}} \\ &= f(\frac{1}{4}(y+\sqrt{y^2-16x}), \frac{1}{4}(y-\sqrt{y^2-16x}))\frac{1}{\sqrt{y^2-16x}} + \\ &\quad f(\frac{1}{4}(y-\sqrt{y^2-16x}), \frac{1}{4}(y+\sqrt{y^2-16x}))\frac{1}{\sqrt{y^2-16x}}, \quad y^2 \geq 16x. \end{aligned}$$

If we choose

$$f(a,b) = \frac{6}{(a_{max}-a_{min})^3}(a-b), \qquad a > b,\ (a,b) \in [a_{min}, a_{max}] \times [a_{min}, a_{max}],$$

the density function $g(x,y)$ will have the form

$$g(x,y) = \frac{3}{(a_{max}-a_{min})^3}$$

for (x,y) such that

$$a_{min}^2 \leq x \leq a_{max}^2,\ 4a_{min} \leq y \leq 4a_{max},$$

$$y \geq 4\sqrt{x},\ y \leq 2a_{min} + \frac{2x}{a_{min}},\ y \leq 2a_{max} + \frac{2x}{a_{max}}.$$

The domain of the density function $g(x,y)$ is shown on Figure 3.1 for $a_{min} = 0.05$ and $a_{max} = 0.15$.

The simulation from the density $f(a,b)$ can be accomplished by the rejection method. But much faster way is to use the fact that $A-B$ has beta distribution and conditionally on $A-B$ the random variable A has uniform distribution on the interval $[a_{min} + A - B, a_{max}]$.

Figure 3.2 shows the comparison of the theoretical survival function $S(x,y)$ and the two-dimensional Kaplan-Meier estimator for $a_{min} = 0.05$, $a_{max} = 0.15$ and $\alpha = 50$.

The latter estimator is based on the rectangles with lexicographic minimum point in the window. Another possibility is to choose a different reference point (e.g. lexicographic maximum point). Averaging of both estimators should lead to the estimator with lower variance.

We would like to investigate whether an additional information from the censored data in the Kaplan-Meier estimator gives a better estimate than if we would use

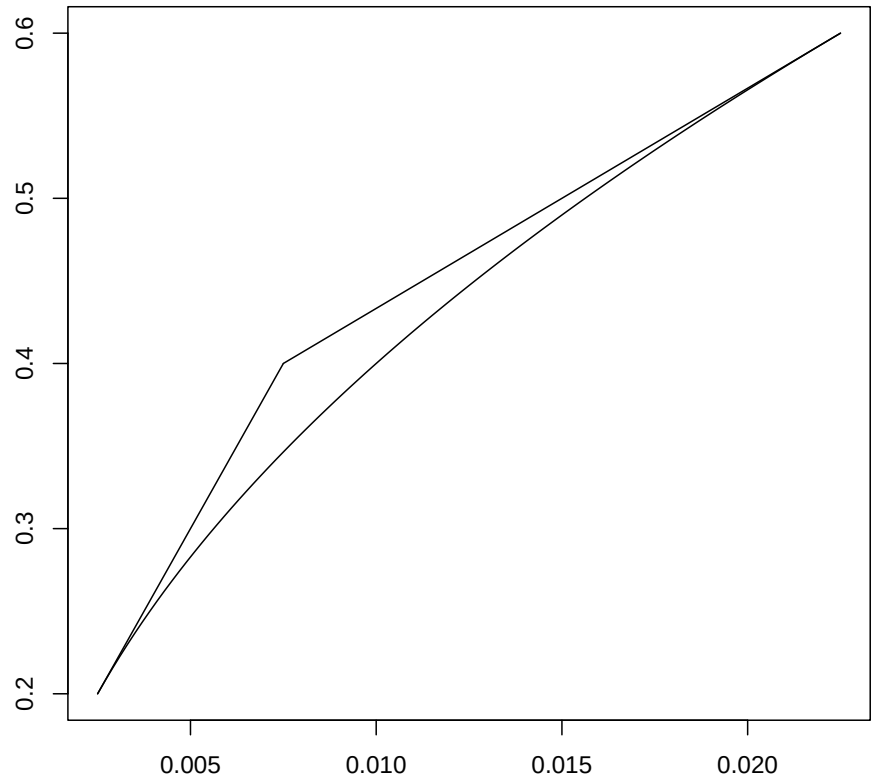


Figure 3.1: The domain of the density function $g(x, y)$ for $a_{min} = 0.05, a_{max} = 0.15$.

only completely observable data (so called minus-sampling, see [2]). For this purpose, we will introduce a Horvitz-Thompson estimator of the distribution function $F(x, y)$ of the vector Y_i. To be able to do this, we define the following set operations:

$$\begin{aligned} \check{V} &= \{-x : x \in V\}, \\ U \ominus V &= \bigcap_{y \in V} (U + y) = \{x : x + \check{V} \subseteq U\} \text{ (Minkowski-subtraction).} \end{aligned}$$

Now, we can define a *Horvitz-Thompson estimator* by

$$\hat{F}_{HT}(x, y) = \frac{1}{\hat{\alpha}} \sum_i \frac{I(X_i + \Xi_i \subseteq W)}{|W \ominus \check{\Xi}_i|} I(Y_i \in (-\infty, x] \times (-\infty, y]),$$

where

$$\hat{\alpha} = \sum_i \frac{I(X_i + \Xi_i \subseteq W)}{|W \ominus \check{\Xi}_i|}$$

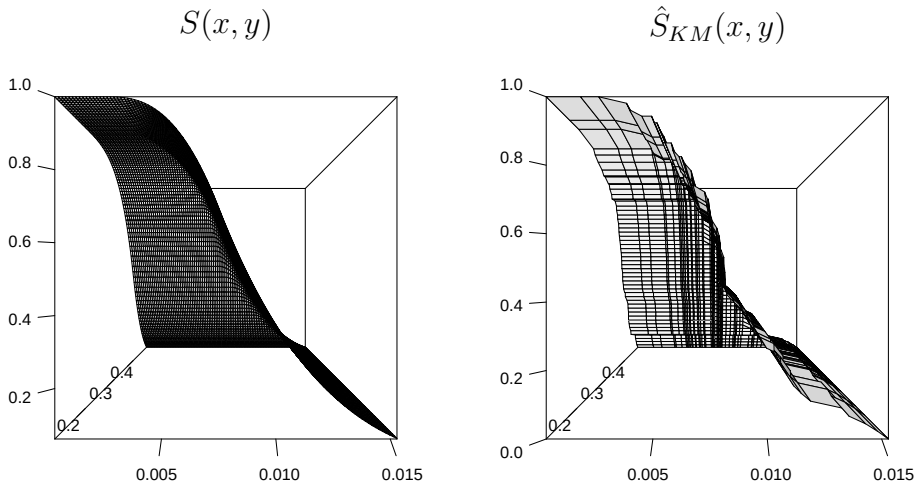


Figure 3.2: Comparison of the survival function and the Kaplan-Meier estimator.

is an unbiased estimator of the intensity α. It follows from Campbell's theorem for stationary marked point processes that $\hat{\alpha}\hat{F}_{HT}(x,y)$ is an unbiased estimator of $\alpha F(x,y)$. Thus, $\hat{F}_{HT}(x,y)$ is a ratio-unbiased estimator of $F(x,y)$. In [5], the asymptotic properties of $\hat{F}_{HT}(x,y)$ were studied.

If we want to get the Horvitz-Thompson estimator of the survival function $S(x,y)$, we cannot use the simple equation

$$\hat{S}_{HT}(x,y) = 1 - \hat{F}_{HT}(x,y),$$

which would be used in one dimension, because according to (2.2), this equation does not hold in our setup. Therefore, we have to put

$$\hat{S}_{HT}(x,y) = \frac{1}{\hat{\alpha}} \sum_i \frac{I(X_i + \Xi_i \subseteq W)}{|W \ominus \check{\Xi}_i|} I(Y_i \in [x,\infty) \times [y,\infty))$$

to get the Horvitz-Thompson estimator $\hat{S}_{HT}(x,y)$ of the survival function $S(x,y)$. Again, $\hat{S}_{HT}(x,y)$ is a ratio-unbiased estimator. The comparison of the survival function and the Horvitz-Thompson estimator for $a_{min} = 0.05$, $a_{max} = 0.15$ and $\alpha = 50$ is shown in Figure 3.3.

For both the Kaplan-Meier estimator and the Horvitz-Thompson estimator, we measure the goodness-of-fit by two commonly used distances between distribution functions (Kolmogorov-Smirnov and Cramer-von Mises). We will use these distance measures for the survival functions instead of the distribution functions. Explicit

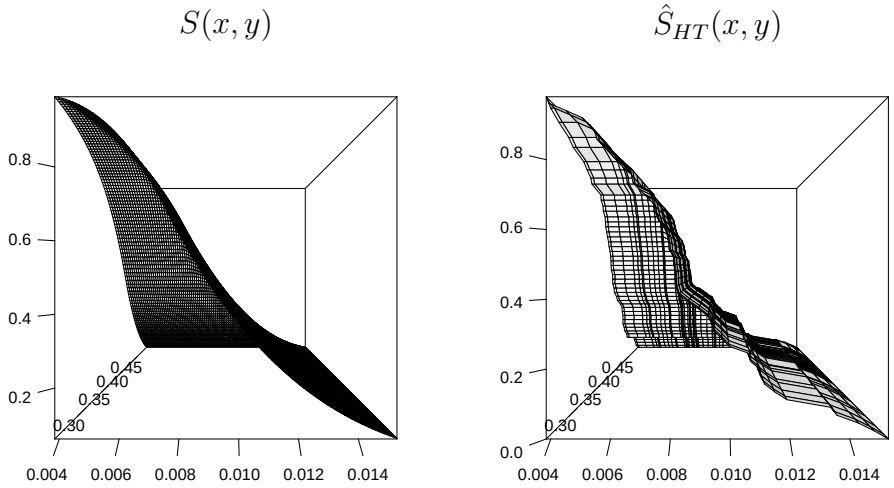


Figure 3.3: Comparison of the survival function and the Horvitz-Thompson estimator.

forms are

$$d_{KS}(\hat{S}, S) = \max_{x,y} |S(x,y) - \hat{S}(x,y)|, \tag{3.1}$$

$$d_{CvM}(\hat{S}, S) = \int_0^\infty \int_0^\infty (S(x,y) - \hat{S}(x,y))^2 g(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.2}$$

We will choose $a_{min} = 0.05$ and $a_{max} = 0.15$. For these parameters we will consider two different intensities $\alpha_1 = 30$ and $\alpha_2 = 50$. This choice of parameters leads to a situation in which we have approximately α_i $(i = 1, 2)$ rectangles that have the reference point inside W and about a fifth of them is censored. We have simulated 50 independent realizations of the process. We calculated both the Kaplan-Meier and the Horvitz-Thompson estimator for each realization. For the calculation of the Kaplan-Meier estimator, we used two different choices of the reference point (lexicographic minimum and maximum). Let us denote these estimators by $\hat{S}_{min}$ and $\hat{S}_{max}$, respectively. We also computed the average of these two estimators. This averaged estimator is denoted by $\hat{S}_{KM}$. The averages of the distances (3.1) and (3.2) over 50 realizations are shown in Table 3.1 for α_1 and in Table 3.2 for α_2.

We see that the Kaplan-Meier estimator gives slightly better results than the Horvitz-Thompson estimator. It is not surprising that the estimators are more precise if the number of observed particles is greater. Similar results can be obtained for the problem of estimation of side lengths which was described in Section 2.4.2.

	$\hat{S}_{min}$	$\hat{S}_{max}$	$\hat{S}_{KM}$	$\hat{S}_{HT}$
$d_{KS}(\hat{S}, S)$	0.2006	0.2021	0.1997	0.2020
$100 \cdot d_{CvM}(\hat{S}, S)$	0.8092	0.8478	0.8099	0.8542

Table 3.1: The comparison of the Kaplan-Meier and the Horvitz-Thompson estimator for $\alpha_1 = 30$.

	$\hat{S}_{min}$	$\hat{S}_{max}$	$\hat{S}_{KM}$	$\hat{S}_{HT}$
$d_{KS}(\hat{S}, S)$	0.1473	0.1480	0.1472	0.1498
$100 \cdot d_{CvM}(\hat{S}, S)$	0.3941	0.4052	0.3958	0.4289

Table 3.2: The comparison of the Kaplan-Meier and the Horvitz-Thompson estimator for $\alpha_2 = 50$.

For the setting of the parameters which we used in the second case ($\alpha_2 = 50$), we will also calculate the rest of our estimators. Figure 3.4 shows the multi-dimensional Nelson-Aalen estimator. Since the function $\hat{H}_A$ is a function of sets in the form $A = [0, t]$, $t \in \mathbb{R}^2_+$, for each set A the value $\hat{H}_A$ is plotted in the point t which is the top right corner of A.

The next two figures show the smoothing of the previous estimator. In Figure 3.5, we used an one-dimensional *Epanechnikov kernel*

$$k(x) = \frac{3}{4}(1 - x^2), \qquad x \in [-1, 1]$$

to create a bivariate kernel function

$$k_e(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2), \qquad x, y \in [-1, 1].$$

The difference between the two graphs in the first figure is induced by different bandwidths. In the first case, we take as a bandwidth an estimate of the mean distance between the points $y_i = (y_i^1, y_i^2)$ representing the realization of the random variables Y_i. But since the area and the perimeter of the rectangles have proportionally different values, we use different bandwidth for each component of Y_i. Therefore, the estimator defined in (2.7) will now have the form

$$\hat{h}(t) = \prod_{i=k}^{d} b_k^{-1} \cdot \sum_{\{j: Y_j \in [t-b, t+b] \cap \xi\}} k\left(\frac{t - Y_j}{b}\right) (Z_0(Y_j))^{-1},$$

where $b = (b_1, \ldots, b_d)$. Then the choice of b in our case has the form

$$\hat{b}_k = \frac{1}{\Phi(W)^2} \sum_{i,j} |y_i^k - y_j^k|, \qquad k = 1, 2.$$

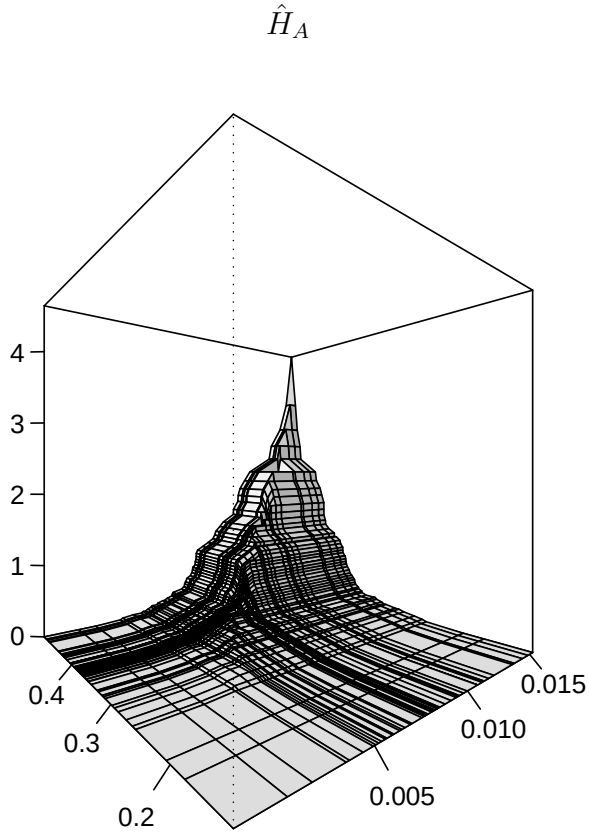


Figure 3.4: The two-dimensional Nelson-Aalen estimator of the cumulative hazard rate.

In the second case, we multiplied the bandwidth by 2. As is perceptible from Figure 3.5, the increase of bandwidth induces that the estimator is more smoothed than in the first case.

By the same construction as for the kernel function k_e, we create a two-dimensional biweight kernel function. An one-dimensional *biweight kernel* is defined as

$$k(x) = \frac{15}{16}(1 - x^2)^2, \qquad x \in [-1, 1].$$

Therefore, the bivariate kernel function based on the biweight kernel will have the form

$$k_b(x, y) = \frac{225}{256}(1 - x^2)^2(1 - y^2)^2, \qquad x, y \in [-1, 1].$$

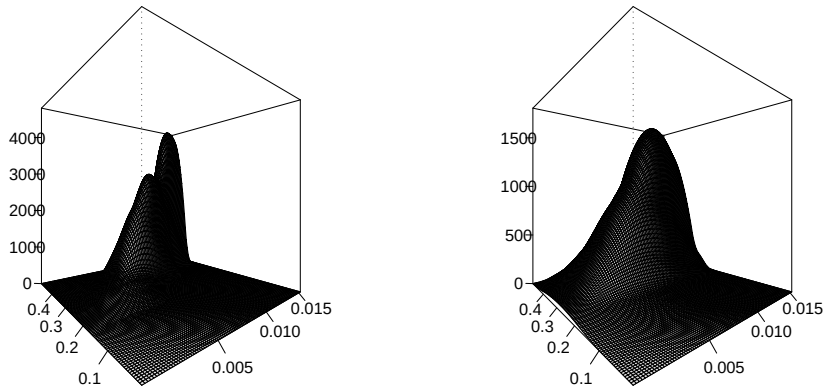


Figure 3.5: Smoothed Nelson-Aalen estimator by using the kernel function $k_e(x, y)$ for two different choices of bandwidths.

The results of smoothing the Nelson-Aalen estimator by the kernel function k_b are shown in Figure 3.6. Again we smoothed the estimator twice, each time with a different bandwidth. The values of the bandwidth were the same as in Figure 3.5.

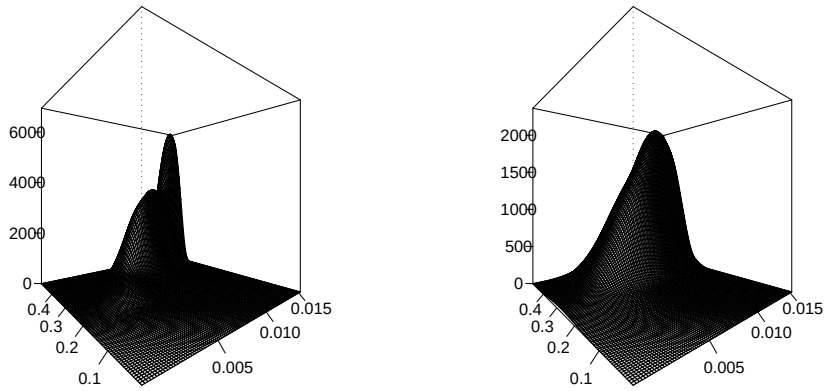


Figure 3.6: Smoothed Nelson-Aalen estimator by using the kernel function $k_b(x, y)$ for two different choices of bandwidths.

Conclusion

In this thesis, we dealt with a problem of estimating the hazard rate and the survival function of the particle process parameters from censored data. For the solution of this problem, the theory of set-indexed random processes was used. It provided us a background for defining the multi-dimensional Nelson-Aalen estimator of the cumulative hazard rate. As an estimator of the multi-dimensional survival function, the Kaplan-Meier estimator was developed. Though, for both estimators we identified a problem of dependence of the censoring mechanism on the data, we used our estimators in a particular case. In Chapter 3 we showed that the difference between the estimators and the theoretical functions was reasonably small. It also turned out that the Kaplan-Meier estimator was more efficient than the Horvitz-Thompson estimator which uses only completely observable data.

Still, there remains an issue to be a subject of further investigations, namely the problem of the influence of dependence between the censoring mechanism and the data on the estimation procedure. Furthermore, it would be also advisable to study the problem for other classes of point processes (cluster or hard-core point processes) than for the Poisson point process. Then we would be in the situation where the censoring variables are no longer independent.

Bibliography

[1] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993): *Statistical Models Based on Counting Processes.* Springer, New York.

[2] Baddeley, A. J. (1999): Spatial sampling and censoring. In *Stochastic Geometry: Likelihood and Computation,* O. E. Barndorff-Nielsen, W. S. Kendall, M. N. M. van Lieshout (eds.), 37–78. Chapman and Hall, London.

[3] Baddeley, A. J. and Gill, R. D. (1997): Kaplan-Meier estimators for interpoint distance distributions of spatial point processes. *Ann. Statist.* **25**, 263–292.

[4] Dabrowska, D. M. (1988): Kaplan-Meier estimate on the plane. *Ann. Statist.* **16**, 1475–1489.

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

[6] Ivanoff, G. and Merzbach, E. (2000): *Set-indexed Martingales.* CRC Press, Boca Ranton, FL.

[7] Ivanoff, G. and Merzbach, E. (2002): Random censoring in set-indexed survival analysis. *Ann. Probab.* **12**, 944–971.

[8] Pons, O. (1986): A test of independence between two censored survival times. *Scand. J. Statist.* **13**, 173–185.

[9] Rataj, J. (2006): *Bodové procesy.* Karolinum, Praha.

[10] Schneider, R. and Weil, W. (2000): *Stochastische Geometrie.* B. G. Teubner, Stuttgart.

[11] Tsai, W.-Y., Leurgans, S. and Crowley, J. (1986): Nonparametric estimation of a bivariate survival function in presence of censoring. *Ann. Statist.* **14**, 1351–1365.