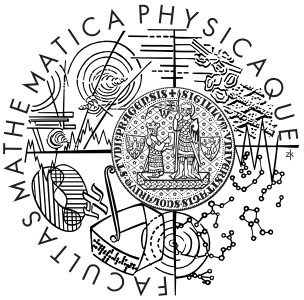
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Markéta Zikmundová

Odhady charakteristik bodových procesů úseček

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.

Studijní program: Matematika
Studijní plán: Teorie pravděpodobnosti a náhodné procesy

2008

Chtěla bych poděkovat svému vedoucímu práce RNDr. Zbyňku Pawlasovi, Ph.D. za všechen ten čas, který mi věnoval a za trpělivost, kterou se mnou měl. Věru to se mnou neměl jednoduché.

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

V Praze dne 8.srpna 2008 Markéta Zikmundová

Obsah

Název práce: Odhady charakteristik bodových procesů úseček
Autor: Markéta Zikmundová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Zbyněk Pawlas, Ph.D.
e-mail vedoucího: zbynek.pawlas@mff.cuni.cz

Abstrakt: Důležitou charakteristikou bodového procesu úseček je rozdělení jejich délek. Kromě známých odhadů Horvitzova-Thompsonova a Kaplanova-Meierova typu pro distribuční funkci délky typické úsečky se práce zabývá také užitím EM algoritmu k přibližnému výpočtu maximálně věrohodného odhadu pro stacionární Poissonův proces. Za parametrické odhady je v práci zmíněn SRE algoritmus, který patří mezi Monte-Carlo algoritmy. Dalším předmětem zkoumání je intenzita průsečíků procesu úseček. Práce se zabývá několika jejími neparametrickými odhady. Pro některé z nich jsou odvozeny střední hodnoty a rozptyly. Porovnání kvality jednotlivých odhadů obou charakteristik bylo provedeno na základě simulací stacionárního izotropního Poissonova bodového procesu úseček v $\mathbb{R}^2$ s rovnoměrným a logaritmicko-normálním rozdělením délek.

Klíčová slova: stacionární Poissonův bodový proces, intenzita průsečíků, Horvitzův-Thompsonův odhad, Kaplanův-Meierův odhad, SRE algoritmus

Title: Estimators of characteristics for segment point processes
Author: Markéta Zikmundová
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Zbyněk Pawlas, Ph.D.
Supervisor's e-mail address: zbynek.pawlas@mff.cuni.cz

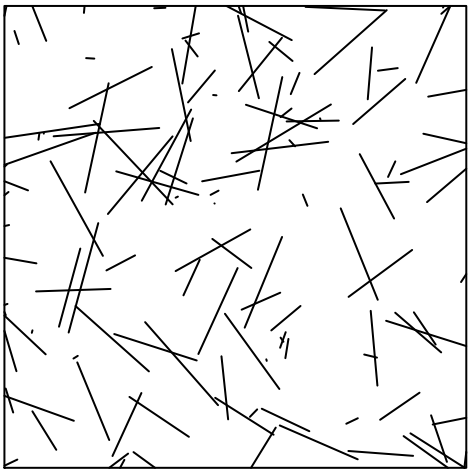
Abstract: One of the important characteristics of segment point process is the length distribution. In addition to known Horvitz-Thompson and Kaplan-Meier type estimators, this work deals with using the EM algorithm for rough computing the maximum likelihood estimator. From parametric estimators we mention SRE algorithm which is an iterative Monte Carlo algorithm. The second object of interest is the intersection intensity of segment process. This work studies several of its non-parametric estimators. The comparison of particular estimators of both characteristics is based on the simulations from a stationary isotropic Poisson segment process in $\mathbb{R}^2$ with the uniform and log-normal distribution of lengths.

Keywords: stationary Poisson point process, intersection intensity, Horvitz-Thompson estimator, Kaplan-Meier estimator, SRE algorithm

Úvod

Bodový proces úseček je speciálním případem bodových procesů kompaktních množin. Každá úsečka v $\mathbb{R}^d$ je jednoznačně určena svým významným bodem (počátečním bodem, koncovým bodem,...), svým směrem a délkou. V praxi můžeme pomocí bodového procesu úseček modelovat například polomy stromů či praskliny v materiálech.

Dvě důležité charakteristiky, které nás v takovém případě zajímají, je délka úseček a intenzita průsečíků. K jejich odhadu máme většinou k dispozici pouze část dat, kterou můžeme pozorovat skrze ohraničené okno pozorování. To samozřejmě má vliv na kvalitu odhadů. Cílem této práce je studium různých odhadů a jejich porovnání na základě numerických výpočtů.



Obrázek 1: Realizace stacionárního Poissonova bodového procesu úseček s rovnoměrným rozdělením délek.

První kapitola práce je věnována základním poznatkům o bodových procesech, zavedení bodového procesu úseček na prostoru $\mathbb{R}^d$ a některým důležitým příkladům bodových procesů.

Následující tři kapitoly jsou věnovány odhadu distribuční funkce rozdělení délek úseček. V práci jsou posuzovány některé typy odhadů založené na různých výběrech úseček, které budou pro odhad použity. Z neparametrických jsou to Horvitzovy-Thompsonovy odhady a Kaplanův-Meierův odhad. Nemalá část kapi-

toly je pak věnována EM algoritmu jako nástroji sloužícímu k přibližnému výpočtu maximálně věrohodného odhadu. Parametrické odhady zde zastupuje SRE algoritmus, jenž je popsán ve třetí kapitole. Ve čtvrté kapitole jsou pak porovnány všechny zmiňované odhady pro stacionární Poissonův proces úseček na základě simulací procesu s rovnoměrným a logaritmicko-normálním rozdělením délek při různé volbě okna pozorování.

Poslední kapitola je zaměřena na intenzitu průsečíků rovinného procesu úseček. Pro uvedené odhady jsou zde (pro stacionární Poissonův proces) vypočteny jejich střední hodnoty a rozptyly. Závěr kapitoly je opět věnován srovnání odhadů na základě simulací.

Simulace a výpočty byly prováděny v programu R [8] s využitím knihovny spatstat [3]. Součástí práce je CD s použitými programy, jež byly pro tento účel napsány.

Kapitola 1

Bodové procesy kompaktních množin

1.1 Základní definice

Mějme separabilní, lokálně kompaktní úplný metrický prostor E opatřený borelovskou σ-algebrou $\mathcal{B} = \mathcal{B}(E)$. *Lokálně konečnou mírou* na E nazveme takovou míru, která je konečná na všech omezených borelovských množinách v E. Množinu všech lokálně konečných měr na $(E, \mathcal{B}(E))$ budeme značit $\mathcal{M}$. Symbol $\mathcal{B}_o = \mathcal{B}_o(E)$ značí systém všech omezených borelovských množin na E.

Definice: Na prostoru $(E, \mathcal{B}(E))$ definujeme množinu $\mathcal{N}$ všech lokálně konečných měr nabývajících pouze nezáporných celočíselných hodnot nebo nekonečna, tedy

$$\mathcal{N} \equiv \{\mu \in \mathcal{M};\ \mu(B) \in \mathbb{N} \cup \{0, \infty\} \text{ pro každou } B \in \mathcal{B}\}.$$

Na prostorech $\mathcal{M}, \mathcal{N}$ definujeme σ-algebry

$$\mathfrak{M} = \sigma\{\mu \mapsto \mu(B) \text{ měřitelné},\ B \in \mathcal{B}\},$$

$$\mathfrak{N} = \{M \cap \mathcal{N} : M \in \mathfrak{M}\},$$

kde $\mathfrak{M}$ je nejmenší σ-algebra na $\mathcal{M}$ vzhledem k níž jsou zobrazení $\mathcal{M} \to \mathbb{R};\ \mu \mapsto \mu(B)$ měřitelná pro $B \in \mathcal{B}$.

Definice: Buď $(\Omega, \Sigma, \mathbb{P})$ pravděpodobnostní prostor. *Bodovým procesem* na E nazveme měřitelné zobrazení $\Phi : (\Omega, \Sigma, \mathbb{P}) \to (\mathcal{N}, \mathfrak{N})$.
Rozdělení bodového procesu je pravděpodobnostní míra Π definovaná vztahem $\Pi(U) = \mathbb{P}[\Phi \in U],\ U \in \mathfrak{N}$.
Řekneme, že bodový proces je *jednoduchý*, jestliže $\mathbb{P}[\Phi \in \mathcal{N}^*] = 1$, kde

$$\mathcal{N}^* = \{\gamma \in \mathcal{N} :\ \gamma(\{x\}) \leq 1;\ \forall x \in E\}.$$

Poznámka: Na jednoduchý bodový proces se dá pohlížet jako na součet náhodného

počtu Diracových měr. Jinými slovy, je-li Φ jednoduchý bodový proces na E, pak existuje posloupnost měřitelných zobrazení $X_i : \Omega \to E$ tak, že

$$\Phi = \sum_{i=1}^{\Phi(E)} \delta_{X_i}.$$

Důkaz lze nalézt v [10], Lemma 3.1.7.

Označení: Jednoduchý bodový proces lze ztotožnit s jeho nosičem. Často proto píšeme $x \in \Phi$ místo $\Phi(\{x\}) > 0$ a $\Phi = \{X_i\}$ místo $\Phi = \sum \delta_{X_i}$.

Definice: Nechť Φ je bodový proces. *Momentovou mírou k-tého řádu* nazveme míru

$$M_k(A) = \mathbb{E} \sum_{X_1,\dots,X_k \in \Phi} \mathbf{1}_{\{(X_1,\dots,X_k)\in A\}}, \qquad A \in \mathcal{B}^k.$$

Speciálně pro $k = 1$ dostáváme

$$\Lambda(A) = \mathbb{E}\Phi(A), \qquad A \in \mathcal{B},$$

míru intenzity procesu Φ.
Faktoriální momentovou mírou k-tého řádu rozumíme míru

$$M_k^!(A) = \mathbb{E} \sum_{X_1,\dots,X_k \in \Phi}^{\neq} \mathbf{1}_{\{(X_1,\dots,X_k)\in A\}}, \qquad A \in \mathcal{B}^k,$$

kde symbol $\sum^{\neq}$ značí sčítání přes všechny k-tice bodů, jejichž složky jsou navzájem různé.

Poznámka: Pro momentovou míru n-tého řádu a $B_1, \dots B_n \in \mathcal{B}$ platí

$$M_n(B_1 \times \cdots \times B_n) = \mathbb{E}[\Phi(B_1) \dots \Phi(B_n)]. \tag{1.1}$$

Pro faktoriální momentovou míru n-tého řádu a $B \in \mathcal{B}$ platí

$$M_n^!(B \times \cdots \times B) = \mathbb{E}[\Phi(B)(\Phi(B)-1)(\Phi(B)-2)\dots(\Phi(B)-n+1)]. \tag{1.2}$$

Velice užitečná je následující věta. Uvedeme si ji v méně obecné formě, která ale dostatečně vyhovuje našim potřebám.

Věta 1.1 (Campbellova) *Nechť Φ je jednoduchý bodový proces na E a buď Λ jeho míra intenzity. Potom pro libovolnou nezápornou měřitelnou funkci $f : E \to \mathbb{R}^+$ je*

$$\mathbb{E}\Big[\sum_{X\in\Phi} f(X)\Big] = \int_E f(x)\Lambda(\mathrm{d}x), \tag{1.3}$$

je-li alespoň jedna strana rovnosti konečná.

Pro libovolnou nezápornou měřitelnou funkci $f : E \times E \to \mathbb{R}^+$ je

$$\mathbb{E} \sum_{X,Y \in \Phi}^{\neq} f(X,Y) = \int_E \int_E f(x,y) M_2^!(\mathrm{d}(x,y)), \tag{1.4}$$

je-li alespoň jedna strana rovnosti konečná.

Důkaz. Pro funkci $f(x) = \mathbf{1}_{\{x \in B\}}$ plyne tvrzení přímo z definice míry intenzity Λ. Z linearity střední hodnoty dostáváme tvrzení pro jednoduché funkce. Nyní pro $f \geq 0$ aproximujeme f pomocí jednoduchých funkcí. Důkaz druhé rovnosti je analogický. □

Definice: Buď $\Lambda \in \mathcal{M}$ a nechť Φ je bodový proces na E takový, že pro všechna $n \in \mathbb{N}$ a $B_1, \ldots, B_n \in \mathcal{B}_o$ po dvou disjunktní množiny jsou náhodné veličiny $\Phi(B_1), \ldots, \Phi(B_n)$ nezávislé a pro všechna $i \in \mathbb{N}$ má $\Phi(B_i)$ Poissonovo rozdělení s parametrem $\Lambda(B_i)$. Potom se proces Φ nazývá *Poissonův proces* na E s mírou intenzity Λ.

Lemma 1.1 *Faktoriální míra Poissonova bodového procesu je $M_k^! = \Lambda^k$.*

Důkaz. Viz [9], str. 26.

Poznámka: Je-li Φ Poissonův proces s difúzní mírou intenzity, pak je Φ jednoduchý (viz [9], Lemma 6.1).

1.2 Stacionární a izotropní procesy

Nechť pro tuto chvíli je $E = \mathbb{R}^d$. Pro bodové procesy na tomto prostoru si zavedeme několik důležitých pojmů.

Definice: Řekneme, že bodový proces Φ na $\mathbb{R}^d$ je *stacionární*, jestliže jeho rozdělení je invariantní vůči posunutí, tj. jestliže proces $\Phi + y = \{x + y;\ x \in \Phi\}$ má stejné rozdělení jako Φ pro každé $y \in \mathbb{R}^d$.
Bodový proces, jehož rozdělení je invariantní vůči rotaci okolo počátku, nazýváme *izotropní*.

Definice: Existuje-li hustota λ míry intenzity Λ vzhledem k Lebesgueově míře, tj. $\Lambda(A) = \int_A \lambda(x)\,\mathrm{d}x,\ A \in \mathcal{B}$, potom se λ nazývá *funkce intenzity*.

Poznámka: Neboť je Lebesgueova míra až na násobek jediná lokálně konečná míra na $\mathbb{R}^d$ invariantní vůči posunutí, je pro stacionární procesy na $\mathbb{R}^d$ s lokálně konečnou mírou intenzity funkce intenzity konstantní a rovnost (1.3) v Campbellově větě má tvar

$$\mathbb{E}\Big[\sum_{X \in \Phi} f(X)\Big] = \lambda \int_{\mathbb{R}^d} f(x)\,\mathrm{d}x. \tag{1.5}$$

Konstanta λ se nazývá *intenzita* stacionárního bodového procesu.

Definice: Nechť Φ je Poissonův proces na $\mathbb{R}^d$. Existuje-li funkce intenzity λ a je konstantní, potom říkáme, že Φ je *homogenní Poissonův proces s intenzitou* λ.

Poznámka: Je-li Φ homogenní Poissonův proces na $\mathbb{R}^d$, pak je stacionární a izotropní.

Důkaz. Rozdělení každého bodového procesu Φ je jednoznačně určeno prázdnými pravděpodobnostmi, tj. pravděpodobnostmi $\mathbb{P}(\Phi(B) = 0)$, $B \in \mathcal{B}^d_o$, viz [9], Důsledek 4.2. Označme $|\cdot|$ Lebesgueovu míru v $\mathbb{R}^d$. Tato míra je translačně i rotačně invariantní mírou na $\mathbb{R}^d$. Proces Φ je homogenní, tedy existuje $\lambda > 0$ takové, že pro míru intenzity procesu je $\Lambda(\cdot) = \lambda|\cdot|$. Počítejme

$$\mathbb{P}(\Phi(B + y) = 0) = \mathrm{e}^{-\Lambda(\mathrm{B}+\mathrm{y})} = \mathrm{e}^{-\lambda|\mathrm{B}+\mathrm{y}|} = \mathrm{e}^{\lambda|\mathrm{B}|} = \mathbb{P}(\Phi(\mathrm{B}) = 0).$$

Nechť ρ_α značí rotaci okolo počátku o úhel α. Potom

$$\mathbb{P}(\Phi(\rho_\alpha(B)) = 0) = \mathrm{e}^{-\Lambda(\rho_\alpha(\mathrm{B}))} = \mathrm{e}^{\lambda|\rho_\alpha(\mathrm{B})|} = \mathrm{e}^{-\lambda|\mathrm{B}|} = \mathbb{P}(\Phi(\mathrm{B}) = 0).$$

Tedy proces je stacionární a izotropní. □

1.3 Kótované bodové procesy

Kótovaný bodový proces dostaneme z bodového procesu, přiřadíme-li každému jeho bodu nějakou náhodnou hodnotu (kótu).

Definice: Mějme separabilní úplný metrický prostor $\mathcal{Z}$. *Kótovaný bodový proces* je bodový proces Φ na $\mathbb{R}^d \times \mathcal{Z}$ takový, že $\Phi(\cdot \times \mathcal{Z})$ je bodový proces na $\mathbb{R}^d$.
Kótovaný proces Φ na $\mathbb{R}^d \times \mathcal{Z}$ je *stacionární*, jestliže je stacionární bodový proces $\Psi(\cdot) = \Phi(\cdot \times \mathcal{Z})$.

Poznámka: Nadále předpokládejme, že proces Ψ je jednoduchý. Vzhledem k poznámce za definicí bodového procesu lze na kótovaný bodový proces Φ pohlížet jako na náhodnou posloupnost $\Phi = \{(X_i, Z_i)\}$ takovou, že $\Psi = \{X_i\}$ tvoří bodový proces na $\mathbb{R}^d$ a $\{Z_i\}$ jsou náhodné elementy s hodnotami v $\mathcal{Z}$.

Uvažujme nyní případ Poissonova procesu. Pokud Φ je stacionární Poissonův proces na $\mathbb{R}^d \times \mathcal{Z}$, potom $\Psi(\cdot) = \Phi(\cdot \times \mathcal{Z})$ je stacionární Poissonův proces na $\mathbb{R}^d$.

Věta 1.2 *Nechť $\Phi = \{(X_i, Z_i)\}$ je stacionární Poissonův kótovaný bodový proces. Potom kóty $\{Z_i\}$ jsou nezávislé stejně rozdělené a nezávislé na $\Psi = \{X_i\}$.*

Důkaz. [10], Věta 3.4.8.

1.4 Proces částic

Nechť $\mathcal{K}' = \mathcal{K}(\mathbb{R}^d) \setminus \{\emptyset\}$ je systém všech neprázdných kompaktních podmnožin v $\mathbb{R}^d$. Prostor $\mathcal{K}'$ opatřený Hausdorffovou metrikou je separabilní úplný lokálně kompaktní – viz [9], Důsledek 11.1. Pro každou $K \in \mathcal{K}'$ označme $c(K)$ její významný bod (např. lexikografické minimum). Předpokládejme, že zobrazení $c : \mathcal{K}' \to \mathbb{R}^d$ je měřitelné a $c(x + K) = x + c(K)$ pro každé $x \in \mathbb{R}^d$. Nechť dále $\mathcal{K}'_0 = \{K \in \mathcal{K}';\ c(K) = 0\}$. Bodový proces Φ na $\mathcal{K}'$ označujeme jako *proces částic*. Můžeme ho zapsat jako kótovaný bodový proces $\Phi = \{(X_i, \Xi_i)\}$, kde $\{X_i\}$ jsou body bodového procesu $\Psi = \sum_{i\geq 1} \delta_{X_i}$ na $\mathbb{R}^d$ a $\{\Xi_i\}$ jsou náhodné kompaktní množiny v $\mathcal{K}'_0$. Potom *proces částic* je kótovaný bodový proces Φ na $\mathbb{R}^d \times \mathcal{K}'_0$. Častěji budeme body tohoto procesu psát ve tvaru $X_i + \Xi_i$.

Nyní si uvedeme verzi Campbellovy věty pro stacionární procesy částic, která se nám bude hodit v kapitole 2.

Věta 1.3 (Campbell-Mecke) *Uvažujme stacionární proces částic Φ na $\mathcal{K}'$. Pro libovolnou měřitelnou funkci $f : \mathcal{K}' \to \mathbb{R}^+$ je*

$$\mathbb{E} \sum_i f(X_i + \Xi_i) = \lambda \int_{\mathbb{R}^d} \int_{\mathcal{K}'_0} f(x + K_0) \Lambda_0(\mathrm{d}K_0)\, \mathrm{d}x, \tag{1.6}$$

kde λ je intenzita procesu $\{X_i\}$ a Λ_0 je rozdělení kót na $\mathcal{K}'_0$.

Definice: Rozdělení Λ_0 z věty 1.3 se nazývá *rozdělení typického zrna*. Náhodná veličina s rozdělením Λ_0 se nazývá *typické zrno*.

Poznámka: Rovnost (1.6) lze přepsat ve tvaru

$$\mathbb{E} \sum_i f(X_i + \Xi_i) = \lambda \int_{\mathbb{R}^d} \mathbb{E}^0 f(x + \Xi_0) \mathrm{d}x = \lambda \mathbb{E}^0 \int_{\mathbb{R}^d} f(x + \Xi_0)\, \mathrm{d}x, \tag{1.7}$$

kde Ξ_0 je typické zrno a $\mathbb{E}^0$ značí střední hodnotu vzhledem k Λ_0.

1.5 Bodové procesy úseček

Označme $\mathcal{S}$ množinu všech úseček kladné konečné délky v $\mathbb{R}^d$. Pro každou úsečku $S \in \mathcal{S}$ nechť $c(S)$ je její lexikografické minimum. Každá úsečka je jednoznačně určena svým význačným bodem $c(S)$, délkou $t \in (0, \infty)$ a směrem $\theta \in \mathcal{U}_d$, kde $\mathcal{U}_d$ je prostor všech lineárních jednodimenzionálních podprostorů v $\mathbb{R}^d$.

Definice: *(Stacionární Poissonův) bodový proces úseček* je (stacionární Poissonův) bodový proces $\Phi : (\Omega, \Sigma, \mathbb{P}) \to (\mathcal{N}, \mathfrak{N})$ na $\mathcal{S}$.

Nadále až do odvolání uvažujme, že Φ je stacionární Poissonův bodový proces úseček. Rozdělení typického zrna Λ_0 je pravděpodobnostní míra na $\mathcal{S}_0 = \{S \in \mathcal{S} : c(S) = 0\}$. Tento prostor je izomorfní prostoru $\mathbb{R}^+ \times \mathcal{U}_d$, budeme proto $\mathcal{S}_0$ ztotožňovat s $\mathbb{R}^+ \times \mathcal{U}_d$. Každá úsečka $S_0 \in \mathcal{S}_0$ je určena svou délkou t a směrem θ. V případě potřeby ji budeme značit $S_0(t, \theta)$. Dále nechť $\mathcal{D}$ značí rozdělení délek typické úsečky S a ρ rozdělení směrů, tedy

$$\begin{aligned} \mathcal{D}(\cdot) &= \Lambda_0(\cdot \times \mathcal{U}_d), \\ \rho(\cdot) &= \Lambda_0(\mathbb{R}^+ \times \cdot). \end{aligned}$$

Pokud ρ je rovnoměrné rozdělení na $\mathcal{U}_d$, řekneme, že proces Φ je *izotropní*.

Někdy je výhodnější pohlížet na Poissonův bodový proces úseček jako na kótovaný bodový proces, kde významné body $\{c(S),\ S \in \Phi\}$ úseček tvoří (stacionární) Poissonův bodový proces Ψ na $\mathbb{R}^d$ a prostor $\mathbb{R}^+ \times \mathcal{U}_d$ tvoří prostor kót.

Kapitola 2

Neparametrické metody odhadu rozdělení délky úseček

Sledujeme-li v praxi realizaci bodového procesu úseček (obecně jakéhokoliv procesu kompaktních množin) na $\mathbb{R}^d$, jsme většinou omezeni na nějaké ohraničené okno W. Potom ty z pozorovaných úseček, které protínají hranici ∂W okna W, nevidíme úplně. Chceme-li pomocí dat pozorovaných ve W odhadnout rozdělení délek úseček tohoto procesu, musíme se rozhodnout, jakým způsobem naložíme s oněmi „neúplnými“ úsečkami. Možností je více.

Jsme-li například schopni zjistit, jak vypadají pozorované úsečky i vně okna W, můžeme zahrnout všechny úsečky pozorované ve W a odhad učinit z celkových délek úseček (tzv. *plus sampling*). Takový odhad ovšem bude vychýlený, neboť upřednostňuje delší úsečky. Nebo naopak, nevíme-li nic o procesu vně výběrového okna, můžeme pro odhad použít pouze úsečky cele obsažené ve W (tzv. *minus sampling*). Ani tento odhad ovšem není nestranný, neboť tentokrát upřednostňuje úsečky menších délek (například úsečky delší než je $\mathrm{diam}(W)$ nemají vůbec šanci být do výběru zahrnuty). Způsob, jakým lze v těchto případech snížit výběrovou odchylku, je upravit „příspěvek“ každé pozorované úsečky pomocí vážení. Tento způsob je typický pro odhady Horvitzova-Thompsonova typu a je ve stochastické geometrii častý (viz [2]).

Další možností je pohlížet na tyto „okrajové efekty“ jako na náhodné censorování a využít při odhadu teorie analýzy přežití. To vede ke Kaplanovu-Meierovu odhadu.

V této kapitole se budeme zabývat některými neparametrickými odhady rozdělení délek úseček založenými na různých výběrových metodách.

Označení: *Výběrovým pravidlem* budeme nazývat měřitelnou funkci $I : \mathcal{K}' \to \{0, 1\}$. To znamená, že částice K je vybrána, právě když $I(K) = 1$.

2.1 Horvitzův-Thompsonův odhad

Předpokládejme nyní, že máme pevně dané konvexní okno $W \subset \mathbb{R}^d$. Symbolem $|W|$ budeme značit Lebesgueovu míru W. V tomto okně pozorujeme realizaci stacionárního procesu částic s intenzitou λ a kompaktním typickým zrnem Ξ_0 s rozdělením Λ_0. Nechť I je některé výběrové pravidlo. Jak už bylo řečeno v úvodu k této kapitole, v závislosti na volbě výběrového pravidla může být náš odhad vychýlený. Definujme nyní váhu

$$\tau(K) = \int_{\mathbb{R}^d} I(x+K)\,\mathrm{d}x, \qquad K \in \mathcal{K}'.$$

Funkce τ je úměrná pravděpodobnosti, že objekt K je pozorován. Z Campbellovy-Meckeho formule (1.7) plyne

$$\begin{aligned}\mathbb{E}\sum_{i\geq 1}\frac{I(X_i+\Xi_i)}{\tau(X_i+\Xi_i)} &= \lambda\int_{\mathcal{K}'_0}\int_{\mathbb{R}^d}\frac{I(x+K_0)}{\tau(x+K_0)}\,\mathrm{d}x\,\Lambda_0(\mathrm{d}K_0)\\ &= \lambda\int_{\mathcal{K}'_0}\frac{1}{\tau(K_0)}\int_{\mathbb{R}^d} I(x+K_0)\,\mathrm{d}x\,\Lambda_0(\mathrm{d}K_0) = \lambda.\end{aligned}$$

Tedy

$$\widehat{\lambda} = \sum_{i\geq 1}\frac{I(X_i+\Xi_i)}{\tau(X_i+\Xi_i)}$$

je nestranný odhad intenzity λ.

Obdobně se dá pro libovolnou nezápornou translačně invariantní (tj. takovou, že $f(x+K) = f(K),\ \forall x,\ \forall K$) měřitelnou funkci f na $\mathcal{K}'$ spočítat

$$\mathbb{E}\sum_{i\geq 1}\frac{I(X_i+\Xi_i)}{\tau(X_i+\Xi_i)}\,f(X_i+\Xi_i) = \lambda\,\mathbb{E}^0 f(\Xi_0),$$

kde $\mathbb{E}^0$ je střední hodnota vzhledem k rozdělení Λ_0.

Podílově nestranný odhad $\mathbb{E}^0 f(\Xi_0)$ Horvitzova-Thompsonova typu je tedy

$$\frac{1}{\widehat{\lambda}}\sum_{i\geq 1}\frac{I(X_i+\Xi_i)}{\tau(X_i+\Xi_i)}\,f(X_i+\Xi_i).$$

Abychom mohli takový odhad učinit, je nutný předpoklad, že $\tau(X_i+\Xi_i)$ známe a že $\tau(\Xi_0)$ je skoro jistě kladná pro typické zrno Ξ_0.

Nyní shrneme tyto poznatky pro rozdělení délek stacionárního (Poissonova) procesu úseček do následujícího tvrzení.

Tvrzení 2.1 *Nechť $\Phi = \{(X_i, S_i)\}$ je stacionární proces úseček, $\Psi = \{X_i\}$. Nechť $L(S)$ značí délku úsečky $S \in \mathcal{S}$. Nechť dále $F(t) = \mathbb{P}[L(S_0) \leq t]$ je distribuční funkce délky typické úsečky S_0. Potom*
(i) (minus-sampling) *uvažujeme-li pouze úsečky cele obsažené ve W, je*

$$\widehat{F}_1(t) = \frac{1}{\widehat{\lambda}_1} \sum_{i \geq 1} \mathbf{1}_{\{X_i + S_i \subseteq W\}} \frac{\mathbf{1}_{(-\infty,t]}(L(S_i))}{|W \ominus \check{S}_i|}, \tag{2.1}$$

kde

$$\widehat{\lambda}_1 = \sum_{i \geq 1} \frac{\mathbf{1}_{\{X_i + S_i \subseteq W\}}}{|W \ominus \check{S}_i|}$$

a

$$W \ominus \check{S}_i = \{x \in W;\ x + S_i \subseteq W\},$$

podílově nestranný odhad distribuční funkce F.
(ii) (plus-sampling) *uvažujeme-li všechny úsečky protínající W, je*

$$\widehat{F}_2(t) = \frac{1}{\widehat{\lambda}_2} \sum_{i \geq 1} \mathbf{1}_{\{(X_i + S_i) \cap W \neq \emptyset\}} \frac{\mathbf{1}_{(-\infty,t]}(L(S_i))}{|W \oplus \check{S}_i|}, \tag{2.2}$$

kde

$$\widehat{\lambda}_2 = \sum_{i \geq 1} \frac{\mathbf{1}_{\{(X_i + S_i) \cap W \neq \emptyset\}}}{|W \oplus \check{S}_i|}$$

a

$$W \oplus \check{S}_i = \{x \in \mathbb{R}^d;\ (x + S_i) \cap W \neq \emptyset\},$$

podílově nestranný odhad distribuční funkce F.
(iii) (unbiased-sampling) *uvažujeme-li všechny úsečky, jejichž význačný bod leží uvnitř okna W, je*

$$\widehat{F}_3(t) = \frac{1}{\widehat{\lambda}_3} \sum_{i \geq 1} \mathbf{1}_{\{X_i \in W\}} \frac{\mathbf{1}_{(-\infty,t]}(L(S_i))}{|W|}, \tag{2.3}$$

kde

$$\widehat{\lambda}_3 = \frac{\Psi(W)}{|W|}$$

podílově nestranný odhad distribuční funkce F.

Poznámka: Asymptotické chování odhadu $\widehat{F}_1(t)$ je studováno v [5].

2.2 Kaplanův-Meierův odhad

Jiný pohled na výběr dat, která použijeme pro odhad délek úseček, nám poskytne teorie přežití. Na úsečky pozorované v okně W budeme nyní pohlížet jako na náhodně censorované. V této podkapitole si uvedeme jen některá potřebná fakta týkající se náhodného censorování. Teorii přežití se budeme více věnovat v následující podkapitole.

Nechť $T_1, \ldots, T_n$ jsou nezávislé, stejně rozdělené kladné náhodné veličiny s distribuční funkcí F. Obvykle jsou to časy výskytu nějaké události, v našem případě se jedná o délky úseček. Tyto náhodné veličiny ovšem nejsou pozorovány. Namísto nich pozorujeme pouze censorovaná data $\widetilde{T}_i = \min(T_i, C_i)$ a indikátory censorování $D_i = \mathbf{1}_{\{T_i \leq C_i\}}$, kde $C_1, \ldots, C_n$, *časové censory*, jsou nezávislé, stejně rozdělené náhodné veličiny nezávislé na $\{T_i\}$. Za těchto předpokladů je Kaplanův-Meierův odhad distribuční funkce F roven

$$\widehat{F}_{KM}(t) = 1 - \prod_{s \leq t} \left(1 - \frac{\sum_{i \geq 1} \mathbf{1}_{\{\widetilde{T}_i = s, D_i = 1\}}}{\sum_{i \geq 1} \mathbf{1}_{\{\widetilde{T}_i \geq s\}}} \right).$$

Nyní se vraťme k procesu úseček. Abychom se vyhnuli výběrovému vychýlení, budeme uvažovat pouze ty úsečky, jejichž lexikografické minimum leží uvnitř okna W. Pozorujeme tedy náhodný počet $\Psi(W)$ nezávislých veličin $L(S_i)$, respektive jejich censorované hodnoty C_i. Indikátor censorování definujeme jako $D_i = \mathbf{1}_{\{L(S_i) \leq C_i\}} = \mathbf{1}_{\{X_i + S_i \subseteq W\}}$, a $\widetilde{T}_i = \min(L(S_i), C_i) = L(S_i \cap (W - X_i))$. Kaplanův-Meierův odhad pro stacionární bodový proces úseček potom vypadá následovně

$$\widehat{F}_{KM}(t) = 1 - \prod_{s \leq t} \left(1 - \frac{\sum_{i \geq 1} \mathbf{1}_{\{L(S_i \cap (W - X_i)) = s\}} \mathbf{1}_{\{X_i + S_i \subseteq W\}}}{\sum_{i \geq 1} \mathbf{1}_{\{L(S_i \cap (W - X_i)) \geq s\}} \mathbf{1}_W(X_i)} \right). \tag{2.4}$$

2.3 Maximálně věrohodný odhad

Dalším klasickým typem neparametrického odhadu je maximálně věrohodný odhad. Je založen na maximalizaci tzv. (logaritmické) věrohodnostní funkce. Ne vždy je ovšem taková maximalizace přímočará na výpočet. Jednou z metod, jak maximálně věrohodný odhad (přibližně) vypočítat, je EM algoritmus. Je to iterativní algoritmus založený na vyjádření maximálně věrohodného odhadu pomocí tzv. *self-consistency equations*.

Nejprve se podíváme na nejjednodušší případ procesu úseček na přímce a odvodíme si pro tuto situaci EM algoritmus. Potom se budeme věnovat situaci v $\mathbb{R}^2$. Budeme vycházet z práce [11].

2.3.1 Proces na přímce

Nechť nyní je okno pozorování W omezený interval $(0, \tau) \subset \mathbb{R}$. Proces úseček na přímce si lze představit jako pobyt pacientů v nemocnici. Pacienti přicházejí

v náhodných časech X_i, které tvoří stacionární Poissonův proces na $\mathbb{R}$ s intenzitou λ. Časy přežití (doba pobytu) jednotlivých pacientů jsou potom kladné i.i.d. náhodné veličiny T_i nezávislé na procesu X_i s distribuční funkcí F. Označme $\mu = \int_0^\infty t \,\mathrm{d}F(t)$ střední délku pobytu. Toto vše definuje Poissonův bodový proces $\Phi = \{(X_i, T_i);\ i \in \mathbb{N}\}$ na $\mathbb{R} \times \mathbb{R}^+$ s mírou intenzity

$$\Lambda_0(B \times C) = \lambda |B| \int_C \mathrm{d}F(t).$$

My pozorujeme pouze ty části dob pobytu pacientů v nemocnici, které protínají interval $(0, \tau)$. Označme

$$A = \{(x,t) \in \mathbb{R} \times \mathbb{R}^+;\ [x, x+t] \cap (0,\tau) \neq \emptyset\}$$

množinu v $\mathbb{R} \times \mathbb{R}^+$, ve které jsme schopni pozorovat body procesu Φ.

Pro ty z pacientů, jež přišli před časem 0, pozorujeme dvojici (W_j, E_j), kde

$$W_j = \min(X_j + T_j, \tau), \quad E_j = \mathbf{1}_{\{X_j + T_j \leq \tau\}},$$

a pro pacienty, kteří přijeli v intervalu $(0,\tau)$, dvojici (Z_i, D_i), kde

$$Z_i = \min(T_i, \tau - X_i), \quad D_i = \mathbf{1}_{\{T_i \leq \tau - X_i\}}.$$

Celkem tedy máme čtyři druhy pozorovaní – zleva censorované, oboustranně censorované, necensorované a zprava censorované.

Nechť nyní je $N = \Phi(A)$ počet bodů procesu Φ, které padnou do množiny A. Protože Φ je Poissonův proces, tak podmíněně při $N = n$, celkovém počtu pozorovaných pacientů, dostáváme n nezávislých pozorování v A. Náhodná veličina N má Poissonovo rozdělení s parametrem $\Lambda(A) = \int_A \lambda \,\mathrm{d}x \,\mathrm{d}F(t) = \lambda\,(\tau + \mu)$. Tedy množina bodů $\Phi \cap A$ má stejné rozdělení jako množina v nezávislém, stejně rozděleném výběru o rozsahu n s pravděpodobnostní mírou

$$\begin{aligned} \mathbf{1}_A(x,t) \cdot \frac{\lambda}{\lambda(\tau+\mu)} \,\mathrm{d}x \,\mathrm{d}F(t) &= \mathbf{1}_A(x,t) \cdot \frac{\tau+t}{\tau+\mu} \,\mathrm{d}F(t) \cdot \frac{1}{\tau+t} \,\mathrm{d}x \\ &= \mathbf{1}_A(x,t) \cdot \mathrm{d}V(t) \cdot \frac{1}{\tau+t} \,\mathrm{d}x, \end{aligned} \tag{2.5}$$

kde

$$\mathrm{d}V(t) = \frac{\tau+t}{\tau+\mu} \,\mathrm{d}F(t). \tag{2.6}$$

Snadno se ukáže, že funkce V je distribuční funkce. Je zřejmé, že stejnou pravděpodobnostní míru jako (2.5) obdržíme pro náhodný výběr (X_i, T_i) na $\mathbb{R} \times \mathbb{R}^+$ o rozsahu n, kde T_i jsou i.i.d. veličiny s distribuční funkcí V a X_i při daném $T_i = t_i$ jsou rovnoměrně rozdělené na $(-t_i, \tau)$. Tím jsme přešli od funkce F k reparametrizaci V. Náhodná veličina T_i má tedy při podmínění počtu bodů $N = n$ v $\Phi \cap A$ distribuční funkci $\int_0^x (\tau+u)/(\tau+\mu) \,\mathrm{d}F(u)$. Důkaz jednoznačné korespondence mezi F a V na $(0,\tau)$ lze nalézt v [11].

Označme nyní po řadě F^l, F^r, F^d a F^u subdistribuční funkce pozorovaných délek pro zleva, zprava, oboustranně censorované a necensorované úsečky. Jednoduchou integrací lze pro $u \in [0, \tau)$ spočítat

$$\begin{aligned} F^l(u) &= \int_0^u \int_{-w}^0 \frac{1}{\tau + w} \,\mathrm{d}x \,\mathrm{d}V(w) + \int_u^\infty \int_{-w}^{-w+u} \frac{1}{\tau + w} \,\mathrm{d}x \,\mathrm{d}V(w) \\ &= \int_0^u \frac{w}{\tau + w} \,\mathrm{d}V(w) + u \int_u^\infty \frac{1}{\tau + w} \,\mathrm{d}V(w), \end{aligned}$$

a tedy

$$\mathrm{d}F^l(u) = \frac{u}{\tau + u} \,\mathrm{d}V(u) - \frac{u}{\tau + u} \,\mathrm{d}V(u) + \int_u^\infty \frac{1}{\tau + w} \,\mathrm{d}V(w) \,\mathrm{d}u.$$

Obdobné výsledky lze obdržet i pro ostatní subdistribuční funkce. Celkem dostáváme

$$\begin{aligned} \mathrm{d}F^l(u) &= \mathbf{1}_{[0,\tau)}(u) \int_u^\infty \frac{1}{\tau + w} \,\mathrm{d}V(w) \,\mathrm{d}u \\ &= \mathbf{1}_{[0,\tau)}(u) \cdot g(u) \,\mathrm{d}u \qquad (2.7) \\ \mathrm{d}F^d(u) &= \int_\tau^\infty \frac{w - \tau}{\tau + w} \,\mathrm{d}V(w) \,\mathrm{d}\delta_\tau(u) \\ &= h \,\mathrm{d}\delta_\tau(u) \qquad (2.8) \\ \mathrm{d}F^r(u) &= \mathbf{1}_{[0,\tau)}(u) \int_u^\infty \frac{1}{\tau + w} \,\mathrm{d}V(w) \,\mathrm{d}u \\ &= \mathbf{1}_{[0,\tau)}(u) \cdot g(u) \,\mathrm{d}u \qquad (2.9) \\ \mathrm{d}F^u(u) &= \mathbf{1}_{[0,\tau)}(u) \frac{\tau - u}{\tau + u} \,\mathrm{d}V(u), \qquad (2.10) \end{aligned}$$

kde g a h jsou definovány jako

$$g(y) = \int_y^\infty \frac{1}{\tau + w} \,\mathrm{d}V(w), \qquad h = \int_\tau^\infty \frac{w - \tau}{\tau + w} \,\mathrm{d}V(w).$$

Integrál h je vlastně pravděpodobnost dvojného censorování. Dále je zřejmé, že subdistribuční funkce délek pro jednostranně (ať už zleva či zprava) censorované úsečky F^s je rovna součtu F^l a F^r a je tedy $\mathrm{d}F^s(u) = \mathrm{d}F^l(u) + \mathrm{d}F^r(u) = 2g(u)\mathrm{d}u,\ u \in [0, \tau)$.

Z předchozích úvah plyne, že

$$1 = F^u(\tau) + 2F^s(\tau) + F^d(\tau) = V(\tau-) + 2\tau g(\tau) + h,$$

kde $V(\tau-)$ značí limitu zleva v bodě τ.

Dále se dají snadno s použitím předchozího odvodit následující rovnosti

$$\begin{aligned}
\frac{\mu}{\tau+\mu} &= \int_0^\tau \frac{w}{\tau+w}\,\mathrm{d}V(w) + \tau g(\tau) + h, \\
\frac{\tau}{\tau+\mu} &= \int_0^\tau \frac{\tau}{\tau+w}\,\mathrm{d}V(w) + \tau g(\tau), \\
g(x) &= \int_x^\tau \frac{1}{\tau+w}\,\mathrm{d}V(w) + g(\tau).
\end{aligned} \tag{2.11}$$

Nyní pomocí rovnic pro subdistribuční funkce odvodíme věrohodnostní funkci pro nemocniční model. Mějme realizaci $\{(x_i, t_i),\ i = 1, \ldots, n\}$ bodového procesu Φ. Pozorovaná data tvoří dvojice $(w_j, e_j),\ j = 1, \ldots, m$ a $(z_i, d_i),\ i = 1, \ldots, n-m$. Nechť $t_1 < t_2 < \cdots < t_r$ jsou uspořádané hodnoty w_j a z_i, pro něž je $e_j = 1$ nebo $d_i = 0, 1$ (tj. jedná se o necensorovaná nebo jednostranně censorovaná pozorování) a nechť ϕ_i a γ_i jsou počty necensorovaných, resp. jednostranně censorovaných hodnot t_i. Potom v řeči funkce V dostáváme věrohodnostní funkci

$$\begin{aligned}
l(V) &\propto \prod_{i=1}^{r} (\mathrm{d}V(t_i))^{\phi_i} \left(\int_{t_i}^{\infty} \frac{1}{\tau+w}\,\mathrm{d}V(w) \right)^{\gamma_i} \cdot \left(\int_{\tau}^{\infty} \frac{w-\tau}{\tau+w}\,\mathrm{d}V(w) \right)^{n-r} \\
&= \prod_{i=1}^{r} (\mathrm{d}V(t_i))^{\phi_i} \left(\int_{t_i}^{\tau} \frac{1}{\tau+w}\,\mathrm{d}V(w) + g(\tau) \right)^{\gamma_i} \cdot h^{n-r}.
\end{aligned} \tag{2.12}$$

Definice: Pro námi zavedený nemocniční model definujeme na intervalu $[0, \tau)$, $\tau < \infty$, empirické subdistribuční funkce následovně

$$\begin{aligned}
F_n^d(t) &\equiv \frac{1}{n}\#\{\text{dvojně censorovaná pozorování} \le t\} \\
F_n^s(t) &\equiv \frac{1}{n}\#\{\text{jednostranně censorovaná pozorování} \le t\} \\
F_n^u(t) &\equiv \frac{1}{n}\#\{\text{necensorovaná pozorování} \le t\}.
\end{aligned}$$

Označení: Nechť $\mathcal{V}$ je množina všech distribučních funkcí na $[0, \tau)$. Symbolem $\mathcal{V}_\tau$ budeme značit množinu

$$\mathcal{V}_\tau = \{(V, h);\ V \in \mathcal{V};\ h \in [0, \infty);\ h \le 1 - V(\tau-) \text{ a „=“ právě když } h = 0\}.$$

Definice: *Logaritmickou věrohodnostní funkcí* funkce V (v kontextu našeho modelu) nazveme funkci

$$\Delta(V, h) \equiv \int_0^\tau \log(\mathrm{d}V(t)) \cdot \mathrm{d}F_n^u(t) + \int_0^\tau \log(g(t)) \cdot \mathrm{d}F_n^s(t) + \log(h) \cdot F_n^d(\tau),$$

kde $(V, h) \in \mathcal{V}_\tau$ a

$$g(t) = \int_t^\tau \frac{1}{\tau+w}\,\mathrm{d}V(w) + g(\tau),$$

a $g(\tau)$ je dáno vztahem

$$V(\tau-)2\tau g(\tau) + h = 1.$$

Dále $(\widehat{V}_n, \widehat{h}_n)$ je *neparametrický maximálně věrohodný odhad* pro (V, h) v $\mathcal{V}_\tau$, jestliže

$$\Delta(\widetilde{V}, \widetilde{h}) - \Delta(V_0, h_0) \leq \Delta(\widehat{V}_n, \widehat{h}_n) - \Delta(V_0, h_0),$$

pro každou dvojici $(\widetilde{V}, \widetilde{h}) \in \mathcal{V}_\tau$ takovou, že $\widetilde{V} \ll V_0$, $\widehat{V}_n \ll V_0$ pro $(V_0, h_0) \in \mathcal{V}_\tau$.

Věta 2.1 *Neparametrický maximálně věrohodný odhad $(\widehat{V}_n, \widehat{h}_n)$ existuje v $\overline{\mathcal{V}}_\tau$, je určen jednoznačně a platí*

$$\begin{aligned} \mathrm{d}\widehat{V}_n(t) &= \mathrm{d}F_n^u(t) + \int_0^t \frac{1}{\widehat{g}_n(v)}\,\mathrm{d}F_n^s(v) \cdot \frac{1}{\tau + t}\,\mathrm{d}\widehat{V}_n(t) \qquad (2.13)\\ &= \mathrm{d}F_n^u(t) + \\ &\quad + \int_{v=0}^{v=t} \left(\int_v^\tau \frac{1}{\tau + w}\,\mathrm{d}\widehat{V}_n(w) + \widehat{g}_n(\tau) \right)^{-1} \mathrm{d}F_n^s(v) \cdot \frac{1}{\tau + t}\,\mathrm{d}\widehat{V}_n(t) \end{aligned}$$

a

$$\widehat{h}_n = F_n^d(\tau) = \frac{n - r}{n}, \qquad (2.14)$$

$$2\tau\widehat{g}_n(\tau) = 1 - \widehat{h}_n - \widehat{V}_n(\tau). \qquad (2.15)$$

Důkaz. viz [11], str. 23, Tvrzení 1.1.5.1.

Uvedením vzorce pro maximálně věrohodný odhad se dostáváme k samotnému EM algoritmu. Tento iterativní algoritmus slouží právě k výpočtu maximálně věrohodného odhadu distribuční funkce V. Předpokládejme, že míra je soustředěna pouze v bodech pozorování $t_1, \ldots, t_r$. Symbol $\mathrm{d}\widehat{V}_n(t_i)$ značí velikost skoku funkce $\widehat{V}_n$ v bodě t_i. Tedy pro $t \notin \{t_1, \ldots, t_r\}$ je $\mathrm{d}\widehat{V}_n(t) = 0$. Nyní zvolíme počáteční nastavení (V_n^0, h_n^0) a to tak, aby soustřeďovalo hmotu ve všech bodech pozorování. Algoritmus pak probíhá takto : v $(k+1)$-ní iteraci dostáváme novou distribuční funkci $\widehat{V}_n^{k+1}$ následovně

$$\begin{aligned} \widehat{g}_n^{k+1}(t) &= \int_t^\tau \frac{1}{\tau + w}\,\mathrm{d}\widehat{V}_n^k(w) + \widehat{g}_n^k(\tau),\\ \mathrm{d}\widehat{V}_n^{k+1}(t) &= \mathrm{d}F_n^u(t) + \int_0^t \frac{1}{g_n^{k+1}(v)}\,\mathrm{d}F_n^s(v) \cdot \frac{1}{\tau + t}\,\mathrm{d}\widehat{V}_n^k(t),\\ 2\,\tau\,\widehat{g}_n^{k+1}(\tau) &= 1 - \widehat{h}_n - \widehat{V}_n^{k+1}(\tau). \end{aligned}$$

Nově vzniklá distribuční funkce v každé iteraci zvyšuje věrohodnost a bylo dokázáno, že EM algoritmus za daných předpokladů konverguje k maximálně věrohodnému odhadu $(\widehat{V}_n, \widehat{h}_n)$.

Původní funkci F dostaneme zpětným vyjádřením $\mathrm{d}F(t)$ z (2.6). Parametr μ je sice neznámý, ale z (2.11) plyne, že

$$\mu + \tau = \left(\int_0^\tau \frac{1}{\tau + w} \, \mathrm{d}V(w) + g(\tau) \right)^{-1}.$$

2.3.2 Dvoudimenzionální proces úseček

V případě dvoudimenzionálního procesu úseček se situace komplikuje, neboť kromě rozdělení délek se nám oproti procesu úseček na přímce nově objevuje také rozdělení směrů. Není-li toto rozdělení známé, může velice zkomplikovat výpočty. V takovém případě lze vše zjednodušit vhodnou volbou okna W. Ideální je kruhové okno, neboť pak často závislost na rozdělení směrů při výpočtech zcela vypadne. Dalšími relativně vhodnými volbami oken jsou jiné různé pravidelné obrazce (např. čtverec). V této sekci budeme pracovat s obecným konvexním oknem W, vzorce pro EM algoritmus si ovšem uvedeme pouze pro volbu kruhového a čtvercového okna.

Nechť $\Phi = \{(X_i, S_i)\}$ je stacionární Poissonův proces úseček na $\mathbb{R}^2$. Body $\{X_i\}$ tvoří stacionární Poissonův proces Ψ s intenzitou λ a každému bodu tohoto procesu je přiřazena úsečka S_i o délce $T_i \in (0, \infty)$ a směru $\Theta_i \in [0, \pi)$. Délky $T_1, T_2, \ldots$ jsou kladné i.i.d. veličiny s distribuční funkcí F a opět označme $\mu = \int t \, \mathrm{d}F(t)$ střední délku úseček. Distribuční funkci veličiny i.i.d. veličin $\Theta_1, \Theta_2, \ldots$ budeme značit K. Každá dvojice (T, Θ) je nezávislá s Poissonovým procesem Ψ. Předpokládejme, že rozdělení délek a směrů jsou nezávislá, tj. $\Lambda_0 = \mathcal{D} \times \rho$.

Realizaci procesu Φ pozorujeme skrz konvexní okno W. Bez újmy na obecnosti předpokládejme, že nejjižnější bod každé úsečky (tj. bod s nejmenší y-souřadnicí) je bodem procesu Ψ. Získaná data se opět dělí na necenzorované, jednostranně cenzorované a oboustranně cenzorované. Opět si definujeme množinu A

$$A = \{(x, t, \theta) \in \mathbb{R}^2 \times \mathbb{R}^+ \times [0, \pi);\ x + S(t, \theta) \cap W \neq \emptyset\}.$$

Náhodná veličina N bude opět značit počet bodů $\Phi \cap A$. Tato veličina má Poissonovo rozdělení s parametrem $\Lambda(A) = \int_A \lambda \, \mathrm{d}x \, \mathrm{d}F(t) \, \mathrm{d}K(\theta)$. Označme $\mathrm{diam}(W, \theta)$ diametr okna W ve směru θ. Integrací se snadno spočte, že

$$\int_A \lambda \, \mathrm{d}x \, \mathrm{d}F(t) \, \mathrm{d}K(\theta) = \lambda(|W| + \mu \, \mathbb{E}_K \, \mathrm{diam}(W)), \tag{2.16}$$

kde $\mathbb{E}_K \, \mathrm{diam}(W) = \int \mathrm{diam}(W, \theta) \, \mathrm{d}K(\theta)$.

Při podmínění $N = n$ mají tedy body množiny $\Phi \cap A$ pravděpodobnostní míru na $\mathbb{R}^2 \times \mathbb{R}^+ \times [0, \pi)$ rovnou

$$\mathbf{1}_A(x, t, \theta) \frac{\lambda \, \mathrm{d}x \, \mathrm{d}F(t) \, \mathrm{d}K(\theta)}{\lambda(|W| + \mu \mathbb{E}_K \, \mathrm{diam}(W))} = \mathrm{d}V(t) \mathrm{d}J(\theta | T = t) \mathrm{d}A(x | T = t, \Theta = \theta),$$

kde

$$\begin{aligned} \mathrm{d}V(t) &= \frac{|W| + t\,\mathbb{E}_K \operatorname{diam}(W)}{|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}F(t), \\ \mathrm{d}J(\theta|T=t) &= \frac{|W| + t \operatorname{diam}(W,\theta)}{|W| + t\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}K(\theta), \\ \mathrm{d}A(x|T=t,\Theta=\theta) &= \frac{1}{|W| + t \operatorname{diam}(W,\theta)}\,\mathrm{d}x \cdot \mathbf{1}_{A_{\theta,t}}(x), \end{aligned} \tag{2.17}$$

kde $A_{\theta,t}$ značí řez množinou A vedený bodem t ve směru θ.

Stejně jako v jednodimenzionálním případě jsme obdrželi stejnou pravděpodobnostní míru, jako kdybychom uvažovali náhodný výběr (X_i, T_i, Θ_i) o rozsahu n na $\mathbb{R}^2 \times \mathbb{R}^+ \times [0,\pi)$, kde T_i jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí V, směry při daném $T_i = t_i$ jsou z rozdělení $J(\cdot|T_i = t_i)$ a X_i jsou při daných $\Theta_i = \theta_i$ a $T_i = t_i$ rovnoměrně rozdělené na A_{θ_i,t_i}.

Další postup při odvozování subdistribučních funkcí a jejich vyjádření pomocí funkce V je obdobný jako v případě jednodimenzionálním. Protože však je odvozování v případě obecného konvexního okna zdlouhavé a pro cíl této práce ne tolik podstatné, omezíme se na uvedení vzorců pro maximálně věrohodný odhad. Podrobný postup výpočtů lze najít v [11], kap. 1.2.5.

Mějme tedy $t_1 < t_2 < \cdots < t_r$ uspořádané hodnoty r navzájem různých pozorování T_i a ϕ_i, γ_i a ζ_i počty necensorovaných, jednostranně censorovaných a dvojně censorovaných pozorování t_i. Potom věrohodnost založená na n nezávislých pozorováních (T_i, Θ_i) je úměrná

$$\prod_{i=1}^{r} \left(\frac{\mathrm{d}V(t_i)}{|W| + t_i\,\mathbb{E}_K \operatorname{diam}(W)} \right)^{\phi_i} (g(t_i))^{\gamma_i} (d(t_i,t_i))^{\zeta_i} \cdot \prod_{j=1}^{n} \mathrm{d}K(\theta_j), \tag{2.18}$$

kde funkce $g(\cdot), d(\cdot,\cdot)$ jsou definovány následovně

$$\begin{aligned} g(x) &= \int_x^\infty \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w), \\ d(x,y) &= \int_x^\infty \frac{w-y}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w). \end{aligned}$$

Přepíšeme-li věrohodnostní funkci pomocí funkce F namísto funkce V, je úměrná výrazu

$$\prod_{i=1}^{r} (\mathrm{d}F(t_i))^{\phi_i} (1-F(t_i))^{\gamma_i} \left(\int_{t_i}^\infty (1-F(w))\,\mathrm{d}w \right)^{\zeta_i} \frac{\prod_{j=1}^{n} \mathrm{d}K(\theta_j)}{(|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W))^n}.$$

Jak je vidět, věrohodnostní funkce závisí i na rozdělení směrů. V případě, že neznáme distribuční funkci K, získáme odhady $\widehat{F}$ a $\widehat{K}$ maximalizací této věrohodnosti. Neboť nelze od sebe tyto dvě věci oddělit, vypočítáme odhady pomocí

následujícího schématu. Nejprve určíme F maximalizací

$$(|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W))^{-n} \prod_{i=1}^{r} (\mathrm{d}F(t_i))^{\phi_i} (1 - F(t_i))^{\gamma_i} \left(\int_{t_i}^{\infty} (1 - F(w))\,\mathrm{d}w \right)^{\zeta_i}$$

pro zvolenou K. Potom pro danou F vypočteme K maximalizací

$$(|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W))^{-n} \prod_{j=1}^{n} \mathrm{d}K(\theta_j).$$

Neparametrický maximálně věrohodný odhad funkce K pro danou F je pak tvaru

$$\widehat{K}_n(\theta) = \frac{\int_{[0,\theta]} (|W| + \mu \operatorname{diam}(W, \eta))^{-1}\,\mathrm{d}L_n(\eta)}{\int_{[0,\pi]} (|W| + \mu \operatorname{diam}(W, \eta))^{-1}\,\mathrm{d}L_n(\eta)},$$

kde L_n je empirická distribuční funkce pozorovaných úhlů.

Nyní tedy můžeme předpokládat, že známe rozdělení směrů. Potom analogicky k jednorozměrnému případu můžeme vyjádřit maximálně věrohodný odhad funkce V. Funkci F dostaneme zpětným vyjádřením z (2.17). K tomu potřebujeme znát $|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W)$. Nechť P značí diametr okna W. Počítejme

$$\begin{aligned}
&\int_0^P \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w) + g(P) \\
&\quad = \int_0^P \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w) \quad + \quad \int_P^\infty \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w) \\
&\quad = \int_0^\infty \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w) \quad = \quad \int_0^\infty \frac{1}{|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}F(w) \\
&\quad = \frac{1}{|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W)}.
\end{aligned}$$

Tedy

$$|W| + \mu\,\mathbb{E}_K \operatorname{diam}(W) = \left(\int_0^P \frac{1}{|W| + w\,\mathbb{E}_K \operatorname{diam}(W)}\,\mathrm{d}V(w) + g(P) \right)^{-1}.$$

Neboť se budeme v simulacích zabývat pouze případem kruhového a čtvercového okna, uvedeme si vzorce pro EM algoritmus pouze pro tyto dvě volby okna. V obecnějším případě jsou rovnice poněkud složitější.

(1) Nechť W je kruh o poloměru $R > 0$. Potom EM algoritmus pro počáteční

nastavení $(V_n^0, h_n^0), g_n^0(2R)$ vypadá následovně

$$\begin{aligned}
\widehat{g}_n^{k+1}(t) &= \int_t^{2R} \frac{1}{|W| + 2uR}\,\mathrm{d}\widehat{V}_n^k(u) + \widehat{g}_n^k(2R) \\
\widehat{d}_n^{k+1}(t,t) &= \int_t^{2R} \frac{u-t}{|W| + 2uR}\,\mathrm{d}\widehat{V}_n^k(u) + \frac{1}{2R}\,\widehat{h}_n^k + (2R-t)\,\widehat{g}_n^k(2R) \\
\widehat{h}_n^{k+1} &= \widehat{d}_n^k\,(2R,2R) \int_0^{2R} \frac{1}{\widehat{d}_n^k(u,u)}\,\mathrm{d}F_n^d(u) \\
\widehat{g}_n^{k+1}(2R)(|W| + 4R^2) &= 1 - \widehat{h}_n^k - \widehat{V}_n^k(2R) \\
\mathrm{d}\widehat{V}_n^{k+1}(t) &= \mathrm{d}F_n^u(t) + \int_0^t \frac{1}{\widehat{g}_n^k(u)}\,\mathrm{d}F_n^s(u) \cdot \frac{1}{|W| + 2tR}\,\mathrm{d}\widehat{V}_n^k(t) \\
&\qquad + \int_0^t \frac{t-u}{\widehat{d}_n^k(u,u)}\,\mathrm{d}F_n^d(u) \cdot \frac{1}{|W| + 2tR}\,\mathrm{d}\widehat{V}_n^k(t).
\end{aligned}$$

(2) Nechť W je čtverec o straně $a > 0$. Potom EM algoritmus pro počáteční nastavení $(V_n^0, h_n^0), g_n^0(P)$ vypadá následovně

$$\begin{aligned}
\widehat{g}_n^{k+1}(t) &= \int_t^{P} \frac{1}{|W| + u\,\mathbb{E}_K\,\mathrm{diam}(W)}\,\mathrm{d}\widehat{V}_n^k(u) + \widehat{g}_n^k(P) \\
\widehat{d}_n^{k+1}(t,t) &= \int_t^{P} \frac{u-t}{|W| + u\,\mathbb{E}_K\,\mathrm{diam}(W)}\,\mathrm{d}\widehat{V}_n^k(u) + d(P,P) + (P-t)\,\widehat{g}_n^k(P) \\
\widehat{h}_n^{k+1} &= 2\,\widehat{d}_n^k\,(P,P) \int_0^{P} \frac{1}{\widehat{d}_n^k(u,u)}\,\mathrm{d}F_n^d(u) \\
\widehat{g}_n^{k+1}(P) &= (2|W| + a_1)^{-1}(1 - \widehat{h}_n^k - \widehat{V}_n^k(P)) \\
\mathrm{d}\widehat{V}_n^{k+1}(t) &= \mathrm{d}F_n^u(t) + \int_0^t \frac{1}{\widehat{g}_n^k(u)}\,\mathrm{d}F_n^s(u) \cdot \frac{1}{|W| + t\,\mathbb{E}_K\,\mathrm{diam}(W)}\,\mathrm{d}\widehat{V}_n^k(t) \\
&\qquad + \int_0^t \frac{t-u}{\widehat{d}_n^k(u,u)}\,\mathrm{d}F_n^d(u) \cdot \frac{1}{|W| + t\,\mathbb{E}_K\,\mathrm{diam}(W)}\,\mathrm{d}\widehat{V}_n^k(t),
\end{aligned}$$

kde

$$P = \mathrm{diam}(W) = a\sqrt{2}, \tag{2.19}$$

$$\mathbb{E}_K\,\mathrm{diam}(W) = 4a \int_0^{\frac{\pi}{4}} \frac{1}{\cos\theta}\,\mathrm{d}K(\theta) \tag{2.20}$$

a

$$a_1 = \int_0^{\pi} aP(|\cos\theta| + \sin\theta)\mathrm{d}K(\theta) - a^2. \tag{2.21}$$

Kapitola 3

Parametrický odhad rozdělení délek úseček

Kromě neparametrických metod odhadu rozdělení délek úseček existují také metody parametrické. Ty se užívají v případě, že typ rozdělení je nám známý, co nevíme je parametr ϑ tohoto rozdělení. SRE (stochastic restoration estimation) algoritmus patří mezi takové metody odhadu. Opět se užívá v případě, kdy máme k dispozici neúplná, censorovaná data. Je to iterativní Monte Carlo algoritmus, jehož výsledkem je homogenní markovský řetězec. Aritmetický průměr tohoto řetězce za určitých podmínek konverguje ke střední hodnotě řetězce.

SRE algoritmus pracuje ve dvou krocích. V prvním kroku každé iterace nejprve obnoví nepozorovanou část dat pomocí simulací s aktuálním odhadem parametru a pak na základě obnovených vytvoří nový odhad parametru. Výsledkem algoritmu je pak aritmetický průměr $\overline{\vartheta}_m$ těchto odhadů pro nějaké dostatečně velké m. V této kapitole si uvedeme obecný SRE algoritmus pro pozorování v (ne nutně konvexním) okně $W \subset \mathbb{R}^2$.

Podrobnější zpracování včetně příkladů pro exponenciální a logaritmicko-normální rozdělení lze nalézt v [4].

Mějme stacionární Poissonův bodový proces úseček $\Phi = \{(X_i, S_i)\}$ na $\mathbb{R}^2$ s intenzitou $\lambda > 0$. Jako významný bod opět volíme nejjižnější bod úsečky. Úsečka S_i je dána délkou $T_i \in [0, \infty)$ a směrem $\Theta_i \in [0, \pi)$. Předpokládejme, že rozdělení $\Lambda_0(\mathrm{d}t \times \mathrm{d}\theta|\vartheta)$ typické úsečky závisí na parametru ϑ.

Realizaci procesu Φ pozorujeme skrz okno $W \subset \mathbb{R}^2$. Označme

$$\begin{aligned} n &= \#\{X_i;\ X_i \in W\}, \\ n_1 &= \#\{X_i \in W;\ X_i + S_i \subseteq W\}, \\ n_2 &= \#\{X_i \in W;\ (X_i + S_i) \cap \partial W \neq \emptyset\}. \end{aligned}$$

Pro danou realizaci procesu označme $s_i = (t_i, \theta_i),\ 1 \leq i \leq n_1$ necensorované úsečky a $\widetilde{s}_j = (\widetilde{t}_j, \widetilde{\theta}_j),\ 1 \leq j \leq n_2$ jednostranně censorované úsečky s významným

bodem uvnitř W. Položme

$$U_0 = \{(t_1, \theta_1), \ldots, (t_{n_1}, \theta_{n_1}), (\widetilde{t}_1, \widetilde{\theta}_1), \ldots, (\widetilde{t}_{n_2}, \widetilde{\theta}_{n_2})\}.$$

Nyní bude naším cílem odhadnout parametr ϑ. Předpokládejme, že pro každé ϑ je známé marginální rozdělení směrů $\rho(\mathrm{d}\theta|\vartheta)$ a podmíněné rozdělení $\mathcal{D}(\mathrm{d}t|\theta, \vartheta)$. Nechť Z je odhad, který bychom použili v případě nezávislého výběru o rozsahu n z necenzorovaných dat. Potom SRE algoritmus probíhá následovně:

(i) Volíme libovolné počáteční nastavení parametru ϑ_0.

(ii) V každé iteraci p provedeme následující dva kroky:

(R)–obnova cenzorovaných dat. Každou cenzorovanou délku $\widetilde{t}_j,\ j = 1, \ldots, n_2$ nahradíme délkou $\widetilde{t}_{j,p+1}$, kterou obdržíme simulací z podmíněného rozdělení $\mathcal{D}(\mathrm{d}t|\widetilde{\theta}_j, t \geq t_j, \vartheta_p)$, kde ϑ_p je aktuální odhad parametru ϑ. Označme

$$U_{p+1} = \{(t_1, \theta_1), \ldots, (t_{n_1}, \theta_{n_1}), (\widetilde{t}_{1,p+1}, \widetilde{\theta}_1), \ldots, (\widetilde{t}_{n_2,p+1}, \widetilde{\theta}_{n_2})\}. \tag{3.1}$$

(E)–učiníme nový odhad parametru $\vartheta_{p+1} = Z(U_{p+1})$.

Posloupnost odhadů (ϑ_p) generovaná SRE algoritmem tvoří homogenní markovský řetězec, který za určitých podmínek má jednoznačné invariantní rozdělení ν_W závislé na pozorování v okně W a pro který platí

$$\overline{\vartheta}_p = \frac{1}{p}\sum_{i=1}^{p} \vartheta_i \longrightarrow \int l \mathrm{d}\nu_W(l) \qquad \text{s.j.} \tag{3.2}$$

Tento řetězec je stabilní ve smyslu, že konvergence (3.2) platí nezávisle na počáteční volbě ϑ_0. SRE odhad je pak definován jako tato limita.

Jestliže lze parametr ϑ rozdělit do dvou podmnožin $\vartheta = (\vartheta_T, \vartheta_\Theta)$ takových, že

$$\Pi(\mathrm{d}t \times \mathrm{d}\theta|\vartheta) = \rho(\mathrm{d}\theta|\vartheta_\Theta) \times \mathcal{D}(\mathrm{d}t|\theta, \vartheta_T),$$

potom, neboť směry $\widetilde{\theta}_j,\ j = 1, \ldots, n_2$ nejsou cenzorovány, můžeme parametr ϑ_Θ odhadnout běžnými typy odhadů (jako je např. metoda nejmenších čtverců či maximálně věrohodný odhad) a soustředit se pouze na odhad parametru ϑ_T.

Poznámka: V algoritmu pracujeme pouze s úsečkami, jejichž nejjižnější bod je obsažen v okně W. Snadno lze ovšem algoritmus použít na úsečky, jejichž nejsevernější bod je obsažen v okně W. Zkombinováním odhadů obou dvou algoritmů (např. aritmetickým průměrem) dostaneme odhad, v němž jsou uvažovány jak necenzorované, tak všechny jednostranně cenzorované úsečky. Nadále však platí, že SRE algoritmus takto postavený nevyužívá oboustranně cenzorovaných pozorování.

Poznámka: Algoritmus sám nevyžaduje, aby okno pozorování W bylo konvexní. Není-li tomu tak, buď můžeme uvažovat pouze úsečky, jejichž význačný bod je obsažen ve W nebo „obnovit“ také místo výskytu tohoto bodu.

Poznámka: SRE algoritmus lze použít i k samotnému odhadu distribuční funkce F délek (hustoty, ...). Pro censorovaná data U_0 volme F_0 např. empirickou distribuční funkci těchto dat. Délky $\widetilde{t}_j$ potom obnovíme simulací z rozdělení s touto distribuční funkcí podmíněně při $\widetilde{t}_{j,p+1} \geq \widetilde{t}_j$. Potom v p-té iteraci je F_{p+1} empirická distribuční funkce pro data U_{p+1}.

Kapitola 4

Porovnání odhadů na základě simulací ze známého rozdělení

V předchozích dvou kapitolách jsme si uvedli některé typy neparametrických a parametrických odhadů pro nezáporné měřitelné funkce náhodných elementů a způsob jejich použití pro odhad rozdělení délek bodových procesů úseček. V této kapitole se budeme věnovat vzájemnému srovnání těchto odhadů pro některé konkrétní volby rozdělení délek stacionárního Poissonova bodového procesu úseček na $\mathbb{R}^2$. Rozdělení směrů úseček volíme rovnoměrné na intervalu $[0;\pi)$. Podle definice je tedy bodový proces úseček stacionární a izotropní. Dále předpokládejme, že rozdělení směrů a délek úseček jsou nezávislé veličiny. Bez újmy na obecnosti volíme z důvodu snazší programovatelnosti za význačné body úseček jejich lexikografická minima resp. maxima vzhledem k druhé souřadnici (tj. „nejjižnější“ resp. „nejsevernější“ bod úsečky). Nejprve provedeme srovnání odhadů pro rovnoměrné rozdělení délek úseček, potom se budeme věnovat logaritmicko-normálnímu rozdělení.

K porovnání kvality odhadů použijeme dvě různé vzdálenosti funkcí. První z nich je Kolmogorovova-Smirnovova vzdálenost definovaná vztahem

$$d_{KS}(\widehat{F}, F) = \sup_{x\in\mathbb{R}} |\widehat{F}(x) - F(x)|.$$

Druhá vzdálenost je Cramérova-von Misesové vzdálenost daná vztahem

$$d_{CVM}(\widehat{F}, F) = \int (\widehat{F}(x) - F(x))^2 \,\mathrm{d}F(x).$$

Pro celou tuto kapitolu nechť F je distribuční funkce délek úseček a nechť $\widehat{F}_1$, $\widehat{F}_2$, $\widehat{F}_3$ značí odpovídající Horvitzovy-Thompsonovy odhady získané užitím tvrzení 2.1, $\widehat{F}_{KM}$ Kaplanův-Meierův odhad získaný z (2.4), $\widehat{F}_{EM}$ maximálně věrohodný odhad vypočtený pomocí EM algoritmu a konečně $\widehat{F}_{SRE}$ odhad získaný pomocí SRE algoritmu.

Srovnání provedeme na základě tisíce simulací, z nichž pak určíme pro obě vzdálenosti výběrový průměr.

4.1 Rovnoměrné rozdělení délek úseček

Nechť Φ je Poissonův proces úseček na $\mathbb{R}^2$ a mějme rozdělení délek rovnoměrné na $(0; M]$, $M > 0$, nezávislé na rozdělení směrů. Uvažujme nejprve, že okno pozorování W je kruh o poloměru $R > 0$. Označme nyní

$$O(x) = \int_{\sqrt{R^2 - \frac{x^2}{4}}}^{R} \sqrt{R^2 - u^2}\,\mathrm{d}u.$$

Potom váhy pro Horvitzův-Thompsonův odhad z tvrzení 2.1 vypadají následovně

$$\begin{aligned} |W \ominus \check{S}(t,\theta)| &= 4O(\sqrt{4R^2 - t^2}) \\ &= \pi R^2 - \frac{t}{2}\sqrt{4R^2 - t^2} - 2R^2 \arcsin\frac{t}{2R}, \\ |W \oplus \check{S}(t,\theta)| &= \pi R^2 + 2R\,t. \end{aligned}$$

Na rovnoměrné rozdělení délek na intervalu $(0; M]$ nebylo možné aplikovat SRE algoritmus, neboť vzhledem k tomu, jak je postaven, mohla nastat (a také nastala) situace, kdy jsme se snažili simulovat náhodnou veličinu z rovnoměrného rozdělení na nějakém intervalu za podmínky, že náhodná veličina sama má být větší než horní mez intervalu.

Následující dvě tabulky udávají aritmetické průměry vzdáleností pro kruhové okno W o poloměru $R = 0{,}5$ resp. $R = 1{,}5$ s parametrem $M = 0{,}25$. Pro každou vzdálenost je v prvním sloupci výběrový průměr a v druhém sloupci výběrový rozptyl vzdáleností d_{KS} a d_{CVM}, oba spočtené na základě 1000 simulací ze stacionárního Poissonova procesu úseček s daným rozdělením délek. Intenzitu λ příslušného procesu Ψ jsme volili 30 a 100.

	$\lambda = 30$				$\lambda = 100$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,1920	0,0042	0,0362	0,0013	0,1042	0,0010	0,0098	0,00007
F_2	0,1631	0,0029	0,0256	0,0006	0,0908	0,0009	0,0078	0,00005
F_3	0,1740	0,0033	0,0293	0,0008	0,0960	0,0008	0,0084	0,00005
F_{KM}	0,1876	0,0040	0,0338	0,0011	0,1023	0,0009	0,0093	0,00006
F_{EM}	0,1934	0,0044	0,0381	0,0014	0,1039	0,0010	0,0099	0,00008

Tabulka 4.1: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Rozdělení délek je rovnoměrné na $(0; 0{,}25]$. Poloměr okna W je $R = 0{,}5$.

	$\lambda = 30$				$\lambda = 100$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,0601	3,3963	0,00327	0,0908	0,03296	1,0493	0,00951	0,0072
F_2	0,0574	3,0900	0,00297	0,0684	0,03124	0,0950	0,00086	0,0062
F_3	0,0585	3,1918	0,00307	0,0754	0,03203	0,0989	0,00090	0,0065
F_{KM}	0,0596	3,2749	0,00319	0,0084	0,03277	1,0089	0,00093	0,0007
F_{EM}	0,0595	3,2028	0,00317	0,0080	0,03278	1,0229	0,00094	0,0008

Tabulka 4.2: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Výběrové rozptyly jsou vynásobeny 10^4. Rozdělení délek je rovnoměrné na $(0; 0{,}25]$. Poloměr okna W je $R = 1{,}5$.

Nyní předpokládejme, že okno W je čtverec o straně $a > 0$. Váhy v Horvitzově-Thompsonově odhadu tedy budou tvaru

$$|W \ominus \check{S}(t,\theta)| = (a - t|\cos\theta|)(a - t\sin\theta), \tag{4.1}$$

$$|W \oplus \check{S}(t,\theta)| = a^2 + at(|\cos\theta| + \sin\theta). \tag{4.2}$$

Pro konstantu a_1 definovanou vzorcem (2.21) platí v tomto případě rovnost

$$a_1 = \frac{4}{\pi}aP - a^2. \tag{4.3}$$

Vzhledem k volbě rovnoměrného rozdělení směrů je $\mathbb{E}_K \operatorname{diam}(W)$ z (2.20) roven

$$\mathbb{E}_K \operatorname{diam}(W) = \frac{4a}{\pi}\int_0^{\frac{\pi}{4}} \frac{1}{\cos\theta}\,\mathrm{d}\theta = \frac{4a}{\pi}\log\tan\frac{3\pi}{8}. \tag{4.4}$$

Výsledky obdržené z tisíce simulací Poissonova procesu úseček s intenzitami $\lambda = 30$, $\lambda = 100$ s rozdělením délek rovnoměrným na $(0; 0{,}25]$ pro čtvercové okno pozorování W o straně $a = 0{,}5$, $a = 1$ uvádějí tabulky 4.3 a 4.4.

	$\lambda = 30$				$\lambda = 100$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,3850	2,3968	0,1743	3,6964	0,2165	0,6393	0,0515	0,2975
F_2	0,2961	1,0975	0,0970	0,7984	0,1874	0,3948	0,0422	0,1394
F_3	0,3123	1,3619	0,1060	1,1616	0,1840	0,4027	0,0365	0,1326
F_{KM}	0,3561	1,6809	0,1375	1,9457	0,2081	0,4790	0,0440	0,1623
F_{EM}	0,3608	1,4518	0,1382	1,4601	0,2237	0,4284	0,0535	0,1707

Tabulka 4.3: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Výběrové rozptyly jsou vynásobeny 10^2. Rozdělení délek je rovnoměrné na $(0; 0{,}25]$. Strana okna W je $a = 0{,}5$.

	$\lambda = 30$				$\lambda = 100$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,1807	24,769	0,0316	4,6841	0,1010	11,884	0,0107	1,0452
F_2	0,1486	20,382	0,0214	4,1494	0,0887	8,7405	0,0083	0,6134
F_3	0,1585	28,899	0,0256	5,5193	0,0938	9,9832	0,0093	0,7692
F_{KM}	0,1725	25,432	0,0285	4,6301	0,0989	10,800	0,0099	0,8490
F_{EM}	0,1688	21,035	0,0261	3,7995	0,0966	9,7074	0,0092	0,6782

Tabulka 4.4: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Výběrové rozptyly jsou vynásobeny 10^4. Rozdělení délek je rovnoměrné na $(0; 0{,}25]$. Strana okna W je $a = 1$.

4.2 Logaritmicko-normální rozdělení délek

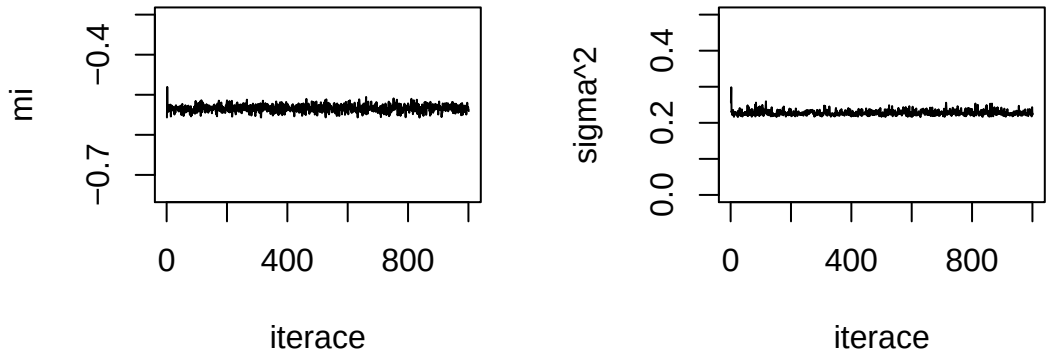
Nyní rozebereme případ, kdy rozdělení délek úseček je logaritmicko-normální s parametry $\mu \in \mathbb{R}$ a $\sigma > 0$. Nechť W je opět čtvercové okno o straně $a > 0$. Váhy pro Horvitzovy-Thompsonovy odhady a konstanty užívané v EM algoritmu jsou stejné jako pro rovnoměrné rozdělení, viz (4.2)–(4.4).

Podívejme se blíže na SRE algoritmus. Neuvedli jsme si žádné předpoklady pro stabilitu markovského řetězce definovaného v předchozí kapitole ani důkaz, že odhad získaný tímto algoritmem je konzistentní. Obecně toto téma zůstává stále otevřeno. Důkaz pro exponenciální rozdělení délek lze nalézt v [4]. Oprávněnost takových předpokladů můžeme posoudit pouze na základě simulací. Ty provedeme pro logaritmicko-normální rozdělení s dvourozměrným parametrem (μ, σ). Počáteční nastavení jsme volili tak, že jsme spočetli výběrový průměr a výběrový rozptyl z nasimulovaných dat a vyjádřili μ_0 a σ_0 ze vzorce pro střední hodnotu, resp. rozptyl tohoto rozdělení. Za statistiku Z použitou k odhadu parametru z obnovených dat jsme zvolili maximálně věrohodný odhad

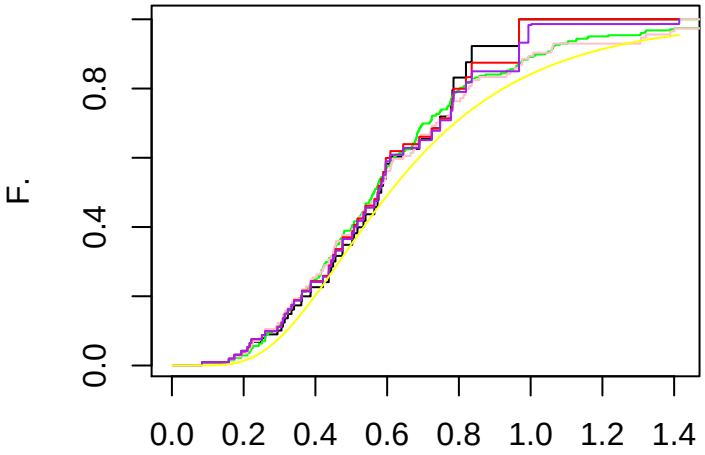
$$\begin{aligned} \widehat{\mu} &= \frac{1}{n_1}\sum_{i=1}^{n_1} \log t_i + \frac{1}{n_2}\sum_{j=1}^{n_2} \log \widetilde{t}_j, \\ \widehat{\sigma}^2 &= \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(\log t_i - \widehat{\mu})^2 + \sum_{j=1}^{n_2}(\log \widetilde{t}_j - \widehat{\mu})^2\right). \end{aligned}$$

Na obrázku 4.1 vidíme markovský řetězec vytvořený SRE algoritmem pro 1000 iterací.

Výběrové průměry a rozptyly vzdáleností d_{KS} a d_{CVM} spočtené na základě tisíce simulací z logaritmicko-normálního rozdělení s parametry $\mu = -0{,}5$ a $\sigma = 0{,}5$ jsou uvedeny v tabulkách 4.5 a 4.6.



Obrázek 4.1: Průběh řetězce získaného SRE algoritmem pro jednu realizaci procesu úseček s $LN(-0{,}5; 0{,}5)$ rozdělením délek úseček.



Obrázek 4.2: Odhady distribuční funkce logaritmicko-normálního rozdělení délek úseček s parametry $\mu = -0{,}5$ a $\sigma = 0{,}5$. $\widehat{F}_1$–černá, $\widehat{F}_2$–zelená, $\widehat{F}_3$–růžová, $\widehat{F}_{KM}$–červená, $\widehat{F}_{EM}$–fialová, $\widehat{F}_{SRE}$–modrá, F–žlutá.

	$a = 1,\ \lambda = 100$				$a = 2,\ \lambda = 25$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,2644	5,3105	0,0223	0,2794	0,1369	2,3308	0,0056	0,0292
F_2	0,1156	0,8199	0,0049	0,0078	0,0985	0,9487	0,0030	0,0058
F_3	0,1294	1,4481	0,0056	0,0156	0,1090	1,2847	0,0036	0,0098
F_{KM}	0,1866	2,3099	0,0082	0,0337	0,1260	1,6673	0,0043	0,0149
F_{EM}	0,1797	1,7630	0,0079	0,0228	0,1241	1,4955	0,0041	0,0117
F_{SRE}	0,0937	2,0704	0,0057	0,0283	0,0658	1,1193	0,0028	0,0078

Tabulka 4.5: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Výběrové rozptyly jsou přenásobeny 10^3. Rozdělení délek je $LN(-0{,}5; 0{,}5)$.

	$a = 3,\ \lambda = 10$				$a = 10,\ \lambda = 1$			
	d_{KS}		d_{CVM}		d_{KS}		d_{CVM}	
F_1	0,1177	1,4999	0,0037	0,0109	0,0889	0,7960	0,0019	0,0030
F_2	0,0921	0,8808	0,0023	0,0043	0,0831	0,6266	0,0016	0,0021
F_3	0,1025	1,1287	0,0028	0,0068	0,0861	0,6952	0,0017	0,0025
F_{KM}	0,1130	1,2908	0,0032	0,0087	0,0883	0,7576	0,0018	0,0028
F_{EM}	0,1108	1,1419	0,0030	0,0067	0,0874	0,7165	0,0018	0,0025
F_{SRE}	0,0585	1,0133	0,0023	0,0059	0,0419	0,6385	0,0012	0,0024

Tabulka 4.6: Výběrové průměry a rozptyly vzdáleností odhadů od distribuční funkce délek spočítané z 1000 simulací. Výběrové rozptyly jsou přenásobeny 10^3. Rozdělení délek je $LN(-0{,}5; 0{,}5)$.

4.3 Diskuze

Provedli jsme simulace stacionárního Poissonova procesu úseček pro různé volby rozdělení délek, intenzity a velikosti a tvaru okna W. Na základě výsledků výpočtů uvedených v tabulkách 4.1–4.6 můžeme vyvodit několik závěrů.

Nezávisle na volbě parametrů, rozdělení a velikosti okna vyšel (z neparametrických odhadů) pro obě vzdálenosti v podstatě pokaždé nejlépe Horvitzův-Thompsonův plus sampling. Jako druhý nejlepší se jevil unbiased sampling. Tento výsledek se dal očekávat, neboť oba tyto odhady uvažují plnou, necensorovanou délku úseček. Plus sampling navíc k odhadu používá všechny úsečky, které okno W protnou (unbiased sampling pouze ty s význačným bodem uvnitř W). Právě pro způsob výběrů úseček se ukázal být nejhorším minus sampling, neboť užívá informace získané pouze od úseček cele obsažených ve W. Proto byl odhad zís-

kaný touto metodou ve všech případech nejhorší. Ale při volbě dostatečně velkého okna pozorování W (viz tabulka 4.6) už byl rozdíl mezi Kaplanovým-Meierovým odhadem a minus samplingem nepatrný.

U rovnoměrného rozdělení si víceméně rovnocenně vedli Kaplanův-Meierův odhad a odhad získaný pomocí EM algoritmu.

Pro logaritmicko-normální rozdělení délek úseček jsme měli ke srovnání navíc ještě parametrický odhad počítaný pomocí SRE algoritmu. Vzhledem k tomu, že u tohoto odhadu se předpokládá znalost rozdělení a odhadují se pouze jeho parametry, zdá se být podle výpočtů nejlepším odhadem. Maximálně věrohodný odhad se v tomto případě jeví lepší typ odhadu než Kaplanův-Meierův odhad.

Kapitola 5

Odhady intenzity průsečíků

Další charakteristikou, která nás u bodového procesu úseček může zajímat, jsou průsečíky a jejich intenzita. Tato informace může být užitečná například v lesnictví při polomech stromů. V této kapitole si definujeme intenzitu průsečíků a budeme se zabývat jejími odhady v různých situacích. Na závěr kapitoly pak provedeme simulace pro různé volby parametrů procesu úseček a okna W. Zabývat se průsečíky úseček má smysl pouze pro $d = 2$, proto budeme v této kapitole pracovat pouze s bodovými procesy úseček na $\mathbb{R}^2$.

Nechť tedy Φ je stacionární Poissonův bodový proces úseček na $\mathbb{R}^2$ s intenzitou λ a rozdělením typického zrna Λ_0.

Definice: *Intenzita průsečíků* N bodového procesu úseček Φ v $\mathbb{R}^2$ je dána vztahem

$$N = \mathbb{E} \sum_{S,S' \in \Phi, S \neq S'} \frac{\mathbf{1}_{\{S \cap S' \cap [0,1]^2 \neq \emptyset\}}}{2}. \tag{5.1}$$

Pro intenzitu průsečíků stacionárního procesu platí následující věta, viz [7], Theorem 1.

Věta 5.1 *Nechť Φ je stacionární Poissonův bodový proces úseček v $\mathbb{R}^2$ s rozdělením směrů ρ nezávislým na rozdělení délek $\mathcal{D}$. Potom*

$$N = \frac{\lambda^2}{2} (\mathbb{E}t)^2 \int \int \sin(|\theta - \theta'|) \rho(\mathrm{d}\theta) \rho(\mathrm{d}\theta'), \tag{5.2}$$

kde $\mathbb{E}t$ značí střední hodnotu délky typické úsečky. Je-li navíc Φ izotropní, potom

$$N = \frac{\lambda^2}{\pi} (\mathbb{E}t)^2. \tag{5.3}$$

Obecně pro libovolný bodový proces úseček na $\mathbb{R}^2$ platí

$$N = \frac{\lambda^2}{2} \int \int tt' \sin(\theta - \theta') \Lambda_0(\mathrm{d}(t, \theta)) \Lambda_0(\mathrm{d}(t', \theta')). \tag{5.4}$$

Důkaz. Bodový proces průsečíků můžeme definovat následovně

$$\Sigma(B) = \sum_{S,S' \in \Phi; S \neq S'} \frac{\mathbf{1}_{\{S \cap S' \cap B \neq \emptyset\}}}{2},$$

kde B je libovolná borelovská množina. Podle věty 1.1 a lemmatu 1.1 můžeme míru intenzity procesu Σ vyjádřit jako

$$\mathbb{E}\Sigma(B) = \int \int \frac{I(S \cap S' \cap B \neq \emptyset)}{2} \Lambda(\mathrm{d}S)\Lambda(\mathrm{d}S').$$

Nyní si definujme míru $f_{S'}(B) = \int_{\mathcal{S}} \frac{I(S \cap S' \cap B \neq \emptyset)}{2} \Lambda(\mathrm{d}S)$. Pro pevné B je funkce $S' \mapsto f_{S'}(B)$ je nezáporná a měřitelná, tedy podle (1.6) máme

$$\begin{aligned}
\mathbb{E}\Sigma(B) &= \int_{\mathcal{S}} f_{S'}(B)\, \Lambda(\mathrm{d}S') = \lambda \int_{\mathcal{S}'_0} \int_{\mathbb{R}^2} f_{S'_0+z'}(B)\, \mathrm{d}z' \Lambda_0(\mathrm{d}S'_0) \\
&= \lambda \int_{\mathcal{S}'_0} \int_{\mathbb{R}^2} f_{S'_0}(B - z')\, \mathrm{d}z' \Lambda_0(\mathrm{d}S'_0) \\
&= \lambda \int_{\mathcal{S}'_0} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \mathbf{1}_{B-z'}(y)\, f_{S'_0}(\mathrm{d}y)\, \mathrm{d}z'\, \Lambda_0(\mathrm{d}S'_0) \\
&= \lambda |B| \int_{\mathcal{S}'_0} f_{S'_0}(\mathbb{R}^2) \Lambda_0(\mathrm{d}S'_0).
\end{aligned}$$

Využili jsme toho, že $f_S(B) = f_{x+S}(x + B)$ pro každé $x \in \mathbb{R}^2$. Odtud již pro $N = \mathbb{E}\Sigma([0,1]^2)$ plyne

$$\begin{aligned}
N &= \lambda \int \int \frac{I(S \cap S'_0 \neq \emptyset)}{2}\, \Lambda(\mathrm{d}S)\, \Lambda_0(\mathrm{d}S'_0) \\
&= \lambda^2 \int \int \int \frac{I((S_0 + z) \cap S'_0 \neq \emptyset)}{2}\, \mathrm{d}z\, \Lambda_0(\mathrm{d}S_0)\, \Lambda_0(\mathrm{d}S'_0) \\
&= \lambda^2 \int\int\int\int\int \frac{I((S_0(t,\theta) + z) \cap S'_0(t',\theta') \neq \emptyset)}{2} \mathrm{d}z \rho(\mathrm{d}\theta)\rho(\mathrm{d}\theta')\mathcal{D}(\mathrm{d}t)\mathcal{D}(\mathrm{d}t') \\
&= \frac{\lambda^2}{2} \int\int\int\int tt' \sin(|\theta - \theta'|)\, \rho(\mathrm{d}\theta)\, \rho(\mathrm{d}\theta')\, \mathcal{D}(\mathrm{d}t)\, \mathcal{D}(\mathrm{d}t') \\
&= \frac{\lambda^2}{2} (\mathbb{E}t)^2 \int \int \sin(|\theta - \theta'|) \rho(\mathrm{d}\theta)\rho(\mathrm{d}\theta').
\end{aligned}$$

Zde jsme využili toho, že $\int \mathbf{1}_{\{(S_0(t,\theta)+z) \cap S'_0(t',\theta') \neq \emptyset\}} \mathrm{d}z$ je plocha rovnoběžníku daného úsečkami $S_0(t,\theta)$ a $S'_0(t',\theta')$. Je-li navíc proces Φ izotropní, potom je $\int\int \sin(|\theta - \theta'|)\rho(\mathrm{d}\theta)\rho(\mathrm{d}\theta') = \frac{2}{\pi}$. □

Budeme-li nyní chtít intenzitu průsečíků odhadovat, jednou z možností je parametrický odhad, kdy odhadneme parametry a použijeme vzorce ve větě 5.1.

Nejjednodušší neparametrický odhad, který se nabízí, je

$$\widehat{N}_c = \sum_{S,S' \in \Phi, S \neq S'} \frac{\mathbf{1}_{\{S \cap S' \cap W \neq \emptyset\}}}{2|W|}. \tag{5.5}$$

Tento odhad je zřejmě nestranný, neboť

$$\mathbb{E}\widehat{N}_c = \frac{1}{|W|}\mathbb{E}\Sigma(W) = \frac{1}{|W|}N|W| = N.$$

Pro izotropní proces Φ s nezávislým rozdělením směrů ρ a délek $\mathcal{D}$ pak užitím nestranného odhadu $\frac{1}{|W|}\sum_{S\in\Phi} L(S \cap W)$ pro délkovou intenzitu $\lambda\,\mathbb{E}t$ dostaneme z (5.3) neparametrický odhad

$$\widehat{N}_l = \frac{1}{\pi}\left(\frac{1}{|W|}\sum_{S\in\Phi} L(S \cap W)\right)^2, \tag{5.6}$$

kde $L(S \cap W)$ je délka úseček pozorovatelná ve W. Odhad (5.6), jak si ukážeme v druhé podkapitole, sice není nestranný, ale při $|W| \longrightarrow \infty$ je asymptoticky nestranný.

V následující podkapitole se budeme věnovat nejlepšímu nestrannému odhadu intenzity průsečíků. Odvodíme tento odhad pro případ, že známe rozdělení typického zrna Λ_0. Odhad pro případ, kdy máme známé rozdělení $\Lambda_0(\vartheta)$ s neznámým parametrem ϑ je uveden v závěrečné poznámce podkapitoly.

5.1 Nejlepší nestranný odhad intenzity průsečíků

Předpokládejme, že máme stacionární Poissonův proces úseček Φ na $\mathbb{R}^2$, kde rozdělení typického zrna Λ_0 je známé. Data pozorujeme skrz ohraničené měřitelné okno $W \subset \mathbb{R}^2$. Nechť $\mathcal{W}$ je měřitelná podmnožina prostoru $\mathcal{S}$. Označme $\mathcal{E}_{\mathcal{W}}$ množinu všech odhadů, které závisí pouze na úsečkách ve $\mathcal{W}$.

Dále nechť $C = \int_{\mathcal{U}^d \times \mathbb{R}^+} |W(t,\theta)|\Lambda_0(\mathrm{d}(t,\theta))$, kde $W(t,\theta) = \{z : z+S(t,\theta) \in \mathcal{W}\}$ je množina všech bodů, které jsou význačnými pro úsečky z $\mathcal{W}$ směru θ a délky t.

Nejlepší nestranný odhad pro izotropní proces nám uvádí následující věta, již si dokážeme v druhé podkapitole.

Věta 5.2 *Nechť Φ je izotropní Poissonův proces úseček na $\mathbb{R}^2$ s ρ a $\mathcal{D}$ nezávislými. Potom je*

$$\widehat{N} = \frac{\Phi(\mathcal{W})^2 - \Phi(\mathcal{W})}{C^2\pi}(\mathbb{E}t)^2 \tag{5.7}$$

nejlepší nestranný odhad intenzity průsečíků N ze všech odhadů v $\mathcal{E}_{\mathcal{W}}$.

Nyní se budeme zabývat dvěma důležitými případy množin $\mathcal{W}$.

1. $\mathcal{W}_1 = \{S \in \mathcal{S} : c(S) \in W\}$ je množina všech úseček, jejichž významný bod leží uvnitř W. Pro tuto volbu $\mathcal{W}$ je konstanta $C = |W|$.
2. $\mathcal{W}_2 = \{S \in \mathcal{S} : S \cap W \neq \emptyset\}$ je množina všech úseček, které protnou okno W.

Je-li okno W čtverec o straně a, potom pro odhad $\widehat{N}$ získaný volbou množiny $\mathcal{W}_2$ je konstanta C rovna

$$C = \int |W \oplus \check{S}_0| \Lambda_0(\mathrm{d}S_0) = a^2 + \frac{4a}{\pi}\mathbb{E}t. \tag{5.8}$$

Jestliže nejsme schopni určit C, můžeme volit případ 1, kde za referenční bod volíme nejprve lexikografické minimum a spočteme (5.7) a následně provedeme totéž pro lexikografické maximum. Výsledný odhad získáme jako

$$\widetilde{N} = \frac{\Phi_1(W)^2 - \Phi_1(W) + \Phi_2(W)^2 - \Phi_2(W)}{2|W|^2\pi}(\mathbb{E}t)^2, \tag{5.9}$$

kde $\Phi_1(W)$ resp. $\Phi_2(W)$ značí počet lexikografických minim resp. maxim obsažených ve W. Tento odhad neuvažuje pouze úsečky, jejichž oba konce leží vně W.

Poznámka: V případě, kdy je parametr ϑ rozdělení typického zrna $\Lambda_0(\vartheta)$ neznámý, je situace složitější. Mějme Φ izotropní proces s rozdělením směrů nezávislým na rozdělení délek. Označme $\Phi(W)$ počet význačných bodů procesu pozorovaných ve W. Dále nechť $T_{\Phi(W)}$ je úplná postačující statistika pro parametr ϑ. Potom je $S_T = (\Phi(W), T_{\Phi(W)})$ úplná postačující statistika pro λ, ϑ, viz [7]. Potom za předpokladu, že známe úplné délky censorovaných úseček, je nejlepší nestranný odhad intenzity průsečíků

$$\widehat{N}^{NNO} = \frac{\Phi(W)^2 - \Phi(W)}{|W|^2\pi} \cdot e_{2,\Phi(W)},$$

kde $e_{2,\Phi(W)}$ značí nejlepší nestranný odhad pro $(\mathbb{E}t)^2$.

5.2 Střední hodnoty a rozptyly odhadů intenzity průsečíků

V této podkapitole se budeme věnovat střední hodnotě a rozptylu některých z uvedených odhadů intenzity průsečíků. Také si dokážeme větu 5.2.

Jak jsme si již ukázali, odhad $\widehat{N}_c$ je nestranný. Pro ostatní odhady se omezíme na izotropní Poissonův proces Φ. Použitím věty 1.1, věty 1.3 a lemmatu 1.1

můžeme pro odhad $\widehat{N}_l$ spočítat

$$\begin{aligned}
\mathbb{E}\widehat{N}_l &= \frac{1}{\pi|W|^2}\,\mathbb{E}\sum_{S\in\Phi} L(S\cap W)^2 \\
&= \frac{1}{\pi|W|^2}\,\mathbb{E}\sum_{S\in\Phi}\sum_{T\in\Phi} L(S\cap W)L(T\cap W) \\
&= \frac{1}{\pi|W|^2}\,(\mathbb{E}\sum_{S,T\in\Phi;S\neq T} L(S\cap W)L(T\cap W) + \mathbb{E}\sum_{S\in\Phi}[L(S\cap W)]^2) \\
&= \frac{1}{\pi|W|^2}\Big(\int\int L(S\cap W)L(T\cap W)M_2^!(\mathrm{d}S\times\mathrm{d}T) \\
&\quad + \lambda\int\int[L((S_0+z)\cap W)]^2\Lambda_0(\mathrm{d}S_0)\mathrm{d}z\Big) \\
&= \frac{1}{\pi|W|^2}\Big(\int\int L(S\cap W)L(T\cap W)\Lambda(\mathrm{d}S)\Lambda(\mathrm{d}T) \\
&\quad + \lambda\int\int[L((S_0+z)\cap W)]^2\Lambda_0(\mathrm{d}S_0)\mathrm{d}z\Big) \\
&= \frac{1}{\pi|W|^2}(\mathbb{E}\sum_{S\in\Phi} L(S\cap W))^2 \\
&\quad + \frac{\lambda}{\pi|W|^2}\int\int[L((S_0+z)\cap W)]^2\Lambda_0(\mathrm{d}S_0)\mathrm{d}z \\
&= N + \frac{\lambda}{\pi|W|^2}\int\int[L((S_0+z)\cap W)]^2\Lambda_0(\mathrm{d}S_0)\mathrm{d}z.
\end{aligned}$$

Odtud je vidět, že odhad $\widehat{N}_l$ není nestranný. Vychýlení má tvar

$$\begin{aligned}
\mathbb{E}\widehat{N}_l - N &= \frac{\lambda}{\pi|W|^2}\int\int L((S_0+z)\cap W)^2\mathrm{d}z\Lambda_0(\mathrm{d}S_0) \\
&= \frac{\lambda}{\pi|W|^2}\int\int L(S_0\cap(W-z))^2\mathrm{d}z\Lambda_0(\mathrm{d}S_0) \\
&= \frac{\lambda}{\pi|W|^2}\int\int\int\int \mathbf{1}_{S_0\cap(W-z)}(u)\mathbf{1}_{S_0\cap(W-z)}(v)L(\mathrm{d}u)L(\mathrm{d}v)\mathrm{d}z\Lambda_0(\mathrm{d}S_0) \\
&= \frac{\lambda}{\pi|W|^2}\int\int\int\int \mathbf{1}_{W-u}(z)\mathbf{1}_{W-v}(z)\mathbf{1}_{S_0}(u)\mathbf{1}_{S_0}(v)\mathrm{d}z\; L(\mathrm{d}u)L(\mathrm{d}v)\Lambda_0(\mathrm{d}S_0) \\
&= \frac{\lambda}{\pi|W|^2}\int\int_{\mathcal{S}_0}\int_{\mathcal{S}_0}|(W-u)\cap(W-v)|L(\mathrm{d}u)L(\mathrm{d}v)\Lambda_0(\mathrm{d}S_0).
\end{aligned}$$

Protože $|(W-u)\cap(W-v)|\le|W|$, dostáváme

$$\mathbb{E}\widehat{N}_l - N \le \frac{\lambda}{\pi|W|^2}|W|\int L(S_0)^2\,\Lambda_0(\mathrm{d}S_0) = \frac{\lambda}{\pi|W|}\mathbb{E}t^2\,.$$

Vychýlení je vždy nezáporné a s $|W| \to \infty$ (za předpokladu konečného druhého momentu délky typické úsečky) klesá k nule. Odhad $\widehat{N}_l$ je tak asymptoticky nestranný.

Z konstrukce odhadu $\widehat{N}_l$ je vidět, že je úměrný druhé mocnině klasického odhadu délkové intenzity procesu úseček. Výpočet $\mathbb{E}\widehat{N}_l$ proto vyžaduje vyjádření stejných integrálů jako při výpočtu rozptylu odhadu délkové intenzity, což je podrobně provedeno v [6]. Pro čtvercové okno W o straně a za předpokladu $L(S_0) \leq a$ s.j. platí

$$\mathbb{E}\widehat{N}_l - N = \frac{\lambda}{\pi a^4}\left(a^2\mathbb{E}t^2 - \frac{1}{3}a\mathbb{E}t^3(\sin\theta + |\cos\theta|) + \frac{1}{6}\mathbb{E}t^4 \sin\theta|\cos\theta|\right),$$

což pro izotropní proces s nezávislým délek a směrů dává

$$\mathbb{E}\widehat{N}_l - N = \frac{\lambda}{\pi a^4}\left(a^2\mathbb{E}t^2 - \frac{4}{3\pi}a\mathbb{E}t^3 + \frac{1}{6\pi}\mathbb{E}t^4\right). \tag{5.10}$$

K tomu, abychom mohli spočítat střední hodnotu a rozptyl odhadu $\widehat{N}$ počítejme

$$\begin{aligned}\mathbb{E}\Phi(\mathcal{W}) &= \int \mathbf{1}_{\mathcal{W}}(S)\,\Lambda(\mathrm{d}S) = \lambda \int\int \mathbf{1}_{\mathcal{W}}(z + S_0)\,\mathrm{d}z\,\Lambda_0(\mathrm{d}S_0)\\ &= \lambda \int |W(t,\theta)|\,\Lambda_0(\mathrm{d}(t,\theta)) = \lambda\, C.\end{aligned}$$

Dále, neboť je proces Φ stacionární Poissonův, máme podle (1.2) a lemmatu 1.1

$$\mathbb{E}\Phi(\mathcal{W})^2 = \lambda C + \lambda^2 C^2.$$

Odtud už plyne nestrannost odhadu $\widehat{N}$, neboť

$$\mathbb{E}\widehat{N} = \frac{\lambda C + \lambda^2 C^2 - \lambda C}{C^2\pi}(\mathbb{E}t)^2 = \frac{\lambda^2}{\pi}(\mathbb{E}t)^2. \tag{5.11}$$

Obdobně opět užitím (1.2) můžeme spočítat $\mathbb{E}\Phi(\mathcal{W})^3$ a $\mathbb{E}\Phi(\mathcal{W})^4$. Celkem tedy je rozptyl $\widehat{N}$ roven

$$\mathrm{var}\widehat{N} = \frac{4\lambda^3 C + 2\lambda^2}{C^2\pi^2}(\mathbb{E}t)^4. \tag{5.12}$$

Optimalita odhadu potom plyne z Raovy-Blackwellovy věty, viz např. [1], věta 7.58.

Označme $\widehat{N}_1$ odhad (5.7) založený na množině $\mathcal{W} = W_1$ všech úseček s lexikografickým minimem uvnitř W a $\widehat{N}_2$ odhad založený na úsečkách s lexikografickým maximem ve W ($\mathcal{W} = W_2$). Z nestrannosti odhadu $\widehat{N}$ okamžitě vidíme, že i odhad $\widetilde{N} = \frac{\widehat{N}_1 + \widehat{N}_2}{2}$ je nestranný. Definujme nyní $D = \int |W \ominus \check{S}_0|\,\Lambda_0(\mathrm{d}S_0)$. Při výpočtu rozptylu $\widetilde{N}$ je třeba vyjádřit

$$\begin{aligned}\mathrm{cov}(\widehat{N}_1, \widehat{N}_2) &= \mathrm{cov}(\Phi(W_1)^2, \Phi(W_2)^2) + \mathrm{cov}(\Phi(W_1), \Phi(W_2))\\ &\quad -\mathrm{cov}(\Phi(W_1)^2, \Phi(W_2)) - \mathrm{cov}(\Phi(W_1), \Phi(W_2)^2).\end{aligned} \tag{5.13}$$

Snadno lze z definice momentové míry odvodit, že pro každou trojici $B_1, B_2, B_3 \in \mathcal{B}^2$ platí

$$\begin{aligned} M_3(B_1 \times B_2 \times B_3) &= \Lambda(B_1 \cap B_2 \cap B_3) + M_2^!((B_1 \cap B_2) \times B_3) \\ &\quad + M_2^!((B_1 \cap B_3) \times B_2) + M_2^!((B_2 \cap B_3) \times B_1) \\ &\quad + M_3^!(B_1 \times B_2 \times B_3). \end{aligned} \tag{5.14}$$

Odtud vzhledem k (1.1) plyne

$$\mathbb{E}[\Phi(W_1)^2\Phi(W_2)] = M_3(W_1 \times W_1 \times W_2) = \lambda D + \lambda^2|W|^2 + 2\lambda^2 D|W| + \lambda^3|W|^3.$$

Je tedy

$$\operatorname{cov}(\Phi(W_1)^2, \Phi(W_2)) = \operatorname{cov}(\Phi(W_1), \Phi(W_2)^2) = \lambda D + 2\lambda^2 D|W|. \tag{5.15}$$

Analogicky jako vztah (5.14) lze odvodit (trochu zdlouhavěji) obdobný vztah i pro M_4, z čehož dostáváme

$$\operatorname{cov}(\Phi(W_1)^2, \Phi(W_2)^2) = 4\lambda^3 D|W|^2 + 4\lambda^2 D|W| + \lambda D + 2\lambda^2 D^2. \tag{5.16}$$

Pro zbývající kovarianci v (5.13) platí

$$\operatorname{cov}(\Phi(W_1), \Phi(W_2)) = \mathbb{E}\sum_{S\in\Phi} \mathbf{1}_{\{S\subseteq W\}} = \lambda \int |W \ominus \check{S}_0|\, \Lambda_0(\mathrm{d}S_0) = \lambda D.$$

Celkem tedy máme pro rozptyl odhadu $\widetilde{N}$ vyjádření

$$\operatorname{var}\widetilde{N} = \frac{2\lambda^3(|W|^3 + |W|^2 D) + \lambda^2(|W|^2 + D^2)}{|W|^4\pi^2}(\mathbb{E}t)^4. \tag{5.17}$$

5.3 Simulace

Vzhledem k tomu, že jsme si většinu vzorců v předchozích podkapitolách odvodili pro stacionární izotropní Poissonův proces úseček na $\mathbb{R}^2$, omezíme se na tento předpoklad i v této podkapitole. Budeme tedy předpokládat, že rozdělení směrů je rovnoměrné na $[0;\pi)$ a nezávislé na (známém) rozdělení délek úseček.

Opět budeme předpokládat, že proces sledujeme skrze čtvercové okno W se stranou a. Pro takové okno pozorování můžeme snadno spočítat konstantu D. Je

$$\begin{aligned} D &= \int\int |W \ominus \check{S}_0(t,\theta)|\rho(\mathrm{d}\theta)\mathcal{D}(\mathrm{d}t) \\ &= \int\int (a - t|\cos\theta|)(a - t\sin\theta)\frac{\mathrm{d}\theta}{\pi}\mathcal{D}(\mathrm{d}t) = a^2 - \frac{4a}{\pi}\mathbb{E}t + \frac{1}{\pi}\mathbb{E}t^2. \end{aligned}$$

Konstanta C pro čtvercové okno je dána vzorcem (5.8).

Budeme opět provádět simulace z Poissonova procesu s intenzitou λ a logaritmicko-normálním rozdělením délek s parametry μ a σ. Připomeňme, že odhady $\widehat{N}_c$, $\widehat{N}$ a $\widetilde{N}$ jsou nestranné a rozptyly pro $\widehat{N}$ a $\widetilde{N}$ jsou dány vzorci (5.12) a (5.17). Symbolem $\widehat{N}_3$ je označen odhad (5.7) pro volbu množiny $\mathcal{W}_2$ (tj. množiny všech úseček protínajících W). V tabulkách 5.1–5.3 jsou shrnuty výsledky pro tisíc simulací provedených z daného rozdělení s parametry $\mu = -0{,}5$ a $\sigma = 0{,}5$. V prvním sloupci jsou uvedeny výběrové průměry m jednotlivých odhadů, ve druhém je odhad střední kvadratické chyby jednotlivých odhadů (medián ze simulací), ve třetím výběrový rozptyl s^2 a v posledním sloupci je skutečná hodnota rozptylu. Ten je uveden pouze u posledních čtyř odhadů, neboť jsme nebyli schopni explicitně vyjádřit teoretický rozptyl odhadů $\widehat{N}_c$ a $\widehat{N}_l$.

	m	med	s^2	var
$\widehat{N}_c$	29,647	72,206	108,27	—
$\widehat{N}_l$	30,496	61,954	100,44	—
$\widehat{N}_1$	35,528	63,861	149,11	137,66
$\widehat{N}_2$	35,780	63,359	163,13	137,66
$\widehat{N}_3$	33,536	35,616	84,785	86,470
$\widetilde{N}$	35,654	49,105	121,72	103,16

Tabulka 5.1: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 15$ a rozdělením délek $LN(-0{,}5; 0{,}5)$. Teoretická hodnota intenzity průsečíků je $N = 33{,}830$. Okno pozorování W je čtverec o straně $a = 1{,}5$

	m	med	s^2	var
$\widehat{N}_c$	83,890	203,79	279,07	—
$\widehat{N}_l$	84,722	192,04	266,09	—
$\widehat{N}_1$	98,448	159,05	372,69	355,01
$\widehat{N}_2$	97,999	139,76	350,87	355,01
$\widehat{N}_3$	93,489	108,14	229,56	246,58
$\widetilde{N}$	98,224	135,35	299,51	285,70

Tabulka 5.2: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 25$ a rozdělením délek $LN(-0{,}5; 0{,}5)$. Teoretická hodnota intenzity průsečíků je $N = 93{,}974$. Okno pozorování W je čtverec o straně $a = 2$.

	m	med	s^2	var
$\widehat{N}_c$	13,771	5,8011	9,6363	—
$\widehat{N}_l$	13,986	4,9371	8,5093	—
$\widehat{N}_1$	15,448	4,3844	9,7872	10,104
$\widehat{N}_2$	15,521	4,3844	10,094	10,104
$\widehat{N}_3$	15,127	3,7590	7,7319	7,8123
$\widehat{N}$	15,485	3,7055	8,7085	8,7330

Tabulka 5.3: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 10$ a rozdělením délek $LN(-0{,}5; 0{,}5)$. Teoretická hodnota intenzity průsečíků je $N = 15{,}036$. Okno pozorování W je čtverec o straně $a = 3$.

	m	med	s^2	var
$\widehat{N}_c$	14,468	0,7155	1,0589	—
$\widehat{N}_l$	14,499	0,6756	0,9416	—
$\widehat{N}_1$	15,121	0,4476	0,8726	0,9048
$\widehat{N}_2$	15,124	0,4091	0,8718	0,9048
$\widehat{N}_3$	15,078	0,3773	0,8079	0,8319
$\widehat{N}$	15,123	0,3986	0,8331	0,8660

Tabulka 5.4: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 10$ a rozdělením délek $LN(-0{,}5; 0{,}5)$. Teoretická hodnota intenzity průsečíků je $N = 15{,}036$. Okno pozorování W je čtverec o straně $a = 10$.

Na závěr této kapitoly si uveďme ještě odhady intenzity pro Poissonův proces úseček s rovnoměrným rozdělením délek na $(0; M]$. Na rozdíl od logaritmicko-normálního rozdělení je zde splněn předpoklad $L(S_0) \leq a$ s.j., tedy střední hodnota odhadu $\widehat{N}_l$ je podle (5.10)

$$\mathbb{E}\widehat{N}_l = N + \frac{\lambda}{\pi a^4}\left(\frac{a^2M^2}{3} - \frac{aM^3}{3\pi} + \frac{M^4}{30\pi}\right). \tag{5.18}$$

V tabulkách (5.5) a (5.6) jsou uvedeny odhady získané při tisíci simulacích s rovnoměrným rozdělením délek s parametrem $M = 0{,}25$. Značení je analogické jako v předchozím případě.

	m	med	s^2	var
$\widehat{N}_c$	12,488	57,245	99,477	—
$\widehat{N}_l$	13,430	22,039	61,575	—
$\widehat{N}_1$	12,453	18,295	50,313	51,452
$\widehat{N}_2$	12,480	18,295	52,998	51,452
$\widehat{N}_3$	12,423	16,814	38,579	38,666
$\widetilde{N}$	12,466	18,637	43,696	43,741

Tabulka 5.5: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 50$ a rovnoměrným rozdělením délek na $(0; 0{,}25]$. Teoretická hodnota intenzity průsečíků je $N = 12{,}4340$, $\mathbb{E}\widehat{N}_l = 13{,}60$. Okno pozorování W je čtverec o straně $a = 0{,}5$.

	m	med	s^2	var
$\widehat{N}_c$	4,3680	2,3219	7,2358	—
$\widehat{N}_l$	4,6104	1,5263	3,6312	—
$\widehat{N}_1$	4,4053	1,2191	2,7893	2,7161
$\widehat{N}_2$	4,4426	1,2191	2,8561	2,7161
$\widehat{N}_3$	4,4121	1,0191	2,3771	2,3379
$\widetilde{N}$	4,4239	1,2301	2,6356	2,5061

Tabulka 5.6: Odhady intenzity průsečíků pro stacionární a izotropní Poissonův proces úseček s intenzitou $\lambda = 30$ a rovnoměrným rozdělením délek na $(0; 0{,}25]$. Teoretická hodnota intenzity průsečíků je $N = 4{,}4762$, $\mathbb{E}\widehat{N}_l = 4{,}6597$. Okno pozorování W je čtverec o straně $a = 1$.

Po porovnání odhadů na základě výpočtů uvedených v tabulkách 5.1-5.6 se zdá, že nejvhodnějším odhadem je odhad $\widehat{N}_3$ a to jak do samotné hodnoty, tak i střední kvadratické chyby a rozptylu. Tento odhad je založený na informaci poskytované všemi pozorovanými úsečkami narozdíl například od ostatních odhadů daných (5.7). Nevýhodou těchto odhadů je, že předpokládáme znalost střední délky typické úsečky. Neparametrické odhady $\widehat{N}_l$ a $\widehat{N}_c$ tuto znalost nevyžadují. Přestože odhad $\widehat{N}_l$ je vychýlený (pro rovnoměrné rozdělení délek je jeho střední hodnota dána (5.18)), jeví se na základě simulací lépe než naivní nestranný odhad $\widehat{N}_c$.

Literatura

[1] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2005.

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

[3] Baddeley A. J., Turner R.: Spatstat: an R package for analyzing spatial point patterns, *Journal of Statistical Software* **12**, 1–42, 2005.

[4] Chadœuf J., Senoussi R., Yao J. F.: Parametric estimation of a Boolean segment process with stochastic restoration estimation, *Journal of Computational and Graphical Statistics* **9**, 390–402, 2000.

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

[6] Honzl O.: Bodové procesy úseček, *bakalářská práce*, MFF UK, Praha, 2006.

[7] Mrkvička T.: Estimation on intersection intensity in a Poisson process of segments, *Commentationes Mathematicae Universitatis Carolinae* **48**, 93–106, 2007.

[8] R Development Core: *R: A language and environment for statistical computing*, Vienna, Austria, 2008.

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

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

[11] Wijers B. J.: *Nonparametric estimation for a windowed line-segment process*, Stichting Mathematisch Centrum, Amsterdam, 1997.