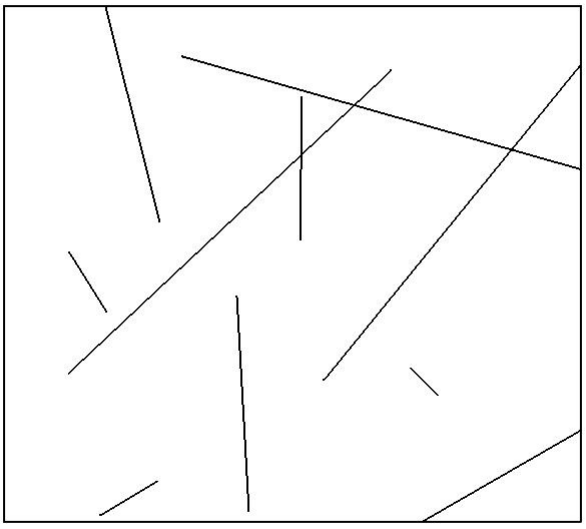
Viivasegmenttiprosessin tunnusten estimointi

Mikko Niilo-Rämä



Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
11. helmikuuta 2011

Mikko Niilo-Rämä: *Viivasegmenttiprosessin tunnusten estimointi*
Pro gradu -tutkielma, 24 sivua, 1 liite
Tilastotiede
11. helmikuuta 2011

Tiivistelmä

Tutkielmassa käsitellään tasossa olevan viivasegmenttiprosessin kahta ensimmäisen kertaluvun tunnusta: intensiteettiä ja pituusjakauman odotusarvoa. Uusina menetelminä esitellään intensiteetin estimoinnissa kahden referenssipisteen käyttö sekä plusotannan aiheuttaman harhan korjaus käyttäen harhalle laskettua odotusarvoa. Pituusjakauman odotusarvon estimoinnissa uutta on sisältymissuhteeseen perustuva estimaattori.

Tilastollisena mallina prosessille käytetään Boolen mallia, joka on erikoistapaus germ–grain-mallista. Lisäksi oletetaan viivasegmenttien pituuden noudattavan eksponenttijakaumaa ja suuntakulman suhteessa x-akseliin tasajakaumaa.

Tunnuslukujen estimoinnissa tarvittava otanta suoritetaan käyttämällä neliön muotoista havaintoikkunaa. Eri otantatekniikoista esitellään plusotanta, joka sisältää kaikki ikkunaa leikkaavat segmentit, sekä miinusotanta, joka sisältää vain kokonaan ikkunaan mahtuneet segmentit. Nämä otantatekniikat sisältävät harhan, jonka suuruus on mahdollista laskea, mikäli viivasegmenttien pituusjakauma tunnetaan. Kolmantena otantatekniikkana esitellään referenssipisteotanta, jossa yksittäisen viivasegmentin mukaantulo otokseen määräytyy siihen liitetyn referenssipisteen perusteella. Tämä otantatekniikka osoitetaan harhattomaksi.

Simulointikokeilla osoitetaan, että Boolen mallin tapauksessa käytettäessä yhden sijaan kahta eri referenssipistettä, saadaan estimaattorin varianssia pienennettyä. Näin käy erityisesti silloin, kun ikkunan koko suhteessa viivasegmenttien keskipituuteen on pieni. Tarkin estimaattori intensiteetille saadaan kuitenkin käyttämällä harhasta korjattua plusotantaa.

Pituusjakauman odotusarvon estimaattoreista tarkastellaan mm. perinteistä suurimman uskottavuuden menetelmää sekä Kaplan–Meier-estimaattoria. Lisäksi esitellään sisältymissuhde-estimaattori, joka perustuu plus- ja miinusotantojen tuottamien otoskokojen suhteeseen eikä edellytä yhdenkään viivasegmentin pituuden tuntemista.

Simulointien perusteella SU-menetelmä osoittautuu erittäin tarkaksi, mutta sisältymissuhteeseen perustuva estimaattori on kuitenkin varsin kilpailukykyinen sen kanssa. Sen sijaan Kaplan–Meier-estimaattori osoittautuu selvästi epätarkemmaksi ja jopa jonkin verran harhaiseksi. Kaikki estimaattorit kuitenkin tarkentuvat varsin nopeasti ikkunan koon kasvaessa.

Avainsanoja: Boolen malli, intensiteetti, miinusotanta, pituusjakauma, plusotanta, referenssipisteotanta, sisältymissuhde-estimaattori, viivasegmenttiprosessi

Sisältö

1 Johdanto

Viivasegmenttiprosessi on suljettu satunnaisjoukko, joka koostuu satunnaisiin paikkoihin sijoittuneista satunnaisen pituisista ja suuntaisista janoista (viivasegmenteistä). Tässä tutkielmassa käsitellään tason $\mathbb{R}^2$ viivasegmenttiprosessia, ja mielenkiinnon kohteena on prosessin kaksi keskeistä ensimmäisen kertaluvun tunnusta: *intensiteetti*, joka tarkoittaa segmenttien lukumäärän odotusarvoa pinta-alayksikköä kohti, sekä segmenttien *pituusjakauman odotusarvo*. Uusia asioita ovat intensiteetin estimoinnissa *kahden referenssipisteen* käyttö ja *plusotannan keskimääräiseen harhaan* perustuva *harhankorjaus* sekä pituusjakauman odotusarvon estimoinnissa *sisältymissuhteeseen* perustuva estimaattori.

Tilastollisena mallina viivasegmenttiprosessille esitellään *merkkinen pisteprosessi* [2] sekä *germ–grain-malli* [5, 8], joka koostuu *pisteprosessin* tuottamista pisteistä (segmenttien sijainnit) sekä tason kompakteista[1] satunnaisjoukoista, joilla on kaksi ominaisuutta: pituus ja suuntaus. Tässä työssä viivasegmenttien pituuden oletetaan noudattavan eksponenttijakaumaa ja suuntakulman suhteessa x-akseliin tasajakaumaa välillä $[0, 2\pi)$. Näitä mallioletuksia tarvitaan myöhemmin mm. *Campbell–Mecken* lausetta sekä simuloinnin toteutusta varten.

Tunnusten estimointi perustuu otokseen, joka saadaan tason konveksista[2] osajoukosta. Osajoukko on tyypillisesti suorakulmio ja sitä kutsutaan usein myös ikkunaksi (Kuva 1). Estimoinnin erityiskysymys liittyy reunaefektiin, eli siihen, että kaikkia viivasegmenttejä ei havaita kokonaan, vaan ne jatkuvat ikkunan ulkopuolelle. Tässä työssä havaintoikkuna oletetaan neliön muotoiseksi ja ideana on tutkia mm. ikkunan koon vaikutusta tunnuslukujen estimointiin.

Rajoitettu havaintoikkuna aiheuttaa estimointiin kaksi perusongelmaa, jotka ovat *otantaharha* ja *sensurointi*. Otantaharha aiheutuu siitä, että pidempi viivasegmentti osuu suuremmalla todennäköisyydellä havaintoikkunaan kuin lyhyempi, jolloin yksinkertaisen satunnaisotannan oletus ei toteudu. Otantaharha saadaan kuitenkin eliminoitua otantatekniikkaa muuttamalla tai havaintoja painottamalla, mikä johtaa ns. *Horvitz–Thompson*-tyyppiseen estimaattoriin [4].

Jos viivasegmentti ei mahdu kokonaan havaintoikkunaan, vaan siitä havaitaan vain osa, kyseessä on sensuroitu havainto. Yksittäinen segmentti saattaa olla joko toisesta tai molemmista päistä sensuroitu. Tällöin havaitaan vain segmentin ja havaintoikkunan leikkaus, jonka perusteella saadaan alaraja kyseisen segmentin pituudelle. Pituusjakaumaa estimoitaessa myös sensuroidut havainnot kannattaa hyödyntää, koska ne antavat tietoa todennäköisyydestä, jolla viivasegmentin pituus on suurempi kuin havaitun sensuroidun segmentin pituus.

Tässä työssä esiteltävät otantatekniikat ovat *plusotanta*, joka sisältää kaikki ikkunaa leikkaavat segmentit, *miinusotanta*, joka sisältää vain kokonaan ikkunaan mahtuvat segmentit, sekä *referenssipisteotanta*, jossa segmentti tulee mukaan otokseen, mikäli sen referenssipiste osuu ikkunaan (Kuvat 2 ja 3). Nämä otantatekniikat tuottavat satunnaisen otoskoon, jota käytetään tunnuslukuna intensiteettiestimaattorissa.

Alun perin kysymyksen sensuroitujen viivasegmenttien pituusjakauman estimoimisesta esittivät geologit, jotka tarvitsivat estimaatin kaivoksen seinämässä olevien

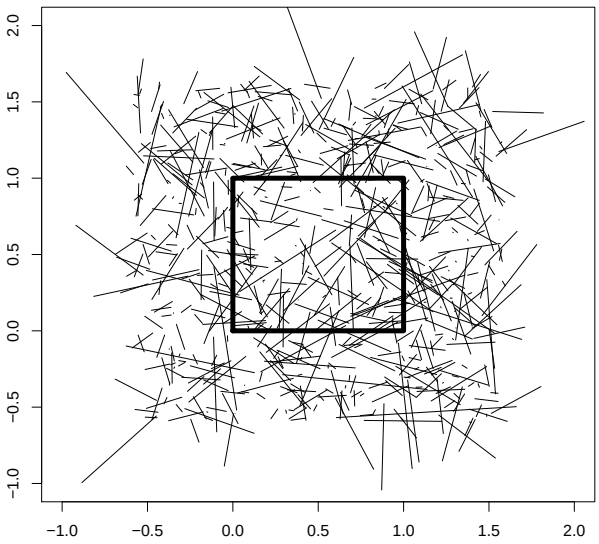
[1] Kompaktilla joukolla tarkoitetaan joukkoa, joka on sekä suljettu, että rajoitettu.

[2] Joukko $A \subset \mathbb{R}^2$ on konveksi, jos kaikille $a, b \in A$ niitä yhdistävä jana $j(a, b) \subset A$.

halkeamien pituusjakaumalle. Kaivoksen seinämä muodosti suorakaiteen muotoisen ”ikkunan”, jolloin osa halkeamista havaittiin vain osittain [1].

Tässä tutkielmassa käsitellään pituusjakauman odotusarvon estimoimiseksi mm. *suurimman uskottavuuden* menetelmää sekä *Kaplan–Meier*-estimaattoria, joita käytetään yleisesti elinaika-analyysissa [10]. Näille estimaattoreille käytetään referenssipisteotantaa. Lisäksi esitellään uusi *sisältymissuhde-estimaattori*, joka perustuu plus- ja miinusotantojen tuottamien otoskokojen suhteeseen. Se ei siis edellytä yhdenkään viivasegmentin pituuden tuntemista.

Tutkielman Luvussa 2 esitellään viivasegmenttiprosessille tilastollinen malli. Luvussa 3 käsitellään spatiaalista otantaa ja osoitetaan Campbell–Mecken lauseen avulla plus- ja miinusotantojen harhaisuus sekä referenssipisteotannan harhattomuus. Neljännessä luvussa testataan simulointikokeita [14] käyttäen em. otantatekniikoiden ominaisuuksia intensiteetin estimoinnissa ja esitellään lisäksi kahden referenssipisteen menetelmä, jolla saadaan pienennettyä estimaattorin varianssia. Luvussa 5 esitellään pituusjakauman odotusarvon estimointimenetelmiä ja vertaillaan niiden ominaisuuksia simulointikokeilla. Luvussa 6 esitellään harhankorjausmenetelmä plusotannalle ja vertaillaan eri intensiteettiestimaattoreiden ominaisuuksia sekä teoreettisesti että simuloimalla. Luvussa 7 on lyhyt yhteenveto tähän mennessä tehdyistä asioista ja pohdintoja siitä, mitä voisi miettiä tulevaisuudessa.



Kuva 1: Simuloitu viivasegmenttiprosessin realisaatio sekä havaintoikkuna.

2 Tilastollisia malleja

Tässä luvussa konstruoidaan viivasegmenttiprosessille tilastollinen malli, jota tarvitaan myöhemmässä vaiheessa teoreettisten tulosten johtamista sekä simulointeja varten.

2.1 Pisteprosesseista

Seuraava esitys perustuu lähteisiin [2, 4, 7]. *Pisteprosessit* ovat stokastisia malleja *pistekuvioille*. Matemaattisesti niitä voidaan tarkastella abstrakteissa avaruuksissa, mutta tässä työssä liikutaan pääosin euklidisessa avaruudessa $\mathbb{R}^2$. Termi "prosessi" voi olla harhaanjohtava, koska siihen yleensä liitetään jokin ajassa tapahtuva muutos. Pisteprosessien avulla käsitellään kuitenkin usein ajasta riippumattomia ilmiöitä, kuten tässä työssä.

Pistekuvio on kokoelma pisteitä jollakin alueella tai joukossa. Tyypillisesti se tulkitaan otokseksi tai realisaatioksi pisteprosessista. Pisteet on usein tapana numeroida merkinnällisistä syistä, mikä ei kuitenkaan tarkoita sitä, että pisteillä olisi jokin looginen järjestys.

Määritelmä 1 *Pisteprosessi Φ avaruudessa $\mathbb{R}^d$ on satunnaismuuttuja, eli mitallinen kuvaus, joka saa arvoja avaruudessa $(\mathbb{N}, \mathcal{N})$, missä $\mathbb{N}$ on kaikkien sellaisten $\mathbb{R}^d$:n pistekokoelmien ϕ perhe, jotka toteuttavat seuraavat ehdot:*

1. *ϕ on* lokaalisti äärellinen*, ts. jokainen rajoitettu $\mathbb{R}^d$:n osajoukko sisältää äärellisen määrän pisteitä*

2. *ϕ on* yksinkertainen*, ts. tietyssä paikassa voi olla korkeintaan yksi piste.*

Määritelmässä $\mathcal{N}$ on pienin σ-algebra avaruudella $\mathbb{N}$, jolle kaikki kuvaukset $\phi \to \phi(B)$ ovat mitallisia, missä B käy läpi kaikki rajoitetut Borelin[3] joukot $\mathbb{R}^d$:ssä.

Pisteprosessi $\Phi(B)$ voidaan tulkita satunnaiseksi *lukumäärämitaksi*, joka antaa joukkoon B sisältyvien pisteiden lukumäärän. Vaihtoehtoisesti pisteprosessi voidaan määritellä suljettuna satunnaisjoukkona, josta kerrotaan lisää Luvussa 2.2.1.

Tässä työssä oletetaan pisteprosessille kaksi tärkeää ominaisuutta, jotka ovat *stationaarisuus* ja *isotrooppisuus*.

Määritelmä 2 *Pisteprosessin Φ sanotaan olevan stationaarinen, jos sekä Φ:llä, että siirretyllä prosessilla Φ_x on sama jakauma kaikilla siirroilla $x \in \mathbb{R}^d$. Siis*

$$\Phi = \{x_1, x_2, \ldots\} \stackrel{D}{=} \Phi_x = \{x_1 + x, x_2 + x, \ldots\}.$$

Määritelmä 3 *Pisteprosessi Φ on isotrooppinen, jos se on invariantti kiertojen suhteen. Tasossa tämä tarkoittaa sitä, että jos α on mielivaltainen kulma väliltä $[0, 2\pi)$, niin*

$$\Phi = \{x_1, x_2, \ldots\} \stackrel{D}{=} R_\alpha \Phi = \{R_\alpha x_1, R_\alpha x_2, \ldots\},$$

missä $R_\alpha x_i = R_\alpha(x_{i1}, x_{i2}) = (x_{i1} \cos\alpha - x_{i2} \sin\alpha, x_{i1} \sin\alpha + x_{i2} \cos\alpha)$.

[3] Borelin σ-algebra $\mathbb{R}^d$:ssä on pienin σ-algebra, joka sisältää kaikki avoimet $\mathbb{R}^d$:n osajoukot [11].

2.1.1 Stationaarinen Poisson-pisteprosessi

Määritelmä 4 *Pisteprosessi Φ avaruudessa $\mathbb{R}^d$ on stationaarinen Poisson-pisteprosessi, jos se toteuttaa seuraavat ehdot:*

1. *Pisteiden lukumäärä $\Phi(B)$ mille tahansa rajoitetulle Borelin joukolle $B \subset \mathbb{R}^d$ noudattaa Poisson-jakaumaa odotusarvolla $\lambda \cdot \nu(B)$, missä $\nu(B)$ on B:n Lebesguen mitta (pinta-ala tapauksessa d=2).*

2. *Pisteiden lukumäärät $\Phi(B_i)$ prosessin Φ erillisille joukoille $B_i, i = 1, \ldots, k$, ovat toisistaan riippumattomia millä tahansa kokonaisluvulla k.*

Parametria λ, joka on pisteiden lukumäärän odotusarvo pinta-alayksikköä kohti, kutsutaan prosessin intensiteetiksi eli pistetiheydeksi.

2.1.2 Merkkinen pisteprosessi

Usein on tarpeellista liittää pisteprosessin pisteisiin jotakin informaatiota. Tätä informaatiota kutsutaan *merkiksi* ja tällöin pisteprosessista tulee *merkkinen pisteprosessi*. Merkit voivat olla yksinkertaisimmillaan esim. reaalilukuja tai vektoreita, mutta niillä voi myös olla hyvinkin monimutkainen rakenne (esim. kompaktit joukot $\mathbb{R}^d$:ssä).

Määritelmä 5 *Olkoon $\Phi = \{x_1, x_2, \ldots\}$ pisteprosessi $\mathbb{R}^d$:ssä sekä S merkkiavaruus. Kun prosessin Φ jokaiseen pisteeseen x_i liitetään merkki $m(x_i) \in S$, saadaan avaruuteen $\mathbb{R}^d \times S$ pisteprosessi $\Psi = \{x_i, m(x_i)\}$, jota kutsutaan merkkiseksi pisteprosessiksi.*

Merkkisellä pisteprosessilla päästään käsiksi viivasegmenttiprosessiin. Oletetaan, että jokaisella viivasegmentillä on yksikäsitteinen referenssipiste x_i, joka määrää segmentin sijainnin. Nyt jokaiseen pisteeseen x_i liittyy *merkki* $m(x_i) = (\theta, r) \in S$, joka sisältää tiedon segmentin suunnasta θ (kulma x-akseliin nähden) sekä pituudesta r. Viivasegmenttiprosessi on siis merkkinen pisteprosessi, joka koostuu pisteistä $(x, s) \in \mathbb{R}^2 \times S$, missä merkkiavaruus $S = (\Theta, R) = ([0, 2\pi), [0, \infty))$.

2.2 Germ–grain-malli

Vaihtoehtoinen lähestymistapa viivasegmenttiprosessille on *germ–grain-malli* [4, 5]. Tämä vaatii hieman tarkempaa tutustumista satunnaisjoukkoihin ja niitä koskeviin oletuksiin.

2.2.1 Satunnaisjoukko ja sen jakauma

Seuraavassa on käytetty lähteitä [5, 7, 8, 9]. Olkoon $\mathbb{F}$ suljettujen $\mathbb{R}^d$:n osajoukkojen perhe. Varustetaan se pienimmällä *osuma-σ*-algebralla $\mathcal{F}$, jonka generoivat joukot $\mathbb{F}_K = \{F \in \mathbb{F} : F \cap K \neq \emptyset\}$, missä K on annettu kompakti "testijoukko" $\mathbb{R}^d$:ssä.

Määritelmä 6 *Olkoon $(\Omega, \mathcal{A}, \mathbb{P})$ todennäköisyysavaruus. Mitallinen kuvaus $\Xi : \Omega \to \mathbb{F}$ on suljettu satunnaisjoukko, jos kaikille kompakteille joukoille $K \subset \mathbb{R}^d$*

$$\{\omega : \Xi(\omega) \cap K \neq \emptyset\} \in \mathcal{A}.$$

Nyt siis on mahdollista asettaa todennäköisyydet tapahtumille ”joukko Ξ leikkaa joukkoa K”. Tämä tuottaa jakauman P avaruuteen $(\mathbb{F}, \mathcal{F})$ ja sitä kutsutaan satunnaisjoukon Ξ jakaumaksi. Siis

$$P(A) = \mathbb{P}(\Xi \in A),$$

missä $A \in \mathcal{F}$.

Jakauman P karakterisointi on helppo ymmärtää *kapasiteettifunktionaalin* T avulla:

$$\begin{aligned} T_\Xi(K) &= \mathbb{P}(\Xi \cap K \neq \emptyset) \\ &= P(\mathbb{F}_K). \end{aligned}$$

Kapasiteettifunktionaali on verrattavissa satunnaismuuttujan kertymäfunktioon ja se määrää satunnaisjoukon jakauman yksikäsitteisesti (Choquet'n lause [7]).

Olkoon $\mathbb{K} \subset \mathbb{F}$ merkkiavaruus, johon kuuluvat kaikki $\mathbb{R}^d$:n kompaktit epätyhjät osajoukot. Määritellään jokaiselle joukolle $K \in \mathbb{K}$ referenssipiste $c(K) \in \mathbb{R}^d$. Oletetaan, että kuvaus $c : \mathbb{K} \to \mathbb{R}^d$ on mitallinen ja siirtoinvariantti eli $c(K+x) = c(K)+x$ kaikille $x \in \mathbb{R}^d$. Rajoitutaan vielä sellaisiin kompakteihin satunnaisjoukkoihin, joiden referenssipiste on origossa, merkitään $\mathbb{K}_0 = \{K \in \mathbb{K} : c(K) = 0\}$. Liitetään $\mathbb{K}_0$:aan vielä sigma-algebra $\mathcal{F}_{\mathbb{K}_0} = \{A \cap \mathbb{K}_0 : A \in \mathcal{F}\}$. Näin saadaan jakauma Q merkkiavaruuteen $(\mathbb{K}_0, \mathcal{F}_{\mathbb{K}_0})$.

Germ–grain-malli on yksi tärkeimmistä satunnaisjoukoista. Se määritellään sellaisen merkkisen pisteprosessin avulla, jossa merkit ovat kompakteja joukkoja.

Määritelmä 7 *Olkoon $\Psi = \{(x_i, K_i)\}$ merkkinen pisteprosessi, missä x_i (germ) on objektin sijainti $\mathbb{R}^d$:ssä ja K_i (grain) on merkki avaruudessa $\mathbb{K}_0$. Kun lisätään prosessin jokaiseen pisteeseen x_i vastaava satunnaisjoukko K_i, saadaan joukko $X_i = x_i + K_i$, eli siirretty grain. Näin saatua yhdistettä*

$$\Xi = \bigcup_i (x_i + K_i)$$

kutsutaan germ–grain-*malliksi.*

Jos oletetaan edellisessä tilanteessa, että germien pisteprosessi Φ on stationaarinen Poisson-pisteprosessi, ja lisäksi satunnaisjoukot K_i ovat riippumattomia toisistaan sekä prosessista Φ, saadaan tärkeä erikoistapaus germ–grain-mallista. Tätä kutsutaan *Boolen malliksi.* Tämän tutkielman tulokset ja simulointikokeet perustuvat Boolen mallin oletukseen.

2.2.2 Germ–grain-malli vs. merkkinen pisteprosessi

Määritelmän mukaan germ–grain-malli on siis satunnaisjoukkojen yhdiste. Näin ollen havaitusta realisaatiosta ei pystytä yleisessä tapauksessa erottelemaan alkuperäisiä joukkoja, mikäli ne leikkaavat toisiaan. Tämä on germ–grain-mallin merkittävin ero merkkiseen pisteprosessiin verrattuna. Merkkisessä pisteprosessissa merkit ovat eri avaruudessa kuin pisteet, jolloin ne pystytään aina yksilöimään. Jos germ–grain-mallille tehdään rotaatio, se vaikuttaa koko yhdisteeseen, eli germien lisäksi jokaiseen grainiin. Merkkisen pisteprosessin tapauksessa rotaatio muuttaa vain pisteprosessia, merkit eivät muutu.

3 Spatiaalinen otanta

Kun spatiaalinen prosessi havaitaan rajoitetussa ikkunassa, prosessin tilastollista päättelyä häiritsee ns. reunaefekti, joka voidaan jakaa kahteen päätyyppiin. Nämä ovat objektien erilaisesta koosta johtuva otantaharha sekä sensurointiefekti. Otantaharha voidaan eliminoida muuttamalla otantatekniikkaa, tai sitä voidaan korjata havaintoja painottamalla. Sensurointiefektiin taas voidaan käyttää elinaika-analyysin metodeja. Edellä mainittuja asioita on käsitelty lähteissä [3], [4] ja [5]. Seuraava esitys perustuu pääosin lähteeseen [4].

3.1 Otantaharha

Otantaharha tulee kyseeseen silloin, kun geometrisen objektin havaitsemistodennäköisyys riippuu sen koosta tai muodosta. Viivasegmenttien tapauksessa voidaan osoittaa, että mitä suurempi on segmentin pituus, sitä suuremmalla todennäköisyydellä se leikkaa annettua joukkoa. Tämä osoitetaan yksikköneliölle Luvussa 5. Havaintoikkunaan perustuvassa otannassa pidemmistä viivasegmenteistä saadaan siis suhteellisesti enemmän havaintoja.

Jos kaikki ikkunaa W leikkaavat segmentit $X_i : X_i \cap W \neq \emptyset$ otetaan mukaan otokseen, puhutaan *plusotannasta*. Toisaalta voitaisiin jättää kaikki ikkunan reunaa leikkaavat segmentit pois, jolloin otoksessa olisivat mukana vain kokonaisina havaitut segmentit $X_i : X_i \subset W$. Tällöin kyseessä on *miinusotanta*. Plus- ja miinusotantoja on havainnollistettu Kuvassa 2.

Molemmat ovat harhaisia otantamenetelmiä sekä intensiteetin että pituuden odotusarvon estimoinnin kannalta. Plusotanta suosii pidempiä viivasegmenttejä, miinusotanta lyhyempiä. Tämän osoittamiseksi tarvitaan seuraava lause [4].

Lause 1 (Campbell–Mecke) *Olkoon $\{(x_i, K_i)\}$ stationaarinen germ–grain-prosessi, missä $x_i \in \mathbb{R}^d$ ja $K_i \in \mathbb{K}_0$. Tällöin mille tahansa mitalliselle kuvaukselle $f : \mathbb{K}_0 \to \mathbb{R}_+$,*

$$\mathbb{E}\left[\sum_i f(X_i)\right] = \lambda \int_{\mathbb{R}^d} \int_{\mathbb{K}_0} f(K + x) dQ(K) dx,$$

missä λ on $\{x_i\}$:n intensiteetti ja Q on grainien jakauma. Toisin sanoen

$$\mathbb{E}\left[\sum_i f(X_i)\right] = \lambda \int_{\mathbb{R}^d} \mathbb{E}^0 \left[f(K_0 + x)\right] dx = \lambda \mathbb{E}^0 \left[\int_{\mathbb{R}^d} f(K_0 + x) dx\right],$$

missä $K_0 \in \mathbb{K}_0$ on "tyypillinen" grain, eli kompakti satunnaisjoukko jakaumasta Q ja $\mathbb{E}^0$ tarkoittaa odotusarvoa jakauman Q suhteen.

3.1.1 Intensiteetin estimointi plus- ja miinusotannoilla

Campbell–Mecken lauseen avulla saadaan laskettua odotusarvo otokseen tulevien objektien määrälle eri otantamenetelmissä. Oletetaan, että ikkuna W on yksikköneliö. Kuvaus $\mathbf{1} : \mathbb{K}_0 \to \{0, 1\}$ on mitallinen, joten plusotannan tapauksessa

$$\begin{aligned}\mathbb{E}\left[\#\{i : X_i \cap W \neq \emptyset\}\right] &= \mathbb{E}\left[\sum_i \mathbf{1}_{\{X_i \cap W \neq \emptyset\}}\right] \\ &= \lambda \mathbb{E}^0\left[\int_{\mathbb{R}^2} \mathbf{1}_{\{(K_0+x)\cap W \neq \emptyset\}} dx\right] \\ &= \lambda \mathbb{E}^0\left[\nu(W \oplus \check{K}_0)\right].\end{aligned}$$

Tässä siis λ on prosessin intensiteetti ja $W \oplus \check{K}_0$ tarkoittaa Minkowskin summaa:

$$W \oplus \check{K} = \{x \in \mathbb{R}^d : (K + x) \cap W \neq \emptyset\},$$

missä $\check{K}$ on K:n reflektio origon suhteen, siis $\check{K} = \{x : -x \in K\}$. Ks. Kuva 5 sivulla 18.

Oletetaan nyt, että K_0 on "tyypillinen" eli satunnaisesti generoitu viivasegmentti jakaumasta Q ja W on yksikköneliö, siis $\nu(W) = 1$. Koska viivasegmentin pituudella on positiivinen odotusarvo, on

$$\mathbb{E}^0\left[\nu(W \oplus \check{K}_0)\right] > 1.$$

Siten $\mathbb{E}\left[\#\{i : X_i \cap W \neq \emptyset\}\right] > \lambda$, joten yksikköneliötä leikkaavien segmenttien lukumäärä on ylöspäin harhainen estimaattori intensiteetille λ. Vastaava lasku antaa miinusotannalle

$$\mathbb{E}\left[\#\{i : X_i \subset W\}\right] = \lambda \mathbb{E}^0\left[\nu(W \ominus K_0)\right] < \lambda,$$

missä $W \ominus K_0$ on Minkowskin erotus:

$$W \ominus K = \{x \in \mathbb{R}^d : (K + x) \subset W\}.$$

Huom. Tämän työn oletuksilla odotusarvot $\mathbb{E}^0\left[\nu(W \oplus \check{K}_0)\right]$ ja $\mathbb{E}^0\left[\nu(W \ominus \check{K}_0)\right]$ on mahdollista laskea. Tähän palataan sisältymissuhde-estimaattorin yhteydessä Luvussa 5.2.2.

3.1.2 Painotettu jakauma ja otantaharhafaktori

Olkoon $f : \mathbb{K}_0 \to \mathbb{R}_+$ mitallinen kuvaus ja $I : \mathbb{K}_0 \to \{0, 1\}$ otantasääntö, missä objekti $X_i \subset \mathbb{R}^2$ tulee mukaan otokseen jos ja vain jos $I(X_i) = 1$. Tällöin

$$\mathbb{E}\left[\sum_{\text{otos}} f(X_i)\right] = \mathbb{E}\left[\sum I(X_i) f(X_i)\right] = \lambda \mathbb{E}^0\left[f(K_0)\pi(K_0)\right], \tag{1}$$

missä $K_0 \in \mathbb{K}_0$ ja

$$\pi(K) = \int_{\mathbb{R}^d} I(K + x) dx \tag{2}$$

on Lebesguen mitta joukolle $\{x \in \mathbb{R}^d : (K + x)$ tulee mukaan otokseen$\}$. Sitä kutsutaan myös *otantaharhafaktoriksi*. Merkitään otantasäännön I generoimaa satunnaista otoskokoa N_I:llä. Otokseen mukaan tulevat objektit ovat π-painotettuja s.e.

$$\frac{\mathbb{E}\left[\sum_{\text{otos}} f(X_i)\right]}{\mathbb{E}\left[N_I\right]} = \frac{\mathbb{E}\left[\sum_i I(X_i) f(X_i)\right]}{\mathbb{E}\left[\sum_i I(X_i)\right]} = \frac{\mathbb{E}^0\left[f(K_0)\pi(K_0)\right]}{\mathbb{E}^0\left[\pi(K_0)\right]}. \tag{3}$$

Tämä on f:n odotusarvo π-painotetun jakauman Q suhteen. Nyt siis plusotannalle $\pi(K) = \nu(W \oplus \check{K}_0)$ ja vastaavasti miinusotannalle $\pi(K) = \nu(W \ominus K_0)$.

Plus- ja miinusotantojen tuottama harha voidaan korjata painottamalla otokseen tulevia objekteja. Plusotannassa objektille X_i annetaan paino $1/\nu(W \oplus \check{X}_i)$ ja miinusotannan tapauksessa $1/\nu(W \ominus X_i)$. Tämä johtaa ns. *Horvitz–Thompson*-tyyppiseen estimaattoriin, josta kerrotaan tarkemmin Luvussa 5.1.1. Toinen mahdollisuus on korvata esim. plusotannan tapauksessa pinta-ala $\nu(W \oplus \check{X}_i)$ sen odotusarvolla $\mathbb{E}^0\left[\nu(W \oplus \check{X}_i)\right]$. Tähän palataan Luvussa 6.

3.1.3 Harhaton otantasääntö

Miten saadaan konstruoitua harhaton otantasääntö? Yhtälöä (3) tarkasteltaessa huomataan, että jos $\pi(K_0)$ on vakio, niin oikealle puolelle jää ainoastaan $\mathbb{E}^0\left[f(K_0)\right]$, joka on siis f:n odotusarvo jakauman Q suhteen, missä Q on painottamaton "tyypillisen" grainin jakauma. Otantasääntö on siis harhaton, jos objektin mukaantulo otokseen ei riipu sen koosta.

Esimerkki harhattomasta otantamenetelmästä on Milesin vuonna 1978 esittämä *referenssipistesääntö* [4]. Siinä jokaiseen objektiin $K \in \mathbb{K}$ liitetään yksikäsitteinen piste $c(K) \in \mathbb{R}^d$. Referenssipiste voidaan valita mielivaltaisesti, kunhan sille pätee $c(K + x) = c(K) + x$ kaikilla $K \in \mathbb{K}$ ja $x \in \mathbb{R}^d$. Objekti X_i tulee mukaan otokseen jos ja vain jos sen referenssipiste $c(X_i)$ osuu ikkunaan W.

Nyt otantaharhafaktori (2) on

$$\begin{aligned}
\pi(K) &= \int_{\mathbb{R}^d} \mathbf{1}_{\{c(K+x)\in W\}} dx \\
&= \int_{\mathbb{R}^d} \mathbf{1}_{\{c(K)+x\in W\}} dx \\
&= \nu(W - c(K)) \\
&= \nu(W),
\end{aligned}$$

joka ei riipu objektista K. Referenssipistesääntö on siis harhaton. Sijoittamalla edellisen yhtälöön (1), saamme yhtälöä (3) käyttäen

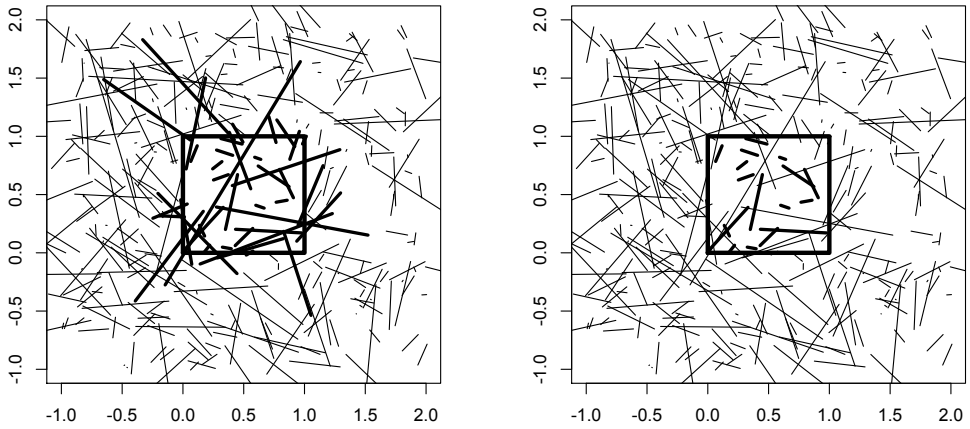
$$\begin{aligned}
\mathbb{E}\left[\sum_{\text{otos}} f(X_i)\right] &= \lambda\nu(W)\mathbb{E}^0\left[f(K_0)\right] \\
\Leftrightarrow \qquad \lambda\nu(W) &= \frac{\mathbb{E}\left[\sum_{\text{otos}} f(X_i)\right]}{\mathbb{E}^0\left[f(K_0)\right]} \\
&= \mathbb{E}\left[N_{\text{ref}}\right],
\end{aligned}$$

missä $N_{\text{ref}} = \sum i : c(X_i) \in W$, joten

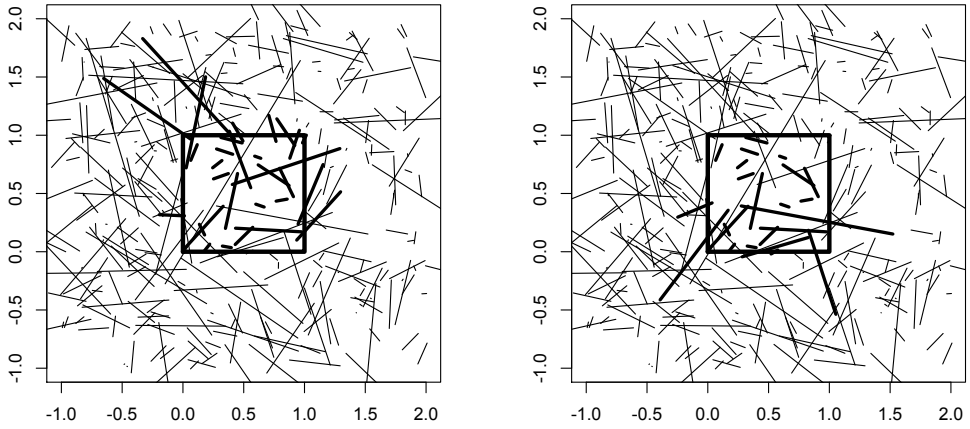
$$\hat{\lambda} = \frac{N_{\text{ref}}}{\nu(W)}$$

on harhaton estimaattori intensiteetille λ.

Viivasegmenttien tapauksessa on luontevaa valita referenssipisteeksi jompi kumpi segmentin päistä. Otantasääntö on havainnollistettu Kuvassa 3, jossa vasemmalla on valittu referenssipisteeksi viivasegmentin "eteläisin" pää (eli se pää, jonka y-koordinaatti on pienempi), oikealla puolestaan "pohjoisin" pää. Huomaa, että otokset eri referenssipisteitä käyttäen voivat olla hyvinkin erilaisia, vaikka alkuperäinen realisaatio on sama.



Kuva 2: Vasemmalla plusotanta ja oikealla samaan aineistoon suoritettu miinusotanta. Otokseen mukaan tulevat viivasegmentit näkyvät tummennettuina.



Kuva 3: Vasemmalla referenssipisteotanta, jossa referenssipisteenä käytetään viivasegmentin eteläisintä päätä. Oikealla puolestaan käytetään pohjoisinta päätä. Huomaa, että molemmissa kuvissa alkuperäinen realisaatio on sama kuin Kuvassa 2.

4 Intensiteettiestimaattoreiden vertailu simulointikokeilla

Tässä luvussa tutkitaan intensiteetin estimoinnin tarkkuutta eri otantatekniikoita käyttäen simulointikokeiden avulla. Edellisessä kappaleessa osoitettiin referenssipisteotannan harhattomuus sekä plus- ja miinusotantojen harhaisuus. Simulointien avulla nähdään harhan suuruus ja lisäksi se, miten ikkunan koon ja intensiteetin muuttuminen vaikuttaa estimaattoreiden variansseihin.

4.1 Tilastollinen malli

Tilastollisena mallina viivasegmenttiprosessille käytetään Boolen mallia, jossa siis segmenttien sijainnit tulevat stationaarisesta Poisson-prosessista ja lisäksi sekä segmentin pituus että suuntaus ovat riippumattomia toisistaan, samoin kuin segmentin sijainnista. Pituudet on simuloitu käyttäen eksponenttijakaumaa.

4.2 Simuloinnin toteutus

Aluksi määrätään prosessin intensiteetti λ sekä parametri μ, joka on havaintoikkunan sivun pituuden suhde viivasegmenttien keskipituuteen. Havaintoikkuna W on yksikköneliö $[0,1]^2$, mutta varsinaiseen simulointialueeseen täytyy kuitenkin lisätä marginaali m, joka on määrätty eksponenttijakauman $Exp(\mu)$ 0.95:n fraktiilina. Simuloinnit suoritetaan siis alueeseen $W^+ = [-m, 1+m]^2$.

Ensin simuloidaan germit, eli viivasegmenttien sijainnit, mikä toteutetaan käyttäen stationaarista Poisson-pisteprosessia. Tällöin pisteiden lukumäärä simuloidaan jakaumasta $Poisson(\lambda\nu(W^+))$, jonka jälkeen simuloidaan pisteille koordinaatit x_1 ja y_1 riippumattomasti tasajakaumasta $Tas[-m, 1+m]$. Seuraavaksi simuloidaan jokaiselle segmentille pituus l jakaumasta $Exp(\mu)$, jolloin pituuden odotusarvo on $\frac{1}{\mu}$. Segmentin kulma θ x-akseliin nähden tulee tasajakaumasta $Tas[0, 2\pi)$, jolloin loppupään koordinaatit ovat $x_2 = x_1 + l\cos\theta$ ja $y_2 = y_1 + l\sin\theta$.

Ikkunan koon säätäminen on simulointiohjelmassa toteutettu kuvaa skaalaamalla. Jos halutaan suurempi havaintoikkuna, niin lyhennetään segmenttejä ja lisätään intensiteettiä. Esim. jos halutaan kasvattaa ikkunan sivun pituutta kaksinkertaiseksi, niin lyhennetään viivasegmentin keskipituutta puoleen ja lisätään intensiteetti nelinkertaiseksi. Simuloimisessa käytettävä havaintoikkuna on siis koko ajan yksikköneliö, vain marginaalit muuttuvat.

Kun realisaatio on simuloitu halutulla intensiteetillä ja ikkunan koolla, valitaan otantatekniikka. Kun otanta on suoritettu, lasketaan estimaatti halutulle tunnukselle, tässä tapauksessa intensiteetille. Estimaattoreiden varianssien estimoimiseksi edellinen toistetaan useita kertoja (esim. 1000 kertaa). Saaduista estimaateista lasketaan otoskeskiarvo $\bar{x}$ ja otosvarianssi s^2.

4.3 Tuloksia

Seuraavassa on simuloitu 1000 viivasegmenttiprosessin realisaatiota intensiteeteillä $\lambda = 10$, $\lambda = 30$ ja $\lambda = 50$, sekä suoritettu intensiteetin estimointi eri otantatekniikoita käyttäen. Havaintoikkunan kokoa on säädetty siten, että ensimmäisessä vaiheessa

ikkunan sivun pituuden suhde viivasegmentin keskipituuteen on 1, toisessa vaiheessa 3 ja kolmannessa vaiheessa 5 (Taulukot 1, 2 ja 3).

Taulukko 1: Intensiteetin estimaattien keskiarvot ja otosvarianssit plusotannalla.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	22.14	21.85	66.53	64.52	111.58	114.74
Ikkuna=3	14.18	1.57	42.49	5.04	70.62	8.23
Ikkuna=5	12.46	0.49	37.46	1.44	62.39	2.35

Taulukko 2: Intensiteetin estimaattien keskiarvot ja otosvarianssit miinusotannalla.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	3.49	3.43	10.50	11.05	17.52	17.67
Ikkuna=3	6.50	0.71	19.54	2.13	32.45	4.17
Ikkuna=5	7.76	0.31	23.17	0.88	38.61	1.43

Taulukko 3: Intensiteetin estimaattien keskiarvot ja otosvarianssit referenssipisteotannalla.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	9.88	10.20	29.65	30.40	49.21	48.74
Ikkuna=3	10.02	1.09	29.83	3.28	49.64	4.95
Ikkuna=5	10.00	0.41	29.92	1.21	49.77	1.81

Taulukoista 1 ja 2 nähdään plus- ja miinusotantojen aiheuttama harha, joka kuitenkin pienenee odotetusti ikkunan koon kasvaessa. Samoin käy variansseille. Referenssipisteotanta, jossa käytetään viivasegmentin eteläisintä päätä, antaa odotusarvoisesti hyviä estimaatteja (Taulukko 3). Sen varianssi on kääntäen verrannollinen ikkunan pinta-alaan; esim. ikkunan sivun kasvaessa kolminkertaiseksi varianssi pienenee yhteen yhdeksäsosaan.

Jokaisen taulukon ylimmältä riviltä huomataan Boolen malliin liittyvä ominaisuus: estimaattorin otosvarianssi on hyvin lähellä sen keskiarvoa. Näin pitääkin olla, koska otokseen tulevien segmenttien lukumäärä noudattaa Poisson-jakaumaa, jolloin odotusarvo ja varianssi ovat samat.

Itse asiassa kaikki Taulukoiden 1–3 keskiarvot ja varianssit on mahdollista laskea teoreettisesti, mikäli viivasegmenttien pituusjakauma tunnetaan. Tähän palataan pituusjakauman estimoinnin jälkeen Luvussa 6.

4.3.1 Kahden referenssipisteen menetelmä

Kuten aikaisemmin osoitettiin, referenssipistettä käyttäen saadaan intensiteetille harhaton estimaattori. Mutta voisiko estimaattorin varianssia pienentää esim. käyttämällä eri referenssipisteitä ja yhdistämällä niiden antamat estimaatit? Toteutetaan seuraavaksi estimointi laskemalla jokaisesta realisaatiosta estimaatti sekä eteläisintä, että pohjoisinta päätä referenssipisteenä käyttäen. Näin saadaan kaksi eri estimaattia $\hat{\lambda}_e$ ja $\hat{\lambda}_p$, joiden keskiarvona saadaan uusi harhaton estimaattori $\hat{\lambda}_{ep}$ (Taulukko 4). Tämä on yleinen varianssin pienentämismenetelmä, jossa käytetään kahden samoin jakautuneen muuttujan keskiarvoa. Jos muuttujat ovat negatiivisesti korreloituneita, puhutaan *antiteettisen muuttujan* menetelmästä [12].

Taulukko 4: Intensiteetin estimaattien keskiarvot ja otosvarianssit käyttäen kahden eri referenssipisteotannan tuottamien estimaattien keskiarvoa.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	9.93	6.81	29.86	21.04	49.80	33.79
Ikkuna=3	9.97	0.92	29.92	2.77	49.80	4.87
Ikkuna=5	9.97	0.40	29.91	1.09	49.79	1.67

Taulukoita 3 ja 4 verrattaessa huomataan, että estimaattorin varianssi todellakin pienenee käytettäessä kahden eri estimaatin keskiarvoa. Näin käy erityisesti pienillä otantaikkunoilla ja intensiteeteillä. Kun ikkunan koko ja intensiteetti kasvavat, ero häviää.

Edellä esitellyn estimaattorin varianssi on

$$Var\left[\hat{\lambda}_{ep}\right] = Var\left[\frac{1}{2}\left(\hat{\lambda}_e + \hat{\lambda}_p\right)\right] = \frac{1}{4}\left[Var[\hat{\lambda}_e] + Var[\hat{\lambda}_p] + 2 \cdot Cov[\hat{\lambda}_e, \hat{\lambda}_p]\right].$$

Tässä ongelmana on kovarianssitermi, jonka laskeminen voi olla haastavaa. Näin ollen varianssin analyyttinen johtaminen sivuutetaan tämän työn osalta. Joka tapauksessa huomataan, että mikäli $Cov[\hat{\lambda}_e, \hat{\lambda}_p] < Var[\hat{\lambda}_e]$, niin estimaattorin $\hat{\lambda}_{ep}$ varianssi on pienempi kuin $\hat{\lambda}_e$:n. Simulointien perusteella näin mitä ilmeisimmin on.

Itse asiassa referenssipiste olisi mahdollista määrittää mielivaltaisen monella tavalla muuttamalla suuntaa, jonka perusteella referenssipiste valitaan. Näin saatujen eri otosten keskiarvona saatava estimaattori olisi teoriassa vieläkin tarkempi (tai vähintään yhtä tarkka) [6]. Käytännössä saavutettava hyöty on kuitenkin varsin vähäinen, kun referenssipisteitä on enemmän kuin kaksi.

Kahta jälkimmäistä taulukkoa katsellessa voi herättää epäilystä se, että kaikki intensiteetin estimaatit ovat yhtä poikkeusta lukuunottamatta alle oikean arvon. Tämä ei varmaankaan voi olla sattumaa. Syy tähän lienee siinä, että simuloinnit on toteutettu äärelliseen alueeseen. Otoksista jää tällöin puuttumaan muutama hyvin pitkä viivasegmentti, jotka yltäisivät ikkunaan, vaikka niiden alkupää olisi varsin kaukana ikkunan reunasta. Teoreettisesti ajatellen simulointialueen pitäisi olla ääretön, kun käytetään äärettömän kantajan omaavaa jakaumaa viivasegmenttien pituudelle.

5 Pituusjakauman estimointi

5.1 Epäparametrisia menetelmiä

Epäparametrisissa menetelmissä ei tehdä oletusta satunnaismuuttujan jakauman muodosta. Seuraavaksi esitellään Horvitz–Thompson-estimaattori, joka perustuu miinusotannan aiheuttaman harhan korjaamiseen havaintoja painottamalla, sekä Kaplan–Meier-estimaattori, joka ottaa huomioon sensuroidut havainnot ja antaa epäparametrisen estimaatin kertymäfunktiolle.

5.1.1 Horvitz–Thompson-estimaattori

Seuraava perustuu lähteeseen [4]. Klassisesta otantateoriasta tiedetään, että otannasta johtuvaa harhaa on mahdollista korjata havaintoja painottamalla [13]. Jos havaintojen X_i otantatodennäköisyydet π_i tiedetään, saadaan oikeat painot otantatodennäköisyyksien käänteislukuina. Tästä saadaan otostotaalin $Y = \sum_{\text{otos}} X_i$ *Horvitz–Thompson*-estimaattori

$$\hat{Y}_{HT} = \sum_{\text{otos}} \frac{X_i}{\pi_i},$$

joka on harhaton, koska

$$\begin{aligned} \mathbb{E}\left[\hat{Y}_{HT}\right] &= \mathbb{E}\left[\sum_{\text{perusjoukko}} \mathbf{1}_{\{\text{otos}\}} \frac{X_i}{\pi_i}\right] \\ &= \sum_{\text{perusjoukko}} \pi_i \frac{X_i}{\pi_i} = Y. \end{aligned}$$

Edellinen tulos edellyttää, että perusjoukko on äärellinen. Horvitz–Thompson-estimaattorin konstruointi spatiaaliselle otokselle on hankalampaa, koska se vaatii keskiarvon ottamista yli äärettömän populaation, jolloin summauksen ja odotusarvon järjestystä ei voi suoraan vaihtaa. Ratkaisu löytyy jälleen Campbell–Mecken lauseesta.

Olkoon Ξ stationaarinen germ–grain-malli, joka koostuu kompakteista graineista K_i jakaumanaan Q ja intensiteettinään λ. Oletetaan, että otos koostuu kaikista objekteista X_i, joille $I(X_i) = 1$. Merkitään viivasegmentin pituutta l:llä. Kuvaus $l : \mathbb{K}_0 \to \mathbb{R}_+$ on siirtoinvariantti, eli $l(K + x) = l(K)$. Plusotannan tapauksessa Campbell–Mecken lause antaa

$$\begin{aligned} \mathbb{E}\left[\sum_{X_i \cap W \neq \emptyset} l(X_i)\right] &= \mathbb{E}^0\left[\sum_i l(X_i)\mathbf{1}_{\{X_i \cap W \neq \emptyset\}}\right] \\ &= \lambda \mathbb{E}^0\left[\int_{\mathbb{R}^2} l(K_0 + x)\mathbf{1}_{\{(K_0+x)\cap W \neq \emptyset\}} dx\right] \\ &= \lambda \mathbb{E}^0\left[l(K_0)\int_{\mathbb{R}^2} \mathbf{1}_{\{(K_0+x)\cap W \neq \emptyset\}} dx\right] \\ &= \lambda \mathbb{E}^0\left[l(K_0)\nu(W \oplus \check{K}_0)\right]. \end{aligned} \tag{4}$$

Vastaavasti miinusotannalle saadaan

$$\mathbb{E}\left[\sum_{X_i \subset W} l(X_i)\right] = \lambda \mathbb{E}^0\left[l(K_0)\nu(W \ominus \check{K}_0)\right]. \tag{5}$$

Plus- ja miinusotanta tuottavat siis harhaisen estimaattorin viivasegmentin pituuden odotusarvolle otantaharhafaktoreiden ollessa $\nu(W \oplus \check{K}_0)$ ja $\nu(W \ominus \check{K}_0)$.

Painotetaan jokaista otokseen tulevaa havaintoa X_i otantaharhafaktorin $\pi(X_i)$ käänteisluvulla olettaen, että $\pi(X_i)$ on tunnettu ja melkein varmasti positiivinen tyypilliselle grainille. Tällöin tuloksia (4) ja (5) käyttäen saadaan

$$\begin{aligned} & \mathbb{E}\left[\sum_{\text{otos}} \frac{1}{\pi(X_i)}\right] &&= \lambda \\ \Leftrightarrow \qquad & \mathbb{E}\left[\sum_{\text{otos}} \frac{1}{\pi(X_i)} l(X_i)\right] &&= \lambda \mathbb{E}^0\left[l(K_0)\right], \end{aligned}$$

joten

$$\frac{\mathbb{E}\left[\sum_{\text{otos}} \frac{l(X_i)}{\pi(X_i)}\right]}{\mathbb{E}\left[\sum_{\text{otos}} \frac{1}{\pi(X_i)}\right]} = \mathbb{E}^0\left[l(K_0)\right].$$

Näin ollen Horvitz–Thompson-estimaattori

$$\frac{\sum_{\text{otos}} \frac{l(X_i)}{\pi(X_i)}}{\sum_{\text{otos}} \frac{1}{\pi(X_i)}}$$

on suhdeharhaton odotusarvolle $\mathbb{E}^0\left[l(K_0)\right]$.

Nyt esimerkiksi miinusotannan tuottama harha saadaan korjattua asettamalla jokaiselle havainnolle X_i paino $1/\nu(W \ominus X_i)$. Harhattoman estimaattorin konstruoimiseksi täytyy vaatia, että $\nu(W \ominus X_i) > 0$ melkein varmasti, eli grain ei saa olla liian suuri, vaan sen on mahduttava aidosti havaintoikkunan sisäpuolelle. Näin ollen viivasegmenttien pituutta ei voi simuloida jakaumasta, jolla on ääretön kantaja (kuten esim. eksponenttijakauma), vaan segmenttien pituudelle on asetettava aito yläraja, joka riippuu ikkunan koosta. Ilman tätä oletusta estimaattori on kuitenkin asymptoottisesti harhaton [7].

Mikäli em. ehto toteutuu ja havaintoikkuna W on suorakulmio, miinusotannan harhankorjaus on helppo tehdä. Jos ikkunan sivujen pituudet ovat a ja b, niin $\nu(W \ominus X_i) = (a-h)(b-v)$, missä h ja v ovat sivujen pituudet pienimmälle mahdolliselle suorakulmiolle, joka sisältää X_i:n.

5.1.2 Kaplan–Meier-estimaattori

Parametrittoman tavan viivasegmenttien pituusjakauman estimoimiseen tarjoaa elinaika-analyysissa yleisesti käytetty Kaplan–Meier-estimaattori. Seuraava perustuu lähteisiin [1, 5].

Oletetaan, että viivasegmenteillä on pituusjakauma, jonka kertymäfunktio on F_l. Olkoon $X_1, X_2, \ldots, X_n$ riippumaton satunnaisotos tästä jakaumasta. Nyt jokaiseen

havaintoon $l(X_i)$ liittyy sensurointiraja C_i, jolloin havaitsemme sensuroidut pituudet $l(\tilde{X}_i) = \min(l(X_i), C_i)$ sekä sensurointi-indikaattorin $D_i = \mathbf{1}_{\{l(X_i)<C_i\}}$.

Kun käytetään referenssipisteotantaa, vältytään otantaharhalta, ja lisäksi havainnot voivat olla vain toisesta päästä sensuroituja. Nyt kertymäfunktion F_l Kaplan–Meier-estimaattori on

$$\hat{F}_{KM}(t) = 1 - \prod_{s \le t} \left(1 - \frac{\sum_{i \ge 1} \mathbf{1}_{\{l(X_i \cap W) = s\}} \mathbf{1}_{\{X_i \subset W\}}}{\sum_{i \ge 1} \mathbf{1}_{\{l(X_i \cap W) \ge s\}}} \right). \tag{6}$$

Tämän estimaattorin avulla voidaan määrittää estimaatti mediaanille, mutta odotusarvon estimaattia ei yleisessä tapauksessa saada. Tämä johtuu siitä, että suurin havainto voi olla sensuroitu, jolloin estimoitu kertymäfunktio ei tavoita ykköstä. Tällöin täytyy tehdä jakaumaoletus. Jos $l \sim Exp(\mu)$, niin $md(l) = \mu^{-1} \log(2)$, jolloin $\hat{\mu} = \log(2)/\hat{md}(l)$.

5.2 Parametrisia menetelmiä

5.2.1 Suurimman uskottavuuden estimaattori

Kuten johdannossa todettiin, viivasegmentti voidaan havaita kokonaan tai se voi olla joko toisesta tai molemmista päistä sensuroitu. Kuitenkin käytettäessä referenssipisteotantaa, voi sensurointeja olla vain yksi. Tällöin havainto on oikealta sensuroitu. Seuraavassa esityksessä on käytetty lähdettä [10].

Oletetaan, että viivasegmenttien pituusjakaumalla on tiheysfunktio f ja kertymäfunktio F. Nämä riippuvat parametrista μ, joka täytyy estimoida. Olkoot $l_1, \ldots, l_m$ kokonaan havaittujen segmenttien pituudet ja $l_{m+1}, \ldots, l_n$ havaintoikkunan ulkopuolelle yltävien segmenttien sensuroidut pituudet. Tällöin koko aineistoon $\mathbf{l} = (l_1, \ldots, l_m, l_{m+1}, \ldots, l_n)$ liittyvä uskottavuusfunktio on

$$L(\mu; \mathbf{l}) = \prod_{i=1}^{m} f(l_i; \mu) \times \prod_{i=m+1}^{n} \{1 - F(l_i; \mu)\}.$$

Eksponenttijakauman tapauksessa saadaan

$$L(\mu; \mathbf{l}) = \prod_{i=i}^{m} \mu e^{-\mu l_i} \times \prod_{i=m+1}^{n} e^{-\mu l_i} = \mu^m e^{-\mu \sum_{i=1}^{n} l_i}.$$

Maksimoimalla uskottavuusfunktio μ:n suhteen saadaan suurimman uskottavuuden estimaattori

$$\hat{\mu}_{SU} = \frac{m}{\sum_{i=1}^{n} l_i}.$$

5.2.2 Sisältymissuhteeseen perustuva estimaattori

Tarkastellaan seuraavaksi, millä todennäköisyydellä havaintoikkunaa leikkaava viivasegmentti mahtuu kokonaisuudessaan ikkunaan, siis $\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset)$. Jotta segmentti X_i mahtuisi kokonaan ikkunaan W, täytyy sen alkupään (germin)

osua pienennettyyn ikkunaan $W \ominus X_i^0$, missä X_i^0 on viivasegmentti, jonka referenssipiste on origossa. Vastaavasti ehdolle "segmentti leikkaa ikkunaa" suotuisa alue on W:n laajennus $W \oplus \check{X}_i^0$.

Oletetaan, että viivasegmentin X_i pituus l ja suuntakulma θ on annettu. Yleisyyttä rikkomatta voidaan olettaa, että havaintoikkuna on yksikköneliö. Tällöin alueiden pinta-alat ovat (Kuva 5)

$$\nu(W \ominus X_i^0) = \mathbf{1}_{\{l|\sin\theta|\leq 1\}}\mathbf{1}_{\{l|\cos\theta|\leq 1\}}(1 - l|\cos\theta|)(1 - l|\sin\theta|)$$

ja

$$\nu(W \oplus \check{X}_i^0) = 1 + l|\cos\theta| + l|\sin\theta|.$$

Koska tapahtumat $X_i \subset W$ ja $X_i \cap W \neq \emptyset$ ovat sisäkkäisiä, saadaan geometrisena todennäköisyytenä

$$\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset; l; \theta) = \frac{\nu(W \ominus X_i^0)}{\nu(W \oplus \check{X}_i^0)}.$$

Oletetaan seuraavaksi, että $\theta \sim Tas[0, 2\pi)$ ja että pituudella l on tiheysfunktio $f_l(l)$. Lisäksi l ja θ ovat riippumattomia. Tällöin alue $W \ominus X_i^0 \in \mathbb{K}_0$ on suljettu satunnaisjoukko, jonka jakauma $Q(K)$ tunnetaan. Odotusarvo tämän satunnaisjoukon pinta-alalle on

$$\begin{aligned}
\mathbb{E}^0[\nu(W \ominus X_0)] &= \int_{\mathbb{K}_0} \nu(W \ominus X_0) dQ(K) \\
&= \int_0^\infty \int_0^{2\pi} f_l(l) f_\theta(\theta)(1 - l|\cos\theta|)^+(1 - l|\sin\theta|)^+ d\theta dl
\end{aligned}$$

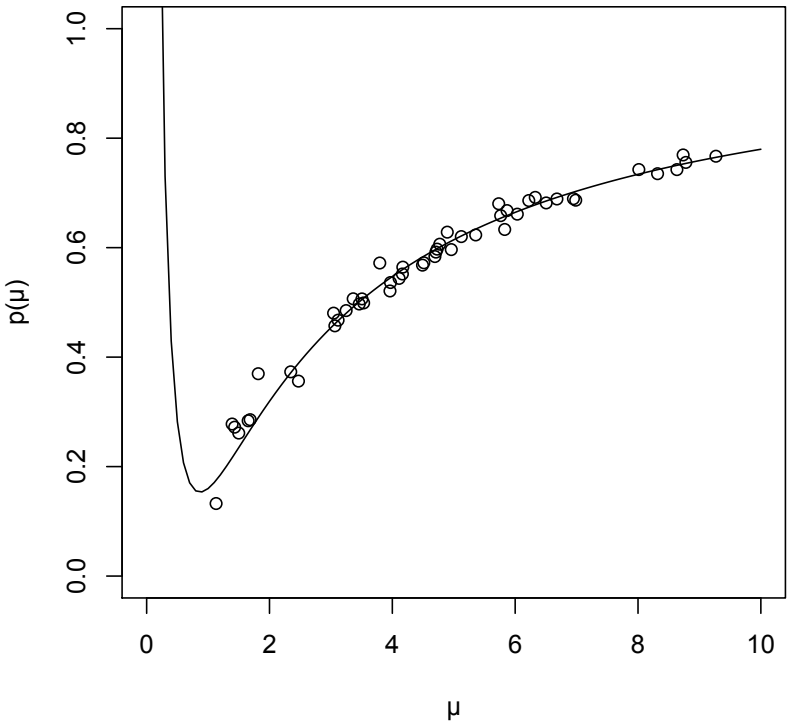
Tämä integraali ei ratkea suljetussa muodossa, joten tyydytään approksimaatioon, jossa ei oteta positiiviosia. Tämä on Monte Carlo -integrointikokeiden perusteella hyvin lähellä oikeata, kunhan l on riittävän[4] pieni:

$$\begin{aligned}
\mathbb{E}^0[\nu(W \ominus X_0)] &\approx \int_0^\infty \int_0^{2\pi} f_l(l) f_\theta(\theta)(1 - l|\cos\theta|)(1 - l|\sin\theta|) d\theta dl \\
&= \int_0^\infty f_l(l) \frac{4}{2\pi} \int_0^{\pi/2} 1 - l\sin\theta - l\cos\theta + l^2 \sin\theta\cos\theta d\theta dl \\
&= \int_0^\infty f_l(l) \frac{2}{\pi} \left(\frac{\pi}{2} - l - l + \frac{1}{2}l^2\right) dl \\
&= \int_0^\infty f_l(l) dl - \frac{4}{\pi}\int_0^\infty l f_l(l) dl + \frac{1}{\pi}\int_0^\infty l^2 f_l(l) dl \\
&= 1 - \frac{4}{\pi}\mathbb{E}[l] + \frac{1}{\pi}\mathbb{E}[l^2].
\end{aligned}$$

Laajennetun ikkunan pinta-alalle saadaan tarkka odotusarvo vastaavalla laskulla:

$$\begin{aligned}
\mathbb{E}^0[\nu(W \oplus \check{X}_0)] &= \int_0^\infty \int_0^{2\pi} f_l(l) f_\theta(\theta)(1 + l|\cos\theta| + l|\sin\theta|) d\theta dl \\
&= 1 + \frac{4}{\pi}\mathbb{E}[l].
\end{aligned}$$

[4] Jos $l \sim Exp(\mu)$, missä $\mu \geq 3$, virhe on $\lesssim 0.005$.



Kuva 4: Funktion $p(\mu)$ kuvaaja sekä hajontakuvio, jossa x-akselilla ikkunan sivun pituuden suhde viivasegmenttien keskipituuteen ja y-akselilla toteutunut kokonaan ikkunaan mahtuneiden segmenttien osuus kaikista ikkunaa leikkaavista segmenteistä. Simuloinnit on toteutettu käyttäen intensiteettiä 30.

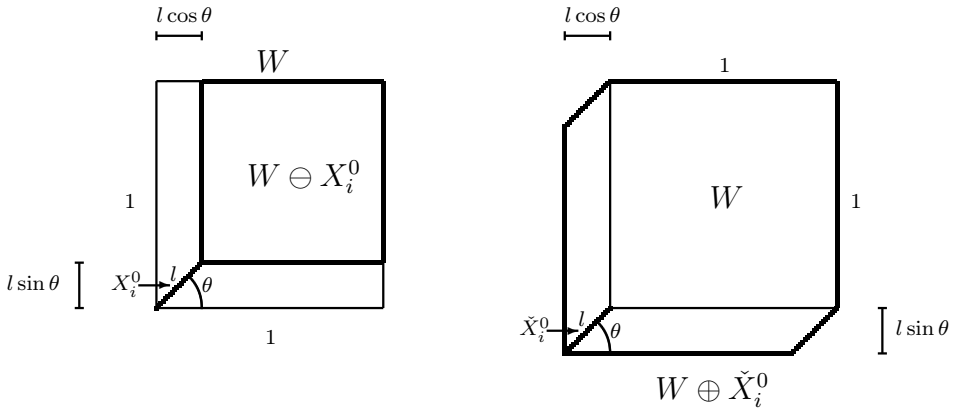
Jos $l \sim Exp(\mu)$, on $\mathbb{E}[l] = 1/\mu$ ja $\mathbb{E}[l^2] = 2/\mu^2$, joten

$$\mathbb{P}(X_i \subset W | X_i \cap W \neq \emptyset) = \frac{\mathbb{E}^0[\nu(W \ominus X_i^0)]}{\mathbb{E}^0[\nu(W \oplus \check{X}_i^0)]} \approx \frac{1 - \frac{4}{\pi\mu} + \frac{2}{\pi\mu^2}}{1 + \frac{4}{\pi\mu}} =: p(\mu). \qquad (7)$$

Parametri μ on nyt ikkunan sivun pituuden suhde viivasegmentin pituuden odotusarvoon. Kuvassa 4 on hajontakuviona eri kokoisilla ikkunansivuilla (μ:n arvoilla) simuloitujen realisaatioiden tuottamia sisältymissuhteita sekä funktion $p(\mu)$ kuvaaja. Kuvasta huomataan, että approksimaatio on varsin hyvä, kun $\mu \geq 1$.

Funktio p on jatkuva ja aidosti monotoninen, kun rajoitutaan tapauksiin, joissa $\mu \geq 1$ (siis oletetaan, että ikkunan sivun pituus on vähintään sama kuin viivasegmentin pituuden odotusarvo). Näin ollen p:llä on käänteisfunktio p^{-1}. Kun simuloidusta otoksesta lasketaan kokonaan ikkunaan mahtuneiden suhde kaikkien ikkunaa leikkaavien segmenttien lukumäärään, saadaan p:lle estimaatti $\hat{p}$. Kun $\hat{p}$ kuvataan p^{-1}:llä, saadaan puolestaan μ:lle estimaatti $\hat{\mu} = p^{-1}(\hat{p})$.

Itse asiassa viivasegmentin sisältymistodennäköisyys on sama asia kuin miinus- ja plusotantojen antamien intensiteettiestimaattien odotusarvojen suhde. Käytämme siis hyväksi kahta harhaista intensiteettiestimaattoria, joiden suhde antaa tietoa



Kuva 5: Vasemmalla vahvennettuna pienennetty ikkuna $W \ominus X_i^0$, oikealla laajennettu ikkuna $W \oplus \check{X}_i^0$.

segmenttien pituusjakaumasta. Luvussa 3.1.1 osoitettiin, että plusotannan tapauksessa yksikköneliötä leikkaavien segmenttien lukumäärän odotusarvo on

$$\mathbb{E}\left[\#\{i : X_i \cap W \neq \emptyset\}\right] = \lambda \mathbb{E}^0\left[\nu(W \oplus \check{K}_0)\right],$$

miinusotannan tapauksessa puolestaan

$$\mathbb{E}\left[\#\{i : X_i \subset W\}\right] = \lambda \mathbb{E}^0\left[\nu(W \ominus K_0)\right].$$

Lausekkeen (7) ensimmäinen yhtälö seuraa näistä.

5.3 Pituusjakauman estimaattoreiden vertailu simuloimalla

Toteutetaan simulointi Luvussa 4.2 kerrotulla tavalla ja vertaillaan kolmen eri estimaattorin antamia tuloksia. Taulukoissa 5, 6 ja 7 on SU-estimaattorin, sisältymissuhde-estimaattorin ja Kaplan–Meier-estimaattorin antamien estimaattien keskiarvot sekä otosvarianssit eri intensiteeteillä. Ensimmäisessä taulukossa ikkunan sivun pituuden suhde viivasegmenttien keskipituuteen on 3, toisessa taulukossa 5 ja kolmannessa taulukossa 10. Simuloitavien viivasegmenttien keskipituudeksi on asetettu 100 ja simulointikierroksia on 1000.

Taulukko 5: SU-estimaattorin, sisältymissuhde-estimaattorin sekä Kaplan–Meier-estimaattorin tuottamien estimaattien keskiarvot ja varianssit eri intensiteeteillä ikkunan sivun pituuden ollessa kolminkertainen viivasegmenttien keskipituuteen nähden.

	$\bar{x}_{SU}$	s^2_{SU}	$\bar{x}_{SS}$	s^2_{SS}	$\bar{x}_{KM}$	s^2_{KM}
$\lambda = 10$	99.67	168.74	103.04	196.39	121.32	597.43
$\lambda = 30$	100.21	64.46	102.22	61.72	121.31	193.18
$\lambda = 50$	99.43	32.41	102.70	38.72	119.38	108.85

Taulukko 6: SU-estimaattorin, sisältymissuhde-estimaattorin sekä Kaplan–Meier-estimaattorin tuottamien estimaattien keskiarvot ja varianssit eri intensiteeteillä ikkunan sivun pituuden ollessa viisinkertainen viivasegmenttien keskipituuteen nähden.

	$\bar{x}_{SU}$	s^2_{SU}	$\bar{x}_{SS}$	s^2_{SS}	$\bar{x}_{KM}$	s^2_{KM}
$\lambda = 10$	99.82	51.85	102.38	93.32	110.63	136.39
$\lambda = 30$	99.63	18.52	101.69	34.42	109.96	46.27
$\lambda = 50$	99.65	10.50	101.74	18.59	109.97	27.67

Taulukko 7: SU-estimaattorin, sisältymissuhde-estimaattorin sekä Kaplan–Meier-estimaattorin tuottamien estimaattien keskiarvot ja varianssit eri intensiteeteillä ikkunan sivun pituuden ollessa kymmenkertainen viivasegmenttien keskipituuteen nähden.

	$\bar{x}_{SU}$	s^2_{SU}	$\bar{x}_{SS}$	s^2_{SS}	$\bar{x}_{KM}$	s^2_{KM}
$\lambda = 10$	99.80	11.44	102.01	50.60	104.53	25.66
$\lambda = 30$	99.79	3.85	101.68	15.07	104.53	9.00
$\lambda = 50$	99.79	2.17	101.45	8.33	104.47	5.36

Simulointitulosten perusteella SU-estimaattori vaikuttaa harhattomalta ja näyttää tarkentuvan nopeasti intensiteetin tai ikkunan koon kasvaessa. Pieni virhe alaspäin johtunee jälleen äärellisestä simulointialueesta. Sisältymissuhde-estimaattori vaikuttaa varsin kilpailukykyiseltä SU-estimaattoriin verrattuna. Sillä näyttäisi olevan pientä harhaa, mutta vain 1–3 %. Varianssi on yleisesti ottaen hieman SU-estimaattorin varianssia suurempi.

Kaplan–Meier-estimaattori näyttää puolestaan antavan selkeästi ylöspäin harhaisia estimaatteja. Virhe on n. 20 % ikkunan sivun pituuden ollessa kolminkertainen ja n. 4.5 % ikkunan sivun pituuden ollessa kymmenkertainen viivasegmenttien keskipituuteen nähden. Varianssi on pienen ikkunan tapauksessa varsin suuri, mutta pienenee suhteellisen nopeasti ikkunan koon kasvaessa.

6 Uusi plusotannan harhasta korjattu intensiteettiestimaattori ja vertailu muihin estimaattoreihin

Palataan vielä intensiteettiestimaattoreihin. Edellisessä luvussa laskettiin odotusarvot plus- ja miinusotantojen tuottamille harhoille intensiteetin estimoinnissa. Näin ollen nämä harhat on mahdollista korjata, mikäli viivasegmenttien pituusjakauma tunnetaan.

Tämän tutkielman oletuksilla siis plusotanta tuottaa viivasegmenttiprosessin intensiteetin estimaattoriin harhakertoimen $1+\frac{4}{\pi\mu}$. Tämä kerroin riippuu parametrista μ, joka on havaintoikkunan sivun pituuden suhde viivasegmenttien keskipituuteen. Miinusotannan tuottama harha puolestaan on asymptoottisesti $1-\frac{4}{\pi\mu}+\frac{2}{\pi\mu^2}$.

Kun plusotannan tuottama estimaatti jaetaan kertoimella $1+\frac{4}{\pi\mu}$, saadaan harhaton estimaattori $\hat{\lambda}_{\text{plus}}$, joka käyttää kaiken mahdollisen havaintoikkunassa olevan informaation hyväkseen. Lasketaan pluskorjatun intensiteettiestimaattorin varianssi.

Otoskoko $\#i : X_i \cap W \neq \emptyset =: N_{\text{plus}}$ noudattaa Poisson-jakaumaa parametrilla $\lambda\nu(W)\left(1+\frac{4}{\pi\mu}\right)$. Näin ollen plusotantaan perustuvan korjatun estimaattorin varianssi on

$$\begin{aligned}
Var\left[\hat{\lambda}_{\text{plus}}\right] &= Var\left[\left(\frac{1}{1+\frac{4}{\pi\mu}}\right)\left(\frac{1}{\nu(W)}\right)N_{\text{plus}}\right] \\
&= \left(\frac{1}{\left(1+\frac{4}{\pi\mu}\right)\nu(W)}\right)^2 Var\left[N_{\text{plus}}\right] \\
&= \left(\frac{1}{\left(1+\frac{4}{\pi\mu}\right)\nu(W)}\right)^2 \lambda\nu(W)\left(1+\frac{4}{\pi\mu}\right) \\
&= \frac{\lambda}{\nu(W)\left(1+\frac{4}{\pi\mu}\right)}.
\end{aligned}$$

Tässä oletettiin, että parametri μ on tunnettu. Vastaavasti esim. referenssipisteotannan varianssi on $\frac{\lambda}{\nu(W)}$, joten korjattuun plusotanta perustuva estimaattori on tarkempi. Taulukkoon 8 on koottu teoreettiset odotusarvot ja varianssit neljälle eri intensiteettiestimaattorille. Miinusotannan kohdalla täytyy huomioida, että tulokset ovat asymptoottisia, eli pätevät μ:n lähestyessä ääretöntä.

Taulukko 8: Intensiteettiestimaattoreiden teoreettiset odotusarvot ja varianssit.

Estimaattori	Plus	Miinus	Ref.piste	Korjattu plus
$\mathbb{E}$	$\lambda\left(1+\frac{4}{\pi\mu}\right)$	$\lambda\left(1-\frac{4}{\pi\mu}+\frac{2}{\pi\mu^2}\right)$	λ	λ
Var	$\frac{\lambda\left(1+\frac{4}{\pi\mu}\right)}{\nu(W)}$	$\frac{\lambda\left(1-\frac{4}{\pi\mu}+\frac{2}{\pi\mu^2}\right)}{\nu(W)}$	$\frac{\lambda}{\nu(W)}$	$\frac{\lambda}{\nu(W)\left(1+\frac{4}{\pi\mu}\right)}$

Nyt voidaan verrata Luvussa 4 tehtyjä simulointeja (sivu 11) Taulukon 8 teoreettisiin arvoihin sijoittamalla parametrien μ ja λ paikalle simuloinneissa käytetyt arvot 1, 3 ja 5 sekä 10, 30 ja 50. Huomataan, että simulointitulokset sopivat yhteen teoreettisten tulosten kanssa.

6.1 Simulointikokeet pluskorjatuille estimaattoreille

Tarkistetaan vielä simuloimalla, että korjattuun plusotantaan perustuvan estimaattorin ominaisuudet ovat sitä mitä pitääkin (Taulukko 9). Sen jälkeen katsotaan, miten parametrin μ estimoiminen vaikuttaa estimaattoriin (Taulukko 10). Toteutetaan simuloinnit jälleen Luvussa 4.2 kerrotulla tavalla.

6.1.1 Pluskorjattu intensiteettiestimaattori, pituusjakauma tunnettu

Oletetaan ensin, että viivasegmenttien pituusjakauman odotusarvo on tunnettu ja estimoidaan intensiteetti korjatulla plusotannalla.

Taulukko 9: Intensiteetin estimointi korjattuun plusotantaan perustuen kun parametri μ oletetaan tunnetuksi.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	9.79	4.36	29.42	12.59	49.18	23.04
Ikkuna=3	9.89	0.75	29.68	2.21	49.54	3.86
Ikkuna=5	9.92	0.32	29.83	1.03	49.70	1.41

Taulukko 9 vahvistaa teoreettiset tulokset, eli harha korjaantuu ja varianssi on selvästi pienempi kuin esim. referenssipisteotannalla.

6.1.2 Pluskorjattu intensiteettiestimaattori, pituusjakauma tuntematon

Mikäli parametria μ ei tunneta, se voidaan estimoida esim. edellisen luvun menetelmillä. Tämä toki lisää epävarmuutta ja kasvattaa intensiteettiestimaattorin varianssia jonkin verran. Suoritetaan seuraavaksi intensiteetin estimointi kahdessa vaiheessa: estimoidaan ensin μ sisältymissuhdemenetelmällä ja sen jälkeen intensiteetti λ korjatulla plusotannalla (Taulukko 10).

Taulukko 10: Intensiteetin estimointi korjatulla plusotannalla kun parametri μ on estimoitu sisältymissuhdemenetelmällä.

	$\lambda = 10$		$\lambda = 30$		$\lambda = 50$	
	$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
Ikkuna=1	9.37	6.18	29.00	22.01	48.79	47.17
Ikkuna=3	9.86	0.91	29.51	2.56	49.33	4.67
Ikkuna=5	9.90	0.36	29.71	1.06	49.59	1.77

Kun parametri μ jouduttiin estimoimaan, kävi juuri niin kuin arveltiinkin, eli estimaattorin varianssi kasvoi hieman. Epävarmuus lisääntyy lähinnä pienen ikkunan kohdalla. Kun ikkunan koko kasvaa, ei variansseissa ole enää merkittävää eroa.

Joka tapauksessa tämäkin estimaattori on tarkempi kuin esim. referenssipisteotanta (vrt. Taulukko 3 sivulla 11). Simulointien perusteella sen varianssi näyttäisi olevan samaa luokkaa kuin kahden referenssipisteen menetelmällä (Taulukko 4 sivulla 12). Todetaan jälleen, että kaikissa keskiarvoissa näkyvä pieni vaje ei anna aihetta huoleen, koska se johtuu äärellisestä simulointialueesta.

7 Yhteenveto

Tässä tutkielmassa esiteltiin tason viivasegmenttiprosessi ja tutkittiin eri estimointimenetelmiä prosessin intensiteetille sekä viivasegmenttien pituusjakaumalle. Oletuksena oli hyvin yksinkertainen Boolen malli, jossa objektit ovat riippumattomia toisistaan ja niiden sijainti on täysin satunnainen. Lisäksi oletettiin segmenttien pituuden noudattavan eksponenttijakaumaa ja suuntauksen tasajakaumaa.

Näillä oletuksilla osoitettiin Campbell–Mecken lausetta käyttäen plus- ja miinusotantojen harhaisuus sekä referenssipisteotannan harhattomuus intensiteetin estimoinnissa. Uutena asiana esiteltiin kahden eri referenssipisteen käyttö, joka pienentää estimaattorin varianssia erityisesti silloin, jos havaintoikkunan koko suhteessa viivasegmenttien keskipituuteen on pieni. Lisäksi laskettiin plusotannan tuottama harhakerroin, jonka käänteisluvulla painottaen saatiin kaikista tarkin harhaton estimaattori intensiteetille. Tämä edellytti tosin viivasegmenttien pituusjakauman tuntemista. Toisaalta, vaikka pituusjakauman odotusarvo jouduttiin estimoimaan, oli korjattuun plusotantaan perustuva estimaattori edelleen tarkempi verrattuna referenssipisteotantaan.

Intensiteettiestimaattoreiden ominaisuuksia tutkittiin aluksi simulointikokeilla, mutta myöhemmin opittiin laskemaan myös teoreettiset odotusarvot ja varianssit, olettaen viivasegmenttien pituusjakauma tunnetuksi. Varianssi kahden referenssipisteen menetelmälle jäi tosin vielä laskematta, mutta todennäköisesti sekin olisi tämän tutkielman oletuksilla laskettavissa.

Pituusjakauman odotusarvon estimoimiseksi esiteltiin neljä eri estimaattoria, joista kolmen ominaisuuksia arvioitiin simuloinneilla. Havaintojen painottamiseen perustuvaa Horvitz–Thompson-estimaattoria ei otettu simulointeihin mukaan, koska plusotannan tapauksessa se vaatisi havaintoikkunan ulkopuolista tietoa ja miinusotannan tapauksessa harhattomuus edellyttäisi rajoitettua pituusjakaumaa.

Simulointitulosten perusteella heikoimmaksi osoittautui Kaplan–Meier-estimaattori, jonka antamat estimaatit olivat liian suuria erityisesti pienen havaintoikkunan tapauksessa. Parempia estimaatteja pituuden odotusarvolle saatiin suurimman uskottavuuden menetelmällä sekä uudella sisältymissuhde-estimaattorilla, joka perustuu plus- ja miinusotantojen tuottamien harhaisten intensiteettiestimaattien suhteeseen. Simulointien perusteella nämä menetelmät olivat kutakuinkin yhtä hyviä havaintoikkunan ollessa pieni. Ikkunan koon kasvaessa SU-estimaattori tarkentui hieman sisältymissuhde-estimaattoria nopeammin. Sisältymissuhdemenetelmä näytti tuottavan muutaman prosentin suuruisen harhan, jonka syy on toistaiseksi epäselvä.

Pituusjakauman estimaattoreiden vertailut tehtiin pelkästään simulointien perusteella, eli teoreettisia variansseja ei laskettu. Uskoakseni ainakin SU- ja sisältymissuhde-estimaattoreille on mahdollista laskea (asymptoottiset) varianssit. Tällöin olisi mahdollista laskea varianssi myös pluskorjatulle intensiteettiestimaattorille, jossa pituusjakauman odotusarvo joudutaan estimoimaan. Näihin palataan toivottavasti myöhemmin.

Esitän kiitokset professori Antti Penttiselle, yliopistonlehtori Dario Gasbarralle sekä lehtori Harri Högmanderille tähän tutkielmaan liittyvistä vihjeistä. Erityiskiitos menee tutkijatohtori Salme Kärkkäiselle erinomaisesta tutkimusaiheesta sekä äärimmäisen intensiivisestä ja motivoivasta ohjaamisesta.

Viitteet

[1] Laslett, G. M. (1982). The survival curve under monotone density constraints with applications of two-dimensional line segment process. *Biometrika*, 69, 153–160.

[2] Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* Wiley, Chichester.

[3] Miles, R. E. (1974). On the elimination of edge effects in planar sampling. In *Stochastic Geometry*, E. F. Harding, D. G. Kendall (eds.), 228–247. Wiley, London.

[4] Baddeley, A. J. (1999). Spatial sampling and censoring. In *Stochastic Geometry: Likelihood and Computation,* O. E. Barndorf-Nielsen, W. S. Kendall, M. N. M. van Lieshout (eds.), 1–78. Chapman and Hall, London.

[5] Pawlas, Z. (2006). Estimation of the distribution in germ-grain models. In *Proceedings of Prague Stochastics 2006*, 1–10. Prague.

[6] Molchanov, I. (1997). *Statistics of the Boolean Model for Practitioners and Mathematicians.* Wiley, Chichester.

[7] Stoyan, D., Kendall, W. S., Mecke, J. (1995). *Stochastic Geometry and its Applications.* 2nd edition, Wiley, Chichester.

[8] Stoyan, D. (1998). Random sets: models and statistics. *International Statistical Review*, 66, 1, 1–27.

[9] Molchanov, I. (2005). *Theory of Random Sets.* Springer, London.

[10] Davison, A. C. (2009). *Statistical Models.* Cambridge University Press, Cambridge.

[11] Taylor, J. C. (1997). *An Introduction to Measure and Probability.* Springer, London.

[12] Ripley, B. D. (1987). *Stochastic Simulation.* Wiley, Chichester.

[13] Thompson, S. K. (1992). *Sampling.* Wiley, Chichester.

[14] R Developement Core Team (2009). *R: Language and Enviroment for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Liitteet

A R-funktioita

```
#############################################################################################
#                              Simuloinnin pääohjelma                                       #
#############################################################################################
#Simuloi viivasegmenttiprosessin realisaation intensiteetillä lambda
#Parametri ikkunanSivu = ikkunan sivun pituus / viivasegmentin keskipituus
#Parametri theta = segmentin keskipituus
#Palauttaa datamatriisin sekä tiedon käytetystä intensiteetistä ja ikkunan sivun pituudesta
sim <- function(lambda=5, theta=1, ikkunanSivu=3) {
marginaali <- qexp(p=0.95, rate=ikkunanSivu) #Kuinka paljon ikkunan ulkopuolelle simuloidaan
alue <- (1+marginaali*2)^2*ikkunanSivu^2 #Simulointialueen pinta-ala
n <- rpois(1,lambda*alue) #pisteiden(segmenttien) lkm Poisson -jakaumasta
A <- matrix(nrow=n,ncol=9)

A[,9] <- c(rep(0,n)) #Otantaindikaattori
for(i in 1:n) {
   #Arvotaan alkupiste tasajakaumasta [-marginaali,(1+marginaali)]^2
   #ja kulma tasajakaumasta(0,2pi) sekä segmentin pituus exponenttijakaumasta
   A[i,1:4] <- c(runif(2,-marginaali,(1+marginaali)), runif(1,0,2*pi), rexp(1,1)/ikkunanSivu)
   #Lasketaan loppupiste
   A[i,5] <- A[i,1]+A[i,4]*cos(A[i,3])
   A[i,6] <- A[i,2]+A[i,4]*sin(A[i,3])
   #Lasketaan sensuroitu pituus
   A[i,7] <- laskeSensuroituPituus(A[i,1],A[i,2],A[i,5],A[i,6])[1]
   #Sensurointien lukumäärä, 0,1 tai 2, NA jos viiva ei leikkaa ikkunaa ollenkaan
   A[i,8] <- laskeSensuroituPituus(A[i,1],A[i,2],A[i,5],A[i,6])[2]
   }

return(list(A,lambda,ikkunanSivu,theta))
}

#############################################################################################
#                            Sensuroidun pituuden laskeminen                                #
#############################################################################################
#Laskee päätepisteiden (x1,y1) ja (x2,y2) määräämän viivasegmentin sensuroidun
#pituuden yksikköneliössä
#Lisäksi palautetaan tieto sensurointien lukumäärästä (0, 1 tai 2)
laskeSensuroituPituus <- function(x1,y1,x2,y2){
     tulos <- NULL
     pisteet <- matrix(nrow=2, ncol=6)
     pisteet[,1] <- c(x1,y1)
     pisteet[,2] <- c(x2,y2)
     pisteet[,3] <- c((-(suora(x1,y1,x2,y2)[1])/(suora(x1,y1,x2,y2)[2])),0)
     pisteet[,4] <- c((1-(suora(x1,y1,x2,y2)[1]))/(suora(x1,y1,x2,y2)[2]),1)
     pisteet[,5] <- c(0,suora(x1,y1,x2,y2)[1])
     pisteet[,6] <- c(1,suora(x1,y1,x2,y2)[1]+suora(x1,y1,x2,y2)[2])
     pisteet

     A <- matrix(nrow=2,ncol=2)
     j <- 1
     i <- 1
     s <- 0
     while(j<3 & i<7){
        x <- pisteet[1,i]
        y <- pisteet[2,i]
        if(x >= 0 & x <= 1 & y >= 0 & y <= 1 & x >= min(x1,x2) & x <= max(x1,x2) &
        y >= min(y1,y2) & y <= max(y1,y2)) {
           A[j,] <- c(x,y)
           j <- j+1
           if(i>2) s <- s+1
           }
        i <- i+1
        }
     if(i==7 & j<2) s <- NULL
     A
```

```
     dist(A)
     tulos <- c(dist(A),s)
     tulos
}

#laskee pisteiden (x1,y1) ja (x2,y2) kautta kulkevan
#suoran y = a + bx parametrit a ja b
suora <- function(x1,y1,x2,y2) {
     b <- (y2-y1)/(x2-x1)
     a <- y1 - b*x1
     c(a,b)
     }
#############################################################################################
#                              Otanta                                                       #
#############################################################################################
#Suorittaa otannan annetusta viivasegmenttisysteemin realisaatiosta halutulla
#otantamenetelmällä ja palauttaa datamatriisin
otanta <- function(data,rule="plus",piirto=FALSE,kierros=1){
A <- data[[1]]
lambda <- data[[2]]
ikkuna <- data[[3]]
theta <- data[[4]]
n <- length(A[,1])
for(i in 1:n){
if(rule=="plus"){
     if(!is.na(A[i,7])) A[i,9] <- 1
     }
   if(rule=="miinus"){
     if(!is.na(A[i,8])){
      if(A[i,8]==0) A[i,9] <- 1
      }
     }
   if(A[i,3] < pi) {
      etel <- c(A[i,1],A[i,2])
      pohj <- c(A[i,5],A[i,6])
      }
   else {
      pohj <- c(A[i,1],A[i,2])
      etel <- c(A[i,5],A[i,6])
      }

   if(rule=="rpetel"){
      if(!is.na(A[i,8])){
         if(min(etel) >= 0 & max(etel) <= 1) A[i,9] <- 1
         }
      }

    if(rule=="rppohj"){
      if(!is.na(A[i,8])){
         if(min(pohj) >= 0 & max(pohj) <= 1) A[i,9] <- 1
         }
     }

    }
    #Piirto
if(piirto==TRUE){
plot(NULL, xlim=c(-1,2), ylim=c(-1,2))
legend(-1,2,lambda)
legend(-1,1.6,ikkuna)
legend(1.5,2,kierros)
for(i in 1:n) {
    x <- c(A[i,1],A[i,5])
    y <- c(A[i,2],A[i,6])
    if(A[i,9]==1){lines(x,y,lwd=3)}
    else lines(x,y, lwd=1)
    }
x <- c(0,1,1,0)
y <- c(0,0,1,1)
polygon(x,y,lwd=3)
legend(-1,-0.75,int.est)
}
```

```
    data[[1]] <- A[A[,9]==1,]
    return(data)
}

##############################################################################################
#                                SU-estimaattori                                             #
##############################################################################################
#Estimoi viivasegmenttien keskipituuden käyttäen suurimman uskottavuuden menetelmää
estSU <- function(data){
ikkunanSivu <- data[[3]]
theta <- data[[4]]
data <- data[[1]]
estimaatti <- (sum(data[,7]))/(sum(I(data[,8]==0)))*ikkunanSivu*theta
return(estimaatti)
}
##############################################################################################
#                            Sisältymissuhde-estimaattori                                    #
##############################################################################################
#Laskee viivasegmenttien sisältymissuhteen ja estimoi pituuden odotusarvon käyttäen
#sisältymistodennäköisyyden approksimaatiota
estSS <- function(otos) {
ikkunanSivu <- otos[[3]]
theta <- otos[[4]] #pituuden odotusarvo
otos <- otos[[1]]
a1 <- length(which(otos[,8]==0))
a2 <- length(which(otos[,8]==1))
a3 <- length(which(otos[,8]==2))
osuus <- a1/(sum(c(a1,a2,a3))) #kokonaan ikkunaan mahtuneiden osuus otoksessa

p1 <- function(x) abs((1 - (4*(1/x))/pi + (2/x^2)/pi)/(1 + (4/pi)*(1/x)) - osuus)
#juuri <- uniroot(p1,c(0.87,30))[[1]] #Haetaan väliltä [0.87,30]
juuri <- optimize(p1,c(0.87,30))[[1]] #Haetaan väliltä [0.87,30]
estimaatti <- juuri*theta/ikkunanSivu
return(estimaatti)
}

#Laskee viivasegmenttien sisältymissuhteen ja estimoi pituuden odotusarvon käyttäen
#Monte-Carlo -integrointia
estMC <- function(otos) {
ikkunanSivu <- otos[[3]]
theta <- otos[[4]] #pituuden odotusarvo
otos <- otos[[1]]
a1 <- length(which(otos[,8]==0))
a2 <- length(which(otos[,8]==1))
a3 <- length(which(otos[,8]==2))
osuus <- a1/(sum(c(a1,a2,a3))) #kokonaan ikkunaan mahtuneiden osuus otoksessa

g <- function(mu) abs(osuus - mc(1000,mu)/(1+4/(pi*mu))) #minimoitava funktio
estimaatti <- optimize(g,lower=0.5,upper=30)[[1]] / ikkunanSivu * theta
return(estimaatti)
}

##############################################################################################
#                              Kaplan-Meier-estimaattori                                     #
##############################################################################################
#Estimoi välttöfunktion oikealta sensuroidulle datalle ja palauttaa eksponenttijakauman
#mukaisen estimaatin keskipituudelle
library(survival)
estKM <- function(otos) {
ikkunanSivu <- otos[[3]]
theta <- otos[[4]]
otos <- otos[[1]]
otos[,4] <- ikkunanSivu*theta*otos[,4] #skaalauksen poisto
otos <- cbind(otos[,4],otos[,8])
otos[,2] <- abs(otos[,2]-1) #muutetaan ykköset nolliksi ja päinvastoin
taulu <- data.frame("Pituus"=c(otos[,1]), "Delta"=c(otos[,2]))
km <- survfit(Surv(Pituus,Delta)~1, data=taulu)
#etsitään mediaani
md <- NULL
S <- 1
```

```
k <- 1
while(S > 0.5){
S <- km$surv[k]
md <- km$time[k]
k <- k+1
}
#keskiarvoestimaatti
est <- md/(log(2))
return(est)
}

#############################################################################################
#                            Korjattu plusotanta                                            #
#############################################################################################
#Estimoi viivsegmenttiprosessin intensiteetin käyttäen korjauskerrointa plusotannan
#aiheuttaman harhan korjaamiseksi
#Ikkunan sivun pituuden suhde segmenttien keskipituuteen oletetaan tunnetuksi
intEstPlus <- function(otos) {
mu <- otos[[3]] #ikkunan sivun pituuden suhde viivasegmenttien keskipituuteen
intEst <- length(otos[[1]][,1])/(mu^2) / (1+4/(pi*mu))
return(intEst)
}

#Estimoi parametrin mu(ikkunan sivun pituuden suhde segmenttien keskipituuteen)
#sisältymissuhdemenetelmällä ja tämän jälkeen intensiteetin korjattua plusotantaa käyttäen
estLambdaMuSS <- function(otos) {
ikkunanSivu <- otos[[3]]
muHat <- ikkunanSivu/estSS(otos)
intEst <- length(otos[[1]][,1])/(ikkunanSivu^2) / (1+4/(pi*muHat))
return(intEst)
}
```